THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *2 février 2022* **par :**

Lucile GELIN

Reconnaissance automatique de la parole d'enfants apprenant·e·s lecteur·ice·s en salle de classe : modélisation acoustique de phonèmes

---

JURY

Lori LAMEL	Directrice de Recherche LIMSI	Présidente
Isabel TRANCOSO	Professeure INESC-ID	Rapportrice
Yannick ESTEVE	Professeur LIA	Rapporteur
Julien PINQUIER	Maître de conférence IRIT	Co-directeur de thèse
Thomas PELLEGRINI	Maître de conférence IRIT	Co-directeur de thèse
Morgane DANIEL	Ingénieure R&D Lalilo	Co-directrice de thèse

École doctorale et spécialité :
MITT : Domaine STIC : Intelligence Artificielle
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse
Directeurs de Thèse :
Julien PINQUIER et *Thomas PELLEGRINI*
Encadrante industrielle :
Morgane DANIEL

Remerciements

Je remercie tout d'abord l'entreprise Lalilo pour avoir accepté de financer cette thèse il y a trois ans, malgré sa jeunesse et son futur incertain, ainsi que l'ANRT pour sa participation au financement. Je remercie également l'équipe SAMoVA de l'Institut de Recherche en Informatique de Toulouse, qui m'a accueillie et permis de mener à bien mes recherches.

Parmi tous mes collègues, trois ont eu une place particulière à mes côtés, car ils ont eu la (lourde) tâche de superviser mes travaux. Je remercie ainsi mes directeurs et directrice de thèse Julien Pinquier, Thomas Pellegrini et Morgane Daniel , pour leur accompagnement technique et moral lors de la première année, où il était frustrant et décevant de ne pas arriver à publier mes travaux, lors de la deuxième année, où ils m'ont aidée à développer mes idées, et enfin lors du marathon d'écriture de la dernière année. Merci à Julien et Thomas pour avoir supporté mon pauvre niveau à la pétanque et aimé (j'y crois) mes gâteaux aux légumes ! Un grand merci à Morgane pour les énormes efforts qu'elle a fourni pour m'intégrer à Lalilo malgré la distance ; pour avoir souvent comblé la confiance en moi que je n'avais pas ; pour sa solidarité sur les sujets écologiques ; pour m'avoir donné envie de continuer à travailler sur la reconnaissance de la parole d'enfant.

Je remercie naturellement l'ensemble des membres de mon jury de thèse pour avoir pris le temps de lire avec attention mon manuscrit de thèse malgré la présence (légère !) d'écriture inclusive et pour leurs retours précieux. Merci donc à Lori Lamel d'avoir présidé ce jury, et à Isabel Trancoso et Yannick Estève pour avoir rapporté mon manuscrit.

Je tiens ensuite à remercier les enseignantes de CP et CE1 (Isabelle, Carole, Cécile, Pascale, Isabelle, Ruth et Pauline) qui ont accepté de m'accueillir dans leurs classes pour ma collecte de données de parole d'enfants, et à leurs élèves que j'ai pu enregistrer. Merci également à Charlotte Sicre, qui a rendu possible cette collecte grâce à son travail sur les autorisations de captation de voix, et à Pierre Puddu qui m'a aidée dans la création du contenu de lecture à voix haute.

J'adresse de plus mes remerciements aux collègues/ami·e·s qui ont fait tout ou une partie du chemin avec moi à Lalilo, et l'ont rendu plus agréable. Merci à celle·eux qui, arrivé·e·s avant moi, m'ont accueillie (Amine, Laurent, Benji, Marjo, Nico, Juju, Lorry, Rodolphe), ainsi qu'à tou·te·s les autres qui nous ont rejoint au fil des ans. Je tiens à remercier tout particulièrement Corentin qui me tire constamment vers le haut grâce à son aide précieuse et son écoute toujours bienveillante ; Juliana qui a assuré ma santé mentale pendant le confinement grâce à ses Voot Camp énergiques ; Lorry dont le courage et la joie de vivre m'ont inspirée au quotidien ; Juju qui a toujours tout fait pour que je me sente bien et à ma place. Merci à Fabian pour les liens professionnels et amicaux que 6 mois difficiles d'encadrement à distance en situation de pandémie mondiale ont renforcés. Je remercie enfin les autres de la team data (Marine, Steve, Thomas, Baptiste, Mikaël) pour leur compréhension et leurs encouragements lors de la longue écriture de ce manuscrit.

N'ayant pas bénéficié d'une, mais de deux équipes, il m'est également nécessaire de remercier toutes celles et ceux qui ont contribué à la super ambiance de l'équipe SAMoVA. Je pense notamment à mes deux adorables co-bureaux pendant ces trois ans : Mathieu, dit l'homme qui agit plus vite que son ombre, qui a bousculé et bouscule encore avec affection la procrastinatrice que je suis ; et Tim, le logopède géant, qui m'aura apporté plus de fous rires et d'amour que je n'aurais pu imaginer. Merci également à l'ancienne locataire du 222, Eugenia, et au petit nouveau, Etienne, pour les échanges amicaux que nous avons eus. J'adresse une pensée à celles et ceux qui sont devenus des ami·e·s : Léo pour son incroyable hospitalité ; Jim, l'éternel stagiaire, pour ses blagues et ses talents de DJ ; Verdi pour son humour et ses talents manuels qui me font rêver ; Lila pour nos discussions autour de nos nombreuses passions communes ; Vincent pour ses conseils bien-être et son rire communicatif ; Robin pour son accent qui nous fait craquer et sa délicieuse tarte au citron ; Sebastião pour sa maîtrise parfaite du français lors de nos discussions philo-politico-écologiques (et pardon de ne toujours pas savoir prononcer ton prénom...) ; Benjamin pour ses conseils culinaires ; ainsi qu'à celle·eux qui ont, à un moment ou à un autre, activement participé aux parties de cartes traditionnelles en salle machine : Evan, Thomas, Estelle, Nicolas, Sébastien, Alice, Baptiste, Amélie, Gautier, Antoine... Merci également aux permanents que je n'aurais pas encore cités (Julie, Isabelle, Christine, Jérôme, Hervé) pour leurs conseils et leur bienveillance, et notamment à Régine pour sa relecture d'une partie de mon manuscrit.

Enfin, je souhaite remercier mes ami·e·s de longue date, de la petite section à la dernière année d'école d'ingénieur·e, qui m'ont soutenue pendant ces trois années de thèse. Un grand merci à ma belle famille pour leur bienveillance, et à mon incroyable famille qui a toujours cru en moi et m'a accompagnée dans cette expérience de vie : à Marie qui a répondu à mes nombreuses questions pédagogiques ; à Vincent et Matthieu qui ont partagé leurs expériences de thésards ; à mes parents, ma soeur et mes frères qui m'ont apporté tout leur amour ; et à tous, qui ont supporté mes tentatives récurrentes pour démystifier mon sujet de thèse. Je finis par remercier infiniment Steven, qui a toujours été présent, et avec qui je grandis chaque jour.

Table des matières

Table des figures

Liste des tableaux

Acronymes

AL	Apprenant·e Lecteur·rice
ANN	*Artificial Neural Network*
Bi-GRU	*Bi-directional Gated Recurrent Unit*
Bi-LSTM	*Bi-directional Long-Short Term Memory*
CE	*Cross-Entropy*
CER	*Character Error Rate*
CMN	*Cepstral Mean Normalisation*
CMVN	*Cepstral Mean and Variance Normalisation*
CNN	*Convolutional Neural Network*
CNRS	Centre National de la Recherche Scientifique
CTC	*Connectionist Temporal Classification*
DBN	*Deep Belief Network*
DNN	*Deep Neural Network*
EAS	Extraction par modèle Auto-Supervisé
FFNN	*Feed-forward Neural Network*
fMLLR	*Feature-space Maximum Likelihood Linear Regression*
GMM	*Gaussian Mixture Model*
GOP	*Goodness Of Pronunciation*
GPU	*Graphics Processing Unit*
GRU	*Gated Recurrent Unit*
HMM	*Hidden Markov Model*
LAS	*Listen, Attend and Spell*
LDA	*Linear Discriminant Analysis*
LF	*Lattice-Free*
LSTM	*Long-Short Term Memory*
L2	Seconde langue
MDD	*Mispronunciation detection and diagnosis*
MFCC	*Mel-Frequency Cepstral Coefficients*
MLP	*Multi-layer perceptron*
MMI	*Maximum Mutual Information*
PASE	*Problem Agnostic Speech Encoder*
PER	*Phoneme Error Rate*

PIRLS *Progress in International Reading Literacy Study*
PISA Programme International pour le Suivi des Acquis des élèves
PLP *Perceptual Linear Prediction*
PS Produit Scalaire
PSP Produit Scalaire Pondéré
QRNN *Quasi-Recurrent Neural Network*
RAP Reconnaissance Automatique de la Parole
RASTA *RelAtive SpecTrAl*
ReLU *Rectified Linear Unit*
RGPD Règlement Général sur la Protection des Données
RNN *Recurrent Neural Network*
RSB Rapport Signal à Bruit
SAT *Speaker Adaptive Training*
TCD Transformée en Cosinus Discrète
TDNN *Time-Delay Neural Network*
TDNNF *Factorised Time-Delay Neural Network*
TFD Transformée de Fourier Discrète
TL *Transfer Learning*
TS Traitement du Signal
ULIS Unités Spécialisées d'Inclusion Scolaire
VTLN *Vocal Tract Length Normalisation*
WCPM *Word Correct Per Minute*
WER *Word Error Rate*
WFST *Weighted Finite-State Transducer*
WSJ *Wall Street Journal*
ZPD Zone Proximale de Développement

Liste des classes de l'école primaire française

CP Cours Préparatoire - 6/7 ans - *First grade*
CE1 Cours Élémentaire 1 - 7/8 ans - *Second grade*
CE2 Cours Élémentaire 2 - 8/9 ans - *Third grade*
CM1 Cours Moyen 1 - 9/10 ans - *Fourth grade*
CM2 Cours Moyen 2 - 10/11 ans - *Fifth grade*

Introduction

Contexte des travaux et problématique de la thèse

Maîtriser la lecture est une étape clef du développement d'un·e enfant afin de devenir autonome au quotidien. Or, l'apprentissage de la lecture est loin d'être facile, et 40,5% des élèves français·e·s de 15 ans ont des difficultés en lecture, qui sont fortement handicapantes pour la moitié d'entre eux [1]. L'apprentissage de la lecture passe par l'oralisation, qui permet à l'enfant de transformer les signes écrits en sons et de construire le sens des mots lus grâce à l'écoute de ces sons [Beaume 1987]. Les professeur·e·s des écoles n'ont toutefois pas le temps de faire lire leurs élèves individuellement, et la lecture orale est ainsi peu pratiquée en classe. Des plateformes pédagogiques et numériques sont d'ores et déjà disponibles pour accompagner l'apprentissage de l'enfant et son acquisition de nouvelles connaissances et/ou compétences dans un contexte ludique. L'entreprise Lalilo vise à favoriser la pratique de la lecture orale en classe, via une plateforme pédagogique éponyme, qui inclut notamment un exercice de lecture orale. Nous nous intéressons dans ce contexte à la possibilité de fournir un retour précis et adapté à l'enfant sur la qualité de sa lecture. Le chapitre 1 sera dédié à la présentation détaillée du contexte industriel dans lequel se place cette thèse. Un système global de détection d'erreurs de lecture, présenté dans [Hembise 2021], est fondé sur un module de reconnaissance automatique de phonèmes, dont la précision influe fortement la capacité du système à correctement classifier lectures correctes et incorrectes.

Les enfants sont de plus en plus amenés à utiliser des technologies vocales et le domaine de la reconnaissance automatique de parole (RAP) d'enfants est en plein essor. La littérature regorge de méthodes et architectures appliquées à la parole d'adultes, mais peu l'ont été à la parole d'enfants. Cette parole est en effet plus difficile à reconnaître que celle d'adultes : le développement de l'enfant induit des gammes de fréquences décalées et élargies et d'importantes variabilités acoustique et prosodique. Pour en modéliser correctement la complexité, une grande quantité de données de parole d'enfants est nécessaire : leurs âges et la tâche de lecture effectuée doivent correspondre à l'âge des futurs utilisateur·rice·s et à la tâche d'application. A défaut, il est indispensable de mettre en œuvre des méthodes pour adapter des données différentes à cette tâche spécifique. De par notre application de la RAP à la lecture orale d'enfants apprenant·e·s lecteur·rice·s (AL) en salle de classe, nos enregistrements sont de plus dégradés par la présence d'erreurs de lecture et de bruit de brouhaha.

Cette thèse CIFRE, financée par l'entreprise Lalilo, s'est focalisée sur le perfectionnement du module de reconnaissance automatique appliqué à la parole lue d'enfants AL. Plutôt que de la RAP classique où des mots sont reconnus, nous choisissons de faire de la reconnaissance automatique de phonèmes : cela nous permet de déceler des erreurs de lecture potentielles à un niveau de granularité très précis. Nous nous sommes ainsi particulièrement concentré·e·s sur les architectures de modélisation acoustique de phonèmes, et proposons diverses stratégies

1. https://colibris.link/pirls-2016

innovantes permettant d'obtenir les meilleures performances pour notre application.

Verrous et défis scientifiques

Dans cette thèse, nous nous interrogeons sur les grandes différences de performance des systèmes de RAP dédiés aux adultes et aux enfants, et questionnons l'applicabilité des méthodes, créées initialement pour la parole d'adultes, à la parole d'enfants. De plus, peu d'équipes de recherche travaillant sur la lecture orale d'enfants AL, nous n'avons que peu d'information quant aux spécificités de cette parole. Nous cherchons donc à comprendre leur influence sur la précision de la reconnaissance, et à contrebalancer de potentiels effets négatifs par des techniques innovantes.

Nos travaux visent ainsi à répondre aux questions suivantes, dont certaines (1 à 4) sont liées à l'application de la RAP à la parole d'enfants, et d'autres (5 et 6) à la spécificité de la parole d'AL en salle de classe :

1. Pourquoi la parole d'enfants est-elle plus difficile à reconnaître que celle d'adultes ? Quel impact ont leurs différences sur les performances de RAP ?
2. Vaut-il mieux adapter à la parole d'enfants un modèle entraîné sur de la parole d'adultes ou bien créer un modèle enfant à partir de rien ?
3. Quelle quantité de données devons-nous utiliser pour une modélisation acoustique acceptable de la parole d'enfants ? Que faire lorsque cette quantité minimale n'est pas à notre disposition ?
4. Quels modèles de la littérature, offrant des performances à l'état de l'art pour la RAP d'adultes, sont pertinents pour la reconnaissance de phonèmes dans la parole d'enfants ? Les nouvelles approches dites *end-to-end* sont-elles en mesure de surpasser les approches classiques hybrides dans notre contexte d'application ?
5. La présence de bruit est un problème persistant dégradant la performance des systèmes de RAP, et le bruit de brouhaha est reconnu comme l'un des plus difficiles à traiter. Quelle est l'influence du bruit de brouhaha d'enfants typique des salles de classe sur la capacité d'un modèle à reconnaître la parole d'un enfant cible ? Est-il possible d'améliorer la robustesse d'un modèle face à ce bruit particulièrement complexe, composé principalement de parole superposée d'enfants ?
6. De façon analogue à d'autres types de parole atypique (bégaiement, apprentissage d'une seconde langue), la lecture orale d'un·e apprenti·e lecteur·rice présente des évènements spécifiques potentiellement néfastes pour la RAP. Quels évènements ont un impact négatif significatif sur la qualité de la reconnaissance, et comment surmonter ces difficultés ?

Nous avons commencé par établir un état de l'art dédié aux caractéristiques acoustiques, linguistiques et prosodiques de la parole d'enfants. Nous avons également étudié les précédents travaux portant sur la RAP d'enfants, et cherché à comprendre le lien entre les caractéristiques de la parole d'enfants et les dégradations des performances observées par rapport à la parole d'adultes. Des analyses qualitatives ont de plus été effectuées tout au long de la thèse pour comprendre quels aspects de la parole d'enfants posent problème aux modèles acoustiques.

Nous nous sommes ensuite lancé·e·s dans l'implémentation de modèles acoustiques pour la reconnaissance de phonèmes dans la parole d'enfants. Axées dans un premier temps sur les approches hybrides, nos recherches ont porté sur le modèle TDNNF-HMM (*Factorised Time-Delay Neural Networks - Hidden Markov Model*) [Povey 2018]. Nous avons exploré différentes stratégies d'entraînement, et en particulier l'utilisation d'apprentissage par transfert, visant à pallier à un manque de données par adaptation d'un modèle entraîné sur parole d'adultes avec de la parole d'enfants.

Nous nous sommes ensuite tourné·e·s vers les architectures *end-to-end*. Nous avons sélectionné différentes méthodes de modélisation : réseaux de neurones récurrents (*Recurrent Neural Networks*, RNN) [Goodfellow 2016], modules CTC (*Connectionist Temporal Classification*) [Graves 2006] et mécanismes d'attention [Graves 2014a]. À partir de ces méthodes, un large panel d'architectures *end-to-end* a été constitué pour fournir une comparaison détaillée entre approches hybrides et *end-to-end* pour la reconnaissance de phonèmes dans la parole d'enfants AL. Cela nous a permis de :

— comparer l'efficacité de l'apprentissage par transfert en fonction du type de modélisation (hybride ou *end-to-end*), et voir si cette efficacité dépend aussi des méthodes utilisées (RNN, CTC, attention) dans les modèles *end-to-end* ;

— comparer les différentes architectures et étudier les avantages et inconvénients des méthodes de modélisation choisies, seules ou combinées, en particulier pour notre parole spécifique ;

— observer le comportement de nos modèles face à deux tâches de lecture orale (mots isolés et phrases), proposées aux enfants en fonction de leur niveau de lecture, et impliquant des longueurs d'enregistrements diverses pouvant affecter certains modèles ;

— déterminer enfin si un modèle *end-to-end* parvient à surpasser notre modèle hybride TDNNF-HMM.

La question 5 a fait l'objet d'analyses détaillées quant aux performances de nos meilleurs modèles en fonction du niveau de bruit de brouhaha de salle de classe. De la même façon, pour répondre à la question 6, nous avons analysé l'influence de la qualité de lecture d'un·e enfant sur la précision de la reconnaissance de phonèmes. Nous nous sommes pour cela appuyé·e·s sur plusieurs indicateurs, choisis à partir de connaissances pédagogiques et d'observation des données : niveau de bruit, vitesse de lecture et présence d'erreurs de fluence et de déchiffrage. Ayant constaté des influences notables, nous avons travaillé sur deux méthodes d'augmentation de données, avec pour objectifs à la fois de pallier au manque de données de parole d'enfants et d'améliorer la robustesse de nos modèles aux spécificités de la parole à traiter :

— une technique d'augmentation par ajout de bruit de brouhaha, utilisant différents types de brouhaha et différents niveaux de bruit ;

— une technique novatrice d'augmentation des données audio par simulation d'erreurs de lecture, imaginée pour améliorer la robustesse du modèle en présence de ces évènements, fréquents dans la lecture orale d'AL. Nous avons fourni une analyse détaillée de l'impact de cette technique en fonction de la sévérité et du type d'erreur : erreur de fluence (répétition de mots) ou erreur de déchiffrage (substitution, insertion ou suppression d'un phonème).

Organisation du manuscrit

Ce manuscrit se compose de 7 chapitres, répartis en 3 grandes parties.

La première partie est constituée des deux premiers chapitres et correspond à la mise en contexte, industriel et scientifique, de cette thèse.

Le chapitre 1 commence par exposer le contexte industriel en présentant l'entreprise Lalilo et son produit, un assistant pédagogique fondé sur de l'intelligence artificielle visant à aider les professeur·e·s à « différencier » leur enseignement de la lecture, c'est-à-dire à procurer à chaque enfant un enseignement personnalisé à ses besoins. Nous étudions ensuite l'importance de l'oralisation pour les apprenant·e·s lecteur·rice·s et justifions l'utilité de systèmes de reconnaissance vocale pour l'apprentissage de la lecture. Nous présentons enfin l'exercice de lecture orale de la plateforme Lalilo qui nous intéressera particulièrement dans cette thèse, puisque fondé sur un tel système, dont l'objectif de notre travail en sera l'amélioration.

Le chapitre 2 établit l'état de l'art dans le domaine de la RAP, incluant les approches classiques hybrides et les approches *end-to-end*, plus récentes, et qui prendront une part importante dans notre travail. Dans une seconde section, nous regroupons les principales études attestant des différences acoustiques, linguistiques et prosodiques entre les paroles d'adultes et d'enfants. Nous présentons enfin les avancées dans le domaine de la reconnaissance de parole d'enfants : mise à disposition de données de parole d'enfants (principalement en langue anglaise), création de tuteurs de lecture et autres projets fondés sur la RAP à visées éducatives, et améliorations notables des modèles acoustiques pour cette parole spécifique.

La seconde partie vise à l'établissement d'un modèle de reconnaissance de phonèmes de référence pour la suite de nos travaux, et contient les chapitres 3 et 4.

Le chapitre 3 expose les méthodes qui seront testées, puis adoptées ou non, pour la création de notre modèle de référence dans le chapitre 4. Nous présentons d'abord plusieurs paramètres acoustiques, postulant que les paramètres classiquement utilisés pour la parole d'adultes ne sont pas forcément les plus pertinents pour la parole d'enfants. Nous décrivons ensuite deux méthodes permettant de pallier à notre manque de données de parole d'enfants. Enfin, nous détaillons les différents éléments composant l'architecture de type hybride de ce modèle, ainsi que les procédures d'entraînement et de décodage.

Le chapitre 4 évalue ces différentes méthodes sur notre corpus originel de lecture orale d'enfants français *Lalil-o-riginel*. Nous décrivons la création de ce corpus : moyens mis en place pour collecter des enregistrements, création du contenu à faire lire aux enfants, déroulement des séances d'enregistrement en école, annotation des données en considérant les différents types d'erreurs de lecture présents dans la parole d'AL. Nos expérimentations amènent finalement au choix des méthodes à appliquer pour obtenir le meilleur modèle acoustique de type hybride, qui servira de référence tout au long de ce manuscrit.

La troisième et dernière partie se compose des chapitres 5, 6 et 7 et présente les principales contributions de cette thèse, axées sur les approches de modélisation *end-to-end* de phonèmes.

Le chapitre 5 présente les méthodes de modélisation et architectures *end-to-end* qui seront testées dans le chapitre 6. Nous exposons d'abord notre choix de méthodes de modélisation pour la transcription phonétique de parole lue d'enfants, puis nous présentons rapidement les techniques d'inférences classiques. Nous décrivons cinq architectures de modèles *end-to-end*, choisies pour leur diversité : chacune utilise une ou plusieurs des méthodes de modélisation choisies.

Le chapitre 6 évalue ces architectures sur un nouveau corpus, nommé *Lalil-o-fficiel*, ayant été rééquilibré avec de nouvelles données. Nous comparons nos modèles sur la parole d'adultes, puis sur la parole d'enfants avec différentes procédures d'entraînement : sur parole d'adultes uniquement, sur parole d'enfants uniquement, puis avec un apprentissage par transfert d'adultes à enfants. Enfin, des analyses détaillées sont proposées, dans lesquelles nous observons l'influence de certaines caractéristiques de nos données, liées à notre application de reconnaissance de la lecture orale d'AL : vitesses de lecture hétérogènes et présence d'erreurs de lecture.

Ayant montré dans le chapitre 6 que la présence d'erreurs de lecture avait une forte influence sur les performances de notre meilleur modèle acoustique, nous introduisons dans le chapitre 7 une méthode novatrice d'augmentation de données par simulation d'erreurs de lecture. Cette méthode vise à améliorer la robustesse quant à ce type d'évènements, très fréquents dans la parole d'AL. Nous reprenons les analyses détaillées du chapitre 6 quant à la présence d'erreurs de lecture en ajoutant notre nouveau modèle augmenté, et étudions l'impact de cette méthode. Nous entraînons également un modèle sur des données augmentées par ajout de bruit de salle de classe, et en évaluons la robustesse au bruit de brouhaha. Enfin, nous combinons ces deux méthodes d'augmentation de données afin d'étudier leur complémentarité.

Première partie

Contexte de la thèse et état de l'art

Chapitre 1

Lalilo, et les enfants apprenant·e·s lecteur·rice·s

Dans ce premier chapitre, nous présentons Lalilo, entreprise ayant financé cette thèse, développant un assistant pédagogique pour les professeur·e·s de CP à CE2 afin de les aider à diversifier leur enseignement de la lecture. Nous étudions également le processus d'oralisation de la lecture, son importance pour l'enfant apprenant à lire, et les pratiques de lecture orale utilisées par des enseignant·e·s d'écoles françaises. À cette étude nous lions l'utilité de la reconnaissance automatique de la parole, et présentons l'exercice de lecture orale de Lalilo. Le système global derrière cet exercice, utilisant la RAP pour fournir un retour à l'enfant sur la qualité de sa lecture, est enfin détaillé.

Sommaire

1.1 Présentation de Lalilo

1.1.1 Lalilo, l'entreprise

En France, les compétences en lecture et compréhension des écolier·ère·s français·e·s en classe de CM1 sont en baisse par rapport au début des années 2000, et la France se situe au 34ème rang sur 50 pays étudiés dans le classement établi par l'étude internationale PIRLS (*Progress in International Reading Literacy Study*) en 2016 [1]. Toujours d'après cette étude, 40,5% des élèves français·e·s de 15 ans ne maîtrisent pas la lecture, et 21,5% sont même en grande difficulté. Les écarts de niveau entre les élèves les plus performant·e·s et les moins performant·e·s sont très importants, d'après une enquête de 2015 du Programme International pour le Suivi des Acquis des élèves (PISA) [2]. L'enquête PISA de 2018 rapporte de plus que les performances des 10% d'élèves les plus avantagé·e·s socio-économiquement sont en moyenne bien supérieures à celles des 10% d'élèves les moins avantagé·e·s, la différence correspondant à environ quatre ans de scolarisation [3]. Un rapport de 2007 de l'Observatoire National de la Lecture [4] indique qu'entre 10 et 15% des élèves sont en grande difficulté à l'entrée au collège selon les études, et que les problèmes de lecture et de compréhension sont causes d'échec dans de nombreuses matières.

Lalilo [5] est née de ces constatations : fondée en 2016 par trois diplômés de l'école Polytechnique de Paris, elle vise à combattre l'illettrisme grâce à des outils numériques d'aide à l'apprentissage de la lecture. Le projet a démarré dans un programme d'accélération de création d'entreprise à San Francisco, puis la start-up a effectué une première levée de fonds en 2017 et s'est installée dans un incubateur à Paris avec les premier·ère·s employé·e·s. Se développant en parallèle sur les marchés des États-Unis et de la France, une branche a été créée à San Francisco, puis à New York. Une deuxième levée de fonds a été faite en 2019, permettant de compléter l'équipe. Lalilo emploie aujourd'hui 40 français·e·s et américain·e·s, réparti·e·s dans des équipes pédagogique, technique, de design, de vente et de marketing. La start-up a été rachetée en mars 2021 par Renaissance Learning [6], une entreprise américaine spécialisée dans les technologies de l'éducation, possédant plusieurs produits pour l'apprentissage et l'évaluation de la lecture et des mathématiques.

Disponible gratuitement depuis le début, la plateforme Lalilo est devenue en partie payante aux États-Unis en 2019. En France, le Ministère de l'Éducation Nationale, de la Jeunesse et des Sports lance le Partenariat d'Innovation et Intelligence Artificielle [7], qui propose un financement pour le développement de solutions au service des apprentissages fondamentaux en français et mathématiques au cycle 2 (CP, CE1, CE2). Lalilo est choisie en 2019 pour être une des lauréates sur la partie Français de ce partenariat.

1. https://colibris.link/pirls-2016
2. https://colibris.link/pisa-2015
3. https://colibris.link/pisa-2018
4. «La lecture au début du collège», téléchargeable ici : https://colibris.link/ONL-2007
5. https://www.lalilo.com
6. https://www.renaissance.com
7. https://colibris.link/p2ia

1.1.2 Lalilo, le produit

Lalilo développe un assistant pédagogique pour les enseignant·e·s de CP, CE1 et CE2, afin de les aider à différencier leur enseignement de la lecture en fonction de l'avancement de chaque élève, grâce à des technologies d'intelligence artificielle. L'assistant est une plateforme en ligne, avec une interface élève et une interface professeur·e. Il est co-construit avec les professeur·e·s des écoles : l'amélioration de la plateforme est continue, prenant en compte les retours des enseignant·e·s. La plateforme a été utilisée en France durant l'année scolaire 2020-2021 par 46 000 professeur·e·s et 271 000 élèves, pour un total de plus de 35 millions d'exercices joués.

