PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Les réseaux de neurones artificiels pour analyser et simuler l'acquisition du langage chez les enfants

Artificial neural networks to analyze and simulate language acquisition in children

Soutenue par

Marvin LAVECHIN

Le 11 Septembre 2023

École doctorale n°158

Cerveau, cognition, comportement

Spécialité

Sciences cognitives

Composition du jury :

Caroline ROWLAND Research director, Max Planck Institute		*Rapporteure*
Thomas HUEBER Directeur de recherche, GIPSA, CNRS		*Rapporteur*
Chloé CLAVEL Professeure, Télécom Paris		*Examinatrice*
Okko RÄSÄNEN Associate professor, Tampere University		*Examinateur*
Afra ALISHAHI Associate professor, Tilburg University		*Examinatrice*
Alejandrina CRISTIA Directrice de recherche, LSCP, CNRS		*Directrice de thèse*
Emmanuel DUPOUX Directeur d'études, ENS, PSL, EHESS		*Co-encadrant*
Hervé BREDIN Chargé de recherche, IRIT, CNRS		*Invité*

ENS | PSL

Meta

Abstract

Lightweight child-worn recorders that collect audio across an entire day allow for a big-data approach to the study of language development. By collecting the child's production and linguistic environment, these recordings offer us a uniquely naturalistic view of everyday language uses. However, such recordings quickly accumulate thousands of hours of audio and require the use of automatic speech processing algorithms. Besides providing ecologically-valid measures of what children hear and say, these recordings can fuel computational models of early language acquisition with what infants truly hear. This opens up new opportunities for running realistic language learning simulations.

A first aspect of my doctoral work is dedicated to developing automatic speech processing algorithms for child-centered long-form recordings. In this manuscript, I first show that current state-of-the-art automatic speech recognition systems fail to capture the complexity of naturalistic speech as found in long-forms. I then present our attempt to propose a free, open-source, and more accurate alternative to the LENA® proprietary software, which is currently the standard tool for obtaining automatic analyses of long-forms. Using supervised learning methods, my collaborators and I built a suite of speech processing tools to detect voice activity, identify voice signal sources (child vocalizations, female or male speech), count the number of linguistic units (phonemes, syllables, or words), and estimate the quantity of background noise and reverberation.

A second aspect of my doctoral work is dedicated to computational models of early language acquisition. I present a first modeling study showing that self-supervised learning algorithms trained on audiobooks can learn phonetic and lexical aspects of their training language. I then show that the same algorithm trained on ecological long-forms needs inductive biases to learn phonetic aspects of its training language reliably and reflect on whether similar inductive biases may guide language learning in infants. Interestingly, there is no evidence for lexical learning on long-forms, contrary to what has been shown in the literature on more curated data. This series of studies illustrates the importance of considering ecologically-valid input data when modeling language acquisition.

Keywords: language development, psycholinguistics, speech processing, deep learning, supervised learning, self-supervised learning, cognitive sciences

Résumé

L'utilisation d'enregistreurs légers portés par les enfants et collectant du son tout au long de la journée ouvre la voie à une approche de 'données massives' pour étudier le développement du langage chez l'enfant. En recueillant la production langagière de l'enfant ainsi que son environnement linguistique, ces enregistrements nous offrent une vision réaliste des usages quotidiens du langage. Cependant, de tels enregistrements constituent rapidement des milliers d'heures d'audio et nécessitent l'utilisation d'outils de traitement automatique de la parole. En plus de fournir des mesures réalistes de ce que les enfants entendent et disent, ces enregistrements peuvent alimenter les modèles computationnels d'acquisition du langage avec une entrée comparable à ce que les enfants entendent réellement, ouvrant ainsi de nouvelles perspectives pour simuler l'apprentissage du langage.

Un premier aspect de mon travail doctoral concerne le développement d'outils de traitement automatique de la parole compatibles avec ces enregistreurs portés par l'enfant. Cette thèse commence par une étude montrant que les outils à la pointe de la reconnaissance automatique de la parole ne parviennent pas à transcrire la parole enregistrée dans des conditions bruitées et non contrôlées. À travers une brève analyse technique et scientifique, j'introduis le logiciel propriétaire LENA, devenu l'outil standard pour l'analyse automatique de ces enregistrements. Je présente nos efforts pour en proposer une version libre, gratuite et plus performante. En collaboration avec d'autres chercheurs, j'ai contribué à développer une série d'outils de traitement automatique de la parole pour détecter l'activité vocale, identifier les sources de signaux vocaux (vocalisations de l'enfant, paroles d'une femme ou d'un homme adulte), compter le nombre d'unités linguistiques (phonèmes, syllabes, mots), et estimer la quantité de bruit et de réverbération.

Un second aspect de mon travail doctoral concerne la modélisation de l'acquisition du langage. Je présente une première étude montrant qu'un algorithme d'apprentissage auto-supervisé entraîné sur des livres audio est capable d'apprendre des aspects phonétiques et lexicaux de sa langue d'entraînement. En revanche, lorsque ce même algorithme est exposé à ce que les enfants entendent réellement, l'implémentation de biais inductifs qui visent à contraindre l'apprentissage est nécessaire pour observer une acquisition de ces mêmes aspects phonétiques. À partir de ce constat, nous réfléchissons à la possibilité que de tels biais inductifs puissent guider l'apprentissage chez les enfants. Il est surprenant de constater que notre algorithme est incapable d'apprendre les aspects lexicaux de sa langue d'entraînement lorsqu'il est exposé à la parole que reçoivent les enfants, contrairement à ce qu'a montré la littérature sur

des données moins bruitées. Cette série d'études illustre l'importance d'utiliser des données d'entrée réalistes lors de la modélisation de l'acquisition du langage.

Mots-clés: développement du langage, psycholinguistique, traitement de la parole, apprentissage profond, apprentissage supervisé, apprentissage auto-supervisé, sciences cognitives

Résumé substantiel

Le langage est le propre de l'humanité. Tandis que l'on trouve des systèmes de communication complexes chez de nombreuses espèces, les langues humaines présentent une complexité et une capacité d'expression inégalées. En combinant un ensemble fini de sons en un ensemble infini de mots, qui peuvent être à leur tour combinés en un ensemble infini de phrases, nous pouvons exprimer un nombre incalculable d'idées. Nous utilisons le langage pour développer des liens avec autrui, collaborer, débattre, apprendre, exprimer nos émotions, plaisanter, réciter de la poésie, écrire des livres et des articles scientifiques – les possibilités sont sans fin.

Devant une telle complexité, la maîtrise du langage est l'une des tâches les plus difficiles qui soient. Malgré cela, la grande majorité des enfants deviennent des utilisateurs compétents du langage, qu'il soit sous forme parlée ou signée. Ceci est d'autant plus surprenant que l'apprentissage du langage chez les enfants se déroule sans effort et en l'espace de quelques années seulement. *Comment les enfants apprennent-ils le langage ? Quelles étapes développementales traversent-ils au cours du processus d'apprentissage ? Comment l'environnement langagier auquel les enfants sont exposés façonne-t-il leurs compétences langagières ?* Ce sont quelques-unes des questions explorées dans la recherche sur le développement du langage.

Dans cette thèse, je propose d'aborder la question du développement du langage au moyen d'une méthode récente de collecte de données : des enregistrements sonores pouvant durer jusqu'à 16 heures et collectant la production langagière et l'environnement linguistique de jeunes enfants. Ces enregistrements longs centrés sur l'enfant (*child-centered long-form recordings*), au cœur de mon travail doctoral, donnent un aperçu réaliste des utilisations quotidiennes du langage. En plus de fournir des mesures sur ce que les enfants entendent et disent en conditions naturelles, dont l'extraction automatique sera le sujet du Chapitre 1, ces enregistrements permettent d'alimenter les modèles computationnels d'acquisition du langage avec des données réalistes, sujet des Chapitres 2 et 3.

Chapitre 1. La mesure est une pierre angulaire de la méthode scientifique. Mesurer ce que les enfants entendent et disent nous fournit de précieuses informations sur l'acquisition du langage, ce qui nous permet de mieux comprendre comment les jeunes apprenants déduisent les règles de leur langue maternelle. Voici quelques-unes des questions auxquelles les chercheurs s'intéressent : *Qui parle à l'enfant ? À quelle fréquence ? Que leur dit-t-on et comment ? Quels sons, mots ou phrases l'enfant produit-t-il ? Comment ces mesures changent-elles avec l'âge de l'enfant ? D'une population à une autre ? D'un individu à l'autre ? Ces mesures sont-elles prédictives des compétences langagières développées par l'enfant ? Peuvent-t-elles être utilisées pour détecter les troubles de l'acquisition du langage ou quantifier l'efficacité des programmes de remédiation langagière ?*

Il est possible de répondre à bon nombre de ces questions en utilisant la technologie des enregistrements longs centrés sur l'enfant. De manière regrettable, résoudre ces questions n'est pas chose aisée. L'information à laquelle on s'intéresse est en effet enfouie dans des milliers d'heures d'enregistrement et nécessite l'utilisation d'outils de traitement automatique de la parole. À cela s'ajoute la nécessité que nos algorithmes soient compatibles avec les conditions non contrôlées caractéristiques des enregistrements centrés sur l'enfant : l'enfant peut être à la maison, au parc ou à la crèche ; la parole enregistrée peut être partiellement masquée par des bruits environnementaux tels que les sons produits par un aspirateur, le trafic routier, la télévision ou la radio.

Dans ce premier Chapitre, nous proposons une analyse des performances d'un système à la pointe de la reconnaissance automatique de la parole, le modèle *Whisper* développé en 2022 par l'entreprise américaine *OpenAI*. Sur des enregistrements longs centrés sur l'enfant collectés dans des familles anglophones vivant aux États-Unis, Whisper obtient un taux d'erreur de mots (*word error rate*) moyen de 47.9%. Il est intéressant de mentionner que des taux d'erreurs bien plus bas sur des corpus couramment utilisés en reconnaissance automatique de la parole ont été documentés. En effet, le taux d'erreur moyen à travers les corpus *LibriSpeech*, *Common Voices*, *Vox Populi* et *Fleurs* est seulement de 7%. Nous attribuons une telle différence de taux d'erreur de mots entre les enregistrements longs centrés sur l'enfant et les corpus couramment utilisés à deux facteurs principaux. Le premier facteur est la présence de parole prononcée par des enfants, très largement absente des données d'apprentissage actuellement utilisées pour entraîner les modèles de reconnaissance automatique de la parole. Le deuxième facteur est la présence de conditions d'enregistrement plus difficiles que celles couramment considérées (voir l'analyse acoustique proposée dans le Chapitre 3). Cette première étude illustre certaines des difficultés liées à l'utilisation des enregistrements longs centrés sur l'enfant.

Nous poursuivons avec une analyse technique et scientifique du logiciel propriétaire LENA (Language Environment Analysis) développé par l'organisme américain à but non lucratif du même nom. L'utilisation combinée d'enregistreurs légers portés par l'enfant et de ce logiciel propriétaire a profondément transformé la recherche sur le développement du langage, le logiciel LENA s'étant rapidement imposé comme l'outil standard pour l'analyse automatique des enregistrements longs centrés sur l'enfant. Entre autres choses, le logiciel LENA segmente l'enregistrement audio en plusieurs catégories en fonction de ce qu'il contient : des vocalisations produites par l'enfant ou de la parole prononcée par un adulte. De cette première étape de segmentation sont extraites plusieurs mesures : 1) le nombre de tours de conversation (*conversational turns*) entre l'enfant et un adulte ; 2) le nombre de mots prononcés par les locuteurs adultes ; 3) le nombre de vocalisations canoniques produites par l'enfant, en ignorant donc les pleurs, les rires, et les sons végétatifs (bruits de respiration ou d'éructation).

Notre analyse révèle trois limitations importantes du logiciel LENA. Premièrement, le logiciel est propriétaire, rendant difficile d'accès les nombreux détails d'implémentation pouvant influer sur les performances du logiciel. Deuxièmement, le logiciel repose sur des technologies de traitement automatique de la parole développées au début des années 2000, et avec deux décennies de progrès, on peut raisonnablement se demander si les performances ne bénéficieraient pas d'une mise à jour. Troisièmement, le logiciel LENA a été entraîné exclusivement sur de l'anglais américain, ce qui ne garantit pas qu'il puisse fonctionner aussi bien sur d'autres langues.

C'est dans ce contexte que nous proposons une alternative libre, gratuite et plus performante au logiciel LENA. En utilisant des méthodes d'apprentissage supervisé, et en collaboration avec d'autres chercheurs, j'ai contribué à développer une suite d'outils de traitement automatique de la parole entraînés sur des corpus multilingues pour détecter l'activité vocale, identifier les sources de signaux vocaux (vocalisations de l'enfant, paroles d'une femme ou d'un homme adulte), compter le nombre d'unités linguistiques (phonèmes, syllabes, mots), et estimer la quantité de bruit et de réverbération. En plus de constituer une alternative libre et gratuite au logiciel LENA, nos résultats montrent que les performances de nos algorithmes sont égales ou supérieures aux performances obtenues par l'algorithme LENA.

Nous concluons ce Chapitre en abordant trois aspects essentiels de cette ligne de recherche, à savoir : 1) la nécessité d'aller au-delà des mesures initialement développées par l'organisation américaine LENA ; 2) l'importance de promouvoir la diversité dans les données d'apprentissage et d'évaluation afin de construire des algorithmes inclusifs avec un minimum de biais ; 3) l'importance de démocratiser

l'usage de ces algorithmes en les rendant accessibles à des acteurs moins familiarisés avec l'informatique et le traitement du signal.

Chapitre 2. Pendant de nombreuses années, des scientifiques de divers domaines, de la linguistique formelle à la psychologie du développement, en passant par l'intelligence artificielle, ont envisagé la possibilité d'exécuter des simulations d'apprentissage du langage sur ordinateur. De telles simulations sont importantes à la fois pour des raisons pratiques et théoriques. Sur le plan théorique, les simulations peuvent aider à prouver ou réfuter certaines hypothèses – et à en formuler de nouvelles – sur la façon dont les nourrissons apprennent leur langue maternelle. Sur le plan pratique, de telles simulations améliorent les compétences langagières des ordinateurs, leur permettant de comprendre le langage et de le parler de manière plus efficace.

Nous commençons ce Chapitre avec une synthèse des principaux résultats d'expériences en laboratoire portant sur les capacités langagières des jeunes enfants. Par exemple, une amélioration des capacités de discrimination des sons de la langue maternelle est observée chez les nourissons entre 6 et 12 mois, tandis que leur capacité à discriminer les sons non natifs diminue. En d'autres termes, la discrimination des sons chez les jeunes enfants se s'adapte à leur langue maternelle. À 4 mois, les nourrissons commencent en général à reconnaître leur propre nom, et à l'âge de 8 mois, la plupart d'entre eux connaissent la signification de nombreux mots. Étonnament, cette connaissance précoce des aspects lexicaux et sémantiques de leur langue maternelle intervient bien avant que les nourrissons ne développent pleinement leur capacité de discrimination sonore, et avant même qu'ils ne produisent leur premier mot, généralement vers la fin de leur première année de vie.

De nombreuses théories ont été proposées afin d'expliquer l'apprentissage du langage chez l'enfant. L'*apprentissage statistique*, sans doute l'une des plus importantes d'entre elles, souligne la capacité des enfants à extraire les propriétés statistiques de leur environnement linguistique afin d'en extraire la structure phonétique, lexicale et grammaticale. Une seconde théorie, appelée *apprentissage inter-situationnel* (*cross-situational learning*), met en avant la capacité des jeunes enfants à agréger des informations à partir d'observations de cooccurrences entre un mot et sa signification, constituant ainsi une explication de l'apprentissage sémantique précoce. Les théories de l'*apprentissage social* mettent l'accent sur le rôle des facteurs sociaux dans l'acquisition du langage et l'importance de l'interaction humaine. Ces théories soulignent l'importance de nombreux mécanismes, y compris l'imitation et le renforcement, l'attention conjointe (c'est-à-dire l'attention coordonnée d'un enfant et

son parent envers un objet ou un événement), les boucles de rétroaction lors de la communication, etc.

Une manière de tester ces théories consiste à les implémenter. En effet, si l'acquisition du langage chez les nourrissons se fait par le biais de certains mécanismes, alors l'implémentation de ces mêmes mécanismes devrait produire des résultats d'apprentissage similaires à ceux observés chez l'enfant. Bien que l'exécution de simulations reflétant toute la complexité du monde réel ne soit pas encore possible aujourd'hui, les simulations d'apprentissage représentent un outil précieux dans l'étude de l'acquisition du langage en nous fournissant des preuves d'apprenabilité sous la forme suivante : "La propriété P peut être apprise à partir de l'entrée E en utilisant le mécanisme M". Ces preuves d'apprenabilité nous éclairent ainsi sur ce que les jeunes enfants peuvent apprendre sur la base exclusive du mécanisme M et de l'entrée E.

Après avoir présenté une vue d'ensemble du paysage méthodologique dans l'exécution de simulations d'apprentissage, nous présentons notre propre approche, au cœur du Chapitre 2 et 3 de cette thèse. Nous proposons de comparer les résultats d'apprentissage de notre simulation aux mesures comportementales faites en laboratoire chez l'enfant ou l'adulte. En particulier, nous proposons l'utilisation d'une tâche de discrimination sonore ABX, un protocole couramment utilisé en psycholinguistique. Durant cette tâche, l'apprenant artificiel reçoit trois triphones A, B et X avec A et X correspondant à différentes occurrences du même triphone (par exemple "bip") et B correspondant à un autre triphone dont le phone central diffère (par exemple "bop"). Si l'apprenant artificiel retourne une distance entre A et B plus petite que la distance entre A et X, alors il réussit, sinon il échoue.

Tandis que cette tâche évalue les capacités de discrimination sonore de l'apprennant artificiel, la seconde tâche proposée évalue ses capacités au niveau lexical. Dans cette tâche lexicale, l'apprenant reçoit un vrai mot (par exemple "dragon") ainsi qu'un pseudo-mot présentant une probabilité phonotactique similaire (par exemple "draton"). Si l'apprenant artificiel associe une probabilité plus élevée au vrai mot qu'au pseudo-mot, alors il réussit, sinon il échoue.

Avant de présenter nos résultats, il convient de mentionner deux caractéristiques essentielles de notre méthodologie. Premièrement, nous adoptons une approche inter-linguistique (*cross-linguistic*) dans laquelle l'apprenant artificiel est exposé soit à de l'anglais, soit à du français, simulant ainsi un enfant apprenant l'anglais ou le français. Durant l'étape d'évaluation, nos deux apprenants sont évalués sur les deux langues. La comparaison entre les scores natifs (apprentissage et évaluation sur la même langue) et non natifs (apprentissage et évaluation sur deux langues

différentes) nous permet de mesurer ce que l'apprenant a appris grâce à l'exposition à sa langue d'entraînement, par opposition à l'exposition à une autre langue. Deuxièmement, nous adoptons une approche développementale dans laquelle nous faisons varier la quantité de parole à laquelle l'apprenant a accès. Ceci nous permet d'étudier l'effet de la quantité de parole sur les résultats d'apprentissage.

Dans cette première étude, nous exposons un modèle d'apprentissage auto-supervisé à de grandes quantités de livres audio en suivant la méthodologie décrite plus haut. Les trajectoires d'apprentissage suivies par notre apprenant artificiel montrent un effet positif de leur langue d'entraînement, c'est-à-dire que l'apprenant natif (par exemple le modèle anglais évalué sur l'anglais) obtient des scores phonétiques et lexicaux plus élevés que l'apprenant non-natif (par exemple le modèle français évalué sur l'anglais), et cela dès l'exposition à 50 heures de parole. Nous observons également un effet important de la quantité de parole, c'est-à-dire que le modèle natif s'améliore sur la tâche de discrimination sonore ainsi que sur la tâche lexicale à mesure que la quantité de parole dans l'ensemble d'apprentissage augmente. Tandis que l'on aurait pu s'attendre à un apprentissage successif durant lequel le modèle apprend à discriminer les sons avant d'apprendre les mots, nous observons une trajectoire parallèle, compatible avec les observations faites chez l'enfant.

Notre simulation fournit une preuve d'apprenabilité. En exposant nos algorithmes auto-supervisés à de la parole, nous avons montré que ces derniers apprennent à discriminer les sons et à discriminer les mots de pseudo-mots. De plus, les trajectoires d'apprentissage de nos apprenants artificiels sont compatibles avec celles observées chez l'enfant. Ainsi, nous démontrons que les théories d'apprentissage statistique, dont nos algorithmes sont l'un des nombreux représentants, suffisent à expliquer certains des aspects de l'apprentissage phonétique et lexical chez l'enfant.

Chapitre 3. Une mission d'importance capitale en sciences en général, et dans les études de modélisation en particulier, est de comprendre dans quelle mesure nos découvertes s'appliquent au monde réel.

Dans la simulation décrite plus haut, nous exposons nos apprenants artificiels à des livres audio. Par ce biais, nous faisons l'hypothèse simplificatrice importante que l'environnement linguistique de l'enfant est constitué de longues phrases articulées, prononcées par le même locuteur, couvrant un large vocabulaire, présentant une syntaxe complexe, et produites dans un environnement acoustique favorable (relativement peu de bruits et peu de réverbération). Ces hypothèses simplificatrices font partie intégrante de l'entreprise de modélisation mais ne correspondent pas aux conditions auxquelles les jeunes enfants sont confrontés. *Comment la complexité de l'environnement langagier réel des enfants affecte-t-elle l'acquisition du langage ?*

Les théories existantes rendent-elles compte de manière adéquate de ce que les enfants entendent vraiment ? Nos modèles computationnels présentent-ils les mêmes résultats d'apprentissage lorsqu'ils sont formés sur des données propres ou réalistes ? Ces questions sont au centre du Chapitre 3. Nous proposons d'alimenter les simulations d'acquisition du langage directement avec ce que les nourrissons entendent en utilisant les enregistrements longs centrés sur l'enfant introduits au Chapitre 1.

Nous commençons par présenter un article expliquant comment un programme de recherche se focalisant sur l'utilisation des enregistrements longs centrés sur l'enfant dans le cadre de la modélisation de l'acquisition du langage pourrait procéder. En particulier, nous présentons comment certaines expériences de perception de la parole en laboratoire chez l'enfant peuvent être adaptées pour évaluer les apprenants artificiels.

À travers une analyse acoustique comparative entre les enregistrements longs centrés sur l'enfant et les livres audio couramment utilisés dans les modélisations d'apprentissage, nous montrons que la parole contenue dans les enregistrements centrés sur l'enfant contient un niveau de bruit bien plus élevé que celle trouvée dans les livres audio. Notre analyse révèle également que contrairement aux livres audio, les enregistrements longs centrés sur l'enfant présentent une grande variété d'environnement réverbérant dégradant fortement la qualité du signal de parole.

À la lumière de ces différences acoustiques, nous proposons de comparer l'apprentissage d'un apprenant artificiel, le même que celui étudié au Chapitre 2, lorsque ce dernier est exposé aux livres audio ou aux enregistrements centrés sur l'enfant. En particulier, nous nous intéressons au processus de synchronisation perceptuelle intervenant entre 6 et 12 mois chez les nourrissons, au cours duquel ces derniers deviennent meilleurs pour discriminer les sons natifs et moins bons pour discriminer les sons non natifs. Adoptant la même approche inter-linguistique et développementale que celle présentée au Chapitre 2, nous montrons que notre apprenant artificiel reproduit en effet le processus de synchronisation perceptuelle lorsqu'il est entraîné sur des livres audio, ce qui n'est pas le cas de l'apprenant artificiel entraîné sur les enregistrements centrés sur l'enfant.

Afin de reproduire le processus de synchronisation perceptuelle à partir de ces enregistrements, il est nécessaire d'équiper notre apprenant avec certains mécanismes visant à guider le processus d'apprentissage. Ces mécanismes, aussi appelés biais inductifs, ont été conçus pour répondre à deux critères : 1) ils doivent être compatibles avec les comportements documentés chez l'enfant ; 2) ils doivent atténuer certaines dégradations du signal identifiées lors de notre analyse acoustique. Puisque l'on observe que ces biais inductifs sont nécessaires pour que notre apprenant artificiel

reproduise le processus de synchronisation perceptuelle, nous explorons la possibilité que des biais inductifs similaires puissent guider l'acquisition du langage chez l'enfant.

Il est important de mentionner que nous n'observons pas de preuves d'un quelconque apprentissage lexical, contrairement à ce qui a été montré dans le Chapitre 2 avec l'apprenant artificiel entraîné sur les livres audio. À travers l'utilisation de microphones portés par l'enfant, nous montrons l'importance d'utiliser des données d'entrée réalistes dans les simulations d'acquisition du langage.

Thesis format

This document follows a thesis-by-publication format (*thèse par articles*), meaning that first-author articles are directly embedded in the document's body. Far from being a mere concatenation of articles, this format allows me to relegate the expert-oriented discourse and technical details to the scientific articles that are published or are in the process of being published. By doing so, I focus on the coherence of our past, present, and future work, highlighting the factors that led us to work on a specific subject, detailing our most important findings, and outlining the potential directions we may pursue in the future.

Throughout Chapters 1, 2, and 3, readers will find sections accompanied by an article. These sections are highlighted with a box containing the article reference, followed by a *Motivation* and a *Paper Summary* subsections. Readers should be able to understand the essence of our work by focusing exclusively on these subsections, which should ease – and considerably speed up – the reading of this document.

Acknowledgements / Remerciements

First of all, I would like to thank the exceptional scientists who agreed to evaluate my work: ***Caroline Rowland***, ***Thomas Hueber***, ***Chloé Clavel***, ***Afra Alishahi***, and ***Okko Räsänen***. I sincerely hope we will cross paths again. Merci à ***Chloé Clavel*** et ***Laurent Besacier*** de m'avoir suivi tout au long de ces trois années en acceptant de faire partie de mon comité de suivi de thèse.

Thanks to ***Hana D'Souza***, ***Okko Räsänen***, ***María Andrea***, ***Khazar***, the participants of the REST-CL 2023 retreat, and all the researchers and students who, whether through an email, a smile, a glance, or a funny joke, contributed to making academia a more inclusive and welcoming place. Sometimes I did not even know your name, but know that you made a real difference in my journey.

Les mots me manquent lorsque vient le moment de remercier mes superviseurs. Je ressors transformé de ces années à travailler à vos côtés. Et vous avez tous les trois contribué, à votre manière, à cette transformation. Merci à ***Alex*** pour ton incroyable leadership (but seriously, how do you do that?). Nos discussions et tes précieux conseils m'ont fait grandir et prendre conscience de ce dont je suis capable. Merci à ***Hervé*** pour ta bienveillance et tes explications d'une limpidité inégalée. Je suis toujours impressionné par la clarté et la rigueur de tes papiers, et par ce que tu as construit (osons le terme de don à l'humanité). Merci à ***Emmanuel*** pour ton abondance d'idées, et pour ton enthousiasme et ton émerveillement communicatifs. J'aimerais avoir un dixième de ta créativité. Merci à vous trois d'avoir tout fait pour me mettre en confiance, et de m'avoir accordé tant de liberté tout en vous montrant si disponibles. Merci à vous trois de m'avoir montré à quoi pouvait ressembler une recherche responsable, libre et engagée. Une leader, un bâtisseur et un visionnaire : je n'aurais pas pu rêver meilleur·es directeur·rices de thèse.

Merci ***Maureen*** de m'avoir montré qu'il n'y avait jamais de problèmes, mais que des solutions (c'est ta deuxième thèse, non ?). Merci aussi pour toutes nos discussions. Tu as été une alliée infaillible, aussi bien sur le plan personnel que professionnel. Le monde académique te regrettera, et moi avec. Je te souhaite tout le bonheur possible pour la suite – tu le mérites amplement – mais je ne me fais pas trop de souci pour toi. Merci à ***Xuan-Nga*** d'avoir été aussi présente et disponible. Merci à ***Juliette*** et ***Camila*** pour vos conseils, vos encouragements et votre bienveillance.

Merci à ***Marianne*** d'avoir travaillé sur tant de choses à la fois sans jamais baisser les bras. On peut être fier·ères de Brouhaha ! Merci à ***Wassim***, ***Ruben*** et ***Lucas E.*** d'avoir accepté d'être mes cobayes en travaillant avec moi alors que je débutais seulement ma thèse. Vous êtes les plus forts ! Merci à ***Angelo***, ***Catherine***, ***Chiara***, ***Hadrien***, ***Jing***, ***Julien***, ***Kasia***, ***Loann***, ***Lucas E.***, ***Lucas G.***, ***Manel***, ***Marianne***, ***Mathieu B.***, ***Mathieu R.***, ***Maxime***, ***Mitja***, ***Nicolas***, ***Philippe***, ***Rachid***, ***Rahma***, ***Robin***, ***Ruben***, ***Sabrina***, ***Salah***, ***Tu Anh***, ***Wassim*** et ***William*** (et tous ceux que j'ai oubliés) pour les rires et pour la science. Merci à toute l'équipe pour nos discussions passionnées et passionnantes et nos soirées jusqu'à pas d'heure.

Merci à ***CosmicDebris*** et ***M. Ledys*** qui m'ont profondément influencé à un très jeune âge et ont changé ma trajectoire de vie, sans même le savoir.

Merci à ***Clémentine***, ***Erwan*** et ***Aurélie*** que j'espère garder à mes côtés encore longtemps.

Merci à ***Kévin***, ***Tim*** et ***Nolan*** de supporter leur frère un peu bizarre, et à ***ma mère*** et à ***mon père*** d'avoir fait tant de sacrifices pour que je puisse réaliser mes rêves.

Merci à ***Laurent*** pour ta patience et ton soutien indéfectibles.

Publications

* denotes shared first authorship.

Included in the main (in order of appearance)

- **Lavechin, M.**, Bousbib, R., Bredin, H., Dupoux, E., Cristia, A. (2020) An open-source voice type classifier for child-centered daylong recordings. *Interspeech*
- **Lavechin, M.,**[*], Métais, M[*], Titeux, H., Boissonnet, A., Copet, J., Rivière, M., Bergelson, E., Cristia, A., Dupoux, E., Bredin, H. (2023) Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation. *ASRU*
- **Lavechin, M.,**[*], de Seyssel, M.[*], Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., Dupoux, E. (2023) Statistical learning bootstraps early language acquisition. *Submitted to Developmental Science*
- **Lavechin, M.**, de Seyssel, M., Gautheron, L., Dupoux, E., Cristia, A. (2022) Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*
- **Lavechin, M.**, de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., Cristia, A. (2023) Statistical learning models of early phonetic acquisition struggle with child-centered audio data. *Submitted to Cognition*
- **Lavechin, M.**, Sy, Y., Titeux, H., Cruz Blandón, M. A., Räsänen, O., Bredin, H., Dupoux, E., Cristia, A. (2023) BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models. *Interspeech*

Included in the appendix

- de Seyssel[*], M., **Lavechin, M.**[*], Dupoux, E. (2023) Simulating early language acquisition: first results and challenges. *Journal of Child Language*

Not included in the manuscript (in reverse order of publication year)

- de Seyssel, M., **Lavechin, M.**, Titeux, H., Thomas, A., Virlet, G., Revilla, A. S., Wisniewski, G., Ludusan, B., Dupoux, E. (2023) ProsAudit, a prosodic benchmark for self-supervised speech models. *Interspeech*
- de Seyssel, M., **Lavechin, M.**, Adi, Y., Dupoux, E., Wisniewski, G. (2022) Probing phoneme, language and speaker information in unsupervised speech representations. *Interspeech*
- Alishahi, A., Chrupała, G., Cristia, A., Dupoux, E., Higy, B., **Lavechin, M.**, Räsänen, O, Yu, C. (2021) ZR-2021VG: Zero-resource speech challenge, visually-grounded language modelling track. *ArXiV*
- Räsänen, O., Seshadri, S., **Lavechin, M.**, Cristia, A., Casillas, M. (2021) ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*
- Cristia, A., **Lavechin, M.**, Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., Bergelson, E. (2021) A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*
- Gautheron, L., **Lavechin, M.**, Riad, R., Scaff, C., Cristia, A. (2020) Longform recordings: Opportunities and challenges. *Linguistique informatique, formelle et de terrain*
- **Lavechin, M.**, Gill, M.-P., Bousbib, R., Bredin, H., Garcia-Perera, L. P. (2020) End-to-end domain-adversarial voice activity detection. *Interspeech*
- Bredin, H., Yin, R., Coria, Gelly, G., Korshunov, P., **Lavechin, M.**, Fustes, D., Titeux, H., Bouaziz, W., Gill, M.P. (2020) Pyannote.audio: neural building blocks for speaker diarization. *International Conference on Acoustics, Speech and Signal Processing*
- Garcia Perera, L. P, Villalba, J., Bredin, H., Du, J., Castan, D., Cristia, A., Bullock, L., Guo, L., Okabe, K., Nidadavolu, P. S., Kataria, S., Chen, S., Galmant, L., **Lavechin, M.**, Sun, L., Gill, M.-P., Ben-Yair, B., Abdoli, S., Wang, X., Bouaziz, W., Titeux, H., Dupoux, E., Lee, K. A., Dehak, N. (2020) Speaker Detection in the Wild: Lessons Learned from JSALT 2019. *Odyssey The Speaker and Language Recognition Workshop*

Contents

Acronyms

ACLEW Analyzing Child Language Experiences around the World

ALICE Automatic LInguistic unit Count Estimator

ANN Artificial Neural Network

APC Adult Phoneme Count

ASC Adult Syllable Count

ASD Autism Spectrum Disorder

ASR Automatic Speech Recognition

AWC Adult Word Count

CDI Child Development Inventory

CHILDES CHIld Language Data Exchange System

CPC Contrastive Predictive Coding

CPU Central Processing Unit

CTC Conversational Turn Count

CVC Child Vocalization Count

ELE Electronic speech

FEM Female adult speech

GPU Graphics Processing Unit

KCHI Key child (wearing the recorder)

LAD Language Acquisition Device

LENA Language ENvironment Analysis

LSCP Laboratoire de Sciences Cognitives et Psycholinguistique

LSTM Long Short-Term Memory

MAL Male adult speech

MDGMM Minimum Duration Gaussian Mixture Model

MFCC Mel-Frequency Cepstrum Coefficients

NLP Natural Language Processing

OCH Other Children

RER Relative Error Rate

RIR Room Impulse Response

SEEDLingS Study of Environmental Effects on Developing Linguistic Skills

SES SocioEconomic Status

SNR Speech-to-Noise Ratio

STELA STatistical learning of Early Language Acquisition

VAD Voice Activity Detection

VTC Voice Type Classifier

WER Word Error Rate

List of figures

List of tables

0 Introduction

Language is unique to humans. While sophisticated communication systems are found across many species, human languages exhibit unrivaled intricacies and expressive power (Hockett, 1960). By combining a finite set of sounds into an infinite set of words, which in turn can be combined into an infinite set of sentences, we can express countless ideas. We use language to bond with others, collaborate, debate, learn, express our feelings, make jokes, recite poetry, and write books and scientific articles – the possibilities are endless.

Given such complexity, mastering one's native language is one of the most challenging tasks a human being will ever face. Nevertheless, the vast majority of infants become proficient language users, whether in spoken or signed form (Goldin-Meadow & Brentari, 2017). What is even more astounding is that they do it within a few years and in an effortless manner. *How do infants learn to comprehend and produce language? What developmental stages do they undergo as the learning process unravels? How does the language infants are exposed to shape their language skills?* Those are some of the key questions explored in language development research.

Attempting to answer these questions takes much collaborative effort from researchers spanning different disciplines such as neuroscience, psychology, linguistics, philosophy, computer sciences, and many others. Despite scientists' best efforts to reveal the hidden workings of infants, numerous secrets continue to elude our understanding. In our quest for comprehension, we must find ingenious ways to collect and describe what children hear and produce, elicit and measure their behavior or brain activity while exposed to language, and so on – much of this methodology will be introduced throughout the thesis.

During my stay at the *Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP)* and *Meta AI* in Paris, under the supervision of Alejandrina Cristia, Hervé Bredin, and Emmanuel Dupoux, I explored how artificial neural networks (ANN) could be used to foster progress in language development research. In particular, I focused on using child-centered long-form recordings that collect everyday language uses from the child's perspective – described in Section 0.2. In this context, my collaborators and I developed a suite of speech processing tools to automatically describe the input afforded to children as well as their language production. This aspect is discussed

in Chapter 1 and illustrates how artificial neural networks can be used *as a tool* to describe children's language environments. A second aspect of my doctoral work focused on modeling early language acquisition. Computer simulations can help us get insights into how infants may acquire their native language. These aspects are discussed in Chapters 2 and 3 and illustrate how artificial neural networks can be used *as a model* of the infant learner.

0.1 The nature versus nurture debate

Abbaye de Royaumont (Val-d'Oise), October 1975. Noam Chomsky, an esteemed American linguist, and Jean Piaget, a distinguished Swiss psychologist, met for the first time to discuss their views on the origins of learning and language (Manesse & Miniac, 1981). The encounter between these two intellectual giants will remain forever engraved in the history of cognitive sciences. At the heart of the discussion was the fundamental question of the relative contribution of genes (nature) and the environment (nurture) in infant language acquisition.

On the one hand, Chomsky's nativist theory posits that language acquisition is primarily driven by innate, universal principles specific to the human brain. Chomsky argues that children possess an innate language acquisition device (LAD), a hypothetical cognitive module containing universal grammatical rules or principles common to all languages and enabling infants to acquire language effortlessly and rapidly (Chomsky, 1959). Central to his claim is the poverty of the stimulus argument, which posits that children are not exposed to rich enough data within their linguistic environments to acquire every feature of their language (Chomsky et al., 1980).

On the other hand, Piaget's constructivist theory argues that language acquisition is intricately linked to cognitive development. According to his theory of cognitive development, children progress through distinct mental stages, constructing knowledge and understanding through their interactions with the environment. In this view, language acquisition is a gradual process that evolves alongside the child's cognitive abilities. Constructivist theories emphasize the proactive role of children in acquiring language from sensorimotor experiences (Piaget, 1935; Vygotsky, 1962; Tomasello & Farrar, 1986).

Although it is widely acknowledged that both nature and nurture play significant roles in infant language acquisition (Karmiloff-Smith, 1998; Rowland, 2013), the relative contribution of each remains the subject of intense and contentious debates (Saffran & Thiessen, 2008; Ambridge & Lieven, 2011).

Recently, the progress of artificial intelligence and machine learning has made available a new source of evidence that can contribute to this debate: simulation. By building an artificial language learner and placing it in an environment similar to that of infants, one could determine, in theory, which aspects of the learning trajectories observed in humans can be learned by the machine (Dupoux, 2018; Lavechin, de Seyssel, Gautheron, et al., 2022). We will dedicate the General discussion to this matter.

As we elaborate in Chapter 4, the key to placing an artificial learner in an environment similar to that of infants comes from child-centered long-form recordings.

0.2 What are child-centered long-form recordings?

Fueled by the advancements in wearable and battery technology, child-centered daylong or long-form recordings[1] allow us to capture what children hear and say across an entire day by collecting audio data directly within the child's environment and from the child's perspective.

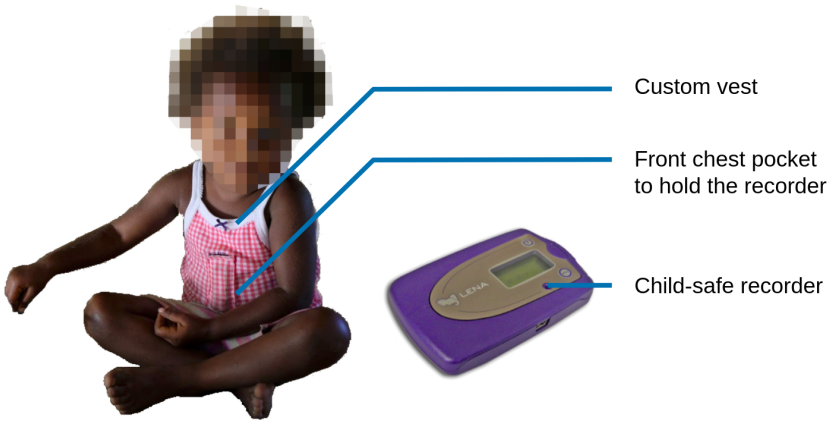

Fig. 1.: Child-centered long-form recordings are typically collected using a custom-made piece of clothing with a front chest pocket to hold a lightweight and child-safe recorder. This Figure shows a LENA® recording device capable of capturing up to 16 hours of audio worn by a little girl growing up in Vanuatu. Photograph taken by Heidi Colleran.

In most settings, infants and young children wear a custom-made piece of clothing with a front chest pocket, within which a lightweight recording device is inserted – see Figure 1. Although researchers sometimes use alternative or custom devices,

[1] From now on, we will use exclusively child-centered long-form recordings (often shortened as long-forms). The 'long-form' adjective better reflects, in opposition to 'day-long', that some researchers do not capture the entirety of the child's waking day, and others may also capture nighttime.

e.g., Cao et al. (2018), the large majority of long-forms are collected with the recorder designed in the late 2000s by the LENA® (Language ENvironment Analysis) nonprofit American organization. As we shall see in Chapter 1, the LENA® device has the advantage of being accompanied by a speech processing software, allowing researchers to extract measures of what children hear and say automatically.

From a scientific perspective, long-forms have many advantages. First, they sample the full range of language input the child is exposed to as well as the language output they produce. Second, they reduce observer effects relative to, e.g., shorter audio or video recordings (Bergelson et al., 2022). Third, they consume relatively little electricity and are compatible with field data collection in remote places – e.g., see Casillas et al. (2021) for a study in a Papuan community. Finally, the technology behind long-forms is becoming increasingly cheap. At the time of writing, one LENA® recorder currently costs between 219\$ and 329\$ ("LENA shop", 2023), while USB lavalier microphones cost less than 20\$.

These combined advantages allow researchers to obtain an ecologically-valid view of children's language environments around the world and to foster progress in language acquisition research. In this thesis, we will adopt a perspective primarily focused on signal processing challenges while trying to remain relevant to language development research[2]. As we delve into the subject, it will become evident that long-form recordings offer many opportunities and pose remarkable challenges.

0.3 Content overview

Having gained a good understanding of the technology discussed in this thesis, we provide a concise content overview.

Chapter 1 presents our contributions to building automatic speech processing systems for child-centered long-forms.

We start with a short study demonstrating the low performance of current state-of-the-art automatic speech recognition systems. We then present the LENA proprietary software, now routinely used to analyze children's language environments collected through wearable recording devices. Finally, we present our efforts in proposing a free and open-source alternative to the LENA® software. In addition to offering better-performing algorithms, our solution expands the range of measures designed

[2]For aspects not covered in this thesis, we will refer to Casillas and Cristia (2019) for a step-by-step guide on collecting and analyzing long-forms and Cychosz et al. (2020) for ethical guidelines.

initially by the LENA® Foundation. This chapter shows how artificial neural networks can be used as a tool to process child-centered long-forms.

Chapter 2 presents our contributions to modeling early language acquisition from audiobooks.

We review some results in the language acquisition literature, presenting key developmental milestones observed in infants and the learning mechanisms proposed to explain how infants acquire their native language. We explain how running language learning simulations in computers can help us better understand infant language acquisition. After presenting the various approaches to modeling language acquisition, we present our own approach, with a short study assessing the phonetic and lexical capabilities of a self-supervised deep learning model trained on audiobooks. This chapter returns to the nature versus nurture debate mentioned above by providing proof of learnability. Furthermore, it sets the stage for the following chapter by proposing a candidate algorithm to learn from ecological long-forms.

Chapter 3 presents our contributions in modeling early language acquisition from ecological child-centered long-forms.

We start with a position paper advocating for the use of ecologically-valid data when modeling language acquisition. After laying out recommendations on how a research program centered around modeling language acquisition from long-forms could proceed, we present our own modeling study of early phonetic acquisition. Concluding this chapter, we introduce a benchmark compatible with the vocabulary typical of children's language environments and identify two challenges that must be addressed to run more realistic simulations of language acquisition. This chapter brings together some of the tools built in Chapter 1 with the candidate learning mechanism proposed in Chapter 2 with the additional constraint that the learning must occur in an environment similar to that of infants.

Chapter 4 consists of a general discussion reflecting on the limitations and the broader implications of our findings.

We first summarize our main contributions. We present how our approach in learning from ecological long-forms can contribute to the nature versus nurture debate. We close this thesis with a hypothetical scenario that outlines our perspective on the trajectory the field must embark upon to make substantial advancements.

1 Automatic analysis of children's language experiences

Measurement is a cornerstone of science, and language development research is no exception. Measurements of what children hear and say provide critical information about early language acquisition, enabling us to gain a deeper understanding of how young language learners infer the rules of their native language. Central questions in language development research include: Who talks to the child, how often, what do they say, and how do they say it? What sounds, words, or sentences does the child produce? How do these measures change with age, across diverse populations, and between individuals? Are these measures predictive of later language skills? Can they be used to detect language disorders or quantify the effectiveness of language remediation programs?

Traditionally, such measures are obtained from short recordings of naturalistic interactions in semi-controlled environments (Hart & Risley, 1995; Bergelson, Amatuni, et al., 2019) or times of in-person observations (Roopnarine et al., 2005). Although informative, these data collection methods provide researchers with a limited perspective on the vast array of communication situations children experience and engage in throughout the day, from morning to night. In the late 2000s, the advent of the LENA® system enabled researchers to capture language use in everyday life directly from the child's perspective, promising an ecologically valid description of children's language experiences.

As we shall see throughout this first chapter, this ecological validity comes with a price, which involves dealing with real-life recordings that are rarely clean, sometimes unintelligible, most often noisy, and span up to 16 hours – audio samples are available on this project page[1]. Child-centered long-form recordings raise many interesting scientific and engineering challenges, some of which are addressed in this first chapter.

[1] https://marvinlvn.github.io/projects/2_project

We begin by addressing some of the challenges of working with child-centered long-form data with a short study assessing the performance of a state-of-the-art automatic speech recognition system on recordings collected in American English-speaking families. We then introduce the LENA® proprietary software dedicated to the automatic analysis of long-forms, whose combined use with the LENA® wearable recorder has had a major impact on language development research. We highlight certain limitations of the LENA® system and present our attempt to propose a free, open-source, and more accurate alternative. Next, we present our voice activity detection model that estimates the background noise level and the quantity of reverberation, both of which are salient in naturalistic speech as found in long-forms. Finally, we reflect on current limitations and potential future work in building automatic speech processing tools compatible with long-forms.

1.1 Automatic speech recognition

Fig. 1.1.: Automatic speech recognition is the task of transcribing a given audio to text, i.e., answering the question: "What is being said?".

.

About a decade ago, the performance of automatic speech recognition (ASR, see Figure 1.1) systems dramatically increased with the adoption of deep neural network-based hybrid approaches (Hinton et al., 2012). More recently, the ASR community has seen another quantum leap forward with the emergence of end-to-end systems supplanting traditional modeling components with a single network (J. Li, 2022). Some claim that modern ASR systems now achieve human-level or even supra-human performance (Xiong et al., 2016; T. S. Nguyen et al., 2020; Radford et al., 2022).

These recent breakthroughs promise to ease the life of language development researchers working with child-centered long-form recordings. Indeed, long-forms can quickly accumulate thousands of hours of audio, making manual transcription

too expensive and time-consuming to be feasible. A fully-functioning ASR system capable of handling the high variability and high quantity of noise inherent in long-form recordings would provide researchers with a precise understanding of what children hear and produce, enabling a level of description that has never been achieved before.

This section proposes a short study of one such ASR system claimed to achieve human-level performance: Whisper from OpenAI (Radford et al., 2022). We aim to demonstrate that current state-of-the-art ASR systems are barely usable on long-forms – and, thus, are far from reaching human-level performance. In doing so, we introduce our readers to some of the difficulties in using long-forms, namely, the challenging recording conditions.

1.1.1 Performance of Whisper on American English long-forms

Experimental protocol

Dataset. For this analysis, we consider the Bergelson corpus (Bergelson, 2017; Bergelson, Casillas, et al., 2019) collected from 44 American English infants, from 6-18 months of age (mean $\mu_{age} = 12.5$ mths and standard deviation $\sigma_{age} = 3.3$ mths), living in Rochester, western New York. This corpus consists of long-form recordings using the same LENA® recorder as presented in the Introduction (single channel, 16-kHz sampling rate). Out of the 44 infants, 10 were selected for manual annotation, chosen to represent the full diversity of the original corpus in terms of maternal education and age range. Fifteen 2-minute non-overlapping sessions were randomly sampled from the selected long-forms, resulting in 5 hours of audio transcribed by expert annotators – see Soderstrom et al. (2021) for the rationale of the sampling and annotation processes.

After filtering out non-speech segments (e.g., cries, laughs, etc.) and utterances transcribed as unintelligible, we are left with 2,174 utterances totaling 51 mn of speech. Among these utterances, 66% are produced by female adults (FEM), 14% by male adults (MAL), 12% by electronic devices such as TVs or toys (ELE), and 6% by children (CHI; either the child wearing the recording device or any other children in the environment).

Model. We automatically transcribe each of the 2,174 utterances using the Whisper large system (Radford et al., 2022). Whisper is a large-scale transformer-based ASR system trained on 680,000 hours of multilingual and multitask audio data collected

from the internet that has been shown to yield competitive performance on a variety of benchmarks, including multiple languages.

Evaluation metric. To evaluate the performance of Whisper, we use the word error rate (WER) computed as:

$$WER = \frac{S + D + I}{N}$$

where S, D, and I are the number of substitutions, deletions, and insertions between the reference (i.e., the human transcription) and the hypothesis (i.e., the automatic transcription), and N is the number of words in the reference.

Results and discussion

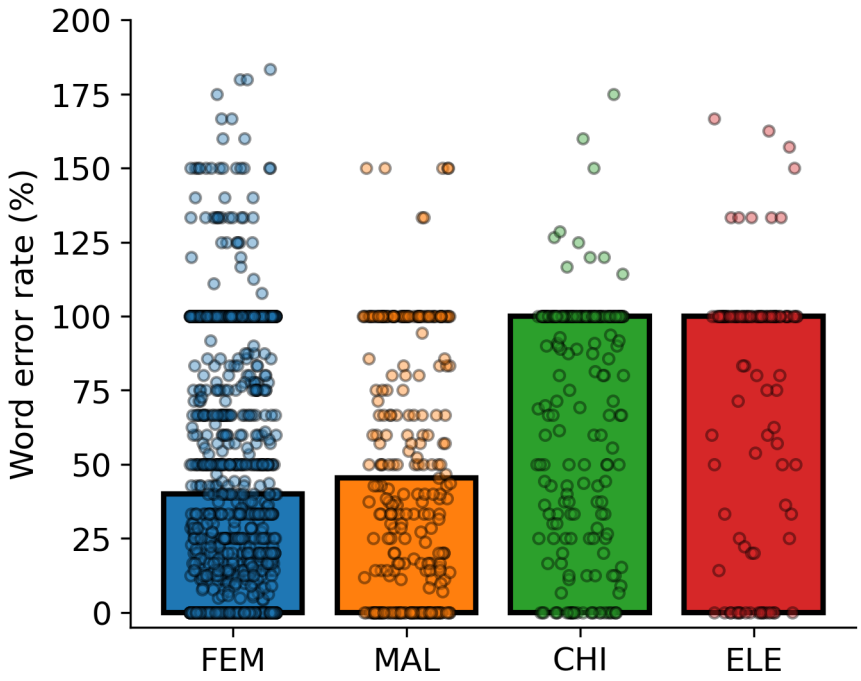

Fig. 1.2.: Word error rate (WER) obtained by Whisper large as a function of voice type. FEM stands for female adult speech, MAL stands for male adult speech, CHI stands for child vocalization, and ELE stands for electronic speech (TV, radio, toys, etc.). Each point is an utterance, and bars indicate the median word error rate across utterances produced by a given voice type. The WER is greater than 100% when there are more errors in the automatic transcript than words in the reference. Only utterances for which Whisper obtained a WER lower than 200% are displayed.

On the Bergelson corpus, Whisper obtains an average WER of 47.9%, suggesting poor transcription capabilities on naturalistic speech utterances found in child-centered long-forms[2]. In comparison, the same model obtains an average WER of 7.0% on four of the most popular benchmarks in the ASR community: LibriSpeech (Panayotov et al., 2015); Common Voice (Ardila et al., 2019); VoxPopuli (Wang et al., 2021); and Fleurs (Conneau et al., 2023). The difficult acoustic conditions found in long-forms might explain the observed 40% difference in WER between them and popular ASR benchmarks. As long-forms collect language use in everyday life, speech is not recorded in a controlled environment. Instead, it is spoken from a distance, it is reverberated and absorbed by surrounding obstacles, and it can be mixed with other speech sounds and background noise. In addition, people do not speak in clear and well-articulated sentences but may mumble, whisper, or laugh while speaking, producing short turns that sometimes overlap. We will return to the impact of the difficult acoustic conditions found in long-forms in Section 1.4.

To a certain extent, the low WER obtained by Whisper relates to the different speech sources found in long-forms, as shown in Figure 1.2. Although this Figure should be interpreted with precaution given the small sample size of our analysis, results suggest similar performance for female and male adult speech (median WER of 40.0% for FEM versus 45.5% for MAL) and worse performance for children's vocalizations and electronic speech (median WER of 100%). Numerous studies indicate that automatic recognition of children's speech is a challenging task due to various factors such as the high variability in speech sounds, the high pitch range, the presence of disfluencies, and the limited amount of data available for training that are primarily focused on adult speakers – see Bhardwaj et al. (2022) for a review. The lower performance obtained on electronic speech may be attributed to electronic devices like TVs or radios typically running in the background, resulting in distorted far-field speech that is more challenging to transcribe accurately. Overall, the data show a high standard deviation (too high to be displayed in Figure 1.2), with a high density of utterances for which Whisper obtains a WER of 0% or 100%. In other words, there are many instances where Whisper produces a transcription that is either completely accurate or completely wrong. This demonstrates the low reliability of Whisper in automatically transcribing utterances found in long-forms.

[2]An alternative experimental protocol consists in running Whisper on the entire audio recordings without using utterance boundaries. In this scenario, we assign a greater weight to the false alarms produced by the model. Following this protocol leads to Whisper obtaining an average WER of 100%, as opposed to the 47.9% WER obtained when using utterance boundaries. Another protocol worth considering consists in extending utterance boundaries to provide the model with additional contextual information. While we have not conducted this analysis, it is reasonable to expect that Whisper may exhibit slightly higher performance with longer contexts.

Human transcription	Automatic transcription
what	what about the laundry
it is time for me to take a shower	it is time for me to take a shower
so you can play with some toys	so you can claim some poison
in the bathroom	with that
i got your house	i got your house
and your baby	and your baby
should we get a couple more toys	should we go get a couple more toys
what else do you like	what else do you like
no remember we did this yesterday	no remember we did this yesterday
it was fun	it was fast
how about your magna doodle	i like your hair too
do you want to take this in	do you want to take this in
so you can color	so you can color
i think we should find one more toy	i think we should find one more toy
how about your ball	how about your ball
where is your ball	where is your ball
here it is	here it is
here you hold this one	here you hold this one
and i will pick up the ball	and i will pick up the box
your father is ridiculous	your father is ridiculous
okay	okay

Tab. 1.1.: Example of a 2-minute session extracted from American English long-forms for which Whisper obtains a word error rate of 8.8%. Insertions and substitutions are indicated in red.

Finally, Tables 1.1 and 1.2 show examples of 2-minute sessions transcribed by Whisper. A closer examination of the transcription produced by Whisper provides valuable insights. We observe transcription errors that seem natural, in the sense that they can easily be attributed to the acoustic similarity between the human and the automatic transcript, e.g., "it was fun" transcribed to "it was fast" or "and I will pick up the ball" transcribed to "and I will pick up the box". However, there also exist errors that can hardly be explained by mere acoustic similarity, e.g., "and a soft rabbit" transcribed to "you know it is all together". Although it is hard to provide precise evidence, we hypothesize that these errors are due to the language modeling task (among other tasks) used to train Whisper. As the model is trained to transcribe the audio and predict the next token, there might be situations where one task dominates the other.

In the same vein, we noticed hallucinations like "thanks for watching" and "please hit the like button and share the video with friends on social media" that are not pronounced in the audio. Although Whisper's authors gave little to no details about

the data source used to train the model (Radford et al., 2022), the observation of such flaws leaves little doubt about how the authors gathered 680k hours of audio.

Human transcription	Automatic transcription
is baby crying	are you crying
why do not you go make him something	are you going to make him do that
hairy dog	there we go
and a soft rabbit	you know it is all together
a velvet mouse	i love it you know
and a furry cat	hurry up
a funky frog	bye bye
you are okay	bye bye
funny frog	bye
sh	bye
sh sh sh	the end
oh who came	oh he is here
you are okay	okay
you are okay	ah
you are okay	ah

Tab. 1.2.: Example of a 2-minute session extracted from American English long-forms for which Whisper obtains a word error rate of 89.6%. Insertions and substitutions are indicated in red.

1.1.2 Challenges and opportunities in building automatic speech recognition systems for long-forms

Child-centered long-form recordings offer a unique ecological view of children's language environment and allow for a big-data approach to the study of language acquisition (Casillas & Cristia, 2019; Gautheron et al., 2020; Cychosz & Cristia, 2021). This comes at the price of handling the complexity of everyday language use: through tedious, time-consuming, and expensive human-made transcription or the development of specialized automatic speech recognition systems.

This section demonstrated Whisper's failure to transcribe American English long-forms. However, it would not be fair to interpret this failure as a lack of progress in automatic speech recognition. There have been breakthroughs – and I consider Whisper one of them. The word error rate is constantly lowering (Baevski et al., 2020; Hsu et al., 2021; Radford et al., 2022; Y. Zhang et al., 2023), bringing us closer to the day when language development researchers will have a fully functional ASR model for long-forms. As of now, it appears evident that transcribed speech found on the

internet and current ASR training sets fail to address the complexity of naturalistic adult and child speech, such as found in long-forms. A direct consequence is that ASR tools developed with their own commercial or research agenda would likely not work on long-forms. Therefore, we are left with one possibility: building our own tools.

We can do so by leveraging existing sharing platforms of child language data. Some of these platforms include the Child Language Data Exchange System (CHILDES), which focuses on child language data in general (MacWhinney, 1996), PhonBank, which focuses on child phonology (Rose & MacWhinney, 2014), and HomeBank, which provides access to multi-hour and real-world recordings like long-forms (VanDam et al., 2016). Although only a small portion of the data is transcribed speech, it can be used to fine-tune existing ASR tools like Whisper, which would likely yield a significant performance boost on long-forms.

While our analysis focused solely on American English, it is crucial to remember that children develop language in diverse cultural and linguistic environments (Scaff, 2019). Therefore, building tools for the more than 7,000 languages that exist worldwide is essential to understand the similarities and differences between the wide variety of children's language experiences (Kidd & Garcia, 2022). The development of inclusive AI tools for low-resource languages, including ASR models, is an active area of research (Besacier et al., 2014; Reitmaier et al., 2022) that will likely be key in building a universal description of children's language experiences.

1.2 The Language ENvironment Analysis (LENA®) software

In addition to providing a wearable recording device capable of capturing up to 16 hours of audio (presented in the Introduction), the LENA® Foundation proposes a proprietary software designed to provide automated quantitative analyses of both the child's vocalizations and their language environment. Since its public release in 2008, the LENA® system has become the industry and research standard for measuring language acquisition in young children. For instance, Warren et al. (2010) used the LENA® technology to compare the vocal production and language environment of a sample of American English children with autism spectrum disorder (ASD) to that of typically-developing children. They report the same amount of adult speech across the two groups but 29% fewer vocalizations and 26% fewer conversational turns for children with ASD than their typically-developing counterparts. Adopting a similar

methodology, Gilkerson et al. (2017) found that, on average, American children from lower socioeconomic status (SES) families produced fewer vocalizations and were exposed to less adult speech than their higher SES peers. The LENA® technology has also been used to measure the effects of language remediation programs by comparing the quantity of speech produced and overheard by children before and after the remediation (Weil & Middleton, 2010; Ota & Austin, 2013; Sacks et al., 2014).

This section aims to provide readers with scientific and technical information on the types of measures extracted by the LENA® software and the models deployed to extract them. Next, we outline the efforts of the language development community to evaluate the system. We conclude by identifying three limitations of the system: 1) its closed-source license; 2) its aging technology; and 3) the fact that it has been optimized for American English only.

Most information concerning the LENA® system was obtained from the various technical reports published by the foundation. See Gilkerson and Richards (2020) for an overview of the system; D. Xu, Yapanel, Gray, et al. (2008) and D. Xu, Yapanel, and Gray (2008) for the models and their performance; Gilkerson and Richards (2008) for the data collection procedure; and Gilkerson et al. (2008) for the annotation procedure.

1.2.1 Extracted measures and models

The full speech processing pipeline, depicted in Figure 1.3, can be described as follows:

1. First, the audio is segmented into broad speaker and non-speaker categories. Categories include female adult speech, male adult speech, key child vocalizations, other child vocalizations, electronic speech, overlapping speech, noise, and silence (the last three are not represented in Figure 1.3). This step is performed using a minimum duration Gaussian mixture model (MDGMM) fed with 36 mel-frequency cepstrum coefficients (MFCCs). A second-pass classification is performed to detect faint sounds via a likelihood-ratio test, i.e., the likelihood of any non-silence segment is compared with the silence-likelihood of the same segment. The segment is considered faint if the ratio is lower than a threshold tuned on a small test set. Detected faint sounds are filtered out from subsequent analyses.

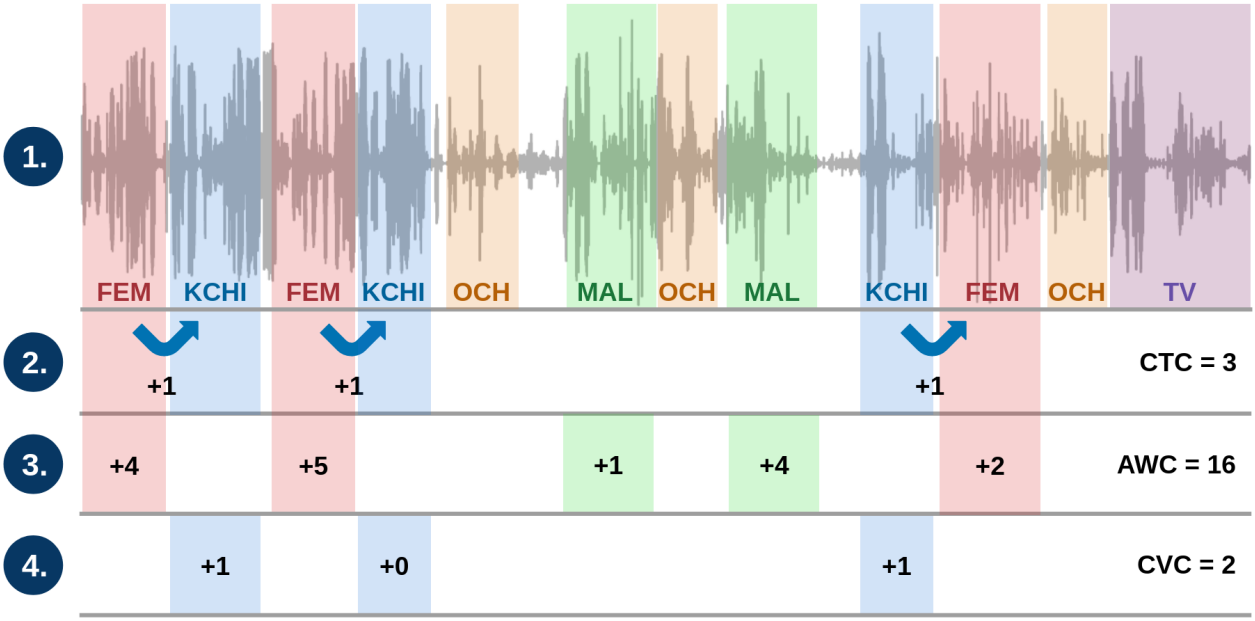

Fig. 1.3.: The LENA® speech processing software. First, the audio recording is segmented into broad speaker and non-speaker categories (FEM: female adult speech; MAL: male adult speech; KCHI: key child vocalization, i.e., the child wearing the microphone; OCH: other child vocalization; ELE: electronic speech). Second, the conversational turn count (CTC) is computed as the number of times an adult speaks and the key child follows with no more than 5 seconds in between. Third, the adult word count (AWC) is estimated as the number of words produced by MAL and FEM categories. Fourth, the child vocalization count (CVC) is estimated as the number of times the key child produces speech-like vocalizations (here, the second vocalization produced by KCHI is a cry and is discarded from the CVC).

2. Second, the conversational turn count, or CTC, is computed as the number of times an adult speaks and the key child follows (or vice versa) with no more than 5 seconds in between.

3. Third, adult speech segments are further processed to estimate the number of words in each segment, resulting in the adult word count (AWC) metric. The AWC is estimated using a least-square linear regression based on the sequence length, the number of consonants and vowels, with the last two variables obtained using the Sphinx phone recognizer optimized for American English adult speech (Lamere et al., 2003).

4. Fourth, segments produced by the key child are further classified into vegetative sounds (e.g., breathing or burping), fixed signals (e.g, crying, screaming, laughing), and speech-like vocalizations. The child vocalization count, or CVC, is estimated as the number of times a speech-like vocalization is encountered. Little information is available about how this classification is performed,

however, D. Xu, Yapanel, Gray, et al. (2008) suggests an approach based on low-level acoustic features discriminative of the type of sounds produced by the child combined with phone-level information (using the same Sphinx phone recognizer as used to estimate AWC).

1.2.2 Training and test data

The LENA® Foundation created a large-scale corpus of long-form recordings across over 300 American infants growing up in monolingual English-speaking families in the Denver metropolitan area. Families were selected to ensure diversity in age (2-48 months) and socioeconomic contexts. In total, 32,000 hours of long-forms were collected (Gilkerson & Richards, 2008). This large-scale corpus has been partially annotated and split into a 155-hour-long training set and a 70-hour-long test following the procedure described in Gilkerson and Richards (2008) and Gilkerson et al. (2008).

1.2.3 Evaluation of the LENA® system

It is essential to establish the reliability of the measures extracted by the LENA® system and document potential biases that may arise when using the system in recording conditions that differ from the ones it was trained on. This is especially critical as the data used to train the models were collected exclusively in American English-speaking families, yet, the system has been used in diverse linguistic, socioeconomic, and cultural contexts (Ganek & Eriks-Brophy, 2018).

Besides internal validation proposed by the designers of the LENA® Foundation (D. Xu, Yapanel, & Gray, 2008), researchers have attempted to characterize the system's reliability in different settings. In a recent meta-analysis, Cristia et al. (2020) found 33 studies reporting on the accuracy of the measures extracted by the LENA® software. While the majority of these studies (N=18) focused on North American English families, the remaining 15 studies included children learning UK English, US Spanish, Dutch, Finnish, Mandarin and Shanghai Chinese, Tsimane' or other languages. Only a few studies (N=8) considered settings matching the LENA® training set. In contrast, the remaining 25 studies included populations not represented in the original training set, including children: diagnosed with or at risk for autism spectrum disorder, of low socioeconomic status, bilingual, etc. Overall, authors report a significant correlation between LENA® AWC and its human-transcribed estimate (average Pearson's R of .79 with N=13 studies) while also

noting that LENA® AWC measures tend to over-estimate their human counterpart (average relative error rate (RER) of 13.8% with N$=14$). The performance obtained by the LENA® software is similar for the CVC measure (average Pearson's R of .77 with N$=5$) with a stronger tendency for under-estimation (average RER of -24.2% with N$=6$). The performance is lower for CTC for which the authors found an average Pearson's R of .36 (N$=6$) with a strong tendency for underestimation (average RER of -34% with N$=4$).

Studies dedicated to specific populations (and their meta-analysis) can inform us about the reliability of the LENA® technology in particular settings. However, these studies vary significantly in design, which can result in widely varying performance measures. Additionally, the number of these studies is still relatively small, making it challenging to detect potential biases of the LENA® algorithm, i.e., systematic errors that may occur in specific populations. In a study that I co-authored (Cristia, Lavechin, et al., 2019)[3] as part of the Analyzing Child Language Experiences around the World project (see "The ACLEW project", 2023), we collected, annotated, and standardized over 800 short clips extracted from child-centered long-forms. These clips amount to 20 hours of audio across five different corpora. Three of these corpora were based on North American children aged 3-36 months, the population for whom the LENA® system was initially designed. One corpus was based on a different dialect of English (UK English), for which we predicted slightly lower performance. And the remaining corpus was collected in a different linguistic and socio-cultural context involving Tsimane' learners aged 15-59 months, living in Northern Bolivia, a rural setting with large families. For this last corpus, we predicted degraded performance, particularly in terms of AWC which the LENA® algorithm estimates using an American English phone recognizer. Although our statistical analyses do not indicate that performance is worse for children who differ from the LENA® original training set, which is encouraging for researchers studying under-represented communities, further follow-up studies with greater statistical power are necessary to confirm this finding. There exists at least one counter-evidence in Räsänen et al. (2019) who found a lower AWC performance on non-American English languages. Finally, Cristia, Lavechin, et al. (2019) comes with a set of recommendations on what research questions can be addressed using the LENA® system and how to evaluate its reliability.

[3]Note that Cristia, Lavechin, et al. (2019) was one of the 33 studies included in the meta-analysis previously mentioned (Cristia et al., 2020).

1.2.4 Limitations of the LENA® system

This section revealed three important limitations of the LENA® system. First, it is a closed-source black box, making it challenging to access information beyond what is provided in the various LENA® technical reports. Second, the system relies on speech processing technologies developed in the early 2000s (MDGMM, Sphinx phone decoder, etc.), and with two decades of progress, one might reasonably question whether the performance could benefit from an update. Third, the LENA® system has been optimized solely for American English and may not work as well on other languages. This bias is particularly evident in the training set and is further reinforced in the LENA® AWC estimate (step 3. of Section 1.2.1), which relies on an American English phone recognizer. It is possible that the second-pass classification over key-child segments (step 4. of Section 1.2.1) also suffers from this same bias, although I could find little information regarding its implementation – which brings us back to the first limitation.

1.3 An open-source alternative to the LENA® speech processing pipeline

The previous section introduced the LENA® software, employed in over 100 peer-reviewed articles involving more than 10,000 children/families ("LENA 15^{th} birthday", 2023). Additionally, we highlighted three important limitations of the software: its closed-source nature, its aging technology, and its inherent bias toward American English.

This section will present our attempt to propose a free, open-source, and more accurate alternative to the LENA® pipeline. In particular, Section 1.3.1 will put forward an alternative approach to the LENA® segmentation algorithm, while Section 1.3.2 will present an alternative approach to estimate adult word counts. Most of the work presented below was done as part of the ACLEW project gathering speech processing and language development experts worldwide and constitutes the starting point of my Ph.D. thesis.

1.3.1 Segmentation into broad speaker categories

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., Cristia, A. (2020) An open-source voice type classifier for child-centered daylong recordings. *Interspeech*

Motivation

The automatic segmentation of adult and child speech (depicted in Figure 1.4) enables language development researchers to study the quantity of speech produced or overheard by children and their possible variations across different populations. Such technology can also help diagnose language delays or disorders early and measure the impact of language remediation programs.

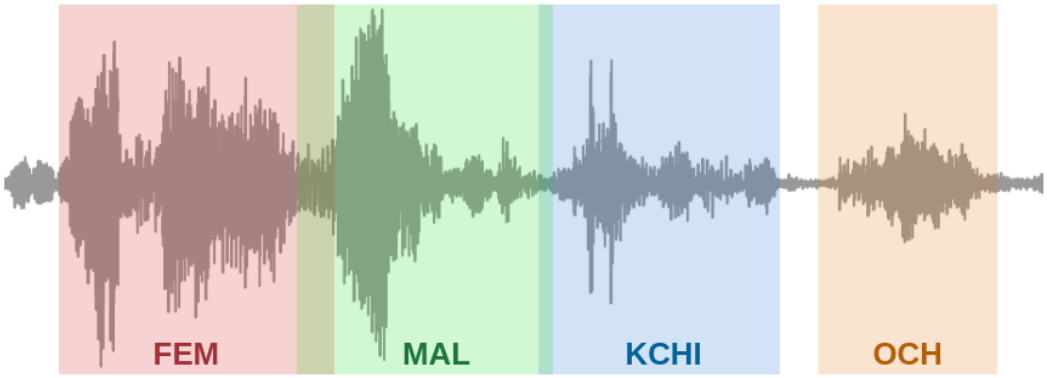

Fig. 1.4.: Voice type classification is the task of identifying voice signal sources in an audio stream. In this example, FEM stands for female adult speech, MAL stands for male adult speech, KCHI stands for vocalizations produced by the key child (wearing the microphone), and OCH stands for vocalizations produced by other children in the environment.

On the speech-processing side, detecting vocal activity segments is often the earliest building block of any speech processing pipeline (see the LENA® pipeline in Figure 1.3). Detected segments can be used as input for downstream tasks such as estimating the number of words produced by adult speakers (Räsänen et al., 2021), classifying whether speech is directed towards an adult or a child (Schuller et al., 2017), and many other tasks relevant for language development research.

By and large, the automatic segmentation of adult and child speech on long-forms is performed using the LENA® proprietary software, developed with early 2000s technology ("LENA release notes", 2023) and optimized exclusively for American

English. Our primary goal in the work presented below was to develop a free, open-source, and more accurate alternative to LENA®'s segmentation algorithm. It was accepted as a proceeding in the Interspeech 2020 conference.

Paper summary

In Lavechin et al. (2020), we propose a model that classifies audio segments into: 1) vocalizations produced by the key child, i.e., the child wearing the recording device (KCHI); 2) vocalizations produced by other children (OCH); 3) adult male speech (MAL); and 4) female adult speech (FEM).

To train our model, we gathered and standardized 260 hours of human-annotated child-centered long-form recordings covering 10 languages. A comparison with the LENA® system reveals that our model performs better. The most notable improvements are obtained in detecting female adult speech (20.8% absolute gain in terms of F-score) and in detecting key-child vocalizations (13.8% absolute gain in terms of F-score). Future work might address the limited amount of training data for the MAL and OCH categories, which resulted in relatively low performance, although still higher than LENA®'s corresponding categories.

Contrary to the LENA® segmentation algorithm, our model can detect overlapping speech, in which case, two speaker categories are activated at the same time. Although our model does not detect electronic speech since our training set was not annotated for this category, additional analyses have revealed that most frames classified as electronic speech by human annotators do not activate any speaker categories in our model. This behavior is desirable for most researchers working on language acquisition as electronic speech is thought not to affect learning outcomes (Kuhl, 2016). See "Voice type classifier: Follow-up analysis", 2023 for performance on electronic and overlapping speech.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

An open-source voice type classifier for child-centered daylong recordings

Marvin Lavechin[1,2], *Ruben Bousbib*[1,2], *Hervé Bredin*[3], *Emmanuel Dupoux*[1,2], *Alejandrina Cristia*[1]

[1]ENS-PSL/CNRS/EHESS, Paris, France;
[2]INRIA Paris, France;
[3]LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France.

marvinlavechin@gmail.com, alecristia@gmail.com

Abstract

Spontaneous conversations in real-world settings such as those found in child-centered recordings have been shown to be amongst the most challenging audio files to process. Nevertheless, building speech processing models handling such a wide variety of conditions would be particularly useful for language acquisition studies in which researchers are interested in the quantity and quality of the speech that children hear and produce, as well as for early diagnosis and measuring effects of remediation. In this paper, we present our approach to designing an open-source neural network to classify audio segments into vocalizations produced by the child wearing the recording device, vocalizations produced by other children, adult male speech, and adult female speech. To this end, we gathered diverse child-centered corpora which sums up to a total of 260 hours of recordings and covers 10 languages. Our model can be used as input for downstream tasks such as estimating the number of words produced by adult speakers, or the number of linguistic units produced by children. Our architecture combines SincNet filters with a stack of recurrent layers and outperforms by a large margin the state-of-the-art system, the Language ENvironment Analysis (LENA) that has been used in numerous child language studies.

Index Terms: Child-Centered Recordings, Voice Type Classification, SincNet, Long Short-Term Memory, Speech Processing, LENA

1. Introduction and related work

In the past, language acquisition researchers' main material was short recordings [1] or times of in-person observations [2]. However, investigating the language phenomenon in this manner can lead to biased observations, potentially resulting in divergent conclusions [3]. More recently, technology has allowed researchers to efficiently collect and analyze recordings over a whole day. By the combined use of a small wearable device and speech processing algorithms, one can get meaningful insights of children's daily language experiences. While daylong recordings are becoming a central tool for studying how children learn language, a relatively small effort has been made to propose robust and bias-free speech processing models to analyze such data. It may however be noticed that some collaborative works that benefit both the speech processing and the child language acquisition communities have been done. In particular, we may cite Homebank, an online repository of daylong child-centered audio recordings [4] that allow researchers to share data more easily. Some efforts have also been made to gather state-of-the-art pretrained speech processing models in DiViMe [5], a user-friendly and open-source virtual machine. Challenges and workshops using child-centered recordings [6, 7], also attracted the attention of the speech processing community. Additionally, the task of classifying audio events has often been addressed in the speech technology literature. In particular, the speech activity detection task [8] or the acoustic event detection problem [9, 10] are similar to the voice type classification task we address in this paper.

Given the lack of open-source and easy-to-use speech processing models for treating child-centered recordings, researchers have been relying, for the most part, on the Language ENvironment Analysis (LENA) software [11] to extract meaningful information about children's language environment. This system will be introduced in more detail in the next section.

1.1. The LENA system

The LENA system consists of a small wearable device combined with an automated vocal analysis pipeline that can be used to study child language acquisition. The audio recorder has been designed to be worn by young children as they go through a typical day. In the current LENA system, after a full day of audio has been been captured by the recorder, the audio files are transferred to a cloud and analyzed by signal processing models. These latter have been trained on 150 hours of proprietary audio collected from recorders worn by American English-speaking children. The speech processing pipeline consists of the following steps [11, 13, 14]:

1 First, the audio is segmented into mutually exclusive categories that include: key child vocalizations (i.e., vocalizations produced by the child wearing the recording device), adult male speech, adult female speech, other child speech, overlapping sounds, noise, and electronic sounds.

2 The key child vocalization segments are further categorized into speech and non-speech sounds. Speech encompasses not only words, but also babbling and pre-speech communicative sounds (such as squeals and growls). Child non-speech sounds include emotional reactions such as cries, screams, laughs and vegetative sounds such as breathing and burping.

3 A model based on a phone decoder estimates the number of words in each adult speech segment.

4 Further analyses are performed to detect conversational turns, or back and forth alternations between the key child and an adult.

This work was performed using HPC resources from GENCI-IDRIS (Grant 2020-A0071011046). It also benefited from the support of ANR-16-DATA-0004 ACLEW (Analyzing Child Language Experiences collaborative project), ANR-17-CE28-0007 (LangAge), ANR-14-CE30-0003 (MechELex), ANR-17- EURE-0017 (Frontcog), ANR-10-IDEX-0001-02 (PSL), ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award.

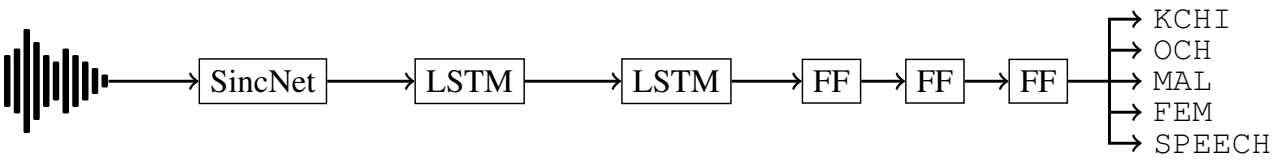

Figure 1: *Proposed architecture. The network takes the raw waveform of a 2s audio chunk as input and passes on to SincNet [12]. The low-level representations learnt by SincNet are then fed to a stack of two bi-directional LSTMs, followed by three feed-forward layers. The output layer is activated by a sigmoid function that returns a score ranging between 0 and 1 for each of the classes.*

The LENA system has been used in multiple studies covering a wide range of expertise including a vocal analysis of children suffering from hearing loss [15], the assessment of a parent coaching intervention [16], and a study of autism spectrum disorders [17]. An extensive effort has been made to assess the performance of the LENA speech processing pipeline [18, 19, 20].

Despite its wide use in the child language community, LENA imposes several limiting factors to scientific progress. First, as their software is closed source, there is no way to build upon their models to improve performance, and we cannot be certain about all design choices and their potential impact on performance. Moreover, since their models have been trained only on American English-speaking children recorded with one specific piece of hardware in urban settings, the model might potentially be overfit to these settings, with a loss of generalization to other languages, cultures, and recording devices.

1.2. The present work

Our work aims at proposing a viable open-source alternative to LENA for classifying audio frames into segments of key child vocalizations, adult male speech, adult female speech, other child vocalizations, and silence. The general architecture is presented in 2.1. Additionally, we gathered multiple child-centered corpora covering a wide range of conditions to train our model and compare it against LENA. This data set is described in further details in 2.2.

2. Experiments

2.1. End-to-end voice type classification

The voice type classification problem can be described as the task of identifying voice signal sources in a given audio stream. It can be tackled as a multi-label classification problem where the input is the audio stream divided into N frames $\boldsymbol{S} = \{s_1, s_2, \ldots, s_N\}$ and the expected output is the corresponding sequence of labels $\boldsymbol{y} = \{\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_N}\}$ where each $\boldsymbol{y_i}$ is of dimension K (the number of labels) with $y_{i,j} = 1$ if the j^{th} class is activated, $y_{i,j} = 0$ otherwise. Note that, in the multi-label setup, multiple classes can be activated at the same time.

At training time, fixed-length sub-sequences made of multiple successive frames, are drawn randomly from the training set to form mini-batches of size M.

As illustrated in Figure 1, these fixed-length sub-sequences are processed by a SincNet [12] that aims at learning meaningful filter banks specifically customized to solve the voice type classification task. These low-level signal representations are then fed into a stack of bi-directional long short-term memory (LSTM) layers followed by a stack of feed-forward (FF) layers. Finally, the sigmoid activation function is applied to the final output layer of dimension K so that each predicted score $\hat{y}_{i,j}$ consists of a number ranging between 0 and 1. The network is trained to minimize the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{KM} \sum_{i=1}^{M} \sum_{j=1}^{K} y_{i,j} \log \hat{y}_{i,j} + (1 - y_{i,j}) \log (1 - \hat{y}_{i,j}) \tag{1}$$

At test time, audio files are processed using overlapping sliding sub-sequences of the same length as the one used in training. For each time step t, and each class j, this results in several overlapping sequences of prediction scores, which are averaged to obtain the final score for class j. Finally, time steps with prediction scores greater than a tunable threshold σ_j are marked as being activated for the class j.

Our use case considers $K = 5$ different classes or sources which are:

- `KCHI`, for key-child vocalizations, i.e., vocalizations produced by the child wearing the recording device
- `OCH`, for all the vocalizations produced by other children in the environment
- `FEM`, for adult female speech
- `MAL`, for adult male speech
- `SPEECH`, for when there is speech

As the LENA voice type classification model is often used to sample audio in order to extract segments containing the most speech, it appeared to us that it was useful to consider a class for speech segments produced by any type of speaker. Moreover, in our data set, some of the segments have been annotated as `UNK` (for unknown) when the annotator was not certain of which type of speaker was speaking (See Table 1). Considering the `SPEECH` class allows our model to handle these cases.

One major design difference with the LENA model is that we chose to treat the problem as a multi-label classification task, hence multiple classes can be activated at the same time (e.g., in case of overlapping speech). In contrast, LENA treats the problem as a multi-class classification task where only one class can be activated at a given time step. In the case of overlapping speech, LENA model returns the `OVL` class (which is also used for overlap between speech and noise). More details about the performance obtained by LENA on this class can be found in [18].

2.2. Datasets

In order to train our model, we gathered multiple child-centered corpora data [21, 22, 23, 24, 25, 26, 27, 28, 29, 30] drawn from various child-centered sources, several of which were not day-long. Importantly, the recordings used for this work cover a

Table 1: *Description of the BabyTrain data set. Child-centered corpora included cover a wide range of conditions (including different languages and recording devices). ACLEW-Random is kept as a hold-out data set on which LENA and our model are compared. DB correpond to datasets that can be found on Databrary, HB the ones that can be found on Homebank.*

					Cumulated utterance duration				
Corpus	Access	LENA-recorded?	Language	Tot. Dur.	KCHI	OCH	MAL	FEM	UNK
			BabyTrain						
ACLEW-Starter	mixture (DB)	mostly	Mixture	1h30m	10m	5m	6m	20m	0m
Lena Lyon	private (HB)	yes	French	26h51m	4h33m	1h14m	1h9m	5h02m	1h0m
Namibia	upon agreement	no	Ju|'hoan	23h44m	1h56m	1h32m	41m	2h22m	1h01m
Paido	public (HB)	no	Greek, Eng., Jap.	40h08m	10h56m	0m	0m	0m	0m
Tsay	public (HB)	no	Mandarin	132h02m	34h07m	2h08m	10m	57h31m	28m
Tsimane	upon agreement	mostly	Tsimane	9h30m	37m	23m	11m	28m	0m
Vanuatu	upon agreement	no	Mixture	2h29m	12m	5m	5m	9m	1m
WAR2	public (DB)	yes	English (US)	50m	14m	0m	0m	0m	9m
			Hold-out set						
ACLEW-Random	private (DB)	yes	Mixture	20h	1h39m	45m	43m	2h48m	0m

wide range of environments, conditions and languages and have been collected and annotated by numerous field researchers.

We will refer to this data set as BabyTrain, of which a broad description is given in Table 1.

We split the BabyTrain data set into a training, development and test sets, containing approximately 60%, 20% and 20% of the audio duration respectively. We applied this split such that files associated to a given key child were included in only one of the three sets, splitting children up within each of the 8 corpora of BabyTrain. The only exception was WAR2, too small to be divided, and therefore put in the training set in its entirety.

In order to ensure that our models generalize well enough to unseen data, and to compare the performance with the LENA system, we kept the ACLEW-Random as a hold-out data set.

2.3. Evaluation metric

For each class, we use the F-measure between precision and recall, such as implemented in `pyannote.metrics` [31] to evaluate our systems:

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where $\text{precision} = \text{tp}/(\text{tp} + \text{fp})$ and $\text{recall} = \text{tp}/(\text{tp} + \text{fn})$ with:

- tp the duration of true positives
- fp the duration of false positives
- fn the duration of false negatives

We select our models by averaging the F-measure across the 5 classes. Note that these 5 metrics have been computed in a binary fashion, where the predictions of our model for a given class were compared to all reference speaker turns such as provided by the human annotations (no matter if the latter were overlapping or not). In diarization studies, the choice of a collar around every reference speaker turns is often made to account for inaccuracies in the reference labels. We chose not to do so, consequently all numbers reported in this paper can be considered as having a collar equal to 0.

2.4. Implementation details

Figure 1 illustrates the broad architecture used in all experiments. For SincNet, we use the configuration proposed by the authors of the original paper [12]. All LSTMs and inner feed-forward layers have a size of 128 and use *tanh* activations. The last feed-forward layer uses a sigmoid activation function.

Data augmentation is applied directly on the waveform using additive noise extracted from the MUSAN database [32] with a random target signal-to-noise ratio ranging from 5 to 20 dB. The learning rate is set up by a cyclical scheduler [33], each cycle lasting for 1.5 epoch.

Since we address the problem in a multi-label classification fashion, multiple classes can be activated at the same time. For the reference turns, the `SPEECH` class was considered to be activated whenever one (or more) of the `KCHI`, `CHI`, `FEM`, `MAL` or `UNK` class was activated. The `UNK` class (see Table 1) corresponds to cases when the human annotator could hear that the audio contained speech or vocalizations, without being able to identify the voice source. This class does contribute in activating the `SPEECH` class, but our model does not return a score for it.

2.5. Evaluation protocol

For all experiments, the neural network is trained for 10 epochs (approximately 2400 hours of audio) on the training set. The development set is used to choose the actual epoch and thresholds $\{\sigma_j\}_{j=1}^{K}$ that maximizes the average F-measure between precision and recall across classes.

We report both the in-domain performance (computed on the test set of BabyTrain) and the out-of-domain performance (computed on the hold-out set, ACLEW-Random). We compare our model with the LENA system on the hold-out set.

3. Results

We evaluate two different approaches, one consisting of 5 models trained separately for each of the class (referred as binary), and one consisting of a single model trained jointly on all the classes (referred as multitask). At first, both in the binary and the multitask scenario, architectures shared the same set of hyper-parameters. Only the dimension of the output layer differed. Results indicated that multitask approaches were significantly better than binary ones, which seems to show that sharing weights during training helps better learn the boundaries between the different classes.

To further improve the performance of our model, we tried

Table 2: *In-domain performance in terms of F-measure between precision and recall. The "Ave." column represents the F-measure averaged across the 5 classes. Numbers are computed on the test set from which the Paido corpora has been removed. Performance on the development set are reported using small font size. We report two variants, the first one is based on 5 binary models trained separately on each of the class, the second one consists of a single model trained in a multitask fashion*

Train/Dev.	System	KCHI	OCH	MAL	FEM	SPEECH	Ave.
without Paido	binary	76.1 79.2	22.5 28.7	37.8 38.9	80.2 83.5	88.0 89.3	60.9 63.9
with Paido	multi	75.8 78.7	25.4 30.3	40.1 43.2	82.3 83.9	88.2 90.1	62.3 65.2
without Paido	multi	77.3 80.6	25.6 30.6	42.2 43.7	82.4 84.2	88.4 90.3	63.2 65.9

multiple sets of hyper-parameters (varying the number of filters, the number of LSTM and FF layers, and their size). However, no significant differences have been observed among the different architectures. The retained architecture consists of 256 filters of length L = 251 samples, 3 LSTM layers of size 128, and 2 FF layers of size 128.

Finally, removing Paido from the training and development set led to improvements on the other test domains, as well as the hold-out set, while the performance on the Paido domain remained high. Indeed, we observed a F-measure of 99 on the KCHI class for the model trained with Paido as compared to 89 for the model trained without it. This difference can be explained by a higher amount of false alarms returned by the model trained without it. The Paido domain is quite far from our target domain since it consists of laboratory recordings of words in isolation spoken by children, and thus it is reasonable to think that removing it leads to better models.

3.1. In-domain performance

Since LENA can only be evaluated in data collected exclusively with the LENA recording device and BabyTrain contains a mixture of devices, we do not report on LENA in-domain performance. Additionally, comparing performance on a domain that would have been seen during the training by our model but not by LENA would have unfairly advantaged us.

Table 2 shows results in terms of F-measure between precision and recall on the test set for each of the 5 classes. The best performance is obtained for the KCHI, FEM, and SPEECH classes, which correspond to the 3 classes that are the most present in BabyTrain (See Table 1). Performance is lower for the OCH class and MAL classes, with an F-measure of 25.6 and 42.2 respectively, most likely due to the fact that these two classes are underrepresented in our data set. The F-measure is lowest for the OCH class. In addition to being underrepresented in the training set, utterances belonging to the OCH class can easily be confused with KCHI utterances since the main feature that differentiates these two classes is the average distance to the microphone.

The multitask model consistently outperforms binary ones. When training in a multitask fashion, increases are higher for the lesser represented classes, namely OCH and MAL. Additionally, removing Paido leads to an improvement of 0.9 in terms of average F-measure on the other domains.

3.2. Performance on the hold-out data set

Table 3 shows performance of LENA, our binary variant, and our multitask variant on the hold-out data set. As observed on the test set, the model trained in a multi-task fashion shows better performance than the models trained in a binary fashion. Removing Paido leads to a performance increase of 4 points on the average F-measure.

Table 3: *Performance on the hold-out data set in terms of F-measure between precision and recall. "Ave." column represents the F-measure averaged across the 5 classes. The hold-out data set has never been seen during the training, neither by LENA, nor by our model.*

Train/Dev.	System	KCHI	OCH	MAL	FEM	SPEECH	Ave.
english (USA)	LENA	54.9	28.5	37.2	42.6	70.2	46.7
without Paido	binary	67.6	23.0	31.6	62.6	77.6	52.5
with Paido	multi	66.4	19.9	39.9	63.0	77.6	53.3
without Paido	multi	**68.7**	**33.2**	**42.9**	**63.4**	**78.4**	**57.3**

Turning to the comparison with LENA, both the LENA model and our model show lower performance for the rarer OCH and MAL classes. Our model outperforms the LENA model by a large margin. We observe an absolute improvement in terms of F-measure of 13.8 on the KCHI class, 4.6 on the OCH class, 5.6 on the MAL class, 20.8 on the FEM class, and 8.1 on the SPEECH class. This leads to an absolute improvement of 10.6 in terms of F-measure averaged across the 5 classes.

4. Reproducible research

All the code has been implemented using `pyannote.audio` [34], a python open-source toolkit for speaker diarization. Our own code, easy-to-use scripts to apply the pretrained model can be found on our GitHub repository [1], which also includes confusion matrices and a more extensive comparison with LENA. As soon as required agreements will be obtained, we plan to facilitate access to the data by hosting them on Homebank.

5. Conclusion

In this paper, we gathered recordings drawn from diverse child-centered corpora that are known to be amongst the most challenging audio files to process, and proposed an open-source speech processing model that classifies audio segments into key child vocalizations, other children vocalizations, adult male speech, and adult female speech. We compared our approach with a homologous system, the LENA software, which has been used in numerous child language studies. Our model outperforms LENA by a large margin and will, we hope, lead to more accurate observations of early linguistic environments. Our work is part of an effort to strengthen collaborations between the speech processing and the child language acquisition communities. The latter have provided data as that used here, as well as interesting challenges [6, 7]. Our paper is an example of the speech processing community returning the favor by providing robust models that can handle spontaneous conversations in real-world settings.

[1] https://github.com/MarvinLvn/voice-type-classifier

6. References

[1] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, 1995.

[2] G. Wells, "Describing children's linguistic development at home and at school," *British Educational Research Journal*, vol. 5, no. 1, 1979.

[3] E. Bergelson, A. Amatuni, S. Dailey, S. Koorathota, and S. Tor, "Day by day, hour by hour: Naturalistic language input to infants," *Developmental science*, vol. 22, no. 1, 2019.

[4] M. VanDam, A. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, P. De Palma, and B. MacWhinney, "HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings," *Seminars in Speech and Language*, vol. 37, no. 02, 2016.

[5] A. Le Franc, E. Riebling, J. Karadayi, Y. Wang, C. Scaff, F. Metze, and A. Cristia, "The ACLEW DiViMe: An Easy-to-use Diarization Tool," in *Interspeech*, 2018.

[6] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Interspeech*, 2019, pp. 978–982. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1268

[7] P. García, J. Villalba, H. Bredin, J. Du, D. Castan, A. Cristia, L. Bullock, L. Guo, K. Okabe, P. S. Nidadavolu *et al.*, "Speaker detection in the wild: Lessons learned from jsalt 2019," *Odyssey*, 2020.

[8] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *Interspeech*, 2013.

[9] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *18th European Signal Processing Conference*, 2010.

[10] M. Mulimani and S. G. Koolagudi, "Acoustic event classification using spectrogram features," in *TENCON*, 2018.

[11] D. Xu, U. Yapanel, S. Gray, and C. T. Baer, "The lena language environment analysis system: the interpreted time segments (its) file," 2008.

[12] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[13] J. Gilkerson, K. K. Coulter, and J. A. Richards, "Transcriptional analyses of the lena natural language corpus," *Boulder, CO: LENA Foundation*, vol. 12, p. 2013, 2008.

[14] H. Ganek and A. Eriks-Brophy, "The language environment analysis (lena) system: A literature review," in *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC*, 2016.

[15] M. Vandam, D. K. Oller, S. Ambrose, S. Gray, J. Richards, J. Gilkerson, N. Silbert, and M. Moeller, "Automated vocal analysis of children with hearing loss and their typical and atypical peers," *Ear and hearing*, vol. 36, 2015.

[16] N. Ferjan Ramírez, S. R. Lytle, and P. K. Kuhl, "Parent coaching increases conversational turns and advances infant language development," *Proceedings of the National Academy of Sciences*, vol. 117, no. 7, 2020.

[17] J. Dykstra Steinbrenner, M. Sabatos-DeVito, D. Irvin, B. Boyd, K. Hume, and S. Odom, "Using the language environment analysis (lena) system in preschool classrooms with children with autism spectrum disorders," *Autism : the international journal of research and practice*, vol. 17, 2012.

[18] A. Cristia, M. Lavechin, C. Scaff, M. Soderstrom, C. Rowland, O. Räsänen, J. Bunce, and E. Bergelson, "A thorough evaluation of the language environment analysis (lena) system," *Behavior Research Methods*, 2019.

[19] J. Gilkerson, Y. Zhang, J. Richards, X. Xu, F. Jiang, J. Harnsberger, and K. Topping, "Evaluating lena system performance for chinese: A pilot study in shanghai," *Journal of speech, language, and hearing research : JSLHR*, vol. 58, 2015.

[20] M. Canault, M.-T. Le Normand, S. Foudil, N. Loundon, and H. Thai-Van, "Reliability of the language environment analysis system (lena) in european french," *Behavior research methods*, vol. 48, no. 3, pp. 1109–1124, 2016.

[21] E. Bergelson, A. Warlaumont, A. Cristia, M. Casillas, C. Rosemberg, M. Soderstrom, C. Rowland, S. Durrant, and J. Bunce, "Starter-ACLEW," 2017. [Online]. Available: http://databrary.org/volume/390

[22] M. Canault, M.-T. Le Normand, S. Foudil, N. Loundon, and H. Thai-Van, "Reliability of the Language ENvironment Analysis system (LENA™) in European French," *Behavior Research Methods*, vol. 48, no. 3, Sep. 2016.

[23] H. Chung, E. J. Kong, J. Edwards, G. Weismer, M. Fourakis, and Y. Hwang, "Cross-linguistic studies of children's and adults' vowel spaces," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, Jan. 2012. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.3651823

[24] J. J. Holliday, P. F. Reidy, M. E. Beckman, and J. Edwards, "Quantifying the Robustness of the English Sibilant Fricative Contrast in Children," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, Jun. 2015.

[25] E. J. Kong, M. E. Beckman, and J. Edwards, "Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese," *Journal of Phonetics*, vol. 40, no. 6, Nov. 2012.

[26] F. Li, "Language-Specific Developmental Differences in Speech Production: A Cross-Language Acoustic Study: Language-Specific Developmental Differences," *Child Development*, vol. 83, no. 4, Jul. 2012.

[27] G. Pretzer, A. Warlaumont, L. Lopez, and E. Walle, "Infant and adult vocalizations in home audio recordings," 2018.

[28] A. Cristia and H. Colleran, "Excerpts from daylong recordings of young children learning many languages in vanuatu," 2007.

[29] C. Scaff, J. Stieglitz, and A. Cristia, "Excerpts from daylong recordings of young children learning tsimane' in bolivia," 2007.

[30] J. S. Tsay, "Construction and automatization of a minnan child speech corpus with some research findings," *IJCLCLP*, vol. 12, 2007.

[31] H. Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech*, 2017.

[32] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[33] L. N. Smith, "Cyclical learning rates for training neural networks," in *Winter Conference on Applications of Computer Vision*, 2017.

[34] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

1.3.2 Phoneme, syllable and word counts estimation

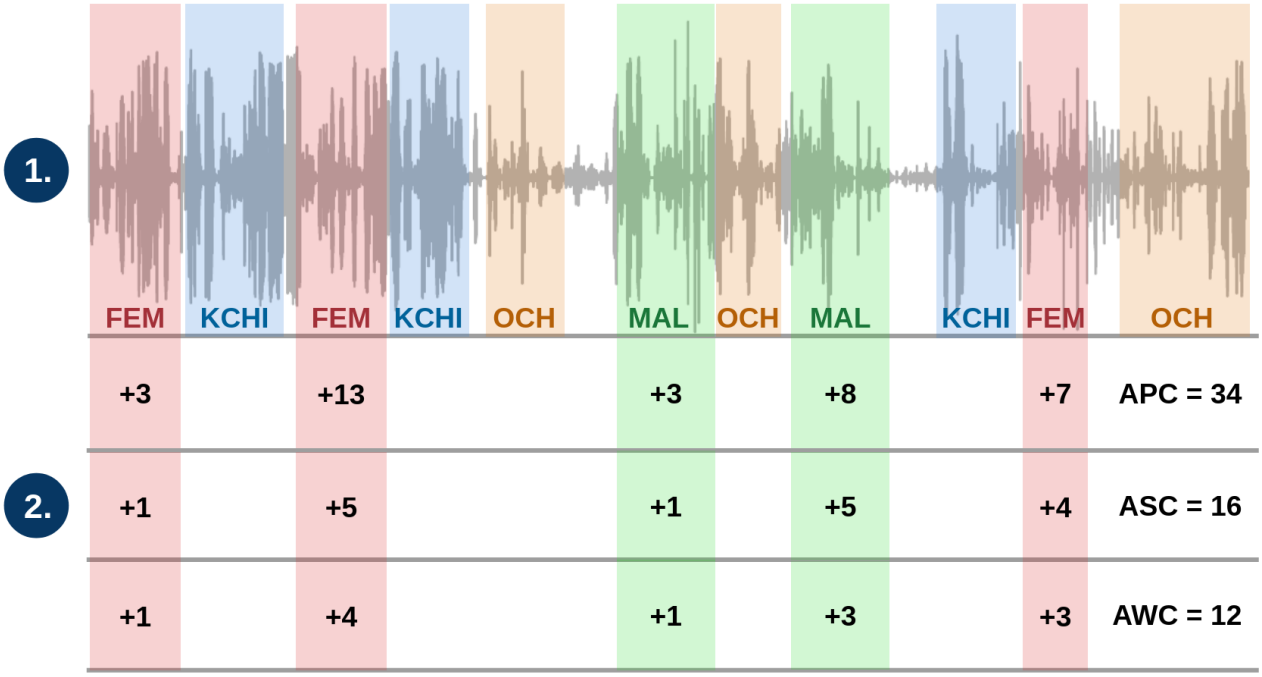

Fig. 1.5.: The ALICE model proposed in Räsänen et al. (2021). In a first stage, speech segments produced by adult speakers are identified using the voice type classifier presented in Section 1.3 (Lavechin et al., 2020). In a second stage, adult speech segments are further processed to estimate the adult phoneme count (APC), adult syllable count (ASC) and the adult word count (AWC).

In the previous section, we presented a neural network trained to identify broad speaker categories on a multilingual corpus of child-centered long-forms and showed that it outperforms the LENA® segmentation model. Can we go a step further and propose an open-source and more accurate alternative to the LENA® adult word count (AWC) estimation model? Our attempt to do so is presented in a co-authored publication in Räsänen et al. (2021), whose model is depicted in Figure 1.5.

As evidenced in Section 1.2, the LENA® speech processing pipeline has been optimized for American English. This is especially true for its AWC estimate, which relies on an American English phone recognizer. Räsänen et al. (2019) suggests that this shortcoming may translate into a lower accuracy on non-English languages. Another important matter that stands on its own is whether the word is a relevant linguistic unit to measure language input in children. Indeed, the composition of words varies greatly among human languages, with some words comprising a single morpheme and others comprising multiple morphemes – e.g., in Japanese, "食べたくなかった", tr. *tabetakunakatta* means "I/he/she/they did not want to eat (it)". With this

consideration in mind, it could be relevant to compute not only the number of words overheard by the child, but also the number of phonemes and syllables.

In Räsänen et al. (2021), we introduce ALICE for Automatic LInguistic unit Count Estimator. ALICE is a model trained to estimate the number of phonemes, syllables, and words produced by adults in child-centered long-forms. After adult segments have been detected by our voice type classifier (Lavechin et al., 2020), we extract various features from each segment. Features include: 1) the estimated number of consonants, vowels, and consonant-vowel or vowel-consonant alternations using Allosaurus (X. Li et al., 2020), a phone recognizer trained on 12 languages; 2) the estimated number of syllables using SylNet (Seshadri & Räsänen, 2019) trained on Estonian and Korean speech; 3) signal-level features such as the utterance duration, the total signal energy, and the number of waveform zero-crossings. All features are then mapped to three separate linear regressions allowing ALICE to correct under- or over-estimation in the feature extraction step. Regression parameters are estimated using the least-square method on 36 hours of annotated audio from long-forms, including Yélî Dnye, Tseltal, Argentinian Spanish, American English, Canadian English, and UK English.

We evaluated the generalization performance of our system to unseen data using a leave-one-corpus-out procedure. Results indicate that ALICE outperforms the LENA® AWC estimate on our American, Canadian, and UK English corpora. Unfortunately, LENA® automatic measures were not available for our Yélî Dnye, Tseltal, and Argentinian Spanish corpora as the latter have been collected with non-LENA® recorders and the LENA® software only accepts audio files collected using their recorder. Consequently, there remains to measure how ALICE performs compared to the LENA® on non-English languages.

Our voice type classifier that segments audio into broad speaker categories (Lavechin et al., 2020), along with ALICE that estimates the number of linguistic units produced by adult speakers (Räsänen et al., 2021), constitute an open-source alternative to steps 1., 2. and 3. of the LENA® pipeline depicted in Figure 1.3. There only remains to implement step 4. which consists in classifying vocalizations produced by the key child into vegetative sounds, fixed signals, and speech-like vocalizations. This would require a first-stage classification to extract vocalizations produced by the key child (using our voice type classifier or a similar model) and a second stage to sub-classify these vocalizations. Evaluating the accuracy of such a pipeline in real-world settings would also require summing up errors acquired in both stages. To the best of my knowledge, there is presently no available open-source pipeline that can be readily used to automatically sub-classify the key-child vocalizations

from long-form recordings, but see Z. Zhang et al. (2018), Al Futaisi et al. (2019), or Anders et al. (2020) for related work.

Throughout this chapter, we presented our contributions to developing automatic speech processing tools to analyze child-centered long-forms. These recordings collect audio in challenging environments and include near-field as well as far-field speech produced by multiple speakers, including children. Besides, the speech can be reverberated and affected by numerous sources of noise.

We close this chapter with a last contribution dedicated to a model that automatically extracts background noise and reverberation measures.

1.4 Background noise and reverberation estimation

Lavechin, M.[*], Métais, M.[*], Titeux, H., Boissonnet, A., Copet, J., Rivière, M., Bergelson, E., Cristia, A., Dupoux, E., Bredin, H. (2023) Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation. *Submitted to ASRU*

Motivation

As long-forms collect everyday language use in naturalistic settings, background noise and reverberation populate the recordings. Background noise most commonly refers to undesired sounds that impede the listeners' perception of more important sounds – speech in most cases. Background noise includes noise generated by heating or ventilating systems, appliances (vacuum cleaner, washer, dishwasher, refrigerator, etc.), outdoor traffic flow, and many others. Reverberation refers to the persistence of sounds as sound waves reflect off obstacles and hard surfaces in the environment. As these acoustic phenomena can strongly degrade the signal of interest, it is crucial to be capable of measuring them at any given moment in the recording.

On the speech-processing side, such measures can be used to evaluate the reliability of automatic tools under noisy or reverberant conditions. On the language-development side, background noise and reverberation measures can be used to assess the wide variety of listening conditions faced by infants. For instance, one could study whether background noise measures extracted from long-forms correlate

with later language development – see Erickson and Newman (2017) for a review on the influence of background noise on infant language learning.

The article presented in this section, for which we give a summary below, proposes a model that automatically extracts background noise and reverberation measures from single-channel audio recordings. It is submitted as a proceeding to the 2023 Automatic Speech Recognition and Understanding (ASRU) workshop.

Paper summary

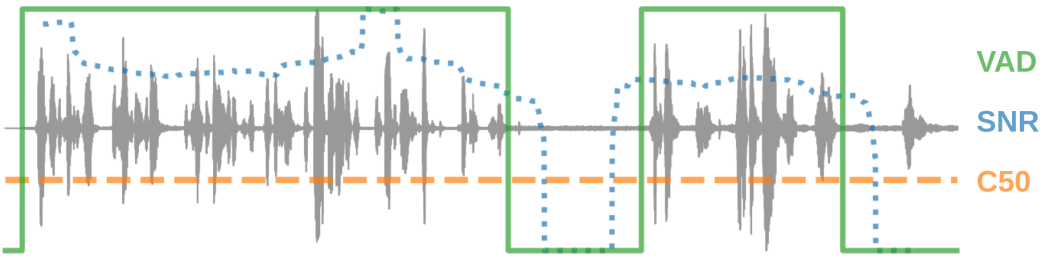

Fig. 1.6.: Background noise and reverberation estimation. Here, we want to automatically measure whether speech is noisy or reverberant in an audio stream. Our objective is to develop a single model that carries out three tasks: voice activity detection (VAD), speech-to-noise ratio (SNR) estimation, and C_{50} estimation.

In Lavechin, Métais, et al. (2022), we introduce *Brouhaha*, depicted in Figure 1.6, a model that extracts: 1) speech/non-speech segments; 2) speech-to-noise ratio (SNR), that measures the relative power of the speech signal as compared to the power of the background noise; and 3) C_{50}, also called speech clarity, that measures the extent to which the environment is reverberant. It does so from single-channel recordings and returns measures at the frame level.

In most cases, SNR and C_{50} measures are not available from the single-channel recordings of interest. Measuring the SNR would require perfectly separating the speech source from the background noise source, and measuring the C_{50}, would require retrieving the room impulse response (RIR) from which this measure is derived (see details in Section 1 of the paper). So, how can we obtain the labels required to train our model?

In this paper, we follow a data-driven approach by which we contaminate clean speech segments with additive background noise and reverberation. By doing so, we generate artificially contaminated speech segments given as input to our model. The model is then trained to predict the 'strength' of the two transformations applied

earlier, namely the SNR and C_{50} measures. Audio examples used during training are available on this this project page[4].

We conducted several experiments to validate our model, including tests on artificially contaminated audio recordings and naturally noisy and reverberant audio. Our results show that the multi-task training regime proposed in the paper improves the model's performance. One particularly relevant result in the context of this thesis manuscript is that of Section 5.6 of the paper, demonstrating that the SNR – and to a lesser extent the C_{50} – impacts the word error rate obtained by Whisper on child-centered long-forms. Using the same American English long-forms and the same ASR system as the ones presented in Section 1.1, our results reveal that Whisper accurately transcribes 83% of the words of utterances whose SNR belongs in the $[12, 23.6]$ dB range. Whisper's accuracy decreases as the SNR decreases until it successfully transcribes only 48% of the words of utterances whose SNR is in the $[-9.4, -4.2]$ dB range.

This exemplifies how *Brouhaha* can be used to assess the reliability of automatic speech processing tools under noisy or reverberant conditions.

[4] `https://marvinlvn.github.io/projects/1_project`

Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation

Marvin Lavechin[⋆,1,2], *Marianne Métais*[⋆,1], *Hadrien Titeux*[1], *Alodie Boissonnet*[2], *Jade Copet*[2], *Morgane Rivière*[2], *Elika Bergelson*[3], *Alejandrina Cristia*[1], *Emmanuel Dupoux*[1,2], *Hervé Bredin*[4]

[1] LSCP, DEC, ENS, EHESS, CNRS, PSL University, Paris, France
[2] Meta AI Research, France [3] Duke University, North Carolina, USA
[4] IRIT, Université de Toulouse, CNRS, Toulouse, France

marvinlavechin@gmail.com

Abstract

Most automatic speech processing systems register degraded performance when applied to noisy or reverberant speech. But how can one tell whether speech is noisy or reverberant? We propose Brouhaha, a neural network jointly trained to extract speech/non-speech segments, speech-to-noise ratios, and C50 room acoustics from single-channel recordings. Brouhaha is trained using a data-driven approach in which noisy and reverberant audio segments are synthesized. We first evaluate its performance and demonstrate that the proposed multi-task regime is beneficial. We then present two scenarios illustrating how Brouhaha can be used on naturally noisy and reverberant data: 1) to investigate the errors made by a speaker diarization model (pyannote.audio); and 2) to assess the reliability of an automatic speech recognition model (Whisper from OpenAI). Both our pipeline and a pretrained model are open source and shared with the speech community.

Index Terms: voice activity detection, speech-to-noise ratio, speech clarity, acoustic environment, reverberation

1. Introduction and related work

Robustness to degraded acoustic environments is a critical factor limiting the impact and adoption of speech technologies. Numerous sources of variations in the audio can degrade or hide the signal of interest and impact the performance of automatic speech processing systems. Be it automatic speech recognition (ASR) [1, 2, 3], speaker identification/diarization [4, 5], or speaker localization [6], most systems exhibit a loss of performance when applied in noisy or reverberant conditions.

While speech processing systems are being improved to handle degraded acoustic environments [7, 8, 9], little work has been devoted to automatically predict the properties of the acoustic environment. A proposed approach involves using synthetic audio generated by applying an audio transformation of interest (e.g., reverberation). A neural network is then trained to extract the 'strength' of this audio transformation. This approach is most commonly used to develop systems that predict room acoustic measures like speech clarity (C_{50}), reverberation time (T_{60}) or direct-to-reverberant ratio (DRR) [10, 11, 12, 13, 14]. In practice, these values can be estimated directly from the room impulse response (RIR, the recording of a high-energy and bursty sound, such as a pistol shot or a balloon popping). However, in most cases, RIRs are not available, and we need to estimate the values of interest from the observed single channel audio recording. A similar approach has been adopted in [15] to automatically estimate the frame-level speech-to-noise ratio (SNR). The authors evaluate the performance of their system on synthetic data, but not on real data. In practice, real SNRs are not available making it impossible to compare the estimated values to the real ones. Thus, it remains unclear if such a system can generalize to real data.

Given the high interplay between noise and reverberation (the SNR may be influenced by how noise and speech sources reverberate, and it is harder to obtain reliable estimates of reverberation parameters in low SNR conditions [16, 17]), can we design a system that tackles both tasks simultaneously? This is one of the questions we address in this work. Our approach is closest to [18] who proposes to train a neural network for jointly estimating room acoustic parameters and the utterance-level SNR. However, the authors use a restrained set of noise segments which cast doubts on the ability of their model to generalize to unseen noises. More importantly, they do not evaluate their system with respect to the SNR, and they do not address the question of whether the proposed multi-task regime is beneficial for the estimation performance.

We propose *Brouhaha*, a model jointly trained on the speech/non-speech classification task and the SNR and C_{50} regression tasks. Our model is trained on $1,250$ hours of synthetic audio generated from clean speech segments contaminated with silence, noise and reverberation. We first demonstrate that the proposed multi-task regime is beneficial and compare the performance of *Brouhaha* against state-of-the-art systems. We then apply *Brouhaha* on real data (under naturally noisy and reverberant conditions) to: 1) analyze the error patterns of a speaker diarization system (*pyannote.audio* [19]); and 2) assess the reliability of an ASR system (Whisper [20]). In addition to showing how *Brouhaha* can be used, these experiments constitute evidence that our system is applicable to real data.

Beyond the scientific interest of exploring the effectiveness of the proposed multi-task training regime and assessing the applicability of the method on real data after training on synthetic ones, we believe our work has a strong practical interest. Unlike previous work [15, 18], *Brouhaha* can be applied to any audio regardless of whether it contains speech, non-speech or both. By using our system, there is no requirement to implement a preliminary voice activity detection system prior to obtaining SNR and C50 values. We believe such advancement, in addition to a simple user interface (one python command!), significantly aids empowering researchers who may not possess expertise in speech processing or machine learning to make the most out of speech technology.

⋆ M. Lavechin and M. Métais equally contributed to this work.

This work was granted access to the HPC resources of GENCI-IDRIS under the allocation 2022-AD011012554. It also benefited from the support of ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017, ANR-19-P3IA-0001; the J. S. Mc-Donnell Foundation; and ERC ExELang grant no 101001095.

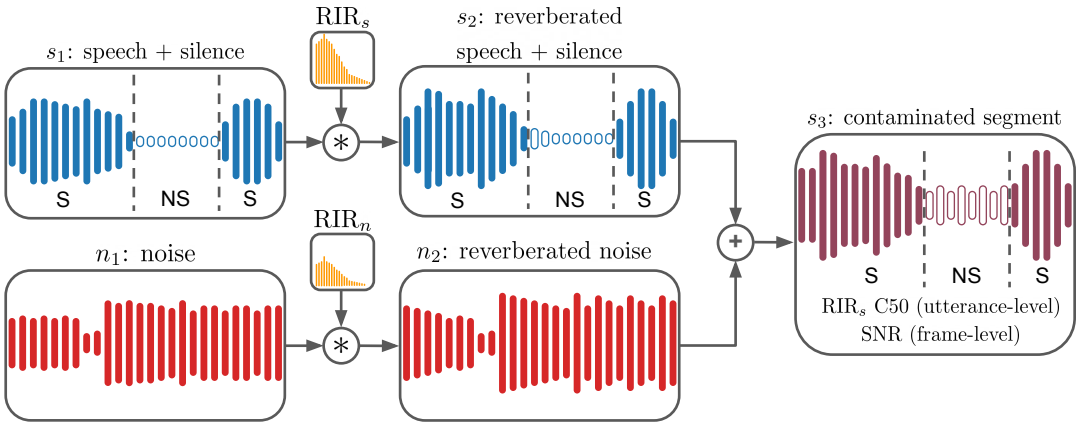

Figure 1: ***Audio contamination pipeline.*** *$s_1 \rightarrow s_2$: With probability $p_{RIR} = 0.9$, the clean speech segment (marked as S) contaminated with silence (marked as NS) s_1 is convolved with a randomly drawn impulse response RIR_s. $n_1 \rightarrow n_2$: With probability p_{RIR}, the randomly drawn noise segment n_1 is convolved with a randomly drawn impulse response RIR_n. $s_2 + n_2 \rightarrow s_3$: The reverberated speech segment s_2 and the reverberated noise segment n_2 are added together to obtain a Speech-to-Noise Ratio (SNR) randomly drawn between 0 and 30 dB. As noises can have a wide dynamic range and the utterance-level SNR captures only global information about the noise level, we recompute SNRs using a 2-second long sliding window shifted every 10 ms over s_2 and n_2. C_{50} is computed as the ratio of early (0 to 50 ms) and late (50 ms to the end of the response) energies of the room impulse response RIR_s. Labels obtained via this pipeline include: speech/non-speech (frame-level), C_{50} measure of RIR_s (utterance-level), and SNR (frame-level).*

2. Audio contamination pipeline

We start from: 1) a set of clean speech segments that will be contaminated; 2) a set of noise segments used to simulate noisy conditions; and 3) a set of RIRs to simulate reverberation. The clean speech segments are contaminated following the steps presented in Figure 1, which we will not repeat here.

3. Multi-task training

We tackled the voice activity detection problem as a classification problem where, for each 16-ms frame, the expected output is 1 if there is speech, 0 otherwise. C_{50} and SNR estimations were tackled as regression problems where, for each 16-ms frame, the expected output is the actual C_{50} or SNR in dB. We tackled the C_{50} estimation at the frame level during training – despite the label being at the utterance level – to allow the model to return smoother transitions when a change in C_{50} is detected at inference time.

At training time, short fixed length sub-sequences are drawn randomly from the training set and gradient-descent is used to minimize the multi-task loss function $\mathcal{L} = \mathcal{L}_{\text{VAD}} + \mathcal{L}_{C_{50}} + \mathcal{L}_{\text{SNR}}$, where $\mathcal{L}_{\text{VAD}}$ is the binary cross-entropy loss, and $\mathcal{L}_{C_{50}}$ and $\mathcal{L}_{\text{SNR}}$ are mean squared error (MSE) losses. Before training, $\mathcal{L}_{C_{50}}$ and $\mathcal{L}_{\text{SNR}}$ are normalized by their maximum value (computed over 10 batches) to ensure all three losses lie between 0 and 1. We computed $\mathcal{L}_{\text{SNR}}$ only over speech frames as the SNR is not defined on non-speech frames.

4. Experiments

4.1. Datasets

Our audio contamination pipeline requires three types of audio data: 1) clean speech segments; 2) noise segments; and 3) RIRs. A pretrained VAD model [19] was applied to find non-speech segments in 1000 hours of clean read-speech, retrieved from the LibriSpeech [21]. Predicted non-speech segments were extended with silence to obtain a ratio of approximately 30 % of non-speech. We used noise segments from AudioSet [22] and discarded human vocalizations. We also downsampled music segments from 38 % to 5 %, leading to a total of 1500 hours of noise segments. Finally, 385 impulse responses were obtained from EchoThief [23] and the MIT Acoustical Reverberation Scene [24] datasets. We used the same train/dev/test split originally proposed in LibriSpeech. Noise segments and impulse responses were randomly split into 80 %, 10 % and 10 % for the training, development and test set, respectively. All files used in this paper consist of 16-kHz single-channel recordings.

4.2. Evaluation metrics

We evaluated *Brouhaha* performance on the VAD task using the F-score between precision and recall, such as implemented in *pyannote.metrics* [25]. SNR and C_{50} predictions were evaluated using the mean absolute error (MAE) at the frame level. Since SNR is not defined on non-speech frames, the SNR was only evaluated across speech frames.

4.3. Architecture, optimization and training procedure

The model consists of SincNet (using the configuration in [26]), followed by a stack of bidirectional long short-term memory (LSTM) and feed-forward layers. Finally, we have three parallel layers: one classification layer (with *softmax* activation) that returns the predicted probability of speech, and two regression layers that return the predicted SNR and C_{50} (with *sigmoid* activation parametrized between -15 and 80 dB for the SNR, and -10 and 60 dB for the C_{50}).

We trained 144 different architectures across different sets of hyperparameters, varying the duration of the input sequences: 4, 6, 8, or 10 seconds; the batch size: 32, 64, or 128 sequences; the size of the hidden LSTM layers: 128 or 256; the number of LSTM layers: 2 or 3; and the dropout proportion: 0, 30 or 50 %. The best architecture was trained with 6-s segments, a batch size of 64 sequences, 3 LSTM layers of size 256, and a dropout proportion of 50 %. The best architecture

was selected on the validation metric: an average of the VAD F-score, SNR and C_{50} MAEs, with the latter two normalized by the maximum error to balance the contribution of each term.

5. Results

5.1. The effect of multi-task training

Table 1: *Performance on unseen synthetic data (our test set) in terms of F-score (VAD) and mean absolute errors (SNR and C_{50}). A checkmark below a given training task indicates that the associated loss is activated during training.*

Training tasks:			VAD	SNR	C_{50}
VAD	SNR	C_{50}	F-score (%)	MAE (dB)	MAE (dB)
✓	✓	✓	**93.7**	**4.1**	**3.5**
✓	✓		**93.7**	4.2	——
✓		✓	93.6	——	3.8
	✓	✓	——	4.3	3.7
✓			93.5	——	——
	✓		——	4.3	——
		✓	——	——	4.2

Table 1 shows performance obtained by models trained to solve either one, two or three of the proposed tasks (VAD, SNR, C_{50}). All models shared the same set of hyper-parameters, only the dimension of the output layer differed. Results indicate that the multi-task training regime is beneficial: the model trained simultaneously on the three tasks obtained better performance than models trained on two tasks which themselves obtained better performance than models trained on a single task. The largest performance gain is observed for the C_{50} estimation, with a decrease of $0.7\,\mathrm{dB}$ in terms of MAE between the single-task and the three-tasks training regime. These results seem to show that sharing weights during training helps better solve the proposed three tasks. Not only does using a single model provide a performance gain, but it is also more convenient and computationally efficient.

5.2. Voice activity detection

Table 2: *Voice activity detection F-score obtained by Brouhaha and pyannote.audio pretrained system [19]. Numbers are reported on synthetic data (our test set) and on real data (BabyTrain [27]).*

Data type	System	VAD F-score (%)
synthetic	Brouhaha (ours)	**93.7**
	pyannote.audio [19]	89.0
real	Brouhaha (ours)	77.2
	pyannote.audio [19]	**80.8**

Table 2 shows voice activity detection performance obtained by *Brouhaha* and a state-of-the-art system (*pyannote.audio* [19]). We consider two evaluation sets: 1) our test set made of unseen synthetic audio data (referred as 'synthetic' in the table); and 2) BabyTrain [27], a corpus of highly naturalistic child-centered recordings (referred as 'real' in the table). Specifically, BabyTrain recordings are acquired via child-worn microphones as they go about their everyday activities and are widely used in language acquisition research [28]. Child-centered recordings are notoriously challenging for speech processing systems as they contain spontaneous and overlapping speech, and a wide variety of noisy and reverberant conditions.

Results show a strong advantage for *Brouhaha* over *pyannote.audio* on unseen synthetic data ($4.7\,\%$ absolute difference in terms of F-score) . This indicates that, on highly noisy and reverberant synthetic audio, our system is competitive on the VAD task. Admittedly, *Brouhaha* has an advantage over *pyannote.audio* as the latter has not been trained on synthetically noisy and reverberant audio. Turning to a performance comparison on real data, numbers reveal that *pyannote.audio* outperforms *Brouhaha* by a $3.6\,\%$ absolute difference in terms of F-score. This result suggests that training a VAD system on LibriSpeech [21] contaminated with reverberation and additive noise might not be optimal, and this is despite the precautions taken in simulating challenging noisy and reverberant conditions. Nonetheless, LibriSpeech is currently the only source of clean speech available in sufficiently large quantities to run our audio contamination pipeline and obtain SNR and C_{50} labels.

5.3. Speech-to-noise ratio estimation

Table 3: *Mean absolute error on the SNR estimation task computed on unseen synthetic data (our test set). All predicted and gold SNRs are brought back to the $[-15, 30]$ dB range as done in [15]. For a given speech utterance, the heuristic estimates the noise (resp. speech) power as the mean power of non-speech (resp. speech) frames within a 6-s window centered around each annotated speech frame (defaulting to the average SNR when no non-speech frames were found within the 6-s window).*

System	SNR MAE (dB)
Brouhaha (ours)	**2.3**
Heuristic	8.4
Li et al. [15]	12.5

Table 3 shows MAE performance on the SNR estimation task computed on our test set made of unseen synthetic audio data for: 1) *Brouhaha*; 2) a heuristic using the oracle VAD that estimates the noise (resp. speech) as the mean power of neighboring non-speech (resp. speech) frames; and 3) the system proposed in [15] (a 4-layer LSTM trained from mel frequency cepstral coefficients).

Results indicate that *Brouhaha* is better at estimating the frame-level SNR than our heuristic, with an absolute difference of $6.1\,\mathrm{dB}$ in terms of MAE (note that both systems use a 6-s window as input, and that our heuristic requires oracle VAD boundaries). Surprisingly, our heuristic performs better than the system proposed in [15] with a $4.1\,\mathrm{dB}$ absolute difference in terms of MAE. This indicates that [15] struggles generalizing to unseen noise or to reverberant environments. Unfortunately, we could not compare systems on the test used in [15] as the latter has not been publicly released.

5.4. C_{50} estimation

We ran *Brouhaha* on the BUT Speech@FIT Reverb dataset [29]. This dataset consists of LibriSpeech test-clean utterances retransmitted by a loudspeaker in 5 different rooms. For each room, the speaker was placed on 5 positions on average and retransmitted utterances were recorded with 31 microphones. RIRs were measured multiple times for each

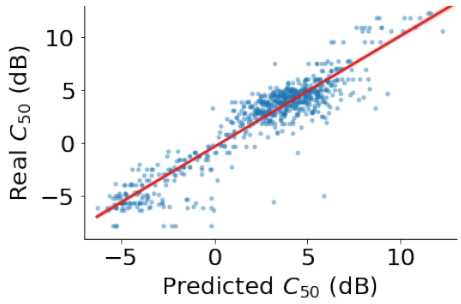

Figure 2: **C_{50} *estimation.*** *Real C_{50} against C_{50} predicted by Brouhaha on 1000 utterances from the BUT Speech@FIT Reverb dataset [29].*

speaker position. Here, we compare the real C_{50} (averaged over between 1 and 9 duplicated RIR measures) to the C_{50} predicted by *Brouhaha* on 1000 randomly drawn utterances.

Figure 2 shows a strong correlation between the real and the predicted C_{50}, with a R^2 of .85 and a mean average error of 1.1 dB. We would have liked to compare the performance of our system on the C_{50} estimation task with other systems, but we could not find any open-source pre-trained C_{50} estimators despite extensive research in this area [11, 12, 14].

5.5. Investigating speaker diarization errors

We ran a pretrained *pyannote.audio* speaker diarization pipeline [19] on the VoxConverse dataset [30] and evaluated its performance at *Brouhaha* frame resolution (16 ms). Each frame can either be classified as: 1) missed detection (when the speaker diarization pipeline incorrectly classifies a speech frame as non-speech): 2) false alarm (the other way around); 3) speaker confusion (when a speech frame is assigned to the wrong speaker); or 4) correct. Figure 3 focuses on speaker confusion (but the same pattern holds for missed detections) and shows the distribution of predicted SNR (left) and C_{50} (right) depending on whether the speech frame was assigned to the correct speaker. There is a clear trend as far as SNR is concerned: *pyannote.audio* is much more likely to confuse speakers in low (predicted) SNR regions. Similarly, the accuracy degrades significantly as we get closer to the lowest predicted C_{50} values.

Exploring the errors made by a pretrained system can provide valuable insights for developing effective strategies. In our case, one might devise strategies to address the issue of high speaker confusion in low SNR conditions: increasing the weight of low-SNR sequences in the training loss, or running speech enhancement algorithms on low SNR areas for instance.

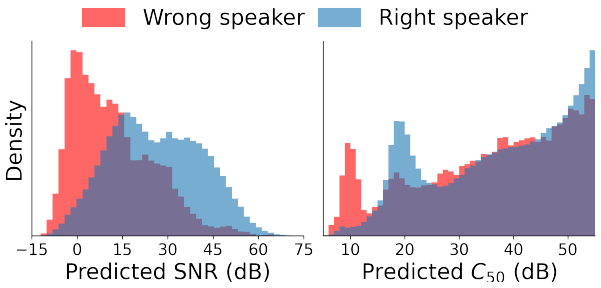

Figure 3: ***Investigating speaker diarization errors.*** *Distribution of SNR (left) and C_{50} (right) predicted by Brouhaha as a function of whether a pretrained speaker diarization system [19] assigns a speech frame to a wrong (red) or to the right speaker (blue).*

5.6. Assessing the reliability of an ASR system

We ran Whisper large ASR system [20] on highly naturalistic speech utterances from the American English Bergelson corpus [31, 32] (child-centered recordings, similar to the ones used in Section 5.2). We evaluate the performance of Whisper using the percentage hits (i.e., percentage of words correctly transcribed). We include a total of 804 utterances at least 5-words long (as short sequences most often led to a score 0 % or 100 %).

Figure 4 shows the average percentage of hits obtained by Whisper for utterances binned according to their predicted SNR (top panel) or C_{50} (bottom panel) decile. On average, Whisper correctly transcribes 83 % of the words on utterances whose SNR belongs in the $[12, 24]$ dB (last SNR decile, top panel). This number decreases as the SNR decreases until Whisper successfully transcribes only 38 % of the words on utterances whose SNR is in the $[-9, -4]$ dB range (first SNR decile). Although utterances whose predicted C_{50} is high tend to be better transcribed by Whisper, the trend with respect to the C_{50} is less clear (bottom panel). In conclusion, by using *Brouhaha*, we demonstrated the low reliability of Whisper on noisy utterances found in child-centered long-forms.

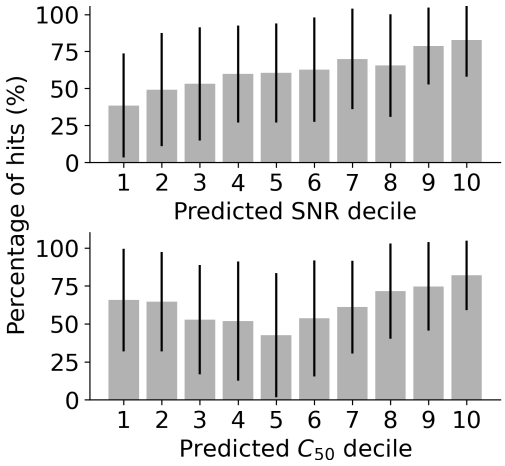

Figure 4: ***Assessing the reliability of an ASR system.*** *Percentage of hits obtained by Whisper large as a function of predicted SNR decile (top panel) and predicted C_{50} decile (bottom panel). Bars represent the percentage of hits averaged across utterances. Thin black lines represent standard errors.*

6. Conclusion and future work

We proposed *Brouhaha*, a model jointly trained on the voice activity detection, SNR, and C_{50} estimation tasks. After evaluating the performance of our system and demonstrating that the multi-task training regime is beneficial, we illustrated two use cases showing how our model can be used on real data. Beyond investigating errors made by speech processing systems or assessing their reliability in noisy and reverberant conditions, we foresee other potential downstream tasks, e.g., SNR- or C_{50}-based microphone selection [33] or SNR-aware speech enhancement [34]. Future work could explore these downstream tasks, the use of spontaneous clean speech to improve VAD performance, or the estimation of other room acoustic parameters, such as T_{60} or DRR. Both a pre-trained model and our audio contamination pipeline are shared with the community[1].

[1] https://github.com/marianne-m/brouhaha-vad

7. References

[1] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5014–5018.

[2] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, pp. 1–19, 2016.

[3] H. Gamper, D. Emmanouilidou, S. Braun, and I. J. Tashev, "Predicting word error rate for reverberant speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 491–495.

[4] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[5] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. W. Church, C. Cieri, J. Du, S. Ganapathy, and M. Y. Liberman, "The third dihard diarization challenge," in *Interspeech*, 2021.

[6] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 136–140.

[7] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[8] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.

[9] A. Narayanan, J. Walker, S. Panchapagesan, N. Howard, and Y. Koizumi, "Learning mask scalars for improved robust automatic speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 317–323.

[10] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ace challenge—corpus description and performance evaluation," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[11] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. A. Naylor, "A single-channel non-intrusive c50 estimator correlated with speech recognition performance," *Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 719–732, 2016.

[12] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1809–1820, 2018.

[13] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1–5.

[14] H. Gamper, "Blind c50 estimation from single-channel speech using a convolutional neural network," in *International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.

[15] H. Li, D. Wang, X. Zhang, and G. Gao, "Frame-level signal-to-noise ratio estimation using deep learning." in *Interspeech*, 2020, pp. 4626–4630.

[16] H. Löllmann, A. Brendel, and W. Kellermann, "Comparative study of single-channel algorithms for blind reverberation time estimation," in *International Congress on Acoustics (ICA)*, 2019.

[17] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[18] D. Looney and N. D. Gaubitch, "Joint estimation of acoustic parameters from single-microphone speech observations," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 431–435.

[19] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.audio: neural building blocks for speaker diarization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," OpenAI, Tech. Rep., 2022. [Online]. Available: https://cdn.openai.com/papers/whisper.pdf

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[23] C. Warren, "Echothief impulse response library."

[24] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[25] H. Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech*, 2017.

[26] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[27] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," in *Interspeech*, 2020.

[28] M. Lavechin, M. de Seyssel, L. Gautheron, E. Dupoux, and A. Cristia, "Reverse engineering language acquisition with child-centered long-form recordings," *Annual Review of Linguistics*, vol. 8, pp. 389–407, 2022.

[29] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[30] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the Conversation: Speaker Diarisation in the Wild," in *Interspeech*, 2020, pp. 299–303. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2337

[31] E. Bergelson, M. Casillas, M. Soderstrom, A. Seidl, A. S. Warlaumont, and A. Amatuni, "What do North American babies hear? A large-scale cross-corpus analysis." *Developmental science*, vol. 22 1, p. e12724, 2019.

[32] E. Bergelson, "SEEDLingS HomeBank corpus," https://homebank.talkbank.org/access/Password/Bergelson.html, 2017.

[33] M. Wolf and C. Nadeu, "Towards microphone selection based on room impulse response energy-related measures," in *Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal*, 2009, pp. 61–64.

[34] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement." in *Interspeech*, 2016, pp. 3768–3772.

1.5 Conclusion

In this first chapter, we discussed how artificial neural networks could provide researchers with automatic speech processing tools to analyze children's language environment captured with child-centered long-form recordings. Using a state-of-art automatic speech recognition system, we highlighted some of the challenges in processing complex and noisy real-world audio recordings and illustrated why off-the-shelf tools built with their own purposes in mind would likely not work on long-forms. Next, we presented the LENA® system, which aims to measure children's language environment and has profoundly impacted language acquisition research. We also presented our collaborative efforts in proposing a free, open-source, and more accurate alternative to the LENA® speech processing software. Lastly, we presented *Brouhaha*, a system to estimate the background noise and reverberation levels in an audio stream.

As we conclude this chapter, I would like to highlight the importance of going beyond the measures designed by LENA® researchers. I will then go on to discuss two important aspects of this line of research, namely: diversity and accessibility. This conclusion will be the opportunity to reflect on the limitations of current approaches and potential future work.

Going beyond LENA® measures. Undeniably, the LENA® system has enabled language development researchers to reach unprecedented scales and obtain a uniquely naturalistic viewpoint of language use in everyday life. Nevertheless, one potential pitfall of the widespread use of the LENA® system is the temptation to rely solely on the limited set of measures designed by its designers. With ALICE, Räsänen et al. (2021) made a step forward in extending the set of measures by introducing phoneme and syllable count estimations. Many other speech processing algorithms hold the potential to provide us with insightful measures once deployed on child-centered long-forms. Language identification, i.e., which language is being spoken and when, would be highly relevant to study children growing up in multilingual environments, e.g., Bartz et al. (2017) or Draghici et al. (2020). As suggested in G. Jones and Rowland (2017), one could also measure lexical diversity, i.e., the number of different word types produced by the caregivers, rather than absolute word count. Beyond looking at *what* speech is being delivered, looking at *how* it is delivered might also be relevant. For instance, child/adult addressee detection, i.e., classifying whether speech is directed towards an adult or a child, was proposed in the Interspeech 2017 computational paralinguistics challenge (Schuller et al., 2017). Certainly more challenging to implement, one could imagine a system to

estimate referential transparency, i.e., how easily individual caregivers' words can be identified from the surrounding linguistic or visual contexts – see Cartmill et al. (2013) for an example of how such a measure has been derived from 218 adult participants using a word masking task strikingly similar to how language models are trained in NLP.

Promoting diversity. Soon after its release, one could read in scientific publications about the promises of the LENA® system to add *objective measures* to the battery of tests used to measure language development (Richards et al., 2008; Oller et al., 2010). Has this goal been met? Regarding objectivity, neither the LENA® algorithm nor any other algorithm can be claimed to be objective. Biases arise throughout the various stages of algorithm development, including during data collection, manual annotation, and algorithm design – see Waseem et al. (2021) for a discussion on the illusion of objectivity in ML algorithms. Acknowledging the limitations of algorithms, particularly machine learning algorithms, is crucial if one wants to start documenting biases and building more inclusive algorithms. The baby step in this direction, presented in Section 1.3, is to incorporate under-represented languages into the training set. Despite more than 7000 languages being spoken worldwide, English dominates language acquisition studies, comprising 54% of current research (Kidd & Garcia, 2022). Building more inclusive speech processing algorithms will likely require a cultural shift within the research community, but see Singh et al. (2023) for guidance on making infant research more representative.

Maximizing accessibility for non-expert users. Another important aspect of this line of work concerns accessibility. Building more efficient algorithms is a laudable goal but has limited impact if targeted non-expert users (i.e., language development researchers) can not run them. My collaborator, Okko Räsänen, and I made a great effort to make our tools accessible to non-expert users. This includes providing: 1) comprehensive documentation; 2) a simple installation; 3) a one-line command to run the system; 4) assistance through emails and Github issues in case of difficulties; and 5) time to fix ever-arising installation bugs, i.e., uninteresting but essential things sometimes neglected in ML research but necessary to allow for accessible, reproducible, and open science. However, despite our efforts, there are still limitations. At the time of writing this manuscript, neither the voice type classifier nor ALICE is compatible with Windows platforms, and inference on CPUs is incredibly slow.

This brings us to an equally important matter that impedes accessibility: hardware requirement. For instance, to process a 12-hour audio recording, the voice type classifier requires 16 minutes on a single Nvidia® Tesla V100 GPU with 32 GB memory

but 3 hours on four Intel® Xeon® Gold 6230 CPUs, and language development researchers do not usually have access to GPUs.

With these considerations in mind, how can accessibility to machine learning models be improved for non-expert users? One proposal put forward in Le Franc et al. (2018) is the use of a virtual machine, which has the advantage of assuming minimal technical skills from users but fails to address the issue of GPUs accessibility. Another potential solution is the development of a cross-platform drag-and-drop user interface, which would allow researchers to deposit audio files and retrieve the output of pre-trained models with ease. However, this solution is costly to develop and would still not solve the issue of GPUs accessibility. A third solution is a GPU cloud computing platform where audio files could be uploaded (possibly encrypted), and inference could be performed on a remote server equipped with GPUs. Although this solution would possibly burden the host laboratory financially, it would enable researchers to perform fast inference with state-of-the-art machine learning algorithms.

To conclude, the large majority of the work presented in this first chapter was carried out in collaboration with various researchers from different fields. Interdisciplinarity and diversity are critical to identifying the right questions, accessing data, designing computer-readable annotation schemes, building robust tools, validating them, and making them accessible.

In this first chapter, we explored the usage of artificial neural networks to analyze children's language experiences. Another application of these networks in the realm of language development research is computational modeling. In the remainder of this manuscript, we will shift our focus to the exciting enterprise of building computational models of infant language learning.

2 Modeling language acquisition from audiobooks

There is a long tradition of modeling in the context of language acquisition. For many years, scientists spanning various fields, from formal linguistics to developmental psychology and artificial intelligence, have contemplated the prospect of running computer-based language learning simulations. Such simulations are important for both theoretical and practical reasons. On the theoretical side, simulations can help prove or disprove hypotheses – and formulate new ones – about how infants learn their native language, thus contributing to building more precise language acquisition theories (Pinker, 1979; Frank, 2011). On the practical side, language learning simulations enhance language skills in machines, allowing them to comprehend and produce language more effectively, potentially yielding, one day, machines that exhibit learning abilities on par with those of young children, as envisioned by Turing (1950).

This chapter aims to acquaint our readers with the literature on infant language development, with a specific focus on early developmental milestones and learning mechanisms in infants – *what we want to model*. We then provide a bird's eye view of the methodological landscape of language learning simulations – *how we approach the modeling process*. To conclude, we present our contribution with a model of early phonetic and lexical learning that reproduces the parallel and gradual learning observed in infants. We also reflect on important areas for future improvement and exploration.

2.1 Early language acquisition in infants

We begin by providing an approximate and simplified timeline of infants' language development. Following that, we review and illustrate three influential learning mechanisms proposed to drive early language acquisition in infants.

2.1.1 A sample of developmental milestones

Language acquisition studies have produced a wealth of results informing us on the language capabilities of young children, some of which are presented in Figure 2.1 – see Ambridge and Rowland (2013) for a review of the experimental methods.

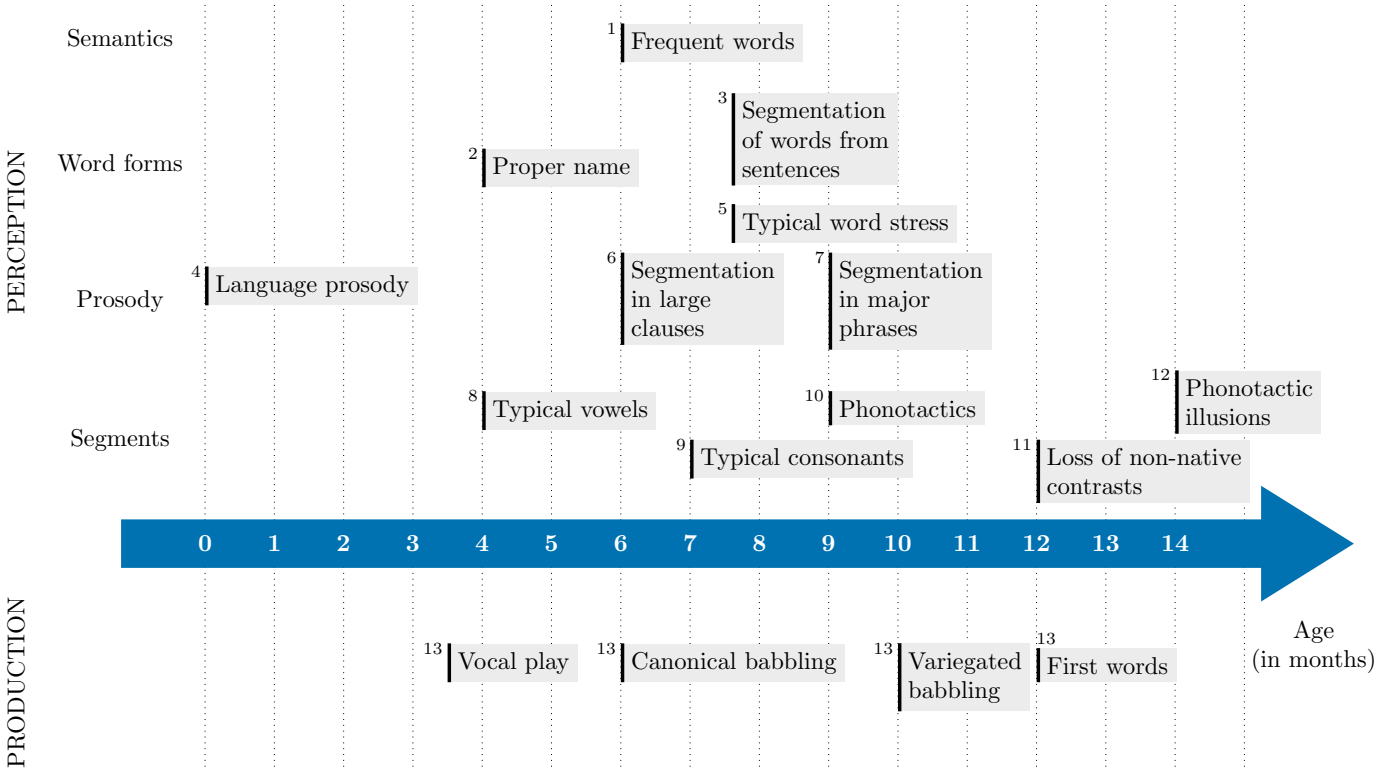

Fig. 2.1.: Sample studies illustrating the timeline of infant's language development. The left edge of each box is aligned to the earliest age at which the result has been documented. [1] Tincoff and Jusczyk, 1999; Bergelson and Swingley, 2012 [2] Mandel et al., 1995 [3] Jusczyk and Aslin, 1995 [4] Mehler et al., 1988 [5] Jusczyk et al., 1999 [6] Hirsh-Pasek et al., 1987 [7] Jusczyk et al., 1992 [8] Kuhl et al., 1992 [9] Eilers et al., 1979 [10] Jusczyk et al., 1993 [11] Werker and Tees, 1984 [12] Mazuka et al., 2011[13] Yeni-Komshian et al., 2014. Figure adapted from Dupoux, 2018.

For instance, between 6 and 12 months, infants show an improvement in discriminating sounds of their native language, while their ability to discriminate non-native sounds declines (Kuhl et al., 1992). In addition to acquiring knowledge about the sounds of their native language, infants also begin learning about words early on. Evidence for word learning starts as early as 4 months when infants begin to recognize their own names (Mandel et al., 1995). By 8 months of age, most infants recognize the auditory form of frequent words (Jusczyk & Hohne, 1997; Carbajal et al., 2021), segment words from fluent speech (Jusczyk & Aslin, 1995), and know the meanings of many common nouns (Bergelson & Swingley, 2012). Remarkably,

this knowledge of lexical and semantic aspects of their native language emerges before infants fully develop their sound discrimination abilities (McMurray et al., 2018), and before they produce their first words, typically around the end of their first year of life (Yeni-Komshian et al., 2014).

Admittedly, the picture is far more complex than depicted in Figure 2.1. First, most studies focused on English-learning infants from relatively high socioeconomic status, but developmental trajectories may change as a function of cultural or socioeconomic variables (Scaff, 2019; Christiansen et al., 2022). Second, beyond group-based differences, strong inter-individual differences have been reported – e.g., Rowe et al. (2005) and Schwab and Lew-Williams (2016); see Kidd et al. (2018) for a review. And finally, results might be confirmed or revised as new studies or meta-analyses are published, e.g., Tsuji and Cristia (2014) and Gasparini et al. (2021).

Nonetheless, the timeline depicted in Figure 2.1 illustrates the gradual and parallel trajectory of infants' language development. Instead of a stage-like developmental trajectory in which learning would unfold sequentially in a hierarchically-organized manner (i.e., from low-level to high-level linguistic structures), we observe that learning occurs simultaneously across all levels. This gradual and parallel developmental trajectory observed in infants will be the focus of Section 2.3.

For now, the developmental timeline observed in infants invites us to return to fundamental questions: How do infants learn so much in so little time? How do they effortlessly unravel the structure and rules of the intricate and hierarchical system that is language? These questions are at the core of the upcoming section.