L'enseignant·e s'inscrit et crée sa classe : chaque élève possède alors un compte personnel, où son avancement est individuel. L'outil est principalement dédié à l'utilisation en classe : l'enseignant·e peut y faire travailler en autonomie un petit groupe d'élèves et s'occuper des autres pendant ce temps. Les enfants peuvent également se connecter à la maison grâce à un code parent, ce qui leur permet de revoir les notions apprises en s'entraînant le soir ou le week-end.

1.1.2.1 Interface élève

Au début de son parcours sur Lalilo, l'enfant effectue un test de positionnement afin de déterminer son niveau dans la progression pédagogique de Lalilo. Des exercices courts lui sont proposés, dont la difficulté s'adapte au fur et à mesure de ses réponses grâce à un système d'intelligence artificielle. Ce système, appelé dans la suite système d'apprentissage adaptatif et décrit dans la section suivante, est utilisé dans toute la progression de l'élève pour lui permettre d'avancer à son propre rythme. La figure 1.1 présente un exemple d'exercice de niveau intermédiaire où l'enfant doit compléter la phrase avec le mot qui lui est lu oralement (« texte »). Les lettres muettes sont grisées pour faciliter la lecture. La jauge sur la droite de l'écran sert à l'enfant à visualiser sa progression.

Afin de rendre l'apprentissage de la lecture ludique, l'enfant progresse sur un chemin qui traverse plusieurs mondes (voir figure 1.2). L'élève gagne également des récompenses (voir figure 1.3) : trésors en rapport avec les mondes traversés, badges représentant les leçons maîtrisées, et petites histoires à lire et à relire.

1.1.2.2 Algorithme d'apprentissage adaptatif

L'objectif de Lalilo, aider les enfants à apprendre à lire à leur propre rythme, repose sur une progression pédagogique fixée, divisée en leçons. Chaque leçon contient plusieurs « objectifs d'apprentissage », qui contiennent à leur tour une large gamme de questions. Par exemple, la leçon « voyelle [a] » contient deux objectifs d'apprentissage : « reconnaître le graphème a » et « décoder le phonème [a] ». Une question pour le premier objectif pourrait être de cliquer sur les graphèmes « a » apparaissant au milieu d'autres graphèmes. Un exercice est formé de trois à sept questions appartenant à la même leçon et au même objectif d'apprentissage. Chaque

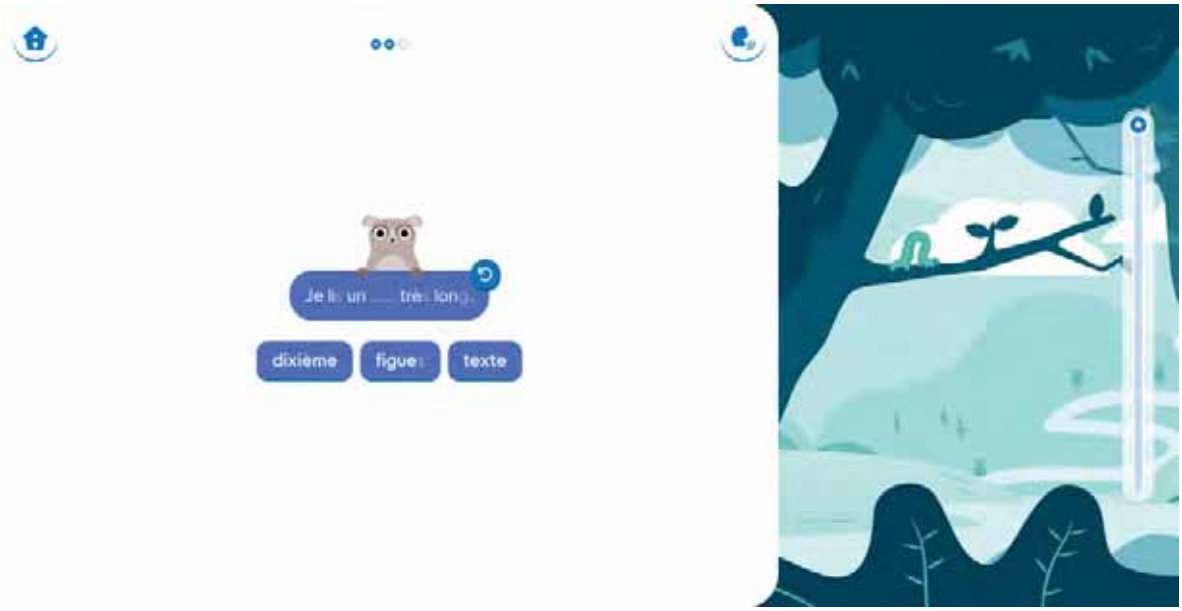


FIGURE 1.1 – Interface élève : dans cet exercice, l'enfant doit compléter la phrase avec le mot qu'il entend (« texte » ici)

FIGURE 1.2 – Mondes de Lalilo : forêt, montagne, océan, désert, campagne, savane, banquise, volcans, jungle et espace

élève travaille sur plusieurs leçons en parallèle, ce qui permet de varier le contenu, de garder l'élève motivé·e, et de favoriser un apprentissage pérenne. Lors d'une session sur Lalilo, l'élève reçoit néanmoins toujours plusieurs exercices de la même leçon à la suite, ce qui permet une compréhension en profondeur. Pour valider un objectif d'apprentissage, l'enfant doit avoir 80% de réussite en moyenne sur un nombre minimum d'exercices (qui varie en fonction de l'objectif). La leçon est validée lorsque tous les objectifs d'apprentissage de cette leçon le sont aussi.

Le système d'apprentissage adaptatif permet de choisir le meilleur exercice pour un enfant à un moment donné en fonction de sa progression et des exercices effectués. Nous utilisons la notion de zone proximale de développement (ZPD), qui correspond à l'intervalle entre le niveau de développement actuel de l'enfant, déterminé par ses capacités à résoudre seul des problèmes, et son niveau de développement potentiel, déterminé par ses capacités à résoudre

FIGURE 1.3 – Récompenses que l'élève peut gagner : trésors, badges et histoires

des problèmes aidé d'un adulte ou de pairs initiés [Vygotsky]. Les experts pédagogiques de Lalilo ont fixé cette ZPD autour de 75%, ce qui signifie qu'un élève doit avoir 75% de chances de réussir la question pour que sa progression soit maximale. La question ne doit donc pas être trop facile pour que l'enfant apprenne quelque chose, mais pas trop dure non plus pour ne pas le décourager. Deux algorithmes sont utilisés pour estimer chacun une probabilité de succès de l'élève à un panel de questions possibles. Le premier algorithme, nommé Catboost [8], est fondé sur du machine learning et prédit le succès d'un·e élève sur une question en apprenant sur des millions d'exercices joués. Le second algorithme, nommé ELO [Pelanek 2016], prédit le succès d'un·e élève sur une question en fonction de deux paramètres, qui se mettent à jour continuellement : le niveau global de l'élève et la difficulté de la question. La moyenne des probabilités obtenus par ces deux algorithmes pour chaque question est calculée, et la question dont la probabilité de succès finale s'approche le plus de 75% est choisie.

1.1.2.3 Interface enseignant·e

L'enseignant·e dispose d'une interface à part contenant toutes les informations nécessaires au suivi de chaque élève individuellement. La figure 1.4 montre le tableau de bord général, permettant de visualiser en un coup d'œil l'avancée de chaque élève dans la progression pédagogique de Lalilo. D'autres tableaux de bord sont également disponibles à des niveaux plus précis : nombre de leçons vues travaillées dans la semaine, temps d'usage, rapport détaillé des exercices et erreurs faites par chaque élève. Les enseignant·e·s peuvent également être alertés lorsqu'un·e ou plusieurs enfants sont en difficulté sur une leçon, ce qui leur permet d'organiser une séance de remédiation sur cette leçon avec ce petit groupe d'élèves. Bien que l'algorithme d'apprentissage adaptatif se charge de fournir à chaque élève des exercices personnalisés, les enseignant·e·s ont la possibilité d'attribuer des leçons spécifiques, par exemple sur la correspondance graphème-phonème [a], pour que les élèves s'entraînent sur la théorie vue en classe.

8. https://catboost.ai/

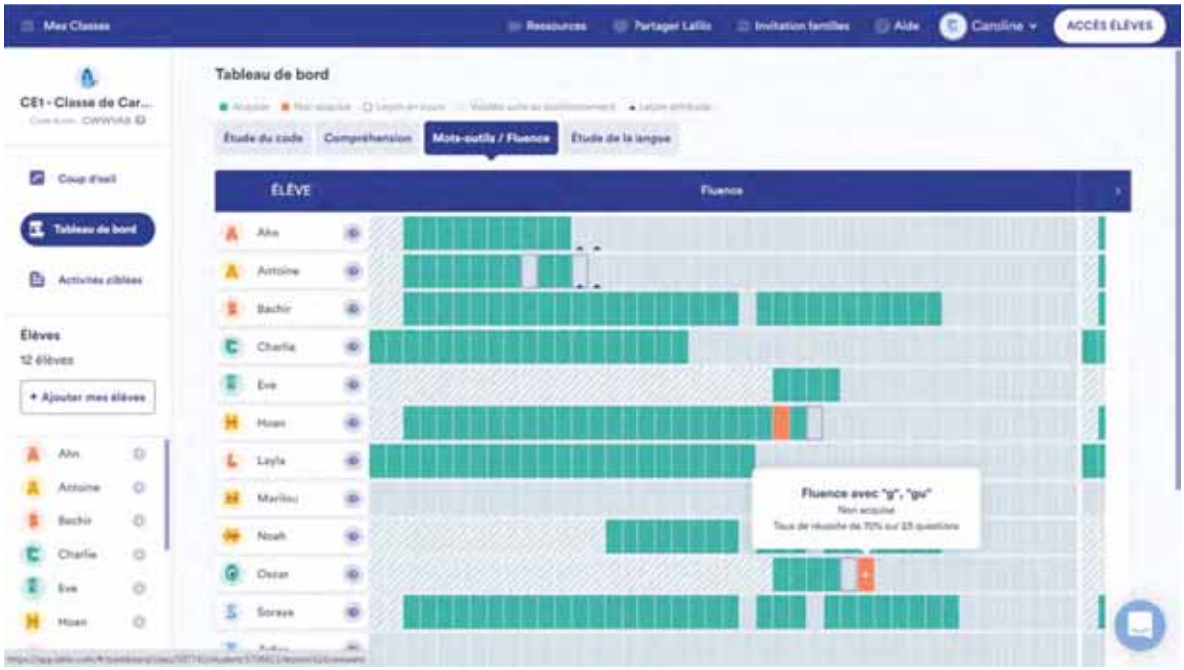


FIGURE 1.4 – Interface enseignant·e : tableau de bord permettant de visualiser l'avancée de chaque élève dans la progression pédagogique de Lalilo

1.2 L'oralisation pour l'apprentissage de la lecture

L'apprentissage de la lecture chez l'enfant passe par l'oralisation. Or, l'oralisation de la lecture en classe est chronophage, car elle nécessite d'être réalisée de manière individuelle et séquentielle, c'est-à-dire un·e élève après l'autre. Elle est donc aujourd'hui très peu pratiquée en classe alors qu'elle est essentielle pour apprendre à lire.

1.2.1 Importance de l'oralisation pour l'enfant

D'après [Beaume 1987], il faut distinguer deux termes : la lecture orale et la lecture à voix haute.

La lecture orale désigne le fait pour l'enfant d'oraliser sa lecture avant d'être capable de lire silencieusement. Elle permet à l'apprenant·e de transformer les signes écrits en sons, et de construire le sens de ce qu'il ou elle lit par l'écoute de ces sons. L'oralisation précède et permet la compréhension. Lorsque l'enfant arrive à déchiffrer seul·e la plupart des mots rencontrés, la lecture orale guidée par un·e adulte permet d'améliorer significativement la capacité de l'enfant à reconnaître un mot, ainsi que sa fluence et sa compréhension, d'après le rapport de 2000 du *National Reading Panel*[9], une institution du département de l'éducation des États-Unis. Cette observation s'applique aussi bien aux bon·ne·s lecteur·rice·s qu'aux élèves en difficulté,

9. National Reading Panel Report 2000 : Teaching children to read (https://colibris.link/NRP-2000)

sur une large fourchette de niveaux de lecture.

La lecture à voix haute n'est possible qu'à un niveau plus avancé, car elle est précédée d'une lecture silencieuse visant à la compréhension du texte, et consiste ensuite à lire ce texte avec restitution de sens, pour soi-même ou pour un auditoire [Beaume 1987]. L'enseignement de la fluence par des exercices de lecture à voix haute a montré des effets positifs sur le niveau global de lecture des élèves en bénéficiant, d'après le rapport de 2005 du *National Reading Panel*[10]. La lecture à voix haute d'un texte permet de travailler la maîtrise du texte, la prononciation, l'articulation, l'intonation, l'interprétation et la restitution du sens du texte [Ros-Dupont 1999]. Enfin, d'après le même ouvrage, la lecture à voix haute joue un rôle important dans la motivation de l'élève à apprendre à lire : être capable de lire un texte à un·e parent·e, un·e ami·e ou un petit frère ou petite sœur est source de satisfaction pour l'enfant.

Concernant les modalités d'enseignement de la lecture orale ou à voix haute, il a été observé que la répétition permet d'améliorer la qualité de la lecture [Levy 1995]. Une amélioration dans la vitesse de lecture et la compréhension a été notée lorsqu'un·e élève lit un passage de façon répétée [Stoddard 1993]. Enfin, il est important pour l'élève de bénéficier d'un retour sur sa lecture, par un·e enseignant·e ou un·e parent·e[10].

Cette étude nous confirme l'importance pour les élèves de s'entraîner à lire oralement, puis à voix haute pour développer leurs capacités de lecture. Nous nous intéressons dans cette thèse principalement à la lecture orale, puisque nos données de parole proviennent d'enfants utilisateurs de Lalilo en classes de CP, CE1 et CE2, dans lesquelles le niveau de lecture des élèves nécessite encore souvent de passer par la lecture orale avant de pouvoir lire silencieusement. Nous étudierons dans notre travail les particularités acoustiques et linguistiques de la lecture orale de jeunes enfants et leur impact sur la précision de la reconnaissance automatique de la parole.

1.2.2 Pratiques de lecture orale en classe

Afin de comprendre les pratiques de lecture orale à l'école en France, et en déduire des applications utiles de la reconnaissance vocale pour l'apprentissage de la lecture, j'ai créé un sondage et l'ai envoyé à plus de 600 professeur·e·s utilisant Lalilo. Nous avons reçu 79 réponses, principalement d'enseignant·e·s de CP (49,4%), de classes mixtes CP-CE1 (13,4%), de CE1 (7,6%) ainsi que de classes ULIS (Unités Spécialisées d'Inclusion Scolaire), des classes spécialisées pour enfants avec un handicap cognitif (8,9%). Nous présentons ici les résultats de ce sondage.

Les graphiques de la figure 1.5 présentent le temps passé en moyenne par semaine à apprendre à lire en classes de CP et CE1. Sachant que les élèves passent environ 30 heures en classe chaque semaine, nous pouvons constater que l'apprentissage de la lecture occupe une part très importante de l'enseignement, notamment en CP. Cela est d'autant plus vrai que ces temps ne représentent que l'enseignement spécifique de la lecture, alors qu'en réalité les élèves

10. National Reading Panel Report 2005 : Practical advice for teachers (https://colibris.link/NRP-2005)

s'entraînent à lire dans chaque matière, par la lecture et compréhension des énoncés.

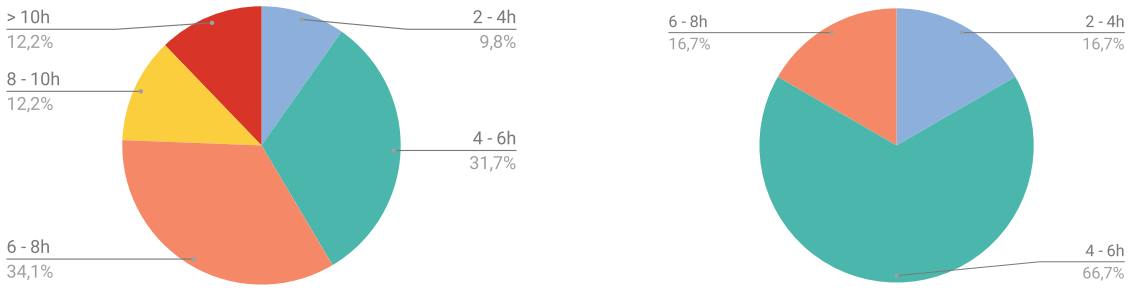


FIGURE 1.5 – Temps (en heures) passé par semaine à l'enseignement spécifique de la lecture en classe de CP (gauche) et CE1 (droite)

Paradoxalement, nous voyons sur la figure 1.6 que la moitié des professeur·e·s évaluent la lecture à voix haute de leurs élèves individuellement moins d'une fois par mois, ce qui ne paraît pas suffisant pour suivre les progrès et difficultés de chaque élève. Une proportion significative (20,3%) des répondant·e·s ont déclaré évaluer leurs élèves continuellement à partir des interventions de chacun dans l'enseignement commun, mais pas forcément de façon individuelle. Ces observations pourraient témoigner d'un manque de temps et d'outils dédiés pour l'évaluation individuelle de la lecture à voix haute.

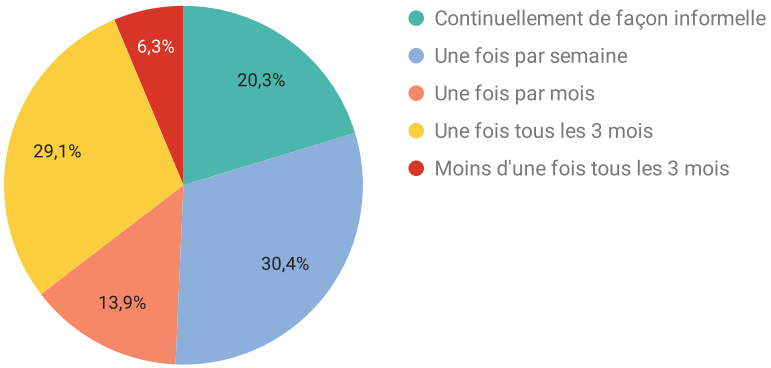


FIGURE 1.6 – Fréquence d'évaluation individuelle de la lecture orale

Les répondant·e·s devaient également choisir les tâches de lecture orale données aux élèves individuellement pour évaluer leur niveau de lecture. Nous voyons sur la figure 1.7 que les tâches les plus courantes sont la lecture orale d'histoire, avec support visuel et sans questions de compréhension, et la lecture orale de liste de mots, chronométrée ou non. La lecture de phrases ou de mots isolés sont également donnés, mais plutôt pour l'entraînement et non pour l'évaluation, qui nécessite un contenu assez long.

Des questions sur la pertinence de faire un retour à l'enfant ont été posées : 86,1% des répondant·e·s trouvent utile de faire un retour à l'enfant à la fin de sa lecture, et 20,3% au moment de l'erreur. Cela dépend évidemment du contenu à lire (mots, phrases, histoire) et du cadre de la lecture : en entraînement, le retour est préféré relativement immédiat (après le

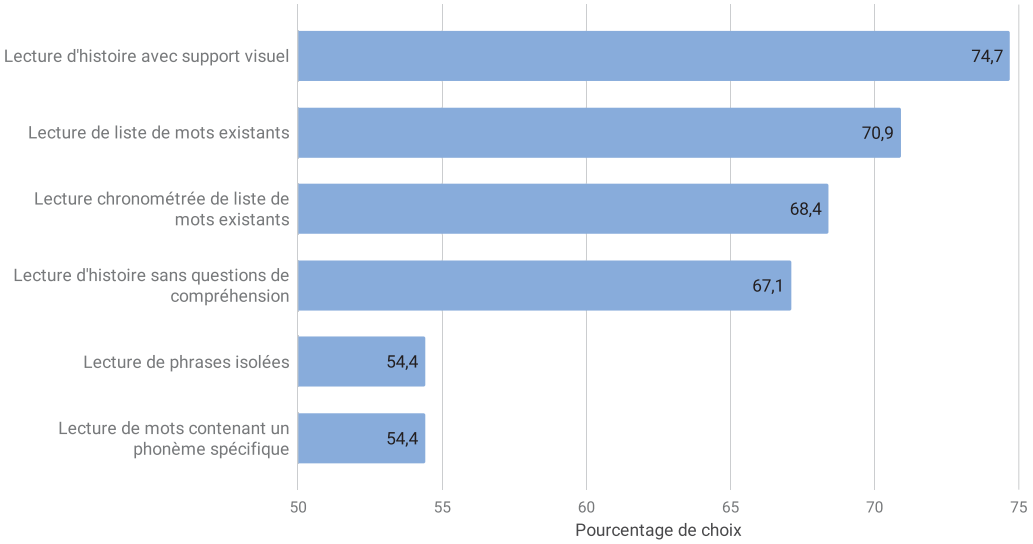


FIGURE 1.7 – Tâches de lecture orale données aux élèves

mot ou la phrase) alors qu'en évaluation le retour se fait généralement en fin de lecture afin de ne pas perturber l'élève dans sa lecture. 85,9% des répondant·e·s souhaitent un retour sonore, avec lecture du mot correct à l'élève, et 71,8% un retour visuel, par exemple en affichant les mots contenant des erreurs en rouge.

Enfin, les enseignant·e·s ont dû choisir, parmi une liste de métriques pour l'évaluation du niveau de lecture orale créée avec l'aide de l'équipe pédagogique de Lalilo, les cinq métriques qu'il·elle·s utilisent le plus souvent. Les réponses, détaillées figure 1.8, montrent que la métrique la plus utilisée est le nombre de mots corrects par minute, suivie de près par le pourcentage de succès (nombre de mots correctement lus divisé par nombre de mots lus). Un score d'expressivité est également beaucoup utilisé, évalué de façon subjective par l'enseignant·e, ainsi que le temps moyen pour lire un mot, mesuré entre la fin du mot précédent et la fin du mot courant afin d'inclure le temps de réflexion entre mots. Arrivent ensuite diverses métriques mesurant un nombre d'évènements : auto-corrections montrant que l'enfant a compris son erreur, substitutions/suppressions de phonèmes ou mots, hésitations intra- ou inter-mots... Les insertions et évènements liés à la fluence (faux départs et répétitions) semblent moins importants pour l'évaluation de la lecture orale. Ces réponses sont bien sûr liées aux classes enseignées par les répondant·e·s : les erreurs de fluence auraient sûrement plus d'importance pour des enseignant·e·s de niveaux plus élevés.

Ce sondage nous confirme l'importance de la lecture orale dans les classes de CP, CE1 et CE2, et suggère que les enseignant·e·s pourraient bénéficier d'outils dédiés pour les aider à faire lire oralement leurs élèves. Il semble également important de donner un retour à l'enfant sur sa lecture, ce que l'enseignant·e n'a pas toujours le temps de faire régulièrement avec chaque enfant de façon individuelle, mais qu'un système de reconnaissance vocale pourrait fournir. Nous apprenons également quelles métriques utilisent les enseignant·e·s pour évaluer le niveau de lecture d'un·e élève, ce qui nous permet de réfléchir à la possibilité de générer ces

métriques à partir d'un système de reconnaissance automatique de la parole.

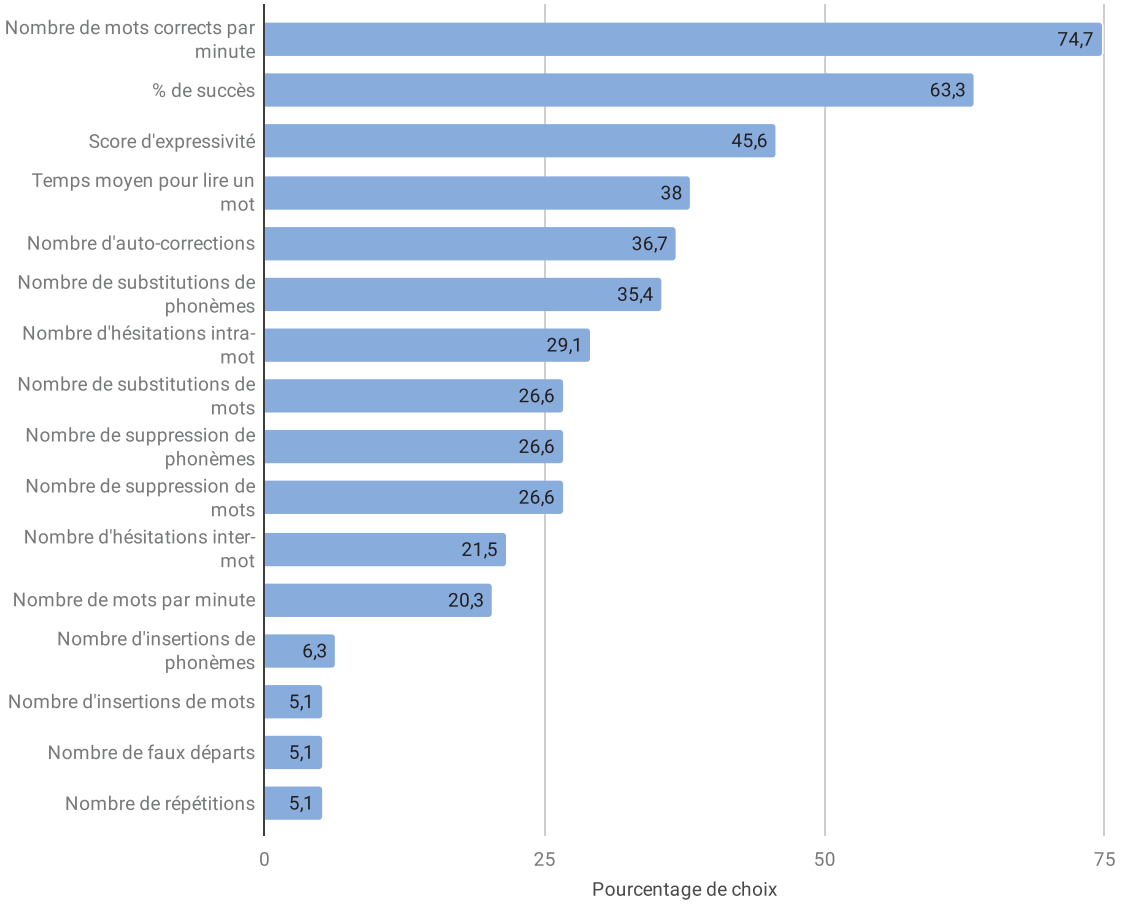


FIGURE 1.8 – Métriques utilisées pour évaluer le niveau de lecture orale des élèves

1.3 La reconnaissance vocale pour l'apprentissage de la lecture

La reconnaissance automatique de la parole paraît apte à répondre aux différents besoins des élèves et enseignant·e·s, étudiés dans les sections précédentes. La plateforme Lalilo contient un exercice de lecture orale, qui nous intéressera tout particulièrement dans cette thèse. Il utilise en effet la reconnaissance automatique de la parole de l'enfant pour lui faire un retour sur la qualité de sa lecture. L'amélioration de la précision de cet exercice, par l'amélioration du système de reconnaissance automatique de la parole, est donc l'objectif premier de cette thèse.

1.3.1 Utilité de la reconnaissance vocale pour la lecture orale en classe

Un exercice de lecture orale permet tout d'abord de fournir un outil aux élèves pour s'entraîner à lire oralement sans subir la pression de la présence d'un public (enseignant·e ou camarades) qui peut être intimidante. Le système de RAP permet de fournir un retour à l'enfant, essentiel d'après le rapport du *National Reading Panel*[10] et les résultats de notre sondage.