2.1.2 Learning mechanisms

Many theories have been proposed to explain how children learn language. Although there might be disagreements in the details of the implementation or how useful each mechanism may be for solving specific problems (Johnson & Tyler, 2010; Lidz & Gagliardi, 2015; Y. Zhang et al., 2019), there is a consensus that infants are prodigious pattern finders capable of integrating cues within a single modality (e.g., the auditory stream) and across multiple modalities (e.g., between the auditory and visual streams). Here, we review three non-mutually exclusive mechanisms thought to drive early language acquisition in infants[1].

[1]More than non-mutually exclusive, these mechanisms are sometimes grouped under the term 'statistical learning', which then becomes another word for 'learning'. Here, we adopt the view whereby statistical learning primarily relates to the auditory stream, cross-referential learning

Statistical learning might be the most prominent learning theory in cognitive and developmental sciences. In the context of language acquisition, it refers to the mechanisms by which infants use "statistical properties of linguistic input to discover structure, including sound patterns, words, and the beginnings of grammar" (Saffran, 2003).

Numerous pieces of evidence corroborate the view that infants are statistical learners – see Saffran and Kirkham (2018) for a review. A first piece of evidence lies in the developmental decline in non-native sound discrimination during the first year mentioned in the previous section, which suggests that infants are sensitive to the statistical cues of speech sounds in their native language (Werker & Tees, 1984; Kuhl et al., 1992; Maye et al., 2008; Tsuji & Cristia, 2014). A second piece of evidence lies in a seminal study from Saffran et al. (1996) who showed, using an artificial language, that 8-month-olds can track transitional probabilities across syllables to identify word boundaries. Beyond phonological and word acquisition, the statistical learning hypothesis has also been proposed to account for syntactic acquisition (Seidenberg, 1997; Gomez & Gerken, 1999; Mintz et al., 2002; Solan et al., 2005).

Since its initial discovery in human infants, statistical learning has been studied in countless experiments across ages, domains, and species (Hauser et al., 2001; Kirkham et al., 2002; Saffran & Kirkham, 2018).

Cross-situational learning refers to the infants' ability to integrate cues across the auditory and visual streams. This mechanism, proposed by many authors like Pinker (1989) or Gleitman (1990), explains how infants associate words with their meanings, namely, by aggregating information from word-referent co-occurrence data. The experimental and computational evidence supporting this mechanism is reviewed by Smith et al. (2014). One example is Smith and Yu's (2008) study in which 12-month-old infants were exposed to pseudowords (e.g., 'bosa', 'kaki', etc.) paired with a particular shape. The results indicate that infants rapidly learn multiple associations between pseudowords and their corresponding shapes.

Social learning theories emphasize the role of social factors in language acquisition (Vygotsky, 1962; Bruner, 1985; Tomasello, 1992) and the importance of human interaction, including imitation and reinforcement (Skinner, 1957), joint attention (caregivers' and children's coordinated attention to each other and to a third object

involves both the auditory and visual streams and social learning involves both the auditory stream and the social context. However, it is important to note that social learning mechanisms can still incorporate visual and statistical cues, and cross-situational learning mechanisms can also integrate statistical cues.

or event, Akhtar and Gernsbacher, 2007), communicative feedback (Goldstein & Schwade, 2008), etc.

Although social learning theories have first been proposed to account for early word learning, evidence suggests that social interaction contributes to an even more basic aspect of language – the learning of sounds. A notable example comes from a study by Kuhl et al. (2003). In a first experiment, the authors found that 9-month-old American English-learning infants who spent time in laboratory sessions with native Mandarin speakers could discriminate between phonemes that occur in Mandarin. Conversely, infants in the control group exposed only to English sessions failed to discriminate the Mandarin phonemes. In a second experiment, the researchers exposed infants to the same Mandarin speakers and materials via audiovisual or audio-only recordings, with no human interaction. In this condition, infants could not discriminate the Mandarin phonemes, suggesting that exposure to recorded Mandarin with no human interaction did not affect the discrimination abilities of the infant participants. According to Kuhl's study, it appears that infants cannot integrate statistical cues from their auditory input without experiencing social interaction. To the fundamental question "What does a live person provide that a DVD cannot?", the authors reply "social cues may be critical".

Despite thousands of laboratory experiments attempting to isolate learning mechanisms in infants and thousands of hours of observations, language acquisition, much like many other cognitive processes, is essentially a black box and can only be studied as such. In other words, we cannot access the learning mechanisms infants use and can only base our theories on indirect cues: the variables that correlate with learning outcomes, the acoustic, visual, or social factors infants are sensitive to, what infants are capable of learning and at which age – much of which have been encountered in this manuscript. However, there is another method, computational modeling or language learning simulations, which we have yet to introduce but plays an important role in the study of language acquisition.

2.2 In-silico language learning simulations

The ultimate test of any language acquisition theory should be that of implementation, as advocated by Dupoux (2018). After all, if language acquisition in infants occurs through learning mechanisms M, then implementing these mechanisms should yield similar learning outcomes as the child. However, given the phenomenon's complexity, running language learning simulations that account for the various aspects

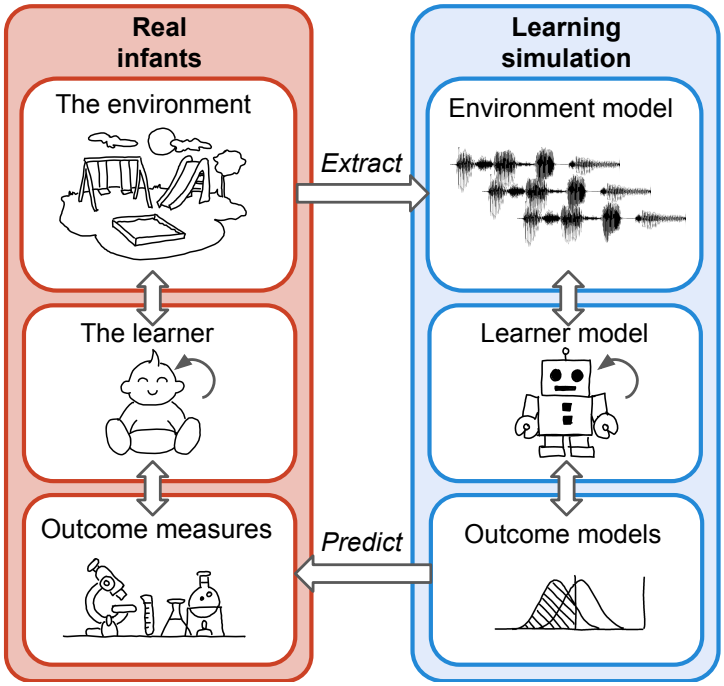

Fig. 2.2.: General outline of a learning simulation in relation to real infants. A simulation consists of 1) an environment model, which should ideally be a subset of the real environment; 2) a learner model, i.e., the mechanisms through which learning occurs in interaction with the environment; and 3) outcome models, i.e., how the learning outcomes are evaluated. The simulated learning outcomes allow us to compare humans to machines, test hypotheses and formulate predictions about how learning occurs in infants. Taken from de Seyssel, Lavechin, and Dupoux (2022).

of human languages is a challenging enterprise. Before we get an artificial learner as good as children themselves, language learning simulations can still provide proofs of learnability under the form of "Learning outcomes O can be acquired from input I using mechanism M", e.g., "Words can be segmented from strings of phonemes from co-occurrence statistics". Before we go any further, this small introductory paragraph invites us to define what we mean by language learning simulations.

As illustrated in Figure 2.2, a learning simulation can be defined as the combination of three components: 1) an environment model, i.e., the learning material available to the learner; 2) a learner model, i.e., the learning algorithm whose parameters are updated based on its interaction with the environment; and 3) a model of the outcome measures, i.e., how the language skills developed by the learner are evaluated.

Adopting this tripartite description, we propose a bird's eye view of the methodological landscape in language learning simulations. Our objective is not to propose an

in-depth literature review, which would be out of the scope of this manuscript, but to narrow down the space of possibilities in order to better outline the work done during this Ph.D. while also acknowledging the work of others.

2.2.1 The environment model: *from what is language learned?*

Regarding the environment model, there is a clear trend towards more complex and naturalistic input. In the early days, the input was kept simple in the form of, for instance, synthetic language in Elman's (1990) syntactic learning model or vowels spoken in isolation in Vallabha et al.'s (2007) simulation of phonetic learning. While written sentences are still routinely used (e.g., Bernard et al., 2020 or Stärk et al., 2022), recent developments have witnessed the emergence of models working with corpora of raw speech (Räsänen et al., 2018; Schatz et al., 2021). This will be the focus of this manuscript.

However, infants do not rely exclusively on speech to learn their native language, as suggested in Section 2.1.2. For studies exploring the contribution of the visual modality, we will refer to Alishahi and Fazly (2010) for models operating on image/caption pairs, or Räsänen and Khorrami (2019) and Nikolaus et al. (2022) for models operating on videos – see also Chrupała (2022) for a recent review. Similarly, embodied or socially grounded language learning agents have been proposed in Yu and Ballard (2003), Hermann et al. (2017), Lair et al. (2019), and Oudeyer et al. (2019).

2.2.2 The learner model: *how is language learned?*

Regarding the model of the learner, rule-based models were the first to emerge (e.g., Anderson, 1975). These models are most commonly used in the context of syntactic acquisition, with rules defining how words can be combined, either hand-crafted or learned from data. A second historical trend emerged with probabilistic or distributional models (e.g., Brent, 1996; de Marcken, 1996; Schatz et al., 2021) that are trained on a relatively large amount of data to learn distributional information of their input, whether it would be written or spoken sentences (see Chater and Manning, 2006 for a review). A third historical trend came with the emergence of connectionist models (e.g., Rumelhart and McClelland, 1986; Elman, 1990) and deep learning models, more recently, that learn complex and non-rule-like patterns from large amounts of data (see Joanisse and McClelland, 2015 for a review). Although this manuscript reflects the connectionist tradition, we do not take a stand

in favor or disfavor of any of the abovementioned approaches. In particular, no approach appears unequivocally more psychologically plausible than others[2].

2.2.3 The outcome models: *what is learned?*

Regarding the model of the outcome measures, there are multiple approaches to assess what has been learned. With my collaborators, I have already laid out most of the arguments below in de Seyssel, Lavechin, and Dupoux (2022) and Lavechin, de Seyssel, Gautheron, et al. (2022). See also Blandón et al. (2021) for an in-depth discussion of the different approaches to which we owe some of the arguments exposed here.

Evaluating models on downstream tasks is a common approach (e.g., phoneme classification accuracy, word error rate, etc.). This approach is relevant in the context of speech processing technologies and can inform us about the artificial learner's capabilities to capture patterns of their input data. However, this approach may not be relevant in the context of language acquisition modeling as the learner needs to receive supervision, e.g., in the form of phonetic or orthographic transcripts, for this evaluation to be possible.

Evaluating models against linguistic theories and abstract representations constitutes a second approach that involves assessing the presence of phonemes, words, and so forth. This approach is most commonly used to evaluate speech segmentation models by computing the proportion of retrieved boundaries, whether it would be phone (Scharenborg et al., 2007), syllable (Räsänen et al., 2018), or word boundaries (Räsänen et al., 2015; Stärk et al., 2022). The same principle is often used in phonetic category learning studies, where the learned clusters are compared to ground-truth phonetic categories, e.g., Vallabha et al. (2007).

This approach has the undeniable advantage of being easy to interpret and provides learnability proofs when a model successfully learns linguistic structures to a satisfactory degree. However, the underlying assumption is that the end goal of learning is to acquire abstract linguistic representations. This assumption can be contentious as linguistic representations are precisely the area in which theories of language (acquisition) diverge most fiercely – for phonemes, we will refer to Feldman et al. (2021) and McMurray (2022), see also Twaddell (1935) for a general criticism on inferring mental entities. In other words, such an assumption may not reflect human processing, neither in infants nor in adults. Although some may disagree, we posit

[2] See Frank's (2023) blog post explaining why psychological plausibility critiques can be harmful and how to adopt an evidence-based approach necessary to move the discussion forward.

that the evaluation of language capabilities in an artificial language learner should be theory agnostic. By not forcing extra assumptions on the learned representations on either the human or the machine learner, the findings have a higher chance of being relevant to a broad range of theories.

Evaluating models against empirical data, as proposed by Dupoux (2018), constitutes a third approach that offers the advantage of being independent of any specific linguistic theory. Indeed, rather than aiming to demonstrate the presence of linguistic representations, our objective is to compare artificial learners with human participants based on observable evidence. Specifically, we discuss two sources of data: brain measures and behavioral measures.

Evaluating models against brain measures involves comparing the activation patterns of neural networks with those of the human brain when performing a similar task (e.g., 'listening' to a story) – see Yamins et al. (2014), Millet et al. (2022), and Caucheteux et al. (2023). While this approach holds promise for gaining insights into how the human brain processes information, it has some limitations in the context of early language acquisition. First, neuroimaging devices that precisely capture brain activities in both time and space are rarely used with infants – e.g., Bosseler et al. (2021). Second, neural activities in infants are typically noisier, and studies necessitate an even larger sample size than those conducted with adults (Cusack et al., 2018; Turner et al., 2018). Therefore, a substantial accumulation of results in the infant neuroimaging literature will likely be necessary before we can use it to establish benchmarks.

Now we turn to the approach followed in this manuscript, namely, evaluating models against behavioral measures. This approach involves comparing responses returned by both machines and humans while undergoing the same psycholinguistics tasks. These tasks are designed to evaluate the subject's ability to process, comprehend, or produce language and encompass a broad range of experimental methods, including sound discrimination, auditory word form recognition, grammaticality judgment, looking-while-listening, language elicitation, etc. Common challenges with neuroimaging experiments include designing tasks that isolate the phenomenon of interest, known as *test validity*, with an appropriate signal-to-noise ratio, known as *test reliability*, both of which are crucial aspects of psychological testing (Gregory, 2004). Regarding the noise inherent to psychological testing, see Blandón et al. (2021) for a proposal to evaluate artificial learners against robust empirical data gathered through meta-analyses.

Experimental methods that probe the subject's language ability at the neural or behavioral level should ideally be administrable to machines, infants, and adults

to allow for direct human/machine comparison and account for developmental trajectories. However, this is rarely possible, especially with infants for whom age-specific test apparatus have to be constructed (high-amplitude sucking or head-turn preference procedure – see Ambridge and Rowland, 2013 for an exhaustive list). As of today, it is also not possible for the artificial learner to undergo a sound discrimination experiment in the laboratory, and specific strategies must be designed to extract the measure of interest. That is why a language learning simulation comprises a model of the outcome that should best approximate experimental methods used with human participants.

Having familiarized ourselves with *what* we aim at modeling and *how* one can approach the modeling process, we now delve into the proposed approach at the core of Chapters 2 and 3.

2.2.4 Proposed approach

The approach we adopt in Chapters 2 and 3 of this manuscript builds upon recent advances in self-supervised learning models that learn from spoken language. These models have demonstrated remarkable linguistic capabilities in various tasks, whether it involves assessing the acceptability of spoken words or sentences (T. A. Nguyen et al., 2020; Dunbar et al., 2021) or generating meaningful and coherent speech (Kharitonov et al., 2021; Lakhotia et al., 2021). By developing linguistic capacities solely from exposure to speech, without the need for human labels, these models promise to advance our understanding of how infants learn language (Lavechin, de Seyssel, Gautheron, et al., 2022; Warstadt & Bowman, 2022).

In particular, we focus on the model used in our STELA (for STatistical Learning of Early Language Acquisition) simulation, which we briefly describe below, adopting the same tripartite description as introduced above – see Section 2.3 for more details.

Our *learner model* consists of two main components: 1) an acoustic model that incorporates a Contrastive Predictive Coding (CPC) algorithm followed by a K-means algorithm, which is in charge of learning discrete representations of the audio; and 2) a language model made of Long Short-Term Memory (LSTM) layers trained on the learned discrete units. The primary learning objective is to predict future audio observations from present and past ones, a process known as auditory predictive coding at the core of the predictive brain hypothesis that has attracted the attention of the neuroscience community (Barlow et al., 1961; Keller & Mrsic-Flogel, 2018;

Hueber et al., 2020). As a statistical learning algorithm, the proposed model can provide us with insights about the aspects of language that can be acquired through statistical learning mechanisms applied to raw speech.

For this chapter, our *environment model* consists of 3,200 hours of raw, unsegmented, multi-speaker, and untranscribed speech. The speech is collected from either English or French audiobooks, simulating the environment of an English-learning or French-learning infant.

Concerning the *outcome models*, we use two behavioral probing tasks to evaluate our learner's language capacities at the phonetic and lexical levels. At the phonetic level, the evaluation consists of an ABX sound discrimination task, a protocol routinely used in psycholinguistics, e.g., Gottfried (1984) or Levy and Strange (2008a). At the lexical level, we use a spot-the-word task in which the model is asked to discriminate between a real word (e.g., 'cookie') and a pseudoword matched in phonotactic probabilities (e.g., 'coonie') – see Baddeley et al. (1993), Yuspeh and Vanderploeg (2000), and Barker-Collo et al. (2008) for examples of studies in adults.

After presenting our approach and situating it within the broader context of current methodologies for modeling infant language acquisition, we present our first contribution to this line of research.

2.3 Can statistical learning bootstrap early language acquisition?

Lavechin, M.[*], de Seyssel, M.[*], Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., Dupoux, E. (2023) Statistical learning bootstraps early language acquisition. *Submitted to Developmental Science*

Motivation

Human languages are intricate systems composed of discrete linguistic categories arranged in a hierarchical structure with interdependent levels. Learning any of these levels depends on the others, thus creating a chicken-and-egg dilemma. As words are composed of phonemes, acquiring words seems to require learning phonemes first (Morgan & Demuth, 2014). But phonemes are defined as the smallest

sound contrast between two words (e.g., 'book' vs. 'look') that make a difference in meaning, therefore suggesting that infants would need words to learn phonemes first (Feldman et al., 2009; Martin et al., 2013). Neither of these learning trajectories is satisfactory, as in-lab experiments suggest that infants do not learn sounds and words separately but jointly, as evidenced in Section 2.1.1. Although widely documented in infants, this gradual and parallel learning constitutes a developmental pattern for which no formal theory has been proposed (Dupoux, 2018). We are therefore left with the following question: *How do infants bootstrap into language?*

This is the central question we explore in Lavechin, de Seyssel, Titeux, et al. (2022), a manuscript under submission to the Journal of Developmental Science, and for which we give a summary below. This work has been done in close collaboration with Maureen de Seyssel, with whom I share co-first authorship.

Paper summary

In Lavechin, de Seyssel, Titeux, et al. (2022), we introduce STELA, a language learning simulation in which we evaluate the language capabilities of our model at the phonetic and lexical levels. It is worth explaining two essential characteristics of our methodology. First, we follow a *cross-linguistic* approach whereby the learner is exposed either to English or French but evaluated on both languages. The comparison between native (training and testing on the same language) and non-native scores (training and testing on different languages) allows us to identify what our model has learned due to exposure to its native language, as opposed to exposure to another language. Second, we follow a *developmental* approach whereby we vary the quantity of speech given to the artificial learner to study the impact of input quantity on the learning outcomes.

The learning trajectories displayed by our artificial learner show a strong positive effect of native language, i.e., the native learner obtains higher phonetic and lexical scores than the non-native learner, and this holds with as few as 50 hours of speech. We also found a strong effect of input quantity, i.e., the native model gets better at discriminating sounds and recognizing auditory word forms as the quantity of speech in the training set increases. Interestingly, we observe a moderate positive correlation between the native phonetic score and the native lexical score obtained by models trained on 50 or 100 hours of speech. In other words, models that are better at discriminating sounds are also better at recognizing auditory word forms – models trained on a higher quantity of data were too few for this to be checked.

Further analyses aimed to understand the nature of the learned representations better and assess the extent to which they were similar to linguistic categories like phonemes and words. Our analyses reveal that linguistic categories structure the learned representations, although our model never learns categories per se. As the quantity of speech increases, phonetic and lexical categories become more linearly separable, which suggests that linguistic categories are not necessary during the learning process but could instead emerge as an end product of learning.

Our simulation is compatible with the gradual and parallel learning trajectory observed in infants and constitutes evidence that statistical learning mechanisms are sufficient to bootstrap early phonetic and lexical learning, such as measured by our sound discrimination and spot-the-word tasks.

Can statistical learning bootstrap early language acquisition? A modeling investigation

Marvin Lavechin[a,b,c,1,2], **Maureen de Seyssel**[a,b,d,1,2], **Hadrien Titeux**[a,b], **Hervé Bredin**[e], **Guillaume Wisniewski**[d], **Alejandrina Cristia**[a], **and Emmanuel Dupoux**[a,b,c]

[a]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France; [b]Cognitive Machine Learning Team, INRIA, Paris, France; [c]Meta AI Research, Paris, France; [d]Laboratoire de linguistique formelle, Université de Paris Cité, CNRS, Paris, France; [e]IRIT, Université de Toulouse, CNRS, Toulouse, France

Before they even produce their first word, infants become attuned to the phonetic properties of their native language, recognize the auditory form of an increasing number of words, and develop a rudimentary knowledge of grammatical categories. What kind of learning mechanism could produce such a puzzling pattern of gradual and overlapping improvement at different linguistic levels ? In-laboratory experiments have shown that young infants are exquisitely sensitive to fine-grained statistical regularities of their language input, leading researchers to propose that *"statistical learning"* could provide such a mechanism. Yet, statistical learning abilities have only been demonstrated in infants with simple artificial languages and remain controversial as a cornerstone for early language bootstrapping. Two questions remain lingering: could statistical learning work at all when fed with the full complexity and variability of natural language? Could it account for overlapping learning at multiple levels? Here, we introduce STELA, a computational model that simulates how infants might bootstrap into language from raw audio signals using statistical learning principles. STELA is built from machine learning algorithms that predict future representations of speech based on past ones. When fed with increasing quantities of raw continuous speech from multiple speakers in French and English (no preprocessing nor human annotation), STELA reproduces the observed pattern of gradual and overlapping specialization to the "native" language across levels: it improves in discriminating sounds, recognizing the auditory form of words, and organizing sounds and words along linguistic dimensions. STELA provides a proof of feasibility that statistical learning from raw speech is sufficient to bootstrap early language acquisition at the sound and word levels. Subsequent analyses indicate that this process occurs without the use of linguistic categories at these levels.

language acquisition | artificial intelligence | self-supervised learning | statistical learning | predictive learning

Infants master critical aspects of the language(s) spoken around them well before they produce their first word. Between 6 and 12 months, infants' discrimination of native sounds shows an improvement, while those of non-native sounds shows a decline (1–4). Not only do infants learn to discover sounds of their native language, they also start learning words very early on. Evidence for word learning starts as early as 4 months, where infants have been shown to recognize their own names (5). At 6-7 months, infants recognize the auditory form of frequent words (6, 7), show a preference for content over function words (8), and segment words from fluent speech (9). For their first birthday, a typically-developing American English infant comprehends around 80 words (10). Evidence suggests, therefore, a scenario of early language acquisition where learning sounds and words develop concurrently. However, it has proved devilishly difficult to understand how infants break into the intricate system that human language is. In other words, it remains unclear how infants manage to bootstrap phonetic and lexical learning from sensory information only.

One mechanism that has been proposed to explain language acquisition is *statistical learning* (11): learning from the statistical regularities of the speech input, i.e. frequency, distribution, variability, transitional probabilities, etc. Concerning phonetic acquisition, in-laboratory experiments suggest that infants use distributional information to discriminate between sounds (12–14). Regarding word learning, in a seminal experiment, Saffran et al. (15) used an artificial grammar to show that infants can track transitional probabilities across syllables to identify word boundaries. Since then, statistical learning has been studied in countless experiments across ages, domains, and species (16). Although there is a consensus among researchers that infants are sensitive to statistical regularities of their speech input, the extent to which statistical learning can explain language acquisition is at the heart of heated debates (17, 18).

One of the most prominent criticisms of the statistical learning hypothesis is that infants are embodied in a much more diverse and complex environment than what is typically present in laboratory experiments. In particular, many experiments use synthetic stimuli and artificial languages, which has the undeniable advantage of isolating the contribution of individual variables, but makes it hard to generalize to real-life language input. Indeed, two critical aspects of natural language are missing in artificial languages used in experiments. First, language is highly variable. However, in word segmentation experiments, it is common to employ artificial languages where every word shares the same length; when more variability in length is introduced, infant's ability to use transitional probabilities to segment words is severely diminished (17). Similarly, sound discrimination experiments use prototypical sounds and cherry-picked contrasts that fail to account for the large variability found in natural languages (19). Second, language is hierarchically organized into linguistic levels. In artificial languages, variability is typically frozen from all levels except the one under study. For instance, in phonetic learning experiments the language introduces phonetic variations (usually along a single dimension) but is made

Author contributions: M.L., M.S, H.B., G.W., A.C. and E.D. designed research; M.L. and M.S. performed research; H. T. created the lexical evaluation set; M.L. and M.S. analyzed data; M.L., M.S, and E.D. wrote the paper with contributions from H.B., G.W., and A.C.

The authors declare no competing interest.

[1]M.L and M.S. contributed equally to this work. Authorship order was decided by a coin flip.

[2]Address for correspondence: marvinlavechin@gmail.com or maureen.deseyssel@gmail.com

only of two monosyllabic utterances. Vice versa, in lexical learning experiments, the language contains more syllables and long "utterances", but syllables are identical copies with no phonetic variability or coarticulation effects. Even though infants have been shown to use statistical learning mechanisms in these simplified languages, could similar mechanisms work when faced with the complexity and variability of real languages and reproduce some of the observed developmental patterns?

In face of the lack of ecological validity of laboratory experiments, one possible answer consists of building computational models of language acquisition, adopting the reverse-engineering approach (20). After all, if language acquisition occurs through statistical learning, algorithms should be able to reproduce behavioral patterns observed in infants when fed with similar input. Unfortunately, the development of language learning algorithms addressing the full complexity and variability of language from raw speech input is not an easy enterprise (see (21) for a review). This is why early attempts at simulating language acquisition through computer models had also to resort to simplifying assumption and/or focus only on one aspect of language at a time. For instance, early statistical models of phonetic learning (22) did not use real continuous speech input but synthetic data generated from average formants measured in isolated syllables. Similarly, statistical models of word learning worked not from real speech, but from phonetic transcription of this input by adults who have already learned the language(23), thereby implicitely assuming that phonetic learning is completed before word learning can take place. These algorithms are useful in advancing our understanding of language acquisition as they provide proofs of learnability under certain hypotheses. However, to the extent that their simplifying assumptions are not met in real life, they do not allow to assess whether statistical learning can really address the full complexity and variability of language, from lower-level sound units to higher-level word units.

Recent advances in machine learning have provided some hope that some of these roadblocks can be lifted. For instance, Schatw et al. (24) proposed a phonetic learning model that, for the first time, learns from raw speech. They showed that a representation learning algorithm based on mixtures of Gaussian applied English or Japanese recordings could reproduce patterns of phonetic attunement as found in infants. Hitczenko et al. (25) showed that, even though language-specific statistical patterns are often obscured by the variability in running speech, such variability can be reduced by taking into account a larger window of analysis incorporating local phonetic context. Both studies constitute substantial evidence in favor of the feasibility of statistical learning hypothesis for early phonetic development. However, both studies are still only addressing one linguistic level in isolation. Would learning algorithms as applied to raw speech result in sufficiently abstract representations to sustain learning at other levels? This question is not a trivial one, given that Schatz et al. (24) found that their model was unable to converge to interpretable phonemic or even phonetic categories. Is it possible to learn words or syntax on top of such non-linguistic representations? In other words, is statistical learning restricted, in practice, to patterns of attunements to the phonology of the native language? Or can higher levels of language acquisition be reached through statistical learning?

Here, we introduce STELA (STatistical Learning of Early Language Acquisition), a learning simulation addressing for the first time the joint learning of phonetic and lexical information from raw speech. Building on recent advances in speech processing and unsupervised representation learning (26, 27), we show in Experiment 1 that a neural network trained to predict the near-future from raw speech signal, and tested with psycholinguistically-inspired discrimination and preference tasks can reproduce gradual and simultaneous learning at both the phonetic and lexical levels. At the phonetic level, the network is increasingly better at discriminating native than non-native sounds, reproducing the so-called perceptual narrowing effect documented in infants (2). At the lexical level, the network reproduces patterns of preference for real words over pseudo-words, i.e. non-existent but plausible words (9). This constitutes the first demonstration that statistical learning is *sufficient* to bootstrap early phonetic and word learning in a simultaneous fashion. In Experiment 2, we investigate whether the learned representations correspond to interpretable linguistic categories. We show that, as the quantity of speech received by the network increases, phonetic and grammatical categories become more linearly separable in the learned representations. However, the learned acoustic representations remain shorter and more variable than phonetic categories (24). A similar phenomenon occurs at the lexical level: the network does not explicitly represent words or word boundaries. Thus, in addition to providing a proof of feasibility to statistical learning potentially explaining multilevel language learning, our STELA simulation further suggests a new hypothesis, i.e., that linguistic categories are not needed to account for patterns of early language development. In the General Discussion, we discuss the consequences of these findings for theories of early language acquisition.

Approach

STELA follows the reverse engineering approach described in (20) whereby a full computational simulation of language acquisition addresses three components of the learning situation: the environment, the learner, and the outcome measures (see Figure 1). Here, we give only a high-level sketch of these components described in more details in the Methods Section.

As in (24, 28), we take the *environment* of the infant to be composed of raw speech input. Here, we extract 3,200 hours of speech from French and English audiobooks, which corresponds to the upper limit of what infants could hear during the first three years of their life (29). For each training language (English or French), we built training sets by randomly splitting the whole set of audio segments into mutually exclusive training sets of 50 hours. These 50-hours training sets were then merged two by two to build the 100-hours training sets. This procedure was repeated until convergence, which left us with 64, 32, 16, 8, 4, 2, and 1 training sets of 50h, 100h, 200h, 400h, 800h, 1,600h and 3,200h of speech.

We simulate the *learner* by using the winning entry of the ZeroSpeech 2021 international challenge on unsupervised representation learning (26). It consists of two components. The *Acoustic Model* takes as input raw audio and outputs a discrete unit every 10ms slice of time. The *Language Model* takes the discretized version of the audio as input and outputs a prediction for the next units, similarly to text-based language

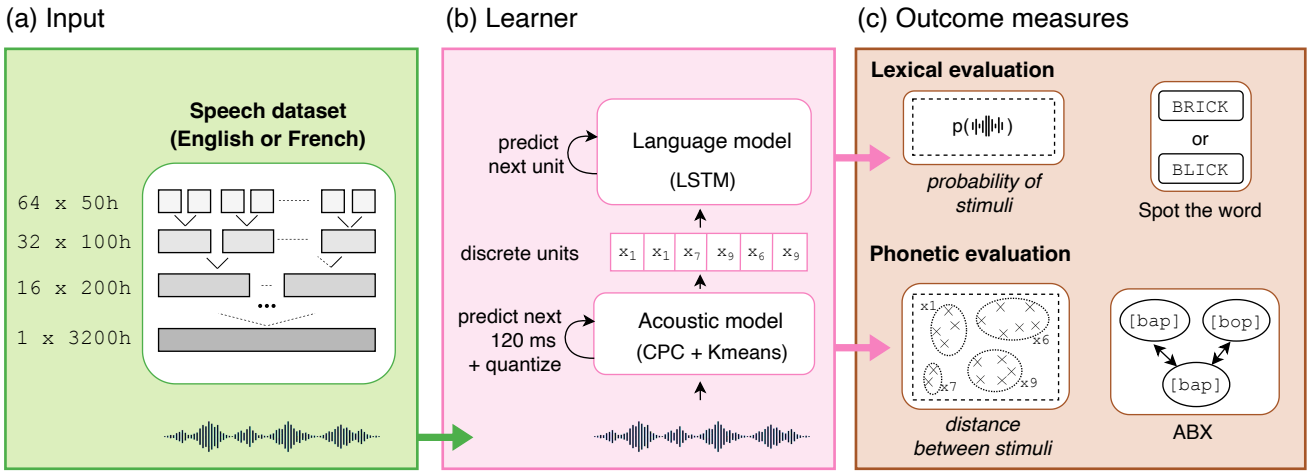

Fig. 1. Overall setup for the training and testing of STELA. a. The audio environment of learners of different 'ages' are modeled using audiobooks segmented and aggregated in increasingly larger sets matched for number hours and of speakers across two languages (See Table 1). b. The learner is composed of an 'Acoustic Model', first trained with predictive coding and followed by a K-means algorithm returning discrete units and a LSTM 'Language Model' trained to predict future units based on past units. c. Outcome measures are obtained by modeling an ABX sound discrimination task at the (discretized) output of the Acoustic Level, and an auditory lexical preference task (Spot-the-Word) by using the ability of the Language Model to compute the probability of stimuli.

models except that the latter are trained on words. The two components are trained by minimizing self-supervised objective functions on the same chunk of data. In other words, the model learns from the raw speech only, without any human annotation intervening in the loop. It thus obeys a critical constraint for modeling infant language development. Children are never explicitly given linguistic knowledge, so neither should computational models. Once trained, a model constitutes a simulation of an infant exposed to a particular language for a given amount of exposure.

We measure our learners' language *outcomes* at two linguistic levels: the phonetic level (sounds) and the lexical level (words), drawing inspiration from psycholinguistic studies (see Section A.3). At the phonetic level, we simulate an ABX auditory discrimination task using phonetic contrasts, e.g. /ɪ/ versus /ɛ/as in "bit" versus "bet". At the lexical level, we simulate a spot-the-word task: the model is asked to identify which of two audio stimuli (e.g., "brick" and "blick") is a word (the former), and which is a pseudo-word (the latter). For each trained model and each target language, we obtain a phonetic and a lexical score, such that 100% and 50% indicate perfect and chance-level accuracy, respectively. We compute the average phonetic and lexical scores in the native condition (the English model evaluated on English, and the French model evaluated on French) and the non-native condition (English model on French, French model on English). Contrary to humans, machines can be presented with thousands of trials for a given stimulus type (words or phonetic contrasts), allowing us to extract robust measures of learning outcomes.

The comparison between native and non-native scores allows us to identify what our model has learned due to exposure to its native language (as opposed to exposure to another language). In other words, the non-native model acts as a control for the native model. By assessing our models' language capabilities as a function of the quantity of speech they have been exposed to, we draw developmental trajectories and ask whether or not learning outcomes exhibited by our model share similarities with infant language development. Finally, we supplement these two tasks with in-depth analysis of the representations learned by the system.

1. Experiment 1 : Can statistical learning bootstrap both phonetic and lexical learning?

The objective of our first experiment is to investigate whether our model demonstrates phonetic and lexical learning outcomes and whether such learning occurs gradually and concurrently, similar to how it does in infants, as aligned with the primary question presented in the introduction.

A. Material and Methods. In this section, we provide a more comprehensive description of the model's implementation, including details on the input data, learner design and outcome measures.

A.1. Training sets. We used 10,000 hours of English audiobooks from the Librivox platform (30) and 10,000 hours of French audiobooks from litteratureaudio (31). We constructed 64 twin chunks of 50 hours of speech (3200 hours total) made of entire book chapters in each language, such that the number of speakers would be as matched as possible across the two languages. To achieve this, we applied a stochastic sampling algorithm that matches across English and French: 1) the cumulated duration, 2) the number of speakers per chunk of 50h, and 3) the number of chunks per speaker. We then randomly aggregated the 64 chunks of 50 hours two by two to obtain 32 chunks of 100, until we obtained one large 3200h chunk in each language. Table 1 provides further statistics that demonstrate the matching between the English and French training sets.

Table 1. Statistics for the French and English training sets varying in quantity of speech. Average number of speakers per training set, average quantity of speech for the least talkative and the most talkative speaker per training set.

Training	French			English		
sets	N	min (h)	max (h)	N	min (h)	max (h)
64x50h	9.7	0.33	16.96	9.7	0.75	15.84
32x100h	17.0	0.19	24.11	17.3	0.55	20.81
16x200h	28.7	0.14	35.61	29.6	0.41	29.90
8x400h	46.9	0.06	58.45	49.1	0.32	45.22
4x800h	73.7	0.05	94.84	74.7	0.23	75.43
2x1600h	107.0	0.04	187.89	108.5	0.19	133.75
1x3200h	147.0	0.01	334.17	147.0	0.17	267.50

A.2. Learner design. We describe below our proposed model learning speech representations from the raw waveform (26). The learner is composed of two components: 1) the Acoustic Model that learns discretized representations of the the raw waveform, and 2) the Language Model that takes the discretized representation as input and returns a probability distribution over the set of discrete units.

The acoustic model. It consists of a Contrastive Predictive Coding (CPC) algorithm (27, 32). The key idea behind CPC is to predict the near future of a sequence given its past context (see Appendix for more details). The learner is given an example that is drawn from the near future up to 120 ms (called positive example), and multiple examples that are not drawn from the near future (called negative examples). Given the past context of a sequence, the learner is asked to maximize the categorical cross-entropy of classifying the positive sample correctly (see Appendix 1.1 for more details). The continuous context-dependent representations output by CPC are then fed to a simple K-means clustering algorithm that returns a discrete representation of the audio.

The language model. The language model takes as input the discrete representation of the audio file returned by the acoustic model. It is trained to predict the next discrete unit via a cross-entropy loss function (see Appendix 1). At test time, the model is simply used to produce a probability of a stimulus $S = q_1, q_2, ..., q_T$ by applying the following formula:

$$P(q_1, ..., q_T) = -\frac{1}{T}\sum_{t=1}^{T} \log p(q_t \mid q_1, ..., q_{t-1})$$

Based on this probability, it becomes possible to simulate a preference task between two stimuli. A stimulus A is preferred over B if its probability, as estimated by the language model, is higher.

A.3. Outcomes measures. **Phonetic evaluation: the machine ABX sound discrimination task**

General principle. The ABX sound discrimination task was first proposed by (33) to offer a way to evaluate models' phonetic discrimination capabilities in a setup comparable to how humans are evaluated. The task consists of generating a wide range of triplets of sounds in the format A, B and X, with A and X corresponding to different variations of the same triphone ('bop') and B to another triphone where the central phone changes ('bap'). Distances between A and X, and B and X are then computed using Dynamic Time Warping (DTW) based on a frame-to-frame cosine distance. A score of 1 is given if $d(A, X) < d(B, X)$, otherwise the score is 0. An average score is finally computed over all possible triplets.

The ABX task can be used on any type of speech representation, and has already proven robust with the CPC+K-means architecture presented here (26). In this paper, we use the discrete representations output from the K-means algorithm to compute the ABX score.

Materials. The triplets are generated over carefully tailored English and French speech test sets, which are subsets of the CommonVoice dataset (34). These test sets, already presented in (35) and (28), consist of 10 hours of read speech balanced between 24 speakers (12 males and 12 females). All utterances from the English and French test sets are tagged as "US accent" and "France accent" respectively in the original CommonVoice dataset. The phone-level alignment was obtained by aligning the audio stream with its transcript using Kaldi recipes (36), eventually allowing us to generate triplets for the ABX task. The ensuing phonetic inventory in International Phonetic Alphabet (IPA) standard for both languages is shown in Table S1.

Lexical evaluation: the spot-the-word task

General principle. The evaluation of lexical knowledge in a recurrent neural network was first proposed in (37) using the spot-the-word task. It consists of presenting the network with a minimal pair of word and non-word (e.g., 'brick' versus 'blick') and evaluating whether the probability given by the network to the word is higher or lower than the probability given to the non-word, yielding an accuracy score, which was averaged across all of the pairs in the test set.

Materials. The pairs are constructed using the Wuggy toolbox (38), which generates lists of nonwords matched for syllabic and phonotactic structure with a given list of words. To build our test set, we first selected the list of words present in our environments and constructed for each word a set of associated non-words using Wuggy and pronunciation dictionaries for French and English (39). We then reduced this list to a single non-word by applying a filter maximizing the frequency of unigrams and bigrams of phonemes between the words and the non words. We then synthesized the words and non-words using the Google text-to-speech API (40) in 4 voices (2 males, 2 females) in each language.

The resulting list of word/non-word pairs was further sorted into frequency bands by intersecting them with the different environments. The highest frequency band was constructed by selecting the words that appeared at least once in each of the 64 50-hours environments. The second highest frequency band was made of words that appeared at least once in each of the 32 100h environments and that were not in the preceding list, and so-forth until we had the corresponding 7 frequency bands. In Figure 2, we only displayed the results of the highest frequency bands. The results for each frequency band can be found in Supplementary Figure S5.

B. Results and discussion. Panels (a) and (b) of Figure 2 show the scores obtained on the phonetic and lexical tasks, for the native and the non-native learners, as a function of input quantity. Results indicate that native models trained on 3,200 hours of speech succeed in discriminating sounds (81.64% phonetic score) and, to a lesser extent, recognize the auditory form of words (62.98% lexical score).

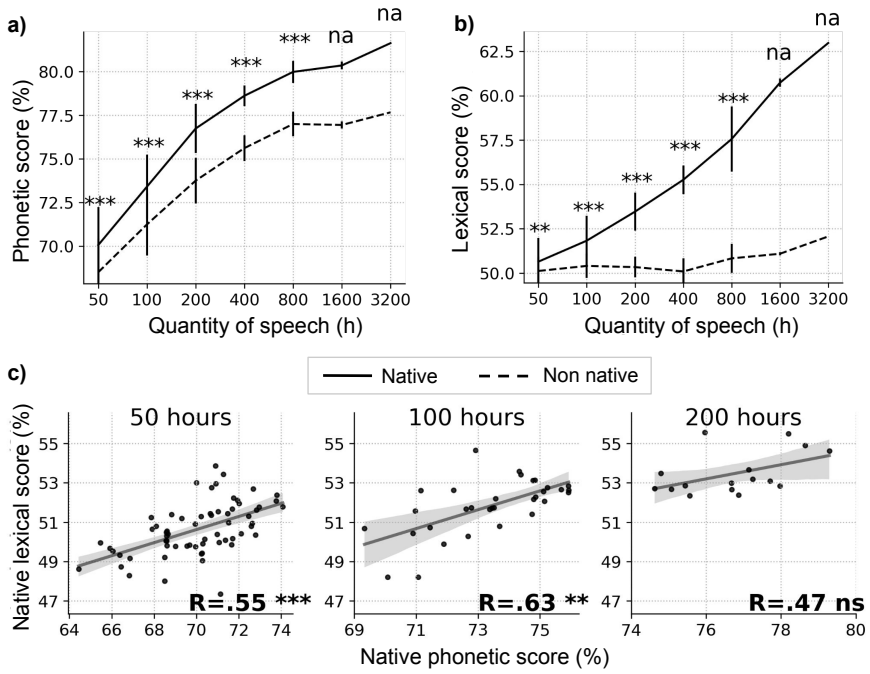

Fig. 2. Gradual and parallel learning across the phonetic and the lexical levels. a) Phonetic score, in terms of ABX accuracy, obtained by the discrete representations for native and non-native input. b) Lexical score, in terms of accuracy on the spot-the-word task, on the high frequency words for native and non-native input. For a) and b), two-way ANOVAs with factors nativeness and training language were carried out for each quantity of speech. Significance scores indicate whether the native models are better than the non-native ones. c) Correlation between the phonetic and lexical scores obtained across individual native models trained for 50h, 100h and 200h in English and French. R is the Pearson correlation coefficient. Significance levels: na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

The developmental aspect of STELA also allows us to assess the evolution of the learning trajectories. In the native condition, both phonetic and lexical scores increase gradually as a function of quantity of speech. Phonetic and lexical scores obtained by the native model are systematically higher than those obtained by the non-native model. This difference increases with input quantity, reaching a relative difference of 5% for the phonetic score, and 18.97% for the lexical score between our native and non-native learners trained on 3,200 hours of speech. Using two-way ANOVAs with factors nativeness and training language, we found that native scores were significantly higher than non-native scores for as little as 50 hours of speech ($F(1, 252) = 18.95$, $p<.001$ for the phonetic score, $F(1, 252) = 15.81$, $p<.0001$ for the lexical score). Significance tests on 1,600 and 3,200 hours of speech are not available due to the low number of models. To summarize, the proposed algorithm learns key aspects of its native language at both the phonetic and the lexical levels in a gradual and simultaneous fashion, consistently with what has been observed in young infants (7, 41–43).

The phonetic score obtained by the non-native model improves with input quantity (as previously noticed in (24)). This developmental pattern might seem to run counter to experiments that report a loss in non-native sound discrimination in infants (2). However, our setup differs from the usual infants experiments, as we systematically average performance over all possible phonetic contrasts in the present study (see Supplementary Table S1 for the list of evaluated phonemes, and Supplementary Section 4 for similar comments on the non-native lexical score). In infant studies, the non-native sound discrimination loss was documented only for a small number of carefully selected phonetic contrasts which are known to be difficult for the non-native language tested (such as the "r" versus "l" as in /rock/ versus /lock/ in Japanese infants). Besides, we know that many non-native contrasts map onto native ones (44), which would explain the phonetic learning even in the non-native condition. For instance, interdental fricatives can map from one language to the other (/s/ and /θ/ in English map to /s/ and /z/ in French). The increase in phonetic score by the non-native model is an interesting observation that could be tested in infants. On the other hand, this non-native learning is not observed in the lexical task. This was expected as, contrary to the phonetic task, there is no overlap between auditory word forms in the two languages.

Further evidence for lexical learning in the native condition is provided by an additional analysis (Supplementary Section 7) showing that the higher the frequency of the evaluated words, the higher the lexical score obtained by the native model. This frequency effect has been widely documented in young infants and has been argued to be an important requirement for any successful account of language acquisition (45). Investigation of a large-scale study of human reaction times in auditory lexical decision (deciding whether a word exists) revealed that word probabilities computed by the native model correlate with linguistic factors shown to influence human lexical decision times (such as the duration, the frequency and the number of phonological neighbors of the word; see Supplementary Section 8). All in all, we found evidence for learning at the phonetic

and lexical levels using an algorithm exclusively based on statistical learning.

Two models exposed to the same quantity of speech can perform differently on the phonetic and lexical tasks. This is due to: 1) the training set itself that may constitute a more or less adequate language experience; and 2) the randomness in the weights' initialization and in the way data is presented, which may advantage or disadvantage the model. With this in mind, we can attempt to characterize the relationship between phonetic and lexical outcomes obtained by our models. Panel (c) of Figure 2 shows significant positive correlations between the scores obtained on the phonetic and lexical tasks, respectively, across models trained on 50h, 100h, or 200h of speech (there were fewer than 8 models trained on larger quantities of speech, not enough to compute meaningful correlations). This result indicates that models that are better at discriminating native sounds are also better at solving the spot-the-word task. This is compatible with infant studies suggesting a positive correlation between native discrimination and vocabulary size at 11 months (46, 47). Similarly, multiple longitudinal studies show that early sound discrimination capabilities predict later language development (48–50). Further work could assess specifically whether there exists a positive correlation between native discrimination and auditory word form recognition.

All the analyses presented in this section can also be found separately across the English and the French model in Supplementary Sections 4 and 5.

2. Experiment 2 : Are linguistic categories required?

In the previous Experiment, we have shown that our models improve in both lexical and phonetic tasks, more so for native than non-native tests, which parallels findings with human infants. In an attempt to better understand the nature of the learned representations, we dedicate the current section to a deeper analysis of how similar these representations are to linguistic categories.

A. Methods. Additional analyses are carried out to compare the model's representations to linguistic categories. Linguistic categories are analyzed at two levels: at the acoustic model for the phonetic categories (phone class/sonority, place of articulation, and voicing) and at the language model for the lexical categories (broad *function vs. content word* differentiation, and content words' part of speech). For these analyses, the same English and French test sets as presented in Section A.3 are used, consisting in speech-to-phones and speech-to-words alignments.

For each category, a qualitative and quantitative analysis is run. The qualitative analysis consists of a 2D visualization of the output speech representations from the model trained on most data (3200h), colored in terms of their linguistic category. We extracted the output acoustic (language) model representation of every test sentence in the language the 3200h model was trained on (the representations are therefore context-dependent). We then extracted the representation for every phone (word) and used a mean-pooling function to obtain a fixed-dimension representation for each of these phones (words). We applied a t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to reduce the representations to 2 dimensions on a subset of the data (N=6,000). Finally, we plotted the resulting dimensions and color-coded the data points based on their target characteristic category (phone class / place of articulation / voicing / part of speech).

The second, quantitative, analysis focuses on the emergence of linguistic categories as a function of input quantity. This allows an understanding of whether the models' speech representations become closer to the linguistic categories when trained on more data. To do so, we split the test set into sub-training and sub-test sets. The sub-training set contains models' representations of all phonemes minus one phoneme. The sub-test set contains models' representations of the final phoneme (the same is done at the lexical level with representations of 50 word types per category chosen out of the 100 most frequent word types - the other 50 being used as a development set - see below). A logistic regression model is then trained for all sub-training set representations, using the desired linguistic categories as targets, before being tested on the sub-test representations. This process is done iteratively with all possible phonemes (word type) being part of the test set, using Leave-One-Out cross-validation. This allows us to retrieve an average classification error for the model on the specified information. This is done on all models of all different training sizes, allowing us to draw developmental curves of these classification errors. The chance classification error and error calculated on raw MFCCs were also computed. Finally, we check the significance of the developmental curve's slope (correlation between classification score and quantity of input) using Spearman's rank correlation.

For the phonetic analyses, representations were extracted from the last hidden layer of the CPC model (these are the same representations used for the ABX task). We use all phones for the 'sonority' analyses, however we only keep consonants for the 'place of articulation' and 'voicing' categories, as vowels are not relevant here*. Regarding the lexical analyses, we chose the hidden layer yielding the best classification error scores on the 3200 hours model, using a development set also formed of 50 word types per category sampled out of the 100 most frequent word types per category. The best hidden layer was the third (last) for both the English and French models (logistic regression scores on all layers are available in Supplementary Table S2).

B. Results and discussion.

B.1. The emergence of phonetic categories. Although our model works with 10-ms frames, the Acoustic Model often assigns the same discrete unit to multiple successive frames. Do these duplicated discrete units share commonalities with phonemes, in terms of duration and perplexity? Our analysis reveals both that these units are much shorter than actual phonemes (see Figure 1), and that a same unit can encode multiple phonemes (see Supplementary Section 10). These conclusions mirror results with a different acoustic model found in (24). We also found that this pattern does not change with the amount of training data. If anything, the learned units become shorter as the amount of data increases (top graphs of panels (b) and (d) of Supplementary Figure S8). At the same time, we observe an opposite trend for the number of units associated to each phoneme: the unit-to-phone perplexity decreases with input quantity. This indicates that the more speech the model receives, the more fine-grained the learned discrete units are.

*We also discard approximants from the place and voicing analyses, as well as the English h and the French ʁ, as they are alone with their place of articulation label.

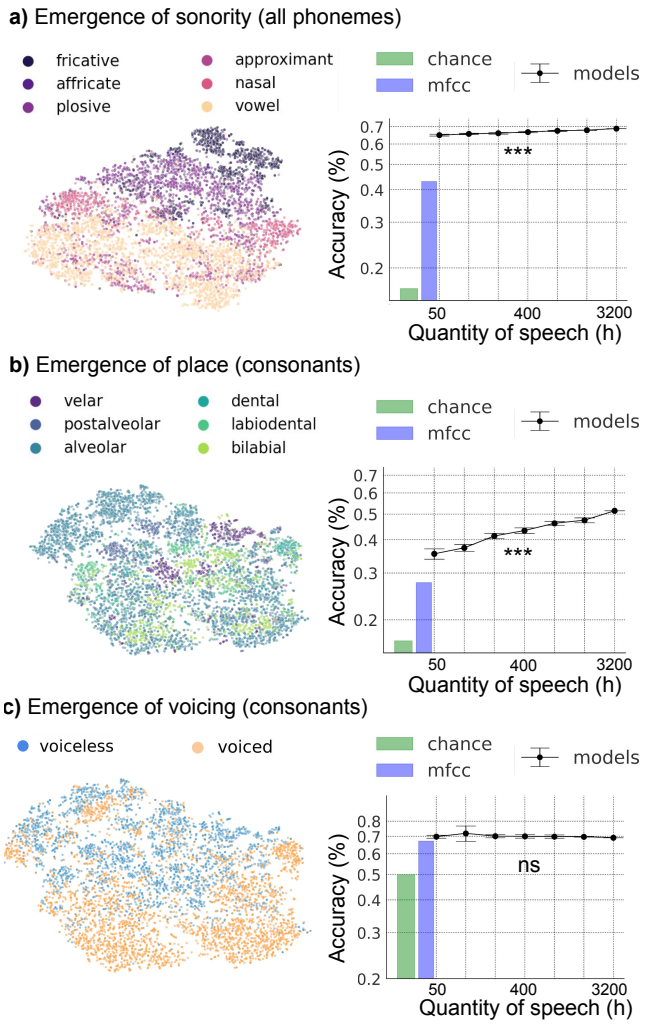

Fig. 3. Emergence of latent linguistic categories at the phonetic level for the English models. Left: tSNEs of the continuous representations of the acoustic model (last layer) pooled within phones in a test set, according to sonority (a), place (b) and voicing (c) for the 3200h English model. Right: developmental curves from a leave-one-phoneme-type-out classification errors as a function of input quantity (taking all 256 dimensions into account). Chance level and MFCCs performances are also given. The asterisks indicate a significant correlation of classification error and input quantity. na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

Although linguistic categories are not hard-coded and therefore do not exist *per se* in our model, could the learned representations encode important linguistic information? Using a t-distributed stochastic neighbor embedding (t-SNE) method on the representations learned by the English acoustic model trained on 3,200h of speech, Figure 3 shows that the learned acoustic representations encode multiple phonetic features. Phone representations are organized along a continuum spanning from sounds that are very sonorous (vowels) to not sonorous (fricatives) (panel (a)). Similarly, consonant representations are clustered by place of articulation (place where the constriction and obstruction of air occur when producing the consonant), and by voicing (whether or not produced with vocal cord vibration) (panels (b) and (c)).

The projection of high-dimensional representations in 2D spaces results in an important loss of information and constitutes only a qualitative analysis of the learned representations. Therefore, we use logistic regressions as probes to analyze quantitatively the information encoded within the models (51, 52). We train a linear classifier on top of the continuous acoustic features to measure the extent to which previously studied phonetic features (sonority, place of articulation, and voicing) are present in the learned representations (Figure 4). We compare the classification scores of our probes with those obtained both by a random linear classifier (representing chance level, in green) and by one trained with mel-frequency cepstral coefficients (MFCCs, representing acoustic representations, in blue). Results indicate that sonority, place of articulation, and voicing are encoded in the learned representations even by the model trained on the smallest quantity of input (50h) of English speech, since all scores are better than both the random baseline in green and the acoustic representations in blue. Classification errors on sonority and place of articulation improve with data quantity, showing a positive effect of exposure. This does not hold for voicing, for which the linear classifier obtains a high classification score. Equivalent analyses on the French model and further details can be found in Supplementary Section 6.

Results presented in this section show that, although the learned representations are too fine-grained to correspond to phonetic categories as defined by linguists, they nonetheless contain information that encodes critical phonetic features. In addition, our study found that for two of the three phonetic features we examined, such a perceptual organization emerges in a gradual fashion, with a positive effect of input quantity.

B.2. The emergence of lexical and grammatical categories. Next, we look at whether lexical and grammatical information is present in the representations learned by the language models. We follow the same procedure as above and analyze word representations in a qualitative way using t-SNE and in a quantitative way using linear classifiers. In particular, we probe two dimensions: 1) the distinction between function and content words; and 2) part-of-speech categories among content words (nouns, verbs, adjectives).

Experimental studies suggest that infants know at least some of the function words of their native language around one year of age (53), and that they use this information to infer part-of-speech categories among content words in their second year of life (54, 55). Mainstream theories like prosodic bootstrapping hold that both the distinction between function and content words, and the part-of-speech categories among content words are crucial cues in early language acquisition, particularly in lexical segmentation and syntactic parsing (56).

A 2D t-SNE projection of the word-level representations learned by the language model does not reveal a clear separation between function and content words (left of panel (a), Fig. 4), although some regions of the space seem specific to each grammatical class. The same conclusion can be drawn when coloring content word representations according to their part of speech categories (left of panel (b), Fig. 4).

However, it is not because t-SNE does not exhibit a clear separation between linguistic categories that the information is not present in the learned representations (as mentioned in the previous section, t-SNE leads to a loss of information). As a matter of fact, linear probing on the learned representations suggests that linguistic information is indeed present. Specifically, it is possible both to classify whether a word is a

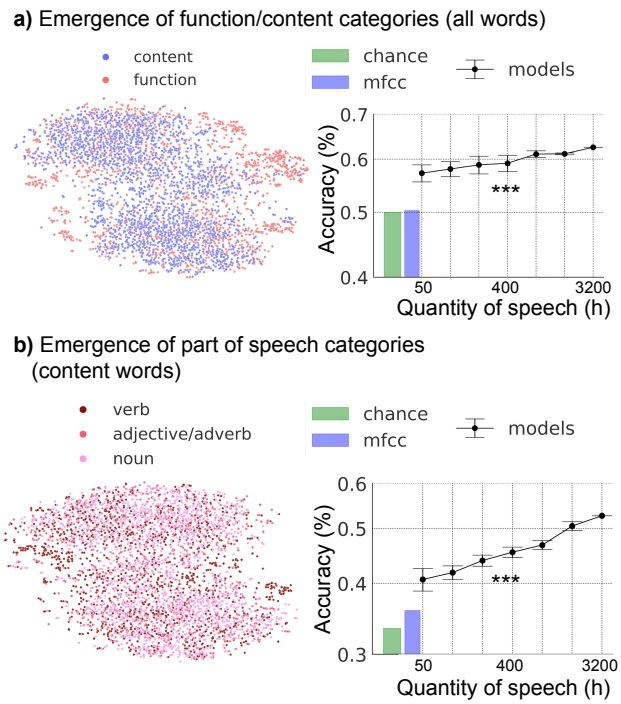

Fig. 4. Emergence of latent linguistic categories at the word level. Left: tSNEs of the continuous representations of the language model (last layer) pooled over words according to (a) function/content distinction and (b) part of speech for the 3200h English model. Right: corresponding developmental curves of leave-one-word-out classification error as a function of input quantity (taking all dimensions into account). Chance level and MFCCs performances are also given. The asterisks indicate a significant correlation of classification error and input quantity. na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

function word or a content word (right of panel (a), Figure 4) and to classify the part-of-speech categories of content words (right of panel (b), Figure 4). Linear classifiers trained on the learned representations of the language model are indeed better than chance and better than the MFCC baseline on both classification tasks. Importantly, accuracy increases when the representations are learned on a larger quantity of data, showing that categories become more linearly separable as the quantity of speech increases, showing a positive effect of exposure (p<.0001).

All in all, results in this section suggest that the learned representations are somewhat structured by word categories. This organization emerges in a gradual fashion, with a positive effect of input quantity.

3. General Discussion

Whether statistical learning can account for infant early language acquisition, despite plethora of experimental infant studies on the topic (see (16) for a review), still remains an open question (17, 57, 58). Besides studies carried out directly on infants, some computational models showed the feasibility of language acquisition in statistical learning, but these models either only focused on a single aspect of language acquisition (phone discrimination (24); word learning, (59)), or they made strong assumptions on the input data (using processed signal or text), making their plausibility as a model of infants questionable. Recent studies (24, 25) provided strong evidence in favor of the statistical learning hypothesis for early phonetic learning. Can statistical learning account for higher-level aspects of early language learning?

STELA constitutes the first proof of feasibility of statistical learning to account for early language acquisition at the phonetic and lexical level. We further showed that phonetic and lexical learning was possible without linguistic categories. More generally, we proposed the first developmental psycholinguistic analysis of a state-of-the-art machine learning model. In this section, we reflect on STELA key findings and limitations.

Statistical learning is enough to bootstrap language learning In our STELA simulation, we have shown for the first time that a self-supervised learning model built within the scope of the statistical learner hypothesis can reproduce developmental patterns of gradual learning at both the phonetic and lexical levels when provided with untranscribed, raw, clean speech data. The model works by implementing a neuro-cognitively motivated statistical learning mechanism (predictive coding) within a linguistically interpretable division of levels (discrete acoustic vs. high-level abstract), trained on raw audio data. We found that linguistic knowledge at these two levels (phonetic and lexical) emerges gradually and in parallel, as attested by psycholinguistic-inspired tests and analyses. Such results constitute strong proof of feasibility for the statistical learning hypothesis, suggesting that such computations are sufficient to bootstrap phonetic and lexical knowledge when provided with raw, clean speech.

However, there are several limits that need to be addressed before claiming that statistical learning alone can bootstrap the entire linguistic system. First, we only analyzed two linguistic levels: phonetic and lexical, the latter being restricted to word forms. Bootstrapping language would require to show the other linguistic levels that have been documented as emerging in young children (prosody, syntax, semantics) also emerge thanks to the same mechanisms. Specific tests inspired by infant psycholinguistics probing these levels would need to be developed and applied to the model[†]. Second, we used as input audiobooks, which are much less noisy than the audio available to infants. It is possible that additional mechanisms besides statistical learning are needed to cope with such variability (28).

Finally, infants are much more than simple statistical learners, and previous studies have found that cross-modal learning and social interactions play a significant role in infants' language acquisition (58, 64–66). We want to point out that our study does not question this, and that these other types of input could well be critical in the development process. Instead, our proof of feasibility shows that relying only on a statistical learning mechanism to start bootstrap language is possible.

Phonetic and lexical learning without linguistic categories The seminal work carried out by Schatz et al. (24) suggested that statistical learning can be used to reproduce developmental patterns in phonetic learning without creating phoneme-like units, therefore questioning the presence of such categories in infants (see also (67)). Analyses carried out on the representations learned by our model point in the same direction:

[†] Work in spoken language modeling (60–63) suggests that these levels can emerge from statistic mechanisms applied to the raw speech, but such models typically require much more input data than is available to infants and it remains to be seen that they can reproduce plausible developmental curves.

the learned units do not correspond to the usual phonetic categories. The discrete level itself (acoustic units) is not linguistically interpretable and does not tend to become more phoneme-like with more input data, but rather to correspond to more fine-grained sub-phonetic units, mirroring Schatz' results with a different model.

Examining the lexical level for the first time, we surprisingly found a similar pattern: the learned representations do not directly map to word-level categories such as part-of-speech. Two main lessons can be drawn from these results: 1) sub-phonetic units are sufficient to learn higher-level aspects of language; and 2) word categories are not required to recognize the auditory form of words. Thus, our work questions the need for any linguistic categories, and not only phonetic ones, in the early stages of language acquisition. Similarly, even though we found evidence for the emergence of lexical and grammatical information, this information does not seem to be grounded into a segmentation of the input into word-like chunks (see Supplementary Section 9).

Gradual and parallel learning in STELA Within the STELA framework, we introduced a carefully designed developmental setup, which allows us to compute the effect of quantity on phonetic and lexical learning, to generate their respective developmental curves, and to compare them to experimental results.

At a qualitative level, the developmental curves show a gradual and parallel increase in both phonetic discrimination and lexical preference. How can we account for such a pattern? Our algorithm works by minimizing quantities called *loss functions*. We use three such functions that are minimized jointly. The acoustic model minimizes the prediction errors over continuous acoustic representations (predictive coding) and then discretize them using a compactness score (discretization). The language model minimizes the prediction error over the discrete units. The two prediction errors are minimized by the stochastic gradient descent algorithm and the compactness score by a variant of the expectation-maximization algorithm. The gradual and parallel aspect of the results is due to the fact that these three loss functions are optimized to lower values as more data is presented (see Supplementary Figure S1).

Does statistical learning actually bootstrap early language acquisition? The core demonstration of our work consists in showing that a statistical learning mechanism can exploit the information present in the raw speech signal and reproduce patterns of early stages of language acquisition, such as measured by our psycholinguistically-inspired evaluation tasks. This shows that infants could rely on statistical learning mechanisms to bootstrap language acquisition, but it does not show that they necessarily do.

In other words, while STELA is valuable in providing a proof of feasibility of the statistical learning mechanism in early language learning, it cannot at present be considered a fully fledged model of the infant because of several limitations. One limitation of the current implementation of STELA is that while it provides a series of cross-sectional predictions (by simulating infants of different ages), it does not allow for longitudinal studies: models are trained anew for every quantity of input, and led to convergence every time. Implementing a longitudinal framework would require larger datasets, with, ideally, each training set representing a single child's input, and a modification of the learning algorithm to yield incremental results for each increasing amount of input.

Regarding the model of the learning outcomes, although heavily inspired by psycholinguistic experiments, it does not directly simulate the experiments as they are run in a laboratory setting: preferential looking, high-amplitude sucking, etc. These procedures have been designed to explore processes at different stages of the infant's speech perceptual development and aim at eliciting specific behavioral responses from the infant. In this regard, the machine evaluation tasks are far simpler and directly interpretable in terms of: 1) distance between sound representations for the phonetic evaluation; 2) prediction error of words and pseudo-words for the lexical evaluation. The next challenge will likely consist in allowing better comparison between infants' language learning outcomes and those obtained via our in-silico simulations, i.e., moving beyond qualitative comparison. The noise inherent to infants' behavioral responses might prevent us from doing that in the near future, but a promising approach might consist in comparing learning outcomes obtained by the machine against large-scale cumulative empirical infant data.

Finally, the model of the environment adopted in the present study used relatively well-articulated speech without background noise. As shown in (28), infants have the additional task of separating speech from noise, which is not taken into account in the present simulation. Once these limitations are addressed, it may be possible to more directly compare the predictions of STELA with actual infant's outcomes, and validate or invalidate it as a possible model for early language acquisition.

4. Conclusion

Overall, this proof of feasibility shows that self-supervised learning models are good *a priori* candidates to help us understand trajectories in infant language development. Machine learning solves deep puzzles in cognitive development and provides quantitative models that make numerical predictions as a function of the amount of input data. While this proof of feasibility shows that phonetic and lexical bootstrapping is possible using only statistical learning mechanisms, there remain many challenges, including going further towards ecological audio data and benchmarking against actual infant experimental results. Even more challenging will be the issue of closing the gap between computational models and the actual cognitive learning processes used by infants: To what extent do infants actually make use of statistical learning mechanisms during language acquisition? And what is the place of other mechanisms (social learning, intrinsic motivation) in the developmental pathway?

STELA offers the potential to simulate the entire language acquisition process in the early years of life using a fully implemented model that operates on real audio input. This could generate a wealth of quantitative predictions that can be compared to data on infants. By open sourcing the model, we hope to inspire a shift towards a more quantitative approach in infant research.

ACKNOWLEDGMENTS. We are especially grateful to Sharon Peperkamp for enlightening discussions and proofreading sessions. We are grateful to LAAC, and CoML members for helpful discussion. All errors remain our own. A.C. gratefully acknowledges financial and institutional support from Agence Nationale de la

Recherche (ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017); the J. S. Mc-Donnell Foundation (Understanding Human Cognition Scholar Award); and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). E.D., in his academic role (EHESS), acknowledges funding from Agence Nationale de la Recherche (ANR-19-P3IA-0001 PRAIRIE 3IA Institute), a grant from CIFAR (Learning in Machines and Brains), and the HPC resources from GENCI-IDRIS (Grant 2020-AD011011829). M.S. acknowledges PhD funding from Agence de l'Innovation de Défense.

1. RE Eilers, WR Wilson, JM Moore, Developmental changes in speech discrimination in infants. *J. Speech Hear. Res.* **20**, 766–780 (1977).
2. PK Kuhl, et al., Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. science* **9**, F13–F21 (2006).
3. FM Tsao, HM Liu, PK Kuhl, Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The J. Acoust. Soc. Am.* **120**, 2285–2294 (2006).
4. Y Sato, Y Sogabe, R Mazuka, Discrimination of phonemic vowel length by japanese infants. *Dev. Psychol.* **46**, 106 (2010).
5. DR Mandel, PW Jusczyk, DB Pisoni, Infants' recognition of the sound patterns of their own names. *Psychol. Sci.* **6**, 314–317 (1995).
6. PW Jusczyk, EA Hohne, Infants' memory for spoken words. *Science* **277**, 1984–1986 (1997).
7. MJ Carbajal, S Peperkamp, S Tsuji, A meta-analysis of infants' word-form recognition. *Infancy* **26**, 369–387 (2021).
8. R Shi, JF Werker, Six-month-old infants' preference for lexical words. *Psychol. Sci.* **12**, 70–75 (2001).
9. PW Jusczyk, RN Aslin, Infants' detection of the sound patterns of words in fluent speech. *Cogn. psychology* **29**, 1–23 (1995).
10. MC Frank, M Braginsky, D Yurovsky, VA Marchman, Wordbank: An open repository for developmental vocabulary data. *J. child language* **44**, 677–694 (2017).
11. AR Romberg, JR Saffran, Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 906–914 (2010).
12. J Maye, JF Werker, L Gerken, Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* **82**, B101–B111 (2002).
13. J Maye, DJ Weiss, RN Aslin, Statistical phonetic learning in infants: Facilitation and feature generalization. *Dev. science* **11**, 122–134 (2008).
14. S Tsuji, A Cristia, Perceptual attunement in vowels: A meta-analysis. *Dev. psychobiology* **56**, 179–191 (2014).
15. JR Saffran, RN Aslin, EL Newport, Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).
16. JR Saffran, NZ Kirkham, Infant statistical learning. *Annu. review psychology* **69**, 181 (2018).
17. EK Johnson, MD Tyler, Testing the limits of statistical learning for word segmentation. *Dev. science* **13**, 339–345 (2010).
18. J Lidz, A Gagliardi, How nature meets nurture: Universal grammar and statistical learning. *Annu. Rev. Linguist.* **1**, 333–353 (2015).
19. D Swingley, Contributions of infant word learning to language development. *Philos. Transactions Royal Soc. B: Biol. Sci.* **364**, 3617–3632 (2009).
20. E Dupoux, Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition* **173**, 43–59 (2018).
21. O Räsänen, Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Commun.* **54**, 975–997 (2012).
22. GK Vallabha, JL McClelland, F Pons, JF Werker, S Amano, Unsupervised learning of vowel categories from infant-directed speech. *Proc. Natl. Acad. Sci.* **104**, 13273–13278 (2007).
23. S Goldwater, TL Griffiths, M Johnson, A bayesian framework for word segmentation: Exploring the effects of context. *Cognition* **112**, 21–54 (2009).
24. T Schatz, NH Feldman, S Goldwater, XN Cao, E Dupoux, Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proc. Natl. Acad. Sci.* **118**, e2001844118 (2021).
25. K Hitczenko, NH Feldman, Naturalistic speech supports distributional learning across contexts. *Proc. Natl. Acad. Sci.* **119**, e2123230119 (2022).
26. TA Nguyen, et al., The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint arXiv:2011.11588* (2020).
27. Avd Oord, Y Li, O Vinyals, Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
28. M Lavechin, et al., Statistical learning models of early phonetic acquisition struggle with child-centered audio data. *PsyArXiv* (2022).
29. A Cristia, A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations (https://osf.io/c86ew/?view_only=f9af0cf7d2574234a8517c38151e4210) (2019).
30. J Kearns, Librivox: Free public domain audiobooks in *Reference Reviews*. (Emerald Group Publishing Limited), (2014).
31. A Brunault, C Pitton, Litterature audio (2007).
32. M Riviere, A Joulin, PE Mazaré, E Dupoux, Unsupervised pretraining transfers well across languages in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE), pp. 7414–7418 (2020).
33. T Schatz, et al., Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*. pp. 1–5 (2013).
34. R Ardila, et al., Common voice: A massively-multilingual speech corpus in *Language Resources and Evaluation Conference (LREC)*. (2020).
35. M de Seyssel, M Lavechin, Y Adi, E Dupoux, G Wisniewski, Probing phoneme, language and speaker information in unsupervised speech representations. *ArXiv* **abs/2203.16193** (2022).
36. D Povey, et al., The kaldi speech recognition toolkit in *Automatic Speech Recognition and Understanding (ASRU) workshop*. (IEEE Signal Processing Society), (2011).
37. G Le Godais, T Linzen, E Dupoux, Comparing character-level neural language models using a lexical decision task in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. pp. 125–130 (2017).
38. E Keuleers, M Brysbaert, Wuggy: A multilingual pseudoword generator. *Behav. research methods* **42**, 627–633 (2010).
39. RL Weide, The cmu pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict* (1998).
40. Avd Oord, et al., Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).
41. PK Kuhl, et al., Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philos. Transactions Royal Soc. B: Biol. Sci.* **363**, 979–1000 (2008).
42. E Bergelson, D Swingley, At 6–9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci.* **109**, 3253–3258 (2012).
43. M Sundara, L Polka, F Genesee, Language-experience facilitates discrimination of/d- /in monolingual and bilingual acquisition of english. *Cognition* **100**, 369–388 (2006).
44. CT Best, , et al., The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development speech perception: The transition from speech sounds to spoken words* **167**, 233–277 (1994).
45. B Ambridge, E Kidd, CF Rowland, AL Theakston, The ubiquity of frequency effects in first language acquisition. *J. child language* **42**, 239–273 (2015).
46. B Conboy, M Rivera-Gaxiola, L Klarman, E Aksoylu, PK Kuhl, Associations between native and nonnative speech sound discrimination and language development at the end of the first year in *Supplement to the proceedings of the 29th Boston University conference on language development*. (2005).
47. BT Conboy, JA Sommerville, PK Kuhl, Cognitive control factors in speech perception at 11 months. *Dev. psychology* **44**, 1505 (2008).
48. FM Tsao, HM Liu, PK Kuhl, Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child development* **75**, 1067–1084 (2004).
49. PK Kuhl, BT Conboy, D Padden, T Nelson, J Pruitt, Early speech perception and later language development: Implications for the" critical period". *Lang. learning development* **1**, 237–264 (2005).
50. TC Zhao, O Boorom, PK Kuhl, R Gordon, Infants' neural speech discrimination predicts individual differences in grammar ability at 6 years of age and their risk of developing speech-language disorders. *Dev. Cogn. Neurosci.* **48**, 100949 (2021).
51. G Alain, Y Bengio, Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).
52. Y Belinkov, Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguist.* **48**, 207–219 (2022).
53. R Shi, Perception of function words in preverbal infants in *10th International Congress for the Study of Child Language, Berlin, Germany*. (2005).
54. C Fisher, SL Klingler, Hj Song, What does syntax say about space? 2-year-olds use sentence structure to learn new prepositions. *Cognition* **101**, B19–B29 (2006).
55. S Bernal, J Lidz, S Millotte, A Christophe, Syntax constrains the acquisition of verb meaning. *Lang. learning development* **3**, 325–341 (2007).
56. A Christophe, S Millotte, S Bernal, J Lidz, Bootstrapping lexical and syntactic acquisition. *Lang. speech* **51**, 61–75 (2008).
57. S Peperkamp, R Le Calvez, JP Nadal, E Dupoux, The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* **101**, B31–B41 (2006).
58. A Seidl, R Tincoff, C Baker, A Cristia, Why the body comes first: Effects of experimenter touch on infants' word finding. *Dev. science* **18**, 155–164 (2015).
59. B Jones, M Johnson, MC Frank, Learning words and their meanings from unsegmented child-directed speech in *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*. pp. 501–509 (2010).
60. K Lakhotia, et al., On generative spoken language modeling from raw audio. *Transactions Assoc. for Comput. Linguist.* **9**, 1336–1354 (2021).
61. E Kharitonov, et al., Text-free prosody-aware generative spoken language modeling in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 8666–8681 (2022).
62. TA Nguyen, et al., Generative spoken dialogue language modeling. *arXiv preprint arXiv:2203.16502* (2022).
63. Z Borsos, et al., Audiolm: a language modeling approach to audio generation. *arXiv e-prints* pp. arXiv–2209 (2022).
64. R Abu-Zhaya, A Seidl, R Tincoff, A Cristia, Building a multimodal lexicon: Lessons from infants' learning of body part words in *Proc. GLU 2017 International Workshop on Grounding Language Understanding*. pp. 18–21 (2017).
65. C Yu, DH Ballard, RN Aslin, The role of embodied intention in early lexical acquisition. *Cogn. science* **29**, 961–1005 (2005).
66. K Nelson, *Young minds in social worlds: Experience, meaning, and memory*. (Harvard University Press), (2007).
67. NH Feldman, S Goldwater, E Dupoux, T Schatz, Do infants really learn phonetic categories? *Open Mind* **5**, 113–131 (2022).