Parmi les métriques couramment utilisées par les enseignant·e·s (voir figure 1.8), une grande partie sont mesurables à partir des sorties d'un système de RAP : nombre de mots corrects par minute, pourcentage de succès, temps moyen pour lire un mot, nombre d'évènements spécifiques... Outre la qualité de lecture de l'enfant, certaines métriques de fluence ont montré des capacités à prédire avec précision des scores de compréhension écrite [Kim 2011, Sabatini 2019]. Elles pourraient ainsi être utilisées pour identifier en amont des élèves risquant d'avoir des difficultés en compréhension [Roehrig 2008] et leur offrir un soutien supplémentaire dès le plus jeune âge.

Dans le cadre de notre sondage pédagogique, et dans la perspective d'améliorer l'exercice de lecture orale de Lalilo pour répondre aux besoins des professeur·e·s, nous leur avons demandé quelles informations et fonctionnalités supplémentaires leur seraient utiles. Nous voyons sur la figure 1.9 que les trois réponses les plus fréquentes relèvent d'analyses détaillées des lectures orales d'un élève à partir des sorties du système de RAP. Ces analyses dépendent de la qualité de la reconnaissance : le système doit non seulement être capable de détecter que l'enfant a fait une erreur, mais aussi de savoir sur quel mot et de reconnaître l'erreur faite (substitution d'un [b] par un [d], lecture de « ent » muet à la fin d'un mot...). 67,1% de répondant·e·s ont également manifesté leur intérêt pour un accès aux enregistrements de leurs élèves, afin de pouvoir écouter leur lecture et en évaluer la qualité ell·eux-mêmes. Cette fonctionnalité a été ajoutée à l'interface professeur·e en 2021 (voir figure 1.10).

1.3.2 L'exercice de lecture orale de Lalilo

L'exercice de lecture orale existe dans la plateforme depuis 2018, et a été amélioré de façon itérative grâce aux retours des utilisateur·rice·s de Lalilo. Le contenu proposé à la lecture orale a également évolué : dans un premier temps uniquement des mots isolés étaient proposés, puis des phrases de quatre à huit mots. Récemment ont été ajoutées des listes de syllabes et de mots et des phrases longues pour correspondre aux besoins des enseignant·e·s (voir figure 1.7), ainsi que des lettres et syllabes isolées pour les plus jeunes enfants. Parmi les 35 millions d'exercices joués en France durant l'année scolaire 2020-2021, plus de 400 000 étaient des exercices de lecture orale.

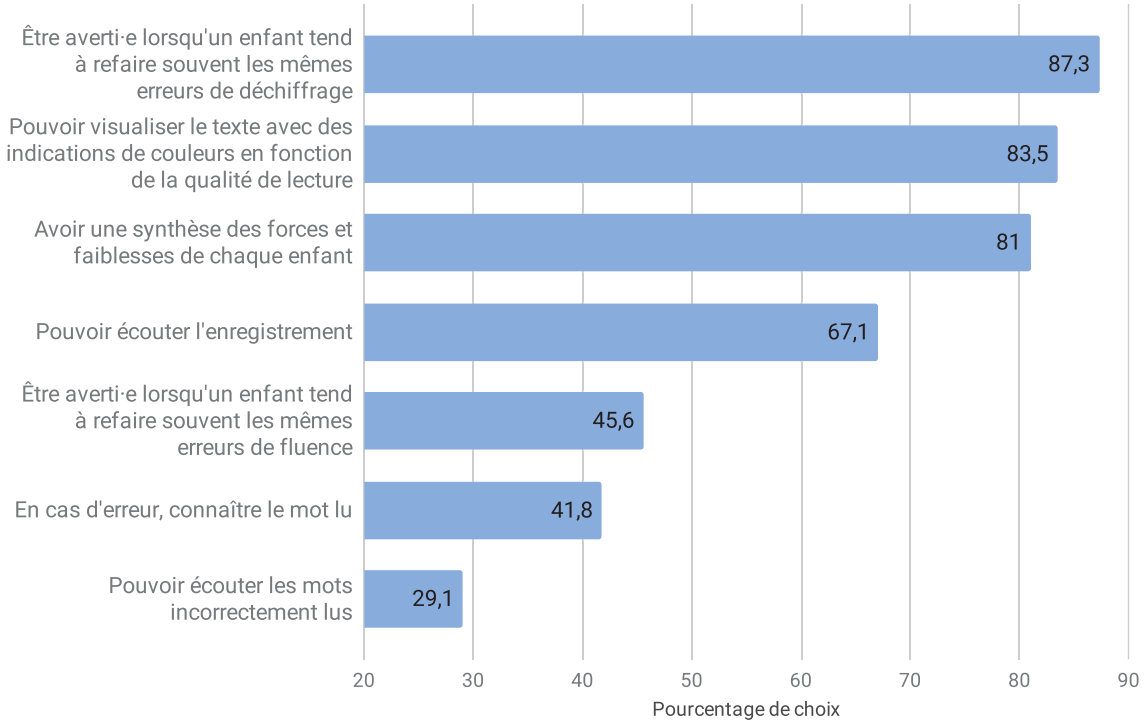


FIGURE 1.9 – Fonctionnalités s'appuyant sur la RAP considérées utiles par les professeur·e·s dans le cadre d'un exercice de lecture orale sur Lalilo

FIGURE 1.10 – Accès aux enregistrements de mots isolés d'un·e élève via le tableau de bord de l'enseignant·e, avec le retour qui lui a été fait (de haut en bas : incertain, incorrect, incorrect, correct)

L'exercice se déroule de la façon suivante :

1. Le mot ou la phrase à lire est affichée pendant que l'enfant écoute l'instruction de l'exercice ;
2. L'enfant clique sur le microphone et lit à voix haute ;
3. L'enfant écoute son enregistrement et indique s'il·elle est content·e de sa lecture ou non. Sinon, l'enfant a la possibilité de se réenregistrer ;
4. Une fois sa réponse validée, que la lecture soit correcte ou non, l'enfant entend la lecture correcte du contenu à lire.

Le contenu à lire est choisi par l'algorithme d'apprentissage adaptatif : en fonction de son niveau, l'enfant doit lire un certain contenu, et le(s) mot(s) à lire correspondent à une leçon en cours d'acquisition par l'enfant.

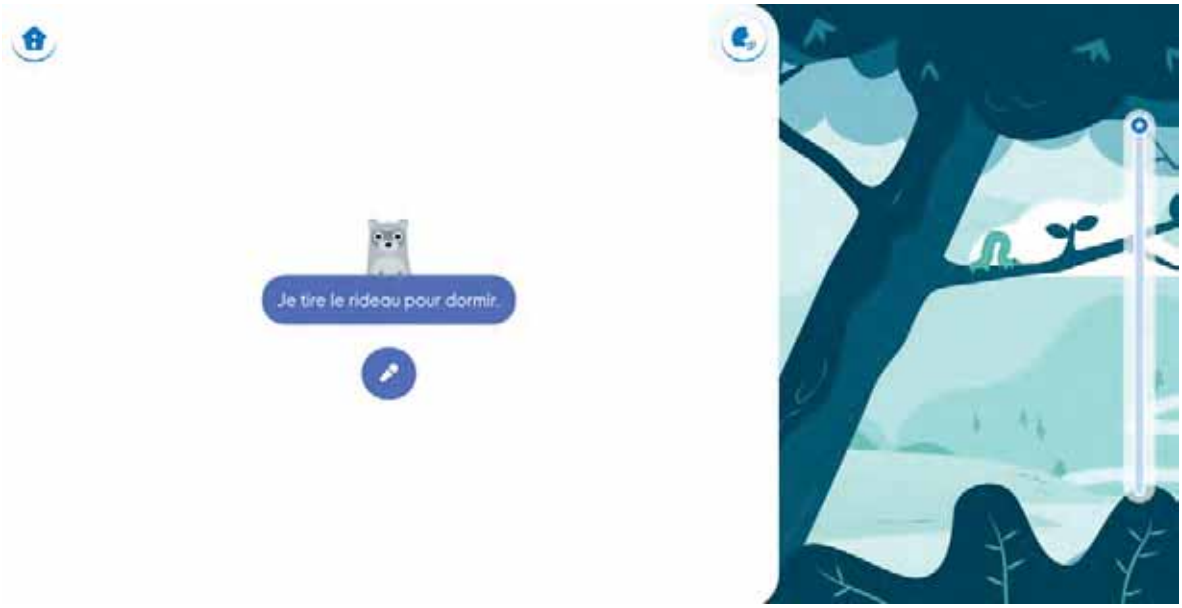


FIGURE 1.11 – Exercice de lecture à voix haute : l'enfant doit cliquer sur le microphone et lire le mot à voix haute

Le fonctionnement du système global de RAP permettant de faire un retour à l'enfant dans l'exercice de lecture à voix haute est décrit sur la figure 1.12. Nous avons accès à deux sources d'informations à l'entrée : le texte à lire et l'enregistrement de la lecture de l'élève. Tout d'abord, une étape de pré-traitement consiste à extraire des paramètres acoustiques du signal audio afin de condenser l'information utile à la reconnaissance. Le processus d'extraction est décrit dans le chapitre 3. Ces paramètres servent ensuite à deux tâches en parallèle : un alignement forcé et une reconnaissance de phonèmes. Cette thèse porte uniquement sur la brique de reconnaissance automatique de phonèmes, encadrée en rouge sur la figure 1.12.

La tâche d'alignement forcé consiste à localiser chaque mot et phonème prononcé sur l'enregistrement de parole grâce à un système de RAP. Le texte effectivement lu par l'enfant pouvant être différent du texte qu'il·elle devait lire (erreurs de lecture, répétitions, omission de mots...), le système utilise un modèle de langage modifié, incorporant les possibles altérations

du texte de référence. En s'appuyant sur l'enregistrement de lecture, le module d'alignement forcé peut donc déjà prédire la présence de potentielles erreurs de fluence. En sortie, nous obtenons les frontières temporelles de chaque mot et chaque phonème supposés lus par l'enfant.

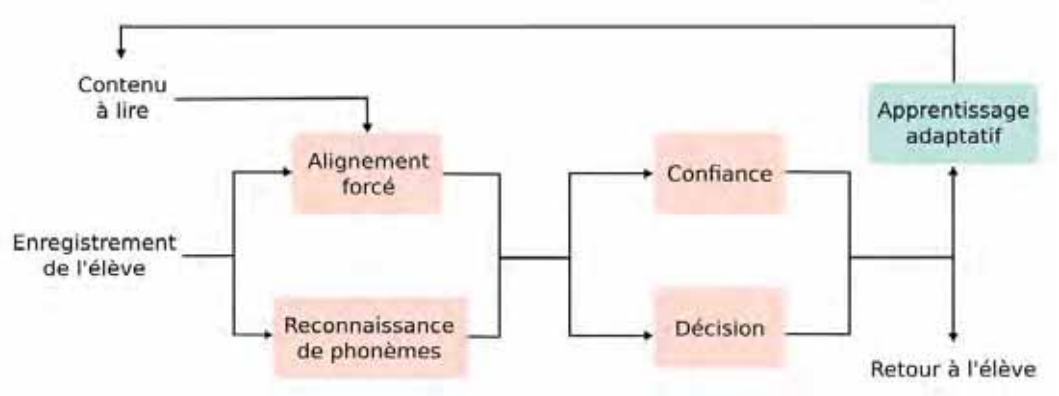


FIGURE 1.12 – Fonctionnement du système de détection d'erreurs de lecture derrière l'exercice de lecture orale

La modélisation acoustique pour la reconnaissance de phonèmes constitue le sujet central de cette thèse et sera bien sûr décrite de façon exhaustive dans la suite de ce manuscrit : nous n'en faisons qu'une description générale dans cette section. La tâche de reconnaissance de phonèmes ne s'appuie pas sur le texte à lire, mais uniquement sur les paramètres acoustiques du signal audio. Son objectif est de détecter le plus précisément possible les phonèmes réellement émis par l'enfant, ainsi que leur position dans le signal de parole. La reconnaissance de phonème est préférée à une reconnaissance classique de mots, moins précise et ne permettant pas de gérer des erreurs de lecture menant à des mots non existants ou incomplets.

À partir des sorties de ces deux modules de reconnaissance vocale est calculé un score de *Goodness of Pronunciation* (GOP) [Witt 2000, Kanters 2009], mesurant la probabilité que l'enfant ait prononcé le phonème correspondant au graphème à lire. Des paramètres dérivés sont également calculés, comme la moyenne et variance du GOP sur un mot. D'autres paramètres sont extraits du texte à lire (nombre de phonèmes, nombre de lettres, difficulté du mot...), de l'enregistrement brut (taux de saturation) ou des sorties du module d'alignement forcé (durée du mot, temps de parole de l'enfant, rapport signal à bruit...) et de reconnaissance de phonèmes (distance de Levenshtein [Levenshtein 1966] entre la séquence de phonèmes prédite et la séquence de phonèmes correspondant aux graphèmes à lire...).

Ces paramètres fournissent de l'information à deux algorithmes : un module de décision, qui calcule une probabilité d'exactitude p_e de la lecture de l'enfant, et un module de confiance, qui calcule une probabilité de confiance p_c du système de décision. Le premier est un algorithme de *machine learning* fondé sur des arbres de décision, nommé XGBoost [11]. Le second utilise le *Trust Score* [Jiang 2018], score qui mesure l'accord entre le module de décision et un second module fondé sur l'algorithme des plus proches voisins. À l'aide de seuils s_{correct} et $s_{\text{confiance}}$ fixés sur la probabilité d'exactitude et la probabilité de confiance, le système rend sa

11. https://xgboost.readthedocs.io/en/latest/

décision finale pour chaque mot du texte à lire. Les seuils sont choisis de façon à minimiser le taux de faux positifs, c'est-à-dire la détection d'une mauvaise lecture alors que l'enfant a correctement lu, qui causent énormément de frustration. Il existe trois cas possibles, définis dans le tableau 1.1. Dans le premier cas, le système est certain que la lecture est correcte : l'enfant reçoit des félicitations, et le système d'apprentissage adaptatif considère la question comme réussie. Dans le second cas, le système est certain que la lecture est incorrecte : l'enfant reçoit des encouragements à poursuivre ses efforts, et le système d'apprentissage adaptatif considère la question comme incorrecte. Enfin, si le système est incertain, l'apprentissage adaptatif ne tient pas compte de cette question, et il est demandé à l'élève de s'auto-évaluer. Cela permet, scientifiquement, d'avoir une indication sur le retour que le système aurait dû fournir, et pédagogiquement d'entraîner l'enfant à prendre conscience de son apprentissage. Une bonne précision du module de reconnaissance de phonèmes est donc primordiale pour fournir un retour adapté à l'enfant, dans lequel le système global a confiance : nous travaillons dans cette thèse à l'amélioration de cette précision.

Table 1.1 – Retour à l'enfant et à l'apprentissage adaptatif en fonction des probabilités d'exactitude et de confiance

Conditions	Retour à l'enfant	Retour au système d'apprentissage adaptatif
$p_e > s_{\text{correct}}$ et $p_c > s_{\text{confiance}}$	« Bravo »	Correct
$p_e < s_{\text{correct}}$ et $p_c > s_{\text{confiance}}$	« Tu y es presque ! »	Not correct
$p_c < s_{\text{confiance}}$	Auto-évaluation	NULL

L'exercice de lecture orale actuel vise à l'entraînement des élèves uniquement. Un outil d'évaluation du déchiffrage ou de la fluence pourrait être imaginé et créé afin de répondre aux besoins des enseignant·e·s. Ce cas d'usage de la RAP nécessiterait cependant un système extrêmement fiable pour fournir des évaluations précises des lectures d'élèves, sans quoi l'utilité pour les enseignant·e·s serait nulle.

1.4 Bilan

Nous avons présenté dans ce chapitre le contexte de la thèse, financée par l'entreprise Lalilo, qui fournit un assistant pédagogique pour les professeur·e·s de CP au CE2 leur permettant de différencier leur enseignement de la lecture. Chaque élève travaille à son rythme grâce à un système d'apprentissage adaptatif qui choisit les exercices maximisant sa progression en fonction de ses réponses aux exercices précédents. Une étude sur l'importance de la lecture orale pour l'apprentissage de la lecture s'est appuyée sur une recherche bibliographique et sur l'analyse d'un sondage pédagogique créé par nos soins auquel 79 enseignant·e·s utilisateur·rice·s de la plateforme Lalilo ont répondu. Nous nous sommes ensuite penché·e·s sur l'utilité de la reconnaissance vocale pour rendre plus fréquente la pratique de la lecture orale en classe. L'exercice de lecture orale de Lalilo a en effet pour objectif de permettre aux élèves de s'entraîner régulièrement, et un système de détection d'erreurs de lecture composé de modules

de reconnaissance automatique de la parole et de classification fournit un retour à l'élève sur la qualité de sa lecture.

Cette thèse porte sur un module de reconnaissance automatique de phonèmes, brique élémentaire d'un système global de détection d'erreurs de lecture dans la parole d'AL. Nous cherchons à améliorer la précision de ce module, objectif primordial : la séquence de phonèmes reconnus constitue la base de la majorité des paramètres fournis au module de détection d'erreurs. Ces améliorations sont essentielles, d'autant plus que nos enregistrements rendent la tâche ardue : la parole d'enfants est acoustiquement très variable, la lecture orale d'AL contient des erreurs de lecture et des disfluences, les enregistrements contiennent du bruit typique des salles de classe... Une séquence de phonèmes correctement reconnus permet d'obtenir de meilleurs paramètres pour la classification de la lecture en tant que correcte ou incorrecte, et d'avoir une plus grande confiance en cette classification. Le retour fait à l'élève est ainsi amélioré et favorise sa progression dans l'apprentissage de la lecture.

Chapitre 2

État de l'art : reconnaissance automatique de la parole d'enfants

Les systèmes de RAP ont gagné en popularité ces dernières années grâce à l'avènement des ordinateurs actuels, avec d'importantes capacités de calcul. Ils sont de nos jours utilisés pour des applications variées : systèmes d'interaction homme-machine, assistants vocaux pour la maison, systèmes technologiques pour l'éducation, applications de divertissement... Étant de plus en plus exposé·e·s aux technologies numériques dans la vie quotidienne, les enfants représentent une proportion importante du public utilisant ces systèmes de reconnaissance vocale. Cependant, si les performances des systèmes de RAP approchent la perfection sur la parole d'adultes, celles sur la parole d'enfants en sont très loin. Aborder les défis que représente la parole d'enfants pour la reconnaissance automatique devient donc essentiel pour atteindre des niveaux de précision similaires.

Dans ce chapitre, nous présentons un état de l'art pertinent pour notre sujet d'étude : la reconnaissance automatique de la parole d'enfants AL. Une première section est dédiée à un court historique des techniques de modélisation acoustique utilisées pour la RAP, dont les approches « génératives » puis « hybrides », et celles, plus récentes, dites « *end-to-end* ». Nous nous focalisons ensuite sur les spécificités de la parole d'enfants afin de comprendre les difficultés des systèmes à reconnaître ce type de parole. Enfin, une restitution des études existantes dans le domaine de la RAP appliquée sur la parole d'enfants est menée.

Sommaire

2.1 Reconnaissance automatique de la parole

La reconnaissance automatique de la parole est une tâche qui fascine depuis longtemps les scientifiques du monde entier. Elle consiste à déterminer, à partir d'un signal audio de parole, les unités de parole (phonèmes, mots, phrases) prononcées par le locuteur ou la locutrice. Les techniques ont beaucoup évolué du 18ème siècle à nos jours. Nous présentons tout d'abord le fonctionnement général d'un système de RAP, puis un état de l'art avec les systèmes encore Ceux-ci se divisent en trois approches, que nous détaillons dans les sections suivantes : génératives, hybrides et *end-to-end*.

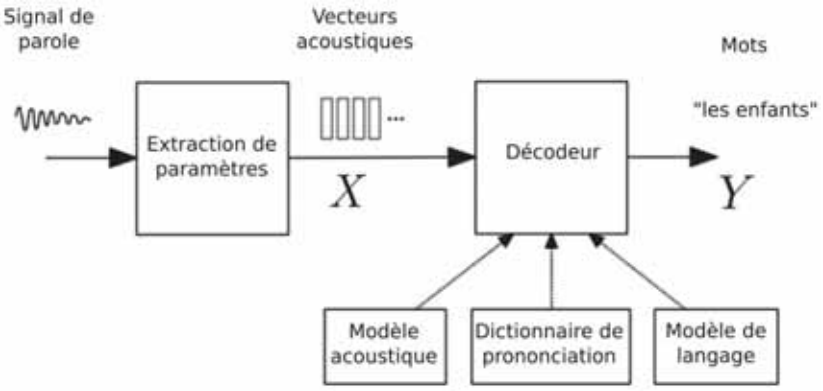


FIGURE 2.1 – Fonctionnement général d'un système de RAP (tiré de [Gales 2008])

Un système de RAP est illustré en figure 2.1. Des paramètres, que nous étudierons au chapitre 3, sont extraits du signal audio et mis sous la forme d'une séquence de vecteurs acoustiques de taille fixe $X_{1:T} = x_1,...,x_T$ avec T le nombre de trames audio. Le décodeur a pour mission de prédire la séquence d'unités de parole (par exemple, des mots) $Y_{1:L} = y_1,...,y_L$, avec L le nombre de mots dans la séquence, ayant le plus de probabilité d'avoir généré la séquence de vecteurs acoustiques X. En d'autres termes, le décodeur doit maximiser la probabilité $P(Y|X)$:

$$\hat{Y} = \arg\max_Y P(Y|X) \tag{2.1}$$

qui peut se décomposer, d'après la loi de Bayes, en :

$$\hat{Y} = \arg\max_Y \frac{P(X|Y)P(Y)}{P(X)} \tag{2.2}$$

La probabilité $P(X)$ peut être ignorée puisqu'elle est constante pour toutes les séquences de mots possibles, donnant ainsi :

$$\hat{Y} = \arg\max_Y P(X|Y)P(Y) \tag{2.3}$$

Le décodeur peut être composé de plusieurs modèles : le modèle acoustique, qui modélise la probabilité $P(X|Y)$ d'observer la séquence de vecteurs acoustiques X pour toute séquence d'unités de parole ; le dictionnaire de prononciation, qui permet la correspondance entre les phonèmes et les mots, et enfin le modèle de langage, qui modélise la probabilité $P(Y)$ d'obtenir une séquence de mots Y en respectant l'enchaînement des mots pour faire une phrase correcte en langue française.

Le modèle acoustique est l'élément le plus important du système, puisqu'il modélise la correspondance entre l'entrée audio et une séquence de phonèmes, à partir de laquelle il est ensuite possible (mais pas nécessaire) de reconstruire des mots et des phrases. L'unité du phonème est classiquement choisie, car c'est la plus petite unité de parole, et chaque mot est décomposable en une séquence de phonèmes. Cela permet au modèle d'apprendre une représentation correcte de chaque phone, qui est la représentation sonore d'un phonème, même avec une quantité de données limitée : apprendre à modéliser les mots directement demanderait une quantité de données bien plus importante pour disposer d'un nombre suffisant d'exemples de chaque mot.

2.1.1 Modèles acoustiques génératifs GMM-HMM

Les ancêtres des modèles actuels ont émergé dans les années 70, grâce à l'utilisation des chaînes de Markov cachées (*Hidden Markov Models*, HMM), formalisées dans [Baum 1967], pour la reconnaissance automatique de la parole [Baker 1975, Jelinek 1976]. Des travaux majeurs introduisant la théorie des HMM et présentant des exemples d'application aux problèmes de la RAP ont contribué à démocratiser leur utilisation [Rabiner 1983, Levinson 1983, Rabiner 1986]. En effet, puisque la parole présente une structure temporelle séquentielle, les HMM fournissent une architecture particulièrement adaptée pour sa modélisation. Les HMM avaient ainsi pour rôle de modéliser l'enchaînement temporel d'unités de la parole, dont la distribution statistique était modélisée par une simple distribution Gaussienne. Cette distribution se fonde sur l'hypothèse que les données à modéliser (ici, les paramètres acoustiques) sont symétriques et uni-modales, ce qui n'est absolument pas le cas, d'après [Gales 2008], pour des données de parole, dans lesquelles le genre, l'accent et les fréquences du locuteur ou de la locutrice forment plusieurs modes.

Au milieu des années 80, des techniques de modélisation statistique plus avancées, telles que les mélanges de lois Gaussiennes (*Gaussian Mixture Model*, GMM) [Juang 1985], sont introduites. Les GMM sont des distributions très flexibles capables de modéliser des distributions asymétriques et multi-modales, et donc mieux adaptées à la modélisation de la parole. Les GMM ont ainsi été combinés aux HMM pour obtenir les systèmes hybrides tels que nous les connaissons aujourd'hui : les GMM-HMM.

Dans un modèle acoustique GMM-HMM dit « monophone », c'est-à-dire traitant chaque phone indépendamment de ses voisins, chaque phone est représenté par un HMM à trois états modélisant le début, le milieu et la fin du phone. Une représentation de ce HMM est affichée en figure 2.2.

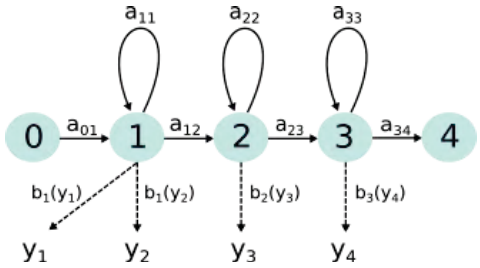


FIGURE 2.2 – Représentation d'un HMM à trois états pour un modèle GMM-HMM monophone

Les probabilités de transition a_{ij} entre deux états s_i et s_j sont apprises par le HMM durant l'entraînement. Lorsque nous entrons dans un état s_j, un vecteur acoustique y est émis grâce à la distribution $b_j()$ associée à cet état. Cette distribution est modélisée par le GMM, mélange de composantes gaussiennes qui sont sommées selon :

$$b_j(y) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(y; \mu^{(jm)}, \Sigma^{(jm)}) \tag{2.4}$$

où c_{jm} est la probabilité *a priori* de chaque composante m pour l'état s_j, et dont la somme sur $m = 1,...,M$ est unitaire. μ et Σ sont les moyennes et variances associées à la gaussienne m pour l'état s_j.

Les modèles GMM-HMM sont soumis à deux hypothèses découlant du HMM. La première considère qu'un état est conditionnellement indépendant de tous les autres états sachant l'état précédent : cela limite donc la dépendance d'un état uniquement à l'état précédent. La seconde concerne les émissions acoustiques : celles-ci sont conditionnellement indépendantes des autres émissions et ne dépendent que de l'état qui les a générées.

Différentes techniques d'amélioration des GMM-HMM ont émergé. Les modèles « triphone » instaurent la prise en compte du contexte et des effets de co-articulation en modélisant des trinômes de phones [Schwartz 1985]. Les triphones rares, ne pouvant être correctement modélisés, ont ensuite été regroupés grâce à des arbres de décision [Bahl 1991]. Des techniques d'adaptation ont ensuite été utilisées pour améliorer les performances de ces modèles, comme l'analyse discriminante linéaire (*Linear Discriminant Analysis*, LDA), technique de sélection de paramètres, la méthode *Speaker Adaptive Training* (SAT) [Anastasakos 1997], qui consiste à modéliser des caractéristiques du·de la locuteur·rice à partir des paramètres acoustiques, ou encore la méthode *Feature-space Maximum Likelihood Linear Regression* (fMLLR) [Gales 1998], visant à adapter les paramètres acoustiques à l'environnement et au·à la locuteur·rice.

Les modèles GMM-HMM restent cependant limités par les problèmes suivants :

— les fonctions de densité de probabilité sont décrites par des mélanges de Gaussiennes, dont la complexité et donc la qualité de modélisation doit être ajustée à la quantité de données disponibles ;

— l'algorithme d'entraînement cherche à maximiser la vraisemblance avec la distribution réelle des données plutôt que les probabilités postérieures, impliquant ainsi une pauvre

capacité de discrimination.

2.1.2 Modèles acoustiques hybrides DNN-HMM

L'idée de remplacer les GMM par des réseaux de neurones artificiels (*Artificial Neural Network*, ANN) est apparue à la fin des années 80. Les ANN qui ne comportaient qu'une seule couche neuronale étaient cependant incapables d'apprendre des représentations complexes comme celle de la parole, et les puissances de calcul disponibles à cette époque n'étaient pas suffisantes pour l'apprentissage automatique de réseaux de neurones avec un grand nombre de couches. Des réseaux de neurones peu profonds (*Deep Neural Networks*, DNN) sous la forme de simples perceptrons multi-couches (*Multi-Layer perceptron*, MLP) ont toutefois été utilisés avec succès pour des tâches primaires [Morgan 1990, Lippmann 1990, Bourlard 1990], par exemple de la reconnaissance de chiffres énoncés oralement. Ces réseaux MLP avaient cependant tendance à mal reconnaître les sons contenant une grande variabilité temporelle, comme les consonnes plosives, de par leur incapacité à gérer la nature séquentielle temporelle des signaux de parole. Pour pallier à ce problème, [Waibel 1989] a alors proposé les réseaux de neurones à délais temporels (*Time-Delay Neural Networks*, TDNN), qui sont encore utilisés aujourd'hui et dont le fonctionnement sera détaillé au chapitre 3.

Ce n'est qu'à partir des années 2000, après le développement des puissances de calcul de nos ordinateurs et l'apparition des GPU, que les DNN se sont démocratisés. Ils sont d'abord utilisés en conjonction avec les GMM : les systèmes « tandem » consistent à entraîner un classifieur DNN à prédire des probabilités postérieures de phonème indépendamment du contexte, et à se servir des représentations extraites comme paramètres d'entrée d'un système GMM-HMM [Hermansky 1999, Ellis 2001]. En 2006, [Hinton 2006] propose un des premiers modèles profonds efficace en remplacement des GMM : ce modèle génératif, nommé *Deep Belief Network* (DBN), est un empilement de sous-réseaux sous forme de machines de Boltzman restreintes [Hinton 1983]. Un modèle DBN a ensuite été utilisé pour initialiser les poids d'un DNN par [Mohamed 2011], dont le modèle DNN-HMM final surpasse un modèle GMM-HMM pour une tâche de RAP.

Les modèles DNN-HMM présentent cependant une tendance à s'adapter trop fortement aux données d'entraînement et à ne pas pouvoir généraliser les connaissances acquises sur des données inconnues et potentiellement légèrement différentes. La technique de *dropout* [Hinton 2012] a permis de résoudre ce problème de sur-apprentissage, et a rendu les modèles DNN-HMM plus populaires. La création de nouvelles fonctions d'activation, par exemple la fonction *Rectified Linear Unit* (ReLU) [Nair 2010] et ses dérivées, remplaçant les fonctions classiques linéaire et sigmoïde, a également apporté des solutions aux problèmes de disparition ou d'explosion de gradients des DNN. Les i-vecteurs, proposés initialement pour la reconnaissance de locuteur·rice [Dehak 2011], encodent de l'information sur l'identité vocale du locuteur ou de la locutrice, et ont amélioré les performances de modèles DNN-HMM [Saon 2013, Senior 2014]. De nouvelles architectures de DNN sont ensuite apparues. Les TDNN, introduits à la fin des années 80, ont refait surface et se sont installés dans les pratiques courantes grâce à leur utilisation dans l'outil Kaldi [Povey 2011]. Les RNN [Rumelhart 1986] se sont ensuite démo-

cratisés pour les signaux de parole [Graves 2013], avec pour objectif d'en saisir la dynamique temporelle, puis les réseaux de neurones convolutionnels (*Convolutional Neural Networks*, CNN) [Abdel-Hamid 2014], initialement créés pour le traitement d'image [Lecun 1995] et permettant de modéliser une invariance temporelle.

Les DNN ont l'avantage, par rapport aux GMM, de modéliser des représentations plus complexes et non fixes grâce à leur capacité à apprendre. Ce sont de plus des modèles discriminants, qui offrent une meilleure distinction entre classes en étant entraînés à modéliser directement les probabilités postérieures et à minimiser l'erreur de reconnaissance.

Un des inconvénients majeurs des modèles fondés sur des HMM reste tout de même la complexité de la procédure d'entraînement. Il est tout d'abord souvent nécessaire de générer des alignements des transcriptions sur les fichiers audio, lorsque ceux-ci ne sont pas pré-alignés manuellement. Une étape de post-traitement, appelée le décodage, est ensuite nécessaire pour générer la séquence de sortie à partir des différents modules : modèle acoustique, dictionnaire de prononciation et modèle de langage. De plus, ces modules étant entraînés indépendamment des autres, il peut y avoir des décalages de comportement entre les différents modules.

2.1.3 Approches *End-to-end*

Les approches *end-to-end* représentent une part importante de cette thèse, et serons présentées en plus grand détail dans le chapitre 5. Nous ne présentons en conséquence dans cette section qu'un bref historique de l'apparition de ces modèles dans l'état de l'art de la reconnaissance automatique de la parole.

En réaction à la complexité d'entraînement des modèles hybrides DNN-HMM, les approches *end-to-end* ont été créées dans un objectif de simplicité. Premièrement, elles se débarrassent de la nécessité de pré-aligner les données avec un modèle supplémentaire, et donc d'entraîner ce modèle. Ensuite, elles visent à unifier le processus d'entraînement grâce à un modèle unique, évitant ainsi les éventuels décalages entre modules. Enfin, également dans un objectif de simplicité, les systèmes *end-to-end* ont introduit la possibilité d'agir au niveau du caractère plutôt que du phonème : cela élimine l'étape préliminaire de conversion des transcriptions textuelles en phonèmes, ainsi que l'étape de post-traitement visant à retrouver les mots à partir des phonèmes prédits.

Un premier pas vers la RAP *end-to-end* a été fait par [Graves 2006] avec la fonction CTC, dans le cadre de la reconnaissance de phonèmes sur le corpus TIMIT [Garofolo 1993b]. Leur modèle est composé d'un simple encodeur à base de couches RNN. Il utilise la fonction CTC pour apprendre l'alignement entre les informations acoustiques et les transcriptions phonétiques, puis pour prédire les séquences de phonèmes. Il obtient un taux d'erreur de phonème (*Phoneme Error Rate*, PER) de 30,5% sur TIMIT. Le système RNN-CTC est perfectionné dans [Graves 2014b], introduisant l'utilisation des caractères plutôt que des phonèmes, ainsi que l'insertion d'un modèle de langage au moment de l'inférence.

Les premières architectures constituées d'un encodeur et d'un décodeur, appelées *sequence-*

to-sequence (*seq2seq*), ont été introduites dans le domaine de la traduction automatique neuronale [Sutskever 2014], puis ont rapidement été intégrées pour la reconnaissance automatique de la parole [Chorowski 2014, Lu 2015]. Dans un système *seq2seq*, l'encodeur a pour rôle d'extraire l'information acoustique des paramètres audio, et le décodeur, celui de prédire une séquence d'unités de parole (caractère ou phonème) à partir de l'information acoustique et des unités déjà prédites. Les deux modules sont généralement constitués de réseaux récurrents, particulièrement adaptés à la modélisation de la parole.

[Chorowski 2014] a ajouté le concept d'attention, précédemment présenté pour de la génération d'écriture manuscrite [Graves 2014a] puis pour de la traduction automatique neuronale [Bahdanau 2014], pour lier l'encodeur et le décodeur du modèle *seq2seq*. Ce mécanisme permet de chercher efficacement la portion d'information acoustique la plus pertinente pour la prédiction d'une unité de parole, étant donné celles déjà prédites. [Chan 2016] met également en avant un modèle *seq2seq* avec de l'attention, baptisé *Listen, Attend and Spell* (LAS), qui est un des premiers modèles *seq2seq* à utiliser des caractères plutôt que des phonèmes. Évalué sur la tâche *Google Voice Search*, il obtient un taux d'erreur mot (*Word Error Rate*, WER) de 10,3%. [Chiu 2018] reprend le modèle LAS et ajoute quelques améliorations, comme l'utilisation de sous-mots (*word pieces*) formés de plusieurs graphèmes au lieu de graphèmes individuels, et de mécanismes d'attention à plusieurs têtes. Cela lui permet de diminuer le WER à 5,6% sur *Google Voice Search*. Les différentes têtes de ce mécanisme portent sur des projections linéaires de la séquence à traiter, dont les sorties sont ensuite concaténées pour former un unique vecteur d'attention.

En comparaison avec un système encodeur-décodeur sans attention, l'utilisation de celle-ci empêche un sur-apprentissage des transcriptions d'entraînement. L'avantage de l'attention par rapport à la fonction CTC est qu'elle ne suppose aucune condition d'indépendance entre symboles ou entre trames audio. L'inconvénient, en revanche, est que l'attention est trop flexible pour la reconnaissance de la parole : le signal de parole étant intrinsèquement séquentiel, l'attention peut faire des erreurs en se dispersant de façon non monotone sur toutes les trames du signal. Pour limiter cet inconvénient, [Chorowski 2015] présente un système à base d'attention améliorée, la contraignant à être plus attentive à la localisation des trames audio, et l'empêchant de se concentrer sur une unique trame. Cette technique d'attention permet de réduire le PER sur TIMIT de 18,7% à 17,6%. [Watanabe 2017] propose par ailleurs une architecture *seq2seq* combinant une fonction CTC et un mécanisme d'attention afin de tirer parti des avantages de chacune de ces méthodes. Ces deux paradigmes sont assemblés pendant l'entraînement grâce à une combinaison linéaire des fonctions de coûts associées, et pendant l'inférence par une moyenne pondérée entre leurs scores de prédiction. Les différentes sorties (CTC, attention et combinée) obtiennent, sur le jeu de test eval92 du corpus Wall Street Journal (WSJ) [Garofolo 1993a], des taux d'erreur caractère (*Character Error Rate*, CER) respectifs de 9,0%, 8,2% et 7,4%.

Les mécanismes d'attention se développent rapidement, et en viennent à remplacer les couches de RNN des encodeurs et décodeurs de systèmes *seq2seq*. [Vaswani 2017] présente pour la traduction automatique neuronale un système entièrement fondé sur de l'attention, nommé Transformer, et introduit ainsi les mécanismes d'auto-attention. Ces mécanismes ont

le même fonctionnement que les mécanismes d'attention classiques, mais au lieu de chercher la correspondance entre informations acoustiques de l'encodeur et informations textuelles du décodeur, ils relient différentes positions d'une même séquence (acoustique ou textuelle) pour en extraire une représentation. Le modèle Transformer utilise des mécanismes d'attention multi-têtes pour améliorer la qualité et la diversité des représentations extraites. Le Transformer est bientôt adapté pour la RAP par [Dong 2018], obtenant un WER de 10,9% sur WSJ eval92 sans utiliser de modèle de langage. Les auteur·rice·s proposent en outre un système d'attention 2D à insérer avant l'encodeur pour extraire des représentations à partir des dépendances temporelles et spectrales d'un spectrogramme. Les modèles Transformer ont ensuite adopté l'architecture combinée CTC/attention [Karita 2019a], qui montre une meilleure précision (WER de 10,2% sur WSJ eval92) et une convergence plus rapide.

De par la grande diversité des modèles à l'état de l'art, et le développement extrêmement rapide de nouvelles architectures, il peut être difficile de sélectionner l'approche la plus appropriée à une application. [Karita 2019b] fournit une étude comparative des approches hybrides (DNN-HMM), *end-to-end* type LAS et *end-to-end* type Transformer, sur une grande sélection de tâches de RAP. Leurs résultats indiquent que l'architecture Transformer surpasse rigoureusement l'architecture LAS sur tous les jeux de données testés. Par exemple, le Transformer obtient un WER de 2,6% et 5,7% sur les jeux test_clean et test_other du corpus Librispeech [Panayotov 2015], contre 3,3% et 10,8% pour le LAS. L'approche hybride reste cependant significativement plus performante sur des données bruitées : elle obtient un WER de 11,4% sur le jeu de test et05_real du corpus CHIME4 [Barker 2017] (contre 14,5% pour le Transformer) et de 81,3% contre 87,1% sur le jeu de test kinect du corpus CHIME5 [Barker 2018]. Elle résiste également face au Transformer sur des jeux de données bien connus tels que WSJ (WER de 2,3% sur eval92, contre 4,4%).

2.2 Particularités de la parole d'enfants

De nombreux projets de recherche portent sur la caractérisation des différences entre la parole d'adultes et la parole d'enfants, tant sur le côté acoustique que phonologique. Des articles de revue fournissent des analyses sur la parole d'enfants : [Kent 1976] réunit les résultats d'un certain nombre d'études acoustiques et les lie au développement anatomique et neuro-musculaire des mécanismes de parole lors de la croissance de l'enfant. [Mugitani 2012] se concentre sur le développement du conduit vocal (depuis les plis vocaux jusqu'aux lèvres) et l'influence de son évolution sur les fréquences de la parole d'enfants. [Potamianos 2007] étudie les caractéristiques acoustiques et linguistiques de la parole d'enfants lorsqu'elle est lue ou spontanée.

2.2.1 Hauteur de la fréquence fondamentale et des formants

Très tôt ont été observées des différences dans la hauteur de la fréquence fondamentale et des formants, ainsi que dans leurs variabilités inter- et intra-locuteur·rice·s [Eguchi 1969, Kent 1976,

Lee 1999]. Étude clé dans le domaine, [Eguchi 1969] se concentre sur trois caractéristiques de la parole : (1) la fréquence fondamentale f_0, (2) les fréquences des premier et deuxième formants f_1 et f_2 des sons vocaliques, et (3) le délai d'établissement du voisement (paramètre temporel défini comme l'intervalle de temps entre le relâchement d'une plosive et la reprise du voisement). Son ouvrage montre tout d'abord que les fréquences f_0, f_1 et f_2 des enfants entre 3 et 13 ans évoluent à la baisse avec l'âge.

Il a fallu attendre que les technologies radiographiques soient remplacées par d'autres comme le scanner ou l'imagerie par résonance magnétique (IRM), moins nocives pour le développement de l'enfant, pour pouvoir étudier l'anatomie de l'appareil de production de la parole chez l'enfant, et faire le lien entre son évolution pendant la croissance de l'enfant et l'abaissement des fréquences. Dans [Mugitani 2012], les auteur·rice·s précisent que la forme et la longueur de ce canal définissent les caractéristiques acoustiques des expressions vocaliques d'une personne, notamment les fréquences f_0, f_1 et f_2. La fréquence fondamentale f_0 dépend de la longueur et du volume des plis vocaux [Hirano 1981], et les fréquences des premier et second formants (f_1 et f_2) diminuent avec l'agrandissement du conduit vocal [Kent 1992]. Les enfants présentent ainsi des fréquences plus élevées de par la plus petite taille de leur conduit vocal et de leurs plis vocaux, et ces caractéristiques évoluent jusqu'à atteindre un niveau adulte vers l'âge de 14-15 ans [Mugitani 2012, Lee 1997]. Les zones de production de chaque voyelle dans l'espace f_1/f_2 sont également décalées par rapport à celles des adultes. Ce décalage est cependant différent pour chaque voyelle à cause du développement non linéaire de la cavité vocale, dont la forme influence la formation des voyelles [Lee 1997]. Cela implique un espace vocalique et une gamme de fréquences plus larges que pour les adultes.

2.2.2 Mécanismes d'articulation non stables

[Eguchi 1969] étudie de plus les déviations standards des fréquences f_0, f_1 et f_2, qui révèlent une diminution uniforme des variabilités intra-locuteur·rice entre 3 et 11 ans. Concernant le délai d'établissement du voisement, une forte variabilité intra-locuteur·rice est également observée chez les jeunes enfants, et décroît graduellement entre 3 et 8 ans. D'après [Kent 1976], les trois caractéristiques étudiées par [Eguchi 1969] sont chacune liées à des mécanismes de production de la parole : (1) l'ajustement du larynx, (2) la position des articulateurs pour la production de voyelles et (3) la coordination articulatoire-laryngienne. L'ensemble des tendances de variabilité intra-locuteur·rice observées corrobore ainsi une amélioration du contrôle moteur des mécanismes de parole entre l'âge de 3 et 11 ans.

[Lee 1999] confirme ces observations, et présente additionnellement des études sur l'évolution des durées de segments de parole (voyelles, consonne fricative [s], phrase), qui diminuent globalement avec l'âge. Dans la totalité de leurs études, les variabilités inter- et intra-locuteur·rice chutent drastiquement entre 8 à 12 ans (selon le paramètre étudié), révélant de fortes variabilités pour les jeunes enfants de 5 à 8 ans. La diminution de la durée segmentale pouvant être reliée à une amélioration de la vitesse d'articulation, les auteur·rice·s suggèrent que le contrôle des mécanismes d'articulation ne se rapproche de la précision d'un adulte qu'à partir de 8-9 ans. Ces conclusions rejoignent celles de l'étude de [Kent 1980] portant sur les

durées des segments de parole d'enfants de 4, 6 et 12 ans récitant des phrases, ainsi que celles de [Lee 1997], qui estime que le niveau adulte est atteint un peu plus tard, vers 11 ans. Sur la base d'études similaires, [Smith 1978] nuance néanmoins les différences entre enfant et adulte, arguant que le système de production de la parole d'un jeune enfant est plus sophistiqué qu'on ne le croit.

2.2.3 Faible capacité de co-articulation

Dans [Lee 1997], une analyse de la variabilité spectrale entre la première moitié et la seconde moitié d'une voyelle révèle d'importantes valeurs pour les enfants de moins de 10 ans, qui sont probablement dûes à des mouvements abrupts de la langue lors de la transition entre la voyelle et la consonne suivante : cela suggère une capacité de co-articulation encore faible chez les jeunes enfants. [Sereno 1985] proposait déjà l'idée que les enfants de 3 à 7 ans présentaient une co-articulation moins précise et plus variable que celle des adultes, et n'étaient pas encore capables d'anticiper la co-articulation entre consonnes et voyelles. Plus récemment, [Gerosa 2006b] étudie les transitions consonne-voyelle et la différence spectrale entre la consonne et la voyelle d'une paire. Ses résultats indiquent que l'effet de la co-articulation est moins marqué chez les jeunes enfants, et que le risque de confusion entre différents phonèmes émis par des enfants est plus grand, de par la faible différence spectrale entre leurs productions de différents phones.

2.2.4 Qualité linguistique et prosodique dégradée

Le lent développement des mécanismes d'articulation peut causer des variantes linguistiques et de prononciation chez les jeunes enfants. D'après [Fringi 2015], qui cite [Lust 2006] et [Cohen 2011], un positionnement non précis des lèvres et de la langue lorsque l'enfant apprend à parler peut entraîner des omissions ou substitutions de phonèmes, qui ont tendance à disparaître avec l'âge, mais peuvent persister jusqu'à l'âge de 6 ans.

Sur de la parole lue, [Lee 1999] relève une décroissance de la durée pour lire une phrase entre 7 et 14 ans, avec une diminution totale de la durée de 45%. Cette mesure est liée au taux de parole (ratio entre le nombre de mots émis par l'enfant et le temps de parole), mais dépend également de la capacité de lecture de l'enfant. Sur de la parole spontanée dans le cadre d'une interaction avec un ordinateur, [Potamianos 1998] trouve également que le taux de parole d'enfants de 11 à 14 ans est 10% plus élevé que celui d'enfants plus jeunes.

[Lee 1999] identifie de plus une présence importante de disfluences (suppression de phonèmes, mauvaises prononciations) chez les enfants de 5-6 ans, mais qui ne sont pas nécessairement causées par de faibles capacités de lecture, puisque les enfants non lecteurs devaient répéter les phrases plutôt que les lire. Sur de la parole spontanée d'enfants légèrement plus âgés (8-10 ans), [Potamianos 1998] observe une fréquence d'occurrence de fautes de prononciation deux fois plus élevée que chez les 11-14 ans, ainsi que 60% plus de bruits de respiration.

Dans le cadre de sa thèse, et de la collecte du corpus de lecture orale LetsRead auprès de jeunes enfants AL portugais, [Proença 2018] analyse les différents types d'erreurs de lecture et disfluences, et leur fréquence d'occurrence en fonction de l'âge des enfants. Les données qu'il récolte auprès d'enfants de 6 à 9 ans (du CP au CM1) sont constituées de phrases et de pseudo-mots, qui sont des mots non-existants créés spécialement pour évaluer la conscience phonémique des enfants en éliminant la possibilité que l'enfant connaisse le mot par cœur. Le pourcentage total d'erreurs est de 18,3% sur les phrases, et de 74,1% sur les pseudo-mots, qui représentent une tâche de lecture bien plus difficile. Le nombre d'erreurs de fluence et de déchiffrage diminue évidemment avec l'âge : sur les phrases, les élèves de CP du jeu de données ont lu 33,4% des mots avec une erreur, ceux de CE1 19,4%, et ceux de CE2 et CM1 16,4% et 12,3%. Les erreurs de déchiffrage les plus fréquentes des élèves de CP et CE1, qui nous intéressent dans cette thèse, sont les faux départs (par exemple, pour le mot « pouce » [pus], l'enfant lit [pu...pus]), les substitutions de mots (« pouce » substitué par « pause » [poz]), les substitutions de phonème ([pas]) et les répétitions de mots ([pus...pus]). Les élèves de CP et CE1 ont de plus tendance à lire plus lentement, à faire plus de pauses intra-mot et d'extensions de phones, tandis que les enfants un peu plus âgés (CE2-CM1) tendent à insérer ou supprimer des phonèmes ou mots à cause d'une lecture trop rapide.

Il convient enfin de signaler que certains types d'erreurs faites par les enfants AL peuvent s'apparenter à des évènements présents dans d'autres types de parole atypique :

— parole pathologique : par exemple, le bégaiement se révèle par des répétitions de mots (entiers ou partiels) et des allongements de phones [Ellis 2009], et certains types d'aphasie, un trouble neurologique du langage, impliquent des substitutions de mots ou des confusions de phonèmes [Alexander 2008] ;

— parole d'apprenant·e de seconde langue (L2) : parmi les « déviances » d'apprenant·e·s L2 identifiées par [Detey 2016], les déviances segmentales (insertions, suppressions, substitutions de phonèmes) et supra-segmentales (liées au rythme et à l'intonation) ressemblent notamment aux erreurs de déchiffrage et de fluence trouvées dans la parole d'enfants AL. Les évènements liés à l'accent et à la distorsion de certains phones ne se retrouvent cependant pas dans notre type de parole.

2.3 Reconnaissance automatique de la parole d'enfants

Avec l'essor des systèmes fondés sur la reconnaissance vocale et leur intégration dans notre vie quotidienne, les enfants deviennent de plus en plus usagers de ces technologies. Le domaine de la RAP pour voix d'enfants s'est donc développé graduellement depuis une trentaine d'années, et gagne en visibilité depuis quelques années seulement.

2.3.1 Jeux de données de parole d'enfants

Le domaine de recherche étant relativement nouveau, les données de parole d'enfants sont encore rares. En langue française, aucun jeu de données n'est actuellement disponible

publiquement : dans cette thèse, nous utiliserons un petit corpus de données récoltées par nos soins, comme nous le décrirons dans le chapitre 4.

[Claus 2013] établit une liste exhaustive des jeux de données existants dans différentes langues et contenant la parole d'enfants d'âges divers. Nous présentons ici les quelques jeux de données de parole d'enfants en langue anglaise sur lesquels s'appuient la plupart des systèmes à l'état de l'art en reconnaissance automatique de la parole d'enfants, que nous décrirons dans la section suivante. Le tableau 2.1 présente des informations générales sur ces jeux de données. Nous voyons que les jeux de données sont extrêmement diversifiés en termes de type de parole, du nombre et de l'âge des locuteur·rice·s, et de la quantité de données. De plus, certains corpus ne sont annotés qu'en partie et ne sont donc pas exploités entièrement actuellement.

Table 2.1 – Récapitulatif des jeux de données de parole d'enfants en langue anglaise

Corpus	Type de parole	# Heures	# Locuteur·rice·s	Âge
CMU Kids [Eskenazi 1996]	lue	~9	76	6-11
CID [Lee 1999]	lue	~2	324	6-14
OGI [Shobaki 2000]	lue	~23	509	6-11
CU Prompted & Read [Cole 2006a]	lue	~26	663	6-11
CSLU [Shobaki 2007]	lue et spontanée	~165	1100	5-16
ChIMP [Potamianos 1998]	spontanée	~10	97	6-14
CU Read & Summarized [Cole 2006b]	spontanée	~33	320	6-11
MyST [Cole 2019]	spontanée	~499	737	8-11
TBall [Kazemzadeh 2005]	dénomination d'images	~40	256	5-8

2.3.2 Systèmes existants

Le développement de la RAP pour voix d'enfants s'est d'abord fait à travers des applications à visée éducative : les « tuteurs » de lecture. L'apprentissage de la lecture est un défi majeur pour le développement intellectuel des jeunes enfants, qui peuvent avoir besoin de toute l'aide possible pour le maîtriser. De nombreux projets sont donc nés dans cet objectif, s'aidant de techniques de reconnaissance vocale pour évaluer des tâches de lecture orale :

— Projet LISTEN (Literacy Innovation that Speech Technology ENables) [Mostow 2001] : projet porté par l'université Carnegie Mellon, aux Etats-Unis. Un tuteur de lecture affiche un texte à lire, écoute l'enfant lire oralement, et propose de l'aide orale et visuelle. Le système de RAP est implémenté avec l'outil Sphinx-II [Ravishankar 1996] et suit une approche hybride avec un modèle acoustique GMM-HMM entraîné sur le corpus CMU Kids [Eskenazi 1996] ;

— Projet CLT (Colorado Literacy Tutor) [Hagen 2003] : projet porté par l'université du Colorado, qui utilise le système de RAP Sonic [Pellom 2001] fondé sur des GMM-HMM et entraîné sur les deux corpus CU [Cole 2006a] et [Cole 2006b] pour reconnaître les résumés d'histoires lues par les enfants.

— Projet TBALL (Technology Based Assessment of Language and Literacy) [Alwan 2007] : projet porté par un consortium d'universités de Californie, qui propose notamment d'évaluer la prononciation des élèves sur des mots isolés, grâce à un réseau Bayésien de classification [Tepperman 2007], ainsi que la compréhension écrite par de simples questions oui/non, grâce à un module de *word spotting* ;

— Projet SPACE (SPeech Algorithms for Clinical and Educational applications) [Duchateau 2009] : projet porté par plusieurs universités belges, proposant une évaluation automatique du niveau de lecture d'un enfant et la génération de retours adaptés au niveau du phonème, de la syllabe, et du mot. Les auteur·rice·s utilisent un modèle acoustique GMM-HMM entraîné sur un corpus de 22 heures de parole d'enfants de 5 à 11 ans lisant en néerlandais ;

— Projet FLORA (FLuent Oral Reading Assessment) [Bolaños 2011] : projet porté par l'entreprise Boulder Learning Technologies et l'université du Colorado, qui se concentre sur l'évaluation automatique de la fluence orale, et propose une application de type karaoké pour suivre la lecture de l'enfant et le débloquer si besoin. Le projet utilise un système de RAP hybride, avec un modèle acoustique GMM-HMM, optimisé pour suivre la lecture de l'enfant en temps réel ;

— Projet LetsRead [Proença 2018] : projet porté par des universités portuguaises et Microsoft, visant à la détection automatique d'erreurs de lecture et de fluence pour les enfants portugais de 6 à 10 ans. Le système utilisé est un réseau de neurones prenant en entrée de long contextes temporels [Schwartz 2009], créé par une équipe de l'université de Brno et entraîné avec 9 heures de portugais ;

— Projet FLUENCE [Godde 2017, Godde 2019] : projet porté par le CNRS et le laboratoire Gipsa-Lab, à Grenoble, qui présente deux outils de type karaoké pour l'entraînement de la fluence. Le premier vise à apprendre aux jeunes enfants à respirer au bon endroit pendant leur lecture, le second à les aider à améliorer leur fluence en les faisant lire en même temps qu'une voix adulte.