Supplementary material

1. Proposed model

In this section, we described the proposed model which corresponds to the low-budget baseline architecture from the zero resource challenge 2021 (1).

A. Acoustic model.

A.1. Training objective. As originally proposed in (2), we used a contrastive loss which forces the latent space to retain information that is useful to predict future samples. Precisely, the input sequence of observations x_t is mapped to a sequence of latent representations through an encoder g_{enc}, such that $z_t = g_{enc}(x_t)$. Then, all $z_{\leq t}$ are aggregated with an auto-regressive model that produces a context-dependent latent representation $c_t = g_{ar}(z_{\leq t})$. Given the past context c_t, a predictor g_{pred} is asked to predict future representations z_{t+k} for $k \in \{1, ..., K\}$. Given a set $X = \{x_1, ..., x_n\}$ of N random samples containing one positive sample from the true positive distribution $p(x_{t+k} \mid c_t)$ and $N-1$ negative samples from the proposal negative distribution $p(x_{t+k})$, we optimize the categorical cross-entropy loss of classifying the positive sample correctly:

$$\mathcal{L} = -\frac{1}{K}\sum_{k=1}^{K}\log\left[\frac{\exp\left(g_{pred}(c_t)^T z_{t+k}\right)}{\sum_{x_j \in X}\exp\left(g_{pred}(c_t)^T z_j\right)}\right]$$

On top of the context-dependent representations c_t, we train a simple K-means algorithm to minimize the within-cluster sum of squares:

$$\mathcal{L} = \sum_{k=1}^{K}\sum_{c_{t,i} \in S_i} d(c_{t,i}, \mu_i)$$

where K is the number of clusters, S_i is the set of points belonging to the i^{th} cluster for $i \in [1..K]$, μ_i is the centroid of points in S_i, d is a distance function defined on the context-dependent representations c_t.

A.2. Implementation details. As proposed in (3), the encoder g_{enc} consists of a 5-layer convolutional neural network with kernel sizes $[10, 8, 4, 4, 4]$ and strides $[5, 4, 2, 2, 2]$ that returns a 256-dimensional vector every 10 milliseconds. The auto-regressive model g_{ar} is a 2-layer long short-term memory network of dimension 256. The model is asked to predict up to $K = 12$ time steps in the future (which is equivalent to 120 ms). The predictor g_{pred} is a single multi-head transformer layer with $K = 12$ heads, each predicting at time step $k \in \{1, ..., 12\}$. Negative samples are drawn from sequences that are temporally close to the sequence the positive sample are drawn from. More precisely, creating a batch consists of selecting 64 successive sequences in the case of the domain-general learner (or 64 successive sequences that have been pronounced by the same speaker for the domain-specific learner). For a current sequence seq_i, negative samples are taken from all other sequences seq_j, with $j \neq i$. All models have been trained on 8 GPUs with batches of 64 sequences, and each sequence has a duration of 1.28 seconds. All models are trained until convergence, and the best epoch is selected according to validation loss (5% of the original training set).

The K-means algorithm was trained with $K = 50$ using a euclidean distance function. All K-means were trained online with 200 sequences of 0.64 seconds using 1 GPU. All models are trained until convergence. At inference time, the input 10ms-frame is assigned the cluster label whose centroid is the closest.

B. Language model: LSTM.

B.1. Training objective. We train a language model on the discretized version of the audio files returned by the Acoustic Model. The Language Model is trained to predict the next unit of a sequence given its past context via a cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} y_{t,k}\log(\hat{y}_{t,k})$$

where T is the length of the input sequence, K is the number of clusters, $y_{t,k}$ is the real cluster at time t, and $\hat{y}_{t,k}$ is the predicted probability at time-step t for cluster k.

B.2. Implementation details. The language model is a 3-layer LSTM with an embedding layer of size 200, hidden layers of size 1024 and a feed-forward output layer of size 200. We used the implementation proposed in (4).

2. Analysis: the training objectives computed on the evaluation set

A. Experimental protocol. We compute the training objectives of the Acoustic model and the Language model on the set of audio files used in the ABX discrimination task. Note that these audio files have never been seen during training.

B. Results. Figure S1 shows the 3 training objectives averaged across the native models (English evaluated on English, French evaluated on French) for the Acoustic Model and the Language Model. Results indicate that the higher the quantity of speech in the training set, the lower the test losses. This indicates a positive effect of exposure on the training objectives. This result is achieved via gradient descent.

3. Analysis: Detailed phonetic and lexical scores

Detailed phonetic and lexical scores are presented in Figure S2. Here, the scores are presented separately for each of the English and French test set.

When tested on English, both phonetic and lexical results follow the trends discussed in the main paper. When tested on French, however, the phonetic scores yielded by the English (non-native) models are closer to the scores yielded by the French (native) models. For the lexical task, the French (native) models do yield higher scores than the English (non-native) models, but the difference between the two curves is lesser than the one observed on the English test set.

Such results could indicate a potential asymmetry between languages. Although it is difficult to provide precise evidence, the fact that the English model tested on the French lexical task (bottom right graph) gets progressively above chance with input quantity might be explained by the high number of cognates and loanwords in English.

Yet, one should stay cautious about such comparisons. As mentioned in the Results Section, such patterns can also be the effect of the training set themselves. For example, (5) found that the presence of non-speech in such models can deteriorate their quality, and it is possible that such differences exist between the French and English training sets (with one being noisier than the other). This is why we recommend focusing on results aggregated symmetrically over the native and non-native conditions, as presented in the results. Still, further work could focus on potential asymmetries between languages.

4. Analysis: Phonetic scores predict lexical scores

A. Experimental protocol. For this analysis, we consider models trained on 50h, 100h, or 200h of English or French speech. We evaluate their phonetic score using the ABX discrimination task, and their lexical score using the spot-the-work task, both described in the Methods Section. Both scores are evaluated in the native condition, i.e. on the training language. Lexical scores are computed either on: 1) words belonging to the 64th frequency band (high frequency words); 2) words belonging to the 1st frequency band (low frequency words) or 3) as the average accuracy across all frequency bands.

B. Results. Figure S3 shows the correlation between the phonetic score and the lexical score obtained by individual models for different training set sizes (column-wise) and for a lexical score computed on different frequency bands (row-wise). Results indicate that, in general, models that are less accurate at discriminating native sounds, are less good in the spot-the-word task. This effect seems more important on high frequency words as shown by the 50h English model that exhibits a Pearson's R correlation coefficient of .52 ($p<.0001$) on high frequency words, .36 ($p<.05$) across frequency bands, and .12 (non-significant) on low frequency words. While the 100-hours and the 200-hours English models seem to exhibit a similar pattern, models trained on French speech show more constant correlation scores across the different frequency bands.

5. The emergence of latent linguistic structure

A. Layer-wise LOO classification scores for the lexical analyses. Leave-One-Out classification scores for the function vs. content (FC) and part-of-speech (POS) categories were computed on a development set for all hidden layers of the 3200h English and French models (see Methods). Results are presented in Table 2. Layer 3 yielded the best scores overall for both the English and French models, and was subsequently chosen to carry out the lexical probing analyses.

B. Results on the French models. In the Results Section, we presented qualitative and quantitative analyses of the emergence of phonetic and lexical categories in the English models. The same analyses on the French models are presented in Figure S4. Experimental methods are the same as described in the Methods).

As for the English model, qualitative analyses carried out on the French 3200h model suggest that this model clearly encodes information about sonority, place and voicing, with the categories being visually well separated (panel a). Moreover, all of these three types of information get progressively better encoded with more training data (panel b). Interestingly, contrary to results on the English models, even the voicing information present in the models gets significantly better.

Regarding the emergence of the lexical and proto-syntactic categories, the patterns are the same as for the English models. No clear categories of function vs content and Part of Speech (POS) can be qualitatively distinguished from the t-SNE(s) on the 3200h French model (panel c). Yet, probing analyses carried out on all models show that this categorisation can be better learnt with models trained on more data, suggesting that this information gets gradually better encoded (panel d). The main simplifying assumption regarding the word segmentation problem in this work is that utterances are represented as strings of phonemes. Any computational model comes with its set of simplifying assumptions, which is fine. However, the authors should discuss this in more detail. In particular, the assumption mentioned above is problematic for two reasons. First, this assumes that children can assign a single phoneme to each phone they hear in an error-free manner. However, evidence suggests that children segment some words way before their perception have reached that of an adult

6. Analysis: the frequency effect

We evaluate the Language model using the spot-the-word task described in the Methods. The lexical score obtained by the native model is displayed in the diagonal of Figure S5 (panels (a) and (d)). The anti-diagonal shows the lexical score obtained by the non-native model (panels (c) and (d)). The number of trials per frequency band is presented in Table 3. In the native condition, results indicate that the higher the quantity of speech, the higher the lexical score, showing a positive effect of exposure. We only observe a slight increase in the non-native condition, which suggests that the non-native model is mostly unable to solve the lexical task. The positive effect of exposure in the native condition seems more important on high-frequency words than low-frequency words (native curves are steeper as the frequency increases).

7. Analysis: the emergence of lexical factors

A. Dataset. We use the Massive Auditory Lexical Decision (MALD) dataset (6) that contains reaction times of human participants on the auditory lexical decision task. In this psycholinguistic task the participant hears an audio stimuli and has to classify it as either a word or a nonword. The MALD contains reaction times for 26,793 words and 9592 nonwords. This sums up in reaction times for 227,179 auditory lexical decisions from 231 unique monolingual English listeners. In addition to reaction times, each stimuli is annotated for various lexical descriptors: the duration of the stimuli, the frequency of the stimuli, the number of phonological neighbors, the phone index of the phonological uniqueness point of the stimuli within the CMU-A dictionary (7), the mean phone-level Levenshtein distance of the item from all entriers within the CMU-A, etc. A detailed description of all descriptors can be found in (6). All data on nonwords were discarded and only words were included in the present analysis.

B. Experimental protocol. We compute the probability of each word of the MALD dataset with the Language Model, and look at which lexical factors are significant predictors of this probability. We do so using a nested linear regression analysis.

We first start with the predictor that leads to the highest R^2. Then, we add the second predictor that increases the R^2 in the most significant way (i.e., the selection criterion is the p-value such as computed by a likelihood-ratio test). We do until the addition of predictor does not yield a significant increase in R^2.

We run the same analysis with human reaction times and then compare the lexical factors for both target: pseudo-probabilities returned by the Language Model, and human reaction times.

C. Results. Panel (a) of Figure S6 shows the various descriptors (duration, frequency, phonological features, part-of-speech categories, etc.) that are: 1) significant predictors of human reaction times (in green); 2) significant predictors of the Language Model probability (in red); 3) both 1) and 2) (intersection of green and red surfaces); or 4) not significant for both human reactions times and pseudo-probabilities (in white).

Results indicate that the duration and the frequency of the word are significant predictors of both the human reaction time and the Language Model probability. PhonND which indicates the number of phonological neighbors (defined as one phone edit away) for the word within the CMU-A dictionary, and PhonUP which indicates the phone index of the phonological uniqueness point of the item within the CMU-A are also significant predictors of both the human reaction time and the Language Model probability. Significant predictors of the Language Model probability capture 31% of its variance, while significant predictors of the human reaction time capture 26% of its variance.

Panel (b) of Figure S6 shows the R^2 obtained by the nested linear regression models as a function of quantity of speech in the training set. The blue curve corresponds to a linear model containing only Duration as a predictor, the orange curve both Duration and PhonLev (the mean phone-level Levenshtein distance to all entries within the CMU-A dictionary), etc. Results indicate that the higher the quantity of speech in the training set, the higher the R^2 obtained by the different nested models. In other words, as the Language Model receives more speech, the abovementioned linguistic factors become more predictive of the probability.

8. Analysis: the emergence of word boundaries

In this analysis, we look into whether the emerging grammatical structure learned by our model is grounded on some notion of words or morphemes as a cohesive sequence of phonemes. In a seminal paper, Elman (8) presented a language model trained on letters and discovered that the probability assigned at each time step gradually increases inside words and sharply decreases between words. This important result suggests that the language model trained on letters implicitly performs word segmentation, although the model is not provided with breaks. In this section, we perform a similar analysis, with, contrary to Elman, our language model that is trained from the raw acoustic input.

A. Experimental Protocol. The analysis shows the probability assigned by the language model as the sentence unfolds over time. We consider either words or sentences from the Common Voice audio files that have also been used in the ABX discrimination task, and that have never been seen during training.

B. Results. Figure S7 present behaviors of the Language Model probability as the audio unfolds over time. The model considered was trained on 3200 hours of English speech.

Panel (a) shows probability curves as a function of length rank (1st rank contains shortest words, 10th rank contains longest words). Probability curves are linearly interpolated so that each word belonging to the same length rank share the same target length (median length for this rank). Results show a length effect, with the probability increasing as the word unfolds over time. Sharp decreases in the beginning and the end of words can be noticed which seems to indicate that the Language Model has a harder time predicting the next token on word boundaries.

Panel (b) shows a similar analysis on sentences. Sentences are sorted depending on their number of words, linearly interpolated so that sentences with the same number of words share the same target duration (median duration)), and averaged. Results show a sharp increase in the probability at the beginning of sentences, then a slight decrease as the sentence unfolds over time. A sharp increase can be noticed at the end of sentences, which indicates that the Language Model is better at predicting the next token at the end of sentences than at the beginning/middle.

Panel (c) shows the probability for the sentence "*I can see a smiling face in the clouds*" (in grey) and the probability estimated by averaging N=500 words of the same size (in red). Results indicate a noisy behavior when considering a single stimuli, despite having applying a moving average of 10 frames (100 ms). However, when considering the average profile of the probability, we notice that the probability slightly increases inside words, and sharply decreases between words.

We draw on the same conclusion than Elman: the language model seems sensitive to word boundaries, but only when averaging across hundreds of inputs. The acoustic variability infants are facing bring a much more difficult problem: normalizing the input across the various acoustic dimensions (speaker, speech rate, etc.)

9. Analysis: the learned units

A. Experimental protocol. Here, we compare the discrete units learned by the K-means algorithm to phones as recognised by an Automatic Speech Recognition (ASR) algorithm. To compute the ASR phones, we use the MLS speech corpus (9), which is an aggregation of read speech taken from the LibriVox project (10). For each language, we select 100h of speech data. We first train a phone bigram language model on each training set using the SRILM toolkit (11). We then train for each language a hybrid GMM-DNN phone recogniser based on a time-delay neural network architecture (12), adapting the s5 librispeech recipe from the Kaldi speech recognition toolkit (13). Finally, we infer ASR phones for our English and French Common Voice test sets using the English and French newly trained phone recognizers respectively*.

A.1. Analyses. We can now compare how K-means units and ASR phones compare to the gold phones from the test set. For each model, we compute $p2u$ (phone-to-unit), the perplexity of gold phones given the ASR phones or the K-means units distribution, and $u2p$ (unit-to-phone), the perplexity of ASR

*The phone accuracy yielded by the phone recognisers on the English and French test sets of 24.7 and 24.6% respectively.

Supplementary references

1. TA Nguyen, et al., The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint arXiv:2011.11588* (2020).
2. Avd Oord, Y Li, O Vinyals, Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
3. E Kharitonov, et al., Data augmenting contrastive learning of speech representations in the time domain in *Spoken Language Technology Workshop (SLT)*. (2021).
4. M Ott, et al., fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038* (2019).
5. M Lavechin, et al., Statistical learning models of early phonetic acquisition struggle with child-centered audio data. *PsyArXiv* (2022).
6. BV Tucker, et al., The massive auditory lexical decision (mald) database. *Behav. research methods* **51**, 1187–1204 (2019).
7. RL Weide, The cmu pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict* (1998).
8. JL Elman, Finding structure in time. *Cogn. science* **14**, 179–211 (1990).
9. V Pratap, Q Xu, A Sriram, G Synnaeve, R Collobert, Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411* (2020).
10. J Kearns, Librivox: Free public domain audiobooks in *Reference Reviews*. (Emerald Group Publishing Limited), (2014).
11. A Stolcke, Srilm-an extensible language modeling toolkit in *Seventh international conference on spoken language processing*. (2002).
12. V Peddinti, D Povey, S Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts in *Sixteenth annual conference of the international speech communication association*. (2015).
13. D Povey, et al., The kaldi speech recognition toolkit in *Automatic Speech Recognition and Understanding (ASRU) workshop*. (IEEE Signal Processing Society), (2011).
14. T Schatz, NH Feldman, S Goldwater, XN Cao, E Dupoux, Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proc. Natl. Acad. Sci.* **118**, e2001844118 (2021).

Table 1. Evaluated phonetic inventory in Metropolitan French and American English in International Phonetic Alphabet (IPA) standard.

Manner of articulation	Metropolitan French	American English
	Consonants	
Stops:	p,b,t,d,k,g	p,b,t,d,k,g
Nasals:	m,n,ɲ	m,n,ŋ
Fricatives:	f,v,s,z,ʃ,ʒ,ʁ	f,v,θ,ð,s,z,ʃ,ʒ,h
Approximants:	j,w,l	j,ɹ,w,l
Affricates:	ʤ	ʤ,ʧ
	Vowels	
Oral	i,y,e,ø,œ,ɛ,a,ə,ɔ,o,u	i,ɪ,ɛ,æ,ɚ,ʌ,ɐ,u,ʊ,ɔ,ɑ
Nasal:	ɑ̃, ɛ̃, œ̃,ɔ̃	
Diphtongs:		aɪ,ɔɪ,aʊ,eɪ,oʊ

Table 2. Leave-One-Out Classification Scores (CS). Scores are computed on the English and French 3200h models using the dev sets for the function vs content (FC) and part-of-speech (POS) categories classification tasks. Best average classification scores are indicated in bold.

Language	Hidden layer	FC CS	POS CS	Average CS
English	1	57.07	46.15	51.61
English	2	58.26	50.57	54.41
English	3	60.42	55.89	**58.16**
French	1	61.37	39.21	50.29
French	2	62.91	47.41	55.16
French	3	66.34	45.43	**55.89**

Table 3. Number of trials in the spot-the-word task. The numbers have to be divided by 4 (number of synthesised voices) to get the number of word/nonword pairs.

	Frequency band						
test language	**1st (rare)**	**2nd**	**4th**	**8th**	**16th**	**32th**	**64th (frequent)**
English	70,136	60,664	49,324	40,204	28,132	17,544	15,108
French	51,956	53,700	42,944	32,032	23,168	16,336	12,976

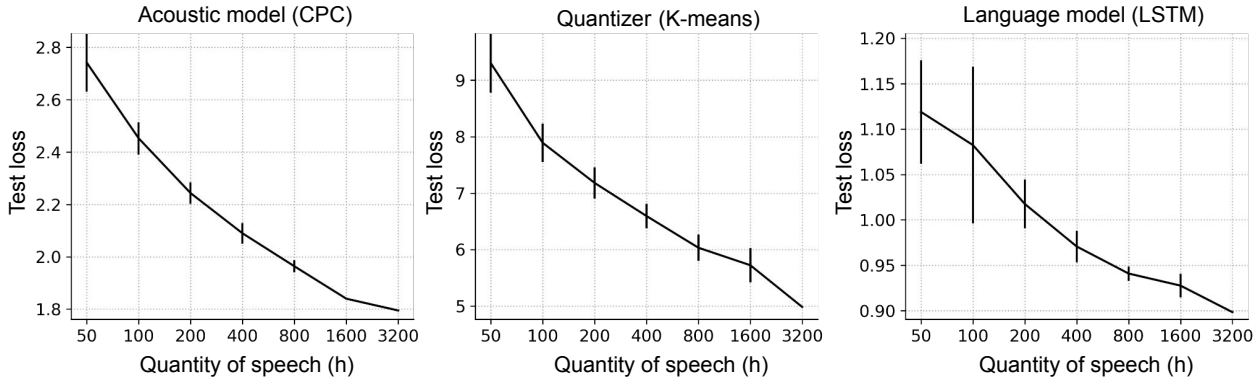

Fig. S1. Graduality and parallelism of the training objectives. The three losses, computed on the test set, for the 2 components of our model: the acoustic model minimizes the cross-entropy of classifying the positive sample correctly (contrastive predictive coding); and the within-cluster sum of squares (K-means); the language model minimizes the cross-entropy of predicting the next token correctly.

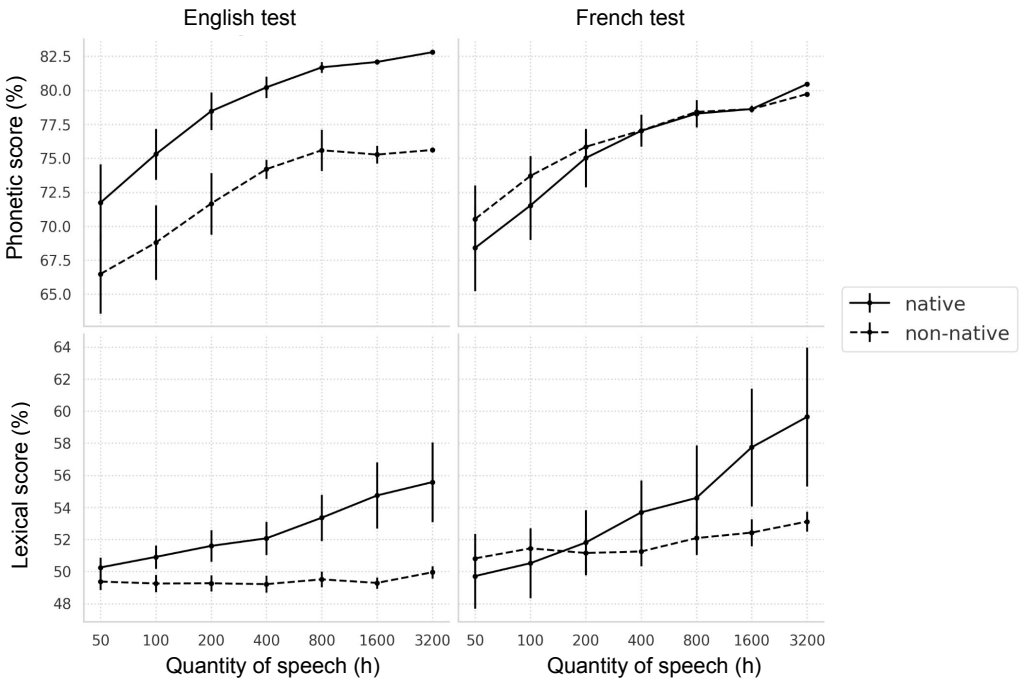

Fig. S2. Phonetic and lexical scores per training and test languages. Phonetic and lexical scores are presented on both English and French test sets, separately for each trained language. Phonetic scores are presented on the top row and lexical scores on the bottom row. On the left column, we show scores calculated on the English test set, and on the right column, scores calculated on the French test set. For the lexical scores, scores are first averaged over each frequency band then per training size. Error bars for the phonetic and lexical scores correspond to the standard deviation between the averaged scores for all models of a same training size and language.

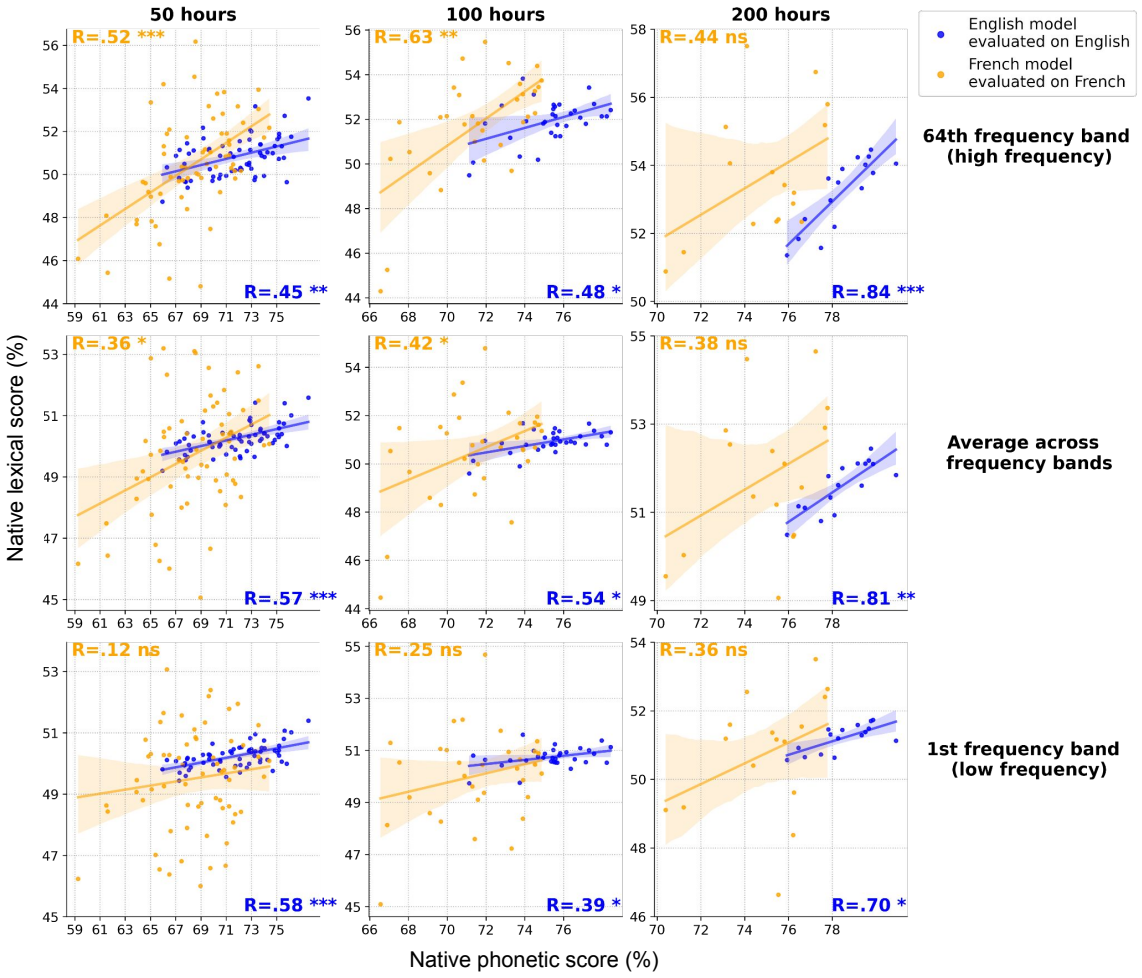

Fig. S3. Phonetic scores predict lexical scores. Correlation between the phonetic and lexical scores across English (in blue) and French (in orange) models trained on 50h (first column), 100h (second column) and 200h (third column) of speech. The lexical score is evaluated on the high frequency words (first row), the average across frequency bands (second row), or low frequency words (third row). R is the Pearson correlation coefficient. Significance levels: na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

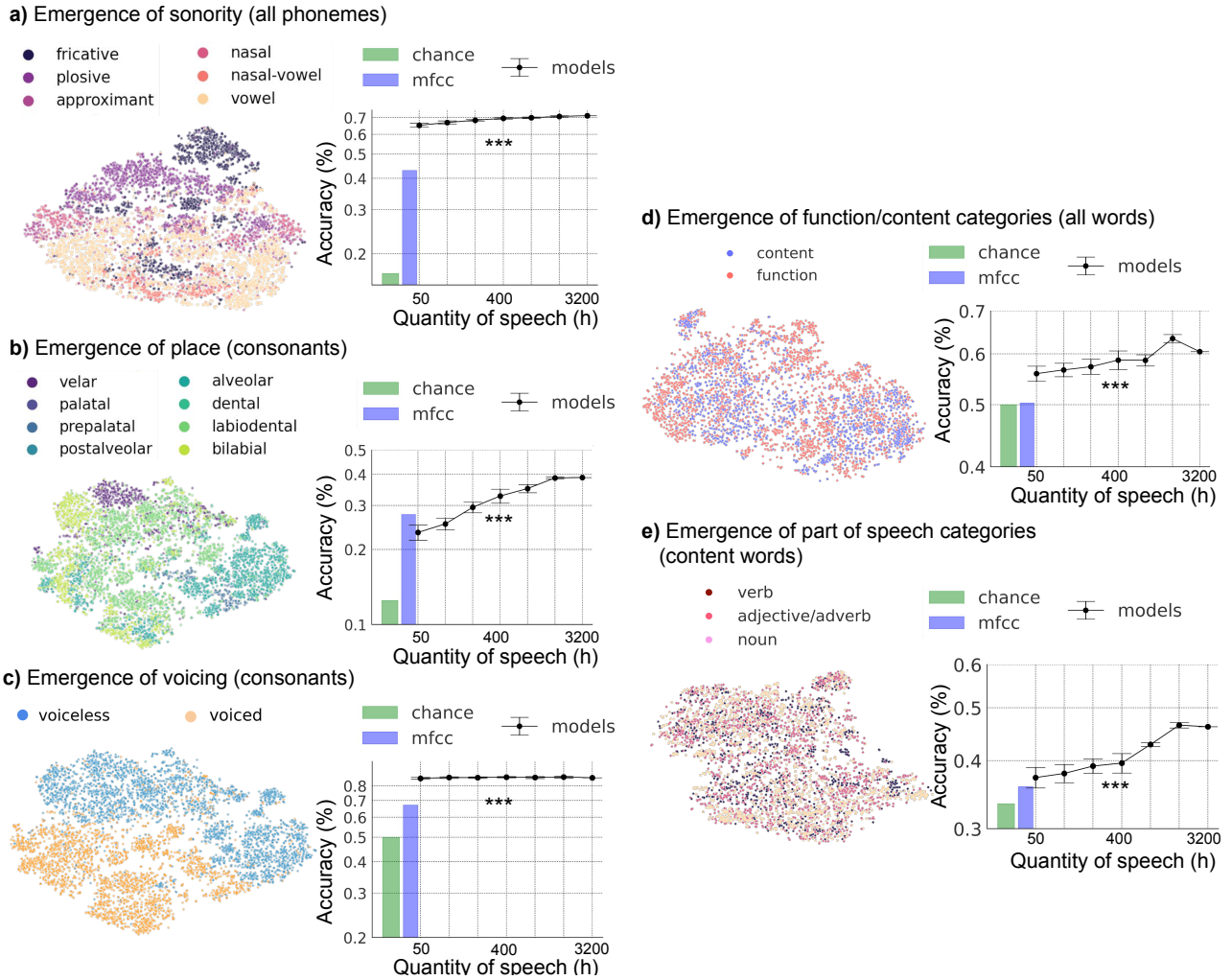

Fig. S4. Emergence of latent linguistic structures at the phonetic level for the French models. Left: tSNEs of the continuous representations of the acoustic model (last layer) pooled within phonetic tokens in a test set, according to sonority (a), place (b) and voicing (c) for the 3200h English model. with their corresponding developmental curves of leave-one-phoneme-type-out classification errors as a function of input quantity (taking all 256 dimensions into account). Chance level and MFCCs performances are also given. Right: tSNEs of the continuous representations of the language model (last layer) pooled over words tokens according to (d) function/content distinction and (e) part of speech for the 3200h English model with their corresponding developmental curves of leave-one-word-out classification error as a function of input quantity (taking all 1024 dimensions into account). Chance level and MFCCs performances are also given. The asterisks indicate a significant correlation of classification error and input quantity. na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

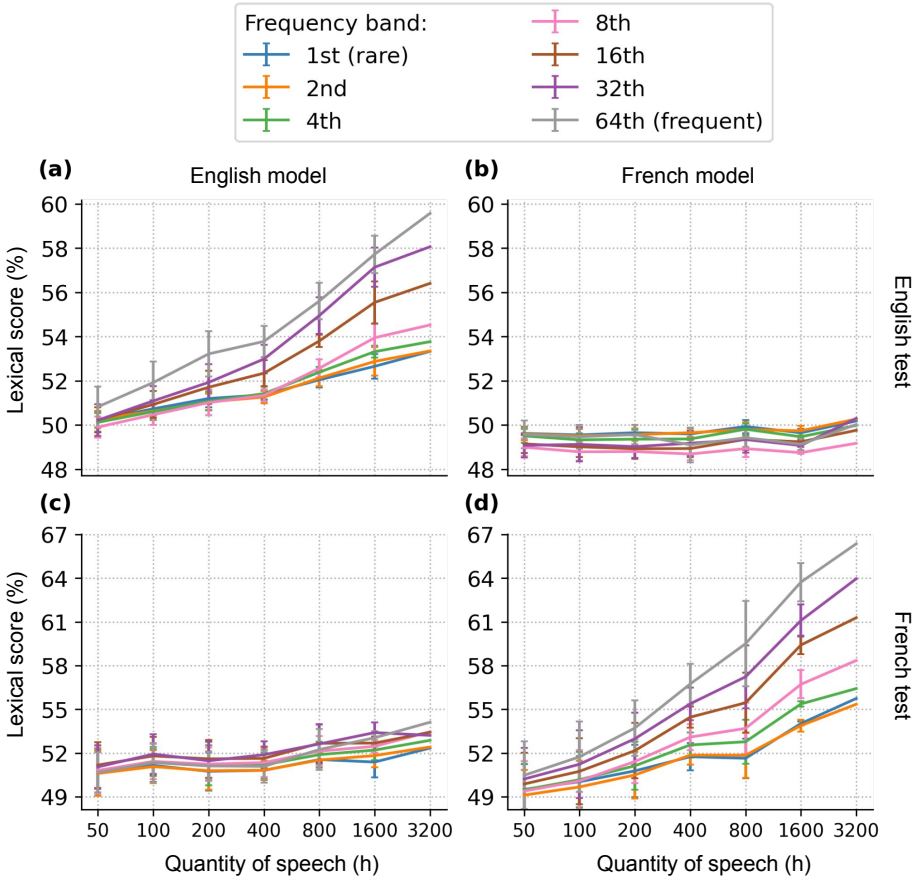

Fig. S5. The frequency effect on the lexical task. Effects of input frequency and input quantity on phonetic and lexical tasks. Left: Right: Lexical scores obtained by English (first column) and French (second column) models on the English (first row) and the French (second row) lexical test. Panels (a) and (d) on the diagonal show lexical scores obtained by native models. Panels (b) and (c) on the anti-diagonal show lexical scores obtained by non-native models. Scores are given as functions quantity of speech available in the training set, and class of frequency of evaluated words. Words in the 64th class of frequency are present at least one time in the 50-hours training sets, two times in the 100 hours, four times in the 200 hours. Words belonging to the 32th class of frequency are present at least one time in the 100 hours training sets, 2 times in the 200 hours, etc. Error bars represent standard errors computed across mutually exclusive training sets whose number depends on the quantity of data available. The last data point along the x-axis is computed on a single learner (trained with all available data), then the number of learners doubles with each step along the x-axis, as the quantity of audio is divided by two.

2.4 Conclusion

In this chapter, we reviewed some of the early developmental milestones in infants' language acquisition and gave a bird's eye view of the methodological landscape in computational models of language acquisition. We presented our simulation, STELA, an implementation of the statistical learning hypothesis at the core of learning theories in developmental sciences (Saffran & Kirkham, 2018). Our simulation constitutes evidence in favor of the statistical learning hypothesis to bootstrap early phonetic and lexical learning.

To conclude, we highlight some limitations of our approach and reflect on potential future work.

Quantitative comparison against empirical data. Where are we with respect to the comparison against empirical data advocated in Section 2.2.3? In our simulation, we evaluate how well the model's language capabilities align with the available empirical data in infants. However, this comparison remains limited to a qualitative level, despite our evaluation framework providing quantitative measures. In other words, we are not trying to predict or capture a certain proportion of variance of infant empirical data. Although this limitation has little bearing on our conclusions, it is important to understand the underlying reasons.

First, infant behavioral data are influenced by various factors in ways that are not fully understood, often resulting in noisy measures. For example, behavioral measures can be influenced by the experimental protocol used to elicit responses and by other cognitive systems not considered in our simulation, like memory or attention – see Ambridge and Rowland (2013) for relevant discussions. Second, infant data are often sparse. For example, the majority of studies about speech sound discrimination in infants only cover a few languages, and within those languages, only a few specific contrasts, mostly consonantal, are examined (Tsuji & Cristia, 2014). It is important to note that researchers are not at fault here, as the methodological difficulties in experimenting with infants are particularly great. Nonetheless, these limitations often result in behavioral studies focusing on statistically significant differences between experimental conditions (typically across two age ranges), which are are then used to draw conclusions on infants' language capabilities. Due to the aforementioned reasons, it remains difficult to make a quantitative comparison with infant empirical data. However, keep an eye on Jing Liu's ongoing work on comparing models against parental questionnaires using the Child Development Inventory (CDI, Ireton, 1992). Another promising direction proposed in Blandón

et al. (2021) consists in comparing models against robust empirical data from meta-analyses.

Although some of the limitations above also apply to adult empirical data, a quantitative comparison with adults appears to be within reach. This is evidenced by numerous studies comparing the performance of humans and machines on the ABX sound discrimination task Millet et al. (2019) and Millet and Dunbar (2020, 2022). While I am unaware of any studies performing a similar comparison on the spot-the-word task, this has been done on the auditory lexical decision task, e.g., Brand et al. (2021) or Nenadić et al. (2022).

Evaluating language capabilities in their full complexity. In STELA, we adopt a methodology inspired by experimental psycholinguistics to evaluate the language capabilities of our learners using an ABX sound discrimination and a spot-the-word task. While these tasks have the advantage of being theory-agnostic, as seen in Section 2.2.3, they only measure a fragment of phonetic and lexical capabilities.

Regarding phonetic capabilities, other tasks could aim at evaluating the invariance of the learned representations with respect to: 1) competing noises, by sampling triphones across background noises; 2) speaker identity, by sampling triphones across speakers as done in T. A. Nguyen et al. (2020); or 3) co-articulation effects, by sampling triphones across phonetic contexts as done in Hallap et al. (2022).

Similar complexities arise to evaluate lexical capabilities. *What does it mean to learn a word?* Is it the ability to differentiate it from pseudowords, as done by Ngon et al. (2013)? Is it the capacity to segment it from continuous speech, as done by Jusczyk and Aslin (1995)? Or is it the capacity to associate it with the correct referent, as done by Bergelson and Swingley (2012)? These experiments conducted with infant participants presumably all measure different aspects of lexical capabilities. Similarly, it may be necessary to implement different tasks when assessing language capabilities in the machine, especially if one wants to understand the interplay between the various possible measures.

In the same light of thought, (early) language acquisition encompasses more than simply acquiring knowledge about the phonetic and lexical aspects of one's native language, and language development is known to occur through a series of developmental cascades with mutually interacting systems (Iverson, 2021; Guo et al., 2023). Although science often proceeds by breaking down large problems into smaller ones, the intricacies and co-dependencies inherent to human languages might well require us to study the problem in its full complexity, as advocated in de Seyssel, Lavechin, and Dupoux (2022). Our simulation, STELA, represents one step forward in this

direction by addressing the joint problem of early phonetic and lexical learning, but joint models are still sparse in the computational modeling literature – see Elsner et al. (2012) for phonetic and lexical learning; Khorrami et al. (2023) for phonetic, lexical, and semantic; or Abend et al. (2017) for syntax and semantic. Assessing speech perception and production, from low-level phonetic to high-level pragmatic aspects, will require the development of a range of psycholinguistic tasks tailored for machines. Several proposals outlining such tasks are discussed in the following chapter, Section 3.1.

Assumptions regarding the environment model. Every modeling study makes different assumptions and compromises regarding the input material available to the artificial learner. For instance, some use phonetically transcribed sentences from children's language environments (B. Jones et al., 2010; Cristia, Dupoux, et al., 2019), abstracting away speech variability (e.g., pace, speaker variability, co-articulation effects, etc.). Others, like us, use raw untranscribed sentences extracted from audiobooks which has the undeniable advantage of being more realistic with respect to speech variability but also deviate further from the actual language input received by children. These considerations give rise to a fundamental question: *How can we be sure that our findings generalize to the actual learning problem solved by infants?*

Children's language experiences are significantly more complex than an error-free string of phonemes and greatly vary from audiobooks recorded under controlled conditions. And there are numerous ways in which assumptions about the learning environment may simplify the learning problem and fail to generalize to the real world.

For instance, modeling the learning process over an error-free string of phonemes is akin to assuming that: 1) infants have access to phone categories and their boundaries, a hypothesis for which there is currently little evidence (Feldman et al., 2021); and 2) perception occurs in an error-free manner, which is unlikely the case, neither for infants nor adults.

While considering raw speech as input undoubtedly helps relax some of the assumptions exposed here, it is not devoid of other simplifying assumptions. Using long stretches of speech recorded under controlled conditions (e.g., in a studio) is difficult to reconcile with what infants truly hear, namely, short utterances that can be produced far from the child and distorted by various background noises. Similarly, in real life, people do not speak in full and well-articulated sentences as in audiobooks but may speak in ways that distort the speech signal: people

may produce short turns that sometimes overlap across speakers, and they may under-articulate, mumble, shout, whisper, sing, or laugh while speaking.

In the best-case scenario, using recordings collected under controlled conditions to abstract away the difficult acoustic conditions that infants typically encounter is akin to placing high computational and perceptual demands on their abilities to normalize the input signal across irrelevant dimensions and discard non-linguistic information. But in the worst case, it can lead us to consider a different learning problem altogether – see Clerkin et al. (2017) for a thought-provoking study on how word-referent statistics differ in real infant egocentric scenes from that of training sets typically used in computational modeling studies.

The remainder of this manuscript is dedicated to revisiting the simplifying assumptions pertaining to the input, which are at the core of the modeling enterprise in infant language acquisition studies. Namely, using child-worn microphones encountered in Chapter 1, we propose to feed our learning algorithms directly with what infants hear.

3

Modeling language acquisition from child-centered long-form recordings

A key mission in science in general, and modeling studies in particular, is to understand the extent to which our findings generalize to the real world. Regarding language acquisition, the conditions infants face differ widely from those typically considered in most modeling studies.

Infants learn their native language through exposure to everyday language use, which is filled with many peculiar phenomena (filled pauses, word repetition, specific vocabulary, ungrammatical constructions, etc.). Furthermore, adults often engage in unique forms of communication when interacting with young children. This type of communication is referred to as child-directed speech, and it follows its own rules with specific prosodic, phonological, lexical, and syntactic properties – see Soderstrom (2007) for a review. Besides, infants face various listening conditions that may facilitate or impede their speech perception. A large proportion of the speech received by infants is not directly addressed to them but to other adults. And even when speech is directed to infants, it can be laced with various background noises, echoes, as well as the speech of other individuals.

How does the complexity of children's real language environment impact language acquisition? Do existing theories adequately account for what children truly hear? Do our computational models exhibit the same learning outcomes when trained on carefully curated or ecological data? These questions are at the core of the present chapter.

Revisiting what might be the most critical simplifying assumption made by the vast majority of computational modeling studies, and by ourselves in Chapter 2, we propose a different approach: directly feeding computational models with the auditory input received by infants. We argue that using ecologically-valid input data allows us to tackle the learning problem in its full complexity and to derive more realistic predictions regarding how the infants' perception adapt to the language(s)

they hear. Additionally, by feeding our artificial language learner with input that faithfully represents the input children are exposed to, we can rule out effects that would reflect the specificities of the input itself. Thus, any match or mismatch between the artificial language learner and the human learner can be explained by other aspects of the computational experiment: 1) the learning mechanism; 2) the amount of data; or 3) the evaluation protocol.

We begin with a manifesto outlining how a research program centered around modeling language acquisition from realistic data could proceed. Putting our words into action, we present a simulation of early phonetic acquisition that illustrates the importance of considering ecologically-valid input data when modeling language acquisition. Next, we present an open-source benchmark to evaluate models trained on realistic input. Our benchmark uses zero-shot probing tasks to evaluate models at the lexical and syntactic levels and has been designed with the vocabulary typical of children's language experiences. To conclude, we present an ongoing project aimed at gaining a deeper understanding of the type of information learned by statistical learning algorithms trained on real-life audio recordings. Furthermore, we take a step back to reflect on the extent to which child-centered long-form recordings effectively capture the full sensory signal available to infants, suggesting potential avenues for improving the recording devices.

3.1 Reverse engineering language acquisition

Lavechin, M., de Seyssel, M., Gautheron, L., Dupoux, E., Cristia, A. (2022) Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*

Motivation

As seen in Chapter 1, child-centered long-form recordings profoundly impacted the field of language acquisition. These recordings collect the full range of communicative situations encompassing a child's entire day, offering a uniquely ecological view of the language input children are exposed to as well as their production. Although long-forms are now commonly used in fieldwork, they are vanishingly rare

in computational modeling studies. Building upon Dupoux (2018), we set recommendations on how a research program centered on modeling language acquisition using long-forms could proceed.

Paper summary

In Lavechin, de Seyssel, Gautheron, et al. (2022), we present how long-forms can be used in a reverse engineering approach to the study of language acquisition. In short, the reverse engineering approach involves designing scalable learning systems fueled with realistic language experiences. By observing what the system learns and how it develops, we can revise our algorithms and formulate hypotheses about how infants learn language.

The envisioned approach brings one question: *Why would one use a black box (the deep learning algorithm) as a proxy to study another black box (the infant)?* Indeed, language development researchers are faced with studying an intricate cognitive process whose inner workings and mechanisms still need to be understood. The same could be said of artificial intelligence researchers studying large models whose responses cannot be explained or interpreted, despite the model's state being fully known at any given time. In Lavechin, de Seyssel, Gautheron, et al. (2022), we argue that artificial learners have two key advantages over their biological counterparts. First, artificial learners are *tireless*. We can study their responses over thousands or millions of stimuli in ways that would be impossible – and unethical – with infants. This allows us to draw statistically robust conclusions on the artificial learner's language capabilities. Second, artificial learners are *adaptable*. We can tweak their input data and/or learning mechanisms and observe how these modifications impact the learning outcomes, running ablation studies that are not feasible with infants.

Combined with the recent advances in self-supervised algorithms that learn from raw speech, these two advantages open up new modeling opportunities that may inspire development researchers to run new infant studies and advance our understanding of how infants acquire language. One way in which we can use the reverse engineering approach is by holding the learning mechanisms constant and controlling the input to measure its downstream effects. Indeed, the literature on infant language development points to the importance of input quantity (Brookman et al., 2020) and quality (Weisleder & Fernald, 2013) and their potential influence on the later language skills developed by children.

For instance, by controlling the quantity of input, we can simulate data deprivation and/or proliferation experiments. Similarly, we can compare the input afforded to

different infants or groups of infants. Instead of simply describing qualitative and quantitative differences between the children or the groups, a reverse engineering approach involves exposing the exact same artificial learner to long-form recordings of these different groups, and then assess whether significant differences in learning outcomes are observed. Likewise, certain research investigates the language input received by children exposed to multiple languages using long-form recordings (Orena et al., 2020). In addition to examining variations in the learning outcomes following exposure to naturally occurring multilingual audio, we can also simulate specific scenarios more accurately by creating bilingual and/or multilingual corpora. This allows us to precisely control for the distribution of these languages during the exposure phase and to use all languages to check for skills in each language during the evaluation phase.

To make the most out of the reverse engineering approach, we advocate that language learning simulations should closely emulate real-life situations. Besides using ecological data to simulate the learning environment, we recommend the use of machine-adapted psycholinguistic tasks to evaluate the learning outcomes developed by the artificial learner, as argued in Section 2.2.3.

ANNUAL REVIEWS

Annual Review of Linguistics

Reverse Engineering Language Acquisition with Child-Centered Long-Form Recordings

Marvin Lavechin,[1,2,3] Maureen de Seyssel,[1,2,4] Lucas Gautheron,[1] Emmanuel Dupoux,[1,2,3] and Alejandrina Cristia[1]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France; email: marvinlavechin@gmail.com, maureen.deseyssel@gmail.com, lucas.gautheron@gmail.com, emmanuel.dupoux@gmail.com, alejandrina.cristia@ens.fr

[2]Cognitive Machine Learning Team, INRIA, Paris, France

[3]Facebook AI Research, Paris, France

[4]Laboratoire de linguistique formelle, Université de Paris, CNRS, Paris, France

Annu. Rev. Linguist. 2022. 8:389–407

First published as a Review in Advance on November 15, 2021

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-031120-122120

Keywords

long-form recordings, LENA, language acquisition, computational studies, ecological validity, reverse engineering

Abstract

Language use in everyday life can be studied using lightweight, wearable recorders that collect long-form recordings—that is, audio (including speech) over whole days. The hardware and software underlying this technique are increasingly accessible and inexpensive, and these data are revolutionizing the language acquisition field. We first place this technique into the broader context of the current ways of studying both the input being received by children and children's own language production, laying out the main advantages and drawbacks of long-form recordings. We then go on to argue that a unique advantage of long-form recordings is that they can fuel realistic models of early language acquisition that use speech to represent children's input and/or to establish production benchmarks. To enable the field to make the most of this unique empirical and conceptual contribution, we outline what this reverse engineering approach from long-form recordings entails, why it is useful, and how to evaluate success.

1. INTRODUCTION

Recent years have seen the rise of data collection through wearable, lightweight, unobtrusive devices that collect audio for tens of hours at a time, allowing a uniquely naturalistic viewpoint of language use as people go about their everyday activities (Oller et al. 2010, Sun et al. 2020, Wu et al. 2018). Although this technique could be used to investigate language use at any age (see **Figure 1**), it has been extensively used with children; as a result, this body of work has important conceptual, methodological, and ethical contributions that are relevant across fields of linguistics. There exist recent systematic overviews of this prior research (see, e.g., Ganek & Eriks-Brophy 2018). Technical aspects of long-form recordings (including step-by-step how-tos and ethical recommendations) have already been largely covered in prior work (see Casillas & Cristia 2019, Cychosz et al. 2020; see also the **Supplemental Appendix**). Thus, after positioning the methodology in the broader context of language acquisition research methods, we devote most of the article to laying out the promise of the technique in the context of a reverse engineering approach to the study of early language acquisition. Indeed, now that it is possible to capture the full complexity of child language experiences, artificial language learners trained on realistic data can help us build better theories about how humans develop their language perception and production skills.

Supplemental Material >

2. POSITIONING LONG-FORM RECORDINGS IN A BROADER METHODOLOGICAL LANDSCAPE

In typical long-form recordings, infants and young children wear a custom-made piece of clothing with a breast pocket, within which a recording device is inserted. This device typically records over many hours. When the full waking day is represented, we may talk of daylong recordings.

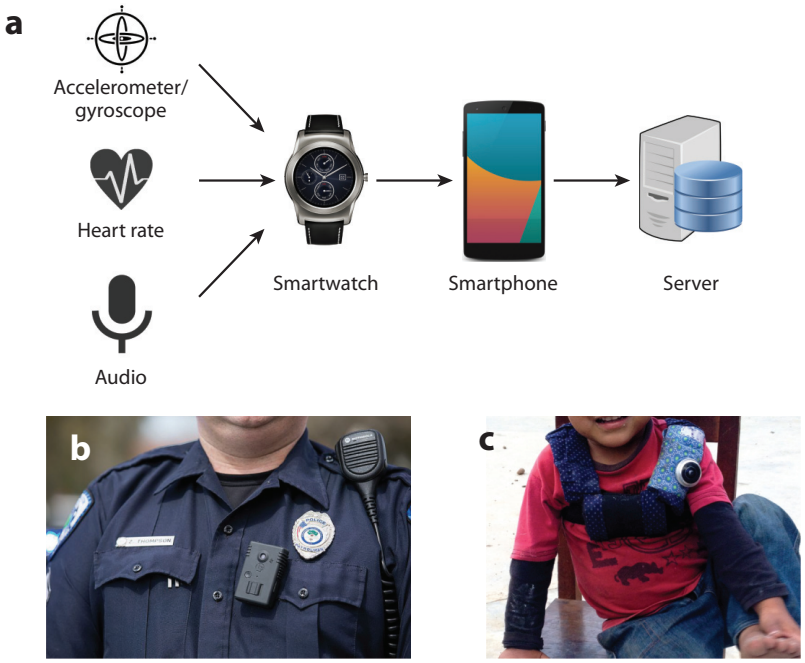

Figure 1

Examples of wearable recorders. (*a*) Smartwatch recording audio, heart rate, and movement. (*b*) Body camera on a South Carolina police officer. (*c*) A small audio recorder and photo camera worn by a Mayan child in Southern Mexico. Panel *a* adapted with permission from Liaqat et al. (2018, figure 1). Photo in panel *b* provided by Ryan Johnson (CC BY-SA 2.0). Photo in panel *c* provided by Marisa Casillas.

Table 1 Main strengths and weaknesses of the four key methods to study language input afforded to children as well as children's language production

Study method	Context sampling	Ecological validity	Reusability	Complexity
Long-form recordings	High	High	High	High
Third-party observations	Medium–high	Medium–high	Low	Low
Parental reporting	Medium	Low	Low	Low
Short audio/video recordings	Variable	Medium	High	Medium

In this review, however, we use the term long-form recordings to highlight the fact that some researchers do not capture the whole waking day but only 8+ hours, whereas others may also capture nighttime.

Language acquisition can be studied by a variety of means, each of which has unique strengths and weaknesses (see **Table 1**). It is most reasonable to reflect on the place the use of long-form recordings can have by comparing it with other methods to study (*a*) children's production and (*b*) the input afforded to children. By and large, productions and input can be studied jointly using one of four methods: (*a*) long-form recordings, (*b*) long observations by third parties (third-party observations, for short), (*c*) parental reporting, and (*d*) shorter audio/video recordings. Next, we define each of these techniques and highlight the relative strengths and weaknesses by comparing the methods according to the following criteria: context sampling (what proportion of the context of the child experiences is sampled), ecological validity (to what extent the acquired data reflect real characteristics of the situations), reusability (how reusable the acquired data are in light of new hypotheses), and complexity (in analyzing the acquired data).

Long-form recordings rely on a sampling of the full range of a child's experiences in one or several days (and sometimes also during nights) and across all the contexts the child may be in (in or out of the house). Among the four methods we discuss here, long-form recordings are therefore closest to third-party observations, such as those that anthropologists employ for time allocation research (Gross 1984) and those that psychologists use for some behavioral observations. For instance, Roopnarine et al. (2005) observed families for 3 h at a time for four separate visits, each time completing a checklist of observed behaviors every 30 s. Long-form recordings and third-party observations have the relative advantage over other techniques that the child's carers do not have to do anything special (not even stay in the same room as the camera). Furthermore, the novelty effect of having an observer should, if anything, decrease over the long observation period, resulting in greater ecological validity for long observations and long-form recordings than the other methods.

Nonetheless, long-form recordings present three advantages compared with third-party observations. First, the observer is less salient, which may result in even lower awareness and fewer perturbations of the natural behavior of participants (high ecological validity). Second, setting aside the potentially longer initial investment to learn the technique and obtain ethical approval, long-form recordings require less effort and time from experts than third-party observations, particularly since the recorder can be mailed. In terms of reusability, recordings, unlike third-party observations, can be consulted, reannotated, and reanalyzed, including to measure behaviors that were not considered before collecting the data. That said, third-party observations have an important advantage over long-form recordings in that the observer has access to multimodal cues and other information, allowing more nuanced interpretations using the full 360° context. Moreover, such third-party observations often rely on standardized checklists, which are then easy to analyze (low complexity), while long-form recordings require the use of manual and/or automatic annotation tools to extract information of interest (high complexity).

Among parental reporting techniques, the method closest to long-form recording is probably the use of smartphone apps or a similar setup to collect reports from caregivers over a long time period. Diaries that ask bilingual parents to report on how frequently they use one or another language, at different moments of the day over a whole week, would fall under this category (e.g., Orena et al. 2020). Although apps are not yet prevalent, they could be useful for sampling behavior at different timescales. For instance, we could ask the caregiver to report who is talking to the child right now (through push notifications that pop up at different points in the day) or to report which words the child says or understands at different child ages (as in Wordful; **http://wordfulapp.com**). Like long-form recordings and third-party observations, such reports could sample from the full range of experiences afforded to children. That said, there could be reporting biases due to relying on caregiver report, lowering both context sampling and ecological validity. As to the former, caregivers may be less able to respond to the app's request about who is talking to the child when they are engaging in hygiene or other hands-on routines. As to the latter, caregivers will be keenly aware of being observed when reporting on their own behaviors and may align their reporting with their beliefs instead of accurately representing what occurs—for instance, a bilingual parent who consistently reports that they use each language 50% of the time. Another limitation of parental reports is that, as with third-party observations, they cannot be revisited to code other behaviors not foreseen in the original design (i.e., low reusability of the method). Nonetheless, parental reports have a key relative advantage over long-form recordings and third-party observations in that caregivers can incorporate their background knowledge about the family and the child in their interpretations. Similarly to third-party observations, data acquired through parental reports are also easier to analyze than long-form recordings (low complexity).

Finally, shorter audio/video recordings are probably the most common way in which psycholinguists have described children's input as well as their production. For example, Bergelson et al. (2019) studied input and production in a longitudinal study on infants aged 6–18 months by setting up a video recorder on a tripod and having each infant wear two head-mounted video recorders. Families were thus recorded at home for 1 h, after which the researchers returned to pick up the equipment. Such shorter audio/video recordings share with long-form recordings the relative advantage that they can be revisited as new hypotheses arise (high reusability). One relative disadvantage of short observations over all other methods discussed so far is that investigators must choose whether they want to keep activity heterogeneity low, by asking all families to record during a specific activity (e.g., mealtime, play, hygiene), or whether they want to represent the full range of experiences, in which case they still need to choose how to sample from each (e.g., whether to record the same number and length of each, or whether to sample them at the frequency at which they occur in a natural day). Such short observations also probably result in increased consciousness of being observed, which potentially affects participants' behavior and lowers ecological validity. This also leads to a key advantage of short recordings over all the other methods, which is that the investigator can purposefully target an activity or setting that is most relevant to their purposes—for instance, by providing a set of toys that leads to an increased use of a relatively rare structure (e.g., eliciting defining and nondefining relative clauses with a purpose-made book or deck of cards). Although shorter audio/video recordings are less complex to analyze than long-form recordings, working on audio and/or videos still requires the use of automatic or manual annotation.

Importantly, no method is perfect, but the various available methods are to a certain extent mutually compatible, such that researchers can try to design data collection using one or more in a complementary scheme. For example, Bergelson et al. (2019) collected one full day's audio recording in addition to the hourlong video recording, on separate days, and found some diverging results (notably a larger quantity of speech in the hourlong videos than in the long-form audios)

as well as considerable convergence (e.g., in terms of who spoke to the child). Additional mixed methods can be devised to serve the researcher's goals: An investigator could ask families to record over two full days and could send in an observer for part of one of those days, or they could ask the families to play an elicitation game and write down the time and day they did so. The investigator could then extract the sections of the recording that contained these dual methods and analyze them further, in order, for instance, to establish the extent to which behaviors are affected by the presence of a third-party observer in the first example, or simply to transcribe and study the speech occurring during the elicitation game in the second example.

Without denying the complementarity of the four methods, we believe that there are three ways in which long-form recordings are unparalleled. First, long-form recordings are a promising technique to collect naturalistic big data in various populations. Nielsen et al. (2017) documented that developmental journals are heavily skewed toward publishing articles with data from WEIRD (Western, educated, industrialized, rich, and democratic) populations. This sampling bias can lead behavioral scientists to wrongly identify culturally specific findings as universal traits. Although researchers using systematic behavioral observations and short recordings have certainly made an attempt to broaden the languages and cultures represented in the literature, both of these methods require so much investment and expertise that, in reality, it is mostly outsiders who document language acquisition in such settings. Therefore, despite our best intentions, we may misrepresent the language and culture, and furthermore, our research questions and output may not have the optimal impact they can have on the population from which the participants are drawn. Admittedly, most current adopters of long-form recordings are from WEIRD societies (Cychosz & Cristia 2022), but we hope that this method will be increasingly used by diverse researchers, including members of underrepresented and underserved linguistic and cultural communities, so that the mainstream literature can better represent their viewpoints and interests, and so that these populations stand a higher chance of benefiting from the research.

Second, long-form recordings may be ideally suited to address current needs for replicable and reproducible research. To begin with, reproducibility is heightened by the use of audio and video archiving and sharing repositories such as HomeBank (VanDam et al. 2016) and Databrary (Simon et al. 2015), in the wake of the CHILDES tradition (MacWhinney 2000). What is more, by capturing a maximally unbiased sample of the child's language experience while also ensuring maximal ecological validity, the use of long-form recordings should, overall, increase the probability of conceptual replications. Additionally, many of the analyses rely on automated methods that are shared across many laboratories. As a result, it becomes easier to quantify (and possibly fix) biases that may be present in measures extracted by these automatic tools than when relying on human annotation.

Third, such recordings may be ideally suited to fostering a new direction of research within the broader field of modeling early language acquisition—namely, a reverse engineering approach to the study of infant language development. We dedicate the rest of this article to laying out this new research direction.

3. REVERSE ENGINEERING LANGUAGE ACQUISITION

There is a long tradition of modeling in the context of language acquisition (e.g., MacWhinney 2005). A complete review would be beyond the scope of this article, but to illustrate the field of possibilities we can cite Anderson (1975) for an example of a syntax-learning model or Brent (1996) for a word-discovery model. While computational models of language acquisition traditionally assumed that speech was represented as an error-free string of adult-like phonemes (which is unlikely the case for infants), more recent studies address the problem of language learning from

raw speech. This line of research can be illustrated with a study by Nguyen et al. (2020), the last iteration of a challenge[1] organized in the speech processing community that revolves around spoken language modeling without annotation or text.

In view of the substantial progress of these past years in algorithms that can learn from raw speech, there is a clear interest in a greater integration of artificial intelligence (AI) and language development studies. Dupoux (2018) set down recommendations for such an enterprise, which we will not repeat here. Like Dupoux, we assume that unsupervised language learning models should be exposed to realistic data and that they should be evaluated on psycholinguistic benchmarks to compare humans' and machines' language capabilities at various linguistic levels. In this new research direction, child-centered long-form recordings play a crucial role by providing artificial language learners with ecologically valid input data in the form of the speech by adults and other children present in the audio recordings, which is then directly comparable to the input afforded to children. Our proposal builds on and extends the work of Dupoux (2018) by spelling out how a research program centered on long-form recordings could proceed, considering such recordings as an information source on not only children's input but also children's production.

In Section 3.1, we describe the reverse engineering approach to the study of infant language acquisition. In Section 3.2, we explain why language acquisition researchers should consider using machines as a proxy to study infants. We then lay out how to study input effects in Section 3.3. Section 3.4 presents our psycholinguistic-driven framework to measuring language skills of artificial language learners. This benchmarking framework is based on behavioral correlates of language learning observed in humans and allows us to compare the language skills of the artificial language learner with those of the human.

3.1. What Does the Reverse Engineering Approach to the Study of First Language Acquisition Include?

Since we are discussing reverse engineering in the context of long-form recordings, by and large the experiences under discussion are unimodal, based only on speech, as the vast majority of long-form recordings being gathered are audio only. That said, recent AI work begins to address the problem of learning language from audiovisual exposure (Alishahi et al. 2021, Chrupała et al. 2017, Harwath et al. 2020), although admittedly these studies do not use realistic data. Using long-form data to train these sighted artificial language learners would require using devices that capture both what children hear and what they see (such a setup has been used in, e.g., Casillas et al. 2020, 2021). Capturing child language experiences across multiple modalities would offer us opportunities to compare audio-only models with audiovisual models and could help us better understand the role of visual experiences during the language acquisition process. Importantly, while audiovisual long-form recordings could be used, touch and smell cannot (yet?) be digitized, particularly at the long-form scale, yet these senses can help in the language acquisition process (Abu-Zhaya et al. 2017). This is a current limitation of the technique, and thus we discuss only audio-based and video-based models in this review.

By and large, the models we discuss in this review are passive learners, in the sense that they cannot affect the input data they receive. This aspect of the models makes them somewhat different from human children, who are able to explore and interact with their environment. For example,

[1] Challenges in the machine learning community are events during which participants (usually researchers, including students) work on improving the performance of a baseline model on one or multiple tasks—from audio classification (e.g., Schuller et al. 2019) to unsupervised language modeling (e.g., Nguyen et al. 2020).

if a learning human child formulates the hypothesis "Cats are those little hairy animals" and wants to check whether this hypothesis is true, the child could interact with their environment to prove or disprove the hypothesis, such as by pointing at the cat while waiting for a caregiver's reaction. This connects with the importance of embodiment in first language acquisition (Yu 2014) and of the child's role in shaping their environmental input (Tamis-LeMonda et al. 2018). Although the artificial learner may benefit from the child's interactions with their environment, if the two types of learner are not at the same learning stage and/or have not formulated the same hypothesis, ultimately the child will profit more from their own experiences than will an artificial learner who is simply reexperiencing the human child's experiences.

With those considerations in mind, the reverse engineering approach to the study of first language acquisition via long-form recordings can be summarized as follows:

1. We design a computer program to have some learning mechanism(s) we believe are useful to learn language (i.e., we control the mechanisms). Those are discussed in the current section.
2. We provide this program with a controlled and realistic language experience (i.e., we control the input; see Section 3.3).
3. We observe what the system learns and how it develops (i.e., we observe the learning outcomes;[2] see Section 3.4).

Let us take the example in which our goal is to understand perceptual development. In such a case, there is often a learning phase and an evaluation phase. The learning phase includes exposing the artificial language learner to a naturalistic and controlled language experience (typically represented by adult speech extracted from child-centered long-form recordings). Then comes the evaluation phase, in which the attuned learner (i.e., the model postexposure) undergoes a battery of psycholinguistic tests. These psycholinguistic tests are conceptually related to experimental protocols used in child studies to assess the language capabilities of infants (e.g., looking-while-listening procedure, conditioned head turning). Behavioral patterns extracted from those tests can then be compared with those observed in humans undergoing the experimental version of the psycholinguistic tests. **Figure 2** provides a diagram illustrating this version of the reverse engineering approach.

Having described the approach, we can now discuss the benefits of using in silico modeling to study infant language acquisition.

3.2. Insights that Artificial Language Learners Can Provide Us

Despite thousands of laboratory experiments that isolate learning mechanisms in babies, thousands of hours of observations, and probably millions of hours of study of both bodies of work, still little is known about the inner mechanisms that underlie human language development in the wild, the causal role of the input afforded to the child, and the best explanation for the perception and production outcomes exhibited over the course of acquisition. This is not truly our research community's fault. Language acquisition, like most cognitive processes, is essentially a black box and can be studied only as such by language development researchers.

It is important to note that the term black box is also used in the AI community and can be defined as the inability to predict an AI model's decisions. Even though the state of the model is fully known at any given time, this state is so complex that its inner processes and workings cannot be fully understood.

[2]Note that we distinguish output (i.e., production) from outcome (i.e., consequences that could be visible in production or perception).

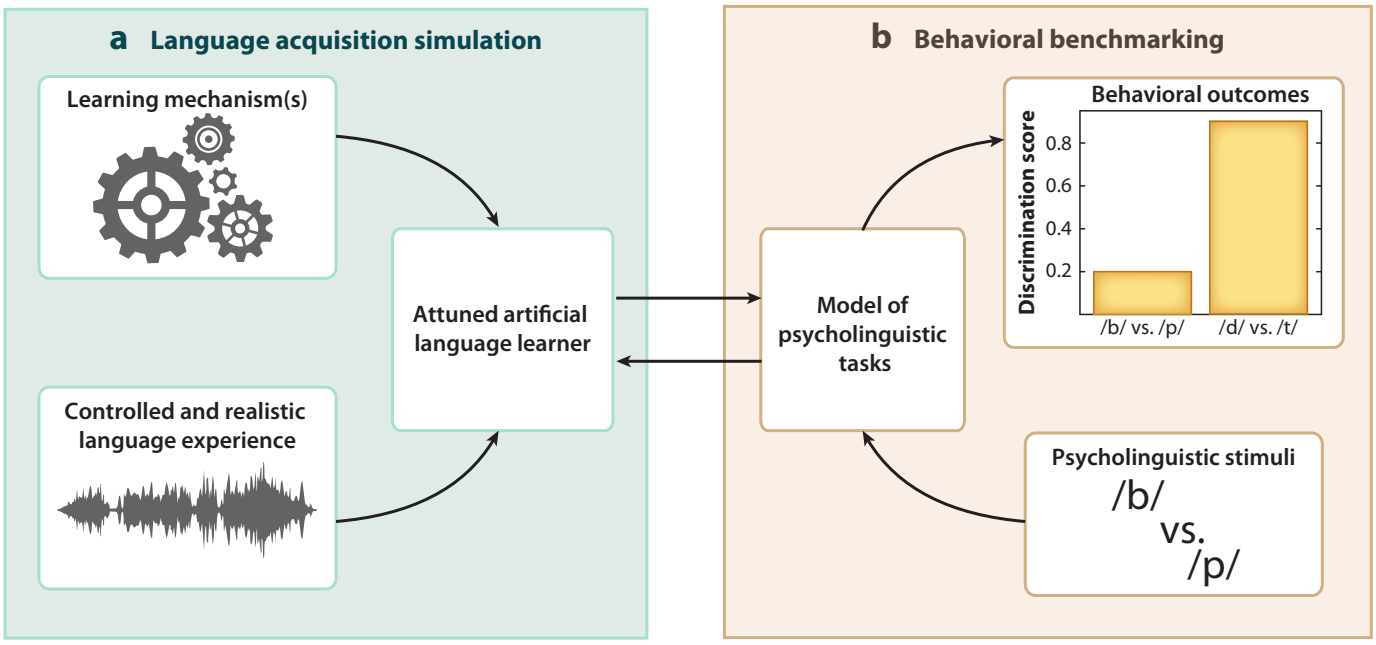

Figure 2

Outline of the reverse engineering approach for language acquisition. (*a*) First, we perform a language acquisition simulation by applying learning mechanisms on a controlled and realistic language experience to derive an attuned artificial language learner. These simulations can be used to check for the effect of variation in those language experiences (e.g., multilingual input, variation in household size; see Section 3.3). (*b*) Second, we use behavioral benchmarking to evaluate learning: The attuned artificial language learner undergoes a battery of psycholinguistic tasks. These tasks are conceptually related to the tasks that are used to study humans, but they are adapted for the machine (see Section 3.4). Note that the psycholinguistic task can also contain a habituation phase that will further modify the state of the attuned artificial language learner (hence the two-sided interaction between the two).

It therefore appears that language acquisition and AI researchers face a common challenge: studying a black box. So one may wonder, Why would one use a black box as a proxy to study another black box?

First, artificial language learners are tireless. We can study their responses over a large quantity of stimuli in ways that would be impossible (or unethical) with infants. Observing behavioral patterns in machines may inspire psycholinguists to run new human studies. Indeed, if a behavior of interest is observed in machines but has never been documented in infants, then it may be worth checking whether infants also exhibit this given behavior. Using thousands or millions of stimuli allows us to generate robust predictions.

The second advantage of the AI black box over the infant's is the adaptability of the former: We can easily tweak an AI model's input data as well as parameters of the learning mechanism(s). Thus, we can observe consequences of the simulated language experience on the language skills of the machine while keeping control over the learning mechanisms, and vice versa.

If there is a mismatch between the learning outcomes of an artificial language learner and the human learner, then we can think of ways to make the machine more human-like. In cases in which we are using input that faithfully represents the input that children are exposed to, we can rule out an issue with the input and thus infer that the mismatch might be due to other aspects of our experiment: namely, (*a*) the learning mechanisms in the model, (*b*) details of the experience presentation (e.g., the quantity of data), or (*c*) the computational implementation of the psycholinguistic tests used to measure learning outcomes.

In the opposite case, if a computer program yields outcomes similar to those of the infants when provided with the same experience infants had, then we are faced with the tantalizing conclusion

that we have found one viable model (probably among many others) of human behavior. Then, based on the results of that viable model, we can suggest new testable predictions regarding how infants' language develops.

3.3. Controlling the Input to Measure Its Downstream Effects

In this section, we assume that we hold mechanisms constant, and we illustrate how the reverse engineering approach can shed light on the causal role of the input, with the goal of measuring learning outcomes through procedures described in Section 3.4.

One way in which we can use this approach is by comparing the input of different infants or groups of infants. Some discussions of, for instance, group variation are based on the idea that some children may be afforded qualitatively (Brookman et al. 2020) or quantitatively (Weisleder & Fernald 2013) different input. Instead of simply describing qualitative and quantitative differences between the children or the groups, a reverse engineering approach would expose the exact same artificial learner to long-form recordings of these different groups and then assess whether there are substantial differences in learning outcomes.

In fact, one advantage of a reverse engineering approach is that we can go beyond attested environments. Since we can control the quantity of data the artificial language learner is exposed to, we can simulate data deprivation and/or proliferation experiments. We can also plot developmental curves that show the evolution of the model's language skills as a function of the quantity of data it has had access to.

Similarly, some research studies the language input of children exposed to multiple languages using long-form recordings (Orena et al. 2020). Along with checking for outcome differences after exposure to naturally occurring multilingual audio, we can more precisely simulate additional cases by creating bilingual and/or multilingual corpora. This allows us to precisely control for the distribution of these languages during the exposure phase and to use all languages to check for skills in each language during the evaluation phase.