Les particularités de la parole d'enfants, décrites en section précédente, ont un fort impact sur les performances des systèmes de reconnaissance automatique de la parole. D'après [Sereno 1985], la parole d'enfants est déjà difficile à comprendre pour des humains : des tests perceptifs ont montré que les voyelles prononcées par des enfants étaient moins bien identifiées que celles prononcées par des adultes. Les auteur·rice·s expliquent ce phénomène par la faible capacité de co-articulation des jeunes enfants, explication soutenue par [Gerosa 2006b], et suggèrent que les mécanismes de co-articulation transmettent beaucoup d'information acoustique utile à l'intelligibilité de la parole. Reconnaître de la parole contenant peu d'information co-articulatoire est ainsi probablement une tâche très difficile pour un système de RAP. [Potamianos 2003] fournit une interprétation complète des conséquences de chaque particularité de la parole d'enfants sur les performances des systèmes de RAP. Tout d'abord, l'évolution de la fréquence fondamentale et des formants avec l'âge implique que les données d'apprentissage et de test doivent contenir des locuteur·rice·s du même âge, au risque d'entraîner de fortes dégradations de performance. Les fortes variabilités temporelles et spectrales présentes dans la parole d'enfants, combinées à une gamme étendue de fréquences, causent un chevauchement des classes phonémiques, ce qui en complique la classification. Additionnellement, il est plus difficile de séparer l'information dépendant du·de la locuteur·rice (la

fréquence fondamentale, par exemple) de l'information dépendant du phonème (les fréquences des formants) lors de l'extraction de paramètres, ce qui implique une moins bonne qualité de l'information acoustique extraite.

De la même façon que pour la reconnaissance de parole d'adultes, dont l'historique a été présenté au début de ce chapitre, la reconnaissance de parole d'enfants a commencé avec des approches génératives et des modèles acoustiques GMM-HMM. Potamianos et son équipe instaurent les premières études sur la RAP pour voix d'enfants et constatent une grande différence de performance par rapport à des systèmes pour adultes [Potamianos 1997, Potamianos 1998, Potamianos 2003]. Dans leur première étude, les auteur·rice·s observent une dégradation drastique des performances de leur modèle GMM-HMM entraîné sur parole d'adultes lorsque testé sur parole d'enfants : selon l'âge, la performance peut être jusqu'à quatre fois pire que sur de la parole d'adultes. Ils observent également que les performances s'améliorent avec l'âge (observation appuyée par [Shivakumar 2014, Fringi 2015, Yeung 2018]) et atteignent un niveau adulte à partir de 13-14 ans, ce qui est en accord avec le processus de développement de l'enfant décrit à la section précédente. Enfin, ils montrent qu'un GMM-HMM triphone, incorporant des informations de co-articulation entre phones, n'a pas apporté d'amélioration significative par rapport à un GMM-HMM monophone, ce qui confirme que les jeunes enfants (5-12 ans) n'ont encore que de faibles capacités de co-articulation. Leur deuxième étude utilise un GMM-HMM entraîné sur de la parole d'enfants, qui obtient un PER de 46,2%, démontrant 25% de différence relative de PER avec un score de référence pour la parole d'adultes sur le corpus TIMIT [Garofolo 1993b]. Enfin, leur troisième étude propose plusieurs méthodes pour pallier aux difficultés causées par la parole d'enfants : une méthode de normalisation par déformation de fréquences, s'apparentant à la technique de *Vocal Tract Length Normalisation* (VTLN), permet de compenser la variabilité dans les longueurs de conduits vocaux des locuteur·rice·s adultes et enfants, et une méthode de *spectral shaping* aide à réduire la variabilité spectrale inhérente à la parole d'enfants. Les techniques d'adaptation VTLN, SAT, fMLLR rencontrent un succès mitigé sur la parole d'enfants : significativement efficaces dans [Shivakumar 2014] avec un GMM-HMM sur de la parole lue (corpus CID), elles n'ont néanmoins qu'un effet limité dans [Gerosa 2006a], toujours avec un GMM-HMM mais sur de la parole spontanée (corpus ChIMP) contenant des disfluences et des hésitations. Des techniques plus avancées pourraient être nécessaires pour traiter de la parole de faible qualité prosodique.

Sur de la lecture orale d'enfants AL portugais, [Proença 2018] observe une influence négative significative de la présence d'erreurs de lecture sur la précision de la reconnaissance de phonèmes. Avec un modèle phonétique utilisant un modèle de langage classique bi-gram, le PER est de 22,8% sur les mots correctement lus et de 36,0% sur les mots contenants une erreur. L'utilisation d'une lattice au vocabulaire phonétique limité aux phonèmes attendus améliore grandement le PER sur les mots corrects (2,6%) au prix d'une dégradation significative de celui sur les mots erronés (41,5%). Des résultats similaires sont observés dans [Gerosa 2006a], sur de la parole spontanée cette fois-ci : en divisant leur ensemble de test en deux sous-ensembles (l'un contenant des disfluences et hésitations, l'autre non), les auteur·rice·s observent une forte dégradation (60% relatifs, 30,1% contre 48,0%) du WER en présence de ces évènements. Les auteur·rice·s constatent de plus que le WER par locuteur·rice est positivement corrélé (+0,65)

avec le nombre d'hésitations et de pauses remplies (uh, um...). Ces résultats contredisent l'étude précédente [Potamianos 2003], où la présence de disfluences ne causait pas de dégradation significative de la performance. Dans notre application, de par le très jeune âge de nos locuteur·rice·s, les erreurs de fluence et de prononciation liées au développement de l'enfant, identifiées par [Lee 1999], pourraient s'avérer contraignantes car difficilement différenciables des erreurs de lecture d'enfants AL.

De par la quantité limitée de données de parole d'enfants dans la plupart des langues, les modèles fondés sur des réseaux de neurones profonds n'ont commencé à être exploités que récemment. Parmi les projets de tuteurs de lecture présentés au début de cette section, seul le projet LetsRead, de 2014 à 2018, utilise des réseaux de neurones pour modéliser la parole d'enfant. Pour une application d'évaluation de l'expression orale en langue anglaise chez des enfants non-natifs de 7-8 ans, [Metallinou 2014] montre qu'un modèle DNN-HMM surpasse un GMM-HMM, même lorsque ce dernier est entraîné sur 8 fois plus de données. Leur meilleur modèle atteint un WER de 19,3% avec 430 heures de parole d'enfants. Pour de la reconnaissance de voyelles d'enfants Malaisiens, [Ting 2004, Yong 2011] observent de meilleures performances avec un TDNN-HMM qu'avec un modèle utilisant un simple MLP. Une étude plus récente introduit l'utilisation d'une version factorisée du TDNN, le TDNNF (*Factorized Time-Delay Neural Network*) pour la parole d'enfants [Wu 2019]. Les expériences de cet article sur les corpus CMU [Eskenazi 1997] et CSLU [Shobaki 2007] (90 heures au total) établissent une hiérarchie claire entre les différentes architectures acoustiques : le TDNNF-HMM, avec un WER de 11,7%, surpasse largement le GMM-HMM (35,6%) et le TDNN-HMM (15,8%). Ce dernier résultat suggère que le TDNNF-HMM est très efficace avec une quantité de données relativement limitée, et est capable de modéliser une parole complexe avec moins de paramètres que le TDNN-HMM.

[Serizel 2014a] utilise deux jeux de données (l'un de parole d'adultes, l'autre d'enfant) en italien, et évalue différentes stratégies d'entraînement d'un modèle hybride DNN-HMM. Ses résultats montrent qu'il est plus intéressant d'entraîner un modèle avec uniquement 7 heures de parole d'enfants qu'avec les deux corpus à la fois (13h20 au total), ce qui confirme que les caractéristiques de parole d'enfants sont en effet très différentes de celles de parole d'adultes. Ré-entraîner le modèle adulte+enfant avec les données d'enfant pour l'adapter de façon spécifique améliore toutefois légèrement les performances. Dans une seconde étude [Serizel 2014b], les auteur·rice·s montrent qu'adapter les paramètres acoustiques avec de la VTLN permet d'obtenir des performances similaires à celle du DNN-HMM adapté avec la parole d'enfants. Dans [Wu 2019], les auteur·rice·s trouvent en revanche que, si l'application de la VTLN améliore légèrement les performances d'un modèle GMM-HMM, elle n'a pas d'effet sur un TDNNF-HMM. Ce résultat, opposé à celui de [Serizel 2014b], peut s'expliquer par une quantité de données d'enfant beaucoup plus importante, permettant au modèle de saisir suffisamment de la variabilité présente dans la parole d'enfants, et donc de ne pas avoir besoin d'adaptation. Une étude de [Yeung 2018] permet d'affiner les résultats en fonction de l'âge : ils trouvent que la technique VTLN est significativement bénéfique uniquement à partir de 11 ans, et dégrade au contraire les scores sur de la parole d'enfants de moins de 8 ans.

Suite aux découvertes de [Serizel 2014a] sur l'adaptation avec de la parole d'enfants de

modèles pré-entraînés, et à l'expansion de la méthode d'apprentissage par transfert (*Transfer Learning*, TL) dans d'autres applications de RAP [Tong 2017b, Abad 2020, Duan 2020], plusieurs études utilisent cette technique pour l'adaptation adulte-enfant [Qian 2016, Tong 2017a, Matassoni 2018, Shivakumar 2020]. [Shivakumar 2020] fournit notamment une analyse complète des stratégies possibles et en infère des recommandations sur lesquelles nous nous sommes largement appuyé·e·s dans cette thèse (voir section 3.2.1 du chapitre suivant). Son étude utilise une combinaison de plusieurs corpus (OGI, ChIMP, CID et les deux corpus CU) pour un total de 91 heures parole lue et spontanée d'enfants, d'âges compris entre 6 et 14 ans. Leur meilleur modèle, un DNN-HMM construit à partir d'un modèle adulte adapté par TL à la parole d'enfants, obtient un WER de 17,8%. Leurs résultats montrent que l'utilisation du TL est beaucoup plus efficace qu'entraîner un modèle sur les 91 heures de parole d'enfants et appliquer les techniques d'adaptation classiques (SAT, LDA, fMLLR, VTLN, i-vecteurs).

Grâce à une décomposition de leurs données par âge, les auteur·rice·s de [Shivakumar 2020] peuvent inférer les besoins des systèmes de RAP en fonction de l'âge. La figure 2.3 affiche l'évolution du WER en fonction de l'âge des enfants et de la quantité de données d'entraînement : il apparaît très clairement que plus les enfants sont jeunes, plus il faut de données et plus le WER pâtit du manque de données. Cette observation s'explique par la très forte variabilité dans la parole des jeunes enfants : il faut ainsi de grandes quantités de données pour qu'un modèle acoustique soit capable de modéliser cette variabilité.

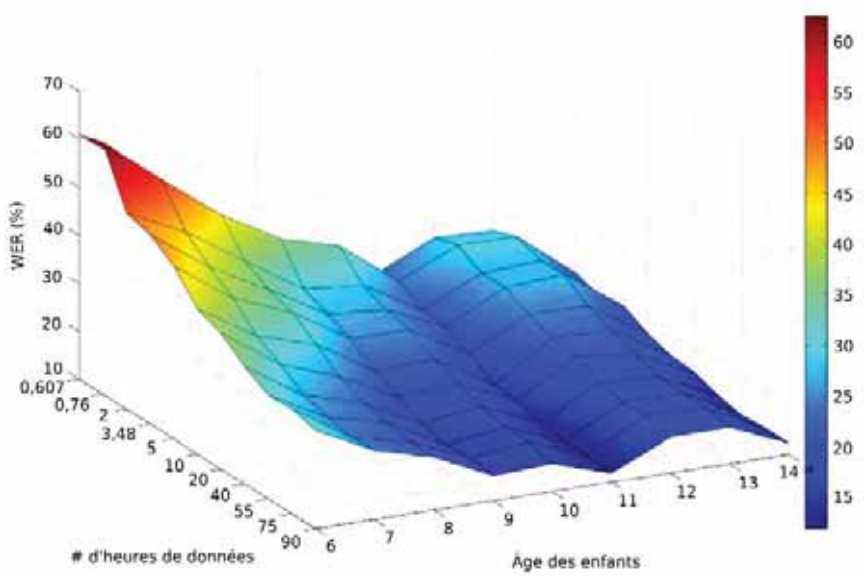


FIGURE 2.3 – WER obtenu sur de la parole d'enfants en fonction de l'âge des enfants et de la quantité de données d'entraînement, tiré de [Shivakumar 2020]

La figure 2.4 montre quant à elle l'évolution du WER en fonction de l'âge des enfants pour des modèles ayant subi une transformation estimée sur des données d'un âge spécifique. Nous observons que pour des enfants de moins de 10 ans, il est indispensable d'utiliser un modèle adapté à leur âge, et que plus les enfants sont jeunes, plus le fait d'utiliser un modèle dépareillé dégrade le WER. Cela confirme que les jeunes enfants ont des caractéristiques acoustiques

particulières et différentes de celles des adultes, et rejoint les conclusions de [Elenius 2005].

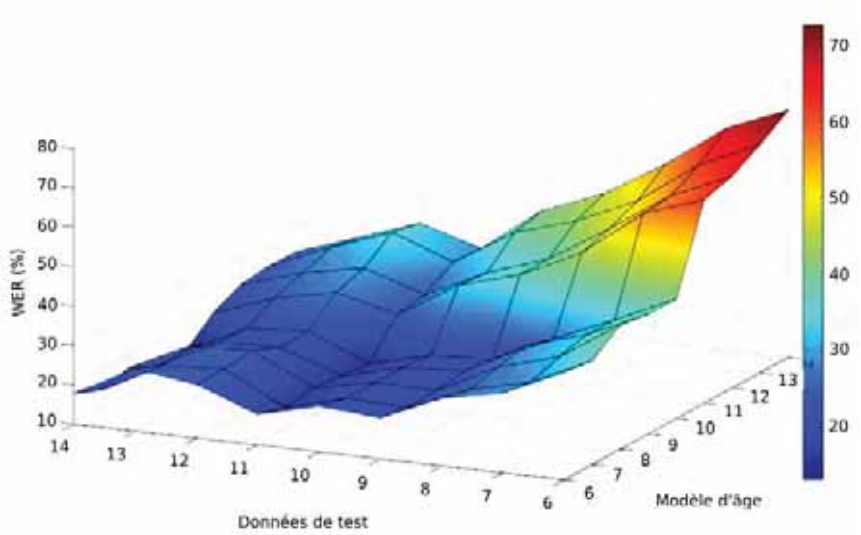


FIGURE 2.4 – WER obtenu sur de la parole d'enfants en fonction de l'âge des enfants des données d'entraînement et de test, tiré de [Shivakumar 2020]

L'étude de [Yeung 2018] observe de la même façon des dégradations flagrantes dans les scores WER sur la parole de très jeunes enfants : 26,9%, 14,6%, 10,5% et 5,1% sont obtenus pour des enfants de 5, 6, 7, et plus de 8 ans, respectivement. Ils obtiennent de plus des résultats prometteurs en entraînant des modèles spécifiques à chaque âge, tant que la quantité de données est suffisamment importante pour pallier à la grande variabilité de la parole d'enfants, ce qui confirme les observations faites sur les figures 2.3 et 2.4.

Les architectures *end-to-end* n'ont que très récemment été appliquées à la parole d'enfants, et cette thèse se distingue par l'utilisation de cette approche pour notre application dans les chapitres 5 et 6. Le travail de [Senior 2015] en 2015 a été précurseur de la tendance, mettant à profit de larges quantités de données de parole d'enfants privées (1,9 millions d'enregistrements, soit plus de 1000 heures) récoltées par Google via Youtube Kids [Liao 2015]. Grâce à un modèle fondé sur des RNN et la fonction CTC, entraîné sur les données d'adulte et d'enfant rassemblées, ils obtiennent un WER de 9,9% et 11,3% sur de la parole d'enfants propre et bruitée (bruit artificiel et réverbération ajoutée), respectivement. En 2020 a été initié un défi pour la reconnaissance de parole d'enfants dans le cadre du IEEE Spoken Language Technology Workshop, pour lequel ont été fournies 59 heures de parole lue d'enfants entre 7 et 11 ans, en mandarin, ainsi qu'un ensemble de test de parole lue et conversationnelle [Yu 2021]. Plusieurs modèles servaient de référence au défi, dont un modèle hybride DNN-HMM et un modèle *end-to-end* Transformer, ce dernier étant un des premiers modèles *seq2seq* appliqué à la parole d'enfants. Leurs modèles de référence obtiennent des CER de 26,0% et 24,7% respectivement. Plusieurs équipes de recherche ont participé, dont deux ont soumis des rapports techniques utilisant des modèles *end-to-end* pour cette tâche : [Chen 2020] entraînent un modèle Transformer augmenté avec des couches convolutionnelles sur des données d'adultes et d'enfants assemblées, et utilisent des techniques d'augmentation de données pour atteindre un CER de 18,8%. [Ng 2020] utilisent de l'apprentissage par transfert avec un modèle *seq2seq* CTC/attention de

type LAS, et obtiennent un CER de 23,6%, puis de 20,1% avec diverses augmentations et l'utilisation d'un modèle de langage. Enfin, une étude encore plus récente de [Shivakumar 2021], soumise au journal *Computer Speech and Language*, propose une comparaison de systèmes, dont un système hybride TDNNF-HMM et trois systèmes *end-to-end* pour la RAP d'enfant. Leurs modèles sont entraînés sur une énorme quantité de parole d'adultes puis adaptés par TL grâce à environ 200 heures de parole spontanée d'enfants de 8 à 11 ans (corpus MyST [Cole 2019]), puis sont testés sur les jeux MyST et OGI. Les auteur·rice·s montrent que dans cette configuration, les systèmes *end-to-end*, et en particulier le Transformer, surpassent le TDNNF-HMM.

2.4 Bilan

Dans ce chapitre, nous avons présenté un état de l'art en adéquation avec cette thèse. Nous avons dans un premier temps détaillé le développement des systèmes de RAP avec l'évolution des modèles acoustiques, des approches génératives GMM-HMM, puis hybrides DNN-HMM aux approches récentes dites *end-to-end*. Le fonctionnement de ces différents systèmes, qui sont au cœur de notre travail, sera développé plus en profondeur dans les chapitres 3 (approche hybride) et 5 (approche *end-to-end*). Nous nous sommes ensuite penché·e·s sur ce qui rend la parole d'enfants différente de celle d'adultes, et extrêmement difficile à reconnaître. De nombreuses études acoustiques établissent les différences suivantes : gammes de fréquences décalées et élargies, mécanismes d'articulation et de co-articulation en cours de développement et qualité linguistique et prosodique dégradée. La parole lue d'apprenant·e·s lecteur·rice·s contient en outre de nombreuses disfluences et erreurs de lecture. Enfin, une revue de la littérature en RAP d'enfants a été effectuée : pour bien comprendre les articles de cette section, nous avons tout d'abord détaillé les caractéristiques des jeux de données les plus utilisés (en langue anglaise). Nous avons listé les applications éducatives utilisant de la RAP d'enfant, ou « tuteurs de lecture », qui ont fait émerger ce domaine de recherche spécifique. Nous avons relevé les conséquences des particularités de la parole d'enfants pour les systèmes de RAP, et avons déroulé l'historique des systèmes de RAP appliqués à la parole d'enfants et noté leurs conclusions les plus importantes. Ces analyses et conclusions, notamment concernant les architectures de modèles (TDNNF-HMM, modèles *end-to-end*) et les techniques d'adaptation (VTLN, i-vecteurs, TL), seront utiles pour le développement de notre travail, et seront retrouvées dans les chapitres suivants.

Deuxième partie

Établissement du modèle de référence pour la reconnaissance de phonèmes sur parole d'enfant

Chapitre 3

Système de référence : paramètres et modèles

Peu d'études traitent spécifiquement de la reconnaissance automatique de la parole d'enfants, et encore moins d'enfants très jeunes (5-8 ans). Pourtant, les techniques utilisées pour la parole d'adultes ne sont peut être pas les plus adaptées pour cette application, à cause de la grande différence acoustique entre parole d'adultes et d'enfants. De plus, la quantité de données d'entraînement peut être limitée en raison de la spécificité de la tâche, ce qui nécessite l'établissement de méthodes pour pallier à ce manque. Ce chapitre présente donc les méthodes aboutissant à l'établissement du système de référence pour la reconnaissance de phonèmes sur parole d'enfants.

Sommaire

3.1 Paramètres acoustiques

Les signaux audio de paroles sont caractérisés par une grande redondance de l'information, ainsi que par la présence d'éléments perturbateurs et non nécessaires à la reconnaissance automatique de la parole : bruit de fond, réverbération, information sur la hauteur de la voix... Pour ces raisons, plutôt que d'entraîner directement les modèles acoustiques sur le signal audio brut, une étape d'extraction de paramètres acoustiques est généralement réalisée. Historiquement, des coefficients cepstraux en fréquences Mel (*Mel-Frequency Cepstral Coefficients*, MFCC) sont utilisés [Davis 1980]. Cependant, ces coefficients ont été créés pour la reconnaissance de parole d'adultes, et pourraient ne pas être les plus adaptés pour l'extraction d'information acoustique dans la parole d'enfants. Cette section présente une sélection de différents paramètres acoustiques, fondés sur des méthodes de traitement de signal et d'extraction auto-supervisée, dont nous avons testé l'efficacité pour la reconnaissance automatique de parole d'enfants.

3.1.1 Paramètres fondés sur du traitement du signal

Tous les paramètres présentés dans cette section sont extraits grâce aux fonctionnalités de Kaldi [Povey 2011], un outil créé pour l'entraînement et l'utilisation de systèmes de reconnaissance automatique de la parole.

3.1.1.1 Coefficients cepstraux en fréquences Mel (MFCC)

Les MFCC sont les paramètres les plus couramment utilisés pour la RAP, et ont pour objectif de séparer l'information de la hauteur de la voix, dépendante du locuteur et inutile pour cette tâche, de l'information provenant du conduit vocal, dont la forme et la longueur, variables en fonction des sons, permettent de discriminer les classes phonétiques. La figure 3.1 montre la procédure d'extraction des MFCC. Le signal de parole est traité en amont : le signal passe par un filtre de pré-emphase, qui rééquilibre le spectre de fréquences en accentuant les hautes fréquences (qui sont souvent de magnitude plus faible). Il est ensuite divisé en trames acoustiques de quelques dizaines de millisecondes afin de conserver l'hypothèse de stationnarité du signal, nécessaire au calcul de transformées de Fourier. Le signal est enfin fenêtré avec une fenêtre de Hamming pour réduire les effets de bords lors de la transformée.

À partir de ce signal fenêtré, une transformée de Fourier discrète (TFD) est calculée pour passer dans le domaine fréquentiel. L'information de la phase est abandonnée en prenant l'amplitude du signal. Un banc de filtres en échelle Mel est appliqué, avec chaque filtre H de forme triangulaire, défini pour une fréquence de coupure f_c selon :

$$H_{f_c}(k) = \begin{cases} \frac{k - f_{c-1}}{f_c - f_{c-1}} & \text{si } f_{c-1} \leq k < f_c \\ \frac{f_{c+1} - k}{f_{c+1} - f_c} & \text{si } f_c \leq k \leq f_{c+1} \\ 0 & \text{sinon} \end{cases} \tag{3.1}$$

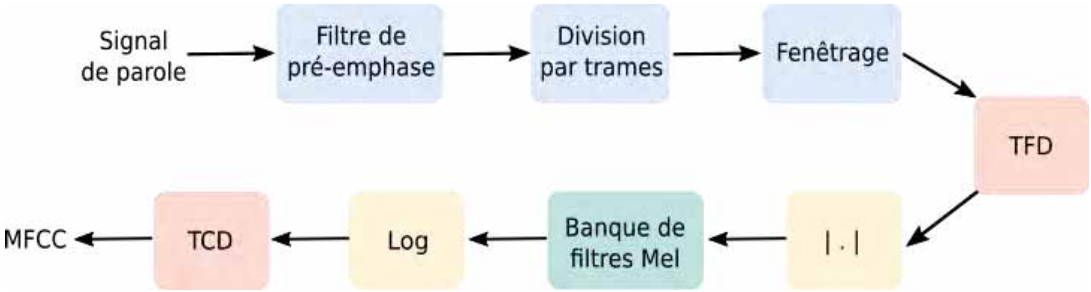


FIGURE 3.1 – Procédure d'extraction des MFCC

avec k tout point sur l'échelle fréquentielle en Hertz, et f_{c-1} et f_{c+1} les fréquences de coupures des filtres précédent et suivant. Ce banc de filtres permet de réduire la dimension des données et d'imiter le comportement de la cochlée. En effet, l'échelle Mel est une échelle perceptive de hauteurs des sons, créée pour imiter ce qu'une oreille humaine entend.

Enfin, le signal est converti en échelle logarithmique, et une transformée en cosinus discrète (TCD) est appliquée : ceci permet de rejoindre le domaine cepstral tout en décorrélant les paramètres. Les paramètres cepstraux 2 à n, avec $n = 13$ ou $n = 40$ en général, sont retenus. Il est commun d'adjoindre à ces coefficients l'énergie du signal en première position, et d'ajouter à la suite les dérivées premières et secondes, calculées sur cinq fenêtres adjacentes. Les paramètres ainsi obtenus sont de dimension $3n$.

3.1.1.2 Coefficients Filterbank

Les coefficients *filterbank* sont obtenus par les mêmes étapes que les MFCC, en s'arrêtant après l'application du banc de filtres Mel. Le fait de ne pas appliquer les dernières étapes permet d'obtenir des informations plus complètes, potentiellement redondantes pour certains modèles acoustiques (notamment les modèles DNN-HMM), mais intéressantes pour d'autres, capables d'extraire de l'information acoustique pertinente complémentaire, comme les systèmes fondés sur des réseaux récurrents ou convolutifs.

3.1.1.3 Paramètres de prédiction linéaire perceptive (PLP)

La prédiction linéaire perceptive (*Perceptual Linear Prediction*, PLP) a été proposée dans [Hermansky 1991], comme amélioration de la prédiction linéaire conventionnelle, avec pour objectifs une meilleur cohérence avec l'audition humaine, une meilleure résistance au bruit, et donc de meilleures performances dans les tâches de RAP.

L'extraction des paramètres PLP est décrite sur la figure 3.2. Comme pour les MFCC, le signal de parole est divisé en trames acoustiques et fenêtré avec une fenêtre de Hamming, et enfin transporté dans le domaine fréquentiel à l'aide d'une TFD. Le spectre de puissance à court terme est calculé en prenant l'amplitude au carré du signal. Les fréquences sont ensuite déformées grâce à un banc de filtres Bark : la différence avec les MFCC réside dans l'utilisation

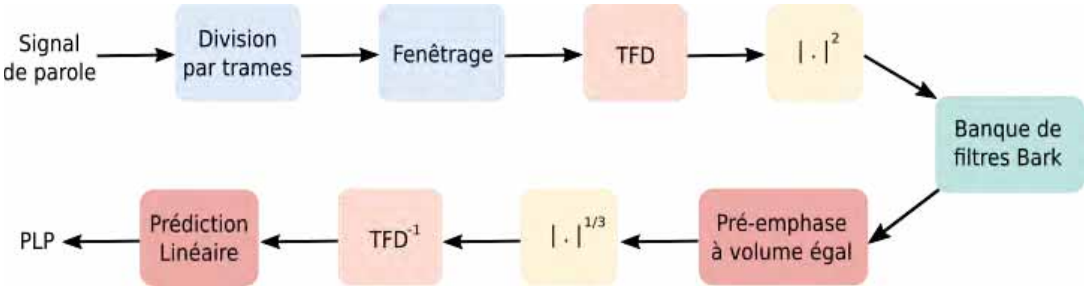


FIGURE 3.2 – Procédure d'extraction des paramètres PLP

de l'échelle Bark [Schroeder 1977] (une échelle perceptive également) et dans la forme du banc de filtres. Suivent l'application d'un filtre de pré-emphase à volume égal, qui simule la sensibilité de l'oreille humaine à un volume donné, et d'une racine cube, qui simule la relation non-linéaire entre l'intensité d'un son et son volume perçu. Ces deux étapes remplacent respectivement les étapes de pré-emphase et de compression logarithmique de l'extraction des MFCC (voir figure 3.1). Enfin, une transformée de Fourier inverse est appliquée pour revenir dans le domaine temporel, et une prédiction linéaire conventionnelle.

3.1.1.4 Paramètres RASTA-PLP

Les paramètres PLP sont fondés sur le spectre à court terme du signal de la parole, et sont peu robustes à la variabilité de ce spectre causée par la réponse en fréquence du canal de communication. Le filtrage RASTA (*RelAtive SpecTrAl*) a été proposé afin de rendre les paramètres PLP plus robuste à ces potentielles distorsions spectrales [Hermansky 1992]. Cette technique consiste à modifier le spectre à court terme initial en appliquant un filtre passe-bande sur chaque bande de fréquences, supprimant ainsi les composantes constantes ou variant très lentement.

3.1.1.5 Coefficients cepstraux en fréquences Gammatone

Les coefficients cepstraux en fréquences Gammatone ressemblent beaucoup aux MFCC : ils sont calculés de façon quasi-identique selon le schéma 3.1, avec un banc de filtres Gammatone à la place du banc de filtres Mel. Plutôt que par une forme triangulaire comme un filtre Mel, un filtre Gammatone, dont la réponse impulsionnelle a été proposée dans [Patterson 1992], est défini pour une fréquence de coupure f_c selon :

$$H_{f_c}(k) = \alpha k^{n-1} e^{-2\pi bk} \cos(2\pi f_c + \phi) \tag{3.2}$$

avec k tout point sur l'échelle fréquentielle en Hertz, α le gain du filtre, n son ordre ($n \leq 4$), ϕ la phase (généralement fixée à zéro) et b la bande passante du filtre en Hertz. Les filtres Gammatone ont le même objectif d'imiter le système auditif humain que les filtres Mel, mais visent de plus à réduire la sensibilité des paramètres au bruit additif, inconvénient des MFCC.

3.1.2 Paramètres extraits par des modèles auto-supervisés

Plusieurs études ont mené à la création de modèles d'extraction de paramètres, entraînés de façon auto-supervisée sur de grandes quantités de données non annotées. Ces modèles sont entraînés à extraire, à partir du signal audio brut, des représentations générales et très robustes contenant l'information nécessaire à différentes tâches de traitement de la parole (reconnaissance de locuteur, détection d'émotions, RAP...). Leur objectif est de fournir des paramètres d'entrée plus pertinents que les approches classiques fondées sur du traitement de signal et ainsi améliorer les performances sur la tâche en question.

Les méthodes auto-supervisées utilisent généralement une tâche « proxy » pour laquelle le modèle est entraîné, et une fonction objectif qui mesure la performance du modèle sur la tâche proxy. La tâche proxy peut consister, dans le cas d'une extraction de paramètres audio, à prédire la forme d'onde de parole, des paramètres MFCC ou toute autre caractéristique de la parole. Les représentations générées par ce type de modèles sont destinées à être fournies en entrée de modèles acoustiques pour remplacer des paramètres classiques tels les MFCC ou les Filterbank. Les modèles peuvent également être ajoutés à l'entrée d'un système de RAP : en les incluant dans la propagation arrière effectuée durant l'entraînement, ils peuvent s'adapter aux données spécifiques de l'application. Certains modèles ont été mis à disposition, gratuitement, afin d'encourager leur utilisation par la communauté scientifique. Nous avons étudié et testé pour l'extraction de représentations de parole d'enfants deux de ces modèles auto-supervisés : Wav2vec [Schneider 2019] et PASE+ [Ravanelli 2020].

3.1.2.1 Modèle Wav2vec

Le modèle Wav2vec, proposé en 2019 dans [Schneider 2019], est un réseau CNN constitué de deux parties : un encodeur et un réseau contextuel. Le premier intègre le signal audio X appartenant à un domaine $\mathcal{X}$, divisé en trames audio, dans un état caché $\mathcal{Z}$, et génère pour une trame i une représentation compacte de paramètres z_i contenant les caractéristiques du signal de parole. Le second est un réseau auto-régressif qui combine v représentations successives passées $z_i ... z_{i-v}$ en un unique vecteur de contexte c_i, appartenant au domaine $\mathcal{C}$.

Le modèle Wav2vec est optimisé à distinguer un échantillon véritable z_{i+k} (k désignant le nombre de pas dans le futur par rapport au pas courant i) d'autres échantillons de distraction $\hat{z}$ tirée d'une distribution p_n. L'optimisation se fait par la minimisation de la fonction *constrative loss* $\mathcal{L}_k$ à chaque pas [Gutmann 2010] :

$$\mathcal{L}_k = -\sum_{i=1}^{T-k} (\log\sigma(z_{i+k}^{\top} h_k(c_i)) + \lambda E_{\hat{z}\sim p_n}(\log\sigma(-\hat{z}^T h_k(c_i)))) \tag{3.3}$$

avec σ la fonction sigmoïde, T la longueur de la séquence, λ le nombre d'échantillons de distraction, $\sigma(z_{i+k}^{\top} h_k(c_i))$ la probabilité que z_{i+k} soit un échantillon correct, et h_k la fonction affine appliquée au vecteur c_i au pas k. La fonction de coût total est la somme des coûts à

chaque pas $\mathcal{L} = \sum_{k=1}^{K} \mathcal{L}_k$.

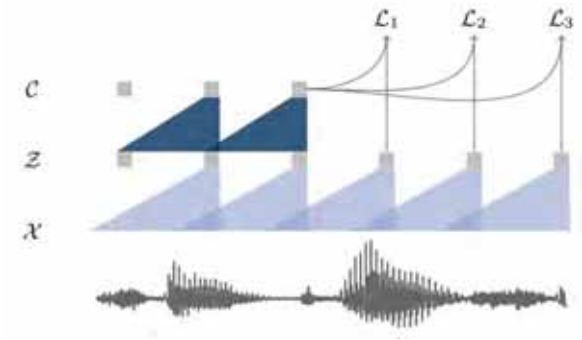


FIGURE 3.3 – Architecture du modèle Wav2vec, tirée de [Schneider 2019]

L'encodeur est constitué de cinq couches CNN à une dimension puis de deux couches affines en sortie, et le réseau contextuel de neuf couches CNN. Chaque couche CNN est composée d'une convolution causale à 512 canaux, d'une couche de normalisation par groupe, et d'une non-linéarité ReLU (*Rectified Linear Unit*).

Le modèle Wav2vec mis à disposition à la communauté scientifique, que nous avons utilisé, a été entraîné sur un mélange de données de parole d'adultes anglaise, composé de 81 heures provenant du corpus WSJ [Garofolo 1993a] et de 960 heures provenant du corpus Librispeech [Panayotov 2015]. Dans les expériences du papier original, évaluées grâce au jeu de test WSJ, remplacer les paramètres Filterbank par des représentations extraites d'un modèle Wav2vec améliore significativement les performances de plusieurs systèmes de reconnaissance de graphèmes. Un de ces modèles, entraîné sur Librispeech, obtient un taux WER de 2,43% et surpasse le modèle Deep Speech 2 [Amodei 2016], meilleur modèle de la littérature pour la reconnaissance de graphème, avec beaucoup moins de données. Les mêmes tendances sont observées pour la reconnaissance automatique de phonèmes sur le corpus TIMIT [Garofolo 1993b].

3.1.2.2 Modèle PASE+

Les noms des modèles PASE [Pascual 2019] et PASE+ [Ravanelli 2020] signifient *Problem Agnostic Speech Encoder* : ces modèles ont pour objectif d'extraire des représentations robustes et agnostiques de la tâche. Ces modèles diffèrent du modèle wav2vec dans les choix de la tâche « proxy » et de la fonction objectif. La tâche proxy est en effet un ensemble de tâches très diverses, qui permettent d'apprendre des représentations transportant de l'information variée, comme l'identité d'un locuteur, les phonèmes prononcés, ou même la présence d'émotions. La fonction objectif est une simple erreur quadratique moyenne.

L'architecture du modèle PASE est schématisée sur la figure 3.4 : il contient un encodeur et sept décodeurs, appelés *workers*. L'encodeur prend en entrée le signal audio brut, et est composé d'un réseau SincNet [Ravanelli 2018] qui se comporte comme un banc de filtre rectangulaires, de couches convolutives, d'une couche linéaire et d'une couche de normalisation par batch qui

génère la représentation finale du modèle.

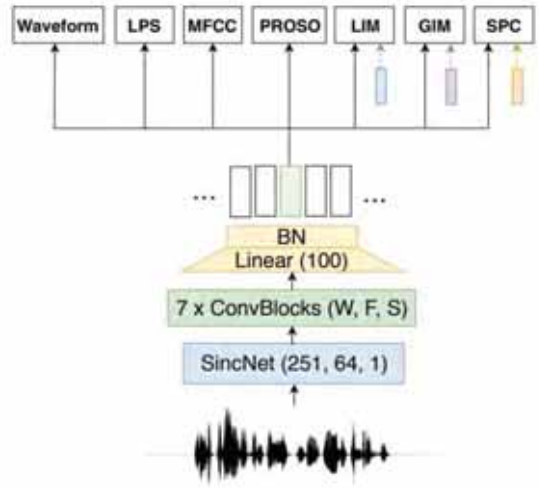


FIGURE 3.4 – Architecture du modèle PASE, tirée de [Pascual 2019]

Les *workers* sont de petits réseaux de neurones avec peu de couches et peu de neurones, pour encourager le modèle à découvrir des paramètres haut-niveau qui pourront être exploitables par des réseaux de toute taille. Parmi les *workers*, trois ont une tâche de régression consistant à extraire un certain type de paramètres et utilisent l'erreur quadratique moyenne entre les paramètres cibles et les prédictions du réseau : LPS (spectre de puissance logarithmique), MFCC et Prosodie. Les paramètres sont calculés à l'aide d'une fenêtre de Hamming sur les représentations de l'encodeur. Le *worker* « Prosodie » regroupe quatre paramètres différents, comme la probabilité d'avoir un phonème voisé/non voisé, le taux de passage à zéro, l'énergie et le logarithme de la fréquence fondamentale. Le *worker* « forme d'onde » utilise également une régression pour prédire la forme d'onde du signal, mais utilise l'erreur moyenne absolue par souci de robustesse aux pics proéminents d'un signal de parole. Enfin, les trois derniers *workers* effectuent une tâche de discrimination binaire, qui permettent d'évaluer des informations de plus haut niveau : le maximum d'information mutuelle locale (LIM) [Ravanelli 2019], le maximum d'information globale (GIM) [Pascual 2019] et le codage de prédiction de séquence (SPC) [Pascual 2019]. Les deux premiers apprennent des informations complémentaires sur l'identité du locuteur. Le dernier a pour objectif de saisir de l'information sur la causalité du signal afin d'établir un contexte large.

Le modèle PASE+ est une extension du modèle PASE, visant à le rendre plus robuste aux environnements bruités et réverbérants. L'architecture PASE+ est présentée sur la figure 3.5, avec les éléments ajoutés à l'originel modèle PASE désignés en bleu. Pour améliorer la robustesse au bruit, un module de distorsion de la parole a été ajouté à l'encodeur, qui déforme des segments de parole non-bruitée à l'aide de réverbération, de bruit additif, de masques temporels/fréquentiels ou encore de parole superposée. L'encodeur fondé sur des CNN est combiné avec un réseau de neurones quasi-récurrent (*Quasi-Recurrent Neural Network*, QRNN) [Bradbury 2017], qui a la capacité d'apprendre des dépendances très long terme sur les signaux de parole. Enfin, de nouveaux ouvriers ont été ajoutés, permettant de prédire de nouveaux paramètres (filterbank, coefficients cepstraux en fréquences Gammatone) et sur de

plus longues fenêtres d'analyse (200 ms au lieu des 25 ms habituelles) afin de modéliser des caractéristiques de parole de plus haut niveau.

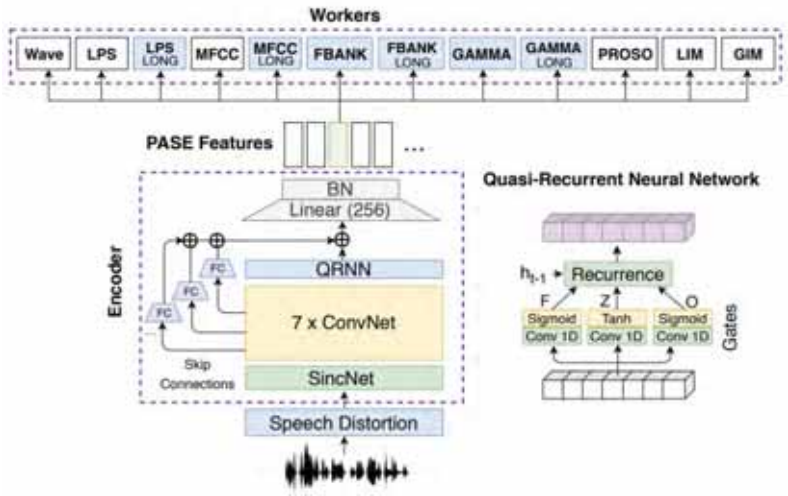


FIGURE 3.5 – Architecture du modèle PASE+, tirée de [Ravanelli 2020]

3.1.3 Méthodes d'adaptation des paramètres

Plusieurs méthodes permettent d'adapter les paramètres acoustiques, dans un objectif d'amélioration de la robustesse à l'environnement sonore ou de réduction de variabilité inter- ou intra-locuteur. Parmi elles, nous utiliserons dans cette thèse la normalisation cepstrale de la moyenne et variance (*Cepstral Mean Variance Normalisation*, CMVN) et la normalisation de la longueur du conduit vocal (*Vocal Tract Length Normalisation*, VTLN).

3.1.3.1 Adaptation aux locuteur·rice·s avec i-vecteurs

Les i-vecteurs sont des vecteurs de paramètres, introduits pour la vérification de locuteur·rice dans [Dehak 2011], qui modélisent une représentation du·de la locuteur·rice. Cette technique vise à améliorer la technique précédente de *Joint Factor Analysis* [Kenny 2007] en combinant la modélisation de la variabilité inter- et intra-locuteur·rice dans un seul espace de dimension réduite. Les i-vecteurs ont ensuite été utilisés pour des tâches de RAP d'adultes, en étant concaténés aux paramètres acoustiques d'entrée : leur utilisation apporte une amélioration relative du WER de 10% dans [Saon 2013] sur le corpus de parole téléphonique Switchboard [Godfrey 1993], et de 9% dans [Senior 2014] sur le corpus *Google Voice Search*.

Les i-vecteurs sont rapidement devenus indispensables pour des performances à l'état de l'art en RAP d'adultes, et sont utilisés par défaut dans toutes les recettes de l'outil Kaldi. En revanche, leurs deux majeurs inconvénients, discutés dans [Verma 2015], pourraient s'avérer néfastes pour notre application :

— la bonne performance de cette technique dépend fortement de la sélection des données utilisées pour entraîner l'extracteur de i-vecteurs, qui peut être difficile avec des données aussi hétérogènes que les nôtres (forte variabilité de la parole d'enfants, différents niveaux de bruit et de qualité de lecture...). A défaut, une grande quantité de données est nécessaire pour correctement modéliser les représentations, ce dont nous ne disposons pas ;

— de moins bonnes performances ont été identifiées dans le cas d'enregistrements courts en raison d'une mauvaise précision des représentations. Cela pourrait affecter l'efficacité des i-vecteurs sur notre corpus de parole d'enfants, constitué d'enregistrements de mots isolés et de phrases courtes.

3.1.3.2 Normalisation cepstrale de la moyenne (CMN) et de la variance (CMVN)

La CMVN est une méthode visant à réduire l'influence des changements sonores de l'environnement sur la performance de RAP, en normalisant les paramètres d'entrée pour une meilleure uniformité indépendamment du bruit environnant [Viikki 1998]. Le niveau de bruit et la qualité de l'environnement sonore dépendant du matériel d'enregistrement, de la salle de classe, du nombre et de l'humeur des élèves présents, la normalisation CMVN paraît adaptée à notre application.

Cette méthode consiste donc à normaliser la moyenne et la variance des coefficients cepstraux à 0 et 1, respectivement. Les vecteurs de paramètres sont normalisés selon :

$$\hat{x}_t(i) = \frac{x_t(i) - \mu_t(i)}{\sigma_t(i)} \tag{3.4}$$

avec $x_t(i)$ le $i^{\text{ième}}$ composant du vecteur de paramètre original au temps t et $\hat{x}_t(i)$ sa version normalisée. Les coefficients de moyenne $\mu_t(i)$ et variance $\sigma_t(i)$ sont calculés sur des fenêtres glissantes de longueur N, selon les équations suivantes :

$$\mu_t(i) = \frac{1}{N} \sum_{n=t-N/2}^{t+N/2-1} x_n(i) \tag{3.5}$$

$$\sigma_t(i) = \sqrt{\frac{1}{N} \sum_{n=t-N/2}^{t+N/2-1} (x_n(i) - \mu_t(i))^2} \tag{3.6}$$

Le calcul de ces coefficients de normalisation peut se faire au niveau de l'enregistrement ou au niveau du locuteur. Dans notre cas, la correspondance locuteur-enregistrement est disponible : les coefficients de normalisation sont calculés au niveau du locuteur afin de maximiser la quantité d'information disponible et donc la précision des coefficients.

La CMVN, de par sa simplicité, est conseillée dans le cas d'enregistrements courts par rapport à des techniques comme l'égalisation d'histogramme, qui nécessite une quantité de données importante pour estimer les paramètres de normalisation [Joshi 2013]. Son utilisation

paraît donc appropriée pour notre corpus, qui est constitué d'enregistrements de lecture orale de mots isolés et de phrases courtes. Les recettes de Kaldi incluent par défaut généralement de la CMN : normalisation de la moyenne, mais pas de la variance.

3.1.3.3 Normalisation de la longueur du conduit vocal (VTLN)

Les fréquences des enfants sont significativement plus hautes que celles des adultes, ce qui constitue une des principales différences. La hauteur des fréquences dépend en grande partie de la longueur du conduit vocal : plus il est court, plus les fréquences seront hautes. Les enfants ont donc un conduit court et des voix aigües, qui descendent vers le grave lors de leur croissance, jusqu'à l'âge adulte.

La normalisation de la longueur du conduit vocal a pour objectif de réduire l'impact de la forme du conduit vocal du locuteur sur la reconnaissance, et donc la variabilité entre locuteurs, en étendant ou compressant sa gamme de fréquences [Lee 1996]. Les fréquences sont transformées suivant l'équation 3.7, où $F^{\alpha}_{\text{vtln}}(f)$ est la fréquence transformée à partir de la fréquence originale f, et α est le facteur de déformation. Ce coefficient α permet de choisir entre une extension ($\alpha > 1$) et une compression ($\alpha < 1$) de la gamme de fréquences, ainsi que de régler l'amplitude de la transformation. f_{min} (resp. f_{max}) est la fréquence fixe en dessous (resp. au dessus) de laquelle les fréquences ne sont pas étendues (resp. compressées).

$$\begin{aligned} \text{si } \alpha \geq 1 : F^{\alpha}_{\text{vtln}}(f) &= \begin{cases} f & \text{si } f \leq f_{\text{min}} \\ \alpha \times f & \text{sinon} \end{cases} \\ \text{si } \alpha < 1 : F^{\alpha}_{\text{vtln}}(f) &= \begin{cases} \alpha \times f & \text{si } f \leq f_{\text{max}} \\ f & \text{sinon} \end{cases} \end{aligned} \tag{3.7}$$

La plupart des études proposent d'utiliser la VTLN sur des corpus contenant des locuteurs différents (homme-femme [Lee 1996] ou homme-femme-enfant [Gray 2014, Serizel 2014b, Shivakumar 2014]) pour normaliser leurs gammes de fréquences et donc réduire les différences acoustiques liées à la longueur du conduit vocal. Nous proposons ici d'adapter les voix d'adultes en rapprochant leurs gammes de fréquences de celles d'enfants, dans l'objectif de les utiliser pour de l'apprentissage par transfert (voir section 3.2.1).

3.2 Gestion du manque de données

Les réseaux de neurones nécessitent une certaine quantité de données pour apprendre à modéliser de la parole correctement. Cependant, lorsque peu de données sont disponibles (langues peu dotées, difficultés pour récupérer et annoter des données très spécifiques...), des techniques existent pour améliorer la performance des modèles acoustiques, en particulier l'apprentissage par transfert et l'augmentation de données, utilisées dans mon travail.

3.2.1 Apprentissage par transfert

L'apprentissage par transfert (TL) est une méthode d'adaptation permettant de pallier à un manque de données dans un domaine d'application. Le principe est d'acquérir des connaissances générales sur une grande quantité de données comparables mais hors du domaine d'application, puis de les adapter au domaine d'application à l'aide d'un petit jeu de données spécifiques. Pour cela, il faut premièrement entraîner un premier modèle, dit modèle source, sur le jeu de données hors-domaine, puis effectuer une deuxième phase d'entraînement avec le jeu de données du domaine pour obtenir le modèle cible.

Cette méthode est généralisable à de nombreuses applications de reconnaissance automatique de la parole : adaptation entre différents langages ([Abad 2020, Tong 2017b, Cho 2018]), types de parole (par exemple entre du discours de télévision et de la parole conversationnelle [Abad 2020]), ou encore entre de la parole de natifs et de non-natifs [Duan 2020]. La procédure peut différer en fonction de l'application, de la quantité de données cibles disponible, ou encore le niveau de similarité entre les données sources et cibles.

Pour l'adaptation entre parole d'adultes et parole d'enfants, le modèle source est entraîné sur de la parole d'adultes, et les données cibles sont des enregistrements de parole d'enfants. L'apprentissage par transfert a démontré son efficacité à plusieurs reprises pour l'adaptation adulte-enfant, notamment avec des modèles hybrides DNN-HMM [Shivakumar 2020, Tong 2017a, Qian 2016]. Cette adaptation spécifique est très sensible à la quantité de données cible, ainsi qu'à l'âge des enfants : comme détaillé au chapitre 2, l'appareil vocal et la qualité de parole se développent continuellement pendant la croissance de l'enfant, ce qui introduit de fortes variabilités selon l'âge, et d'importantes différences acoustiques avec la parole d'adultes [Shivakumar 2020].

Plusieurs stratégies de TL sont envisageables : la première, dénotée *Ré-init*, consiste à réinitialiser la couche de sortie du modèle source avant l'application du transfert, afin d'adapter spécifiquement les sorties aux données cibles, comme dans [Tong 2017a, Tong 2017b]. La seconde, dénotée *Gelée*, a été imaginée en inversant la réflexion de la première et à partir d'une hypothèse de [Shivakumar 2020], qui suggère que la variabilité acoustique de la parole d'enfants est apprise par les premières couches du réseau. Nous faisons l'hypothèse que la couche de sortie contient majoritairement des informations décisionnelles de classification phonétique et non acoustiques, qui ne dépendent pas de la parole d'enfants et dont la qualité pourrait être dégradée par une trop grande variabilité dans les données cibles. Cette méthode consiste donc à geler les poids de la couche de sortie pour conserver uniquement les connaissances apprises sur la parole d'adultes, et à adapter les autres couches. La troisième stratégie, dénotée *Complet*, s'appuie sur les résultats de Shivakumar et al. [Shivakumar 2020]. Sur leur corpus de parole d'enfants âgés de 6 à 14 ans, les auteurs concluent que deux stratégies sont équivalentes : adapter le réseau neuronal tout entier ou uniquement les deux premières et deux dernières couches de neurones. Cependant, pour de jeunes enfants (6 à 8 ans), ils montrent que modéliser les caractéristiques acoustiques et prosodiques de leur parole, très complexe et variable, nécessite plus de paramètres (et donc plus de couches). Cela revient donc à favoriser la première option : adapter tout le réseau. Une quantité minimum de données d'adaptation est

néanmoins requise (~10 heures) : sans cela, adapter le réseau entier dégraderait la performance à cause du biais introduit par la très forte variabilité de la parole des jeunes enfants. En revanche, si le jeu de parole d'enfants est suffisamment grand et contient suffisamment de locuteurs, le réseau est capable de modéliser cette variabilité, et adapter toutes les couches fonctionne mieux qu'adapter uniquement quelques couches. Nous verrons dans les prochains chapitres que notre jeu de données correspond au scénario détaillé ci-dessus : composé de légèrement plus de 10 heures de parole d'enfants âgés de 5 à 8 ans. La stratégie d'apprentissage par transfert *Complet* consistera donc à adapter toutes les couches du modèle source avec les données cibles.

3.2.2 Augmentation de données avec bruit de brouhaha

Le bruit de brouhaha (*babble noise* ou encore *cocktail party noise*) désigne le bruit engendré par un grand nombre de personnes parlant en même temps. Lorsque l'application Lalilo est utilisé en salle de classe, les enfants sont en autonomie, avec une supervision réduite de l'enseignant·e qui peut utiliser ce temps pour suivre d'autres élèves. Les enfants sont plus libres de discuter entre eux, ce qui engendre inévitablement un niveau important de bruit de brouhaha sur certains enregistrements de parole d'enfants. Le bruit est d'autant plus présent que les microphones utilisés sont souvent non directifs et de mauvaise qualité. Dans ce cadre applicatif, notre système de RAP doit être robuste au bruit de brouhaha afin de donner des retours les plus précis possibles aux enfants, et les aider dans leur apprentissage.

La robustesse au bruit est un sujet récurrent dans le domaine de la RAP, et beaucoup de travaux traitant de ce problème sont disponibles dans la littérature. Dans une revue de littérature sur le sujet [Li 2014a], nous trouvons des méthodes s'appuyant sur de nouveaux paramètres acoustiques ou sur des méthodes d'adaptation de modèles acoustiques. L'entraînement multi-conditions est une technique d'augmentation de données qui consiste à superposer différents types de bruit, à différents niveaux de bruit, aux enregistrements de parole du jeu d'entraînement. Cela permet au modèle acoustique d'apprendre à reconnaître la parole dans le bruit environnant lors de son entraînement. Des jeux de données de bruit, comme DEMAND [Thiemann 2013] ou NOISEX [A.Varga 1993], sont disponibles librement, et contiennent notamment des enregistrements de bruit de brouhaha. Il est également possible de créer artificiellement du bruit de brouhaha en superposant la parole de plusieurs locuteurs [Wu 2019]. De nombreuses études ont trouvé que l'entraînement multi-condition avec de la réverbération et différents types de bruit permettait d'améliorer la robustesse au bruit des modèles [Rajnoha 2009, Ko 2017, Braun 2017, Gibson 2018, Airaksinen 2019].

3.3 Architecture du système de référence

Les architectures hybrides DNN-HMM pour la modélisation acoustique se sont démocratisées ces dernières décennies, grâce à la disponibilité de larges quantités de données et l'augmentation des puissances de calcul. Les performances de ces architectures ont dépassé

les GMM-HMM dans l'état de l'art. Notre architecture de référence correspond donc à une architecture hybride DNN-HMM, avec pour DNN un réseau de neurones factorisé à délais temporels (*Time-Delay Factorised Neural Network*, TDNNF).

3.3.1 Architecture du TDNNF

Le TDNNF est une amélioration du modèle TDNN, introduit pour la reconnaissance de phonèmes dans [Waibel 1989]. Ce modèle a été utilisé avec succès pour la reconnaissance de voyelles dans de la parole d'enfants avec de faibles quantités de données [Yong 2011]. Une modification de ce traditionnel TDNN a donné lieu à l'architecture TDNNF, qui a démontré une plus grande efficacité, notamment sur des applications avec peu de ressources [Povey 2018]. Cette architecture a également obtenu de meilleures performances sur un petit corpus de parole d'enfants en langue anglaise [Wu 2019].

3.3.1.1 Réseau de neurones à délais temporels (TDNN)

Un réseau de neurones classique calcule une somme pondérée des entrées données, qu'il fait ensuite passer dans une fonction non-linéaire (sigmoïde par exemple). Le TDNN introduit des délais D_1,...D_N dans cette somme pondérée, comme nous pouvons le voir sur la figure 3.6, où w_{i+n} est le poids associé à la trame d'entrée i différée de n trames. En pratique, l'instauration de ces délais correspond à l'extension spatiale du contexte d'entrée d'une couche sur un nombre N de trames d'entrée. Une couche TDNN peut ainsi être considérée comme une couche de convolution à une dimension entièrement connectée. Le nombre N de trames varie en fonction des couches du réseau, ce qui permet de profiter de différentes largeurs contextuelles : les couches proches de l'entrée utilisent généralement un N faible pour apprendre les caractéristiques acoustico-phonétiques de courte durée, alors que les couches proches de la sortie modélisent des caractéristiques plus complexes de longue durée avec un N plus important.