The approach can be generalized to many other aspects of language exposure that are currently hard to control and tease apart, although for some it may be harder to do so well. For instance, we think it is technically possible to vary speaker distribution in terms of household size, number of siblings, and gender distribution because we can start with recordings done in households with only one adult caregiver and only one child and then mix them together to create families with one to ten caregivers and one to ten "siblings," since in the original setting we can trust that the adult and child sections will have been correctly attributed to one adult and one other child, respectively. Even so, the siblings in these simulated large families will be easier to tell apart than siblings found in natural recordings of large families because they will not necessarily sound similar to each other (since they are originally drawn from different families). Other dimensions may be hard to create because the tools they rely on are not necessarily mature yet. For instance, varying vocabulary size would require an accurate automatic speech-to-text model, which is beyond the state of the art at present; varying the frequency of book readings would require automatically extracting book-reading interactions, but there is no automatic tool to detect such interactions to date. Nonetheless, and given the fast pace of AI research, it is useful to think about this feature of in silico modeling in general terms.

3.4. Evaluating Language Skills of the Artificial Language Learner

Let us start by assuming that we want to check what was learned. In this case, one may be tempted to ask, "Do models learn phonemes, nouns, and so forth?" This is akin to attempting to establish

whether the model has a given linguistic representation. We see two problems with this option. First, it is devilishly difficult to demonstrate that a representation is in place; doing so involves, for instance, additional levels of agreement in terms of how to prove its presence (e.g., for phonemes, see Jaeger 1980; for a general criticism on inferring mental entities, see Twaddell 1935). Second, representations are precisely the area in which theories of language (acquisition) diverge in the fiercest manner (e.g., Ambridge & Lieven 2015). Although some may disagree with us, we would like to posit that psycholinguistic benchmarking tasks should be theory-agnostic. By not forcing extra assumptions of representations on either the human or the machine learner, the findings have a higher chance of being relevant to a broad range of theories.

A second definition of learning may be to ask about representations in terms of their neural implementation. In the avenue of human–machine comparison, an approach that has interested the AI and neuroscience communities involves comparing activation patterns of neural networks with those of the human brain while being asked to perform a similar task (e.g., Yamins et al. 2014). While this approach promises interesting scientific insights about human brain information processing, it also shows some limitations in the context of early language acquisition. First, data acquisition devices that locate activity precisely in both time and space are very seldom used with infants (but see, e.g., Bosseler et al. 2021). Second, these techniques require large sample sizes for reproducible results even among adults (Turner et al. 2018), and given that measurements in infancy are typically even noisier than those gathered in adulthood, we can infer that it is possible that the body of literature on infant neuroimaging will require substantial accumulation of results before we can employ it for our benchmarks.

Therefore, given the current theoretical and methodological landscape, we propose a behavior-oriented approach in which we study the behavior of infants in parallel to the behavior of machines.[3]

To extract behavioral patterns for the machine, we additionally need to reflect on how numbers returned by the machine can be related to the kinds of behaviors elicited and/or observed among human learners. By numbers, we mean either the output representations[4] of the input stimuli for perception models or the vocalizations produced by production models.

With that in mind, we turn to the following question: What behaviors do infants and adults exhibit through the course of language learning, and which could serve as benchmarks for artificial learners? We discuss these issues in two subsections, targeting perception (Section 3.4.1) and production (Section 3.4.2).

3.4.1. Measuring perception. Much of the language acquisition literature has attempted to look at perception, mainly through infants' reactions to specific stimuli in clearly defined laboratory tasks. This work suggests that much goes on in the child's mind even before there are obvious changes in production. In this section, we reflect on how long-form recordings paired with computational modeling can help us understand the development of these markers.

---

[3] In this review, we are simplifying matters by not going into detail about the fact that, in reality, a perfectly comparable artificial learner would develop not only language but also all of its other cognitive systems, which would allow the influence of attention, memory, and other systems orthogonal to language to be accurately represented in the model. In other words, this would entail not only modeling language development but also modeling how the child approaches the psycholinguistic task. For relevant discussion, we refer readers to Robinaugh et al. (2021).

[4] Representations learned and returned by the model should not be confused with linguistic representations. Output representations returned by the machine are numbers describing and/or organizing the input that was given to the model. By linguistic representations, we mean hypothesized mental units that represent elements of language (e.g., phonemes, morphemes).

As described in Section 3.1, there are two phases in the reverse engineering approach. The first one involves exposing the artificial language learner to a controlled and realistic language experience extracted from long-form recordings, which can be selected to vary certain dimensions parametrically, as explained in Section 3.3. As described in this section on perception, the second phase is one of evaluation of perception skills. Since perception cannot be directly investigated in either human or artificial agents, in this section we rely on behavior that is elicited in controlled conditions. Thus, the AI learner is presented with stimuli like the ones submitted to children in the laboratory, to elicit numbers that can be interpreted as behavioral perceptual patterns in the artificial learner. In other words, we create a computational implementation of infant perceptual benchmarks, to which the artificial language learner is submitted. In **Table 2**, we draw examples mainly from studies on infants aged 0–12 months, but we trust that our reasoning can be generalized to perception tasks beyond 1 year of age. We want to highlight that this is not an exhaustive

Table 2 A sample of human behavioral correlates of language skills that have been reported in the literature along with their computational implementation

	Age (months)	Task	Data set	Literature
Sound-only behaviors				
Discriminate across rhythmically distinct languages	0	Distance-based	Bilingual set of stimuli	Gasparini et al. 2021
Discriminate native and nonnative consonants	6–8	Distance-based	Phonetically aligned clean speech	Werker & Tees 1984
Accept novel content words more easily than novel function words	6	Few-shot learning + probability-based	Jabberwocky sentences	Shi et al. 2006
Prefer high over low phonotactics	9	Probability-based	Made-up words varying in phonotactics	Jusczyk et al. 1994
Prefer high- over low-frequency content words	11	Probability-based	Real words varying in frequency	Jusczyk et al. 1994
Do not discriminate nonnative consonants	12	Distance-based	Phonetically aligned clean speech	Jusczyk et al. 1994
Cross-modal behaviors				
Treat words and monkey vocalizations, but not beeps or coughs, as possible labels	3	Few-shot learning + distance-based	Images paired with words, monkey vocalizations, beeps, or coughs	Ferry et al. 2010
Treat words but not monkey vocalizations as possible labels	6	Few-shot learning + distance-based	Images paired with words or monkey vocalizations	Ferry et al. 2010
Treat content but not function words as possible labels	6	Few-shot learning + distance-based	Images paired with function words or content words	Hochmann et al. 2010
Know the meanings of many common nouns	6–9	Distance-based	Images paired with common nouns	Bergelson & Swingley 2012
Few-shot learning of new word–object pairings	9	Few-shot learning + distance-based	Images paired with words	Yeung & Werker 2009
Treat words with native but not nonnative sounds as possible labels	10	Few-shot learning + distance-based	Images paired with first- and second-language words	May & Werker 2014

The Task column describes the task that is meant to be submitted to the artificial language learner. Distance-based tasks consist of computing the distance between the output representations of the input stimuli. Probability-based tasks consist of computing the probability of the output representations. Few-shot learning tasks involve a learning phase during which the model is given some examples. The Data Set column describes the test stimuli that need to be gathered to submit the task of interest. Labels are audio stimuli consistently paired with a visual stimulus—for instance, a monkey vocalization systematically followed by a fish picture. The Literature column suggests entry points in the psycholinguistic literature.

list, and we would be delighted if other researchers created computational implementations of other infant perceptual benchmarks.

The top section of **Table 2**, dedicated to sound-only behaviors, represents experiments where the audio is key for eliciting infants' responses, whereas the second half of the table shows cross-modal behaviors where both audio and visual stimuli are used.

Test tasks may be of two types: distance-based and probability-based. Distance-based benchmarks rely on computing the distance between different stimuli, whereas probability-based benchmarks require the machine to compute the probability of each stimulus. We illustrate these concepts with examples below.

Let us start with a distance-based example. Newborn humans can discriminate across rhythmically distinct languages (Nazzi et al. 1998). One can measure a discriminability score in the machine by submitting to it three audio stimuli, A, B, and X, such that A comes from a first language $\mathcal{L}_1$, B comes from a second language $\mathcal{L}_2$, and X comes from $\mathcal{L}_1$ but is different from A. Under a distance function d, one may expect that $d(A, X) < d(B, X)$ as A and X have been drawn from the same language. While this might not be true for a given stimulus, repeating the procedure across thousands of stimuli allows us to extract robust patterns in the artificial language learner. In other words, we use here the computational implementation of the *ABX* discrimination task used in psychology. An example of a computational study using this task can be found in the work of de Seyssel & Dupoux (2020), who tested the language discrimination capabilities of an i-vector model to assess the role of monolingual versus bilingual exposure, akin to studies on monolingual and bilingual human infants. The same distance-based method can be used to benchmark the machine's phoneme discrimination capabilities. In this setup, A, B, and X are triphones with A and B differing only in their center phone (/bet/ versus /bat/) and X being the same triphone as A (but another occurrence). In the same way, one may expect that $d(A, X) < d(B, X)$ as A and X represent the same triphone. Phoneme discrimination capabilities have been evaluated on a Gaussian mixture model by Schatz et al. (2021), who notably showed that a model exposed to Japanese exhibits a lower discrimination score on the [ɹ] versus [l] contrast than does a model exposed to American English. Incidentally, note that Schatz et al. (2021) concluded from their results that the AI learner solves this task without phonemic representations per se, exemplifying one way in which not making assumptions about representations may facilitate cross-pollination of findings between developmental science and computational research.

Next, we turn to a probability-based example. At the age of 11 months, infants have been shown to prefer high- over low-frequency content words (Carbajal et al. 2021). Checking that this behavior is present in the machine would consist of submitting to it two stimuli, A and B, with A drawn from high-frequency content words and B drawn from low-frequency content words. One should observe that the probability the model returns for A is higher than the probability returned for B, as the first one is supposed to be more frequent than the second one in the training set.

The same two benchmarking approaches (probability- and distance-based) can be adapted to a cross-modal setting in which decisions need to be taken by integrating information from the auditory and another (typically visual) modality. As for the visual modality, a task used to test infants' comprehension of words and sentences that seems particularly easy to adapt for the machine is the looking-while-listening task (Fernald et al. 2008). In this task, the child sits in front of a screen that shows two images, only one of which corresponds to the audio the child is concomitantly presented with. During the computational implementation of this test, the machine would receive an audio stimulus A_1 (e.g., "nose") and would be presented with two images, one of them representing the audio stimulus I_1 (i.e., a picture of a nose) and the other one, I_2, representing something else (e.g., a picture of a mouth). The machine would then be asked to output the representations of all three stimuli. To check if the machine was able to map the audio stimuli to the right image,

we would consider a distance function d and check that $d(A_1, I_1) < d(A_1, I_2)$ as A_1 and I_1 share the same semantic content. Note that, alternatively, the joint probability distribution between a word and an image might be used instead.

In **Table 2**, we highlight the fact that such a test phase can be preceded by a learning phase during which the machine is presented with some examples. This phase is called few-shot learning; few-shot means that only a small number of exposure instances are used before the evaluation. For instance, these exposures can take the form of nonsense words in the audio-only setting or image–word pairs in the cross-modal setting.

Finally, let us note that the audio stimulus does not have to be a real word. For instance, a fish picture can be paired with a monkey vocalization, in which case we would evaluate the ability of the machine to learn this new word–object pairing. Indeed, a sizable literature in developmental research investigates the (presumably innate) biases infants bring to word-learning tasks, and it has been found that young infants exposed to words or monkey vocalizations systematically paired with a visual category (e.g., dinosaurs) will generalize the "label" to a new exemplar of the same visual category, whereas the same behavior is not observed when dinosaur pictures are systematically paired with beeps or coughs (for an overview of this line of research, see Vouloumanos & Waxman 2014). To our knowledge, similar biases have not been investigated in artificial agents.

In sum, the probability- and distance-based evaluation paradigms are extremely powerful. They have already been used in the ZeroSpeech 2021 challenge for computational implementations of human psycholinguistic benchmarks across multiple linguistic levels, including phonetics, lexicon, semantics, and syntax (Nguyen et al. 2020). Nonetheless, it is important to bear in mind that most of the previous computational studies used relatively manicured recordings (such as audiobooks), and to our knowledge, none of them has tried to tackle language learning after exposure to audio from long-form recordings.

3.4.2. Measuring production. In this section, we turn to production, although we start by admitting that this line of research is a great deal less developed than the perception one, and it may take considerable time to make progress in this area. As with perception, the development of production involves learning mechanisms that are still the subject of active debate (Long et al. 2020).

There are two key differences regarding how perception and production reverse engineering approaches can work. First, models of perception development require the extraction from long-form recordings of only the speech that represents children's input, whereas to model production development it is worthwhile to extract both the input and the child's output. In fact, when considering production development, there is a strong case to be made about children's production being shaped by their own output—and thus their output may, under some accounts, be considered as input too. Second, while perception models are only required to return representations of their input, production models need to integrate those evolving representations of the input, with (*a*) (potentially changing) biophysical constraints on production (due to the fact that the child's body, including their tongue, is changing with age), (*b*) mechanisms for learning-related changes in production, and (*c*) some system for actually generating vocalizations. That said, not all extant models of children's language production consider all of these aspects, and others actually consider additional constraints, such as social constraints (Pagliarini et al. 2021). To take a specific example, Warlaumont et al.'s (2011) model has an articulatory component (which generates the child's output in terms of gestures) as well as a perceptual auditory component (which captures patterns in the input as well as the auditory consequences of the child's production)—but note that the articulatory component does not include biophysical constraints per se. Approaches with realistic models of the developing vocal tract are rare (but see Philippsen 2021).

At least in principle, then, a reverse engineering approach to production development that builds on long-form recordings would proceed as follows: If one believes that input speech can affect production, then the model should use the input to hone perceptual representations (i.e., Section 3.4.1); in all cases, one can use children's production as a benchmark against which to compare the model learner's production. As discussed in Section 3.1, one can control the learning mechanisms and the input in order to measure outcomes.

In reality, however, we see a considerable gap between how production is typically modeled and long-form recording data, for both the perception and the production aspects of production development. In current work aiming to model production development, input is most typically represented in a simplified manner (e.g., with first and second formant values representing vowels), and output is similarly reduced to such summary representations (although exceptions exist; see, e.g., Rasilo & Räsänen 2017). What is more, it is not uncommon to see evaluations akin to an elicited imitation task, where the system is provided with an adult vocalization as input and is asked to imitate them (i.e., produce the articulatory activations corresponding to this auditory input). Such a benchmark does not seem realistic for children under 2 years of age given that eliciting repetition is methodologically challenging even at around 20 months (albeit possible; see Hoff et al. 2008). Moreover, it is unclear that studies are actually referring to what human children's performance is in actual imitation tasks. Instead, much of this work appears to operate under the assumption that imitation is prevalent in real-life interaction. However, laboratory observations suggest that imitation is vanishingly rare (Athari et al. 2021), even in a setting where parents may be driven to increase their interactions with their child as a consequence of being observed. Convergence, which is a form of imitation, was not found to be systematic in the analysis of long-form recordings (Seidl et al. 2018).

A second issue standing in the way of relating long-form recordings and production models concerns the fact that researchers of production development typically specialize in certain development phases and phenomena. For instance, some study the emergence of syllable structure (Warlaumont & Finnegan 2016), and others study vowel targets (Rasilo & Räsänen 2017). This specialization entails that models of production developed at present do not generate the whole range of vocalizations observed in long-form recordings, but are instead dedicated to features of vocalizations.

When we look at production development more broadly, we see many aspects that should be accounted for, including the following:

- The presence of both speech-like and non-speech-like vocalizations (Long et al. 2020)
- The increase in canonical vocalizations (having at least one adult-like consonant–vowel or vowel–consonant transition) with age (Cychosz et al. 2021)
- The appearance of meaningful utterances, starting with single words (de Boysson-Bardies & Vihman 1991)
- The appearance and increased prevalence of word combinations (Braine & Bowerman 1976)

Not only are these aspects out of reach for any extant model learner, but also the basic description of these phenomena in long-form recordings is rare. In fact, it was only recently found that speech-like vocalizations are prevalent in long-form recordings even among newborns (Long et al. 2020) and that the proportion of vocalizations containing canonical transitions appears to continually increase well beyond the first year of age, according to long-form data (Cychosz et al. 2021). Information from long-form recordings on the other phases—namely, the appearance of meaningful speech and of word combinations—has been documented in only two studies (Casillas et al. 2021, 2020), both of which employed human annotation. It would be ideal to develop automated techniques so that they can be applied at scale in multiple languages. Moreover, further

development is needed to create computational implementations of these human psycholinguistic benchmarks if we are to succeed in our goal of comparing humans and systems.

In sum, reverse engineering the development of production is farther away on the horizon than the study of perception, awaiting conceptual advances in modeling approaches, which may also necessitate important changes in the way we do descriptive analyses of production data gleaned from long-form recordings. As briefly noted, some work in production also takes a social stance, incorporating caregivers in the developmental loop. Progress in such conceptual settings will require even more work, as they necessitate reverse engineering the caregiver as well.

3.4.3. Limitations of the human–machine behavioral comparison. Before concluding the review, we want to highlight some limitations of our benchmarking approach. All of it relies on the psycholinguistic human data being empirically solid and unbiased—and we believe progress on both of these fronts is necessary to support the backward and forward loop between humans and machines.

To begin with, **Table 2** may continue to perpetuate the illusion that infants' skills can be described with simple statements. In reality, conclusions drawn from child studies are rarely as clear as "children do *X*" or "they don't do *Y*." Results may vary depending on the sample size, the methodology used, and the age of the participants, as reported, for instance, in a study of 12 meta-analyses by Bergmann et al. (2018). Ultimately, we should probably instead look at the distribution of effect sizes emerging from meta- or mega-analyses rather than an arbitrary yes/no binomial assessment. This point applies both to perception and to production benchmarks.

Another issue with evaluating AI learners against extant human infant literature comes from the fact that this literature is biased toward specific populations and languages. A study of three leading developmental journals (where perception experiments are often published) showed that over three-quarters of their papers bore on North American and/or European infants (Nielsen et al. 2017), and a study in the *Journal of Child Language* (which often publishes articles on children's naturalistic production) showed that a shocking 69% of the papers bore on English learners, with a mere 15% bearing on non-Indo-European languages (Slobin 2014). And although there is less evidence about this, it is likely that the samples from, for instance, North American infants are not representative of the greater populations. Thus, a characteristic observed in an American infant growing up in a high-socioeconomic-status setting may or may not be observed in infants growing up in other communities. In the absence of systematic observations across cultures, the AI learner seems doomed to reproduce bias found in the language acquisition literature. Several researchers are working hard to collect long-form recordings from more diverse populations (see, e.g., Cychosz et al. 2021), and we hold out hope that, at least in terms of long-form audio, the bias may be weakened in years to come. However, for our perceptual benchmarks we require something like laboratory experiments, and there are currently very few researchers collecting perception data from more diverse communities [but look out for Marisa Casillas's output in coming years (Casillas et al. 2020) and the ManyBabies efforts].

Setting aside these two issues of robustness and bias in the data, we foresee that further work is needed to reflect on which tasks we want to incorporate in our benchmark. Being able to solve a given task does not tell us whether solving this task is required for language learning. For instance, divergent discrimination responses to rhythmically similar versus different languages may be neither necessary nor sufficient for language learning. Alternatives include that this difference in behavior is an acquired response bias or a side effect of auditory development as affected by ambient sounds (similar to how infants prefer their mother's voice but not their father's voice at birth; Lee & Kisilevsky 2014)—in other words, behaviors that emerge but that are neither necessary nor

sufficient for language learning. If such a task is used to evaluate the AI learner, observations in humans and machines may be divergent for uninteresting reasons.

That said, we think that this problem is less worrisome than the other two. In fact, deciding which behaviors necessarily occur as a function of language development could be a problem with which computational models can help. The intuition is that if a behavior is necessary and sufficient for language development, then it should be systematically observed for any and all artificial agents that do acquire language. Large-scale cross-linguistic studies assessing language skills of AI learners across different learning mechanisms and language experiences may indeed help us gain insight into which behaviors are merely side effects and which are necessary for language learning. In the opposite case, a single computational model solving task *T* without exhibiting behavior *B* would be enough to conclude that *B* is not necessary for *T*—as in the case of Schatz et al.'s (2021) model, which shows perceptual attunement to the exposure language without phonemic representations. To take another example not yet attested, we can imagine a model that could learn some level of semantics while being unable to detect word boundaries. This would constitute a proof of principle of the computational tractability of semantic learning without word boundaries.

4. CONCLUSION

Long-form recordings offer an ecological view of language use in everyday life. Aside from capturing child language experiences in an ecologically valid way, they offer new and exciting research opportunities in reverse engineering infant language development. Building upon the work of Dupoux (2018), we have defined two key aspects of the reverse engineering approach: (*a*) the language acquisition simulation, or how to use controlled and realistic data to create simulated language experiences; and (*b*) the behavioral benchmarking, or how to assess language skills of the artificial language learner with psycholinguistic tests. This two-sided approach has the potential to increase our understanding of how language is acquired and how it develops through exposure, both in humans and in machines. The more we understand language acquisition in humans, the more human-like artificial language learners we can create. Similarly, the closer artificial language learners are to humans, the more we understand how language outcomes are shaped by exposure.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful to DARCLE (Daylong Audio Recordings of Children's Linguistic Environments), LAAC (Language Acquisition Across Cultures), and CoML (Cognitive Machine Learning) members for helpful discussion. Any errors remain our own. A.C. gratefully acknowledges financial and institutional support from the Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017) and the J.S. McDonnell Foundation (Understanding Human Cognition Scholar Award). This work was also partly funded by l'Agence de l'Innovation de Défense. Early study of unsupervised phonetic learning from child-centered long-form recordings was performed using high-performance computing resources from GENCI-IDRIS (Grant 2020-AD011011829). Their computational support largely contributed in instantiating our methodological approach.

LITERATURE CITED

Abu-Zhaya R, Seidl A, Tincoff R, Cristia A. 2017. Building a multimodal lexicon: lessons from infants' learning of body part words. In *Proceedings of the GLU 2017 International Workshop on Grounding Language Understanding*, pp. 18–21. Grenoble, Fr.: Int. Speech Commun. Assoc.

Alishahi A, Chrupała G, Cristia A, Dupoux E, Higy B, et al. 2021. *ZR-2021VG: Zero-Resource Speech Challenge, Visually-Grounded Language Modelling track, 2021 edition*. arXiv:2107.06546 [cs.CL]

Ambridge B, Lieven E. 2015. A constructivist account of child language acquisition. In *The Handbook of Language Emergence*, ed. B MacWhinney, W O'Grady, pp. 478–510. Chichester, UK: Wiley-Blackwell

Anderson JR. 1975. Computer simulation of a language acquisition system. In *Information Processing and Cognition: The Loyola Symposium*, ed. RL Solso, pp 295–349. Hillsdale, NJ: Lawrence Erlbaum

Athari P, Dey R, Rvachew S. 2021. Vocal imitation between mothers and infants. *Infant Behav. Dev.* 63:101531

Bergelson E, Amatuni A, Dailey S, Koorathota S, Tor S. 2019. Day by day, hour by hour: naturalistic language input to infants. *Dev. Sci.* 22(1):e12715

Bergelson E, Swingley D. 2012. At 6–9 months, human infants know the meanings of many common nouns. *PNAS* 109(9):3253–58

Bergmann C, Tsuji S, Piccinini PE, Lewis ML, Braginsky M, et al. 2018. Promoting replicability in developmental research through meta-analyses: insights from language acquisition research. *Child Dev.* 89(6):1996–2009

Bosseler AN, Clarke M, Tavabi K, Larson ED, Hippe DS, et al. 2021. Using magnetoencephalography to examine word recognition, lateralization, and future language skills in 14-month-old infants. *Dev. Cogn. Neurosci.* 47:100901

Braine MD, Bowerman M. 1976. Children's first word combinations. *Monogr. Soc. Res. Child Dev.* 41(1):1–104

Brent MR. 1996. Advances in the computational study of language acquisition. *Cognition* 61(1–2):1–38

Brookman R, Kalashnikova M, Conti J, Xu Rattanasone N, Grant KA, et al. 2020. Depression and anxiety in the postnatal period: an examination of infants' home language environment, vocalizations, and expressive language abilities. *Child Dev.* 91(6):e1211–30

Carbajal MJ, Peperkamp S, Tsuji S. 2021. A meta-analysis of infants' word-form recognition. *Infancy* 26(3):369–87

Casillas M, Brown P, Levinson SC. 2020. Early language experience in a Tseltal Mayan village. *Child Dev.* 91(5):1819–35

Casillas M, Brown P, Levinson SC. 2021. Early language experience in a Papuan community. *J. Child Lang.* 48(4):792–814

Casillas M, Cristia A. 2019. A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychol.* 5(1):24

Chrupała G, Gelderloos L, Alishahi A. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 1: *Long Papers*, pp. 613–22. Stroudsburg, PA: Assoc. Comput. Linguist.

Cychosz M, Cristia A. 2022. Using big data from long-form recordings to study development and optimize societal impact. In *Advances in Child Development and Behavior*, ed. JJ Lockman, R Gilmore, Vol. 62. Cambridge, MA: Academic. In press

Cychosz M, Cristia A, Bergelson E, Casillas M, Baudet G, et al. 2021. Vocal development in a large-scale crosslinguistic corpus. *Dev. Sci.* 24(5):e13090

Cychosz M, Romeo R, Soderstrom M, Scaff C, Ganek H, et al. 2020. Longform recordings of everyday life: ethics for best practices. *Behav. Res. Methods* 52:1951–69

de Boysson-Bardies B, Vihman MM. 1991. Adaptation to language: evidence from babbling and first words in four languages. *Language* 67(2):297–319

de Seyssel M, Dupoux E. 2020. Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors. In *CogSci - 42nd Annual Virtual Meeting of the Cognitive Science Society*, pp. 2791–97. **https://cogsci.mindmodeling.org/2020/papers/0683/0683.pdf**

Dupoux E. 2018. Cognitive science in the era of artificial intelligence: a roadmap for reverse-engineering the infant language learner. *Cognition* 173:43–59

Fernald A, Zangl R, Portillo AL, Marchman VA. 2008. Looking while listening: using eye movements to monitor spoken language. In *Language Acquisition and Language Disorders*, Vol. 44: *Developmental Psycholinguistics: On-line Methods in Children's Language Processing*, ed. IA Sekerina, EM Fernández, H Clahsen, pp. 97–135. Amsterdam: John Benjamins

Ferry A, Hespos S, Waxman S. 2010. Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Dev.* 81:472–79

Ganek H, Eriks-Brophy A. 2018. Language ENvironment Analysis (LENA) system investigation of day long recordings in children: a literature review. *J. Commun. Disord.* 72:77–85

Gasparini L, Langus A, Tsuji S, Boll-Avetisyan N. 2021. Quantifying the role of rhythm in infants' language discrimination abilities: a meta-analysis. *Cognition* 213:104757

Gross DR. 1984. Time allocation: a tool for the study of cultural behavior. *Annu. Rev. Anthropol.* 13:519–58

Harwath D, Hsu WN, Glass J. 2020. *Learning hierarchical discrete linguistic units from visually-grounded speech*. Paper presented at the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiop., Apr. 26–30

Hochmann J-R, Endress A, Mehler J. 2010. Word frequency as a cue to identify function words in infancy. *Cognition* 115:444–57

Hoff E, Core C, Bridges K. 2008. Non-word repetition assesses phonological memory and is related to vocabulary development in 20- to 24-month-olds. *J. Child Lang.* 35(4):903–16

Jaeger JJ. 1980. Testing the psychological reality of phonemes. *Lang. Speech* 23(3):233–53

Jusczyk PW, Luce PA, Charles-Luce J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *J. Mem. Lang.* 33(5):630–45

Lee GY, Kisilevsky BS. 2014. Fetuses respond to father's voice but prefer mother's voice after birth. *Dev. Psychobiol.* 56(1):1–11

Liaqat D, Wu R, Gershon A, Alshaer H, Rudzicz F, de Lara E. 2018. Challenges with real-world smartwatch based audio monitoring. In *WearSys '18: Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, pp. 54–59. New York: Assoc. Comput. Mach.

Long HL, Bowman DD, Yoo H, Burkhardt-Reed MM, Bene ER, Oller DK. 2020. Social and endogenous infant vocalizations. *PLOS ONE* 15(8):e0224956

MacWhinney B. 2000. *The CHILDES Project: The Database*, Vol. 2. New York: Psychol. Press

MacWhinney B. 2005. A unified model of language acquisition. In *Handbook of Bilingualism: Psycholinguistic Approaches*, ed. JF Kroll, AMB de Groot, pp. 49–67. Oxford, UK: Oxford Univ. Press

May L, Werker J. 2014. Can a click be a word?: Infants' learning of non-native words. *Infancy* 19(3):281–300

Nazzi T, Bertoncini J, Mehler J. 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. *J. Exp. Psychol.: Hum. Percept. Perform.* 24(3):756–66

Nguyen TA, de Seyssel M, Rozé P, Rivière M, Kharitonov E, et al. 2020. *The Zero Resource Speech Benchmark 2021: metrics and baselines for unsupervised spoken language modeling*. Paper presented at NeurIPS 2020 Virtual Workshop on Self-Supervised Learning for Speech and Audio Processing, Dec. 11

Nielsen M, Haun D, Kärtner J, Legare CH. 2017. The persistent sampling bias in developmental psychology: a call to action. *J. Exp. Child Psychol.* 162:31–38

Oller DK, Niyogi P, Gray S, Richards JA, Gilkerson J, et al. 2010. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *PNAS* 107(30):13354–59

Orena AJ, Byers-Heinlein K, Polka L. 2020. What do bilingual infants actually hear? Evaluating measures of language input to bilingual-learning 10-month-olds. *Dev. Sci.* 23(2):e12901

Pagliarini S, Leblois A, Hinaut X. 2021. Vocal imitation in sensorimotor learning models: a comparative review. *IEEE Trans. Cogn. Dev. Syst.* 13(2):326–42

Philippsen A. 2021. Goal-directed exploration for learning vowels and syllables: a computational model of speech acquisition. *KI - Künstliche Intell.* 35:53–70

Rasilo H, Räsänen O. 2017. An online model for vowel imitation learning. *Speech Commun.* 86:1–23

Robinaugh DJ, Haslbeck JMB, Ryan O, Fried EI, Waldorp LJ. 2021. Invisible hands and fine calipers: a call to use formal theory as a toolkit for theory construction. *Perspect. Psychol. Sci.* 16(4):725–43

Roopnarine JL, Fouts HN, Lamb ME, Lewis-Elligan TY. 2005. Mothers' and fathers' behaviors toward their 3- to 4-month-old infants in lower, middle, and upper socioeconomic African American families. *Dev. Psychol.* 41(5):723–32

Schatz T, Feldman NH, Goldwater S, Cao X-N, Dupoux E. 2021. Early phonetic learning without phonetic categories: insights from large-scale simulations on realistic input. *PNAS* 118(7):e2001844118

Schuller B, Batliner A, Bergler C, Pokorny FB, Krajewski J, et al. 2019. The INTERSPEECH 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2019)*, pp. 2378–82. Grenoble, Fr.: Int. Speech Commun. Assoc.

Seidl A, Cristia A, Soderstrom M, Ko ES, Abel EA, et al. 2018. Infant–mother acoustic–prosodic alignment and developmental risk. *J. Speech Lang. Hear. Res.* 61(6):1369–80

Shi R, Werker JF, Cutler A. 2006. Recognition and representation of function words in English-learning infants. *Infancy* 10(2):187–98

Simon DA, Gordon AS, Steiger L, Gilmore RO. 2015. Databrary: enabling sharing and reuse of research video. In *JCDL '15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 279–80. New York: Assoc. Comput. Mach.

Slobin DI. 2014. Before the beginning: the development of tools of the trade. *J. Child Lang.* 41(S1):1–17

Sun J, Harris K, Vazire S. 2020. Is well-being associated with the quantity and quality of social interactions? *J. Personal. Soc. Psychol.* 119(6):1478–96

Tamis-LeMonda CS, Kuchirko Y, Suh DD. 2018. Taking center stage: infants' active role in language learning. In *Active Learning from Infancy to Childhood*, ed. MM Saylor, PA Gane, pp. 39–53. Cham, Switz.: Springer

Turner BO, Paul EJ, Miller MB, Barbey AK. 2018. Small sample sizes reduce the replicability of task-based fMRI studies. *Commun. Biol.* 1:62

Twaddell WF. 1935. On defining the phoneme. *Language* 11(1):5–62

VanDam M, Warlaumont AS, Bergelson E, Cristia A, Soderstrom M, et al. 2016. HomeBank: an online repository of daylong child-centered audio recordings. *Semin. Speech Lang.* 37(2):128–43

Vouloumanos A, Waxman SR. 2014. Listen up! Speech is for thinking during infancy. *Trends Cogn. Sci.* 18(12):642–46

Warlaumont AS, Finnegan MK. 2016. Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PLOS ONE* 11(1):e0145096

Warlaumont AS, Westermann G, Oller DK. 2011. *Self-production facilitates and adult input interferes in a neural network model of infant vowel imitation*. Paper presented at AISB 2011: Study of Artificial Intelligence and Simulation of Behaviour, York, UK, Apr. 4–7

Weisleder A, Fernald A. 2013. Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychol. Sci.* 24(11):2143–52

Werker JF, Tees RC. 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* 7(1):49–63

Wu R, Liaqat D, de Lara E, Son T, Rudzicz F, et al. 2018. Feasibility of using a smartwatch to intensively monitor patients with chronic obstructive pulmonary disease: prospective cohort study. *JMIR mHealth uHealth* 6(6):e10046

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111(23):8619–24

Yeung HH, Werker J. 2009. Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition* 113(11):234–43

Yu C. 2014. Linking words to world: an embodiment perspective. In *The Routledge Handbook of Embodied Cognition*, ed. L Shapiro, pp. 139–49. New York: Routledge

3.2 On the importance of inductive biases for early phonetic learning

Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., Cristia, A. (2023) Statistical learning models of early phonetic acquisition struggle with child-centered audio data. *Submitted to Cognition*

Motivation

Putting our words into action, we present here a modeling investigation of early phonetic learning from ecological child-centered long-form recordings. Let us first describe the behavior we aimed to model.

Certain non-native speech contrasts pose a challenge for adult listeners. For instance, while most French speakers can effortlessly distinguish between [u] and [y] (as in 'dessous', *below* versus 'dessus', *above*), American English struggle to hear the difference (Levy & Strange, 2008b). Interestingly, this is not what we observe in young infants who discriminate not only native contrasts but also non-native contrasts to which they have never been exposed (Trehub, 1976; Werker et al., 1981). During their first year of life, infants transition from general listeners to specialized listeners whose discrimination patterns closely reflect the phonetic system of their native language(s) (Maye et al., 2002; Kuhl et al., 2006). This process is known as *perceptual attunement* or *perceptual narrowing*.

One influential mechanism proposed to explain these early perceptual changes is *statistical learning*, whereby infants track distributional cues from their native language(s) (Kuhl et al., 2008). These distributions would then cause a language-specific perceptual space to develop, which in turn alters infants' perception of native speech sounds (increasing discrimination) as well as non-native speech sounds (decreasing discrimination).

Support for the statistical learning hypothesis arises from two sources of evidence. The first source comes from laboratory experiments demonstrating that infants' perceptual abilities are sensitive to manipulated linguistic materials with altered distributional properties (Maye et al., 2002; Reh et al., 2021). The second source of evidence comes from modeling studies showing that it is possible to reproduce some

developmental results in speech perceptual learning using self-supervised algorithms that learn from raw speech, e.g., Schatz et al. (2021).

However, an important limitation of laboratory experiments and modeling studies is their ecological validity. Laboratory experiments rely on simplified stimuli (typically isolated syllables) that vary only in a few dimensions (typically the first formants). Similarly, modeling studies use highly manicured speech as input, with even the most realistic computational approaches using studio recordings of conversations or audiobooks (De Boer & Kuhl, 2003; Coen, 2006; Vallabha et al., 2007; Miyazawa et al., 2010; Schatz et al., 2021). Consequently, it remains unclear if perceptual attunement can emerge through statistical learning mechanisms applied to real children's language environments.

In Lavechin, De Seyssel, et al. (2023), currently under review in Cognition, we evaluate the presence of perceptual attunement – or absence thereof – in self-supervised learning algorithms trained on curated audiobooks or ecological child-centered long-forms.

Paper summary

We begin with a comparative analysis between audiobooks and long-forms. Analyzing the acoustic environments of both sources of data using the voice type classifier along with Brouhaha presented in Chapter 1, we show that speech utterances extracted from long-forms have a higher level of background noise ($\mu_{\text{SNR}} = 10$ dB in long-forms versus 47 dB in audiobooks). Our analysis also reveals that contrary to audiobooks, long-forms present a wide variety of reverberant environments with a C_{50} varying from -5 to 57 dB. Besides marked differences in terms of acoustic environments, the speech found in audiobooks differs greatly from that found in long-forms. Contrary to audiobooks, which consist of nearly 100% speech content, long-forms only contain approximately 20% speech, with the remainder comprising a variety of environmental noises such as vacuum cleaner sounds, music, traffic noise, and moments of silence. Genders are also more balanced in audiobooks than in long-forms. Finally, while audiobooks contain long stretches of speech read by the same speaker (estimated median turn duration of 17 minutes), the speech found in long-forms consists of short turns spoken by different speakers (estimated median turn duration of 2.4 seconds).

In view of the significant differences between well-articulated clean speech and the signal available to infants, we propose three *inductive biases*, i.e., mechanisms

guiding the learning process in our algorithms. These inductive biases are designed with two considerations in mind: 1) they should align with documented or plausible behavior found in infants; and 2) they should alleviate the signal degradations documented above. These inductive biases take the form of 1) a voice activity detection mechanism that filters out non-speech segments and forces the system to learn exclusively on the speech signal; 2) a pseudo-speaker separation mechanism that helps the system learns speaker-invariant representations; and 3) a data augmentation mechanism that induces invariance with respect to the various reverberant environments and the multiple voice fundamental frequencies found in long-forms (see details in the paper enclosed below).

We train our artificial learner on either Metropolitan French or American English, simulating the learning process of an infant acquiring either American English or Metropolitan French. The audio data used for training is obtained from either audiobooks or ecological long-forms. Furthermore, our learner comes in two flavors: with or without inductive biases. Using the same machine ABX sound discrimination task as used in Section 2.3, we consider two outcome measures: the *native discrimination* (to what extent the learner discriminates between native contrasts) and the *native advantage* (to what extent it does so better than a non-native learner). The native advantage is a measure of perceptual attunement. A positive native advantage indicates that the native learner is better at discriminating the sounds of its native language than the non-native learner is on the same sounds. A negative native advantage indicates the opposite, and a native advantage of 0 indicates that the native and the non-native learners are equally good at discriminating the same sounds.

Our results indicate that reproducing perceptual attunement when training on audiobooks is possible, regardless of whether the learner is pre-equipped with inductive biases. This constitutes a replication of past studies, e.g., Schatz et al. (2021). However, on ecological long-forms, the picture is drastically different. Only the learner equipped with inductive biases reproduces perceptual attunement, and this remains true regardless of the amount of data available for training.

In an ablation study, we evaluate the individual contribution of our three inductive biases. Although, on ecological long-forms, the most important contribution is brought by the voice activity detection mechanism, all three inductive biases help the learner develops a higher native discrimination and native advantage. Finally, we show that the learning outcomes developed by our system are highly sensitive to 1) the proportion of non-speech present in the training set; and 2) the quantity of background noise and reverberation.

By design, statistical learning algorithms learn the underlying structure of their input data. In children's language environments, speech exhibits an intricate structure impacted by various factors, including the speaker's identity, the way sounds propagate in the environment, and the various sources of background noises. Structure can also be found in non-speech sounds, which constitute the large majority of the infants' auditory stream. Yet, infants attune only to speech. Similarly to our artificial learner, infants likely come pre-equipped with inductive biases that may guide the language acquisition process. We suspect that these inductive biases are, to a certain extent, inherited from our evolutionary past, and to another extent, learned. Achieving a deeper comprehension of how these biases come to be active in human infants will likely necessitate a collaborative endeavor involving theoretical and empirical efforts from computational, developmental, and behavioral scientists.

Finally, we showed that the learning outcomes of our artificial learner were exquisitely sensitive to the details of the input signal, with drastically different behaviors when training on audiobooks or long-forms. This illustrates how considering input data that do not reflect characteristics of children's language environments can lead us to underestimate the complexity of a given learning problem (that is, we may believe a problem has been solved when it has only been solved on unrealistic data), or overestimate the power of a proposed mechanism (that is, we may believe a mechanism is sufficient to explain the phenomenon that is being modeled when it is not).

Statistical learning models of early phonetic acquisition struggle with child-centered audio data

Marvin Lavechin[a,b,c,*], Maureen de Seyssel[a,b,d], Marianne Métais[a,b], Florian Metze[c], Abdelrahman Mohamed[e], Hervé Bredin[f,1], Emmanuel Dupoux[a,b,c,1] and Alejandrina Cristia[a,b,1]

[a]*Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France*
[b]*Cognitive Machine Learning Team, INRIA, Paris, France*
[c]*Meta AI Research, Paris, France*
[d]*Laboratoire de linguistique formelle, Université de Paris, CNRS, Paris, France*
[e]*Rembrand, Palo Alto, California, United States*
[f]*Institut de Recherche en Informatique de Toulouse, Université de Toulouse, CNRS, Toulouse, France*

ARTICLE INFO

Keywords:
language acquisition
computational modeling
statistical learning
phonetic learning
perceptual attunement
child-centered long-forms

ABSTRACT

Infants learn their native language(s) at an amazing speed. Before they even talk, their perception adapts to the language(s) they hear. However, the mechanisms responsible for this *perceptual attunement* remain unclear. The currently dominant explanation for perceptual attunement posits that infants apply a *statistical learning* mechanism consisting in learning regularities from the speech stream they hear, and which may be found in learning across domains and species. Critically, the feasibility of the statistical learning hypothesis has only been demonstrated with computational models on unrealistic and simplified input. This paper presents the first attempt to study perceptual attunement with 2,000 hours of ecological child-centered recordings in American English and Metropolitan French. We show that, when applied on ecologically valid data, generic learning mechanisms develop a language-relevant perceptual space but fail to show evidence for perceptual attunement. It is only when supplemented with inductive biases, in the form of data filtering, sampling, and augmentation mechanisms that computational models show a significant attunement to the language they have been exposed to. As inductive biases are necessary for our model to become attuned to their native language, we reflect on whether similar inductive biases may shape early phonetic learning in infants. More generally, we show that *what* our model learns, and *how* it develops through exposure to speech, depends exquisitely on details of the input signal. By doing so, we illustrate the importance of considering ecologically valid input data when modeling language acquisition.

1. Introduction

1.1. Perceptual attunement to sounds

Our ability to discriminate things tends to increase for frequently encountered stimuli and to decrease for infrequently encountered ones. For instance, while most French speakers have no difficulties distinguishing between [u] and [y] (as in 'dessous', *below* versus 'dessus', *above*), American English speakers struggle hearing the difference (Levy and Strange, 2008). Similarly, Japanese native speakers often confuse [ɹ] and [l] (as in 'right' versus 'light') where American English speakers do not (Miyawaki, Jenkins, Strange, Liberman, Verbrugge and Fujimura, 1975). The way adults perceive speech sounds is therefore shaped by the language(s) they have been exposed to. In other words, their perception is attuned to their native language(s).

Meta-analytic evidence suggests that the process of tuning in to certain sounds, known as *perceptual attunement*, begins in early childhood and gives rise to one of the earliest language-specific effects in infant speech development (Tsuji and Cristia, 2014; Singh, Rajendra and Mazuka, 2022). According to the perceptual attunement account, young infants can distinguish between different sounds regardless of the language they are exposed to. However, as they age, infants' ability to distinguish between sounds that are relevant to their native language improves while their ability to distinguish non-native sounds deteriorates (Kuhl, 2004). As an example, at 6-8 months, American English and Japanese infants show similar discrimination scores for the [ɹ] versus [l] pair. However, by 10-12 months, American English infants' discrimination score for the [ɹ] versus [l] pair increases, while Japanese infants' discrimination score deteriorates (Kuhl, Stevens, Hayashi, Deguchi, Kiritani and Iverson, 2006), resulting in a *native advantage* (higher scores for the infants for whom the contrast is native).

1.2. Can statistical learning account for early phonetic learning?

These early perceptual changes, which occur before infants talk, have been linked to the *statistical learning hypothesis* (Maye, Werker and Gerken, 2002; Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola and Nelson, 2008). This hypothesis states that infants "use statistical properties of linguistic input to discover structure, including sound patterns, words, and the beginnings of grammar" (Saffran, 2003). In the case of early phonetic acquisition[1], the idea is

*Corresponding author at Département d'Etudes Cognitives, Ecole Normale Supérieure, Paris, France
E-mail address: marvinlavechin@gmail.com (M. Lavechin)
[1]H.B., E.D., and A.C. contributed equally to this work.

[1]Note that early phonetic learning/acquisition as employed in this manuscript should not be confused with phonetic category learning. Evidence suggests that sound perception continues to develop well beyond

that infants track the statistical distribution of sounds in their native language (Maye et al., 2002; Kuhl et al., 2008). These distributions would then cause a language-specific perceptual space to develop, which in turn alters infants' perception of native speech sounds (increasing discrimination) as well as non-native speech sounds (decreasing discrimination).

A first source of evidence in favor of the statistical learning hypothesis for phonetic learning is found in laboratory experiments demonstrating that exposing infants to linguistic materials whose distributional properties have been manipulated alters infants' perceptual abilities. Specifically, when exposed to a bimodal distribution of speech sounds, infants are able to distinguish between the sounds, but when exposed to a unimodal distribution, they are not (Maye et al., 2002; Reh, Hensch and Werker, 2021). A key limitation of these studies, however, is that they rely on simplified stimuli (typically isolated syllables) that vary only in a few dimensions (typically first formants). In addition, a prerequisite of the statistical learning hypothesis is that native versus non-native contrasts are cued by one versus two modes along some acoustic dimension, but this is not the case for many sound contrasts, including vowel length Bion, Miyazawa, Kikuchi and Mazuka (2013). Thus, from in-laboratory experiments alone, it remains unclear whether early phonetic learning from ecological input can occur through statistical learning.

A second source of evidence in favor of the statistical learning hypothesis for early phonetic learning lies in computational modeling studies. After all, if infants develop a language-specific perceptual space via statistical learning, computers should be able to reproduce it, and they may be able to do it relying on more complex approaches than noticing unimodal versus bimodal distributions. Hitczenko and Feldman (2022) shows that duration cues are sufficient to recover when vowel length is contrastive, as long as contextual cues are considered. Using spectrographic input representations, Schatz, Feldman, Goldwater, Cao and Dupoux (2021) reproduce some developmental results in speech perceptual learning using a rather simple self-supervised learning algorithm based on mixtures of Gaussians. Within the field of machine learning, a myriad of powerful self-supervised learning algorithms have been recently developed and tested in several languages (Versteegh, Thiollière, Schatz, Cao, Miró, Jansen and Dupoux, 2017; van den Oord, Li and Vinyals, 2019; Schneider, Baevski, Collobert and Auli, 2019), re-instilling hope that statistical learning is indeed a plausible mechanism to account for early phonetic learning in infants. An important limitation of this line of work, however, is that up to now, it has only been applied to highly curated inputs, with even the most realistic computational approaches using studio recordings of conversations or audiobooks (De Boer and Kuhl, 2003; Coen, 2006; Vallabha, McClelland, Pons, Werker and Amano, 2007; Miyazawa, Kikuchi and Mazuka, 2010; Schatz et al., 2021; Millet, Chitoran and Dunbar, 2021).

the first year (McMurray, Danelz, Rigler and Seedorff, 2018) and that well-defined discrete phonetic categories may only be acquired later in life (Feldman, Goldwater, Dupoux and Schatz, 2022; McMurray, 2022b).

Corpus	Audio dur.	Speech prop.	Women ratio	Median turn dur.
		Audiobooks		
R-Eng.	1071h	94.9%	51.0%	14.5mn
R-Fr.	1089h	94.0%	57.3%	20.4mn
		Long-forms		
E-Eng.	1054h	17.6%	67.6%	2.3s
E-Fr.	1008h	19.0%	69.9%	2.5s

Table 1
Audio duration, proportion of adult speech, proportion of adult speech pronounced by women, and median turn duration for our 4 corpora. R-<language> indicates read-speech corpora from audiobooks. E-<language> indicates ecological audio corpora from child-centered long-form recordings. Proportions of speech are estimated using a pretrained model (Lavechin, Bousbib, Bredin, Dupoux and Cristia, 2020). The median turn duration is estimated as 1) the median duration of chapters for audiobooks (each chapter is read by a single speaker); and 2) the median cumulated duration of successive sequences produced by the same voice type for long-forms. Data are taken from (Kahn, Rivière, Zheng, Kharitonov, Xu, Mazar'e, Karadayi, Liptchinsky, Collobert, Fuegen, Likhomanenko, Synnaeve, Joulin, rahman Mohamed and Dupoux, 2020; Bergelson, 2017; Canault, Normand, Foudil, Loundon and Thai-Van, 2016a; Cristia, 2021; Lavechin and Cristia, 2021).

1.3. What infants truly hear

Neither studio recordings of isolated syllables nor audiobooks reflect what infants truly hear, constituting crucial limitations of experimental and modeling studies alike. First, the large majority of what children hear during their wake time is not speech at all, but various ambient sounds, noise, music, or non-linguistic vocalizations. Second, the speech sections, are not isolated syllables but multi-words utterances. Additionally, speech signals are far from being recorded in a studio but may be distorted as they are spoken far from the child and are multiply reverberated and absorbed by the surrounding obstacles in the environment. Speech signals may also be covered by a variety of background noises or crowded by other concurrent speech sounds. Finally, in real life, people do not speak in full and well-articulated sentences as in audiobooks, but may speak in ways that distort the clear articulation of phonemes: people may produce short turns that sometimes overlap across speakers, and they may under-articulate, mumble, shout, whisper, sing, or laugh while speaking. Given what infants truly hear, can we assume that the statistical learning mechanisms posited from laboratory study will work in real life?

From this point on, we focus on the difference between audiobooks commonly used in computational modeling studies and the audio children are exposed to. In Table 1, we compare the statistics of open-source audiobooks, self-recordings of books by volunteers using home equipment (Kearns, 2014) and long-forms that are captured using lightweight recorders worn by the infant, which continuously

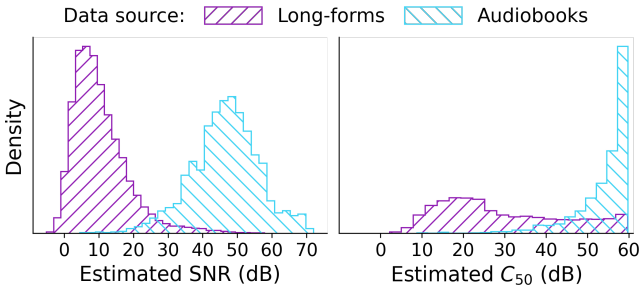

Figure 1: The quantity of noise and reverberation in child-centered long-forms and audiobooks. Speech-to-Noise Ratio (SNR) and C_{50} distributions on 16 hours of speech utterances in English extracted from long-forms (in purple) or audiobooks (in blue). Both measures are automatically extracted using the pretrained model proposed in Lavechin et al. (2022).

collect audio over long periods of time (typically 8+ hours of the infant's waking time). Whereas audiobooks contain nearly 100% of speech, long-forms contain only about 20% of speech, the rest being various environmental noises (vacuum cleaner, music, traffic noise, silence, etc.). Genders are also more balanced in audiobooks than in long-forms, and while audiobooks contain long stretches of speech read by the same speaker (median turn duration of 17mn), long-forms contain relatively short turns spoken by different speakers (median turn duration of 2.4s).

Figure 1 displays signal characteristics of the two sources of audio, notably: 1) the Speech-to-Noise Ratio (SNR) which measures the strength of the speech signal relative to the strength of background noise (the higher the SNR the lower the amount of noise); and 2) the C_{50} measure which quantifies the level of reverberation (a higher C_{50} indicates less reverberation). Both measures are computed at the utterance level (i.e., speech only) on either English long-forms or audiobooks using the pretrained model proposed in (Lavechin, Métais, Titeux, Boissonnet, Copet, Rivière, Bergelson, Cristia, Dupoux and Bredin, 2022). Our analysis reveals that, on average, speech utterances extracted from audiobooks have an SNR of 46 dB, while those from long-forms have an SNR of 10 dB, indicating, therefore, a higher level of background noise in long-forms. Similarly, audiobooks contain utterances with a high C_{50} indicating a low reverberation level. On the contrary, utterances extracted from long-forms span a C_{50} between -5 and 57 dB indicating a wider variety of reverberant conditions. In total, audiobooks that are recorded in a quiet environment, close to the microphone, tend to be of higher quality than long-form recordings, that are recorded in uncontrolled environment, with the source of speech being often far away from the listener child. Note that similar results for both the SNR and the C_{50} were found in our French corpora.

1.4. Inductive biases to guide early phonetic learning

The difference between clean and well-articulated speech and the signal available to infants is so large that one might rightfully wonder: would statistical learning lead infants to discover the structure of sounds, or would it rather lead them to focus on the dominant non-linguistic structure of the audio?

It remains possible that infants' early phonetic learning involves more than simply statistical learning. Indeed, among other astonishing capacities, human infants have a preference for listening to speech from birth (Cooper and Aslin, 1990; Vouloumanos and Werker, 2007), come pre-equipped with an auditory system capable of source separation (Bregman, 1994), and display an early ability to discriminate human voices (Decasper and Prescott, 1984; Floccia, Nazzi and Bertoncini, 2000). Could these abilities shape or guide statistical learning in order to improve the acquisition of species-relevant communication signals?

In machine learning, mechanisms that constraint or guide the learning process by forcing the algorithm to learn on specific parts of the signal, or to extract specific information are called *inductive biases*(Hüllermeier, Fober and Mernberger, 2013). In this study, we equip a computational model with inductive biases, and study how they affect the learning of our model from ecological signals. We propose inductive biases that have been designed with two considerations in mind. First, they must be designed to address the signal degradations found in long-forms, as documented in Section 1.3. Second, they should align with documented or plausible behaviour found in infants (more details to be found in Experiment 1). To the extent that our inductive biases are effective in coping with the signal degradations in long-form recordings, this provides support to the hypothesis that infants too are using such inductive biases, in addition to or in conjunction with, statistical learning.

1.5. Key questions

The aim of this paper is to study the effect of input realism (audiobooks versus long-forms) and inductive biases (absence or presence) in a computer simulation of learning. This is relevant for both theoretical and practical reasons. On the theory side it helps us assess the kind of learning mechanisms infants likely need to bring to the task. On the practical side, it indicates the level of evidence we can obtain from a given modeling experiment. That is, if results from audiobooks can not be replicated with truly naturalistic data (long-forms), then this casts important doubts on any work relying on audiobooks.

In this study, we examine one of the earliest language-specific phenomenon, perceptual attunement, and ask:

1. Can statistical learning lead to perceptual attunement when given such sparse, variable and noisy signals as found in ecological data?
2. Can inductive biases help overcoming difficult conditions found in long-forms?
3. What characteristics of ecological long-forms impact the phonetic learning outcomes of our learner?

In Experiment 1, we evaluate whether our base learner (trained without inductive biases) can reproduce perceptual

attunement when exposed to either audiobooks or long-forms. In addition, we measure the impact of our inductive biases in the long-form condition. In Experiment 2, we examine the impact of varying the quantity of audio to determine if conclusions drawn from our first Experiment remain consistent. In Experiment 3, we conduct a complete ablation study examining the individual contribution of each inductive bias. In a similar vein, Experiment 4 zooms in on one of our three inductive biases (Voice Activity Detection, presented below), progressively degrading it to study its contribution. In Experiment 5 we check whether our learner trained with inductive biases is still sensitive to noise and reverberation, two salient characteristics of long-forms. Finally, in Experiment 6, we explore the possibility to simulate ecological long-forms with audiobooks artificially contaminated with noise and reverberation.

2. Experiment 1: the impact of inductive biases

In this first Experiment, we ask how native discrimination and native advantage differ as a function of the learner and broad features of the environment. Whereas (Schatz et al., 2021) chose a learning algorithm (mixture of Gaussians) related to the hypothesis that infants track the modes in distributions of acoustic properties Maye et al. (2002), but as a result only builds representations in short-time windows (10ms), we chose a learning algorithm related to the predictive coding hypothesis Huang and Rao (2011), that better captures the temporal dynamics of speech. In a nutshell, the algorithm learns by attempting to predict future representations of speech based on past ones (details in the Methods section).

We first train our base learner on audiobooks, which constitutes a conceptual replication of past studies (Schatz et al., 2021) with a more powerful algorithm and on a different pair of languages. We separately train a learner with the same algorithm on realistic child-centered long-forms: If pure statistical learning suffices, then we should observe perceptual attunement in this condition as well. In addition, we introduce a modified learner incorporating three inductive biases aiming at guiding the learning process in our algorithm. Based on newborns' preference for speech (Cooper and Aslin, 1990; Vouloumanos and Werker, 2007), we propose an inductive bias that helps our learner deal with the high quantity of non-speech found in long-forms (Table 1) by restricting learning to speech segments. Based on infants early ability to discriminate human voices (Mehler, Bertoncini, Barriere and Jassik-Gerschenfeld, 1978; Decasper and Fifer, 1980), we propose an inductive bias that leverage the speaker information to guide our learner deal with the frequent speaker change and the gender imbalance (Table 1); this is done by restricting learning to segments that plausibly come from the same speaker. Our third inductive bias helps our learner in achieving better perceptual constancy which is an important feature of the human auditory system (Kuhl, 1979; Beeston, Brown and Watkins, 2014); this is done by nudging the learning algorithm to be invariant with respect to changes in pitch and to audio modifications induced listening conditions. By comparing the learning outcomes obtained by our base learner with those of our speech-biased learner, also trained on long-forms, we measure the impact of the inductive biases with which our biased learner is endowed.

2.1. Methods

Our simulation approach consists in implementing three key components of language learning (Dupoux, 2018): the environment, the learner and the outcome measure (see Fig. 2). Regarding the environment, we build on 2,000 hours of child-centered, ecologically valid audio recordings of infants learning either American English or Metropolitan French, which represents an essential step forward as compared to previous modeling studies. Regarding the learner, we employ one of the best self-supervised learning algorithms as a key component of our learner (Dunbar, Bernard, Hamilakis, Nguyen, de Seyssel, Rozé, Rivière, Kharitonov and Dupoux, 2021). The learner's model we built for this paper comes in two flavors: the *base learner* implements a predictive coding mechanism (introduced below), whereas the *speech-biased* learner filters and processes its input before applying the predictive coding mechanism. Finally, regarding the outcome, we employ the same outcomes as previous work (Schatz et al., 2021). In a nutshell, we will study *native discrimination* (to what extent the learner discriminates between native contrasts) and *native advantage* (to what extent it does so better than a non-native learner).

2.1.1. Environment: training datasets

Audiobooks. The English and French read-speech corpora were built from audiobooks using the LibriVox platform (Kearns, 2014; Kahn et al., 2020).

Long-forms. The long-form training set was built from 156 recordings from 80 children (40 female) aged 2-48 months from four studies: the American English SEEDLingS study (Bergelson, Casillas, Soderstrom, Seidl, Warlaumont and Amatuni, 2019; Bergelson, 2017), and three Metropolitan French studies (Canault, Normand, Foudil, Loundon and Thai-Van, 2016b; Canault et al., 2016a; Cristia, 2021; Lavechin and Cristia, 2021). All recordings were collected using a LENA audio recorder (single channel, 16 kHz; see Ford, Baer, Xu, Yapanel and Gray (2008) for full specification). Half of the recordings came from English infants and the other half from French infants. The French and English sets of recordings were matched as much as possible for the age of the child wearing the recording device, within the limitations of independently collected corpora. This resulted in a set of Metropolitan French recordings acquired from 40 children (21 female) whose age varies from 2 to 41 months ($\mu_{FR} = 17.3$ mths, $\sigma_{FR} = 10.9$ mths); and a set of American English recordings acquired from 40 children (19 female) whose age varies from 6 to 17 months ($\mu_{EN} = 12.5$ mths, $\sigma_{EN} = 3.3$ mths).

Creation of training sets. For each language (English or French) and each audio source (audiobooks or long-forms),

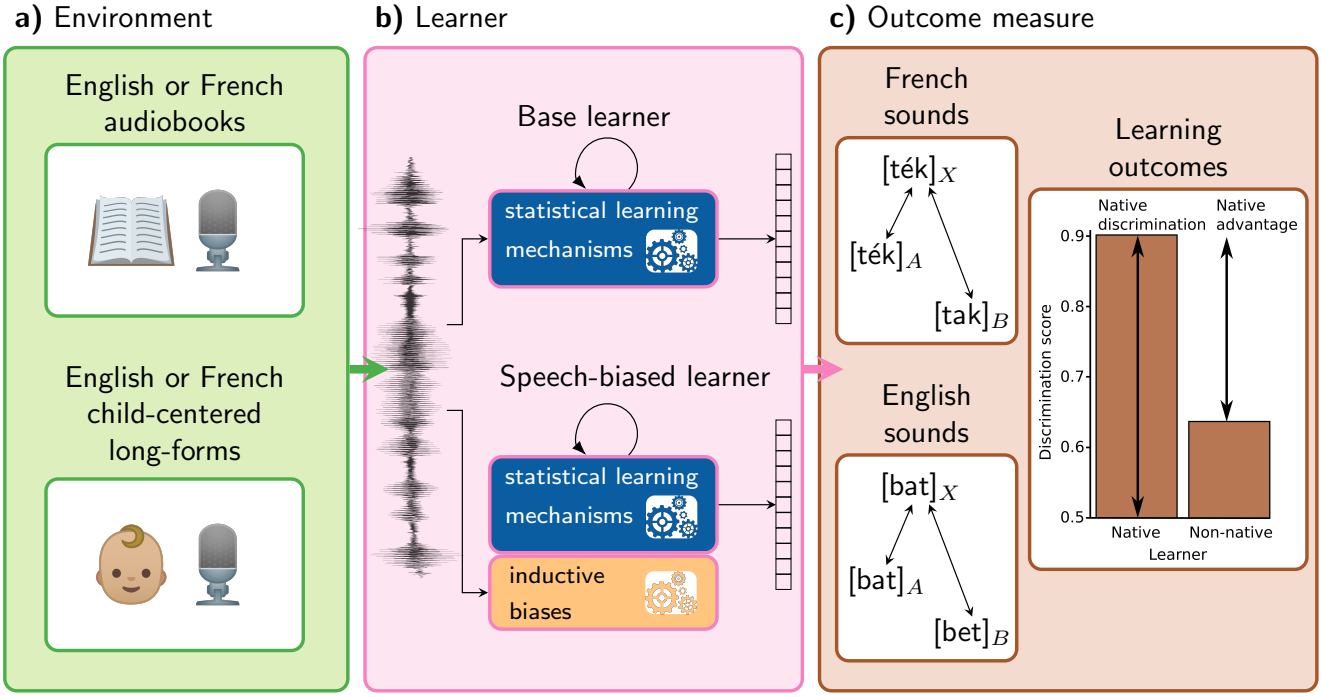

Figure 2: Phonetic learning simulation. **(a)** Simulated language environments in American English and Metropolitan French built from (top) *audiobooks*, or (bottom) child-centered *long-forms*. **(b)** Simulated learners built from (top) *statistical learning mechanisms* (contrastive predictive coding applied on nearby audio sequences), (bottom) with the addition of *inductive biases* (filtering of non-speech, speaker-invariant and pitch-/room-resistant training). **(c)** Outcome measures obtained via an ABX auditory discrimination task in English and French. *Native discrimination* measures the ability of the learner to discriminate sounds in its native language. *Native advantage*, measures perceptual attunement, i.e., the extent to which the native learner is better at discriminating sounds of its native language than the non-native learner.

we built mutually exclusive training sets of 128 hours so as to measure the robustness of the learning outcomes across different language exposures. As our base learner receives both speech and non-speech, while our speech-biased learner receives exclusively speech segments, the number of training sets depends both on the source of the data (audiobooks or long-forms) and the type of learner considered (base or speech-biased). We created 8 training sets from audiobooks and 7 from long-forms for our base learner (containing both speech and non-speech). We were only able to create one training set of 128 hours of speech from long-forms, as that was the amount of speech present in our 1,000 hours of long-forms, reflecting the scarcity of speech in these naturalistic data. (see Appendix A for more details about the creation of the training set).

2.1.2. Learners and baseline

Statistical learning from raw audio. In this section, we present the mechanism at the heart of both our base and speech-biased learners.

We chose *Contrastive Predictive Coding* (CPC) (van den Oord et al., 2019) as the heart of our learners. In machine learning, CPC has been shown to be powerful in a wide variety of modalities ranging from audio and images to natural language and reinforcement learning (van den Oord et al., 2019). In the ZeroSpeech 2021 international challenge on unsupervised representation learning, CPC was the best system to learn representations that accurately discriminate speech sounds (Dunbar et al., 2021). The key idea behind CPC is to predict the future states of a sequence given its past context. The learner is given an example that is drawn from the near future up to 120 ms (called positive example), and multiple examples that are not drawn from the near future (called negative examples). Given the past context of a sequence, the learner maximizes the categorical cross-entropy of classifying the positive sample correctly (see Contrastive Predictive Coding panel, top right of Fig. 3).

As originally proposed in (van den Oord et al., 2019), we used a contrastive loss which forces the latent space to retain information useful to predict future samples. Precisely, the input sequence of observations x_t is mapped to a sequence of latent representations through an encoder g_{enc}, such that $z_t = g_{enc}(x_t)$. Then, all $z_{\leq t}$ are aggregated with an autoregressive model that produces a context-dependent latent representation $c_t = g_{ar}(z_{\leq t})$. Given the past context c_t, a predictor g_{pred} is asked to predict future representations z_{t+k} for $k \in \{1, ..., K\}$. Given a set $X = \{x_1, ..., x_n\}$ of N random samples containing one positive sample from the true positive distribution $p(x_{t+k} \mid c_t)$ and $N-1$ negative samples from the proposal negative distribution $p(x_{t+k})$, we optimize the categorical cross-entropy loss of classifying the positive sample correctly:

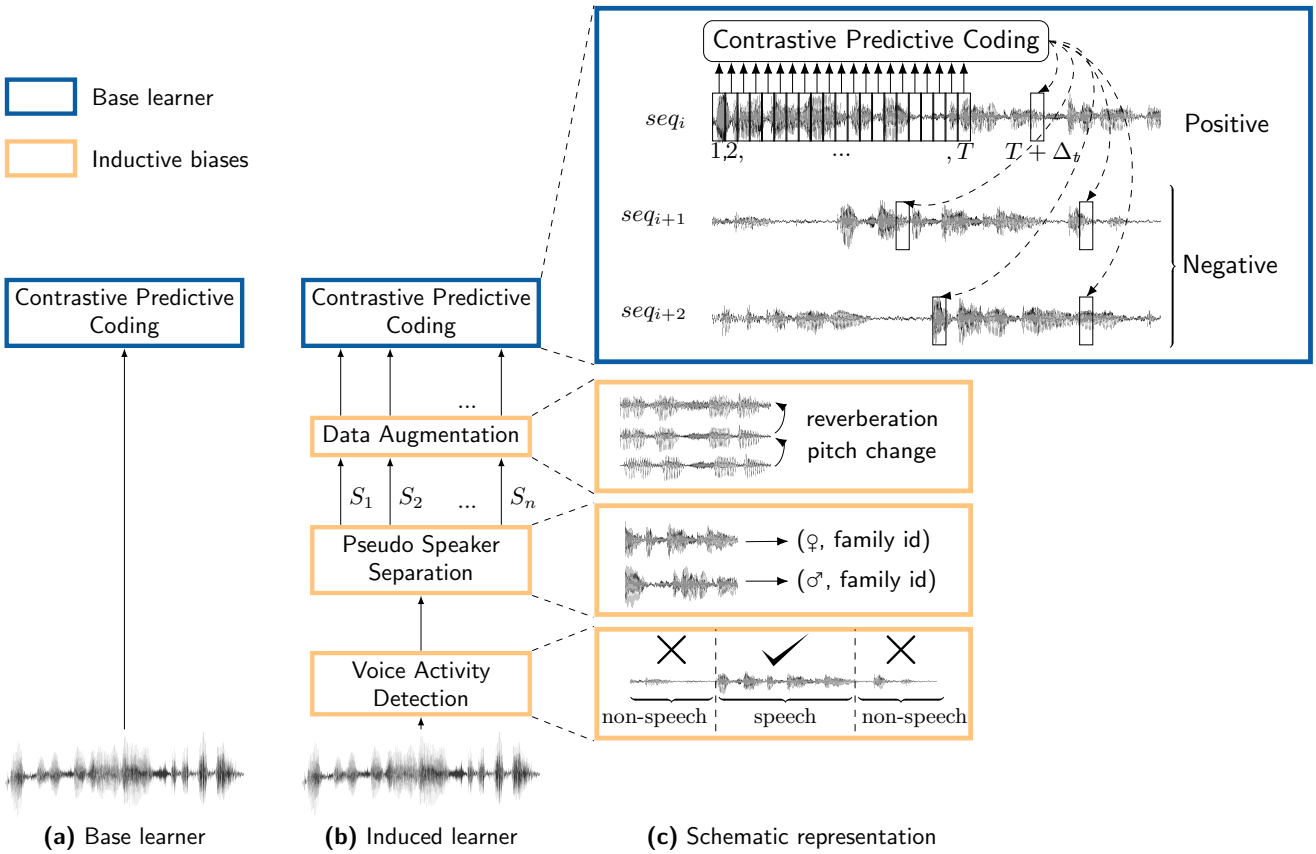

Figure 3: Base learner versus speech-biased learner. Schematic structure of **(a)** our base learner (trained without inductive biases); and **(b)** our speech-biased learner (trained with inductive biases). *Contrastive Predictive Coding (CPC)* learns audio representations from predicting the near future; *Voice Activity Detection* filters out non-speech segments of the audio. *Pseudo Speaker Separation* is used to sort the different speech segments according to by whom they have been pronounced. *Data Augmentation* leads to both channel and pitch mismatch between past and future samples.

$$\mathcal{L} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{\exp\ g_{pred}(c_t)^T z_{t+k}}{\sum_{x_j \in X} \exp\ g_{pred}(c_t)^T z_j}\right]$$

A key difference compared to previous implementations (Rivière, Joulin, Mazaré and Dupoux, 2020; Kharitonov, Rivière, Synnaeve, Wolf, Mazaré, Douze and Dupoux, 2021) lies in the fact that we sample negatives from temporally close sequences from the time series of interest. This strategy ensures that the positive/negative examples used in the contrastive task takes place in a short time span, therefore reducing mismatch between both types of examples in terms of their local environment. Implementation details can be found in Appendix B.

Inductive biases In this section, we present inductive biases designed to help our speech-biased learner to overcome difficulties found in realistic audio (see Section 1.3). For the present study, these mechanisms are fixed, but other work could assess whether they may be learned through exposure or whether they may have been selected throughout evolution.

Voice Activity Detection. First, based on newborns' preference for speech (Cooper and Aslin, 1990; Vouloumanos and Werker, 2007), we propose an inductive bias that discards non-speech segments and induces the learner to learn only from the speech signal, alleviating the issue of the high proportion of non-speech found in long-forms (see Table 1). This is achieved via a pretrained *Voice Activity Detection* (VAD) model Lavechin et al. (2020).

Pseudo Speaker Separation. Second, a *Pseudo Speaker Separation*[2] inductive bias groups the speech segments according to by whom they have been produced. By way of illustration, without this bias, positive and negative examples could be spoken by different speakers, inducing the learner to focus on low-level acoustic differences between speakers. Pseudo Speaker Separation decreases the probability of having a speaker mismatch between the positive and the negative examples, inducing the learner to learn speaker-invariant representations (van den Oord et al., 2019). This bias incorporates the idea that infants have an early ability to discriminate human voices (Mehler et al., 1978; Decasper

[2]The prefix 'pseudo' refers to the fact that we approximate speaker identity by combining the gender information with the recording unique identifier.

and Fifer, 1980), including unfamiliar ones (Decasper and Prescott, 1984; Floccia et al., 2000), but that they also progressively learn representations that are invariant with respect to a change in speaker (see Seidl, Onishi and Cristia (2014); Bergmann, Cristia and Dupoux (2016) for behavioral evidence found in infants, and Choi and Shukla (2021) for a review). This bias aims at alleviating the issues of the short turns and the gender imbalance found in long-forms (see Table 1).

Data Augmentation. Third, the *Data Augmentation* inductive bias consists in applying acoustic transformations to the past context of each segment, but not to the segments from which the positive and the negative examples are drawn (as proposed in (Kharitonov et al., 2021)), therefore inducing CPC to learn representations that are invariant with respect to these transformations and achieve better perceptual constancy (Kuhl, 1979; Beeston et al., 2014). Here we use two such transformations: artificial reverberation and pitch modification. Artificial reverberation induces invariance with respect to the distance of the speaker to the 'ears' of the artificial learner and to the presence of various sound-reflective objects in the environment. Pitch modification, on the other hand, induces invariance with respect to voice fundamental frequency, which varies across speakers as a function of their vocal cord anatomy. Pitch modification aims at alleviating the short turn duration and the gender imbalance found in long-forms (Table 1) while artificial reverberation aims at alleviating the wide variety of reverberant environments (Fig. 1).

Auditory baseline Following recommendations to include a control to check for the effects of learning, we compare the performance of our learners against a simple auditory baseline. We first slice the speech signal into 25ms-long frames sampled every 10ms. Descriptors of each frame consist of mel-frequency cepstral coefficients (MFCCs) with the first and second derivatives, resulting in a 39-dimensional feature vector for each 10ms-frame.