Ces délais temporels instaurent inévitablement de la redondance dans les évènements acoustiques vus par le réseau. Les poids correspondants à ces copies doivent être identiques pour ne pas troubler l'information de position temporelle de chaque unité. Ainsi, plutôt que de mettre à jour les poids séparément, le poids de chaque unité décalée de n trames est calculé comme la moyenne des changements sur toutes les versions différées de cette unité. Par exemple, si $N = 2$, le poids $W_{(i)+0}$ de l'unité i décalée de 0 trame sera mis à jour selon :

$$W_{(i)+0} \leftarrow W_{(i)+0} + \frac{\delta W_{(i)+0} + \delta W_{(i-1)+1} + \delta W_{(i-2)+2}}{N+1}$$

avec $\delta W_{(i-n)+n}$ la mise à jour à appliquer au poids de l'unité $(i-n)$ décalée de n trame(s). Le vecteur de sortie d'une couche TDNN contenant I unités avec un délai N est donc de taille $I \times (N+1)$.

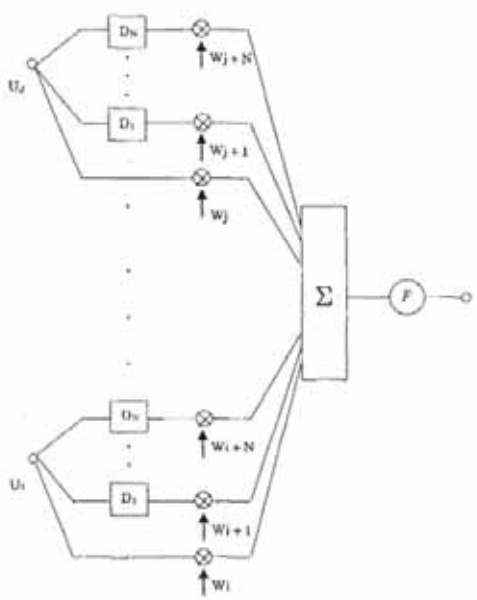


FIGURE 3.6 – Schéma d'une unité de réseau de neurones à délais temporels, tiré de [Waibel 1989]

3.3.1.2 Réseau de neurones à délais temporels factorisé (TDNNF)

La prise en compte de larges contextes de trames permet de modéliser efficacement les caractéristiques complexes de la parole, mais le chevauchement des trames rend les calculs particulièrement non-efficients. Le TDNNF améliore l'efficience de calcul en réduisant le nombre de paramètres grâce à la méthode de décomposition en valeurs singulières. La matrice de poids W de chaque couche est décomposée en une approximation du produit de deux matrices de rang inférieur en négligeant les plus petites valeurs singulières [Povey 2018] :

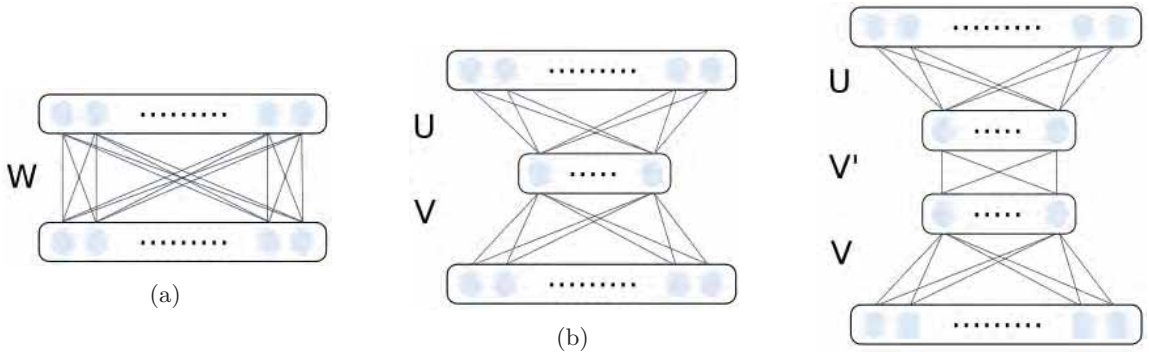
$$W = UV \tag{3.8}$$

où $W \in \mathbb{R}^{u \times v}$, $U \in \mathbb{R}^{u \times k}$ et $V \in \mathbb{R}^{k \times v}$.

En choisissant une valeur $k \leq \min(m,n)$, cela revient à insérer une couche d'étranglement (*bottleneck*) additionnelle dans une couche traditionnelle de TDNN, réduisant le nombre de paramètres de la couche. La figure 3.7b illustre l'insertion d'une couche *bottleneck* entre deux couches classiques de TDNN (figure 3.7a).

Afin de rendre l'entraînement stable, il est nécessaire de contraindre la matrice V à être semi-orthogonale, c'est à dire respectant une des conditions $VV^T = I$ ou $V^TV = I$. Cette contrainte est appliquée toutes les quelques mises à jour des paramètres du réseau complet, sous la forme d'une fonction objectif additionnelle s'assurant que la matrice reste proche de la semi-orthogonalité [Povey 2018] :

$$V \leftarrow V - 4\nu(VV^T - I)V \tag{3.9}$$



(a) (b) (c)

FIGURE 3.7 – Couches de (a) TDNN standard (b) TDNNF avec une couche *bottleneck* (c) TDNNF avec deux couches *bottleneck*

avec I la matrice identité, et $\nu = \frac{1}{8}$ un taux d'apprentissage, dont la valeur $\frac{1}{8}$ permet d'assurer la convergence quadratique de la fonction.

Les auteurs de la méthode ont par ailleurs démontré qu'il était encore plus efficace d'insérer deux couches *bottleneck* de dimensions égales, comme sur la figure 3.7c, avec les matrices $V \in \mathbb{R}^{k \times v}$ et $V' \in \mathbb{R}^{k \times k}$ contraintes à la semi-orthogonalité. Enfin, certaines couches sont reliées par des *skip connections* : elles reçoivent non seulement la sortie de la couche précédente $i-1$, mais également les sorties concaténées d'une sélection de couches précédentes $\leq i-2$. Cette structure de couche TDNNF a été introduite dans l'outil Kaldi et est celle que nous utilisons dans cette thèse.

Notre réseau TDNNF est formé de 12 couches TDNNF, chacune contenant une couche TDNN de dimension 1024 et deux couches *bottleneck* de dimension 128, comme illustré sur la figure 3.7c. La factorisation des couches TDNN en TDNNF a permis de diminuer le nombre de paramètres de 12,5 à 7,6 millions.

3.3.2 Paramètres et entraînement du TDNNF-HMM

Les processus d'extraction de paramètres, d'entraînement et de décodage avec le TDNNF-HMM sont entièrement réalisés avec Kaldi.

3.3.2.1 Alignements et paramètres d'entrée

Un modèle GMM-HMM monophone à 3 états fournit les alignements pour entraîner le modèle TDNNF. Comme nous l'avons vu au chapitre 2, des études sur le développement de la parole chez l'enfant montrent de fortes variabilités intra-locuteur·rice temporelle et spectrale dans les transitions consonne-voyelle, suggérant que les mécanismes de co-articulation des jeunes enfants ne sont pas encore stables [Gerosa 2006b, Zharkova 2015]. L'utilisation de triphones ne transporte ainsi peut être que peu d'information pertinente pour la reconnaissance

de parole de jeunes enfants. De plus, les enfants présentant une forte variabilité inter- et intra-locuteur dans leurs prononciations [Lee 1999], un nombre d'occurrences de chaque triphone relativement élevé est nécessaire pour une correcte modélisation, ce qui n'est pas le cas dans un petit corpus de parole d'enfants. Enfin, la présence d'erreurs de déchiffrage dans le corpus de test introduit des triphones rares ou non-existants dans la langue française, qui ne seraient pas représentés pendant l'apprentissage, et donc pas détectés en phas de test. Ces hypothèses ont été vérifiées expérimentalement, et de meilleurs résultats ont été obtenus avec un modèle monophone.

Les paramètres d'entrée des modèles GMM-HMM et TDNNF-HMM sont des coefficients MFCC de dimension 13 et 40 respectivement, sur lesquels est appliqué une normalisation CMVN, comme dans [Bayerl 2019], ainsi que dans les recettes Kaldi conventionnelles. Les dérivées premières et secondes sont ajoutées à la suite de ces paramètres. Un sous-échantillonnage de facteur 3 est effectué sur les trames d'entrée, ce qui accélère les calculs, et permet en outre d'appliquer de la perturbation de vitesse aux facteurs 0,9 et 1,1 pour augmenter la quantité de données.

3.3.2.2 Fonction objectif

Nous utilisons un modèle DNN-HMM dit « chaîne » [Povey 2016], qui diffère d'un modèle DNN-HMM conventionnel par l'utilisation d'une fonction objectif au niveau de la séquence plutôt qu'au niveau de la trame : la fonction de maximisation de l'information mutuelle (*Maximum Mutual Information* en anglais, MMI). Cette fonction vise à maximiser l'information mutuelle entre les distributions de la séquence observée et de la séquence de référence, ce qui revient à minimiser l'erreur attendue sur la séquence. Elle est définie selon :

$$F_{\text{MMI}} = \sum_u log \frac{p(O_u|S_u)^\kappa P(S_u)}{\sum_S p(O_u|S_u)^\kappa P(S)} \tag{3.10}$$

où O_u et S_u sont les séquences observée et de référence sur l'énoncé u, et κ est un facteur pour corriger la sur-estimation. La somme du dénominateur est faite sur toutes les séquences S de la lattice de décodage générée pour l'énoncé u.

De façon intuitive, maximiser ce critère correspond à maximiser le numérateur (c'est-à-dire augmenter la probabilité que le modèle prédise une séquence similaire à la référence) et à minimiser le dénominateur (diminuer la probabilité des autres séquences). La procédure d'entraînement des modèles TDNNF-HMM chaînes est identique à un entraînement LF-MMI (*Lattice-Free Maximum Mutual Information*) [Veselý 2013], qui consiste à utiliser l'algorithme rétro-progressif pour calculer le numérateur et le dénominateur de la fonction objectif F_{MMI} plutôt que des lattices.

Nous avons conservé des paramètres par défaut dans Kaldi : le taux d'apprentissage de 5e-4 et le taux de régularisation l_2 de 1e-2. Les modèles adulte, enfant et TL ont été entraînés sur 990, 89 et 89 epochs respectivement, avec un temps d'entraînement moyen de 46, 2,1 et 2,1 heures sur un unique GPU GTX 2080 Ti.

3.3.3 Décodage avec le TDNNF-HMM

Les modèles DNN-HMM sont généralement décodés à l'aide d'un graphe de décodage formé d'une combinaison de transducteurs pondérés à états finis (*Weighted Finite State Transducer*, WFST) [Mohri 2008]. Les WFST sont des automates finis dont les transitions entre états sont étiquetées avec un symbole d'entrée, un symbole de sortie et un poids. Un chemin traversant un WFST permet ainsi d'établir la correspondance entre une séquence de symboles d'entrée et une séquence de symboles de sortie. Le graphe de décodage modélise les quatre composants du système [Mohri 2008] :

— Les chaînes de Markov cachées (H) ;

— Le modèle de dépendance au contexte (C) ;

— Le dictionnaire de prononciation, ou lexique (L) ;

— Le modèle de langage, ou grammaire (G).

L'opération « composition » des WFST permet de combiner ces composants modélisant différents niveaux de représentations pour l'obtention d'un graphe de décodage complet, nommé HCLG d'après le nom de ses composants [Mohri 2008].

Chaînes de Markov cachées (HMM)

Chaque phone est modélisé par une chaîne de Markov cachée, généralement à trois états comme sur la figure 3.8a, où a_{ij} est la probabilité de transition de l'état i à l'état j, et $b_j(y_k)$ la probabilité d'émission de l'observation y_k par l'état j. À l'entrée d'un état j, un vecteur acoustique y_k est émit suivant la distribution associée $b_j(y_k)$. La distribution $b_j()$ est définie par le modèle acoustique, qui peut être un GMM ou un DNN (ici, un TDNNF). Les états 0 et 4 du HMM sur la figure 3.8a sont les états initiaux et finaux, non-émetteurs, présents pour simplifier la combinaison des HMM entre eux.

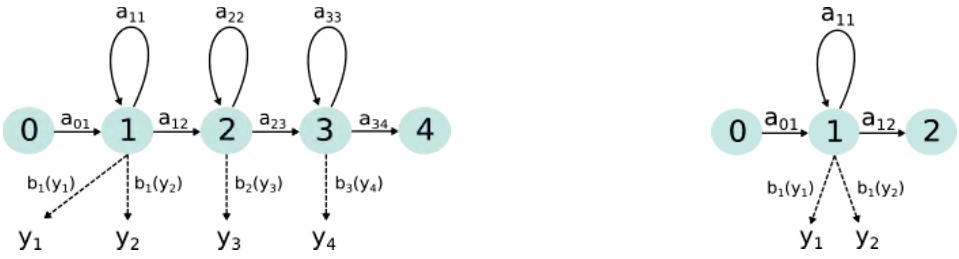


FIGURE 3.8 – Topologies de HMM (a) standard (b) chaîne

Les modèles chaînes ont la particularité d'utiliser un signal sous-échantillonné avec un facteur 3, ce qui réduit la quantité de calculs à effectuer et simplifie le décodage en temps réel. En conséquence, le HMM utilisé dans un modèle chaîne présente une topologie différente des HMM à trois états classique en RAP : il ne comprend qu'un état, afin d'être potentiellement traversé en une seule transition. Les deux topologies de HMM (standard et chaîne) sont

comparées sur la figure 3.8. Puisque le HMM chaîne ne contient qu'un seul état par phone, ses probabilités de transitions sont fixées à l'initialisation et non entraînées, leur rôle étant identique à celui des probabilités de sortie du réseau de neurones. Sur la figure 3.8b, les probabilités pour l'unique état émetteur sont fixées et équiprobables : $a_{11} = a_{12} = 0{,}5$.

Modèle de dépendance au contexte

Ce graphe modélise les dépendances entre trois phones successifs pour favoriser les combinaisons, appelées *triphones*, les plus courantes dans la langue française. Nous n'utilisons pas de modèle de dépendance au contexte car son effet risquerait de gommer des erreurs de lecture des enfants, qui peuvent potentiellement donner des combinaisons de phones rarement ou non existantes en français. Tous nos modèles hybrides, GMM-HMM ou DNN-HMM, sont des systèmes « monophones » et non « triphones » comme la plupart des systèmes de la littérature [Wu 2019, Serizel 2014a, Yeung 2018, Shivakumar 2014].

Dictionnaire de prononciation

Le lexique permet de combiner une suite de phonèmes en un mot existant dans la langue française. Pour notre tâche de reconnaissance de phonème, le dictionnaire de prononciation correspond à l'identité et lie chaque phonème à lui-même. Notre dictionnaire comporte 33 phonèmes :

— Voyelles orales : [a/ɑ, e, ɛ, ə/œ, i, o, ɔ, u, y]
— Voyelles nasales : [ã, ɛ̃/œ̃, ɔ̃]
— Consonnes plosives : [p, t, k, b, d, ɡ]
— Consonnes fricatives : [f, s, ʃ, v, z, ʒ]
— Consonnes nasales : [m, n, ŋ, ɲ]
— Consonnes liquides : [l, ʁ]
— Semi-voyelles : [j, w, ɥ]

Dans cette liste, l'utilisation d'une barre oblique entre deux phonèmes signifie que ces phonèmes ont été fusionnés dans notre vocabulaire, dans les annotations phonétiques et dans les séquences de phonèmes reconnues par le système de RAP. Ces fusions ont plusieurs motivations : premièrement, notre quantité de données limite la correcte modélisation de ces phonèmes acoustiquement proches (proximité renforcée par les fréquences étendues et la variabilité inhérente à la parole d'enfants). Deuxièmement, la confusion entre ces phonèmes proches n'est généralement pas comptée comme une erreur de lecture par les enseignant·e·s, ce qui implique que le système de reconnaissance vocale n'a pas nécessairement besoin de savoir les distinguer.

Modèle de langage

La grammaire modélise l'enchaînement des mots pour former des phrases dans un système de reconnaissance classique. Les modèles les plus utilisés sont les $N-grammes$, qui prédisent

un mot en prenant en compte ses $N-1$ prédécesseurs. Dans notre cas, le modèle de langage est un uni-gramme au niveau du phonème, modélisant l'apparition de chaque phonème individuellement, avec une probabilité calculée sur le corpus d'apprentissage de Lalilo. Utiliser un ordre plus élevé aurait également un effet de lissage des erreurs de lecture des apprenant·e·s lecteur·rice·s, qui ne suivent pas forcément les lois grammaticales de la langue française.

3.4 Bilan

Ce chapitre présentait les diverses méthodes qui seront testées dans le chapitre 4, visant l'établissement de notre modèle acoustique de référence pour la suite de nos travaux. Nous avons tout d'abord exploré les paramètres acoustiques, postulant que les MFCC, couramment utilisés pour le traitement de la parole, ne sont peut-être pas les paramètres les plus adaptés pour la parole d'enfants. Nos recherches ont ciblé deux types de paramètres acoustiques : paramètres fondés sur des techniques de traitement du signal (Filterbank, PLP, RASTA-PLP, Gammatone), et paramètres obtenus grâce à des modèles auto-supervisés pré-entraînés pour l'extraction de paramètres acoustiques (Wav2vec, PASE+). Les premiers, plus classiques, pourraient révéler une meilleure robustesse à la variabilité acoustique de la parole d'enfants et au bruit présent sur les enregistrements. L'étude des seconds nous permettra d'évaluer la performance de ces modèles à l'état de l'art, supposés agnostiques à la tâche, sur de la parole atypique hors domaine d'entraînement. Nous avons enfin détaillé deux techniques de normalisation (CMVN et VLTN) classiquement appliquées sur les paramètres pour améliorer la robustesse du modèle à différents environnements et locuteurs.

Nous avons présenté dans une seconde partie plusieurs techniques choisies pour pallier à notre manque de données de parole d'enfants. L'apprentissage par transfert, consistant à entraîner un premier modèle sur un large corpus de données hors domaine et de l'adapter à nos données cibles, peut être utilisé pour de l'adaptation adulte-enfant lorsque peu de données sont disponibles. Nous planifions d'évaluer trois stratégies de transfert différentes : ré-initialisation de la couche de sortie, gelée de cette dernière, et enfin entraînement du modèle complet. L'augmentation par ajout de bruit de brouhaha nous permettrait, en plus d'augmenter la quantité de données d'entraînement, de présenter au modèle des enregistrements bruités pendant l'apprentissage pour améliorer sa robustesse au bruit.

Enfin, nous avons détaillé l'architecture du modèle qui nous servira de référence tout au long de cette thèse : le modèle TDNNF-HMM. Ce modèle, de type hybride DNN-HMM, est couramment utilisé et reste à l'état de l'art sur de nombreuses tâches malgré sa relative ancienneté. Le TDNNF est particulièrement efficace pour modéliser la parole avec un nombre restreint de paramètres, et notamment avec une quantité de données limitée. Nous avons décrit le processus de décodage grâce à un graphe WFST et précisé la forme et le contenu des quatre éléments (HMM, modèle de dépendance au contexte, lexique et grammaire) de ce graphe, qui sont différents pour de la reconnaissance de phonèmes de ceux classiquement utilisés pour de la reconnaissance de mots.

Chapitre 4

Système de référence : résultats et analyses

Le chapitre précédent détaillait l'architecture, l'entraînement et le décodage d'un modèle acoustique hybride TDNNF-HMM, ainsi que les potentiels paramètres acoustiques d'entrée et les méthodes permettant de pallier au manque de données. Ce chapitre est dédié à l'expérimentation de ces paramètres, modèles et méthodes afin d'aboutir à un système de référence pour la reconnaissance de phonèmes sur parole d'enfants. Pour cela, nous décrivons les données de parole utilisées et la procédure d'acquisition d'enregistrements de parole d'enfants. Les métriques d'évaluation sont présentées, suivies des résultats et analyses de diverses expériences.

Sommaire

4.1 Présentation des données de parole

Nous utilisons deux sources de données de parole française : le corpus adulte Common Voice et des données de parole d'enfants internes à l'entreprise. Le corpus initial de parole d'enfants, utilisé pour les expériences de ce chapitre, est nommé ci-après *Lalil-o-riginel*. Il sera légèrement altéré et rééquilibré grâce à de nouvelles données annotées dans le chapitre 6 pour former le corpus *Lalil-o-fficiel*.

4.1.1 Données de parole d'adultes : Common Voice

Le corpus Common Voice est créé via une plateforme participative en ligne [1], où chacun peut s'enregistrer en train de lire des phrases. Composé d'enregistrements réalisés avec différents équipements dans différents environnements, il est donc particulièrement adapté à notre tâche de lecture en classe. Néanmoins, comme les locuteur·rice·s s'enregistrent généralement seul·e·s dans leur bureau, les données ne contiennent pas de bruit de brouhaha et ont un rapport signal sur bruit (RSB) élevé en moyenne. De plus, chaque enregistrement est validé par deux annotateur·rice·s, toujours avec un système participatif : le corpus ne contient donc qu'une très petite quantité d'erreurs de lecture. Le corpus français a évolué de façon continue durant les trois années de cette thèse : en 2019, seulement 20 heures de paroles étaient enregistrées et validées, puis environ 150 heures début 2020, et depuis fin 2020, plus de 620 heures.

En français, les ensembles Train, Valid et Test que nous avons utilisés pour ces expériences contiennent respectivement 148,9, 2,4 et 7,2 heures de parole (voir Tableau 4.1).

Table 4.1 – Information sur le jeu de données de parole d'adultes Common Voice

Jeu de données	Train	Valid	Test
Durée (h)	148,9	2,4	7,2
Locuteur·rice·s	1276	372	1113
Temps moyen (s)			
Par énoncé	4,1	4,2	4,2
Par locuteur	420,1	23,5	23,4
Nombre d'énoncés	133k	1,8k	5,5k
Nombre de mots	1M	15k	44k
Nombre de phonèmes	11M	54k	161k
RSB (dB)			
Moyenne	34,4	34,3	34,3
Std	14,7	14,5	14,7

Les trois sous-ensembles ne contiennent pas de locuteur·rice·s en commun. Une sélection a été faite sur les enregistrements, en mettant de côté les énoncés pour lesquels le locuteur ou la locutrice avait déclaré un accent (québécois, belge, anglais, américain...). Les participant·e·s au

1. https://commonvoice.mozilla.org/fr

corpus n'étant pas obligé·e·s de s'inscrire pour s'enregistrer, et encore moins de déclarer leur nationalité, l'information de l'accent n'est pas connue pour la plupart des enregistrements. Le corpus peut donc tout de même contenir des enregistrements de parole avec accent. L'ensemble de test a été conçu pour maximiser le nombre de locuteur·rice·s distinct·e·s tout en gardant la même durée moyenne par locuteur·rice que dans l'ensemble de validation.

4.1.2 Données de parole d'enfants : Lalilo

Les données de Lalilo comprennent des enregistrements d'enfants de la grande section au CE2, âgés de 5 à 8 ans, lisant à voix haute des mots isolés et des phrases. Ces deux tâches de lecture sont habituellement assignées par les enseignant·e·s aux lecteur·rice·s débutant·e·s, en fonction de leur niveau de lecture.

Les enregistrements ont été recueillis soit via une première collecte auprès des familles des employés de Lalilo, soit par le biais d'un exercice de lecture orale sur la plateforme Lalilo, soit par mes soins directement dans les écoles. Les protocoles de collecte sont détaillés dans les sections suivantes, ainsi que la procédure d'annotation. La politique de Lalilo sur la collecte de données personnelles est stricte et respecte le RGPD (Règlement Général sur la Protection des Données). Ainsi Lalilo s'engage auprès des enseignants et des parents à ne partager aucune des données collectées [2], ce qui implique que ces données restent privées. Le corpus initial de parole d'enfants utilisé dans ce chapitre, *Lalil-o-riginel*, est ensuite présenté.

4.1.2.1 Première collecte auprès des familles

En 2018, avant le début de cette thèse, a été lancée une première collecte de données auprès des familles des employés (peu nombreux à l'époque) de Lalilo. Quelques histoires (exemples en annexe A) ont été fournies, et il a été laissé le soin aux familles d'enregistrer elles-mêmes leurs enfants en classes de CP, CE1 ou CE2. 23 enregistrements ont été récupérés dans le cadre de cette collecte, pour une durée totale d'environ 63 minutes. Les familles ayant enregistré avec leurs propres appareils, de qualité parfois médiocre, certains enregistrements contiennent de la saturation et du bruit blanc. Les enregistrements ne contiennent néanmoins pas de bruit de salle de classe, puisqu'ils ont généralement été effectués à la maison.

4.1.2.2 Collecte de données via la plateforme Lalilo

La plateforme Lalilo, utilisable gratuitement depuis 2018 par les enseignant·e·s, inclue un exercice de lecture orale, où l'enfant doit lire à voix haute le mot ou la phrase affichée sur son écran. Si cet exercice constitue aujourd'hui notre source majeure de données de parole, au début de la thèse, en 2019, le nombre d'utilisateur·rice·s était petit, le contenu pédagogique était peu varié et l'exercice marchait mal, conditions qui ne permettaient donc pas de récupérer beaucoup d'enregistrements de bonne qualité.

2. https://ressources.lalilo.com/privacy-fr.pdf

La plateforme est faite pour que les enseignant·e·s puissent laisser un petit groupe d'élèves jouer sur la plateforme en autonomie, avec une supervision réduite. Cela implique inévitablement la présence, à fréquence très variable, d'évènements sonores typiques des salles de classe, comme de la parole superposée, des raclements de chaises ou des claquements de portes sur les enregistrements. De plus, les écoles ne bénéficiant pas toujours de casques-micro, les microphones intégrés des ordinateurs sont parfois utilisés, enregistrant la majeure partie de ces bruits ambiants. Contrairement aux données de la collecte, enregistrées dans des conditions de bruit faible et avec un micro-casque de qualité, les données de la plateforme peuvent ainsi être considérablement bruitées. Enfin, certains enregistrements ne sont pas exploitables à cause d'un problème technique, lorsqu'un enfant ne comprend pas qu'il doit lire oralement, ou encore lorsqu'un·e professeur·e essaye l'exercice de lecture orale.

4.1.2.3 Collecte de données en école

Au début de cette thèse, les données de parole d'enfants enregistrées et annotées par Lalilo étaient très limitées : seulement trois heures dans le jeu d'entraînement, et une dizaine de minutes dans le jeu de test. Ces données contenaient les histoires de la première collecte auprès des familles, et des mots isolés provenant de la plateforme Lalilo. Une collecte de données en école a donc été organisée, dans le cadre légal du projet de thèse avec le laboratoire IRIT. Les documents officiels mis en place pour la collecte sont consultables en annexe B : autorisation d'enregistrement de la voix à faire signer à l'enfant et aux parents, et fiches d'explications à l'attention des enseignant·e·s et des parents.

La création du contenu a été faite avec l'aide de l'équipe pédagogique de Lalilo. Afin d'obtenir une base de données complète, permettant de travailler sur plusieurs tâches de RAP dans un futur plus ou moins proche (détection d'erreurs de lecture sur différents contenus, évaluation de la fluence...), nous avons sélectionné des mots, phrases et histoires courtes dans le contenu délivré par la plateforme Lalilo aux enfants. Le contenu est divisé en 4 à 6 niveaux afin de pouvoir adapter la difficulté à chaque enfant. En effet, le contenu proposé à un enfant ne doit pas être trop facile pour qu'il soit stimulé intellectuellement et que l'on obtienne des erreurs de lecture et de fluence dans les enregistrements, mais ne doit pas être trop difficile pour ne pas le frustrer et lui faire perdre confiance. Les tableaux 4.2 et 4.3 décrivent les différents niveaux de contenu pour les mots et phrases/histoires.