2.1.3. Outcomes: the machine ABX discrimination test

The ABX discrimination test is commonly used to study human perceptual system (Schatz, Peddinti, Bach, Jansen, Hermansky and Dupoux, 2013). In this test, the machine is given three triphones: A, B, and X, with A and X two different occurrences of the same triphone (e.g, /bup/) and B another triphone differing only in its center phone (e.g., /bœp/). Representations of each stimulus (R_A, R_B and R_X) are extracted and pairwise distances $d(R_A, R_X)$ and $d(R_B, R_X)$ are computed. As representations can have different lengths (depending on the duration of the input stimuli), the distance between two representations is computed along their dynamic time-warped alignment. Following (Versteegh et al., 2017), we used the cosine distance function to measure dissimilarity between individual frames (Supplementary Fig. 9). The machine is considered to be right if $d(R_A, R_X) < d(R_B, R_X)$ as A and X represent the same triphone. Context, as defined by the preceding and the following sound, is controlled as the same sound can be pronounced differently in different contexts. Thus, for each contrast (e. g., /u/ vs /œ/) and each context (e.g., [b_p]), a discrimination accuracy is computed as the amount of times the learner is right. If we note, $S(C_1)$ the set of sounds from category C_1 (e.g., all /bup/ triphones of the evaluation set) and $S(C_2)$ the set of sounds from category C_2 (e.g., all /bœp/ triphones of the evaluation set), the non-symmetric discrimination accuracy between category C_1 and C_2 can be computed as:

$$\theta(C_1, C_2) := \frac{1}{m(m-1)n} \sum_{A \in S(C_1)} \sum_{B \in S(C_2)} \sum_{\substack{X \in S(C_1) \\ X \neq A}} \mathbb{1}_{d(R_A,R_X)<d(R_B,R_X)} + \frac{1}{2}\mathbb{1}_{d(R_A,R_X)=d(R_B,R_X)}$$

with m and n the cardinal of $S(C_1)$ and $S(C_2)$ respectively, $\mathbb{1}$ the indicator function, and R_A, R_B and R_X the representations of sounds A, B and X respectively. A symmetric discrimination accuracy is obtained by averaging $\theta(C_1, C_2)$ and $\theta(C_2, C_1)$.

Representations are extracted from the last layer of the auto-regressive model, which results in a 256-dimensional feature vector for each 10ms frame.

The ABX discrimination test was built from 10 hours of speech downloaded from Common Voice (Ardila, Branson, Davis, Henretty, Kohler, Meyer, Morais, Saunders, Tyers and Weber, 2020) for each of the two target languages, English and French. More information about the test data can be found in Appendix C. The evaluated phonetic inventory for each of the two target languages can be found in Supplementary Table 4.

Based on this ABX discrimination test, we consider two outcome measures, which we call native discrimination and native advantage. *Native discrimination* is the average ABX accuracy of both learners when tested on all contrasts of the language they have been exposed to, measuring how good the learner is at discriminating sounds of their native language. *Native advantage* is the relative difference between native and non-native learners averaged across the two languages. It measures to which extent our learners develop a language-specific perceptual space. A positive native advantage indicates that the native learner is better at discriminating the sounds of its native language than the non-native learner is on the same sounds, similarly to what we observe in infants towards the end of the first year of life Kuhl et al. (2006). A negative native advantage indicates the opposite, and a native advantage of 0 indicates that the native and the non-native learners are equally good at discriminating the same set of sounds.

2.2. Results and discussion

Panel (a) of Fig. 4 shows the *native discrimination* of our base learner (in blue) and our speech-biased learner (in orange). In all conditions, learners outperformed the baseline (the dashed black line), indicating performance improvements with language exposure over generic auditory processing. That said, performance was not equivalent throughout.

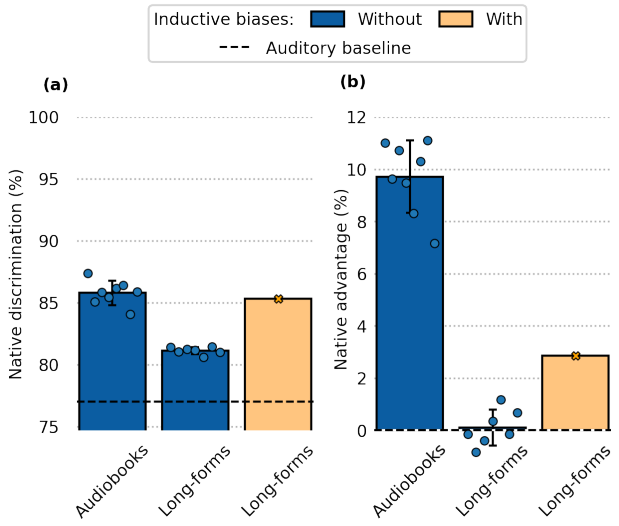

Figure 4: Results of Exp. 1: The impact of inductive biases. (a) Native discrimination and (b) native advantage obtained by our base learner (trained without inductive biases, in blue) and our speech-biased learner (trained with inductive biases, in orange). Both learners are trained on 128 hours of audio in each of the two target languages (American English and Metropolitan French) as a function of data type (audiobooks, long-forms). The dashed black line corresponds to the score obtained by a spectral representation typically used for speech recognition (39 mel-frequency cepstral coefficients). Each dot represents an independent replication on a separate training set of 128 hours. The number of independent replications, i.e., number of points, for each condition depends on the amount of data available.

A comparison between the two blue bars in this panel indicates that our base learner is better at discriminating sounds of its native language when trained on highly curated audio from audiobooks (result that conceptually replicates Schatz et al., 2021) than when trained on ecological audio from long-forms, with a drop in performance of 5%.

Interestingly, our speech-biased learner (in orange), endowed with inductive biases, performs much better in the same condition, reaching the same level of native discrimination as the base learner trained on curated data. These results suggest that inductive biases help overcome difficult learning conditions found in long-forms.

Panel (b) of Fig. 4 shows the *native advantage* obtained by both types of learners. First, let us consider our base learner. When trained on audiobooks, our base learner is 9.7% better at discriminating native sounds than the same non-native learner (first blue bar; a conceptual replication of Schatz et al., 2021). That is, the American English model has learned to discriminate American English sounds better than the Metropolitan French model (and vice-versa). However, our base learner trained on ecological long-forms fails to show any native advantage (second blue bar). In other words, when exposed to realistic data, the American English learner discriminates American English sounds as well as the Metropolitan French learner (and vice versa). Therefore, our base learner trained on long-forms has failed to develop a language-specific perceptual space.

Turning to our speech-biased learner in orange, we observe that the inductive biases allow the learner to exhibit a native advantage when faced with long-forms. However, unlike native discrimination results, performance is not matched to that with curated data, which suggests that native advantage is relatively more fragile to noisy environments than native discrimination.

3. Experiment 2: the effect of data quantity

Experiment 1 suggested that inductive biases helped overcome difficulties found in long-forms; in particular, inductive biases allowed our learner trained on 128 hours of audio to develop a language-specific perceptual space. Before concluding that inductive biases are truly *necessary* for our learner to develop a language-specific perceptual space, we must ask: could difficulties found in long-forms disappear when learning on *more* audio? After all, in machine learning, one can often replace specially engineered mechanisms with generic ones fed with enough data.[3].

3.1. Methods

In Experiment 1, learners were provided with 128 hours of audio. In the present Experiment, learners were provided between 8 and 1,024 hours, depending on the condition. We generated mutually exclusive training sets of varying lengths. Given limitations in data availability, we could provide our base learner up to 1,024h from audiobooks or 512h from long-forms (containing both speech and non-speech). As for our speech-biased learner, it receives up to 1,024h from audiobooks, or up to 128h from long-forms (containing exclusively speech). We vary the quantity of data by splitting larger training sets into two mutually exclusive subsets. Hence, our base learner exposed to audiobooks is trained separately on 2 training sets of duration 512h, 4 of duration 256h, etc. Our speech-biased learner exposed to long-forms is trained separately on 2 training sets of duration 64h, 4 of duration 32h, etc (see Appendix A for more details).

Apart from the training data quantity, we employ the same auditory baseline, learners, and ABX discrimination test as in Experiment 1.

3.2. Results

Let us first focus on the top panels (a) and (b) of Fig. 5 showing the native discrimination and native advantage obtained by learners trained on audiobooks. Panel (a) shows that the native discrimination obtained by learners supplemented with inductive biases plateaus at 89.4% after 128 hours of data, resulting in native learners that are 16.6% better at discriminating native sounds than non-native learners (solid orange lines of panels (a) and (b)). Without inductive biases (blue curve), the base learner exhibits significantly worse performance, both in terms of native discrimination

[3]The Bitter Lesson of Rich Sutton: http://incompleteideas.net/IncIdeas/BitterLesson.html

and native advantage. However, on the 1024-hour-long training set, both types of learners eventually catch up. This can be explained by the highly curated nature of the data our learners are trained on, as we will see in subsequent Experiments.

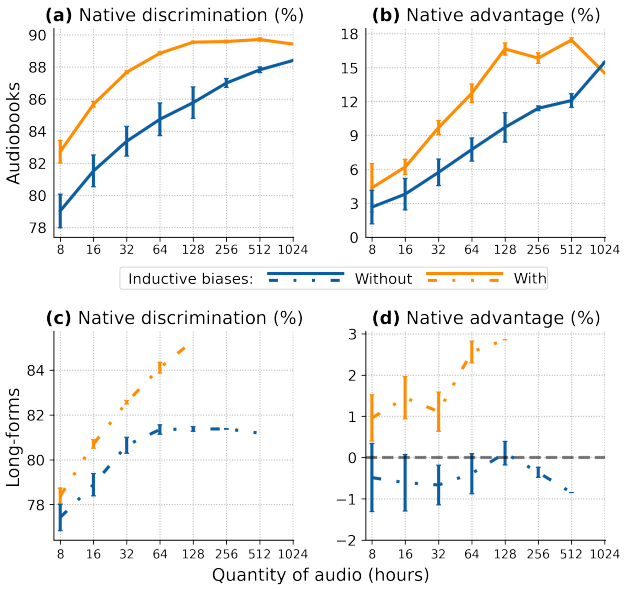

Figure 5: Results of Exp. 2: Effect of the quantity of data. Comparison of our base learner (trained without inductive biases, in blue) and our speech-biased learner (trained with inductive biases, in orange). Top panels (a) and (b) show the native discrimination and the native advantage obtained by learners trained on audiobooks. Bottom panels (c) and (d) correspond to learners trained on long-forms. Each learner is trained on either American English or Metropolitan French. Error bars represent standard errors computed across mutually exclusive training sets whose number depends on the quantity of data.

The bottom panels (c) and (d) illustrate radically different trends when learners are exposed to ecological long-forms. Starting with the base learner (blue dot-dot-dashed lines), its native discrimination plateaus at 81.3% after 64 hours of data, with no benefit from additional exposure; and its native advantage remains relatively flat and close to 0% (panel (d)). In contrast, our speech-biased learner exhibits a higher discrimination performance and develops a positive native advantage with increasing exposure. The contrast across learners is even more striking when we consider quantity of speech, rather than quantity of audio. Indeed, the base learner trained on 512 hours of long-forms will receive approximately 100 hours of speech (Table 1). That said, the base learner trained on 512 hours of long-forms does not even reach performance comparable to that of the speech-biased learner trained on 32 hours of long-forms. Thus, quantity of speech alone is not sufficient to explain the difference in behavior we observe between our learners trained with or without inductive biases. Instead, it suggests that the inductive biases themselves are responsible for the radically different behavior we observe.

3.3. Discussion

This Experiment shows that it does not take a great deal of highly curated audiobooks for our base learner to develop a native advantage. However, on long-forms, our base learner plateaus after 64 hours of audio and fails to develop a language-specific perceptual space, unlike what we observe with our speech-biased learner. Interestingly, a systematic review suggests infants accumulate between 60 and 1,000h of infant-directed speech in the first year of life (Cristia, 2019), and there is about 128h of speech in our 1,024h-long long-form dataset. Although the present manipulation covers only the low end of this range, the plateau observed by 64h of audio (roughly 12h of speech) in the long-form condition suggests that further increasing the training set size would not result in the emergence of perceptual attunement. Thus, from this work, it does appear to be the case that, when faced with *ecological long-forms*, inductive biases are *necessary* for our learner to develop a language-specific perceptual space and to reproduce perceptual attunement.

One may wonder if a more powerful statistical learning algorithm with no inductive biases would succeed in reproducing perceptual attunement from child-centered long-form recordings – perhaps with even more data. In our view, a more powerful pure statistical learning algorithm may pick up on more of the structure of noise and distortions, leading to a more general and non-language-specific perceptual space. Nonetheless, we may be proven wrong by further work.

4. Experiment 3: ablation study

Experiments 1 and 2 strongly suggest that our three inductive biases are crucial for perceptual attunement to emerge from long-form recordings. But what is their individual contribution? One way to assess this is via *an ablation study*, i.e., one in which each component is turned off. This is what we do in the present Experiment.

4.1. Methods

In addition to using the same base and speech-biased learner as in Experiment 1, we also create other learners in which each of the inductive biases (Voice Activity Detection, Pseudo Speaker Separation, Data Augmentation) can be turned on or off. Note that Pseudo Speaker Separation cannot be activated without Voice Activity Detection also being activated. This results in six learners. We do not include a baseline since it is the same as in Experiment 1. The same ABX discrimination test is also used as in Experiment 1.

4.2. Results and Discussion

Inspection of Table 2 suggests that for the audiobooks, the Voice Activity Detection and the Pseudo Speaker Separation mechanisms have little effect on the performance obtained by our base learner (lines 1, 2, and 5). This is most certainly due to the high proportion of speech and the long speaker turns found in audiobooks (see Table 1). It appears that Data Augmentation mechanism is the most

Corpora	#	Inductive biases	Native discrimination (%)	Native advantage (%)
Audiobooks	1	none	87.37	11.01
	2	VAD	87.73	12.95
	3	DA	88.98	16.00
	4	VAD + DA	89.37	14.87
	5	VAD + PSS	87.37	10.70
	6	VAD + PSS + DA	89.51	16.53
Long-forms	7	none	81.41	-0.15
	8	VAD	84.64	1.32
	9	DA	83.96	0.63
	10	VAD + DA	84.64	1.50
	11	VAD + PSS	84.53	1.32
	12	VAD + PSS + DA	85.09	2.45

Table 2
Results of Exp. 3: Ablation study. All models are trained on 128 hours of audio taken from audiobooks (top panel) or long-forms (bottom panel). VAD stands for Voice Activity Detection. PSS stands for Pseudo Speaker Separation. DA stands for Data Augmentation. The PSS mechanism can not be activated without the VAD mechanism. Lines numbered 1 and 7, where none of the inductive bias is activated, correspond to the base learner. Lines numbered 6 and 12, where all of the inductive biases are activated, correspond to the speech-biased learner.

impactful mechanism, with a drop of up to 5% absolute native advantage when it is not included (lines 1 and 3). However, note that results of Fig. 5 suggest that the Data Augmentation mechanism has a beneficial impact only for low data quantities.

The results obtained by the learner trained on long-forms suggest that, among all three inductive biases, Voice Activity Detection plays the most crucial role, with a drop of up to 3.2% absolute native discrimination when it is not included (lines 7 and 8). The learner trained with both the Voice Activity Detection and the Pseudo Speaker Speaker Separation mechanisms reaches a similar performance than the learner trained with the Voice Activity Detection mechanism alone (lines 8 and 11). However, the best native discrimination and the best native advantage are obtained by our speech-biased learner, i.e., trained with all three inductive biases, which suggests that all three biases contribute to the emergence of a native advantage (line 12).

5. Experiment 4: the effect of the speech/non-speech ratio

In this Experiment, we focus on the largest difference between long-forms and audiobooks: the proportion of speech as opposed to non-speech in the recordings (20%-80% versus 100%-0%, respectively; Table 1). This difference is the primary reason why we introduced a filtering mechanism (the voice activity detection, VAD, Fig. 3) to shield our learner from the large quantity of non-speech present in long-forms. In the current Experiment, we ask: What is the real impact of this particular inductive bias? We study this by providing our speech-biased learner with input containing varying amounts of non-speech, as if their VAD mechanism was dysfunctional.

5.1. Methods

To better control for audio quality and quantity, we employ the same 128h audiobooks training set as in Experiment 1. We simulate a dysfunctional VAD mechanism that allows non-speech to leak in by adding to the training set varying amounts of non-speech sections extracted from long-forms. Thus, the learner always receives 128h of speech, but the total quantity of input audio data varies, so that the proportion that is speech varies across 4 conditions, from 100% speech to an eighth of speech (i.e., 128h of speech in 1024h of audio). We employ the same speech-biased learner, auditory baseline and machine ABX discrimination test as in Experiment 1.

5.2. Results and discussion

Focusing first on panel (a) of Fig. 6, results indicate that the lower the proportion of speech, the lower the native discrimination. Regarding native discrimination, the speech-biased learner trained exclusively on speech (128 hours) scores 4.2% higher than the same learner trained on only an eighth of speech, with the rest being non-speech (corresponding to 128 hours of speech and 896 hours of non-speech).

Importantly, results show that our speech-biased learner is robust to a certain amount of non-speech, as shown by the native discrimination of 88.4% developed by learners trained on half of speech, half of non-speech (128 hours of each). More surprisingly, the learner trained on an eighth of speech still exhibits a higher native discrimination than the auditory baseline (dashed black line). This shows that despite the important quantity of non-speech segments in the training set, the learner still develops a perceptual space in which sounds are accurately discriminated.

Turning to panel (b), the effect of the proportion of non-speech segments over speech segments is even greater on

Figure 6: Results of Exp. 4: Simulating a dysfunctional voice activity detection mechanism. Native discrimination (panel (a)) and native advantage (panel (b)) obtained by our learner supplemented with inductive biases, depending on the proportion of speech (in green) and non-speech (in pink) in the training set. Learners are trained on 128 hours of English and French audiobooks (first column) to which we added a fixed quantity of non-speech segments drawn from long-forms: 128 hours (second column), 384 hours (third column), and 896 hours (fourth column). The dashed black line corresponds to the score obtained by a spectral representation typically used for speech recognition (39 mel-frequency cepstral coefficients).

the native advantage. Indeed, learners trained exclusively on speech exhibit a native advantage of 16.5%, while this goes down to 2.5% for learners trained on an eighth of speech.

Together, these results suggest that the proportion of non-speech segments in the input harms both the native discrimination and the native advantage. Although it is difficult to provide precise evidence on this, we hypothesize that the perceptual space the learner develops during training becomes less relevant to discriminate between speech sounds as the subspace allocated to discriminate between non-speech sounds grows.

6. Experiment 5: sensitivity to noise and reverberation

In this Experiment, we turn to another important difference between audiobooks and long-forms: the *quality* of the speech signal itself. More specifically, how do the different degradations that occur in ecological data affect the perceptual space developed by our learner? To answer this question, we trained our learner on audiobooks that we corrupted by manipulating two major factors: additive background noise and reverberation, which are prevalent in long-forms. In this Experiment, we examine performance obtained by our speech-biased learner because: 1) simulating non-speech found in naturalistic data (required by the base learner) from audiobooks is not trivial as audiobooks contain almost exclusively speech; 2) if any detrimental effect is observed on the performance obtained by our speech-biased learner, then it is very likely that an even bigger effect will be observed on our base learner, as the latter does not include mechanisms aimed at overcoming difficulties found in naturalistic data.

6.1. Methods

To study the effects of audio degradations, we created a new dataset by contaminating the audiobook data used in Experiment 2 with additive noise and reverberation. For the additive noise, we extracted various domestic noises from long-forms, as detailed in Appendix D. For the reverberation, we used two sets of impulse responses, the MIT Acoustical Reverberation set (Traer and McDermott, 2016) and the EchoThief impulse response library (Warren, 2013), whose combination resulted in 385 impulse responses acquired in a wide variety of places. We then applied these two sources of contamination as follows.

We started with speech segments extracted from the English and French audiobooks. For each speech segment, we built a noise segment of the same duration by crossfading successive noise sequences with a crossfade duration of 50 ms. To simulate a realistic acoustical scene, we randomly chose two impulse responses: the first was convoluted with the speech segment, and the second with the noise segment. It is only when the speech and the noise segments have been convoluted that they are normalized and added together. This effect makes the speech segment seem to exist in one location and the noise segment in another. Using this pipeline, we examine the audiobook corpora corrupted with: 1) reverberation only; 2) additive noise only for a signal-to-noise ratio uniformly sampled between 0 and 15 dB; and 3) both reverberation and additive noise (which we called simulated long-forms condition, and for which a comparative analysis in terms of SNR and C_{50} can be found in Appendix E). We employ the same speech-biased learner and the same machine ABX discrimination test as in previous experiments.

6.2. Results

In general terms, Fig. 7 shows that challenging acoustic conditions negatively impact the learning outcomes of our speech-biased learner, yielding lower native discrimination and lower native advantage for all conditions compared to the clean one.

Applying additive noise to the training set causes only a slight decrease in the performance obtained by our speech-biased learner. The drop in performance is more important when reverberation is applied. And the worst performance is observed when both additive noise and reverberation are applied to the training set.

May this negative effect be overcome by adding more data? Inspection of the curves as a function of data quantity suggests that our learner's native discrimination and native advantage both plateau after 128 hours of speech, regardless of the acoustic condition, and thus more data is unlikely to solve the problem. Therefore, it appears that additive noise

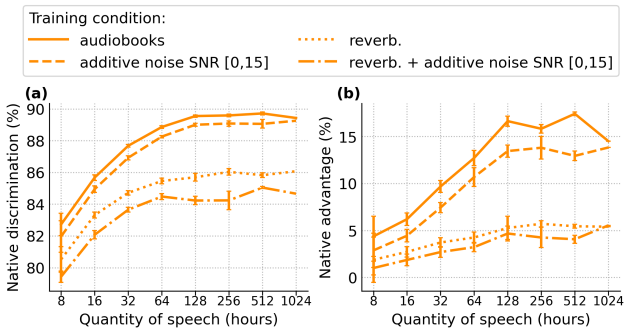

Figure 7: Results of Exp. 5: Sensitivity to noise and reverberation. Panel (a) and (b) indicate the native discrimination and the native advantage, respectively, obtained by our speech-biased learner (trained with inductive biases) as a function of both the quantity of speech, and varying acoustic conditions. Learners are trained on clean read-speech segments extracted from audioboks (solid line). These clean read-speech segments are corrupted with 1) additive noise for a signal-to-noise ratio (SNR) randomly chosen between 0 and 15 dB (dashed line); 2) reverberation (dotted line); 3) both reverberation and additive noise (dash-dotted line). The latter corresponds to the simulated long-form condition. Error bars represent standard errors computed across mutually exclusive training sets whose number depends on the quantity of data.

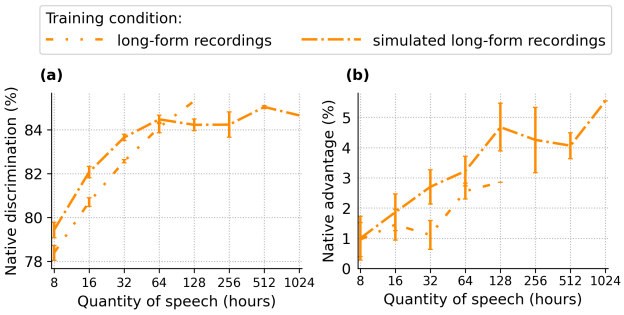

Figure 8: Results of Exp. 6: Comparing real against simulated long-forms Comparison of native discrimination (panel (a)) and native advantage (panel (b)) of learners trained with inductive biases on long-forms (dot-dot-dashed line) or simulated long-forms (dash-dotted line). Simulated long-forms are generated using read-speech retrieved from audiobooks whose speech segments are corrupted with additive noise and reverberation. Error bars represent standard errors computed across mutually exclusive training sets whose number depends on the quantity of data.

and reverberation strongly degrade representations learned during training, leading to a less language-specific perceptual space in situations where signal and noise compete with each other than in cleaner conditions.

6.3. Discussion

Results in this Experiment replicate those in Experiment 2: An increase in the quantity of data yields a higher native discrimination and a higher native advantage, regardless of the acoustic condition, although performance eventually plateaus. The present Experiment adds an important caveat: the perceptual space developed by our learner becomes less language-specific (i.e., lower native advantage) when it is exposed to more challenging acoustic conditions, with reverberation having a massive negative effect, and additive noise a substantially smaller one. Thus, our speech-biased learner is sensitive to the lower acoustic quality found in ecological recordings, and the ensuing negative effect cannot be solved by increasing the quantity of data.

7. Experiment 6: using audiobooks to simulate long-form recordings

In Experiments 4-5, we covered the impact of some of the most important differences between audiobooks and long-forms, namely the proportion of speech, and their quality in terms of presence of additive noise and reverberation. In this final Experiment 6, we ask a slightly different question connected to the importance of understanding the differences between audiobooks and long-forms for the purposes of modeling. Audiobooks are easier to collect and share than long-forms, and simulated long-forms could be a useful tool to calibrate the amount of data collection needed to obtain enough statistical power in a computational modeling study. So, can we use *simulated long-forms* as a proxy for real long-forms? To answer this, we carry out a direct comparison between our learner trained with inductive biases on long-forms, and the same learner trained on simulated long-forms.

7.1. Methods

We focus on the simulated long-forms that are most challenging, those with reverberation and additive noise for a random SNR between 0 and 15 dB (dash-dotted orange line of Fig. 7). As in Experiments 2 and 5, we vary the quantity of data. In this Experiment, we employ only the speech-biased learner and the same ABX discrimination test as in Experiment 1.

7.2. Results and discussion

Fig. 8 shows the native discrimination and native advantage obtained by our speech-biased learners when trained on varying quantity of speech extracted from long-forms, or simulated long-forms. Learners have essentially identical native discrimination when trained on long-forms or simulated long-forms (panel (a)). Overall, the close match between learners trained on real and simulated data indicates that noise and reverberation capture the essential components of ecological recordings and can be used to simulate the performance obtained by our learner.

One could expect a similar match in terms of native advantage. Surprisingly, however, learners trained on long-forms exhibit a substantially lower native advantage than learners trained on simulated long-forms (panel (b)). This mismatch in terms of native advantage must then be due to other characteristics of ecological speech that are not captured in our simulated long-forms. Beyond acoustic characteristics, audiobooks likely have lower speech rate, greater hyperarticulation, and are primarily composed of complex

Experiment 1	Our base learner trained on audiobooks reproduces perceptual attunement in English and French The same learner fails to do so when trained on long-forms Inductive biases allow for full recovery of native discrimination, and partial recovery of native advantage
Experiment 2	Our base learner never reproduces perceptual attunement on long-forms, even with more data as performance plateaus
Experiment 3	Inductive biases work in concert to achieve perceptual attunement, showing a supra-additive effect The higher performance gain is brought by the voice activity detection mechanism
Experiment 4	When non-speech is added, native discrimination diminishes and native advantage vanishes
Experiment 5	When background noise and reverberation are applied, native discrimination diminishes and native advantage vanishes, despite the data augmentation mechanism, and regardless of data quantity
Experiment 6	Long-forms and simulated long-forms yield similar native discrimination, but different native advantage, i.e., there remain other differences between long-forms and simulated long-forms (e.g., hyperarticulation found in read speech)

Table 3
Summary of key results

utterances with diverse lexical items, features that in the language acquisition literature have been argued to be beneficial for infant learning (Rowe, 2012; Weisleder and Fernald, 2013; Anderson, Graham, Prime, Jenkins and Madigan, 2021). In contrast, ecological data contains infant-directed speech which has been claimed to facilitate learning (Eaves Jr, Feldman, Griffiths and Shafto (2016); Adriaans and Swingley (2017), but see McMurray, Kovack-Lesh, Goodwin and McEchron (2013) for a different view). Further investigation is needed to verify the contributions of these additional characteristics on artificial learners.

8. General discussion

We carried out six experiments investigating how the design of the learner and the type of data it receives affect perceptual attunement. These are summarized in 3. We discuss below limitations of our study and ways in which future modeling work can improve upon it. Before doing so, we point out three predictions for human infant data drawn from the present work.

Predictions for human infant studies and future experimental work. The first prediction is that quantity of input is beneficial to infants' ability to discriminate native-language phonetic contrasts (Fig. 5 and 7, and Li, Schatz, Matusevych, Goldwater and Feldman (2020); Schatz et al. (2021) for relevant previous results). Our results suggest that the beneficial effect of input quantity still holds in challenging acoustic environments, although with different slopes (Fig. 7), for both types of learners (Fig. 5). This prediction is largely compatible with current views of language acquisition and empirical evidence correlating input quantity with vocabulary and standardized language tests (Weisleder and Fernald (2013); Gilkerson, Richards, Warren, Oller, Russo and Vohr (2018), but see Sperry, Sperry and Miller (2019) for a different view). The evidence for input quantity effects on native advantage specifically is scarce and equivocal (Cristia, 2020), but at least one study has found compatible results (Marklund, Schwarz and Lacerda, 2019).

The second prediction is that both acoustic and speech quality positively impact infants' ability to discriminate native-language phonetic contrasts. We have shown, in Fig. 7, that poor signal quality of speech segments yields lower discrimination accuracy and poorer language attunement. We found a similar result when speech-biased learners were partially trained on non-speech segments (Fig. 6). To the best of our knowledge, input quality thus defined has not been linked to infants' discrimination abilities of phonetic contrasts yet. That said, it is possible that infants also have access to additional input filters, leading them to preferentially learn on clean and well-formed speech sounds, rather than noisy segments (non-speech, far-field speech, etc.). Indeed, a prior computational study found that prosodic exaggeration (higher pitch, greater duration of some sounds, etc.) found in infant-directed speech may facilitate vowel learning by enhancing separability between vowel categories (Adriaans and Swingley, 2017).

A third prediction is that inductive biases are *necessary* to drive language attunement in the wild (panels (c) and (d) of Fig. 5). We showed that experiencing clean read speech was sufficient for either learner, trained with or without inductive biases, to become attuned to its language of exposure. However, the base learner fails to show any perceptual attunement when exposed to child-centered long-forms. It is only when supplemented with inductive biases that the learner exhibits a significant level of attunement (panel (d) of Fig. 5). We are not certain of how precisely this prediction could be verified in human infants, but we hope developmental scientists will reflect on this.

We also recommend additional studies to check whether human infants have filters like the ones with which we endow our biased learner. Concerning the voice activity detection mechanism, there is evidence suggesting that humans are born with a preference for listening to speech (Cooper and Aslin, 1990; Vouloumanos and Werker, 2007), demonstrating, therefore, an early ability to discriminate between speech and non-speech segments. As for the pseudo speaker

separation mechanism, evidence suggests that newborns can discriminate their mother's voice (Mehler et al., 1978; Decasper and Fifer, 1980) but also voices of unfamiliar speakers (Decasper and Prescott, 1984; Floccia et al., 2000). Infants also progressively learn representations that are invariant with respect to a change in speaker (Seidl et al., 2014; Bergmann et al., 2016; Choi and Shukla, 2021). However, to our knowledge, there is no experimental evidence documenting something similar to our data augmentation mechanism, and in particular robustness to reverberation in infants (see Beeston et al. (2014) for evidence in adults).

Further work is needed to shed light on how infants may come to have such biases. We suspect to a certain extent they are inherited from our evolutionary past, and to another extent they are learned. This is similar to theoretical proposals for face perception. Inspired by work on chicks (and other social vertebrates), Morton and Johnson (1991) proposed two mechanisms that work in concert to explain human infants' preference for faces from birth, together with strong changes over the course of development, CONSPEC and CONLERN. CONSPEC is an inductive bias that guides human newborns' preference for human face-like patterns, and CONLERN allows them to learn facial features of specific individuals. Extrapolating to speech, we suspect that a better understanding of how these (or other) speech biases come to be active in human infants will minimally require a concerted theoretical and empirical effort by modelers, developmental scientists, and ethologists.

Finally, we firmly believe that considering input data that do not reflect characteristics of children language environments can lead us to underestimate the complexity of a given learning problem – that is, we may believe a problem has been solved when it has only been solved on unrealistic data –, or overestimate the power of a proposed mechanism – that is, we may believe a mechanism is sufficient to explain the phenomenon that is being modeled when it is not –. This is not to say that simplification of the input data in computational modeling studies is an insightless enterprise. In fact, checking computational models against data, which may not necessarily reflect real-world properties can help us gain a better understanding of the phenomena being modeled. However, we argue here that a computational model of language acquisition should not only be evaluated based on its ability to reproduce infants' developmental trajectories, but also on its ability to learn from input that is as close as possible to the sensory signals infants actually experience.

Limitations and further modeling work. Our study is, to our knowledge, the first fully-implemented proof of a statistical learning account additionally including inductive biases for early phonetic learning. But we are certain it will not be the last one. Despite the benefits of the proposed inductive biases in improving native discrimination and enabling the emergence of perceptual attunement, our speech-biased learner trained on long-forms still did not achieve the same level of native advantage as the base learner trained on audiobooks (as shown in Figures 4 and 5). If we want to tackle the complexity inherent to long-forms, it is very likely that we should aim at finding what other inductive biases might guide phonetic learning in our algorithms.

Indeed, as shown in Experiment 5, our speech-biased learner is still greatly sensitive to additive background noise and reverberation that are ubiquitous in children learning environments. This background noise issue could be alleviated with an additional mechanism filtering out segments with a low speech-to-noise ratio. Such a filter remains plausible as evidence shows that 4.5-month-olds prefer listening to speech in a quiet than in a noisy environment, suggesting that they are aware of the presence of noise (Newman and Hussain, 2006). Similarly, overlapping speech, found in long-forms but not in audiobooks, may necessitate an additional separation mechanism. Indeed, overlapping speech represents a challenging source of data for phonetic learning as sounds produced by different speakers blend together. Such a source separation mechanism may take place in infants as evidence for the so-called 'cocktail party effect' (our ability to focus our auditory attention on a particular source while filtering out other competing sources) has been found at 7.5 months (Newman, 2005). Finally, long-forms contain both hyperarticulated infant-directed speech and spontaneous speech that may sometimes be hypoarticulated. It is possible that a mechanism forcing the algorithm to learn preferentially on better articulated speech segments may help improve both native discrimination and native advantage. These three types of biases could meaningfully adopt by and large the same approach we had here, including minimally audiobooks and long-forms as input, and utilizing an ablation approach to assess the contribution of individual components.

In this study, we examined if our learners could reproduce perceptual attunement such as measured by the native advantage. However, most research on perceptual attunement in humans only covers a few contrasts and it is unclear to which extent human infants or adults exhibit a native advantage on other contrasts that remain to be studied. In the absence of systematic data across contrasts and populations, we do not know the size of infants' native advantage and are limited to a quantitative comparison between infants and artificial language learners. However, further work could aim at comparing the effect size of the native advantage obtained by our learners with those of infants such as provided by meta-analytic studies (see methodology proposed in Blandón, Cristia and Räsänen, 2021).

Other aspects not addressed here will require substantially different modeling decisions. Beyond inductive biases that apply to the auditory stream, infants may use more than just audio to learn about the phonemes of their native language. Visual cues, such as lip movements, may complement the audio channel and help infants achieve better and possibly faster learning. Modeling research combining other modalities typically uses even more manicured input than speech-only work (with the exception of Alishahi, Chrupała, Cristià, Dupoux, Higy, Lavechin, Räsänen and Yu, 2021). Furthermore large scale naturalistic audio-visual long-form data does not currently exist. It would be necessary for these

data to be similarly ego-centered as our long-forms, since modeling development from a third- versus a first-person perspective implies distinct challenges. Recent discussions also highlight the potential role of supervisory signals McMurray (2022a), considering the overwhelming evidence that perception continues to develop beyond the first year. While promising, incorporating supervisory feedback in our learner would require substantially different modeling approaches than employed here, both in terms of input and of mechanisms.

9. Conclusion

Recent advances in machine learning algorithms that learn from raw speech, together with ongoing efforts to acquire large-scale datasets of infants' language environments, make building ecologically valid models of language acquisition feasible. Through six experiments, we argued that statistical learning leads to language-specific perception provided it is focused on clean speech signal, either as input or thanks to processing mechanisms; and that such biases are required given the uniquely difficult input data afforded to human infants. Evidently, our work does not constitute a proof of infeasibility. Perceptual attunement to sounds may be feasible with purely statistical algorithms that are yet to be discovered. However, we strongly believe that for researchers to discover such mechanisms, computational models of early language learning must address the problem in its full complexity, learning from the same data than that available to infants. In that regard, we believe that our work paves the way for a deeper understanding of mechanisms involved in early language acquisition.

Data and code availability

We used audiobooks – publicly available from LibriVox (Kearns, 2014) – and child-centered long-forms[4] to train our learners. Long-forms are available on HomeBank and via ChildProject repositories (Bergelson, 2017; Canault et al., 2016a; Cristia, 2021; Lavechin and Cristia, 2021). Access to long-forms requires ethical training and full approval by the principal investigator(s) in charge. Code to reproduce the experiments will be made available in a public GitHub repository upon publication.

Acknowledgements

We are grateful to DARCLE, LAAC, and CoML members for helpful discussion. All errors remain our own. A.C. gratefully acknowledges financial and institutional support from Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017); the J. S. Mc-Donnell Foundation (Understanding Human Cognition Scholar Award); and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). E.D., in his academic role (EHESS), acknowledges funding from Agence Nationale de la Recherche (ANR-19-P3IA-0001 PRAIRIE 3IA Institute), a grant from CIFAR (Learning in Machines and Brains), and the HPC resources from GENCI-IDRIS (Grant 2020-AD011011829). M.S. acknowledges PhD funding from Agence de l'Innovation de Défense.

A. Training sets

In this section, we first describe our data sets, and the differences between long-forms and audiobooks. Next, we describe how the ABX evaluation set has been built to assess phonetic discrimination of our models, and in particular, which phonemes have been retained for each of the two target languages.

For each language (English or French), and each audio source (audiobooks or long-forms), we built separate training sets by randomly splitting the whole set of audio segments into mutually exclusive training sets of 8 hours. 8-hours training sets were then merged two by two to build the 16-hours training sets. This procedure was repeated until a single training set covering the entirety of the dataset is obtained. Therefore, the number of subsets depends on the total amount of data available.

On audiobooks, 128, 64, 32, 16, 8, 4, 2 and 1 training sets were built whose duration was 8, 16, 32, 64, 128, 256, 512 and 1,024 hours respectively, which contain both speech and non-speech segments in the case of our base learner; and only speech as detected by the VAD pretrained model in the case of our speech-biased learner.

For long-forms, the procedure resulted in 16, 8, 4, 2 and 1 training sets whose duration was 8, 16, 32, 64 and 128 hours respectively containing only speech as detected by the pretrained VAD model. Similarly 32, 16, 8, 4, 2, 1 training sets were built whose duration was 8, 16, 32, 64, 128, 256, and 512 hours respectively (containing both speech and non-speech). In Experiment 1, only the 128 h sets were used. The other sets were created for use in Experiments 2 and 4.

B. Learners: implementation details

B.1. Contrastive predictive coding

As proposed in (Kharitonov et al., 2021), the encoder g_{enc} consists of a 5-layer convolutional neural network with kernel sizes $[10, 8, 4, 4, 4]$ and strides $[5, 4, 2, 2, 2]$ that returns a 256-dimensional vector every 10 milliseconds. The auto-regressive model g_{ar} is a 2-layer long short-term memory network of dimension 256. The model is asked to predict up to $K = 12$ time steps in the future (which is equivalent to 120 ms). The predictor g_{pred} is a single multi-head transformer layer with $K = 12$ heads, each predicting at time step $k \in \{1, ..., 12\}$. Negative samples are drawn from sequences that are temporally close to the sequence the positive sample are drawn from. More precisely, creating a batch consists of selecting 64 successive sequences in the case of the domain-general learner (or 64 successive

[4] Child-centered long-forms were accessed and stored only by authors outside of Meta AI Research using the GENCI-IDRIS (Grant 2020-AD011011829) infrastructure.

sequences that have been pronounced by the same speaker for the domain-specific learner). For a current sequence seq_i, negative samples are taken from all other sequences seq_j, with $j \neq i$. All models have been trained on 8 GPUs with batches of 64 sequences, and each sequence has a duration of 1.28 seconds. All models are trained until convergence, and the best epoch is selected according to validation loss (5% of the original training set).

B.2. Inductive biases

Voice Activity Detection. Speech segments were identified using a voice type classification model (Lavechin et al., 2020) that has been specifically trained on a multilingual corpus of long-forms. This pretrained model partitions the audio stream into segments belonging to one or more of the following classes: 1) KCHI, for vocalizations produced by the child wearing the recording device; 2) OCH, for vocalizations produced by any other children; 3) FEM, for female adult speech; 4) MAL, for male adult speech and 5) SPEECH, for when there is speech. When none of the 5 classes were activated, the segment was considered to be non-speech. Only segments belonging to the adult (MAL or FEM categories) were selected; thus, segments with speech by children or containing only non-speech were filtered out.

For long-forms, a total of about 200 hours of speech was extracted this way. Approximately 1000 hours of speech were extracted from the audiobooks.

Pseudo Speaker Separation. For long-forms, the category information (MAL, FEM) was combined with the long-form unique identifier to approximate speaker identity. We would have liked to use a speaker diarization system but no such system performs reasonably well on long-forms (García, Villalba, Bredin, Du, Castán, Cristia, Bullock, Guo, Okabe, Nidadavolu, Kataria, Chen, Galmant, Lavechin, Sun, Gill, Ben-Yair, Abdoli, Wang, Bouaziz, Titeux, Dupoux, Lee and Dehak, 2020). During the CPC training procedure, this results in positive and negative examples that have been pronounced by: 1) the same category (MAL or FEM); and 2) within the same long-form. Note that the second constraint is further reinforced by the Temporal Proximity Sampling mechanism which ensures positive and negative examples are temporally close to one another. For audiobooks, and contrary to long-forms, the speaker identity was retrieved from the metadata, because it is constant throughout each recording. We used this information as-is in the Pseudo Speaker Separation mechanism. Consequently, positive and negative examples during the CPC training procedure were drawn exactly within speakers (and not within pseudo speakers as in the case of long-forms).

Data Augmentation. A distinction needs to be made about the way we 1) simulated long-forms by adding additive noise and reverberation on clean-read speech taken from audiobooks (section 6.1); and 2) apply data augmentation methods during training. When simulating long-forms, the training set is contaminated and then stored on disk. In other words, a clean segment of speech will always be associated to the same transformation (additive noise, and reverberation using impulse responses), which has for effect to make the learning process more challenging. In contrast, in the Data Augmentation mechanism, the modifications applied to the audio sequences are performed in an online fashion. This has for effect that a given audio sequence can be associated to multiple different transformations (chosen randomly), therefore forcing the model to extract features that are invariant with respect to the applied transformations.

Furthermore, following (Kharitonov et al., 2021), only past sequences were augmented with pitch modification and artificial reverberation. Modifications applied to the past sequences are: pitch shifting by a random integer between 300 and -300 (measured by 1/100 of a tone), and artificial reverberation using a random room size.

C. The machine ABX discrimination test

The logic of the ABX is represented in Figure 9. To build the ABX evaluation set, 10 hours of read-speech in American English and Metropolitan French were downloaded from Common Voice (Ardila et al., 2020). Segments were pronounced by 24 speakers whose gender was balanced in both languages. The phone-level alignment was obtained by aligning the audio stream with its transcript using Kaldi recipes (Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlicek, Qian, Schwarz et al., 2011). The ensuing phonetic inventory in International Phonetic Alphabet (IPA) standard for both languages is shown in Table 4.

D. A database of noises to use as additive noise

We extracted all non-speech segments from SEEDLingS' long-forms. For each non-speech segment, the pretrained voice type classification model (Lavechin et al., 2020) returns a probability that can be interpreted as how confident the model is that the segment is indeed non-speech (computed as 1-p_s where p_s is the probability of the segment being speech). All non-speech segments whose probability of being non-speech was lower than the median non-speech probability were discarded. We did so to minimize the amount of segments that have been wrongly identified as being non-speech when in fact they do contain speech. Finally, 1,024 hours of non-speech segments were kept among those having the highest energy to ensure that segments containing only white noise (which are closer to silence) were discarded. Based on manual inspection of a small subset, selected non-speech segments contained various noises including running water, animal sounds, vacuum cleaner, heartbeats of the key child, etc.

E. Noise and reverberation in long-forms and simulated long-forms

We used the same pretrained model Lavechin et al. (2022) as in Fig. 1 to study the distributions of the SNR and the C_{50} of utterances extracted from English long-forms or simulated long-forms (i.e. audiobooks contaminated with noise and reverberation). While Fig. 1 revealed an important

mismatch in terms of SNR and C_{50} between long-forms and audiobooks, we observe in Fig. 10 that this difference has been reduced between long-forms and simulated long-forms.

On average, utterances extracted from long-forms have an SNR of 10.4 dB while those from simulated long-forms have an SNR of 10.6 dB. The mismatch in terms of C_{50} has also been reduced with an average of 29.4 dB for long-form utterances and 15 dB for simulated long-form utterances (as opposed to 54 dB for audiobooks). The lower C_{50} obtained on simulated long-forms indicate those contain more reverberant environments than real long-forms. Similar results were found on French.

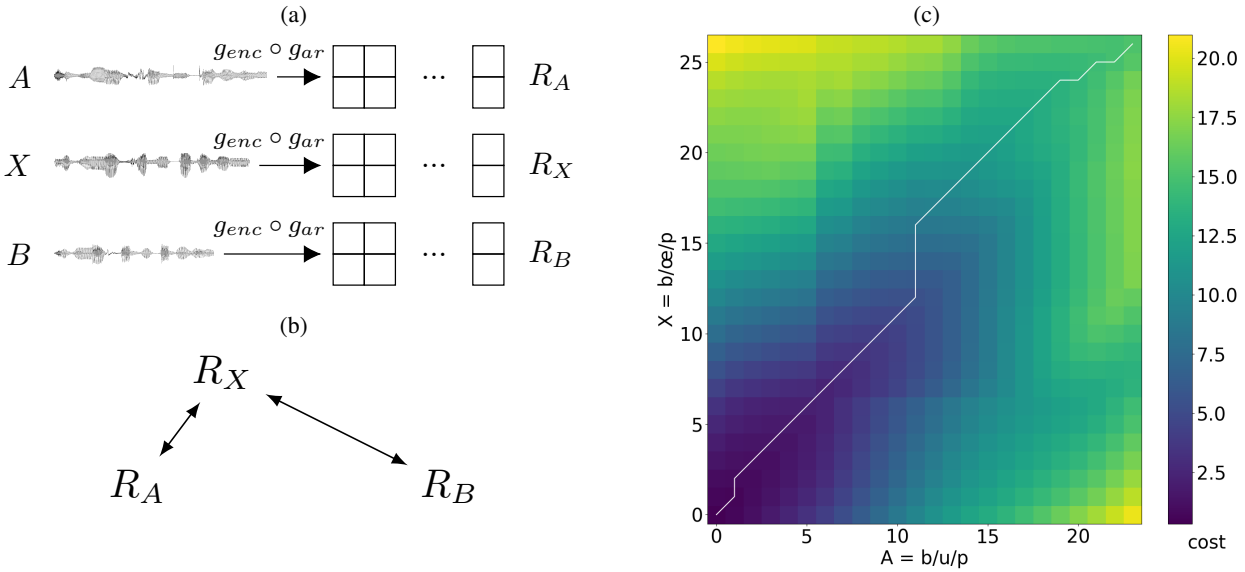

Figure 9: (The machine ABX discrimination test) a) The machine is given 3 stimuli A, B and X whose representations R_A, R_B and R_X are extracted by the composition $g_{enc} \circ g_{ar}$ of the encoder and the autoregressive model of CPC; b) Distances between R_A and R_X and R_B and R_X is computed with dynamic time warping (DTW). In the example, DTW cost matrix and shortest path (in white) is shown for A = [bup] and X = [bœp]; c) If A and X are different occurrences of the same triphone (e.g. [bup]), and B an other triphone differing only in its center phone (e.g. [bœp]), the machine is right if R_A is closer to R_X than R_B, wrong otherwise.

Manner of articulation	Metropolitan French	American English
	Consonants	
Stops:	p,b,t,d,k,g	p,b,t,d,k,g
Nasals:	m,n,ɲ	m,n,ŋ
Fricatives:	f,v,s,z,ʃ,ʒ,ʁ	f,v,θ,ð,s,z,ʃ,ʒ,h
Approximants:	j,w,l	j,ɹ,w,l
Affricates:	ʤ	ʤ,ʧ
	Vowels	
Orals:	i,y,e,ø,œ,ɛ,a,ə,ɔ,o,u	i,ɪ,ɛ,æ,ɚ,ʌ,e,u,ʊ,ɔ,ɑ
Nasals:	ɑ̃, ɛ̃, œ̃,ɔ̃	
Diphtongs:		aɪ,ɔɪ,aʊ,eɪ,oʊ

Table 4
Evaluated phonetic inventory in Metropolitan French and American English in International Phonetic Alphabet (IPA) standard.

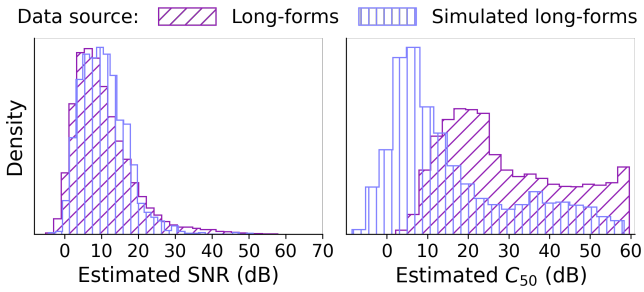

Figure 10: (The quantity of noise and reverberation in long-forms and simulated long-forms) Speech-to-Noise Ratio (SNR) and C_{50} distributions on 16 hours of speech utterances extracted from long-forms (slanting hatches) or simulated long-forms (vertical hatches). Both measures are automatically extracted using the pretrained model proposed in Lavechin et al. (2022)

References

Adriaans, F., Swingley, D., 2017. Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. The Journal of the Acoustical Society of America 141 5, 3070.

Alishahi, A., Chrupała, G., Cristià, A., Dupoux, E., Higy, B., Lavechin, M., Räsänen, O., Yu, C., 2021. ZR-2021VG: Zero-resource speech challenge, visually-grounded language modelling track. arXiv preprint arXiv:2107.06546 .

Anderson, N.J., Graham, S.A., Prime, H., Jenkins, J.M., Madigan, S., 2021. Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. Child Development 92, 484–501.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common voice: A massively-multilingual speech corpus, in: Language Resources and Evaluation Conference (LREC).

Beeston, A.V., Brown, G.J., Watkins, A.J., 2014. Perceptual compensation for the effects of reverberation on consonant identification: evidence from studies with monaural stimuli. The Journal of the Acoustical Society of America 136 6, 3072.

Bergelson, E., 2017. SEEDLingS HomeBank corpus. https://homebank.talkbank.org/access/Password/Bergelson.html. doi:10.21415/T5PK6D.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A.S., Amatuni, A., 2019. What do North American babies hear? A large-scale cross-corpus analysis. Developmental science 22 1, e12724.

Bergmann, C., Cristia, A., Dupoux, E., 2016. Discriminability of sound contrasts in the face of speaker variation quantified, in: CogSci.

Bion, R.A., Miyazawa, K., Kikuchi, H., Mazuka, R., 2013. Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. PloS one 8, e51594.

Blandón, M.A.C., Cristia, A., Räsänen, O., 2021. Evaluation of computational models of infant language development against robust empirical data from meta-analyses: what, why, and how? .

Bregman, A.S., 1994. Auditory scene analysis: The perceptual organization of sound. MIT press.

Canault, M., Normand, M.T., Foudil, S., Loundon, N., Thai-Van, H., 2016a. LYON HomeBank corpus. https://homebank.talkbank.org/access/Password/Lyon.html. doi:21415/T58P6Q.

Canault, M., Normand, M.T.L., Foudil, S., Loundon, N., Thai-Van, H., 2016b. Reliability of the Language ENvironment Analysis system (LENA™) in European French. Behavior Research Methods 48, 1109–1124.

Choi, M., Shukla, M., 2021. A new proposal for phoneme acquisition: Computing speaker-specific distribution. Brain Sciences 11, 177.

Coen, M., 2006. Self-supervised acquisition of vowels in American English, in: Association for the Advancement of Artificial Intelligence (AAAI).

Cooper, R.P., Aslin, R.N., 1990. Preference for infant-directed speech in the first month after birth. Child development 61 5, 1584–95.

Cristia, A., 2019. A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. https://osf.io/c86ew/?view_only=f9af0cf7d2574234a8517c38151e4210.

Cristia, A., 2020. Language input and outcome variation as a test of theory plausibility: The case of early phonological acquisition. Developmental Review 57, 100914.

Cristia, A., 2021. PhonSES: socioeconomic status effects on infants' word and sound processing. https://gin.g-node.org/LAAC-LSCP/phonSES-public.

De Boer, B., Kuhl, P.K., 2003. Investigating the role of infant-directed speech with a computer model. Acoustics Research Letters Online 4, 129–134.

Decasper, A.J., Fifer, W.P., 1980. Of human bonding: newborns prefer their mothers' voices. Science 208 4448, 1174–6.

Decasper, A.J., Prescott, P., 1984. Human newborns' perception of male voices: preference, discrimination, and reinforcing value. Developmental psychobiology 17 5, 481–91.

Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Dupoux, E., 2021. The zero resource speech challenge 2021: Spoken language modelling, in: Interspeech.

Dupoux, E., 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. Cognition 173, 43–59.

Eaves Jr, B.S., Feldman, N.H., Griffiths, T.L., Shafto, P., 2016. Infant-directed speech is consistent with teaching. Psychological review 123, 758.

Feldman, N.H., Goldwater, S., Dupoux, E., Schatz, T., 2022. Do infants really learn phonetic categories? Open Mind 5, 113–131.

Floccia, C., Nazzi, T., Bertoncini, J., 2000. Unfamiliar voice discrimination for short stimuli in newborns. Developmental Science 3, 333–343.

Ford, M., Baer, C.T., Xu, D., Yapanel, U., Gray, S., 2008. The LENA^{TM} language environment analysis system .

García, P., Villalba, J., Bredin, H., Du, J., Castán, D., Cristia, A., Bullock, L., Guo, L., Okabe, K., Nidadavolu, P.S., Kataria, S., Chen, S., Galmant, L., Lavechin, M., Sun, L., Gill, M.P., Ben-Yair, B., Abdoli, S., Wang, X., Bouaziz, W., Titeux, H., Dupoux, E., Lee, K.A., Dehak, N., 2020. Speaker detection in the wild: Lessons learned from JSALT 2019, in: Odyssey.

Gilkerson, J., Richards, J.A., Warren, S.F., Oller, D.K., Russo, R., Vohr, B.R., 2018. Language experience in the second year of life and language outcomes in late childhood. Pediatrics 142.

Hitczenko, K., Feldman, N.H., 2022. Naturalistic speech supports distributional learning across contexts. Proceedings of the National Academy of Sciences 119, e2123230119.

Huang, Y., Rao, R.P., 2011. Predictive coding. Wiley Interdisciplinary Reviews: Cognitive Science 2, 580–593.

Hüllermeier, E., Fober, T., Mernberger, M., 2013. Inductive bias. Encyclopaedia of systems biology. Springer, New York .

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazar'e, P.E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., rahman Mohamed, A., Dupoux, E., 2020. Libri-light: A benchmark for ASR with limited or no supervision, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Kearns, J., 2014. Librivox: Free public domain audiobooks, in: Reference Reviews, Emerald Group Publishing Limited.

Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.E., Douze, M., Dupoux, E., 2021. Data augmenting contrastive learning of speech representations in the time domain, in: Spoken Language Technology Workshop (SLT).

Kuhl, P.K., 1979. Speech perception in early infancy: perceptual constancy for spectrally dissimilar vowel categories. The Journal of the Acoustical Society of America 66 6, 1668–79.

Kuhl, P.K., 2004. Early language acquisition: cracking the speech code. Nature Reviews Neuroscience 5, 831–843.

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., Nelson, T., 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). Philosophical Transactions of the Royal Society B: Biological Sciences 363, 979–1000.

Kuhl, P.K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., Iverson, P., 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Developmental science 9 2, F13–F21.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., Cristia, A., 2020. An open-source voice type classifier for child-centered daylong recordings, in: Interspeech.

Lavechin, M., Cristia, A., 2021. Babylogger versus LENA microphones study. https://gin.g-node.org/LAAC-LSCP/babylogger-vs-lena-data-public.

Lavechin, M., Métais, M., Titeux, H., Boissonnet, A., Copet, J., Rivière, M., Bergelson, E., Cristia, A., Dupoux, E., Bredin, H., 2022. Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and c50 room acoustics estimation. arXiv preprint arXiv:2210.13248 .

Levy, E.S., Strange, W., 2008. Perception of French vowels by American English adults with and without French language experience. Journal of phonetics 36, 141–157.

Li, R., Schatz, T., Matusevych, Y., Goldwater, S., Feldman, N.H., 2020. Input matters in the modeling of early phonetic learning, in: CogSci.

Marklund, E., Schwarz, I.C., Lacerda, F., 2019. Amount of speech exposure predicts vowel perception in four-to eight-month-olds. Developmental Cognitive Neuroscience 36, 100622.

Maye, J., Werker, J.F., Gerken, L., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition 82, B101–B111.

McMurray, B., 2022a. The acquisition of speech categories: Beyond perceptual narrowing, beyond unsupervised learning and beyond infancy. Language, Cognition and Neuroscience .

McMurray, B., 2022b. The myth of categorical perception. The Journal of the Acoustical Society of America 152, 3819–3842.

McMurray, B., Danelz, A., Rigler, H., Seedorff, M., 2018. Speech categorization develops slowly through adolescence. Developmental psychology 54, 1472.

McMurray, B., Kovack-Lesh, K.A., Goodwin, D., McEchron, W., 2013. Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? Cognition 129, 362–378.

Mehler, J., Bertoncini, J., Barriere, M., Jassik-Gerschenfeld, D., 1978. Infant recognition of mother's voice. Perception 7, 491 – 497.

Millet, J., Chitoran, I., Dunbar, E., 2021. Predicting non-native speech perception using the perceptual assimilation model and state-of-the-art acoustic models, in: Conference on Computational Natural Language Learning (CONLL).

Miyawaki, K., Jenkins, J.J., Strange, W., Liberman, A.M., Verbrugge, R., Fujimura, O., 1975. An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. Perception & Psychophysics 18, 331–340.

Miyazawa, K., Kikuchi, H., Mazuka, R., 2010. Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model, in: Interspeech.

Morton, J., Johnson, M.H., 1991. CONSPEC and CONLERN: a two-process theory of infant face recognition. Psychological review 98, 164.

Newman, R.S., 2005. The cocktail party effect in infants revisited: listening to one's name in noise. Developmental Psychology 41, 352.

Newman, R.S., Hussain, I., 2006. Changes in preference for infant-directed speech in low and moderate noise by 4.5-to 13-month-olds. Infancy 10, 61–76.

van den Oord, A., Li, Y., Vinyals, O., 2019. Representation learning with contrastive predictive coding. arXiv:1807.03748.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The kaldi speech recognition toolkit, in: Automatic Speech Recognition and Understanding (ASRU) workshop, IEEE Signal Processing Society.

Reh, R.K., Hensch, T.K., Werker, J.F., 2021. Distributional learning of speech sound categories is gated by sensitive periods. Cognition 213, 104653.

Rivière, M., Joulin, A., Mazaré, P.E., Dupoux, E., 2020. Unsupervised pretraining transfers well across languages, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Rowe, M.L., 2012. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. Child development 83, 1762–1774.

Saffran, J.R., 2003. Statistical language learning: Mechanisms and constraints. Current Directions in Psychological Science 12, 110–114.

Schatz, T., Feldman, N.H., Goldwater, S., Cao, X.N., Dupoux, E., 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. Proceedings of the National Academy of Sciences 118.

Schatz, T., Peddinti, V., Bach, F.R., Jansen, A., Hermansky, H., Dupoux, E., 2013. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline, in: Interspeech.

Schneider, S., Baevski, A., Collobert, R., Auli, M., 2019. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 .

Seidl, A., Onishi, K.H., Cristia, A., 2014. Talker variation aids young infants' phonotactic learning. Language Learning and Development 10, 297–307.

Singh, L., Rajendra, S.J., Mazuka, R., 2022. Diversity and representation in studies of infant perceptual narrowing. Child Development Perspectives 16, 191–199.

Sperry, D.E., Sperry, L.L., Miller, P.J., 2019. Reexamining the verbal environments of children from different socioeconomic backgrounds. Child development 90, 1303–1318.

Traer, J., McDermott, J.H., 2016. Statistics of natural reverberation enable perceptual separation of sound and space. Proceedings of the National Academy of Sciences 113, E7856 – E7865.

Tsuji, S., Cristia, A., 2014. Perceptual attunement in vowels: A meta-analysis. Developmental psychobiology 56, 179–191.

Vallabha, G.K., McClelland, J.L., Pons, F., Werker, J.F., Amano, S., 2007. Unsupervised learning of vowel categories from infant-directed speech. Proceedings of the National Academy of Sciences 104, 13273 – 13278.

Versteegh, M., Thiollière, R., Schatz, T., Cao, X.N., Miró, X.A., Jansen, A., Dupoux, E., 2017. The zero resource speech challenge 2017, in: Automatic Speech Recognition and Understanding (ASRU) workshop.

Vouloumanos, A., Werker, J.F., 2007. Listening to language at birth: evidence for a bias for speech in neonates. Developmental science 10 2, 159–64.

Warren, C., 2013. EchoThief impulse response library. http://www.echothief.com/.

Weisleder, A., Fernald, A., 2013. Talking to children matters: Early language experience strengthens processing and builds vocabulary. Psychological science 24, 2143–2152.

3.3 Lexical and syntactic acquisition from long-forms

Lavechin, M., Sy, Y., Titeux, H., Cruz Blandón, M. A., Räsänen, O., Bredin, H., Dupoux, E., Cristia, A. (2023) BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models. *Interspeech*

Motivation

In section 3.1, we presented a position paper advocating the use of developmentally plausible training sets in language learning simulations, a view shared by many in the field (Dupoux, 2018; Warstadt & Bowman, 2022; Warstadt et al., 2023). We presented one of our modeling investigations in Section 3.2 but limited ourselves to evaluating the sound discrimination capabilities developed by our artificial learner. Yet, self-supervised representation learning models trained on large quantity of clean audio or audiovisual input have been shown to learn lexical and syntactic aspects of their training language (T. A. Nguyen et al., 2020; Alishahi et al., 2021; Lavechin, de Seyssel, Titeux, et al., 2022). The question that arises is whether the same approach is applicable to our endeavor of modeling language acquisition from ecological data?

Current benchmarks have been designed for models trained on curated training sets, which most often consist of audiobooks in the case of spoken language models (Kahn et al., 2020; T. A. Nguyen et al., 2020; Hsu et al., 2021). For instance, the spot-the-word task proposed by T. A. Nguyen et al. (2020) covers a large vocabulary specific to books, including words like 'rhapsodize', 'zirconium', or 'tercentenary', which are vanishingly rare in children's language environments. Similarly, grammatical acceptability judgment tasks probe complex syntactic paradigms that are rare, even in spontaneous adult-adult conversations. One example from BLIMP (Warstadt et al., 2020) includes '*Who is Bill's nephew that won't attack Janice listening to?*' (grammatical) versus '*Who is Janice listening to Bill's nephew that won't attack?*' (ungrammatical). Consequently, current benchmarks do not reflect the characteristics of children's language environment and cannot be used to evaluate models trained on developmentally plausible training sets.

Paper summary

In Lavechin, Sy, et al. (2023), we introduce a language-acquisition-friendly benchmark to evaluate written or spoken language models at the lexical and syntactic levels. To ensure this benchmark is compatible with the vocabulary typical of children's language experiences, we used transcripts from various child-centered scenarios sourced from the CHILDES database (MacWhinney & Snow, 1985). Examples of stimuli are available on this project page[1] for the most curious of our readers.

To demonstrate the applicability of our benchmark, we use it to evaluate speech-based and text-based models trained on developmentally plausible training sets. In increasing order of data plausibility, we consider:

1. BabyBERTa (Huebner et al., 2021), a transformer-based language model trained on a 5M word corpus of various child-centered situations built from the CHILDES database (MacWhinney, 1996).

2. LSTM language models trained on the Providence corpus of spontaneous infant-parent interactions in phonetic, orthographic, or audio form.

3. STELA models (CPC+K-means+LSTM) trained on child-centered long-forms (the same model as used in Section 2.3 of this manuscript).

All models are trained on American English speech for a duration varying from 128 to 1,024 hours (equivalent to 1.2M and 9.6M words).

Our results indicate that speech-based language models trained directly on child-centered long-forms perform at chance level. When abstracting away the high variability of spontaneous speech and the difficult acoustic conditions of real-life recordings by training on phonemes or words, the scores on the spot-the-word and grammatical acceptability judgment tasks go up. Our LSTM model trained on phonemes reaches 75.4% accuracy on the spot-the-word task. BabyBERTa obtains the highest performance on the grammatical acceptability judgment task with an accuracy of 70.4%.

In summary, our findings indicate that considering truly ecological data as input remains beyond the capabilities of current models. To drive further advancements in the field, we identify two outstanding challenges that must be addressed. First, we show an important performance gap between models trained on speech and those trained on phonetic or orthographic transcripts. This might indicate that the acoustic units discovered by the speech-based language model lack the level

[1] `https://marvinlvn.github.io/projects/3_project`

of abstraction found in phonemes, a lesson compatible with one of our previous analyses (Lavechin, de Seyssel, Titeux, et al., 2022), included in Section 2.3 of this thesis. Second, we show another important performance gap between models trained on curated audiobooks and those trained on realistic long-forms. This suggests that the high variability of spontaneous speech, in contrast with read speech, and the challenging acoustic conditions found in ecological recordings severely impede the model's performance.

BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models

Marvin Lavechin[1,2], *Yaya Sy*[1], *Hadrien Titeux*[1], *María Andrea Cruz Blandón*[3], *Okko Räsänen*[3], *Hervé Bredin*[4], *Emmanuel Dupoux*[1,2,5], *Alejandrina Cristia*[1]

[1]LSCP, ENS, EHESS, CNRS, PSL University, Paris, France [2]Meta AI Research, France
[3]Unit of Computing Sciences, Tampere University, Finland [4]IRIT, CNRS, Toulouse, France
[5]Cognitive Machine Learning Team, INRIA, France

marvinlavechin@gmail.com

Abstract

Self-supervised techniques for learning speech representations have been shown to develop linguistic competence from exposure to speech without the need for human labels. In order to fully realize the potential of these approaches and further our understanding of how infants learn language, simulations must closely emulate real-life situations by training on developmentally plausible corpora and benchmarking against appropriate test sets. To this end, we propose a language-acquisition-friendly benchmark to probe spoken language models at the lexical and syntactic levels, both of which are compatible with the vocabulary typical of children's language experiences. This paper introduces the benchmark and summarizes a range of experiments showing its usefulness. In addition, we highlight two exciting challenges that need to be addressed for further progress: bridging the gap between text and speech and between clean speech and in-the-wild speech.

Index Terms: spoken language modeling, language acquisition, self-supervised learning, child language

1. Introduction and related work

Machine learning for Natural Language Processing (NLP) has led to models that develop linguistic competence from exposure to written or spoken language. On text, Language Models (LMs) now achieve impressive performance on a wide variety of natural language understanding tasks [1]. More recently, speech-based LMs have also shown impressive linguistic competence on lexical or grammatical acceptability judgment tasks [2, 3], or spoken language generation [4, 5]. Since these models develop linguistic competence without the need for human labels, they promise to advance our understanding of how infants learn language[6, 7, 8]. However, if we want to maximize the impact of the evidence obtained from LMs, it is essential to ensure that our simulations closely emulate real-life situations – as advocated for syntactic acquisition in text-based LMs in [8, 9].

How can we do so? First, we should match the *quantity* of data available to young infants. Although large differences exist across cultures [10] and socioeconomic contexts [11], current estimates of yearly speech input vary between 300 and $1,000$ hours for American English-learning children [6, 12]. This means that by age 3, American English-learning children would have been exposed to approximately 3,000 hours of speech – for those who received the most speech input. Yet, by then, infants know many words and already engage in simple conversations [13]. Second, we should match the *quality* of data available to young infants. Contrary to LMs, infants do not learn language by scraping the entire web or through exposure to a large quantity of audiobooks. Instead, infants' input is speech – not text –, and it contains a relatively small vocabulary arranged in simple and short sentences, sometimes overlapping across speakers and laced with various background noises [7, 14].

Evaluating LMs trained on quantitatively and qualitatively plausible corpora requires the creation of adapted benchmarks, but none exists for speech-based LMs – see [9] or the BabyLM challenge [15] for text-based LMs. Current benchmarks using zero-shot probing tasks, although inspired by human psycholinguistics (e.g., spot-the-word or grammatical acceptability judgment tasks), have been designed for models trained on audiobooks [2]. As a result, these benchmarks use a large vocabulary specific to books (including words like 'rhapsodize', 'zirconium', or 'tercentenary') and probe syntactically complex sentences that are vanishingly rare even in spontaneous adult-adult conversation.

Here, we propose *BabySLM*, the first language-acquisition-friendly benchmark to probe speech-based LMs at the lexical and syntactic levels, both of which are compatible with the vocabulary typical of children's language experiences. Our benchmark relies on zero-shot behavioral probing of LMs [2] and considers a spot-the-word task at the lexical level and a grammatical acceptability judgment task at the syntactic level. To show the utility of our benchmark, we first use it to evaluate text-based and speech-based LMs trained on developmentally plausible training sets. The text-based LM is a long short-term memory (LSTM) trained on phonemes or words. The speech-based LM is the low-budget baseline used in the ZeroSpeech 2021 challenge on unsupervised representation learning of spoken language [2]. Both systems are trained on Providence [16], a dataset of spontaneous parent-child interactions. The comparison between text-based and speech-based LMs shows an important gap that future work should address. Next, *BabySLM* enables us to compare the performance of speech-based LMs when trained on $1,000$ hours of speech extracted from 1) audiobooks, a source of training data commonly used [17, 18]; or 2) child-centered long-form recordings acquired via child-worn microphones as people go about their everyday activities [19]. Our results reveal that speech-based LMs are overly sensitive to the differences between clean speech and in-the-wild speech.

2. Methods

2.1. Metrics

2.1.1. Lexical evaluation: the spot-the-word task

General principle. In the lexical task, the system is presented with minimal pairs of an existing word and a pseudo-word that

We thank HPC resources of GENCI-IDRIS (2022-AD011012554); ANR-19-P3IA-0001; J. S. McDonnell Foundation; ERC (ExELang, 101001095).

Table 1: ***Lexical task.*** *Minimal pairs of real and pseudo-words. Phonetic (Phon.) transcriptions are given in International Phonetic Alphabet (IPA) standard. Orthographic (Orth.) transcriptions of pseudo-words are proposed for ease of reading.*

Word	Pseudo-words Phon.	Orth.	Word	Pseudo-words Phon.	Orth.
hello h ə l oʊ	l ə l oʊ	lello	thanks θ æ ŋ k s	θ ɛ ŋ k s	thaynks
	p ə l oʊ	pello		θ ɔ ŋ k s	thoanks
	s ə ɹ oʊ	sero		θ ɪ s k s	thisks
	d ə l oʊ	dello		θ æ m p s	thamps
	s ə l oʊ	sello		θ æ n t s	thants
cookie k ʊ k iː	k ʊ t iː	kootie	jump dʒ ʌ m p	dʒ æ m p	jamp
	k ʊ n i	koonie		dʒ ʌ l k	julk
	ɹ ʊ d iː	roodie		dʒ ʌ s k	jusk
	ɹ ʊ t iː	rootie		dʒ ʌ f t	juft
	b ʊ n iː	boonie		dʒ ʌ b s	jubs

is phonologically plausible but does not actually exist [2, 20] (examples in Table 1). The system gets a score of 1 if it returns a higher probability for the former, and 0 otherwise. Contrary to [2], we generate multiple pseudo-words per word. Scores are first averaged across pseudo-words to yield per-word accuracy, which are then averaged across all words to yield a measure of *lexical accuracy*.

Task generation. We first listed all words in the American English CHIld Language Data Exchange System (CHILDES) database [21]. This database contains human-annotated transcripts of various child-centered situations (play sessions, storytelling, etc.), making it a valuable source of vocabulary in real children's input. After excluding items not found in either the Celex [22] or CMU dictionary [23] (e.g., mispronounced, incorrectly annotated or made-up words: 'insectasaurus', 'hiphippopotamus'), we obtained $28,000$ word types. Pseudo-words were produced using the Wuggy pipeline [24], which generates, for a given word, a list of candidate pseudo-words matched for syllabic and phonotactic structure. We applied the same post-processing steps used in [2]. Contrary to [2], to ensure that there is no bias from phone-based unigrams or bigrams, we balanced the count of pseudo-words that had higher (or lower) phonemes unigram and bigram probabilities compared to those computed for the actual word. If a given word had only pseudo-words with higher (or lower) unigram or bigram possibilities, it was discarded from the evaluation set. The resulting $> 90,000$ minimal pairs across $18,000$ words were each synthesized using Google Text-To-Speech (TTS) system using 10 voices (5 males, 5 females).

2.1.2. Syntactic evaluation: grammatical acceptability

General principle. In the syntactic task, the system is presented with minimal pairs of grammatical and ungrammatical sentences across six syntactic phenomena [2, 9] (examples in Table 2), giving the system a score of 1 when it assigns a higher probability to the former, and 0 otherwise. We average scores within each syntactic phenomenon, then across phenomena to obtain our measure of *syntactic accuracy*.

Task generation. We generated templates for each of the six syntactic phenomena explored. For instance, for the noun-verb agreement phenomenon, we used templates such as "The <noun$_1$> <3rd person verb> <noun$_2$>" versus "The <noun$_1$> <1st person verb> <noun$_2$>". Contrary to [2], we restricted this benchmark to simple syntactic phenomena and short sentences which better reflect the type of input children are exposed to. We filled the templates using high-frequency words from CHILDES [21]. For instance, selected animate nouns include words like 'mom', 'girl', or 'cat'; selected adjectives include words like 'good', 'little', or 'big'; and selected verbs include words like 'see', 'know', or 'need'. The resulting $10,800$ minimal pairs were each synthesized using Google TTS system using the same 10 voices (5 males, 5 females).

Table 2: ***Syntactic task.*** *Minimal pairs of grammatical (✓) and ungrammatical (✗) sentences from each of the six syntactic phenomena included in our benchmark. N is the number of* $1,000$ *minimal pairs within each category.*

Phenomenon	N	Sentence example
Adjective-noun order	1.6	✓ *The good mom.* ✗ *The mom good.*
Noun-verb order	1	✓ *The dragon says.* ✗ *The says dragon.*
Anaphor-gender agreement	2	✓ *The dad cuts himself.* ✗ *The dad cuts herself.*
Anaphor-number agreement	1	✓ *The boys told themselves.* ✗ *The boys told himself.*
Determiner-noun agreement	3.6	✓ *Each good sister.* ✗ *Many good sister.*
Noun-verb agreement	1.6	✓ *The prince needs the princess.* ✗ *The prince need the princess.*

2.1.3. Development and test split

For both our lexical and syntactic evaluation sets, we randomly selected one male and one female voice for the development set and the 8 remaining ones for the test. We randomly selected $20\,\%$ of the lexical and syntactic minimal pairs for the development set and the remaining $80\,\%$ for the test.

2.2. Training sets

We built a first training set by extracting human-annotated speech utterances from Providence [16], a publicly available corpus containing transcribed recordings of six American children during spontaneous interactions with their parents. Available utterance-level timestamps were refined with a pretrained voice activity detection (VAD) system [25]. We converted human orthographic transcripts into phonetic transcripts using [26]. This procedure resulted in 128 hours of highly naturalistic infant-parent interactions in audio, orthographic, and phonetic form, allowing us to compare LMs trained on speech, phonemes, or words.

We built a second training set by extracting $1,024$ hours of adult speech utterances – using the same VAD system [25] – from SEEDLingS [19], a corpus of child-centered long-form recordings collected in 61 American English families. This training set enables us to train speech-based LMs in maximally plausible conditions, i.e., directly on what infants hear.

2.3. Models

STELA (speech-based). STELA is a speech-based LM originally proposed in [2, 27]. It comprises an acoustic model that learns discrete representations of the audio and a language

Table 3: ***The BabySLM benchmark.*** *Lexical and syntactic accuracies obtained by different language models trained on developmentally plausible corpora of speech, phonemes, or words. Numbers are computed on the test set, and performances on the development set are reported using small font size. The starred cumulated duration and number of words are estimates based on the* 1.2 M *of words present in the* 128 *hours of speech from Providence. Data plausibility indicates the extent to which the training set is close to the real sensory signal available to infants.*

System	Input	Training set	Cumulated duration (h)	Number of words (M)	Data plausibility	Lexical acc. (%)	Syntactic acc. (%)
Random baseline	—	—	0	0	—	49.2 52.5	49.3 50.0
STELA [27]	speech	SEEDLingS	1024	9.6*	+++	49.5 45.4	50.3 50.5
STELA [27]	speech	Providence	128	1.2	++	56.8 47.1	50.3 51.1
LSTM	phonemes	Providence	128	1.2	+	75.4 75.2	55.1 55.9
LSTM	words (BPE)	Providence	128	1.2	+	—	65.1 65.3
BabyBERTa [9]	words (BPE)	AO-CHILDES	533*	5	+	—	70.4 70.4

model trained on top of the learned discrete representations. The acoustic model is built from a Contrastive Predictive Coding (CPC) model followed by a K-means clustering algorithm. The language model consists of LSTM layers. We used the same architecture and hyper-parameters as the low-budget baseline proposed in [2]. Contrary to [2] who trained CPC by sampling the positive and negative examples from the same speaker, we applied a second constraint: negative examples were drawn from temporally close speech sequences to reduce mismatch between the positive and negative examples in terms of their local environment as this was found to be helpful when training on long-forms [14].

LSTM (text-based). We include LSTM LMs trained on words – using byte-pair encoding – or on phonemes, using the same architecture and hyper-parameters than [2].

BabyBERTa (text-based). BabyBERTa [9] is a transformer-based LM trained on a 5 M word corpus of American English child-directed input built from the CHILDES database [21].

3. Results and discussion

3.1. The BabySLM benchmark

Results obtained on our *BabySLM* benchmark are reported in Table 3. Rows are sorted according to the plausibility of the training data. Child-centered long-form recordings (SEEDLingS) have the highest plausibility score as these recordings faithfully capture children's everyday language experiences. In particular, long-forms collect audio data over a whole day – or several – and therefore sample the full range of language experiences across all possible contexts: the child may be in or out of the house, the speech may be directed to the child or others, etc. The audio extracted from in-home recordings of spontaneous infant-parent interactions (Providence) is slightly less plausible as it fails to capture the full range of language experiences: fewer speakers than in a real-life setting, most of the speech is directed to the child, etc. Finally, words and phonemes extracted from AO-CHILDES or Providence have the lowest plausibility score since infants do not learn language from orthographic or phonetic transcriptions but from the continuous signal that is speech.

Results indicate no evidence of lexical and syntactic knowledge for STELA trained on 1,024 hours of speech from SEEDLingS. This contrasts, in appearance, with what has been found in the ZeroSpeech challenge [2], but this is due to the large variability of speech found in long-forms as we will see in Section 3.3. Results are no different for STELA trained on 128 hours of speech extracted from Providence whose lexical and syntactic accuracies remain close to chance level. However, we hypothesize that the lexical accuracy obtained by STELA might increase with more audio data from semi-controlled recordings of infant-parent interactions as these contain cleaner speech than what is typically found in long-forms. Contrary to speech-based LMs, text-based LMs perform largely above chance level. As expected, the LSTM model trained on words reaches higher syntactic accuracy than the LSTM trained on phonemes. The highest syntactic accuracy is obtained by BabyBERTa, which is a transformer-based LM and has been trained on a larger quantity of data than our LSTM LMs.

Performances on *BabySLM* show a clear gap between text-based and speech-based LMs. Another important finding is that, as of now, spoken language modeling from children's real language experiences seems out of reach, as evidenced by the chance-level lexical and syntactic accuracies obtained by STELA trained on SEEDLingS. We dedicate the remaining sections to illustrating these two challenges: bridging the gap between text and speech and between clean speech and in-the-wild speech.

3.2. Language modeling: from text to speech

Figure 1 shows lexical and syntactic accuracies obtained by text-based (words or phonemes) or speech-based LMs as a function of quantity of data. The LSTM trained on phones requires at least 16 hours of speech, equivalent to 150,000 words, to start performing above chance level. Once lexical knowledge has emerged, the model follows a logarithmic trend (note the log-scale x-axis), initially improving rapidly and then slowing down. In other words, we need to double the amount of data to obtain the same gain in lexical accuracy. The same patterns hold for the syntactic accuracy obtained by the LSTM model trained on words[1]. For STELA, the lexical accuracy remains close to chance level, although the curve seems to increase between 32 and 128 hours of speech, and there is no evidence for syntactic knowledge.

All in all, the lexical and syntactic accuracy slopes show very different patterns when training from raw speech or phonemes or words. This is despite receiving the same data

[1]Note, however, that the syntactic accuracy obtained by the LSTM model trained on words decreases to 45 % (below chance level) between 0 and 8 hours (= 75,000 words). This effect was found to be driven by co-occurrence statistics in the noun-verb order task. The same pattern was found with a 3-gram model, with a slight decrease between 0 and 8 hours and an increase between 8 and 128 hours.

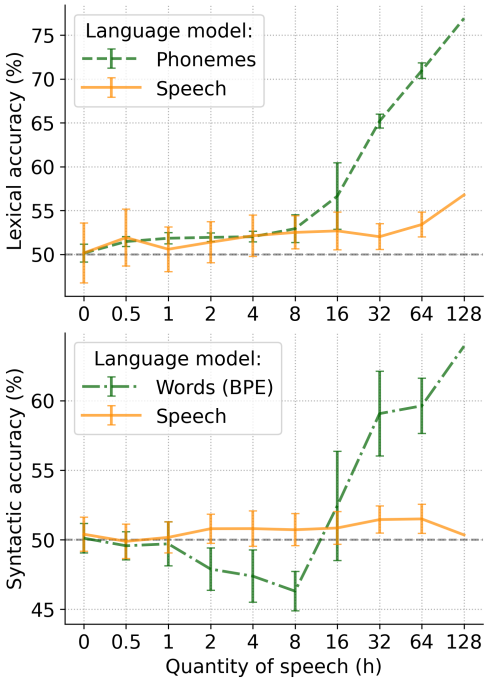

Figure 1: ***Language modeling from text to speech.*** *Top panel shows the lexical accuracy obtained by language models trained on audio (STELA) or phonemes (LSTM). Bottom panel shows the syntactic accuracy obtained by language models trained on audio (STELA) or byte-pair-encoded (BPE) words (LSTM). All models are trained on the Providence corpora in audio, phonetic, or orthographic form. Numbers are computed on the test set. Error bars represent standard errors computed across mutually exclusive training sets.*

in different forms. Admittedly, the speech-based LM faces a more challenging task as it must learn its own discrete units, while text-based LMs must not. Future work might investigate how these slopes change with more data, particularly for the speech-based LM for which 128 hours seems insufficient.

3.3. Language modeling: from clean to in-the-wild speech

So far in the paper, we have little evidence that lexical or syntactic knowledge can emerge in speech-based LMs. To address this concern, we ran one more experiment, this time training STELA under more controlled recording conditions: on up to $1,024$ hours of speech extracted from audiobooks – commonly used to train speech-based LMs [17]. Figure 2 compares this experiment against the performance obtained by STELA when trained on child-centered long-forms (SEEDLingS, Table 3).

Results are unequivocal: we observe a strong improvement on the lexical task for the model trained on audiobooks, while the same model trained on long-forms remains at chance level. On the syntactic task (not shown above), STELA trained on $1,024$ hours of audiobooks obtains an accuracy of $52.8\,\%$ compared to $50.3\,\%$ on long-forms. This is in line with the results in [2] showing that more powerful architectures are necessary to learn at the syntactic level.

Why do we observe chance-level performance when training on long-forms? First, the speech signal found in long-forms is much more challenging than the one found in audiobooks: the speech might be distorted as it is being spoken far from the child; it might overlap with various background noises; and it is

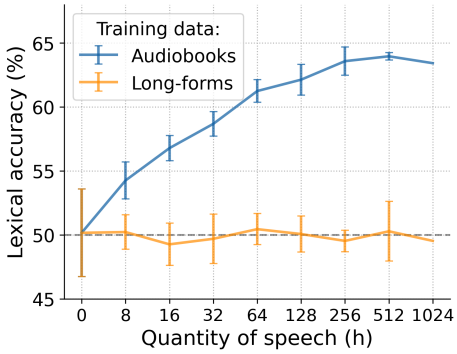

Figure 2: ***Language modeling from clean to in-the-wild speech.*** *Lexical accuracy obtained by STELA trained on audiobooks (Libri-light, in blue) or child-centered long-forms (SEEDLingS, in orange) as a function of speech quantity. Numbers are computed on the test set. Error bars represent standard errors computed across mutually exclusive training sets.*

often produced in short turns that might be under-articulated – see [14] for a comparative analysis. Another essential factor to consider is the domain mismatch between the training and test sets. While the training set contains far-field under-articulated speech as well as close-field storytelling, the test set consists of well-articulated synthesized stimulus to which STELA fails to generalize. However, infants show no difficulties generalizing from uncontrolled real-life conditions to more controlled ones (in-laboratory conditions). We advocate here that generalization is part of the language acquisition problem, and LMs should be evaluated accordingly.

We hypothesize that the discrete units learned by STELA might be too dependent on the various non-linguistic factors found in long-forms, as suggested in [14]. This dependency could prevent the LSTM LM from learning long-term dependencies necessary to solve the lexical or syntactic tasks.

4. Conclusion

Benchmarks are instrumental in allowing cumulative science across research teams. In this paper, we have described how BabySLM has been carefully designed to be adapted to the kinds of words and sentences children hear. We have shown how it can be used to evaluate LMs trained on developmentally plausible text or speech corpus. By doing so, we revealed two outstanding challenges that the community must solve to build more plausible cognitive models of language acquisition. First, we need to reduce the gap between text-based and speech-based LMs, as the latter performed close to chance level on BabySLM. Second, we need to reduce the gap between LMs trained on clean and in-the-wild speech, as evidenced by the striking difference we obtained on the lexical task when training on clean audiobooks versus ecological long-forms.

Future work might consist in evaluating speech-based LMs grounded in the visual modality [28], or linking performances obtained on *BabySLM* with behavioral measures in infants – e.g., age of acquisition as in [29]. A crucial limitation of our benchmark is that it focuses on English, which already accounts for a whopping $54\,\%$ of language acquisition studies [30]. We hope that this paper, together with shared scripts[2], will facilitate the creation of similar benchmarks in other languages.

[2] `https://github.com/MarvinLvn/BabySLM`

5. References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," in *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[3] E. Dunbar, M. Bernard, N. Hamilakis, T. Nguyen, M. de Seyssel, P. Roz'e, M. Rivière, E. Kharitonov, and E. Dupoux, "The zero resource speech challenge 2021: Spoken language modelling," in *Interspeech*, 2021.

[4] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[5] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T. Nguyen, M. Rivière, A. rahman Mohamed, E. Dupoux, and W.-N. Hsu, "Text-free prosody-aware generative spoken language modeling," in *Annual Meeting of the Association for Computational Linguistics*, 2021.

[6] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43–59, 2018.

[7] M. Lavechin, M. de Seyssel, L. Gautheron, E. Dupoux, and A. Cristia, "Reverse engineering language acquisition with child-centered long-form recordings," *Annual Review of Linguistics*, vol. 8, pp. 389–407, 2022.

[8] A. Warstadt and S. R. Bowman, "What artificial neural networks can tell us about human language acquisition," in *Algebraic Structures in Natural Language*. CRC Press, 2022, pp. 17–60.

[9] P. A. Huebner, E. Sulem, F. Cynthia, and D. Roth, "Babyberta: Learning more grammar with small-scale child-directed language," in *Proceedings of the 25th conference on computational natural language learning*, 2021, pp. 624–646.

[10] A. Cristia, E. Dupoux, M. Gurven, and J. Stieglitz, "Child-directed speech is infrequent in a forager-farmer population: A time allocation study," *Child Development*, vol. 90, no. 3, pp. 759–773, 2019.

[11] S. Dailey and E. Bergelson, "Language input to infants of different socioeconomic statuses: A quantitative meta-analysis," *Developmental science*, vol. 25, no. 3, p. e13192, 2022.

[12] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing, 1995.

[13] E. Hoff and M. Shatz, *Blackwell handbook of language development*. John Wiley & Sons, 2009.

[14] M. Lavechin, M. de Seyssel, M. Métais, F. Metz, A. Mohame, H. Bredin, E. Dupoux, and A. Cristia, "Statistical learning models of early phonetic acquisition struggle with child-centered audio data," Mar 2022. [Online]. Available: psyarxiv.com/5tmgy

[15] A. Warstadt, L. Choshen, A. Mueller, A. Williams, E. Wilcox, and C. Zhuang, "Call for papers – The BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus," 2023. [Online]. Available: https://arxiv.org/abs/2301.11796

[16] B. Börschinger, M. Johnson, and K. Demuth, "A joint model of word segmentation and phonological variation for English word-final/t/-deletion," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 1508–1516.

[17] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for ASR with limited or no supervision," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673.

[18] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *Journal of Selected Topics in Signal Processing*, 2022.

[19] E. Bergelson, M. Casillas, M. Soderstrom, A. Seidl, A. S. Warlaumont, and A. Amatuni, "What do North American babies hear? A large-scale cross-corpus analysis." *Developmental science*, vol. 22 1, p. e12724, 2019.

[20] G. Le Godais, T. Linzen, and E. Dupoux, "Comparing character-level neural language models using a lexical decision task," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 125–130.

[21] B. MacWhinney and C. Snow, "The child language data exchange system," *Journal of child language*, vol. 12, no. 2, pp. 271–295, 1985.

[22] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "Celex2," *Linguistic Data Consortium, Philadelphia*, 1996.

[23] R. Weide *et al.*, "The Carnegie Mellon pronouncing dictionary," *release 0.6, www.cs.cmu.edu*, 1998.

[24] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior research methods*, vol. 42, pp. 627–633, 2010.

[25] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," in *Interspeech*, 2020.

[26] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.

[27] M. Lavechin, M. de Seyssel, H. Titeux, H. Bredin, G. Wisniewski, A. Cristia, and E. Dupoux, "Statistical learning bootstraps early language acquisition," Dec 2022. [Online]. Available: psyarxiv.com/rx94d

[28] M. Nikolaus, A. Alishahi, and G. Chrupała, "Learning English with Peppa Pig," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 922–936, 2022.

[29] E. Portelance, J. Degen, and M. C. Frank, "Predicting age of acquisition in early word learning using recurrent neural networks." in *CogSci*, 2020.

[30] E. Kidd and R. Garcia, "How diverse is child language acquisition research?" *First Language*, vol. 42, no. 6, pp. 703–735, 2022.

3.4 What is going on with child-centered long-form recordings?

At this point of the document, it is useful to take a step back and summarize some of the results encountered through Chapters 2 and 3. Section 2.3 demonstrated that, when exposed to clean recordings of speech, the STELA algorithm (CPC+K-means+LSTM) successfully learns phonetic and lexical aspects of its training language. In Section 3.2, we found that the same algorithm exposed to ecological long-forms needs inductive biases in the form of data augmentation or filtering mechanisms to reproduce perceptual attunement – i.e., discriminate sounds in a language-specific manner. Finally, Section 3.3 showed that, despite these inductive biases, no evidence for learning at the lexical level is found when the language exposure consists of child-centered long-forms.

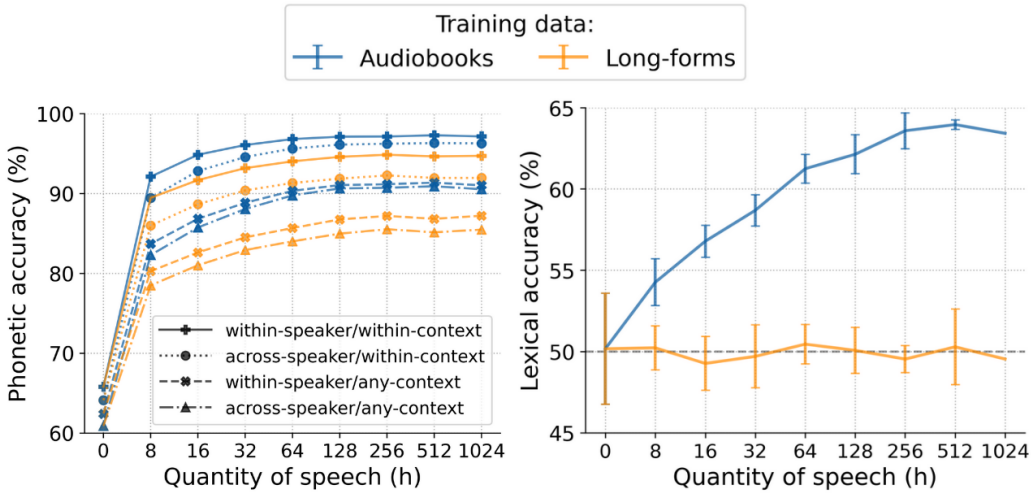

Fig. 3.1.: Phonetic and lexical accuracies obtained by STELA (CPC+K-means+LSTM) models trained on American English audiobooks (in blue) or child-centered long-form recordings (in orange) as a function of quantity of speech. The phonetic accuracy is computed using the ABX sound discrimination task (from Hallap et al., 2022) and the lexical accuracy is computed using the spot-the-word task (from BabySLM, Lavechin, Sy, et al., 2023. Numbers are computed on the test set. Error bars represent standard deviations computed across mutually exclusive training sets. Standard deviations on the phonetic accuracy are are too small to be displayed (e.g., $\mu_{ABX} = 94.04\%$ and $\sigma_{ABX} = 0.21\%$ in the within-speaker/within-context condition for models trained on 64 hours of audiobooks).

The observed failure observed when training on long-forms prompts us to undertake a more thorough evaluation of the information learned by these models, which we do in Figure 3.1. This Figure shows the phonetic accuracy on the ABX sound discrimination task (left graph) and the lexical accuracy on the spot-the-word task

(right graph) obtained by STELA models trained on curated audiobooks or ecological long-forms.

Focusing on the invariance of the learned representations, we use the ABX sound discrimination task proposed by Dunbar et al. (2021). It covers two speaker conditions: the within-speaker condition in which A, B, and X are drawn from the same speaker; and the across-speaker condition in which A and B are drawn from the same speaker and X from a different speaker. The performance gap between these two conditions enables us to assess the invariance of the learned representations with respect to the speaker information. Similarly, it covers two context conditions: the within-context condition in which A, B, and X are drawn within the same phonetic context (e.g., A='bip', B='bop', X='bip'); and the any-context condition in which A, B, and X are drawn from any phonetic context (e.g., A='bip', B='tol', X='bil'). The comparison between these two conditions enables us to assess the invariance of the learned representations with respect to the phonetic context.

The left graph of Figure 3.1 indicates that, regardless of the speaker or context condition, the phonetic accuracy of models trained on long-forms is lower than those trained on audiobooks. The results show that the learned representations lack invariance with respect to the speaker identity and the phonetic context. This is true both for models trained on audiobooks and those trained on long-forms. However, in the within-context condition, models trained on long-forms seem more sensitive to the speaker information than models trained on audiobooks (solid versus dotted lines). Compared to Dunbar et al. (2021), this graph adds important information, indicating that the sensitivity to the speaker and context information is independent of the quantity of data. In other words, adding more data to the training set does not allow the model to learn representations that are more robust to the speaker information and the phonetic context (with the exception, perhaps, of models trained on audiobooks, which seem to become more robust to the speaker information as the quantity of data increases, see the gap between the solid blue line and the dotted blue line slightly reduces when adding more data).

The lexical task is the language-acquisition-friendly version presented in BabySLM Lavechin, Sy, et al., 2023. By using the BabySLM version, designed with the vocabulary typical of children's language experiences, we advantage models trained on long-forms over those trained on audiobooks. Despite this advantage, models trained on long-forms exhibit a chance-level performance, whereas those trained on audiobooks do not. In contrast with the performance obtained on the sound discrimination task, here, we are faced with the stark lesson that models trained on ecological long-forms utterly fail on the lexical task.

As advocated multiple times in this manuscript, this is likely due to the challenging acoustic conditions and the high variability of spontaneous speech found in children's language environments. But how do these characteristics affect the outcomes of our artificial learner? How do infants succeed in acquiring their native language given such a sparse and noisy input? How can we revise our theories to better account for the complexity of children's language experiences? There remain many open questions that we have only just begun to explore in our submission to *Cognition*, presented in Section 3.2.

3.5 Conclusion

In this chapter, we presented our approach to simulating language acquisition from ecologically-valid data in the form of child-centered long-forms. We showed that dedicated mechanisms, called inductive biases, were necessary to guide the learning process in our algorithm and reflected on whether similar inductive biases could shape language acquisition in infants. Although these inductive biases enable the CPC algorithm to learn representations that better discriminate sounds, they do not appear sufficient for STELA (CPC+K-means+LSTM) to learn at the lexical and syntactic levels. Interestingly, when trained on curated audiobooks, the same algorithm effectively learns at the lexical and syntactic levels, as demonstrated in Section 3.3. More broadly, our results indicate that the learning outcomes developed by computational models are exquisitely sensitive to the details of the input signal, potentially casting doubts on modeling studies relying on manicured input data.

Although some of the limitations discussed in Chapter 2 also apply to Chapter 3, we will refrain from repeating them here. We first reflect on why we observe no evidence of learning at the lexical level when training on long-forms and argue that this shortcoming observed in STELA may extend to other statistical learning models. Finally, we discuss an essential aspect of our approach: the ecological validity of child-centered long-form recordings.

Statistical learning algorithms are universal pattern finders. One key guiding principle in self-supervised learning algorithms trained on a large quantity of data is "*the more you put in, the more you get out*". The task used during training, like next word or next frame prediction, is optimized over the entire training set without considering whether attempting the prediction is meaningful or allows the model to learn anything useful. As a consequence of this, such models are capable of learning various types of information.

In particular, self-supervised learning models applied to speech have been shown to encode phonetic, prosodic, lexical, and syntactic information (T. A. Nguyen et al., 2020; Lavechin, de Seyssel, Titeux, et al., 2022; de Seyssel et al., 2023), but also gender (de Seyssel, Lavechin, Adi, et al., 2022) and speaker identity (van Niekerk et al., 2021). The astonishing ability of self-supervised models to learn everything and anything of their input data may come at our disadvantage if our goal is to build computational models of language acquisition.

In fact, a more desirable property for these models would be to acquire abstract representations unaffected by the identity of the speaker producing the sound, word, or sentence and by the acoustic conditions in which it is produced. To clarify, we do not suggest that computational models should develop categorical perception entirely independent of the speaker or acoustic information. In this regard, our views align with the recent and expanding line of work questioning our interpretations that infants acquire phonetic categories (Feldman et al., 2021; Schatz et al., 2021; McMurray, 2022). Rather, we advocate that the speech perceptual capabilities developed by the model (e.g., its sound discrimination capabilities) should be robust to speaker variations and variations due to the acoustic environment.

The research community is actively working on finding ways to disentangle the representational subspaces encoding the phonetic and speaker information, e.g., Liu et al. (2023). However, much less attention has been dedicated to disentangling the information pertaining to the acoustic conditions (e.g., background noise and reverberation). I believe this presents a considerably more challenging problem that greatly contributes to the observed failure of STELA when being trained on long-forms. Indeed, additive noise and reverberation significantly impede the performance of our artificial learner, with an absolute decrease of 4.4% in terms of native discrimination and 13.5% in terms of native advantage (see Section 3.2).

Going back to the infant literature, behavioral evidence suggests that infants progressively learn representations that are invariant with respect to a change in speaker (Seidl et al., 2014; Bergmann et al., 2016; Choi & Shukla, 2021). However, there is limited research examining the influence of background noise and reverberation. For an exception rather than the rule, see Newman and Hussain (2006), who showed that 5-month-olds fail to recognize their own name when the SNR is 10 dB. The findings from our modeling investigation reveal an intriguing prediction: the presence of background noise and reverberation should strongly impede infants' perceptual abilities, with a greater loss under reverberant conditions. There remains to know if this prediction is accurate, and if so, to assess how robustness to these factors changes as infants develop.

That being said, Marianne Métais, an engineer on our team, and I are developing an ABX sound discrimination task on noise sequences. Our objective is to understand better the sensitivity of computational models to background noises. In this version of the ABX task, the model is, for instance, asked to discriminate the sound emitted by a vacuum cleaner from that emitted by an air-conditioning system. Initial findings indicate that models trained on audiobooks exhibit lower performance than those trained on long-forms, which aligns with the cleaner recording conditions found in audiobooks. Interestingly, when the proportion of speech sounds in the training set is varied relative to the proportion of non-speech sounds, there appears to be a trade-off between performance on the triphone and the noise version of the task. It remains to be checked if a similar trade-off exists when speech and background noises overlap. I hypothesize that models trained predominantly on low-SNR utterances will perform well on the noise version of the task and poorly on the triphone version of the task (and vice-versa for high-SNR utterances). Once we have measures to assess the model's capability to discriminate between speech sounds and between noises, we can begin designing strategies to make them more robust to the various acoustic conditions found in real-life audio recordings.

Capturing the sensory signal available to infants. In this chapter, we advocate that language learning simulations should be fueled with ecologically-valid input data. *But are child-centered long-form recordings truly ecological?*

This question is akin to asking: *To what extent can one capture the sensory signal available to infants?* Throughout this manuscript, we employed the LENA® microphone device, which the child wears. This approach offers the undeniable advantage of directly collecting language environments from the perspective of the infant learner. Nevertheless, there exist immediate ways to improve the hardware. For instance, the LENA® microphone only collects single-channel recordings, making it hard to apply any source separation algorithm without prior knowledge of the number of sources and their characteristics. One way to reduce the gap between infants and our artificial learners would be to equip the latter with 'ears' using binaural recordings. Indeed, the remarkable ability of the human auditory system to separate mixtures of signals likely plays a critical role during language acquisition (Bregman, 1994).

Beyond the auditory stream, one can attempt to collect other sources of information. For instance, capturing the infant's visual and social environment would necessitate head-mounted cameras, as done by Long et al. (2022). However, this method remains too invasive to be used over an entire day, but see Casillas et al. (2020), who propose a more lightweight setup in which the infant wears a recording vest including both a microphone and a miniature camera with a fish-eye lens. Certainly

cheaper to implement, an accelerometer could provide critical information on the child's body movements.

While it is currently not possible to digitize touch, smell, and taste at the long-form scale, the integration of sensors capable of capturing these sensory streams could offer artificial learners valuable information – see Seidl et al. (2015) for a study where human infants learn better from tactile-speech than visual-speech co-occurrences.

Although largely unaddressed in the research community – see Cao et al. (2018) for one of the few available alternatives to the LENA® recording device –, the choices underlying the hardware design are critical for describing children's language environments. What is captured and what is not delineates the field of possibilities in language learning simulations. So to the question: *Are child-centered long-form recordings truly ecological?* The answer could be: *Yes, but they are not perfect.* Nonetheless, alternatives in language acquisition modeling often involve running algorithms on strings of words or highly manicured read speech. Without a doubt, child-centered long-forms offer, by far, a more ecological way to capture children's language environment.

4 General discussion

We carried out a body of work illustrating how artificial neural networks can be used to analyze and simulate language acquisition in children. Regarding the analysis, we presented a suite of automatic speech processing tools to extract some of the interesting information bits contained in child-centered long-form recordings, yielding a more accurate view of children's language environment. Regarding the simulation, we found that the learning outcomes developed by our model were exquisitely sensitive to the details of the input signal, showing the importance of considering ecologically-valid input data when modeling language acquisition. As we observed that inductive biases were necessary for our learners to reproduce perceptual attunement when exposed to ecological input data, we reflected on whether similar inductive biases may play a role in infant language acquisition.

Rather than repeating the implications of our work, its limitations, and the potential avenues for future research covered in their respective chapters, we discuss two new matters here. First, highlighting the synergy between the different chapters of this thesis, we explain how our work in building automatic speech processing tools to analyze long-form recordings offered us new modeling opportunities. By assessing the effect of built-in capabilities in artificial learners, we might find new ways to explore the age-old nature versus nurture debate (Piaget, 1935; Chomsky, 1957). Second, we reflect on the core question explored in this thesis: *What can artificial neural networks tell us about infant language acquisition?* Through a thought experiment, we outline what we think is the necessary trajectory for the field to make substantial advancements.

4.1 Summary of our main contributions

Before delving deeper into the discussion, we present a summary of our main contributions in Table 4.1. Since these contributions have already been outlined in the preceding chapters, we will refrain from repeating them here.

Chapter	Contributions	Conclusions
	Artificial neural networks as a tool	
1	• Short study assessing the performance of a state-of-the-art automatic speech recognition system on long-forms collected in American English families • Development of a suite of speech processing tools to detect voice activity, identify voice signal sources, count the number of linguistic units, and estimate the quantity of background noise and reverberation	• Off-the-shelf tools do not currently work on long-forms, so we must develop our own through stronger collaborations between language and speech processing communities • Our speech processing suite provides a free, open-source, and better-performing alternative to the LENA® proprietary software to obtain automatic analyses of long-forms
	Artificial neural networks as a model	
2	• Design of a developmental cross-linguistic and psycholinguistic framework to assess artificial learners' learning trajectories • Application of our framework to study the phonetic and lexical learning trajectories of a self-supervised learning algorithm trained from clean recordings of read speech	• Statistical learning theories are a priori sufficient to instantiate a gradual and parallel phonetic and lexical learning (from clean recordings of read speech) • It is possible that linguistic categories are not necessary during language acquisition and instead emerge as a result of the learning process
3	• Advocacy for modeling language acquisition from ecological data: training, evaluation, and new research directions • Assessment of some of the effects of learning from curated recordings compared to ecological long-form recordings • Modeling investigation of early phonetic acquisition from ecological long-form recordings • Creation of zero-shot lexical and syntactic probing tasks compatible with the vocabulary typical of children's language experiences	• Learning outcomes are exquisitely sensitive to the details of the input signal, and our theories and models inadequately account for what children truly hear • Inductive biases might be necessary to reproduce perceptual attunement, potentially offering new ways to explore the nativist versus constructivist debate • No evidence of lexical learning is found suggesting that more work is needed in engineering the right inductive biases and/or more robust learning mechanisms

Tab. 4.1.: Summary of the main contributions presented in this thesis.

4.2 New ways for exploring the nativist versus constructivist debate?

Chapter 1 was dedicated to using artificial neural networks *as a tool* to automatically analyze children's language environments. Chapters 2 and 3 were dedicated to using artificial neural networks *as a model* of the infant learner. Although these two research areas may appear largely independent, they are far from being so.

Throughout this thesis, developing new tools that extract interesting information bits from child-centered long-form recordings always opened up new modeling opportunities. By equipping our artificial learner with new built-in capabilities, the voice type classifier (Lavechin et al., 2020) enabled us to filter out non-speech segments and children's vocalizations from the training set. While non-speech segments strongly degrade the model's performance on the ABX sound discrimination task, as seen in Section 2.1.1, this is not the case for children's vocalizations which have no effect. Driven by the idea that the quality of the input speech signal may have an effect, we tried to remove utterances whose speech probability – such as estimated by the voice type classifier – was the lowest, assuming that these utterances were of lower acoustic quality. This had no effect either.

Two years later, when we built Brouhaha (Lavechin, Métais, et al., 2022), we could investigate the issue of signal acoustic quality on the performance of our artificial learner in a more thorough way, studying the impact of the speech-to-noise ratio and the C_{50} reverberation measure. Although we observed only a slight performance improvement, we found that it was possible to filter out 70% of the utterances in the training set without impeding the ABX sound discrimination performance (keeping only the 30% of utterances whose C_{50} was the highest, i.e., very little reverberation). We did not observe any performance improvement on the spot-the-word task, which remained at the chance level, suggesting that data augmentation and filtering mechanisms may not be enough to learn at the lexical level from ecological long-forms.

Among other investigations, one that seemed particularly promising consisted in using a speech enhancement model to help our model learn better on ecological long-forms. We tried the model proposed by Défossez et al. (2020). After all, if some speech segments found in long-form recordings are noisy and the acoustic quality of the input speech signal affects the performance, removing the background noise should help our artificial learner to learn better. It did not help either, regardless

of whether the speech enhancement model was applied to the training set, the evaluation set, or both.

Undeniably, the results might have been different with better-performing algorithms or algorithms specifically designed for long-form recordings (which is not the case for the speech enhancement model we used or for Brouhaha). However, no automatic speech processing algorithm is error-free. The same applies to human speech perception, especially while it is still developing, and finding strategies to deal with difficult acoustic conditions is part of the learning problem.

In Chomsky's words (1957), our approach to exposing artificial learners to children's real language environment involved engineering the right 'Language Acquisition Device' (LAD). One thing for sure is that there remains much work to be done to engineer better and more performant LADs that are both compatible with the actual input received by children and allow artificial learners to discover the rules and structure of their native language.

One of the only approaches I can think of to investigate whether some hypothesized built-in capabilities may contribute to infant language acquisition is by equipping artificial language learners with these same capabilities. In that sense, our modeling approach represents a promising opportunity to shed new light on the nativist versus constructivist debate in language acquisition (Piaget, 1935; Chomsky, 1957; Tomasello, 2005) – see Ambridge and Lieven (2011) for an overview.

4.3 What can artificial neural networks tell us about infant language acquisition?

Modeling studies provide precious learnability proofs. For instance, using a statistical learning algorithm applied to clean recordings of speech, Schatz et al. (2021) showed that it was possible to reproduce some developmental results in speech perceptual learning. In this thesis, we showed that our own statistical learning algorithm applied to clean recordings of speech was capable of learning phonetic and lexical aspects of its training language in a gradual and parallel fashion, which is also how infants do as argued with the developmental timeline presented in Figure 2.1. The same algorithm applied to children's real language environments collected via child-worn microphones requires extra built-in capabilities to learn at the phonetic level, and we found no evidence for learning at the lexical level.

Under the simplifying assumptions made in these studies, these proofs are indisputable[1], akin to theorems in mathematics. However, these proofs are all about the input signal, not the infant. So one may rightfully wonder: *What can artificial neural networks tell us about infant language acquisition?*

At the risk of disappointing my supervisors, I have yet to find a precise answer to this question, even after dedicating an entire thesis to the subject – which represents too little time to learn about a new field. Proofs about the input signal constitute essential cues that can help us validate or reject our theories of infant language acquisition, but can we do better and go beyond proofs about the input signal? To this question, I would answer a definite yes. However, this will likely require datasets with increased:

1. *density* – we need the full input received by children to reproduce language acquisition

2. *size* – we need a high enough number of children for our conclusions to be statistically robust

3. *diversity* – participating children should be drawn from diverse linguistic, cultural, and socioeconomic contexts as we do not want our conclusions to depend on these factors

To illustrate the envisioned strategy, we can use a thought experiment to describe an ideal dataset. Let us imagine that we can collect the whole language experience of many children, let us say 200 children, from birth until age 3. Although collecting only audio would likely yield significant advances in the field, one might imagine collecting videos too. This is our input I_k for $k \in [\![1, 200]\!]$.

In parallel, the language capabilities developed by these 200 children should regularly be measured through a battery of psycholinguistic and standardized language tests, let us say the full battery every month. This is our output O_k for $k \in [\![1, 200]\!]$.

Then, it becomes possible to feed 200 artificial language learners with the individual experience of the 200 participating children I_k and correlate the predicted learning outcomes developed by the machine $\hat{O}_k$ with those of the participating children O_k. We can study how these correlation coefficients change as a function of the child's age, the presence – or absence thereof – of some built-in capacities in the machine, the access to visual or social cues, the learning mechanism, and many other variables.

[1]Of course, for this to be true, the study must first be shown reproducible.

For instance, an interesting modeling experiment could vary the proportion of the child's individual language experience to which the artificial learner has access. We would expect the predicted learning outcomes $\hat{O}_k$ to better fit O_k when our artificial learner is exposed to the entirety of the child's individual language experience than when exposed to only half of it.

If statistical learning algorithms capture a significant proportion of the observed learning outcomes variability O_k, this will constitute strong evidence that these algorithms do indeed capture something essential about language acquisition. More importantly, this would inform us of the extent to which infant language acquisition relies on learned (the proportion of captured variability) or innate behaviors (the uncaptured proportion).

While it is true that our hypothetical experiment does not provide causal evidence that infants learn via mechanism M, an approach based on modeling developmental trajectories from the child's individual language environment – in opposition to modeling an abstract average infant – would take us one step closer to our goal. Instead of providing proofs about the input signal contained in the training set, we would provide proofs about the child's unique language environment.

This hypothetical research program may appear ambitious, but collecting such a dense, large-scale, longitudinal dataset seems within reach, given enough time, money, and sweat. As a matter of fact, a few large-scale longitudinal datasets have already been collected or are in the process of being collected (see Table 4.2). While most initiatives have been concentrated to American English speakers, see the Casillas HomeBank corpus (Casillas et al., 2017) for a large-scale study of Tseltal Mayan children and the Warlaumont HomeBank corpus (Warlaumont et al., 2016) for a longitudinal study of English- and/or Spanish-learning children.

Corpus	Number of children	Age (mo)	Recording frequency	Language	Audio?	Video?	Child centered?
Human Speechome	1	$0-36$	daily	Am. E.	✓	✓	✗
SEEDLingS	44	$6-18$	monthly	primarily Am. E.	✓	✓	✓
SAYCam	3	$6-32$	weekly	Am. E. Au. E.	✓	✓	✓
First 1,000 Days	~ 20	$0-36$	daily	Am E.	✓	✓	✗

Tab. 4.2.: A sample of large-scale longitudinal datasets collected to study language development in infants. The Human Speechome project (Roy et al., 2006) is one of the earliest initiatives in this direction, using a dozen cameras and microphones to record 10h of audio and video on a daily basis. SEEDLingS (Bergelson, Amatuni, et al., 2019; Bergelson, Casillas, et al., 2019), without which this thesis would not have been possible, used LENA® microphones to collect up to 14h of audio on a monthly basis. Besides long-form recordings, each participating child was recorded at home for 1h at a time every month using head-mounted cameras. SAYCam (Sullivan et al., 2021) consists of audio and video recordings collected via head-mounted cameras for approximately 2h per week. The ongoing First 1,000 Days project ("The First 1000 Days", 2023) strives to collect children's language experiences during their first three years of life using a similar setup to that used in the Human Speechome project. Am. E. stands for American English, Au. E. stands for Australian English.

4.4 Conclusion

Driven by recent advances in artificial neural networks that learn from raw speech on the one side and in lightweight wearable recording devices on the other, our work sought to advance our understanding of how infants acquire language.

We used child-centered long-form recordings, not only to build a suite of automatic speech processing tools allowing us to analyze children's language environment, but also for simulating infant language acquisition. Contrary to previous modeling studies relying on critical simplifying assumptions regarding the learning material, we set ourselves the objective of fueling computational models of language acquisition with what young children truly hear. In doing so, we explored certain mechanisms infants may need to bring in to acquire their native language and showed how the ecological validity of the learning material could profoundly transform the learning outcomes developed by artificial language learners.

In questioning the generalizability of our measures, models, and theories to the real world, our work invites us to attribute greater importance to ecological validity and opens new opportunities to progress on some of the longstanding controversies in language acquisition.

Bibliography

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143 (cit. on p. 75).

Akhtar, N., & Gernsbacher, M. A. (2007). Joint attention and vocabulary development: a critical look. *Language and linguistics compass*, *1*(3), 195–207 (cit. on p. 45).

Al Futaisi, N., Zhang, Z., Cristia, A., Warlaumont, A., & Schuller, B. (2019). Vcmnet: weakly supervised learning for automatic infant vocalisation maturity analysis. *2019 International Conference on Multimodal Interaction*, 205–209 (cit. on p. 29).

Alishahi, A., Chrupała, G., Cristia, A., Dupoux, E., Higy, B., Lavechin, M., Räsänen, O., & Yu, C. (2021). Zr-2021vg: zero-resource speech challenge, visually-grounded language modelling track. *arXiv preprint arXiv:2107.06546* (cit. on p. 125).

Alishahi, A., & Fazly, A. (2010). Integrating syntactic knowledge into a model of cross-situational word learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32) (cit. on p. 47).

Ambridge, B., & Lieven, E. V. (2011). *Child language acquisition: contrasting theoretical approaches*. Cambridge University Press. (Cit. on pp. 2, 142).

Ambridge, B., & Rowland, C. F. (2013). Experimental methods in studying child language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 149–168 (cit. on pp. 42, 50, 73).

Anders, F., Hlawitschka, M., & Fuchs, M. (2020). Automatic classification of infant vocalization sequences with convolutional neural networks. *Speech Communication*, *119*, 36–45 (cit. on p. 29).

Anderson, J. R. (1975). Computer simulation of a language acquisition system: a first report (cit. on p. 47).

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common voice: a massively-multilingual speech corpus. *International Conference on Language Resources and Evaluation* (cit. on p. 11).

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The spot-the-word test: a robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology*, *32*(1), 55–65 (cit. on p. 51).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: a framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460 (cit. on p. 13).

Barker-Collo, S., Bartle, H., Clarke, A., van Toledo, A., Vykopal, H., & Willetts, A. (2008). Accuracy of the national adult reading test and spot the word estimates of premorbid intelligence in a non-clinical new zealand sample. *New Zealand Journal of Psychology*, *37*(3), 53–61 (cit. on p. 51).

Barlow, H. B., et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01), 217–233 (cit. on p. 50).

Bartz, C., Herold, T., Yang, H., & Meinel, C. (2017). Language identification using deep convolutional recurrent neural networks. *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part VI 24*, 880–889 (cit. on p. 37).

Bergelson, E. (2017). SEEDLingS HomeBank corpus. (Cit. on p. 9).

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: naturalistic language input to infants. *Developmental science*, *22*(1), e12715 (cit. on pp. xxvii, 7, 145).

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental science*, *22 1*, e12724 (cit. on pp. xxvii, 9, 145).

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C., Ramirez-Esparza, N., Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., et al. (2022). Everyday language input and production in 1001 children from 6 continents (cit. on p. 4).

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258 (cit. on pp. 42, 74).

Bergmann, C., Cristia, A., & Dupoux, E. (2016). Discriminability of sound contrasts in the face of speaker variation quantified. *CogSci* (cit. on p. 136).

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., et al. (2020). Wordseg: standardizing unsupervised word form segmentation from text. *Behavior research methods*, *52*, 264–278 (cit. on p. 47).

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: a survey. *Speech communication*, *56*, 85–100 (cit. on p. 14).

Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., & Hamam, H. (2022). Automatic speech recognition (ASR) systems for children: a systematic literature review. *Applied Sciences*, *12*(9), 4419 (cit. on p. 11).

Blandón, M. A. C., Cristia, A., & Räsänen, O. (2021). Evaluation of computational models of infant language development against robust empirical data from meta-analyses: what, why, and how? (Cit. on pp. 48, 49, 73).

Bosseler, A. N., Clarke, M., Tavabi, K., Larson, E. D., Hippe, D. S., Taulu, S., & Kuhl, P. K. (2021). Using magnetoencephalography to examine word recognition, lateralization, and future language skills in 14-month-old infants. *Developmental Cognitive Neuroscience*, *47*, 100901 (cit. on p. 49).

Brand, S., Mulder, K., ten Bosch, L., & Boves, L. (2021). Models of reaction times in auditory lexical decision. rtonset versus rtoffset (cit. on p. 74).

Bregman, A. S. (1994). *Auditory scene analysis: the perceptual organization of sound*. MIT press. (Cit. on p. 137).

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, *61*(1-2), 1–38 (cit. on p. 47).

Brookman, R., Kalashnikova, M., Conti, J., Xu Rattanasone, N., Grant, K.-A., Demuth, K., & Burnham, D. (2020). Depression and anxiety in the postnatal period: an examination of infants' home language environment, vocalizations, and expressive language abilities. *Child development*, *91*(6), e1211–e1230 (cit. on p. 79).

Bruner, J. (1985). Child's talk: learning to use language. *Child Language Teaching and Therapy*, *1*(1), 111–114 (cit. on p. 44).

Cao, X.-N., Dakhlia, C., Del Carmen, P., Jaouani, M.-A., Ould-Arbi, M., & Dupoux, E. (2018). Baby cloud, a technological platform for parents and researchers. *LREC 2018-11th edition of the Language Resources and Evaluation Conference* (cit. on pp. 4, 138).

Carbajal, M. J., Peperkamp, S., & Tsuji, S. (2021). A meta-analysis of infants' word-form recognition. *Infancy*, *26*(3), 369–387 (cit. on p. 42).

Cartmill, E. A., Armstrong III, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283 (cit. on p. 38).

Casillas, M., Brown, P., & Levinson, S. C. (2017). Casillas HomeBank corpus. (Cit. on p. 144).

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a tseltal mayan village. *Child Development*, *91*(5), 1819–1835 (cit. on p. 137).

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a papuan community. *Journal of Child Language*, *48*(4), 792–814 (cit. on p. 4).

Casillas, M., & Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology*, *5*(1) (cit. on pp. 4, 13).

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 1–12 (cit. on p. 49).

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, *10*(7), 335–344 (cit. on p. 47).

Choi, M., & Shukla, M. (2021). A new proposal for phoneme acquisition: computing speaker-specific distribution. *Brain Sciences*, *11*(2), 177 (cit. on p. 136).

Chomsky, N. (1957). Syntactic structures. Mouton de Gruyter. (Cit. on pp. 139, 142).

Chomsky, N. (1959). Review of B.F. Skinner, Verbal Behavior. *Language*, *35*, 26–58 (cit. on p. 2).

Chomsky, N., et al. (1980). On cognitive structures and their development: a reply to piaget. *Language and learning: the debate between Jean Piaget and Noam Chomsky*, 35–54 (cit. on p. 2).

Christiansen, M. H., Contreras Kallens, P., & Trecca, F. (2022). Toward a comparative approach to language acquisition. *Current Directions in Psychological Science*, *31*(2), 131–138 (cit. on p. 43).

Chrupała, G. (2022). Visually grounded models of spoken language: a survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, *73*, 673–707 (cit. on p. 47).

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160055 (cit. on p. 76).

Coen, M. (2006). Self-supervised acquisition of vowels in American English. *Association for the Advancement of Artificial Intelligence (AAAI)* (cit. on p. 102).

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., & Bapna, A. (2023). Fleurs: few-shot learning evaluation of universal representations of speech. *2022 IEEE Spoken Language Technology Workshop (SLT)*, 798–805 (cit. on p. 11).

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the language environment analysis system segmentation and metrics: a systematic review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105 (cit. on pp. 17, 18).

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: a time allocation study. *Child development*, *90*(3), 759–773 (cit. on p. 75).

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C. F., Räsänen, O. J., Bunce, J. P., & Bergelson, E. (2019). A thorough evaluation of the language environment analysis (lena) system. *Behavior Research Methods*, *53*, 467–486 (cit. on p. 18).

Cusack, R., McCuaig, O., & Linke, A. C. (2018). Methodological challenges in the comparison of infant fmri across age groups. *Developmental Cognitive Neuroscience*, *33*, 194–205 (cit. on p. 49).

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., Casillas, M., De Barbaro, K., Bang, J. Y., & Weisleder, A. (2020). Longform recordings of everyday life: ethics for best practices. *Behavior research methods*, *52*, 1951–1969 (cit. on p. 4).

Cychosz, M., & Cristia, A. (2021). Using big data from long-form recordings to study development and optimize societal impact (cit. on p. 13).

De Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134 (cit. on p. 102).

Défossez, A., Synnaeve, G., & Adi, Y. (2020). Real time speech enhancement in the waveform domain. *ArXiv*, *abs/2006.12847* (cit. on p. 141).

de Marcken, C. (1996). Unsupervised language acquisition. *ArXiv*, *cmp-lg/9611002* (cit. on p. 47).

de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., & Wisniewski, G. (2022). Probing phoneme, language and speaker information in unsupervised speech representations. *arXiv preprint arXiv:2203.16193* (cit. on p. 136).

de Seyssel, M., Lavechin, M., & Dupoux, E. (2022). Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language*, *49*(4), 714–740 (cit. on pp. xxiv, 46, 48, 74).

de Seyssel, M., Lavechin, M., Titeux, H., Thomas, A., Virlet, G., Revilla, A. S., Wisniewski, G., Ludusan, B., & Dupoux, E. (2023). Prosaudit, a prosodic benchmark for self-supervised speech models. *Interspeech* (cit. on p. 136).

Draghici, A., Abeßer, J., & Lukashevich, H. (2020). A study on spoken language identification using deep neural networks. *Proceedings of the 15th International Audio Mostly Conference*, 253–256 (cit. on p. 37).

Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T., de Seyssel, M., Roz'e, P., Rivière, M., Kharitonov, E., & Dupoux, E. (2021). The zero resource speech challenge 2021: spoken language modelling. *Interspeech* (cit. on pp. 50, 134).

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: a roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59 (cit. on pp. 3, 42, 45, 49, 52, 79, 125).

Eilers, R. E., Gavin, W., & Wilson, W. R. (1979). Linguistic experience and phonemic perception in infancy: a crosslinguistic study. *Child Development*, 14–18 (cit. on p. 42).

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211 (cit. on p. 47).

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a unified model of lexical and phonetic acquisition. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 184–193 (cit. on p. 75).

Erickson, L. C., & Newman, R. S. (2017). Influences of background noise on infants and children. *Current directions in psychological science*, *26*(5), 451–457 (cit. on p. 30).

Feldman, N., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do infants really learn phonetic categories? *Open Mind*, *5*, 113–131 (cit. on pp. 48, 75, 136).

Feldman, N., Griffiths, T., & Morgan, J. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31) (cit. on p. 52).

Frank, M. C. (2011). Computational models of early language acquisition. *Current Opinion in Neurobiology*, *21*(3), 381–386 (cit. on p. 41).

Ganek, H., & Eriks-Brophy, A. (2018). Language environment analysis (lena) system investigation of day long recordings in children: a literature review. *Journal of Communication Disorders*, *72*, 77–85 (cit. on p. 17).

Gasparini, L., Langus, A., Tsuji, S., & Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in infants' language discrimination abilities: a meta-analysis. *Cognition*, *213*, 104757 (cit. on p. 43).

Gautheron, L., Lavechin, M., Riad, R., Scaff, C., & Cristia, A. (2020). Longform recordings: opportunities and challenges. *LIFT 2020-2èmes journées scientifiques du Groupement de Recherche" Linguistique informatique, formelle et de terrain"*, 64–71 (cit. on p. 13).

Gilkerson, J., Coulter, K. K., & Richards, J. A. (2008). Transcriptional analyses of the LENA natural language corpus. *LENA Foundation* (cit. on pp. 15, 17).

Gilkerson, J., & Richards, J. A. (2008). The LENA natural language study. *Boulder, CO: LENA Foundation. Retrieved March*, *3*, 2009 (cit. on pp. 15, 17).

Gilkerson, J., & Richards, J. A. (2020). A guide to understanding the design and purpose of the LENA® system. *LENA Foundation: Boulder, CO* (cit. on p. 15).

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, *26*(2), 248–265 (cit. on p. 15).

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55 (cit. on p. 44).

Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign, and language: the coming of age of sign language and gesture studies. *Behavioral and brain sciences*, *40*, e46 (cit. on p. 1).

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological science*, *19*(5), 515–523 (cit. on p. 45).

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135 (cit. on p. 44).

Gottfried, T. L. (1984). Effects of consonant context on the perception of french vowels. *Journal of Phonetics*, *12*(2), 91–114 (cit. on p. 51).

Gregory, R. J. (2004). *Psychological testing: history, principles, and applications*. Pearson Education India. (Cit. on p. 49).

Guo, L. X., Pace, A., Masek, L. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2023). Cascades in language acquisition: re-thinking the linear model of development. *Advances in Child Development and Behavior*, *64*, 69–107 (cit. on p. 74).

Hallap, M., Dupoux, E., & Dunbar, E. (2022). Evaluating context-invariance in unsupervised speech representations. *arXiv preprint arXiv:2210.15775* (cit. on pp. 74, 133).

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children*. Paul H Brookes Publishing. (Cit. on p. 7).

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53–B64 (cit. on p. 44).

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., et al. (2017). Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551* (cit. on p. 47).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82–97 (cit. on p. 8).

Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269–286 (cit. on p. 42).

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97 (cit. on p. 1).

Hsu, W.-N., Tsai, Y.-H. H., Bolte, B., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: how much can a bad teacher benefit asr pre-training? *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6533–6537 (cit. on pp. 13, 125).

Hueber, T., Tatulli, E., Girin, L., & Schwartz, J.-L. (2020). Evaluating the potential gain of auditory and audiovisual speech-predictive coding using deep learning. *Neural Computation*, *32*(3), 596–625 (cit. on p. 50).

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). Babyberta: learning more grammar with small-scale child-directed language. *Proceedings of the 25th conference on computational natural language learning*, 624–646 (cit. on p. 126).

Ireton, H. (1992). Child development inventory. *Clinical Pediatrics* (cit. on p. 73).

Iverson, J. M. (2021). Developmental variability and developmental cascades: lessons from motor and language development in infancy. *Current Directions in Psychological Science*, *30*(3), 228–235 (cit. on p. 74).

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(3), 235–247 (cit. on p. 47).

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental science*, *13*(2), 339–345 (cit. on p. 43).

Jones, B., Johnson, M., & Frank, M. C. (2010). Learning words and their meanings from unsegmented child-directed speech. *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, 501–509 (cit. on p. 75).

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive psychology*, *98*, 1–21 (cit. on p. 37).

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, *29*(1), 1–23 (cit. on pp. 42, 74).

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of memory and language*, *32*(3), 402–420 (cit. on p. 42).

Jusczyk, P. W., Hirsh-Pasek, K., Nelson, D. G. K., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive psychology*, *24*(2), 252–293 (cit. on p. 42).

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, *277*(5334), 1984–1986 (cit. on p. 42).

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive psychology*, *39*(3-4), 159–207 (cit. on p. 42).

Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al. (2020). Libri-light: a benchmark for ASR with limited or no supervision. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7669–7673 (cit. on p. 125).

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in cognitive sciences*, *2*(10), 389–398 (cit. on p. 2).

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron*, *100*(2), 424–435 (cit. on p. 50).

Kharitonov, E., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T., Rivière, M., Mohamed, A.-r., Dupoux, E., & Hsu, W.-N. (2021). Text-free prosody-aware generative spoken language modeling. *ArXiv*, *abs/2109.03264* (cit. on p. 50).

Khorrami, K., Blandón, M. A. C., & Räsänen, O. (2023). Computational insights to acquisition of phonemes, words, and word meanings in early language: sequential or parallel acquisition? (Cit. on p. 75).

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, *22*(2), 154–169 (cit. on p. 43).

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735 (cit. on pp. 14, 38).

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42 (cit. on p. 44).

Kuhl, P. K. (2016). Language and the social brain: the power of surprise in science. *Scientists Making a Difference: One Hundred Eminent Behavioral and Brain Scientists Talk about their Most Important Contributions*, 206 (cit. on p. 21).

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000 (cit. on p. 101).

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, *9 2*, F13–F21 (cit. on p. 101).

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096–9101 (cit. on p. 45).

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608 (cit. on pp. 42, 44).

Lair, N., Colas, C., Portelas, R., Dussoux, J., Dominey, P. F., & Oudeyer, P. (2019). Language grounding through social interactions and curiosity-driven multi-goal learning. *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019* (cit. on p. 47).

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., et al. (2021). On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, *9*, 1336–1354 (cit. on p. 50).

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., & Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. *International Conference on Acoustics, Speech and Signal Processing ICASSP*, *1*, 2–5 (cit. on p. 16).

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *Interspeech* (cit. on pp. xxiv, 21, 27, 28, 141).

Lavechin, M., De Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., & Cristia, A. (2023). Statistical learning models of early phonetic acquisition struggle with child-centered audio data (cit. on p. 102).

Lavechin, M., de Seyssel, M., Gautheron, L., Dupoux, E., & Cristia, A. (2022). Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*, *8*, 389–407 (cit. on pp. 3, 48, 50, 79).

Lavechin, M., de Seyssel, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., & Dupoux, E. (2022). Can statistical learning bootstrap early language acquisition? A modeling investigation. *PsyArXiv preprint PsyArXiv:rx94d* (cit. on pp. 52, 125, 127, 136).

Lavechin, M., Métais, M., Titeux, H., Boissonnet, A., Copet, J., Rivière, M., Bergelson, E., Cristia, A., Dupoux, E., & Bredin, H. (2022). Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation. *arXiv preprint arXiv:2210.13248* (cit. on pp. 30, 141).

Lavechin, M., Sy, Y., Titeux, H., Blandón, M. A. C., Räsänen, O., Bredin, H., Dupoux, E., & Cristia, A. (2023). Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models. *Interspeech* (cit. on pp. 126, 133, 134).

Le Franc, A., Riebling, E., Karadayi, J., Wang, Y., Scaff, C., Metze, F., & Cristia, A. (2018). The aclew divime: an easy-to-use diarization tool. *INTERSPEECH*, 1383–1387 (cit. on p. 39).

Levy, E. S., & Strange, W. (2008a). Perception of french vowels by american english adults with and without french language experience. *Journal of phonetics*, *36*(1), 141–157 (cit. on p. 51).

Levy, E. S., & Strange, W. (2008b). Perception of French vowels by American English adults with and without French language experience. *Journal of phonetics*, *36*(1), 141–157 (cit. on p. 101).

Li, J. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(1) (cit. on p. 8).

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., Anastasopoulos, A., Mortensen, D. R., Neubig, G., Black, A. W., et al. (2020). Universal phone recognition with a multilingual allophone system. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8249–8253 (cit. on p. 28).

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: universal grammar and statistical learning. *Annu. Rev. Linguist.*, *1*(1), 333–353 (cit. on p. 43).

Liu, O., Tang, H., & Goldwater, S. (2023). Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces. *arXiv preprint arXiv:2305.12464* (cit. on p. 136).

Long, B., Goodin, S., Kachergis, G., Marchman, V. A., Radwan, S., Sparks, R., Xiang, V., Zhuang, C., Hsu, O., Newman, B., et al. (2022). The babyview camera: designing a new head-mounted camera to capture children's early social and visual environment (cit. on p. 137).

MacWhinney, B. (1996). The CHILDES system. *American Journal of Speech-Language Pathology*, *5*(1), 5–14 (cit. on pp. 14, 126).

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of child language*, *12*(2), 271–295 (cit. on p. 126).

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, *6*(5), 314–317 (cit. on p. 42).

Manesse, D., & Miniac, C. B.-d. (1981). Centre royaumont pour une science de l'homme. — théories du langage, théories de l'apprentissage : le débat entre jean plaget et noam chomsky (cit. on p. 2).

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive science*, *37*(1), 103–124 (cit. on p. 52).

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental science*, *11*(1), 122–134 (cit. on p. 44).

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111 (cit. on p. 101).

Mazuka, R., Cao, Y., Dupoux, E., & Christophe, A. (2011). The development of a phonological illusion: a cross-linguistic study with japanese and french infants. *Developmental science*, *14*(4), 693–699 (cit. on p. 42).

McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, *152*(6), 3819–3842 (cit. on pp. 48, 136).

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental psychology*, *54*(8), 1472 (cit. on p. 43).

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178 (cit. on p. 42).

Millet, J., Caucheteux, C., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., King, J.-R., et al. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, *35*, 33428–33443 (cit. on p. 49).

Millet, J., & Dunbar, E. (2020). Perceptimatic: a human speech perception benchmark for unsupervised subword modelling. *Interspeech* (cit. on p. 74).

Millet, J., & Dunbar, E. (2022). Do self-supervised speech models develop human-like perception biases? *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7591–7605 (cit. on p. 74).

Millet, J., Jurov, N., & Dunbar, E. (2019). Comparing unsupervised speech learning directly to human performance in speech perception. *CogSci 2019-41st Annual Meeting of Cognitive Science Society* (cit. on p. 74).

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive science*, *26*(4), 393–424 (cit. on p. 44).

Miyazawa, K., Kikuchi, H., & Mazuka, R. (2010). Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. *Interspeech* (cit. on p. 102).

Morgan, J. L., & Demuth, K. (2014). *Signal to syntax: bootstrapping from speech to grammar in early acquisition*. Psychology Press. (Cit. on p. 51).

Nenadić, F., Tucker, B. V., & Ten Bosch, L. (2022). Computational modeling of an auditory lexical decision experiment using diana. *Language and Speech*, 00238309221111752 (cit. on p. 74).

Newman, R. S., & Hussain, I. (2006). Changes in preference for infant-directed speech in low and moderate noise by 4.5-to 13-month-olds. *Infancy*, *10*(1), 61–76 (cit. on p. 136).

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (non) words,(non) words,(non) words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34 (cit. on p. 74).

Nguyen, T. S., Stueker, S., & Waibel, A. H. (2020). Super-human performance in online low-latency recognition of conversational speech. *Interspeech* (cit. on p. 8).

Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., & Dupoux, E. (2020). The zero resource speech benchmark 2021: metrics and baselines for unsupervised spoken language modeling. *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing* (cit. on pp. 50, 74, 125, 136).

Nikolaus, M., Alishahi, A., & Chrupała, G. (2022). Learning english with peppa pig. *Transactions of the Association for Computational Linguistics*, *10*, 922–936 (cit. on p. 47).

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., & Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354–13359 (cit. on p. 38).

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2020). What do bilingual infants actually hear? evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental science*, *23*(2), e12901 (cit. on p. 80).

Ota, C. L., & Austin, A. M. B. (2013). Training and mentoring: family child care providers' use of linguistic inputs in conversations with children. *Early Childhood Research Quarterly*, *28*(4), 972–983 (cit. on p. 15).

Oudeyer, P.-Y., Kachergis, G., & Schueller, W. (2019). Computational and robotic models of early language development: a review. (cit. on p. 47).

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210 (cit. on p. 11).

Piaget, J. (1935). *La naissance de l'intelligence chez l'enfant*. Delachaux et Niestlé Neuchatel-Paris. (Cit. on pp. 2, 139, 142).

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283 (cit. on p. 41).

Pinker, S. (1989). Learnability and cognition: the acquisition of argument structure (cit. on p. 44).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (cit. on pp. 8, 9, 13).

Räsänen, O., Doyle, G., & Frank, M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *Sixteenth Annual Conference of the International Speech Communication Association* (cit. on p. 48).

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*, 130–150 (cit. on pp. 47, 48).

Räsänen, O., & Khorrami, K. (2019). A computational model of early language acquisition from audiovisual experiences of young infants. *Interspeech* (cit. on p. 47).

Räsänen, O., Seshadri, S., Karadayi, J., Riebling, E., Bunce, J., Cristia, A., Metze, F., Casillas, M., Rosemberg, C., Bergelson, E., et al. (2019). Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech. *Speech Communication*, *113*, 63–80 (cit. on pp. 18, 27).

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2021). ALICE: an open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*, *53*, 818–835 (cit. on pp. xxiv, 20, 27, 28, 37).

Reh, R. K., Hensch, T. K., & Werker, J. F. (2021). Distributional learning of speech sound categories is gated by sensitive periods. *Cognition*, *213*, 104653 (cit. on p. 101).

Reitmaier, T., Wallington, E., Kalarikalayil Raju, D., Klejch, O., Pearson, J., Jones, M., Bell, P., & Robinson, S. (2022). Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–17 (cit. on p. 14).

Richards, J. A., Gilkerson, J., Paul, T., & Xu, D. (2008). *The lenatm automatic vocalization assessment* (tech. rep.). Technical Report LTR-08-1). LENA Foundation. (Cit. on p. 38).

Roopnarine, J. L., Fouts, H. N., Lamb, M. E., & Lewis-Elligan, T. Y. (2005). Mothers' and fathers' behaviors toward their 3-to 4-month-old infants in lower, middle, and upper socioeconomic african american families. *Developmental psychology*, *41*(5), 723 (cit. on p. 7).

Rose, Y., & MacWhinney, B. (2014). The PhonBank project. *The Oxford Handbook of Corpus Phonology* (cit. on p. 14).

Rowe, M. L., Pan, B. A., & Ayoub, C. (2005). Predictors of variation in maternal talk to children: a longitudinal study of low-income families. *Parenting: Science and Practice*, *5*(3), 259–283 (cit. on p. 43).

Rowland, C. (2013). *Understanding child language acquisition*. Routledge. (Cit. on p. 2).

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., et al. (2006). The human speechome project. *Symbol Grounding and Beyond: Third International Workshop on the Emergence and Evolution of Linguistic Communication, EELC 2006, Rome, Italy, September 30–October 1, 2006. Proceedings*, 192–196 (cit. on pp. xxvii, 145).

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs (cit. on p. 47).

Sacks, C., Shay, S., Repplinger, L., Leffel, K. R., Sapolich, S. G., Suskind, E., Tannenbaum, S., & Suskind, D. (2014). Pilot testing of a parent-directed intervention (project ASPIRE) for underserved children who are deaf or hard of hearing. *Child Language Teaching and Therapy*, *30*(1), 91–102 (cit. on p. 15).

Saffran, J. R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114 (cit. on p. 44).

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928 (cit. on p. 44).

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual review of psychology*, *69*, 181–203 (cit. on pp. 44, 73).

Saffran, J. R., & Thiessen, E. D. (2008). Domain-general learning capacities (cit. on p. 2).

Scaff, C. (2019). *Beyond WEIRD: an interdisciplinary approach to language acquisition* (Doctoral dissertation). PhD thesis. (Cit. on pp. 14, 43).

Scharenborg, O., Ernestus, M., & Wan, V. (2007). Segmentation of speech: child's play? *Interspeech* (cit. on p. 48).

Schatz, T., Feldman, N., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories: insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, *118*(7), e2001844118 (cit. on pp. 47, 102, 103, 136, 142).

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., et al. (2017). The interspeech 2017 computational paralinguistics challenge: addressee, cold & snoring. *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 3442–3446 (cit. on pp. 20, 37).

Schwab, J. F., & Lew-Williams, C. (2016). Language learning, socioeconomic status, and child-directed speech. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(4), 264–275 (cit. on p. 43).

Seidenberg, M. S. (1997). Language acquisition and use: learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603 (cit. on p. 44).

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker variation aids young infants' phonotactic learning. *Language Learning and Development*, *10*(4), 297–307 (cit. on p. 136).

Seidl, A., Tincoff, R., Baker, C., & Cristia, A. (2015). Why the body comes first: effects of experimenter touch on infants' word finding. *Developmental science*, *18*(1), 155–164 (cit. on p. 138).

Seshadri, S., & Räsänen, O. (2019). Sylnet: an adaptable end-to-end syllable count estimator for speech. *IEEE Signal Processing Letters*, *26*(9), 1359–1363 (cit. on p. 28).

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. M. (2023). Diversity and representation in infant research: barriers and bridges toward a globalized science of infant development. *Infancy* (cit. on p. 38).

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts. (Cit. on p. 44).

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word–referent learning. *Trends in Cognitive Sciences*, *18*, 251–258 (cit. on p. 44).

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568 (cit. on p. 44).

Soderstrom, M. (2007). Beyond babytalk: re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532 (cit. on p. 77).

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing a cross-cultural annotation system and metacorpus for studying infants' real world language experience. *Collabra: Psychology*, *7*(1), 23445 (cit. on p. 9).

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, *102*(33), 11629–11634 (cit. on p. 44).

Stärk, K., Kidd, E., & Frost, R. L. (2022). Word segmentation cues in german child-directed speech: a corpus analysis. *Language and Speech*, *65*(1), 3–27 (cit. on pp. 47, 48).

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). Saycam: a large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, *5*, 20–29 (cit. on pp. xxvii, 145).

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological science*, *10*(2), 172–175 (cit. on p. 42).

Tomasello, M. (1992). The social bases of language acquisition. *Social development*, *1*(1), 67–87 (cit. on p. 44).

Tomasello, M. (2005). *Constructing a language: a usage-based theory of language acquisition*. Harvard university press. (Cit. on p. 142).

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child development*, 1454–1463 (cit. on p. 2).

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child development*, 466–472 (cit. on p. 101).

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: a meta-analysis. *Developmental psychobiology*, *56*(2), 179–191 (cit. on pp. 43, 44, 73).

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460 (cit. on p. 41).

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fmri studies. *Communications Biology*, *1*(1), 62 (cit. on p. 49).

Twaddell, W. F. (1935). On defining the phoneme. *Language*, *11*(1), 5–62 (cit. on p. 48).

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*, 13273–13278 (cit. on pp. 47, 48, 102).

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: an online repository of daylong child-centered audio recordings. *Seminars in speech and language*, *37*(02), 128–142 (cit. on p. 14).

van Niekerk, B., Nortje, L., Baas, M., & Kamper, H. (2021). Analyzing speaker information in self-supervised models to improve zero-resource speech processing. *ArXiv*, *abs/2108.00917* (cit. on p. 136).

Vygotsky, L. (1962). *Thought and language*. Mit Press Cambridge, MA. (Cit. on pp. 2, 44).

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J. M., & Dupoux, E. (2021). VoxPopuli: a large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *Annual Meeting of the Association for Computational Linguistics* (cit. on p. 11).

Warlaumont, A., Pretzer, G., Mendoza, S., & Walle, E. A. (2016). Warlaumont HomeBank corpus. (Cit. on p. 144).

Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2010). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of autism and developmental disorders*, *40*, 555–569 (cit. on p. 14).

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17–60). CRC Press. (Cit. on pp. 50, 125).

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., & Zhuang, C. (2023). Call for papers–the babylm challenge: sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796* (cit. on p. 125).

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). Blimp: the benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392 (cit. on p. 125).

Waseem, Z., Lulz, S., Bingel, J., & Augenstein, I. (2021). Disembodied machine learning: on the illusion of objectivity in NLP. *arXiv preprint arXiv:2101.11974* (cit. on p. 38).

Weil, L. W., & Middleton, L. (2010). Use of the LENA tool to evaluate the effectiveness of a parent intervention program. *Perspectives on Language Learning and Education*, *17*(3), 108–111 (cit. on p. 15).

Weisleder, A., & Fernald, A. (2013). Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychological science*, *24*(11), 2143–2152 (cit. on p. 79).

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child development*, 349–355 (cit. on p. 101).

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, *7*(1), 49–63 (cit. on pp. 42, 44).

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256* (cit. on p. 8).

Xu, D., Yapanel, U., & Gray, S. (2008). Reliability of the LENA™ language environment analysis system in young children's natural language home environment (technical report LTR-05-2). (Cit. on pp. 15, 17).

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., & Hansen, J. (2008). Signal processing for young child speech language development. *First Workshop on Child, Computer and Interaction* (cit. on pp. 15, 17).

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624 (cit. on p. 49).

Yeni-Komshian, G. H., Kavanagh, J. F., & Ferguson, C. A. (2014). *Child phonology: volume 1, production* (Vol. 1). Academic Press. (Cit. on pp. 42, 43).

Yu, C., & Ballard, D. H. (2003). A computational model of embodied language learning. *Computer Science Deptartment, Un. of Rochester, Rochester, New York*, *32* (cit. on p. 47).

Yuspeh, R. L., & Vanderploeg, R. D. (2000). Spot-the-word: a measure for estimating premorbid intellectual functioning. *Archives of clinical neuropsychology*, *15*(4), 319–326 (cit. on p. 51).

Zhang, Y., Chen, C.-h., & Yu, C. (2019). Mechanisms of cross-situational learning: behavioral and computational evidence. *Advances in child development and behavior*, *56*, 37–63 (cit. on p. 43).

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., et al. (2023). Google USM: scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037* (cit. on p. 13).

Zhang, Z., Cristia, A., Warlaumont, A., & Schuller, B. (2018). Automated classification of children's linguistic versus non-linguistic vocalisations (cit. on p. 29).

Webpages

The ACLEW project. (2023, April 1). https://sites.google.com/view/aclewdid/home. (Cit. on p. 18)

The First 1000 Days. (2023, June 27). https://wellcomeleap.org/1kd/. (Cit. on pp. xxvii, 145)

Frank, M. C. (2023, May 11). "*Psychological plausibility*" *considered harmful*. http://babieslearninglanguage.blogspot.com/2014/02/psychological-plausibility-considered.html. (Cit. on p. 48)

LENA 15^{th} *birthday*. (2023, April 1). https://www.lena.org/15-years. (Cit. on p. 19)

LENA release notes. (2023, April 1). https://cdn.shopify.com/s/files/1/0596/9601/files/ReleaseNotes.html?12770507767426244063. (Cit. on p. 20)

LENA shop. (2023, June 21). https://shop.lena.org/products/lena-device. (Cit. on p. 4)

Voice type classifier: follow-up analysis. (2023, May 4). https://github.com/MarvinLvn/voice-type-classifier/blob/new_model/docs/evaluations.md. (Cit. on p. 21)

A

Appendix: *Realistic and broad-scope learning simulations:first results and challenges*

Journal of Child Language (2023), 1–24
doi:10.1017/S0305000923000272

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Realistic and broad-scope learning simulations: first results and challenges

Maureen de SEYSSEL[1,2,†], Marvin LAVECHIN[1,3,†] and Emmanuel DUPOUX[1,3]

[1]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Études Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[2]Laboratoire de Linguistique Formelle, Université Paris Cité, CNRS, Paris, France
[3]Meta AI Research, Paris, France
Corresponding authors: Maureen de Seyssel and Marvin Lavechin; Emails: maureen.deseyssel@gmail.com; marvinlavechin@gmail.com

(Received 04 October 2022; revised 24 April 2023; accepted 04 April 2023)

Abstract
There is a current 'theory crisis' in language acquisition research, resulting from fragmentation both at the level of the approaches and the linguistic level studied. We identify a need for integrative approaches that go beyond these limitations, and propose to analyse the strengths and weaknesses of current theoretical approaches of language acquisition. In particular, we advocate that language learning simulations, if they integrate realistic input and multiple levels of language, have the potential to contribute significantly to our understanding of language acquisition. We then review recent results obtained through such language learning simulations. Finally, we propose some guidelines for the community to build better simulations.

Keywords: Language acquisition; computational modelling; phonetic learning; word learning; phonetic categories

What is needed and why?

Theory in crisis

The field of language acquisition is prolific, with an extensive range of high-quality research published every year. However, there has been surprisingly slow progress in solving some long-standing controversies regarding the basic mechanisms that underlie language acquisition. For instance, do infants learn language primarily from extracting statistics over speech inputs (Romberg & Saffran, 2010; Saffran & Kirkham, 2018), from examining cross-situational correlations over multisensory inputs (Smith & Yu, 2008; Suanda, Mugwanya & Namy, 2014; Yu & Smith, 2017; Zhang, Chen & Yu, 2019), or by relying on social interactions and feedback (Tomasello, 2003; Tsuji, Cristia & Dupoux, 2021; Yu & Ballard, 2007)? Do they learn by leveraging discrete linguistic categories or

†M.S and M.L. contributed equally to this work. Authorship order was decided by a coin flip.

continuous sensory representations (Kuhl et al., 2008; McMurray, 2021)? Do they rely on language-specific or domain-general learning mechanisms (Elman, Bates & Johnson, 1996; Karmiloff-Smith, 1994; Pinker, 1994)? Such a lack of headway may be due in part to the 'replication crisis': the experimental study of human cognition in general and infant cognition, in particular, is inherently noisy and difficult (Frank, Bergelson, Bergmann, Cristia, Floccia, Gervain, Hamlin, Hannon, Kline & Levelt, 2017), slowing down cumulative progress. Here, we explore the possibility that there is, in addition, a 'theory crisis'. To say it bluntly, perhaps, current theories have shortcomings that prevent us from even finding the right experimental setup to make progress on basic questions about learning mechanisms.

Several papers have already been devoted to the theory crisis in psychology in general; psychological theories have been claimed to be mere statistical model fitting (Fried, 2020), too descriptive or fragmented (Muthukrishna & Henrich, 2019), or to not contribute in cumulative theory building (McPhetres et al., 2021). In developmental psychology, Kachergis, Marchman, and Frank (2021) called for a 'standard model' that would allow integration of results in a cumulative fashion. In this paper, we explore the possibility proposed in Dupoux (2018) that recent advances in machine learning could help address the theory crisis through systems that realistically simulate how infants learn language in their natural environment. Such learning simulations are computer models that would ideally learn from similar inputs as the ones available to infants (raw sensory data), and reproduce the broad spectrum of outcome measures as obtained in laboratory experiments or corpus studies. To the extent that these new computer models are powerful enough to address the complexity and variability of data available to infants during language development, they could help us make progress in some of the aforementioned controversies. At best, such learning simulations can provide proof of principle that a given hypothesis (e.g., the statistical learning hypothesis) can account for learning outcomes as observed in infants. In addition, they can help us go beyond said long-standing controversies by providing new insights into the learning process and a wealth of associated quantitative predictions.

In this paper, we first discuss how these new types of learning simulations are complementary to more familiar theoretical approaches in cognitive development and argue that they provide one step towards the needed cumulative integrative theories or standard models. We then present STELA, a recent learning simulation implementing the hypothesis that infants are statistical learners, and show how it provides insights into some long-standing controversies.

Varieties of theories in language acquisition

The theoretical landscape of language development is vast and complex. Even if one focuses on early language development, there are wild varieties of theoretical approaches that differ not only in scope (the range of phenomena they cover) but also in style (verbal, statistical, formal, computational). Here, far from making a comprehensive survey of these approaches, we attempt to classify them into types and sort them along dimensions that outline their respective strengths and weaknesses with regard to addressing basic questions/controversies about learning mechanisms. Familiar types are verbal frameworks (among others: The competition model: MacWhinney & MacWhinney, 1987; WRAPSA: Jusczyk, 1993; Usage-based theory: Tomasello, 2005; NML-e: Kuhl et al., 2008; PRIMIR: Werker & Curtin, 2005), which weave a narrative around a large body of experimental research using verbally defined concepts, sometimes complemented by

box-and-arrow schemas (e.g., the ScALA framework from Tsuji et al., 2021). Correlational approaches (e.g., Fernald, Marchman & Weisleder, 2013; Hart & Risley, 1995; Swingley & Humphrey, 2018) aim to identify the main variables that predict language development outcomes through statistical models. Formal models (e.g., Jain, Osherson, Royer & Sharma, 1999; Tesar & Smolensky, 2000) and computational models (e.g., Brent, 1997) aim to study how algorithms can learn language through mathematical proofs or empirical study of the learning outcomes. All theoretical approaches of early language development recognise that infants receive inputs from their environment, and have a learning mechanism, which produces a linguistic competence that can be accessed through outcome measures. The differences between these theoretical approaches lie in the simplifying assumptions and degree of specifications they make about inputs, learning mechanisms and outcome measures. We distinguish four dimensions or axes to sort these theoretical approaches: Causal versus Correlational, Quantitative versus Qualitative, Realistic versus Abstract, and Broad Scope versus Narrow Scope.

Causal/Correlational

A theory is causal when it provides a specification/implementation of the learning mechanism underlying language acquisition; it is correlational when it focuses on the input/outcome relationship without specifying a learning mechanism. A correlational model can outline the important factors that drive learning and therefore provide insights into the development of learning mechanisms. However, only a causal model can provide proof of principle that a postulated learning mechanism is sufficient to reproduce a developmental outcome given an input. As a result, to the extent that they can be effectively implemented, causal models are better positioned to resolve disagreements about learning mechanisms than correlational models.

Quantitative/Qualitative

A theory is quantitative if it can produce numerical outcomes that can be compared to human performance. It is qualitative when it produces predictions about the possible presence of a significant effect without a numerical prediction about its strength. Qualitative models are useful to inspire novel experimental paradigms, and provide insights about learning mechanisms, but are hard to refute and difficult to compare to one another. Quantitative theories make very precise predictions and can be compared to one another by computing the degree of fit of their predictions against some observed outcome. As a result, they are better positioned to solve disagreements about learning mechanisms than qualitative theories.

Realistic/Abstract

A theory is realistic when its model of the environment is as close as possible to the actual sensory/motor environment of the child. It is abstract when the environment is specified through synthetic data, or human/categorical annotations of observed environments (e.g., textual transcriptions). Abstract theories are useful because they enable a high degree of control and interpretability and provide insights into what type of input information can yield particular outcomes. However, they cannot prove that their conclusions apply to real-world data as perceived by infants and are therefore not very informative when it comes to solving long-standing controversies. Realistic theories, in

contrast, to the extent that they can be effectively implemented, are better positioned: because they directly reproduce the learning outcomes associated with a given input and learning mechanism.

Broad/Narrow Scope

A theory has a broad scope if it encompasses not one single linguistic level (phonetic, morphological, syntactic, semantic, etc.) or phenomenon but several at once. Narrow Scope theories are useful in focusing on learning specific representations, assuming all other representations are fixed. However, many controversies about learning mechanisms arise because of co-dependencies between linguistic levels, making it problematic to assume all levels are fixed except one. Being able to account for how infants can learn jointly all of these levels is at the heart of solving so-called 'bootstrapping' problems that are integral to language learning.

In Table 1, we position some familiar theoretical approaches in terms of these four axes. This characterisation may seem overly simplistic or reductionist, but we hope it will help outline the specific contribution of learning simulations. Verbal frameworks typically have a broad scope and embrace the complexities of the child's real environment. They are causal to the extent that they mention specific learning principles but are not on the quantitative side. They are still the single most influential theoretical approach for infant language learning, providing insight into large quantities of experimental results. However, they resist empirical refutations or amendments because of their qualitative nature. Correlational models are on the quantitative side and integrate many variables and levels. When informed by a corpus of infant/caretaker interactions, they can reveal the relationships between input quality, quantity, and language outcome (Fernald et al., 2013; Hart & Risley, 1995). However, because they are not causal and rely on abstract variables derived from the input, they cannot directly speak to learning mechanisms. Computational/formal models (henceforth called learning simulations*)* are both causal and quantitative, but their ability to significantly impact controversies about learning mechanisms depends on the breadth of their scope and their degree of realism or abstraction. We discuss such models in more detail in the next section.

A brief history of learning simulations

For a long time, scientists with various backgrounds, from formal linguistics to developmental psychology and artificial intelligence, have contemplated the possibility of building mathematical models or computer simulations of language learning in infants. The hope was that building a simulated ersatz of the infant would reveal the formal conditions

Table 1. Four dimensions along which theoretical approaches of language acquisition can be sorted

Properties	Verbal Framework	Correlation Model	Learning Simulation
Causal	✗ / ✓	✗	✓
Quantitative	✗	✓	✓
Realistic	✗	✗	✗ ↔ ✓
Broad Scope	✗	✓	✗ ↔ ✓

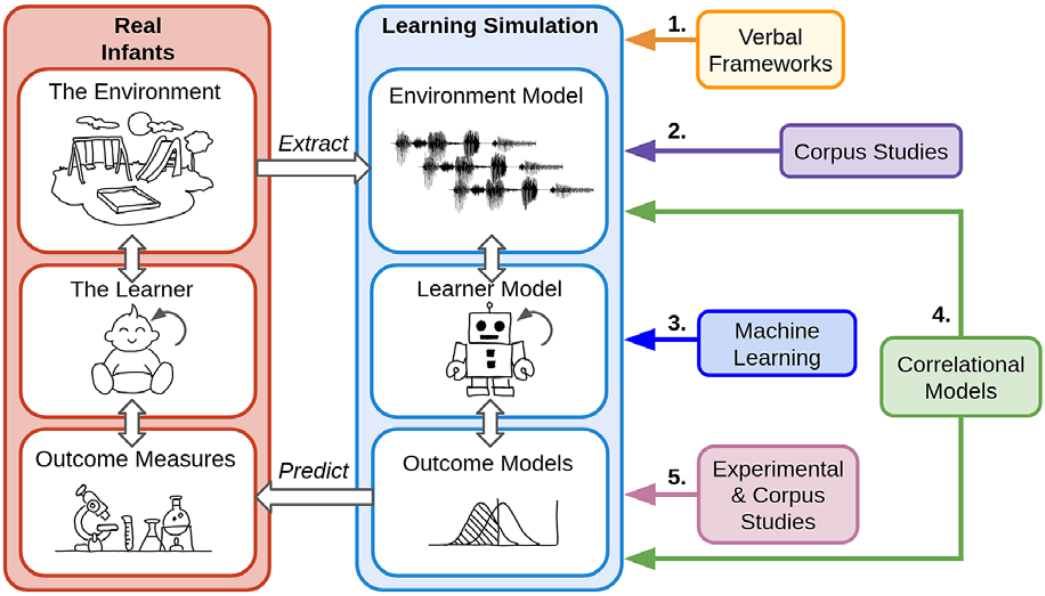

Figure 1. General outline of a realistic learning simulation (centre) in relation to real infants (left) and traditional theoretical approaches (right). 1. Verbal frameworks inspire and help set up the entire language learning simulation by describing the environment, learner, and outcome models; 2. Corpus studies of children's input help us build realistic models of the environment. In the best case, the model of the environment is a subset of a real environment, obtained through child-centred long-form recordings, for instance; 3. Machine learning provides effective artificial language learners. The learner model is relatively unconstrained as learning mechanisms used by the real learner (i.e., infants) remain largely unobservable; 4. Correlational models describe how the input should relate to the outcome measures; 5. Experimental and corpus studies of children's outcomes show how we can evaluate learning outcomes of the artificial learner. The real versus predicted outcome measures allow us to compare humans to machines and provide new predictions for correlational models that relate input to outcomes in infants.

for learning (Pinker, 1979), would allow us to better formulate hypotheses about how infants actually learn (Frank, 2011; Meltzoff, Kuhl, Movellan & Sejnowski, 2009) or would yield machines that learn in a graceful and robust fashion (Turing, 1950). Here again, the diversity of the proposed models is too large to be reviewed (see Dupoux, 2018, for an attempt). Instead, we classify the approaches based on the dimensions which we claim are central to answering key questions about learning mechanism: realism and scope.

As illustrated in Figure 1, all learning simulations consist of three components: a model of the environment, a model of the learner, and a model of the outcome measure. The model of the environment specifies the type of inputs/interactions available to the learner. The learner updates itself using a learning algorithm based on its interaction with the environment. The outcome measures of the learner are measured after exposure to speech. Where learning simulations differ is how they implement these three components.

Focusing on AI-inspired models, the most visible trend historically has been on how to implement the learner. Early models (e.g., Anderson, 1975; Kelley, 1967) were rule-based. The second phase was probabilistic models (e.g., Brent, 1996; de Marcken, 1996), followed by connectionist and deep learning models (Brown et al., 2020; Elman, 1990), each phase replacing hand-wired components with more and more powerful learning systems. As far as we are concerned, the way in which the learner is implemented is irrelevant. What counts is whether the learning mechanism actually reproduces the learning outcome or

not, given infants' input[1]. More relevant to our argument, another trend can be seen regarding the model of the environment, moving from synthetic data (e.g., Elman, 1990; Vallabha, McClelland, Pons, Werker & Amano, 2007) to transcribed corpora (e.g., Bernard et al., 2020) and, more recently, to raw audio and images or video recordings (Räsänen & Khorrami, 2019; Schatz, Feldman, Goldwater, Cao & Dupoux, 2021). Finally, the first models were focused on learning a single linguistic level (e.g., phonetic categories: Vallabha et al., 2007; word forms: Brent, 1999; word meanings: Roy & Pentland, 2002; syntax: Pearl & Sprouse, 2013), and more recent approaches would learn several levels jointly (phonemes and words: Elsner, Goldwater & Eisenstein, 2012; syntax and semantics: Abend, Kwiatkowski, Smith, Goldwater, Steedman, 2017; phonetics, words and syntax: Nguyen et al., 2020).

In other words, thanks to recent progress in machine learning and AI (Bommasani et al., 2021), learning models that are simultaneously of broad scope and able to ingest realistic data are around the corner. Obviously, a complete model that would feature maximal scope (integrating all relevant input and output modalities for language and communication) and maximal realism (using sensory data indistinguishable from what infants experience) is still out of reach. In the next section, we examine STELA, a recently proposed model (Lavechin, de Seyssel et al., 2022c) and argue that even though it is limited both on scope and realism, this work can help us make nontrivial progress on some of the long-standing controversies regarding language learning mechanisms.

Before moving on, let us clarify that we are not claiming that broad-scope realistic simulations are the only valuable approach. Narrow-scope abstract models still have valuable contributions to make (e.g., Frank, Goodman & Tenenbaum, 2009; Kachergis et al., 2021). First, contrary to many realistic and broad-scope models, abstract and narrow models are interpretable and therefore allow building bridges with verbal frameworks. They are also more tractable and can be easily modified and experimented on in a way which is more difficult with larger models. Finally, one can view abstract learning simulations as "control" experiments: by comparing an abstract and a realistic learning simulation implementing a similar learning mechanism, we can gain knowledge on the role of specific abstractions made by the learner.

What has been achieved so far?

Among the competing hypotheses regarding the learning mechanisms that underlie early language learning, the one that seems the most natural to approach with learning simulations is the statistical learning hypothesis (Pelucchi, Hay & Saffran, 2009; Saffran, Aslin & Newport, 1996). It posits that infants learn at least some linguistic levels (phonetic, lexical and morphosyntactic) through a statistical or distributional analysis of their language inputs. The idea has a long history (Rumelhart, McClelland & MacWhinney, 1987; Skinner, 1957) and has generated many controversies (Chomsky, 2013; Fodor & Pylyshyn, 1988) and mathematical investigation (Gold, 1967; Jain et al., 1999). But it is also the simplest hypothesis to implement in a learning simulation. If one equates language input to the auditory modality, the corresponding learning simulation would simplify the environment to audio recordings, and the learner to a probabilistic model

[1]Many developmental scientists worry about the so-called 'psychological plausibility' of these various kinds of models. Following Frank (2014), we believe that issues of plausibility have either to be formulated as outcome measures that the model should reproduce, or should be disregarded.

that accumulates statistics paying no attention to other modalities or context, nor interacting with its environment.

Here, we present recent work on simulating a statistical learner for language acquisition (Lavechin et al., 2022b; Lavechin, de Seyssel et al., 2022c). We present the simplifying assumptions made in these simulations and reflect on how simulated learners compare to infants. Then, we go over different use cases of such a simulation by showing how some of the skills the simulated learner has acquired through exposure can help shed light on some long-standing controversies in our understanding of language acquisition in infants.

We focus on a high-level description of this simulation as we believe it makes it easier to appreciate its lessons. However, readers interested in the technical details can refer to the original paper (Lavechin, de Seyssel et al., 2022c). We will also list specific research use cases that the framework helped deepen. By doing so, we illustrate concretely how such realistic learning simulations can help future research, both in terms of proof of feasibility and inspiration for research.

Introducing STELA

Lavechin, de Seyssel et al. (2022c) introduced STELA (STatistical learning of Early Language Acquisition), a language learning simulation that tackles the problem of discovering structure in the continuous, untranscribed, and unsegmented raw audio signal. As said above, the scope of this simulation is restricted to the statistical learning hypothesis, where infants learn passively and uniquely by extracting statistical cues from what they hear (see Table 2). In this section, we present the model of the environment, the model of the learner, and the model of the outcome measures used in STELA.

The environment

STELA specifies the environment as raw audio speech recordings. For this to remain relevant, we need to restrict the quantity of speech within a plausible range of data. Current estimates of cumulative speech experiences by one year of age vary from around 60 hours (Cristia, Dupoux, Gurven & Stieglitz, 2019) to approximately 1,000 hours (Cristia, 2022). In STELA, the data comes either from open-source audiobooks with quantities varying from 50 to 3,200 hours covering the observed range. Admittedly, the infant's language environment is different from audiobooks. On the one hand,

Table 2. Non-exhaustive list of language learning assumptions for infants and whether they are included within the STELA simulation

Assumption	STELA
Infants are statistical learners (Bulf, Johnson & Valenza, 2011; Romberg & Saffran, 2010; Saffran et al., 1996)	✓
Quantity of speech input predicts language outcome (Newman, Rowe & Ratner, 2016)	✓
Modalities other than speech can be useful in language learning (Abu-Zhaya, Seidl, Tincoff & Cristia, 2017; Seidl, Tincoff, Baker & Cristia, 2015).	×
Infants learn by **interacting** with peers – reinforcement learning (Kuhl, Tsao & Liu, 2003; Nelson, 2007; Snow, 1989; Yu, Ballard & Aslin, 2005)	×

audiobooks contain clearly articulated speech (read speech) and relatively good audio conditions, potentially facilitating learning for the model compared to the spontaneous and noisy speech available to infants (see Lavechin et al., 2022b). On the other hand, audiobooks may use larger vocabularies and more complex sentences than infants' input, potentially putting the model in a more challenging situation than infants (Gleitman, Newport & Gleitman, 1984). Nevertheless, this type of input is in the range of what infants could plausibly hear or overhear and is relatively easier to access in large quantities across languages than long-form recordings. Therefore, they are a good starting point, offering controlled conditions and replicability for the deployment and analysis of such simulations. Long-form recordings represent the extreme in realism that can be achieved in such simulations, but they are less accessible than audiobooks due to privacy concerns (Lavechin, de Seyssel, Gautheron, Dupoux & Cristia, 2022a).

The learner

Elman (1990) was perhaps the first to introduce a practical implementation of a system that learns non-trivial linguistic representations by extracting regularities from language inputs: a simple recurrent neural network trained to predict future words or characters based on past ones. Since then, this idea has been expanded with more complex and larger neural networks trained on increasingly larger datasets. The resulting so-called "language models" can be viewed as models of the probability distribution of sentences and have been shown to generalise beyond the sentences in the training set (Baroni, 2020), reaching near human performances on many language tasks (Liu, He, Chen & Gao, 2019). One major limitation of these models – as models of the infant learner – is that they only take as input words or characters, which are not entities accessible to a learning infant. However, recent breakthroughs in representation learning have made it possible to expand these models to work with raw audio inputs (Borsos et al., 2022; Dunbar et al., 2021; Lakhotia et al., 2021). In a nutshell, these so-called 'Generative Spoken Language Models' replace text with their own discrete representations learnt from the audio and learn a probabilistic model of speech directly from raw inputs.

In Figure 2a, we present the model used in STELA, which has been selected from the class of Generative Spoken Language Models (Dunbar et al., 2021) for its simplicity. From a high-level perspective, the learner can be described as the combination of two components, which are named according to the current practices in machine learning 1) an '*acoustic model*' and 2) a '*language model*'[2]. The acoustic model is fed with raw, continuous waveforms and trained using a form of predictive coding. It learns a vector representation for each slice of 10ms of signal by attempting to predict each of the twelve upcoming slices based on past ones, yielding a prediction over a 120ms time window. An exciting outcome of such a learning procedure is that the model learns representations that successfully abstract away from acoustic details and encode phonetic information. In STELA, we discretise these representations using clustering, yielding a discrete code each 10ms, which is passed onto the language model. This model is similar to Elman's

[2]Although the term 'language model' can sound counterintuitive in the context of phonological and lexical acquisition, as no language-related or language-specific heuristics are integrated into the model, which learns on its own to discover structures in the speech input, we view it from the machine learning point of view, where a language model is simply an algorithm which learns to predict, from a sequential input, the next representation (let it be text, speech or other) based on the previous representations.

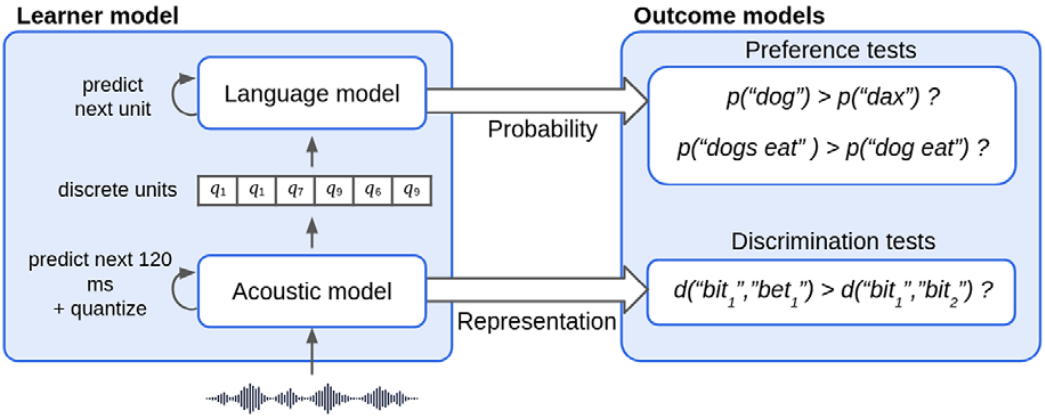

Figure 2. Overview of the STELA learner and outcome measures. a. (left): model of the learner; b. (right): add-on models for two types of outcome measures.

recurrent language model, only using an improved architecture (LSTMs) and more parameters. This model is trained to predict the next code based on past ones. Because the model's output is not a single code, but a probability distribution over all the discrete codes, one can compute the probability of an utterance as the product of the conditional probabilities of each successive code (see Appendix A).

The outcome measures

Several types of outcome measures are used in infant development. Some are provided by caregivers (like the Child Development Inventory, or CDI: Fenson, 2007), who assess whether a word is known or produced by the child, some are linked to the production of the child as attested through transcription of naturalistic corpora (mean length of utterance such as used in Miller & Chapman, 1981 for instance), and some are obtained via in-lab experiments. Here we concentrate on the last type of measure. In principle, a maximally broad language learning simulation would include all linguistic and non-linguistic components (attention, memory, eye movement, etc.) and the artificial learner could just be virtually seated in a virtual lab and be subjected to the same experiments as real babies (Leibo et al., 2018). Here, STELA only simulates a subpart of infants' linguistic competence and therefore has to specify a special add-on module to generate the equivalent of experimental outcome measures. Fortunately, experimental paradigms in infants are relatively simple and can be sorted into two main types: discrimination experiments and preference experiments[3], yielding two types of add-on modules.

Discrimination experiments can vary in how they are conducted in the lab (ABX, AXB, AX, etc.). Still, they all rely on the ability of the learner to compute a perceptual distance between two stimuli (such as 'bit' versus 'bet'). An add-on for ABX discrimination will just need to (a) extract a representation of a stimulus from the learner (typically the activation pattern of some layer) and (b) compute a distance over two representations (typically, the normalised dot product, or the angle between the vectors). In STELA (Lavechin, de

[3]This is a non-exhaustive list. Some experiments use a more complex design where infants are familiarised to some materials (for instance, an artificial language) and then tested using preference or discrimination metrics. This would require the learner to memorise or learn from the familiarisation phase, which has not been implemented in STELA so far.

Seyssel et al., 2022c), this is used to measure phonetic knowledge through a machine ABX sound discrimination task (Schatz et al., 2013) in which the learner has to choose two occurrences of, *e.g., 'bop'* as being closer than one occurrence of 'bop' and one occurrence of *'bip'*. The test is done over thousands of trials and over all possible contrasts of phonemes[4].

Preference experiments rely on the ability to compute a 'preference' or 'probability' associated with an input stimulus. Most learning algorithms learn by minimising an objective function, such as the error made in predicting the future based on the past. We can use the same objective function and apply it to test stimuli: if the stimulus is well represented or considered probable by the model, then the error should be low. Totally novel or unexpected stimuli should give a high error.

In STELA, this is used through the spot-the-word task developed in Nguyen et al. (2020). Here, the model receives a spoken word (*e.g.,* 'apple') and a spoken non-word (*e.g.,* 'attle') matched for syllabic and phonotactic structure. We then look at the model's probability of generating both words. The model is considered correct for the trial if the probability of generating the correct word is higher than the non-word. The same logic can be applied at the syntactic level using pairs of grammatical and ungrammatical sentences (i.e., 'the brother learns' versus 'the brothers learns'), in which the model has to assign a higher probability to the grammatical sentence.

In the next section, we present case studies illustrating how meeting the four above-mentioned properties in a single simulation can help us make theoretical advances.

Results

Learning simulations can either be used as "proof of concept" for particular hypotheses about learning mechanisms or to offer novel predictions, never tested experimentally. Here, we focus on the first use case by addressing three long-standing controversies on language learning mechanisms as applied to the phonetic and lexical levels. In each instance, we use a design which enables us to conduct experiments that are both developmental (obtained by training the same learner on increasing quantity of speech, from 50 hours up to 3200 hours) and cross-linguistic (obtained by training and testing the models on two languages, French and English, deriving scores for the native and non-native language).

Could infants rely exclusively on statistical learning over speech inputs to bootstrap into language?

One of the major conceptual difficulties in accounting for early language acquisition is understanding how the young learner can learn several interdependent linguistic levels simultaneously and gradually. Statistical learning (Saffran et al., 1996) seems a good hypothesis to address this, since it posits that infants gather information about the distribution of sounds. This would naturally yield gradual learning. As for simultaneous learning across levels, it could rest on the idea that probabilities can be gathered at several levels of descriptions simultaneously. Now, the evidence in favour of statistical learning is

[4]It is worth pointing out at this point that the sound contrasts presented in this task are extracted from read speech across many different contexts, while stimuli used in laboratory experiments are more controlled. Potential coarticulation effects make the machine sound discrimination task harder than typical in-lab phone discrimination tasks.

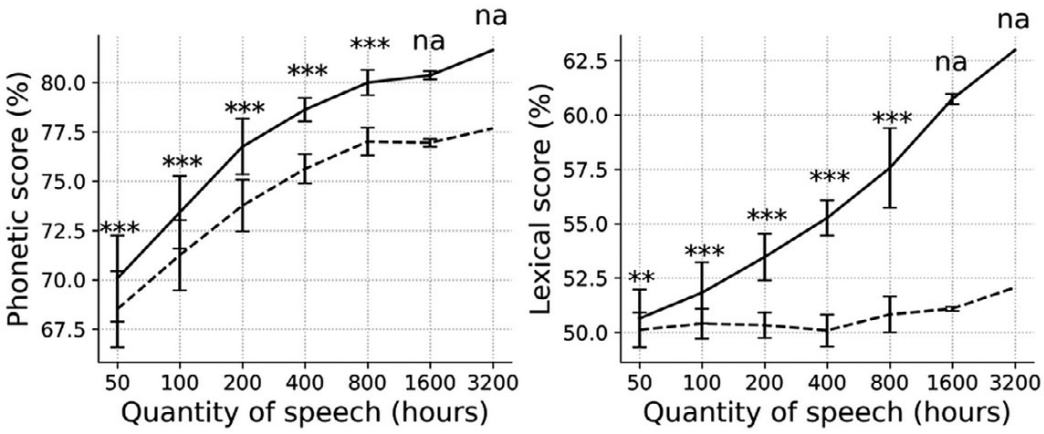

Figure 3. Phonetic (left) and Lexical (right) scores for native and non-native input at different quantities of training data. Phonetic score is expressed in terms of ABX accuracy, obtained by the discrete representations for native and non-native inputs. Lexical score is expressed in terms of accuracy on the spot-the-word task, on the high frequency words for native and non-native inputs. Error bars represent standard errors computed across mutually exclusive training sets. Two-way ANOVAs with factors of nativeness and training language were carried out for each quantity of speech. Significance scores indicate whether the native models are better than the non-native ones. Significance was only computed when enough data points were available to run sensical comparisons. Significance levels: na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001. Figure taken from Lavechin, de Seyssel et al. (2022c).

itself debated. Experimental evidence in infants only rests on simplified artificial languages (synthetic stimuli, small number of sounds), and it is not clear that this would translate to audio data in which speech sounds are highly variable according to phonetic context, speaker, speaking style and rate, in addition to being potentially contaminated by non-speech background sounds.

In Figure 3, we highlight a few key results obtained by STELA when presented with raw audio from audiobooks (Lavechin, de Seyssel et al., 2022c) and tested at the phonetic level (ABX discrimination) and lexical level (spot-the-word) using the tasks presented in a previous section. The results clearly show above-chance performance on native test stimuli and gradual and parallel learning at both phonetic and lexical levels, with the system being able to discriminate sounds better, and prefer words over nonwords more, as more data is presented to the model. This improvement is weaker when tested on a non-native language (actually, not present at all for the lexical task). Further tests (not shown in Figure 3) using a syntactic task (which is also carried out on the language model component presented in Figure 2) in which the system has to show a preference for legal versus illegal sentences revealed much weaker learning. Only the model trained on the largest quantity of speech available (that is, 3200 hours) was able to show preference on an adjective-noun order task ('the nice rabbit' versus 'the rabbit nice'), with a slightly-above-chance 55% accuracy.

In brief, the STELA simulation suggests that raw speech input only, combined with statistical learning, and more precisely predictive learning, is: 1) sufficient to bootstrap the phonetic, the lexical and only very weakly the syntactic levels; 2) sufficient to reproduce the gradual and overlapping developmental trajectory observed in infants at the phonetic and lexical levels[5]. It is the first time a simulation reproduces the gradual and multilevel learning observed in infants from audio signals, at least when audiobooks are used as input.

[5]Larger models, trained with more audio data are able to pass more complex syntactic tests, and show the beginning of semantic abilities as well (Dunbar et al., 2021), suggesting that the structure of the model can itself learn at several levels beyond phonetic and lexical levels.

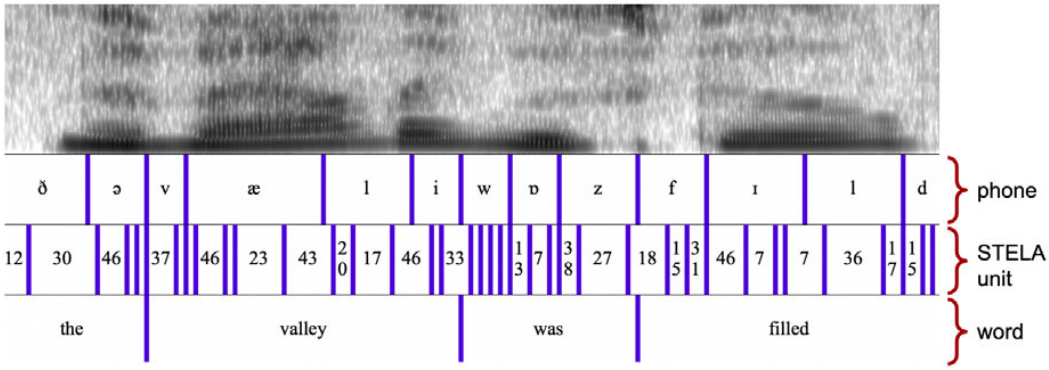

Figure 4. An example spectrogram of an English utterance, along with the corresponding phonemes (top tier) and the units discovered by a STELA model trained on 3200 hours of English. Transcription: "The valley was filled"

Do infants learn and perceive language in terms of linguistic categories?

A second debate concerns whether linguistic categories (phones, words) are necessary building blocks in early language acquisition. On the one hand, linguistic theories describe adult competence in terms of such categories. On the other hand, these categories are language-dependent and therefore need to be learned by infants, who have only access to continuous sensory information at the beginning. Schatz et al. (2021) recently proposed a learning simulation of phonetic learning from raw audio signals based on a probabilistic model using Mixtures of Gaussians. While reproducing observed native advantage effects in phonetic discrimination between Japanese and English phonemes, the learner used in this simulation did not learn phonemes or units that could be described linguistically. These results suggest that phonetic learning can occur without the existence of phonetic categories.

The STELA simulation reproduces this conclusion using a totally different learning algorithm, supporting once again the idea that phonetic categories are not necessary for phonetic learning (see also Feldman, Goldwater, Dupoux & Schatz, 2022). To dive further into this, it is interesting to reflect on how the acoustic model behaves during training concerning the duration of the learnt representations. Pre-exposure (i.e., before the model has received any input) speech is represented within the model as a string of random units. As the model receives speech, it learns to structure this discrete representation: discrete units start repeating themselves, and the sound discrimination accuracy increases. An analysis of the duration of the discrete learnt units revealed that the latter are too short to correspond to phones (43 ms for the learnt units, versus 90 ms for a typical English phone), similarly to what has been found in Schatz et al. (2021). An example of how the discovered units compare to the original phones is presented in Figure 4, where units are clearly shorter than the phones. More surprisingly, the more speech the model receives, the lower the duration of the discrete units. It is essential to note that no constraint is applied to the duration of these units. The model could, in principle, converge to phone-length discrete units, but does no such thing. In other words, the model does not converge to phone-like representations, yet it can still pass phonetic, lexical and, to a certain extent, syntactic tests for which phoneme representations are still often considered a prerequisite[6].

[6]Probing experiments using linear separation revealed however that the representations learned by the acoustic model become more and more structured according to phonetic dimensions like phonetic category

In STELA, it is also possible to ask the question of linguistic categories at higher linguistic levels. Surprisingly, even though the model can distinguish words from non-words, we could not find an indication that the model represents words as such, or would represent the boundaries between words. Yet, the continuous activations found in the hidden layers of the recurrent model contained some approximate linguistic information, as a trained linear classifier was able to classify test words into function versus content words or verb versus adjective/adverb versus noun better than chance, and the separation increased with more input data. These results show that, although the model does not learn discrete and interpretable linguistic categories internally, linguistic information increasingly structures the learnt representations (for more in-depth analyses of the types of units yielded by such models, see de Seyssel, Lavechin, Adi, Dupoux & Wisniewski, 2022; Nguyen, Sagot & Dupoux, 2022; Sicherman & Adi, 2023). Thus, our simulation promotes the view that linguistic categories could be the end product of learning, not their prerequisite.

Can statistical learning alone account for early phonetic acquisition from ecological audio?

One of the largest controversies in language learning orbits around the poverty of the stimulus argument (Chomsky, 1980). This argument states that the input available to infants is too scarce and too noisy to warrant language learning through a general-purpose learning algorithm. Therefore, only a learning algorithm with strong inductive biases would be able to reproduce human language learning. For a long time, this controversy has remained unsolved for lack of learning algorithms that can work even on rather simple inputs. With STELA, at last, we are able to address this controversy, at the level of phonetic and lexical learning. The preceding sections show that a relatively general-purpose system based on predictive coding is able to learn at both levels when fed with audiobooks, but this kind of input may not be realistic enough to correspond to the learning problem faced by infants. Indeed, the audio environment of infants, first of all, contains a majority of non-speech noises, and the little amount of speech that is heard may be under-articulated, reverberated and absorbed by the surrounding obstacles in the environment, and overlaid with various background noises. Could the relatively generic learner of STELA handle such noisy inputs?

One way in which one can revisit this simplifying assumption is by using child-centred long-form recordings, i.e., daylong recordings collected via child-worn microphones as people go about their everyday activities. Lavechin et al. (2022b) exposed the STELA contrastive predictive coding algorithm to such ecological recordings of children's language experiences and found that the discrimination gap between the native and the non-native models vanishes. It is only when supplemented with inductive biases in the form of filtering and augmentation mechanisms (restricting learning to speech parts, taking into account speaker invariance, and making the system resistant to reverberant noise) that the model could exhibit some form of perceptual attunement again (see Figure 5). In addition to this result, Lavechin et al. (2022b) showed that, even in the

(vowels, fricatives, approximants, plosives, etc.), place of articulation for consonants (bilabial, labiodental, dental, etc.), and voicing (voiced or voiceless) as a function of amount of input data. This suggests that the model is learning some phonetic structure from the data even though it is not learning interpretable categories like phonemes.

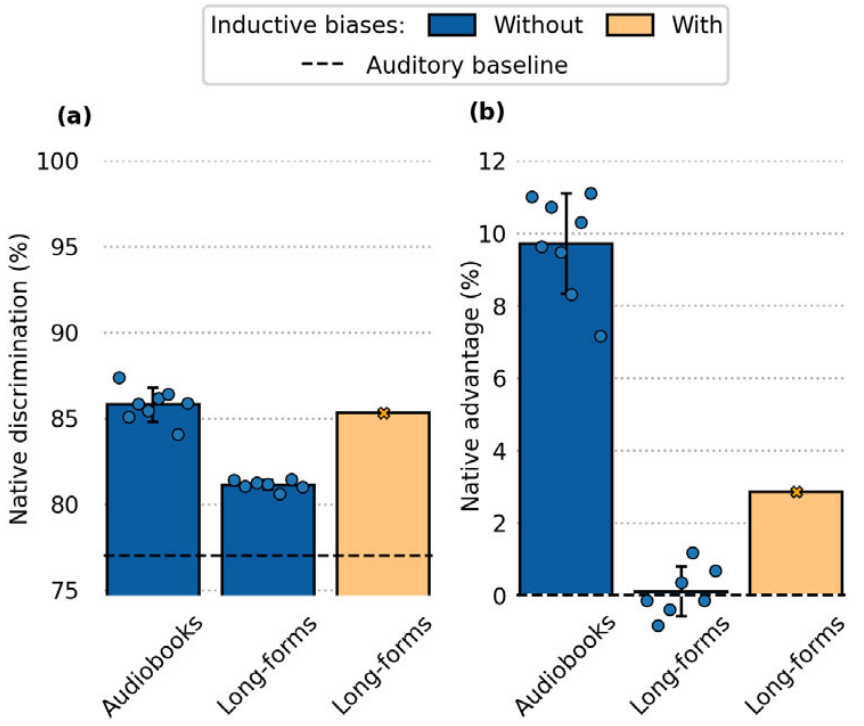

Figure 5. Panel (a) shows native discrimination accuracy, as measured in an ABX discrimination task, obtained by American English and Metropolitan French CPC models (both models are evaluated on phonemes of their native languages). Panel (b) shows native advantage, computed as the average relative difference of the native model and the non-native model, obtained by the same pairs of models (a positive native advantage indicates that the native model is better at discriminating native sounds than the non-native model). Figure adapted from Lavechin et al. (2022b).

presence of inductive biases, the learning speed of the learner was still negatively impacted by the presence of additive noise and reverberation in the training set and that this loss could not be recovered by adding more data.

Given the sparse, variable and noisy nature of the speech overheard by children, this simulation suggests that a statistical learning algorithm alone might not be sufficient to account for early phonetic acquisition. Given that linguistic input represents a small fraction of the audio environment of the child, and that even speech is itself overlapped with non-speech signals, any statistical learning algorithm will devote its resources to discovering the structure of the entire audio, thereby failing to capture the structure of speech sounds.

The three types of inductive biases that were introduced in this study are plausible and independently motivated by experimental evidence in infants: infants show an early preference for attending to speech versus non-speech sounds, and it is plausible that they would learn preferentially on such sounds. In addition, there is evidence that infants distinguish speakers and associate speakers to their voices at an early age; it is therefore plausible that their learning algorithm would be speaker-specific. Finally, the human learner has the benefit of an auditory system that has been fine-tuned by millions of years of evolution to accurately perceive sound sources in complex auditory scenes, and it is plausible that learning operates not on raw sensory data, but rather on sensory streams organised according to source and therefore resist additive noise and reverberation. It is important to note, however, that the inductive biases we implemented are not sufficient;

as subsequent testing at the lexical level showed that, even with them, no lexical learning is evidenced in STELA when fed with long-form recordings. This indicates that, as far as phonetic and lexical learning is concerned, some form of poverty of the stimulus argument is valid, and that generic learning algorithms (at least the ones we tested) need to be supplemented with strong inductive biases.

In brief

We showed that realistic learning simulations could help address some of the key controversies within language acquisition. For instance, STELA shows that statistical learning can be sufficient to reproduce some key findings in infants (phonetic attunement, preference for words over nonwords) from raw audio inputs in the total absence of multimodal grounding or social feedback. It also shows that such learning patterns can arise in the total absence of interpretable linguistic categories. However, it also shows that it has to be supplemented with inductive biases in order to deal with the noise present in naturalistic recordings that are representative of what infants really hear. Of course, these findings are only theoretical results: and, as such, can demonstrate that mechanism A is sufficient (or not needed) to observe outcome B. Whether infants really use similar mechanisms remains to be further established.

What lies ahead?

So far, we have presented evidence that learning simulations, when scaled to incorporate realistic inputs and to model more than one linguistic level, can address some long-standing controversies regarding learning mechanisms in infants. However, our demonstration was limited to testing one hypothetical learning mechanism: statistical learning, and a particularly narrow version of it that is restricted to audio inputs. While STELA could perhaps be counted as the first successful learning simulation of early language acquisition in infants when trained on audiobook data, it struggles to learn with ecological data, even with inductive biases. This suggests two directions of future work: (1) improving STELA with more inductive biases; (2) build a model that incorporates other learning mechanisms (e.g., cross-situational learning, social feedback, etc.). Either way, there is work to be done for both the psycholinguistic and AI communities, which we review below.

Guidelines for psycholinguistics and AI communities

Modelling the environment

Concerning the learning environment, we believe that one challenge that lies ahead consists of collecting and characterising more ecological data. As demonstrated above, results are quite different when models are presented with audiobooks or long-form recordings. We foresee that moving towards more naturalistic training sets will increase the impact and relevance of language learning models.

As data is the crux of any language learning simulation, we believe constant efforts must be put in place to collect and share ecological learning environments. On this front, we would like to highlight important initiatives such as the privacy-preserving sharing platforms for long-form audio recordings (VanDam et al., 2016) or video data (Simon,

Gordon, Steiger & Gilmore, 2015), and the DARCLE (DAylong Recordings of Children's Language Environments, DARCLE.org, n.d.) community. We believe these initiatives must become standard practices as they can transform our understanding of language development by enabling incremental and reproducible science and fueling language learning simulations with realistic data.

In addition, most of what we know concerning language development comes from Western, Educated, Industrialised, Rich, and Democratic (WEIRD) populations (Henrich, Heine & Norenzayan, 2010; Scaff, 2019), and this bias toward WEIRD populations reflects in the type of data computational modellers have access to. Current large-scale audio datasets – whether they contain child-centred recordings or audiobooks – are primarily collected in American English (Kearns, 2014; VanDam et al., 2016). We believe this represents a significant limitation for language realistic learning simulations that can – and should – be run considering diverse socioeconomic and cultural backgrounds. Doing so would help us extract and understand universal constants taking place in the course of language development.

Finally, another challenge is to enrich the nature of the data provided to the learner by incorporating ecologically collected multimodal data, in order to address the importance of cross-situational learning in real life. Also, quantifying the nature and prevalence of social feedback (some of which is nonverbal) is very important as a first step towards building interactive models of the learning environment (Tsuji et al., 2021)

Modelling the learner

One key challenge on the learner side relates to the quantity of data needed to reach a certain level of linguistic performance. Today's most performant text-based language models are trained on roughly one thousand times the amount of linguistic input afforded to a typical child (Warstadt & Bowman, 2022). Therefore, current language models are confronted with a data efficiency problem that is doomed to be even more critical when learning from the raw audio, where other sources of variations have to be considered (speaker's identity, speech rate, acoustic conditions, etc.). Future research should focus on implementing algorithms that can reach human-like performances with the same input data available to an infant – that is, that can map the input and the output measures to those of the modelled human.

Related to this question is the challenge of improving perceptual constancy (on the difficulty of obtaining speaker-invariant representations, see van Niekerk, Nortje, Baas & Kamper, 2021) for state-of-the-art learners of audio representations. As stated above, speech sounds, words and sentences can be realised in numerous ways depending on the speaker's identity, the speech rate, or the acoustic environment. This problem is bypassed when considering the text as input, although text brings other simplifying assumptions irrelevant in the context of language acquisition. We believe normalising audio representations along all dimensions irrelevant to language represents one crucial step to bridging the performance gap between audio-based and text-based language models.

Finally, it is important to develop learners that go beyond the statistical learning hypothesis (Erickson & Thiessen, 2015; Romberg & Saffran, 2010; Saffran et al., 1996). Comparing this hypothesis with alternative ones (cross-modal grounding, social constructivism, etc.) will require developing learners with other learning mechanisms to play a more critical role. Reinforcement learning may, for instance, integrate social and interactive rewards, whereas supervised learning may integrate corrective feedback from

caregivers. Admittedly, integrating multiple learning mechanisms and modalities in a single learning simulation requires collaborative work across fields, as has been analysed in Tsuji et al., 2021.

Modelling the outcome measures

The ultimate test of any language learning simulation is the comparison to humans. Dupoux (2018) proposed to aim at cognitive indistinguishability in that setup: "a human and machine are cognitively indistinguishable with respect to a given set of cognitive tests when they yield numerically overlapping results when run on these tests". This critically assumes that cognitive tests that can be applied to the infant and the learner alike are available.

This is not an easy task, and much more can be done in this regard. As discussed above, outcome measures come in several flavours. Laboratory experiments require infants to cooperate with the setting, which is not a given. As a result, the outcome measures are loaded with non-linguistic factors. Infants' performance depends on various factors that most simulations do not currently consider (e.g., memory or fatigue). This problem is even worse when considering babies for which measures are noisier (but see Blandón, Cristia & Räsänen, 2021, who propose evaluations against meta-analyses). This measurement noise needs to be integrated into the outcome model before direct comparisons between infants and simulations can be done. We refer to this problem as the calibration problem. Some outcome measures are more ecological, and extracted directly from the speech of infants. This requires a learner that can also speak, which has not yet been developed. Other measures, like the CDI, depend on the judgement of a caretaker, which here again needs to be modelled specifically. Ultimately, the calibration of measures extracted from the machine to those extracted from the human (or vice versa) will have to be dealt with one measure at a time.

Similarly to HomeBank (VanDam et al., 2016) or Databrary (Simon et al., 2015), we believe both the AI and the psycholinguistics communities would greatly benefit from a privacy-preserving platform to share stimuli – as well as responses – used in psychology experiments. Such a platform would allow researchers to 1) re-use stimuli as new hypotheses arise; 2) revisit stimuli – or responses – to control for confounding factors, or in the context of meta-analytic studies; and 3) create benchmarks that aim at comparing humans and machines. Concerning the last point, we believe there are still too few works that directly compare human and machine performance on a common benchmark (but see Millet & Dunbar, 2020 for a sound discrimination capability study). A stimuli-sharing platform would accelerate collaborative works across the AI and the psycholinguistics community and could also extend to other domains of psychology (including decision-making or social experiments, for instance).

Conclusion

The article's main aim was to provide an extensive description of an emerging theoretical approach in the field of language acquisition: learning simulations, and especially realistic and broad-scope learning simulations. We proposed four criteria we believe are essential for such a simulation to address the current theory crisis and act as a cumulative and unifying theory of language acquisition. We then presented STELA, one such simulation, and showed how it could help shed light on long-standing controversies. Realistic

Appendix A Appendix: *Realistic and broad-scope learning simulations:first results and challenges*

learning simulations can – and should – integrate the large body of knowledge acquired by the different approaches that comprise the field of language acquisition. Such realistic learning simulations are by no means replacements for other approaches, as all are needed to reach a unified theoretical landscape. Indeed, verbal frameworks can inspire the design of artificial learners, computational models can provide hands-on algorithms, statistical models can exhibit relationships between input and learning outcomes, and corpus studies help describe the characteristics of language environments. Of course, there remain challenges ahead of us to build more complete realistic learning simulations, and we dedicated the last section to address some of them.

Acknowledgements. We are particularly thankful to Dr. Daniel Swingley, Dr. Michael Frank and Dr. Abdellah Fourtassi for helpful insights on previous versions of the manuscript. We are grateful to CoML members for helpful discussions. All errors remain our own. E.D., in his academic role (EHESS), acknowledges funding from Agence Nationale de la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute), a grant from CIFAR (Learning in Machines and Brains), and the HPC resources from GENCI-IDRIS (Grant 2020-AD011011829). M.S. acknowledges PhD funding from Agence de l'Innovation de Défense. M.S and M.L. contributed equally to this work. Authorship order was decided by a coin flip.

Competing interest declaration. The authors declare none.

References

Abend, O., **Kwiatkowski, T.**, **Smith, N. J.**, **Goldwater, S.**, & **Steedman, M.** (2017). Bootstrapping language acquisition. *Cognition*, **164**, 116–143.

Abu-Zhaya, R., **Seidl, A.**, **Tincoff, R.**, & **Cristia, A.** (2017). Building a multimodal lexicon: Lessons from infants' learning of body part words. *GLU 2017 International Workshop on Grounding Language Understanding*, 18–21. https://doi.org/10.21437/GLU.2017-4

Anderson, J. R. (1975). Computer simulation of a language acquisition system: A first report.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, **375**(1791), 20190307.

Bernard, M., **Thiolliere, R.**, **Saksida, A.**, **Loukatou, G. R.**, **Larsen, E.**, **Johnson, M.**, **Fibla, L.**, **Dupoux, E.**, **Daland, R.**, & **Cao, X. N.** (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, **52**(1), 264–278.

Blandón, M. A. C., **Cristia, A.**, & **Räsänen, O.** (2021). *Evaluation of computational models of infant language development against robust empirical data from meta-analyses: What, why, and how?* PsyArXiv. https://doi.org/10.31234/osf.io/yjz5a

Bommasani, R., **Hudson, D. A.**, **Adeli, E.**, **Altman, R.**, **Arora, S.**, **von Arx, S.**, **Bernstein, M. S.**, **Bohg, J.**, **Bosselut, A.**, **Brunskill, E.**, **Brynjolfsson, E.**, **Buch, S.**, **Card, D.**, **Castellon, R.**, **Chatterji, N.**, **Chen, A.**, **Creel, K.**, **Davis, J. Q.**, **Demszky, D.**, … **Liang, P.** (2021). *On the opportunities and risks of foundation models.* ArXiv Preprint ArXiv:2108.07258

Borsos, Z., **Marinier, R.**, **Vincent, D.**, **Kharitonov, E.**, **Pietquin, O.**, **Sharifi, M.**, **Teboul, O.**, **Grangier, D.**, **Tagliasacchi, M.**, & **Zeghidour, N.** (2022). *AudioLM: A language modeling approach to audio generation.* ArXiv Preprint ArXiv:2209.03143.

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, **61**(1-2), 1–38.

Brent, M. R. (Ed.). (1997). *Computational approaches to language acquisition*. Cambridge, MA: MIT Press.

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, **3**(8), 294–301.

Brown, T. B., **Mann, B.**, **Ryder, N.**, **Subbiah, M.**, **Kaplan, J.**, **Dhariwal, P.**, **Neelakantan, A.**, **Shyam, P.**, **Sastry, G.**, **Askell, A.**, **Agarwal, S.**, **Herbert-Voss, A.**, **Krueger, G.**, **Henighan, T.**, **Child, R.**, **Ramesh, A.**, **Ziegler, D. M.**, **Wu, J.**, **Winter, C.**, **Hesse, C.**, **Chen, M.**, **Sigler, E.**, **Litwin, M.**, **Gray, S.**, **Chess, B.**, **Clark, J.**, **Berner, C.**, **McCandlish, S.**, **Radford, A.**, **Sutskever, I.**, & **Amodei, D.** (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

Bulf, H., **Johnson, S. P.**, & **Valenza, E.** (2011). Visual statistical learning in the newborn infant. *Cognition*, **121**(1), 127–132. https://doi.org/10.1016/j.cognition.2011.06.010

Chomsky, N. (1980). *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Chomsky, N. (2013). 4. A review of BF Skinner's verbal behavior. *Volume I Readings in Philosophy of Psychology, Volume* I, 48–64.

Cristia, A. (2022). A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science*, e13265.

Cristia, A., **Dupoux, E.**, **Gurven, M.**, & **Stieglitz, J.** (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, **90**(3), 759–773.

DARCLE.org. (n.d.). Retrieved 9 September 2022, from https://darcle.org/.

de Marcken, C. (1996). *Unsupervised language acquisition*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

de Seyssel, M., **Lavechin, M.**, **Adi, Y.**, **Dupoux, E.**, & **Wisniewski, G.** (2022). *Probing phoneme, language and speaker information in unsupervised speech representations*. In *Proc. Interspeech 2022*, doi:10.21437/Interspeech.2022-373.

Dunbar, E., **Bernard, M.**, **Hamilakis, N.**, **Nguyen, T. A.**, **De Seyssel, M.**, **Rozé, P.**, **Rivière, M.**, **Kharitonov, E.**, & **Dupoux, E.** (2021). *The zero resource speech challenge 2021: Spoken language modelling*. In *Proc. Interspeech 2021*, 1574-1578, doi: 10.21437/Interspeech.2021-1755.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, **173**, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, **14**(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Elman, J. L., **Bates, E. A.**, & **Johnson, M. H.** (1996). *Rethinking innateness: A connectionist perspective on development* (Vol. **10**). Cambridge, MA: MIT press.

Elsner, M., **Goldwater, S.**, & **Eisenstein, J.** (2012, July). Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 184–193).

Erickson, L. C., & **Thiessen, E. D.** (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, **37**, 66–108. https://doi.org/10.1016/j.dr.2015.05.002

Feldman, N. H., **Goldwater, S.**, **Dupoux, E.**, & **Schatz, T.** (2022). Do infants really learn phonetic categories? *Open Mind*, **5**, 113–131.

Fenson, L. (2007). *MacArthur-Bates communicative development inventories*. Baltimore, MD: Brookes.

Fernald, A., **Marchman, V. A.**, & **Weisleder, A.** (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science*, **16**(2), 234–248.

Fodor, J. A., & **Pylyshyn, Z. W.** (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, **28**(1–2), 3–71.

Frank, M. C. (2011). Computational models of early language acquisition. *Current Opinion in Neurobiology*, **21**(3), 381–386.

Frank, M. C. (2014). "Psychological plausibility" considered harmful. *Babies learning language*. http://babieslearninglanguage.blogspot.com/2014/02/psychological-plausibility-considered.html

Frank, M. C., **Goodman, N. D.**, & **Tenenbaum, J. B.** (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, **20**(5), 578–585.

Frank, M. C., **Bergelson, E.**, **Bergmann, C.**, **Cristia, A.**, **Floccia, C.**, **Gervain, J.**, **Hamlin, J. K.**, **Hannon, E. E.**, **Kline, M.**, & **Levelt, C.** (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, **22**(4), 421–435.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, **31**(4), 271–288.

Gleitman, L. R., **Newport, E. L.**, & **Gleitman, H.** (1984). The current status of the motherese hypothesis. *Journal of Child Language*, **11**(1), 43–79. https://doi.org/10.1017/S0305000900005584

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, **10**(5), 447–474.

Hart, B., & **Risley, T. R.** (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.

Henrich, J., **Heine, S. J.**, & **Norenzayan, A.** (2010). Most people are not WEIRD. *Nature*, **466**(7302), 29–29.

Jain, S., **Osherson, D.**, **Royer, J. S.**, & **Sharma, A.** (1999). *Systems that learn: An introduction to learning theory*. Cambridge, MA: MIT press.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, **21**(1-2), 3–28.

Kachergis, G., **Marchman, V. A.**, & **Frank, M. C.** (2021). Toward a "standard model" of early language learning. *Current Directions in Psychological Science*, **31**, 20–27.

Karmiloff-Smith, B. A. (1994). Beyond modularity: A developmental perspective on cognitive science. *European Journal of Disorders of Communication*, **29**(1), 95–105.

Kearns, J. (2014). Librivox: Free public domain audiobooks. *Reference Reviews.*

Kelley, H. H. (1967). Attribution theory in social psychology. In: D. Levine (Ed.) *Nebraska symposium on motivation* (Vol. **15**, pp. 192–240). Lincoln: University of Nebraska.

Kuhl, P. K., **Tsao, F.-M.**, & **Liu, H.-M.** (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, **100** (15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

Kuhl, P. K., **Conboy, B. T.**, **Coffey-Corina, S.**, **Padden, D.**, **Rivera-Gaxiola, M.**, & **Nelson, T.** (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**(1493), 979–1000.

Lakhotia, K., **Kharitonov, E.**, **Hsu, W.-N.**, **Adi, Y.**, **Polyak, A.**, **Bolte, B.**, **Nguyen, T.-A.**, **Copet, J.**, **Baevski, A.**, **Mohamed, A.**, & **Dupoux, E.** (2021). Generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, **9**, 1336–1354.

Lavechin, M., **de Seyssel, M.**, **Gautheron, L.**, **Dupoux, E.**, & **Cristia, A.** (2022a). Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics.* https://doi.org/10.1146/annurev-linguistics-031120-122120

Lavechin, M., **de Seyssel, M.**, **Métais, M.**, **Metze, F.**, **Mohamed, A.**, **Bredin, H.**, **Dupoux, E.**, & **Cristia, A.** (2022b). *Statistical learning models of early phonetic acquisition struggle with child-centered audio data.* PsyArXiv. https://doi.org/10.31234/osf.io/5tmgy

Lavechin, M., **de Seyssel, M.**, **Titeux, H.**, **Bredin, H.**, **Wisniewski, G.**, **Peperkamp, S.**, **Cristia, A.**, & **Dupoux, E.** (2022c). *Can statistical learning bootstrap early language acquisition? A modeling investigation.* PsyArxiv. https://doi.org/10.31234/osf.io/rx94d

Leibo, J. Z., **d'Autume, C. de M.**, **Zoran, D.**, **Amos, D.**, **Beattie, C.**, **Anderson, K.**, **Castañeda, A. G.**, **Sanchez, M.**, **Green, S.**, & **Gruslys, A.** (2018). *Psychlab: A psychology laboratory for deep reinforcement learning agents.* ArXiv Preprint ArXiv:1801.08116.

Liu, X., **He, P.**, **Chen, W.**, & **Gao, J.** (2019). *Improving multi-task deep neural networks via knowledge distillation for natural language understanding.* ArXiv Preprint ArXiv:1904.09482.

MacWhinney, B., & **MacWhinney, B.** (1987). The competition model. *Mechanisms of language acquisition*, 249–308. London, United Kingdom: Routledge.

McMurray, B. (2021). Categorical perception: Lessons from an enduring myth. *The Journal of the Acoustical Society of America*, **149**(4), A33–A33.

McPhetres, J., **Albayrak-Aydemir, N.**, **Mendes, A. B.**, **Chow, E. C.**, **Gonzalez-Marquez, P.**, **Loukras, E.**, **Maus, A.**, **O'Mahony, A.**, **Pomareda, C.**, & **Primbs, M. A.** (2021). A decade of theory as reflected in psychological science (2009–2019). *PloS One*, **16**(3), e0247986.

Meltzoff, A. N., **Kuhl, P. K.**, **Movellan, J.**, & **Sejnowski, T. J.** (2009). Foundations for a new science of learning. *Science*, **325**(5938), 284–288.

Miller, J. F., & **Chapman, R. S.** (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research*, **24**(2), 154–161.

Millet, J., & **Dunbar, E.** (2020). The perceptimatic English benchmark for speech perception models. *CogSci 2020-42nd Annual Virtual Meeting of the Cognitive Science Society.*

Muthukrishna, M., & **Henrich, J.** (2019). A problem in theory. *Nature Human Behaviour*, **3**(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nelson, D. G. K., **Jusczyk, P. W.**, **Mandel, D. R.**, **Myers, J.**, **Turk, A.**, & **Gerken, L.** (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, **18**(1), 111–116.

Nelson, K. (2007). *Young minds in social worlds: Experience, meaning, and memory*. Cambridge, MA: Harvard University Press.

Newman, R. S., **Rowe, M. L.**, & **Ratner, N. B.** (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, **43**(5), 1158–1173. https://doi.org/10.1017/S0305000915000446

Nguyen, T. A., **de Seyssel, M.**, **Rozé, P.**, **Rivière, M.**, **Kharitonov, E.**, **Baevski, A.**, **Dunbar, E.**, & **Dupoux, E.** (2020). *The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling.* ArXiv Preprint ArXiv:2011.11588.

Nguyen, T. A., **Sagot, B.**, & **Dupoux, E.** (2022). Are discrete units necessary for spoken language modeling? *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1415–1423.

Pearl, L., & **Sprouse, J.** (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, **20**(1), 23–68.

Pelucchi, B., **Hay, J. F.**, & **Saffran, J. R.** (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, **80**(3), 674–685.

Pinker, S. (1979). Formal models of language learning. *Cognition*, **7**(3), 217–283.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York, NY: Harper Collins.

Räsänen, O., & **Khorrami, K.** (2019). *A computational model of early language acquisition from audiovisual experiences of young infants.* ArXiv Preprint ArXiv:1906.09832.

Romberg, A. R., & **Saffran, J. R.** (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, **1**(6), 906–914.

Roy, D. K., & **Pentland, A. P.** (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, **26**(1), 113–146.

Rumelhart, D., **McClelland, J.**, & **MacWhinney, B.** (1987). *Mechanisms of language acquisition*. In B. MacWhinney (Ed.), (pp. 195–248). Erlbaum Hillsdale, NJ.

Saffran, J. R., **Aslin, R. N.**, & **Newport, E. L.** (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928. https://doi.org/10/fcqz9d

Saffran, J. R., & **Kirkham, N. Z.** (2018). Infant statistical learning. *Annual Review of Psychology*, **69**, 181–203.

Scaff, C. (2019). *Beyond WEIRD: An interdisciplinary approach to language acquisition* [PhD Thesis]. PhD thesis.

Schatz, T., **Peddinti, V.**, **Bach, F.**, **Jansen, A.**, **Hermansky, H.**, & **Dupoux, E.** (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1781–1785.

Schatz, T., **Feldman, N. H.**, **Goldwater, S.**, **Cao, X.-N.**, & **Dupoux, E.** (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, **118**(7), e2001844118.

Seidl, A., **Tincoff, R.**, **Baker, C.**, & **Cristia, A.** (2015). Why the body comes first: Effects of experimenter touch on infants' word finding. *Developmental Science*, **18**(1), 155–164. https://doi.org/10.1111/desc.12182

Sicherman, A., & **Adi, Y.** (2023). *Analysing discrete self supervised speech representation for spoken language modeling.* ArXiv Preprint ArXiv:2301.00591.

Simon, D. A., **Gordon, A. S.**, **Steiger, L.**, & **Gilmore, R. O.** (2015). Databrary: Enabling sharing and reuse of research video. *Proceedings of the 15th Acm/Ieee-Cs Joint Conference on Digital Libraries*, 279–280.

Skinner, B. F. (1957). *Verbal behavior* (pp. xi, **478**). Acton, MA: Copley Publishing Group

Smith, L., & **Yu, C.** (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, **106**(3), 1558–1568.

Snow, C. E. (1989). Understanding social interaction and language acquisition; sentences are not enough. In *Interaction in Human Development* (pp. 83–103). Lawrence Erlbaum Associates, Inc.

Suanda, S. H., **Mugwanya, N.**, & **Namy, L. L.** (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, **126**, 395–411.

Swingley, D., & **Humphrey, C.** (2018). Quantitative linguistic predictors of infants' learning of specific English words. *Child Development*, **89**(4), 1247–1267. https://doi.org/10.1111/cdev.12731

Tesar, B., & **Smolensky, P.** (2000). *Learnability in optimality theory*. Cambridge, MA: Mit Press.

Tomasello, M. (2003). The key is social cognition. *Language in mind: Advances in the study of language and thought*, pp47–57. Cambridge, MA: MIT press.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tsuji, S., **Cristia, A.**, & **Dupoux, E.** (2021). SCALa: A blueprint for computational models of language acquisition in social context. *Cognition*, **213**, 104779. https://doi.org/10.1016/j.cognition.2021.104779

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, **59**(236), 433.

Vallabha, G. K., **McClelland, J. L.**, **Pons, F.**, **Werker, J. F.**, & **Amano, S.** (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, **104**(33), 13273–13278.

VanDam, M., **Warlaumont, A. S.**, **Bergelson, E.**, **Cristia, A.**, **Soderstrom, M.**, **De Palma, P.**, & **MacWhinney, B.** (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language*, **37**(02), 128–142.

Van Niekerk, B., **Nortje, L.**, **Baas, M.**, & **Kamper, H.** (2021). *Analyzing speaker information in self-supervised models to improve zero-resource speech processing.* ArXiv Preprint ArXiv:2108.00917.

Warstadt, A., & **Bowman, S. R.** (2022). *What artificial neural networks can tell us about human language acquisition*. ArXiv Preprint ArXiv:2208.07998.

Werker, J., & **Curtin, S.** (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, **1**(2), 197–234. https://doi.org/10.1207/s15473341lld0102_4

Yu, C., **Ballard, D. H.**, & **Aslin, R. N.** (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, **29**(6), 961–1005. https://doi.org/10.1207/s15516709cog0000_40

Yu, C., & **Ballard, D. H.** (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, **70**(13–15), 2149–2165.

Yu, C., & **Smith, L. B.** (2017). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*, **41**, 5–31.

Zhang, Y., **Chen, C.**, & **Yu, C.** (2019). Mechanisms of cross-situational learning: Behavioral and computational evidence. *Advances in Child Development and Behavior*, **56**, 37–63.

Appendix A: How to derive a probability from a Language Model?

Head-turn preference experiments (Nelson et al., 1995) provide a wealth of results regarding the type of stimuli infants prefer to listen to. However, mechanisms underlying this preference remain unobservable. Computational modelling provides complementary information by assessing hypotheses about HOW statistical information might be used to exhibit similar preference patterns as those exhibited by infants, or *what* underlying information processing problem is being solved. Language models, and probabilistic models in general, offer a natural way to extract a preference measure from an artificial learner: a stimulus A is preferred to a stimulus B if A is more probable than B.

But how does one compute the probability of a stimulus from a Language Model? First, the waveform goes through the Acoustic Model which returns a discrete representation of the audio: $q_1, q_2 \ldots, q_T$. Then, the Language Model, which has been trained to predict the next discrete unit of a sequence given its past context, assigns a probability to the discrete sequence using the following chain-rule:

$$P(q_1, \ldots, q_T) = \prod_{t=1}^{T} P(q_t | q_1, \ldots, q_{t-1})$$

We compute the logarithm of the resulting probability which has the effect of increasing the difference between probabilities assigned to a minimal pair of stimuli (e. g., a word and a non-word that differ in a single phoneme). The logarithm is then normalised by the length of the input stimuli to enforce the model to not show a constant preference for the longest stimuli.

Appendix B: Overview of the learner used in STELA

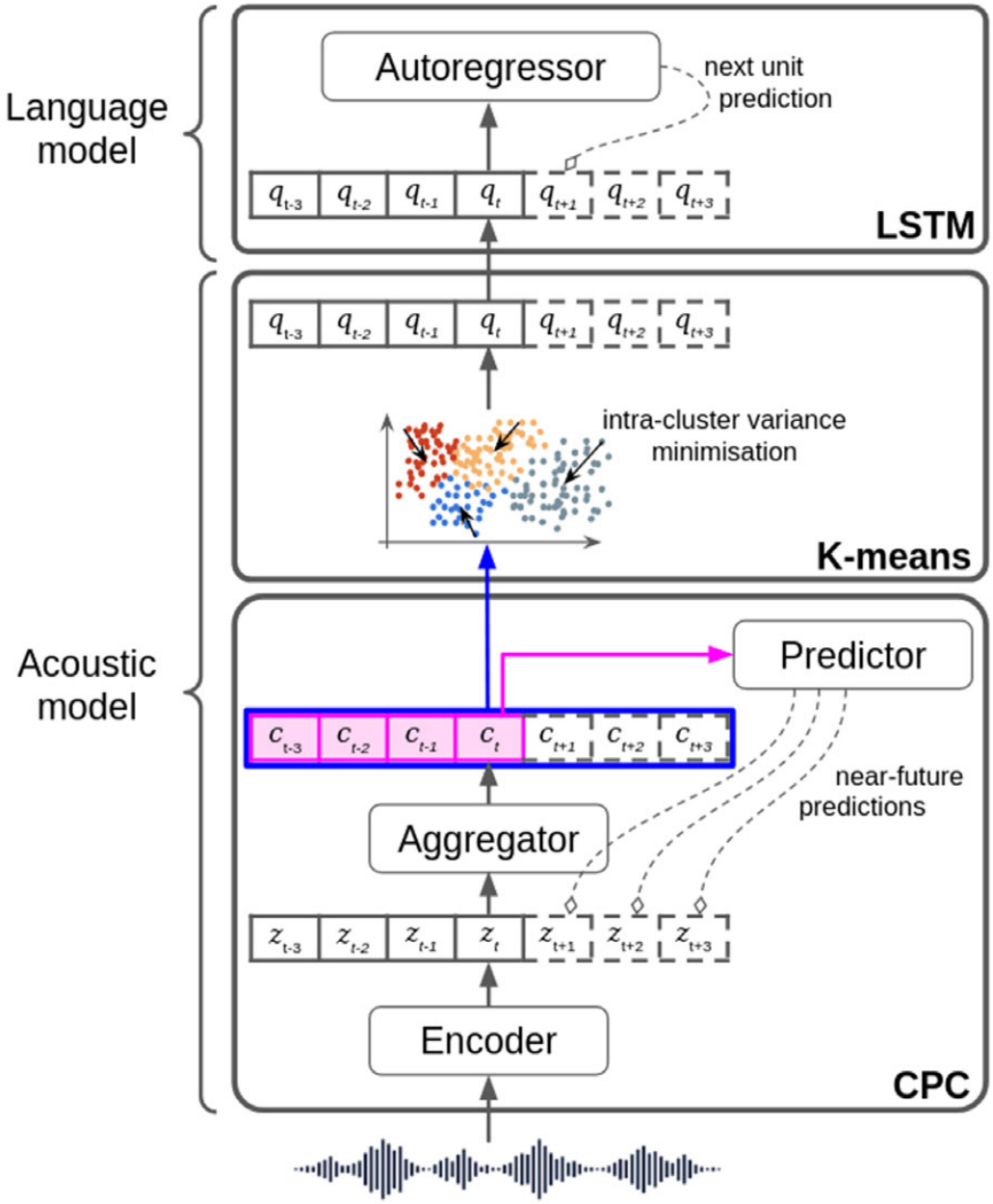

Figure A1. Model of the learner used in STELA. The Acoustic model is composed of a convolutional encoder which delivers a vector of continuous values z_t every 10ms. This is sent to a recurrent network aggregator that integrates context and delivers vectors with the same time step. Contrastive Predictive Coding is trained to predict the outputs of the encoder in the near-future (up to 120 ms). The output of the aggregator is sent to a K-means algorithm that discretise the continuous representations c_t into q_t. Then, a language model (long-short term memory (LSTM) network) is trained to predict the next q_t unit based on past ones.

Cite this article: de Seyssel M., Lavechin M., & Dupoux E. (2023). Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language* 1–24, https://doi.org/10.1017/S0305000923000272

RÉSUMÉ

L'utilisation d'enregistreurs légers portés par les enfants et collectant du son tout au long de la journée ouvre la voie à une approche de 'données massives' pour étudier le développement du langage chez l'enfant. En recueillant la production langagière de l'enfant ainsi que son environnement linguistique, ces enregistrements nous offrent une vision réaliste des usages quotidiens du langage. Cependant, de tels enregistrements constituent rapidement des milliers d'heures d'audio et nécessitent l'utilisation d'outils de traitement automatique de la parole. En plus de fournir des mesures réalistes de ce que les enfants entendent et disent, ces enregistrements peuvent alimenter les modèles computationnels d'acquisition du langage avec une entrée comparable à ce que les enfants entendent réellement, ouvrant ainsi de nouvelles perspectives pour simuler l'apprentissage du langage. Nous présentons d'abord nos contributions au développement d'algorithmes de traitement automatique de la parole pour les enregistrements longs centrés sur l'enfant. À travers une série d'études, nous montrons ensuite comment le caractère réaliste des données d'entrée affecte les résultats d'apprentissage des modèles computationnels d'acquisition précoce du langage, démontrant ainsi l'importance d'exécuter des simulations d'apprentissage du langage qui reflètent étroitement les caractéristiques de la vie réelle.

MOTS CLÉS

développement du langage, psycholinguistique, traitement de la parole, apprentissage profond, apprentissage supervisé, apprentissage auto-supervisé, sciences cognitives

ABSTRACT

Lightweight child-worn recorders that collect audio across an entire day allow for a big-data approach to the study of language development. By collecting the child's production and linguistic environment, these recordings offer us a uniquely naturalistic view of everyday language uses. However, such recordings quickly accumulate thousands of hours of audio and require the use of automatic speech processing algorithms. Besides providing ecologically-valid measures of what children hear and say, these recordings can fuel computational models of early language acquisition with what infants truly hear. This opens up new opportunities for running realistic language learning simulations. We first present our contributions to developing automatic speech processing algorithms for child-centered long-form recordings. Through a series of studies, we then show how the ecological validity of the input data affects the learning outcomes of computational models of early language acquisition, demonstrating the importance of running language learning simulations that closely emulate real-life situations.

KEYWORDS

language development, psycholinguistics, speech processing, deep learning, supervised learning, self-supervised learning, cognitive sciences