Le déroulement d'une séance d'enregistrement consistait à passer une dizaine de minutes avec chaque enfant (ayant rempli l'autorisation) individuellement. L'enfant était installé dans la mesure du possible dans une pièce à part (mais obligatoirement proche de la salle de classe, avec la porte ouverte) sur un bureau individuel, avec le texte imprimé sur papier face à el·lui, et un casque-micro relié à un ordinateur. La séance commençait par une série de questions visant à connaître le nom, prénom, l'âge de l'enfant, ainsi que sa langue maternelle et dans le cas échéant, les différentes langues parlées en fonction de l'environnement (maison, école...). L'enfant se voyait ensuite expliquer de façon pédagogique le fonctionnement d'un casque-micro, la façon de le positionner, et le principe d'enregistrement de la voix via un ordinateur. En fonction d'indications potentielles données par l'enseignant·e sur le niveau de l'enfant, une

Table 4.2 – Explication détaillées des niveaux de contenu pour les mots isolés

Niveau	Description du contenu	Exemple
1	Pseudo-mots uni-syllabiques facilement déchiffrables	pa, li
2	Pseudo-mots bi-syllabiques Mots uni-syllabiques facilement déchiffrables Mots outils simples	icha, obi roi, mer elle, que
3	Mots uni-syllabiques Mots bi-syllabiques facilement déchiffrables Mots tri-syllabiques facilement déchiffrables	chien, ville mouton, café caméra, tartelette
4	Mots uni-syllabiques avec irrégularités Mots bi-syllabiques avec irrégularités Mots tri-syllabiques et plus	nerf, sœur tournis, sueur multiplication, chatouiller

Table 4.3 – Explication détaillées des niveaux de contenu pour les phrases et histoires

Niveau	Description du contenu
1	Phrases de 4-5 mots connus et facilement déchiffrables Histoires très courtes avec mots connus et phonèmes de début d'apprentissage Lettres muettes grisées moins visibles pour aider l'enfant *Exemple : Elle joue au ballon.*
2	Histoires courtes avec mots connus et phrases répétant la même structure Mise en page séparant les phrases les unes des autres *Exemple :* *J'aime les bananes mais j'adore les ananas.* *J'aime les fruits mais j'adore les frites.*
3	Histoires courtes moins structurées, avec vocabulaire plus difficile Mise en page séparant les phrases les unes des autres *Exemple :* *Le cheval hennit dans le pré.* *Le chat miaule sur le lit.*
4, 5, 6	Histoires avec longueur et difficulté croissantes Une phrase par ligne, puis deux, puis petits paragraphes Introduction de la ponctuation : guillemets, tirets...

série de 20 mots était choisie à un niveau correspondant. L'enfant devait ensuite lire une seconde série de 20 mots (plus ou moins difficile que la première, en fonction de la qualité de sa lecture), puis deux histoires : une à son niveau et une légèrement plus difficile. L'enfant était libre de poser des questions entre les séries, ou d'arrêter la séance à tout moment.

J'ai effectué la collecte en me rendant dans quatre écoles de l'agglomération toulousaine, auprès de trois enseignantes de CP et quatre enseignantes de CE1, pour un total de 58 élèves de CP et 76 élèves de CE1. Les informations de cette collecte sont affichées dans le tableau 4.4 : j'ai récolté approximativement 90 minutes de mots isolés et 65 minutes de phrases/histoires d'enfants en CP, et 130 minutes de mots isolés et 140 minutes de phrases/histoires d'enfants en CE1, pour un total de 425 minutes d'enregistrement, soit environ 7,1 heures.

Métrique	CP	CE1
Nombre d'élèves	58	76
Quantité de mots isolés (min)	90	130
Quantité de phrases/histoires (min)	65	140
Quantité totale (heures)	2,6	4,5

Table 4.4 – Informations sur les données collectées en école

4.1.2.4 Annotations des données

Transcrire phonétiquement avec exactitude ce qui a été lu par des enfants non-lecteur·rice·s est une tâche difficile. Leur jeune âge implique des difficultés à articuler, de par leurs mécanismes faciaux en cours de développement, et à utiliser le langage : présence d'hésitations, de bégaiements, ou encore production de sons n'existant pas dans la langue française. Leur apprentissage de la lecture cause de plus la présence d'erreurs de lectures, d'hésitations ou de commentaires hors de propos.

Bien qu'il ne s'agisse pas d'une tâche impossible, l'annotation au niveau du phonème prend beaucoup de temps et est très « coûteuse », et nous avons choisi de ne pas investir dans des transcriptions phonétiques pour l'ensemble du corpus pour le moment. Par conséquent, seuls les énoncés correctement prononcés et lus avec un certain degré de fluidité, identiques au texte demandé, ont été inclus dans les données d'entraînement et de validation des modèles acoustiques et linguistiques. Seules les données de test, contenant des mots et des phrases avec des erreurs de lecture, ont été transcrites phonétiquement pour permettre le calcul des scores de performance et évaluer la reconnaissance sur des erreurs de lecture. Les transcriptions ont été effectuées manuellement par deux juges humain·e·s, et les énoncés ont été mis de côté en cas de désaccord.

La procédure de transcription suivie par les annotateur·rice·s consiste d'abord à écouter chaque enregistrement et en rejeter certains, suivant les catégories suivantes :

— « Problème technique » : le son n'a pas été enregistré à cause d'un problème technique ;

— « Sans parole » : l'enfant ne parle pas, mais l'enregistrement peut contenir du bruit ;

— « Chuchotement » : l'enfant chuchote (souvent par timidité) ;

— « Adulte » : un adulte lit à la place d'un enfant. Cela arrive lorsque les professeur·e·s testent la plateforme avant de l'utiliser, ou qu'un élève en difficulté utilise la plateforme avec l'aide d'un·e éducateur·rice spécialisé·e ;

— « Commentaire » : l'enfant ne lit pas ce qu'il doit lire, mais parle, chante, crie...

Sur chaque enregistrement non rejeté, l'annotateur·rice se voit proposer le texte que l'enfant devait lire, et doit modifier la transcription pour qu'elle corresponde exactement à ce que l'enfant a lu :

— Enlever les mots non lus
Le chat est gris → *Le chat gris* ;

— Ajouter les mots additionnels lus (pas les commentaires !)
Le chat est gris → *Le chat le beau chat est gris* ;

— Indiquer la prononciation exacte en alphabet phonétique[3] en cas d'erreur de lecture (confusion de phonèmes, faux départ...)
Le chat est gris où *gris* est lu *gros* → *Le chat est [G R o]:gris* ;

— Indiquer la prononciation et la position du ou des silences en cas d'hésitation intra-mot
Le chat est gris avec une hésitation dans *gris* → *Le chat est [G R sil i]:gris* ;

— Ajouter les commentaires faits par l'enfant cible en tant que mots hors lecture grâce au symbole #. Si le mot n'existe pas, l'annotateur doit indiquer au mieux la prononciation
Attends pfff → *attends:# [P F]:#* ;

— Indiquer un accent régional modifiant la prononciation d'un mot avec le suffixe _a
La fée rose → *la fée rose_a* ;

— Indiquer la prononciation d'un e muet. Si cela résulte en une erreur de lecture, la prononciation est indiquée comme pour une erreur classique. Sinon, on l'indique avec le suffixe _e.
La fée rose (où lire le e de *fée* est considéré comme une erreur, mais celui de *rose* est accepté) → *la [F e AE]:fée rose_e* ;

— Indiquer la liaison entre deux mots grâce au suffixe _l sur le premier mot
Le chat est mon ami → *le chat est mon_l ami*.

Les enregistrements et leur transcription sont ensuite enregistrés en base de données, où chaque mot reçoit automatiquement une catégorie en fonction du texte de référence et de la transcription exacte parmi les catégories du tableau 4.5. La gestion de nos jeux d'enregistrements et de leurs transcriptions se fait grâce à un fichier json dont la forme est détaillée en annexe C.

Une première phase d'annotation a été effectuée à l'été 2019, visant uniquement des mots isolés, car l'exercice de lecture de la plateforme ne présentait que ce type de contenu à l'époque. Des mots isolés de la plateforme et de la collecte ont donc été soit classés « correct » ou « non correct » pour servir à l'entraînement des modèles acoustiques, soit transcrits phonétiquement pour tester la reconnaissance automatique de phonèmes. Une deuxième phase d'annotation, à

3. Nous utilisons un alphabet phonétique non conventionnel. Voici les correspondances en alphabet IPA pour les symboles donnés en exemple : G = [g], R = [ʁ], o = [o], i = [i], P = [p], F = [f], e = [e], AE = [ə/œ]

Table 4.5 – Catégories pour les mots d'un enregistrement

Catégorie	Description
OK	Correctement lu
MISPRON_1	Contient une erreur de lecture sur un seul phonème
MISPRON_2	Contient une erreur de lecture sur plusieurs phonèmes
SUB	Remplacé par un autre mot existant
FS	Est un faux départ (l'enfant n'a pas lu le mot en entier)
HESIT	Contient une hésitation
COMMENT	Est un commentaire de l'enfant
SKIP	Non lu par l'enfant

l'été 2020, a permis d'obtenir plus de phrases pour l'entraînement et le test. Une troisième phase (été 2021) a été précédée d'une phase de sélection des données (de façon automatique à partir du niveau de lecture global des élèves sur la plateforme Lalilo) afin de choisir des énoncés contenant des erreurs de lecture.

4.1.2.5 Analyse détaillée des données

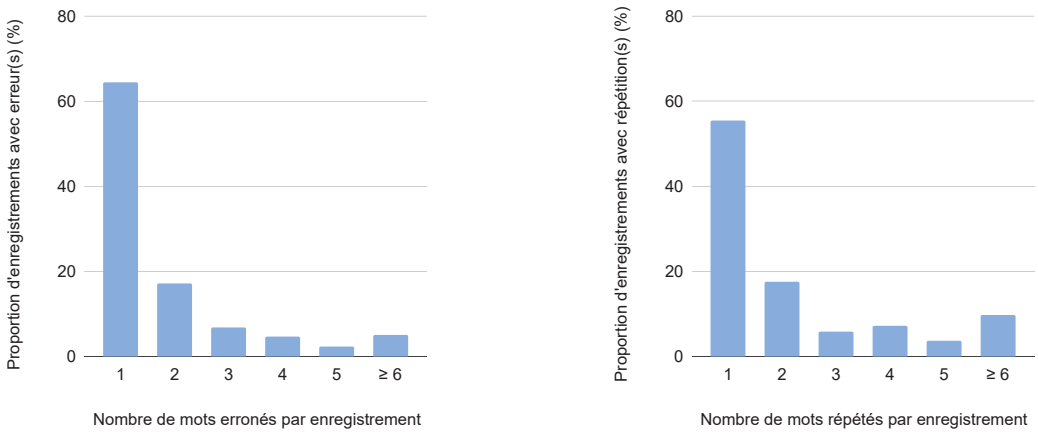
Dans le cadre de la deuxième phase d'annotation, nous avons prélevé un échantillon d'enregistrements de phrases, récoltés sur la plateforme Lalilo pendant le mois de février 2020, et l'avons annoté selon le protocole détaillé en section précédente. Nous présentons ici une analyse détaillée des erreurs de lecture rencontrées dans ces enregistrements. Cet échantillon aléatoire est représentatif de notre application, et nous permet d'établir des statistiques fiables sur les proportions des différents types d'erreur de lecture. Les résultats de cette analyse doivent être replacés dans leur contexte temporel : le mois de février se trouvant au milieu de l'année scolaire, les proportions d'erreurs de lecture devraient être plus faibles qu'en septembre 2019, mais plus élevées qu'en juin 2020. Ce phénomène est cependant probablement nivelé par l'utilisation du système d'apprentissage adaptatif, qui choisit les exercices de façon à ce qu'ils tombent dans la zone proximale de développement de l'enfant.

Dans nos annotations, dont le format est présenté en annexe C, les mots sont classés selon les catégories du tableau 4.5. Nous utilisons ces annotations pour calculer différentes métriques. La métrique principale est le taux d'occurrence pour chaque catégorie d'erreur, calculé comme le ratio entre le nombre de mots appartenant à cette catégorie et le nombre total de mots à lire. Le tableau 4.6 présente les taux d'occurrence des mots sans erreur (correspondant à la catégorie « OK » du tableau 4.5), des mots supprimés (catégorie « SKIP »), des mots substitués (contenant une ou plusieurs substitutions de phonèmes, étant substitué complètement par un autre mot, ou étant un faux départ du mot à lire) et des mots répétés (qu'ils contiennent ou non une erreur de lecture). Il est important de noter que les mots répétés ne correspondent pas à une catégorie d'annotation du tableau 4.5 et sont également comptés dans les « sans erreur » et les « substitutions ». C'est pourquoi il ne faut pas les inclure dans la somme des pourcentages pour obtenir 100% des mots. Nous sélectionnons de plus les enregistrements

contenant au moins un mot erroné (présentant n'importe quelle erreur de lecture), et mesurons le nombre de mots erronés par enregistrement. De la même façon, nous comptons le nombre de mots répétés sur les enregistrements ayant au moins un mot répété. Ces métriques sont représentées sur les histogrammes 4.1a et 4.1b.

TABLE 4.6 – Taux d'occurrence (%) pour chaque type d'erreur

Type d'erreur	Catégorie(s) d'annotation	Taux d'occurrence
Sans erreur	OK	94,3
Suppression	SKIP	1,3
Substitution	MISPRON_1, SUB MISPRON_2, FS	4,4
dont mot résultant existant		*2,2*
Répétition	Toutes catégories	3,5
dont sans erreur	*OK*	*2,8*



(a) (b)

FIGURE 4.1 – Histogrammes des nombres de (a) mots erronés (tous, y compris répétés) et de (b) mots répétés dans les enregistrements contenant au moins (a) un mot erroné ou (b) un mot répété

Nous constatons tout d'abord dans le tableau 4.6 que la majorité (94,3%) des mots ne contiennent pas d'erreur de lecture. L'histogramme 4.1a montre de plus que plus de 60% des énoncés contenant au moins un mot erroné n'en contienne qu'un seul. Cela est dû au fonctionnement du système d'apprentissage adaptatif, qui choisit pour les élèves des exercices ni trop faciles, ni trop difficiles : les élèves recevant des phrases à lire oralement ont généralement un niveau relativement élevé, et font peu d'erreurs de lecture. Nous voyons ensuite qu'une faible proportion de mots sont entièrement supprimés (1,3%). Par une analyse des enregistrements contenant des suppressions de mots, nous observons deux cas distincts : dans le cas d'un·e bon·ne lecteur·rice, le(s) mot(s) supprimé(s) se trouvent en milieu de phrase car il·elle lit trop vite et saute un mot. Dans le cas d'un·e lecteur·rice avec plus de difficultés, ils se situent plutôt en fin de phrase, lorsque l'enfant considère la lecture trop difficile et abandonne.

Les substitutions représentent une part plus importante (4,4%), et la moitié des substitutions résultent en un mot existant dans la langue française. L'autre moitié correspond aux mots dans lesquels des phonèmes ont été substitués, insérés ou supprimés, formant des combinaisons parfois insolites. Afin de mieux comprendre le type de substitutions que font les enfants, nous proposons quelques exemples de substitution de mots ou de phonèmes, tirés de notre échantillon de phrases :

— Substitution de temps : l'enfant modifie le temps du verbe de la phrase. Souvent l'enfant choisit un temps plus facile, déjà assimilé, ou contenant moins de mots
Exemple : « je vais m'asseoir sur le banc » *est lu* « je m'assoie sur le banc » ;

— Substitution de champ lexical : l'enfant substitue un mot par un mot du même champ lexical, même si le mot ne ressemble pas du tout. Ce phénomène est exacerbé par le contenu pédagogique de Lalilo, qui contient des phrases identiques à l'exception d'un mot, et que l'enfant peut avoir rencontré dans un précédent exercice
Exemple : « elle est dans le champ » *est lu* « elle est dans la forêt » ;

— Substitution par mot connu : l'enfant lit trop vite en essayant de deviner, et substitue un mot inconnu par un mot connu ressemblant
Exemple : « c'est un champion » *est lu* « c'est un champignon » ;

— Lecture de lettre(s) muette(s) : l'enfant lit à tort une ou des lettres muettes, apparaissant souvent en fin de mot dans la langue française
Exemple : « Léo et Léa rient [ʁi] » *est lu* « Léo et Léa ri**ent** [ʁiɑ̃t] » ;

— Substitution de phonème : l'enfant substitue une voyelle ou une consonne de la phrase par une autre. Cela arrive souvent avec les graphèmes qui correspondent à un phonème différent en fonction des graphèmes environnants, ou avec les graphèmes avec accents
Exemple : « il faut des gouttes [gut] de sirop » *est lu* « il faut des [jut] de sirop »
Exemple : « il mange la salade [saladə] » *est lu* « il mange la [salade] » ;

— Inversion de phonèmes : l'enfant inverse deux phonèmes, souvent dans les mots n'en contenant que deux, et souvent lorsque les deux graphèmes se ressemblent
Exemple : « il [il] a une moto » *est lu* « [li] a une moto » ;

— Suppression de phonème : l'enfant supprime un phonème. Cela arrive notamment avec les groupes de consonnes, qui sont difficiles à lire
Exemple : « je déte**st**e [detɛstə] la chaleur » *est lu* « je [detɛtə] la chaleur » ;

— Insertion de phonème : l'enfant insère un phonème, soit parce qu'il·elle lit trop vite et insère un graphème environnant, soit parce qu'il·elle n'arrive pas à appréhender un graphème ou groupe de graphèmes
Exemple : « il a rit avec [a ʁi avɛk] elle » *est lu* « il arrive avec [aʁiv avɛk] elle »
Exemple : « je vais dehors [dəɔʁ] » *est lu* « je vais [dəʒɔʁ] ».

Enfin, les répétitions représentent également une part non négligeable (3,5%), et 80% des mots répétés ne contiennent pas d'erreur de lecture au niveau du phonème. De la même façon que pour les mots erronés, nous observons sur l'histogramme 4.1b que plus de la moitié des phrases contenant au moins un mot répété n'en contiennent qu'un seul. Une proportion non négligeable (9,8%) contient cependant six mots répétés ou plus. Après analyse, cinq cas ressortent :

(1) l'enfant est vraiment en difficulté et répète plusieurs mots individuellement en butant dessus

(2) l'enfant commence à lire, fait une erreur, s'en rend compte, et recommence au début

(3) l'enfant lit une première fois la phrase de façon hésitante, puis la relit entièrement de façon plus confiante

(4) l'enfant comprend à la fin de la phrase ce qu'il·elle a lu et recommence avec une meilleure fluidité

(5) l'enfant ne comprend pas le fonctionnement de l'exercice, et répète la phrase entière alors qu'il·elle l'avait correctement lu, pensant que sa lecture n'a pas été enregistrée.

Dans les deux premiers cas notamment, les mots répétés contiennent souvent des erreurs de lecture. Nous identifions deux types de répétitions :

— Répétition individuelle (cas 1) : un ou plusieurs mots sont répétés individuellement, avec erreur de déchiffrage ou non
Exemple : « il lit sur une île » *est lu* « il lit sur une li... île » ;

— Répétition par motif (cas 2 à 5) : l'enfant répète un motif de plusieurs mots à la suite, parfois la phrase entière
Exemple : « il a une hache » *est lu* « il a une hache il a une hache »
Exemple : « elle lui dira de lire un livre » *est lu* « elle lui dira de livre... de lire un livre ».

Ces analyses seront utiles dans le chapitre 7 pour le développement d'une méthode de simulation d'erreurs de lecture dans le but d'améliorer la robustesse des systèmes face à la parole d'apprenant·e·s lecteur·rice·s.

4.1.2.6 Le corpus initial de parole d'enfants : *Lalil-o-riginel*

Au commencement de cette thèse n'étaient disponibles que trois heures de parole d'enfants. Le premier véritable ensemble de données de parole d'enfants a été créé après l'été 2019, et sera désigné ci-après par le nom *Lalil-o-riginel*. L'exercice de la plateforme Lalilo ne proposant à l'époque que la lecture de mots isolés, le jeu de test ne comporte que ce type d'enregistrements. Le jeu d'entraînement contient en outre quelques-unes des histoires courtes récoltées et annotées avant le début de la thèse, lors de la première collecte auprès des familles.

Le tableau 4.7 présente des informations générales sur le corpus *Lalil-o-riginel*. Le jeu d'entraînement contient 13,2 heures, avec plus de 26 000 mots et 99 000 phonèmes. Le jeu de test est formé de 1172 mots (4116 phonèmes), pour environ 50 minutes de parole. Le nombre de locuteur·rice·s dans le jeu d'entraînement est presque deux fois plus élevé que celui de Common Voice (1276), pour un nombre d'heures 10 fois plus petit, ce qui implique un temps moyen de parole par locuteur beaucoup plus faible (19,6 comparé à 420,1). Le temps de parole par locuteur·rice est encore plus court dans le jeu de test puisqu'il ne contient que des mots isolés et que le nombre de locuteur·rice·s est également élevé. Ce grand nombre de locuteur·rice·s est dû à l'utilisation pratique de Lalilo : un grand nombre d'enfants travaillent très peu de temps sur l'exercice à voix haute. Nous observons également que les valeurs de RSB moyen des jeux de données d'enfant (25,9 dB pour le train, 22,2 dB pour le test) sont significativement plus

faibles que celle de Common Voice (environ 34 dB), à cause de l'utilisation de la plateforme Lalilo en salle de classe.

Table 4.7 – Information sur le corpus *Lalil-o-riginel* de parole d'enfants

Jeu de données	Train	Test M
Durée (h)	13,2	0,85
Locuteurs	2307	425
Temps moyen (s)		
Par énoncé	2,6	2,6
Par locuteur	19,6	7,0
Nombre d'énoncés	22k	1172
Nombre de mots	26k	1172
Nombre de phonèmes	99k	4,1k
RSB (dB)		
Moyenne	25,9	22,2
Std	14,2	13,2

4.2 Métriques d'évaluation

La performance de nos systèmes de reconnaissance de phonème est mesurée grâce à un taux d'erreur phonème (*Phoneme Error Rate*, PER), obtenu en comparant la séquence de phonèmes prédite par le modèle à la séquence de phonèmes effectivement prononcée par l'enfant. Il est calculé à partir du nombre de phonèmes substitués, insérés et supprimés par le modèle :

$$\text{PER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Suppression}}{\text{Nombre de phonèmes de référence}} \tag{4.1}$$

Dans les sections et chapitres suivants, nous pourrons mesurer la performance en termes de PER soit sur les jeux de test globaux, soit sur diverses catégories de niveau de bruit, de qualité de lecture ou d'âge afin d'étudier un effet spécifique d'une méthode sur ces aspects.

Nous introduisons également la notion de PER-Oracle, qui représente le meilleur PER atteignable par la combinaison de deux modèles, et sert ainsi à mesurer la complémentarité de ces modèles. Il est calculé en comparant la sortie de chaque modèle avec la référence : pour chaque phonème prononcé par l'enfant, si au moins l'un des deux modèles prédit le bon phonème, alors nous considérons la prédiction correcte. Une erreur (substitution, insertion ou suppression) n'est donc comptée que si les deux modèles l'ont faite. Cette méthode permet également de calculer le pourcentage de différence de deux modèles dans leurs réponses correctes et incorrectes, indicateurs de leur complémentarité.

4.3 Expériences et évaluations

Cette section présente les expériences liées aux techniques détaillées dans le chapitre précédent et effectuées sur le corpus *Lalil-o-riginel*. Ces expériences serviront à établir notre modèle de référence pour la suite de nos travaux.

4.3.1 Comparaison des modèles TDNN-HMM et TDNNF-HMM

Afin de valider le choix d'un TDNNF-HMM comme système de référence plutôt que son aîné le TDNN-HMM, nous établissons une rapide comparaison entre deux modèles équivalents au niveau de de l'entraînement (modèles chaînes avec une fonction objectif LF-MMI, entraînés sur le corpus *Lalil-o-riginel* uniquement) et du décodage (graphe HCLG uni-gramme). Nous nous appuyons sur les recettes Kaldi pour le corpus WSJ [Garofolo 1993a] : les structures diffèrent légèrement entre l'architecture TDNN [4] (8 couches de dimension 448) et TDNNF (12 couches de dimension 1024) [5]. Les paramètres acoustiques sont des MFCC extraites sur des fenêtres de 25 ms décalées de 10 ms. Nous appliquons la CMN uniquement lors de l'entraînement du GMM-HMM et de l'alignement des données, comme dans les recettes Kaldi utilisées. Alors que la grande majorité des recettes ajoutent des i-vecteurs aux paramètres acoustiques, nous avons fait le choix de ne pas en utiliser. Des études avec un TDNN-HMM sur un corpus préliminaire nous ont permis de :

— vérifier que cette configuration de CMN était la plus efficace. L'application de CMN est effectivement bénéfique sur le GMM-HMM et pour la génération des alignements sur lesquels est entraîné le TDNN-HMM (amélioration relative du PER de 32,6%) mais dégrade légèrement les performances lorsqu'appliquée lors de l'entraînement du TDNN-HMM (dégradation relative de 6,6%).

— observer que l'utilisation des i-vecteurs était néfaste lorsque la quantité de parole par locuteur·rice était trop faible : ils apportent une dégradation relative du PER de 11,5% ;

Le tableau 4.8 dévoile la comparaison des scores de PER de ces deux modèles. Nous voyons distinctement que le TDNNF-HMM, en plus d'une plus grande efficience de calcul par l'utilisation d'un nombre de paramètres inférieur, démontre de meilleures performances pour la reconnaissance de parole d'enfants avec peu de données.

Table 4.8 – Comparaison entre les modèles TDNN-HMM et TDNNF-HMM : nombre de paramètres (en millions) et PER (%)

Modèle	Nombre de paramètres	PER
TDNN-HMM	12,5	36,7
TDNNF-HMM	7,6	32,6

4. https://colibris.link/script-TDNN-WSJ
5. https://colibris.link/script-TDNNF-WSJ

4.3.2 Choix des paramètres audio

Les systèmes de reconnaissance de la parole fondés sur des approches hybrides DNN-HMM utilisent habituellement des coefficients MFCC pour extraire l'information pertinente des signaux audio. Nous étudions ici des paramètres alternatifs, présentés dans le chapitre 3, afin de déterminer si certains sont mieux adaptés pour les spécificités de la parole d'enfants. Des études ont porté il y a quelques années sur l'évaluation de paramètres fondés sur du traitement du signal et de leurs configurations pour la reconnaissance de parole d'enfants, mais avec des modèles GMM-HMM [Shivakumar 2014, Li 2001] ou des modèles DNN-HMM entraînés sur parole d'adultes [Shahnawazuddin 2016]. Le tableau 4.9 présente les scores PER obtenus avec le TDNNF-HMM sur le corpus *Lalil-o-riginel* avec ces différents paramètres alternatifs, fondés sur des techniques de traitement du signal (TS) ou d'extraction via un modèle auto-supervisé (EAS). Le score du modèle TDNNF-HMM entraîné sur des MFCC (32,1%) est légèrement différent de celui affiché dans le tableau 4.8 (32,6%). Les résultats de cette section proviennent de modèles entraînés sur une machine différente, ce qui explique la légère variabilité de ce score.

Table 4.9 – PER (%) avec le TDNNF-HMM entraîné et testé sur le jeu de données d'enfant *Lalil-o-riginel* pour différents paramètres audio TS et EAS

Méthode	Paramètres audio	PER
TS	MFCC *(référence)*	**32,1**
	PLP	**32,1**
	RASTA-PLP	35,6
	Fbank	
	alignements avec Fbank	66,7
	alignements avec MFCC	32,3
	Gammatones	40,7
EAS	Wav2vec	34,9
	PASE+	33,8

4.3.2.1 Paramètres fondés sur du traitement du signal (TS)

Tous les paramètres sont calculés sur des fenêtres de 25 ms décalées de 10 ms. Les MFCC, PLP, RASTA-PLP et Fbank sont de dimension 40, et les Gammatones de dimension 80. Les PLP et RASTA-PLP sont calculés avec une prédiction linéaire d'ordre 40. Pour les MFCC, PLP, RASTA-PLP et Gammatones, le TDNNF-HMM est entraîné sur des alignements générés par un GMM-HMM entraîné sur les mêmes paramètres, ce qui montre que ces paramètres sont adaptés aux deux sortes de modèles. Pour les Fbank, en revanche, utiliser des alignements d'un GMM-HMM entraîné sur des Fbank donne un PER très haut de 66,7%. Cette mauvaise performance fait écho à plusieurs études [Abka 2015, Yuliani 2017, Shivakumar 2014] où les performances obtenues par des GMM-HMM entraînés sur des Fbank sont significativement en dessous de celles obtenues avec d'autres paramètres (MFCC, PLP...). Cela pourrait signifier que les

GMM-HMM n'arrivent pas à modéliser ces paramètres qui contiennent une grande redondance d'information, contrairement aux MFCC auxquels une transformée en cosinus discrète est appliquée pour les décorréler [Yuliani 2017]. Dans [Abka 2015], les auteurs observent également une forte dégradation de la performance des modèles entraînés sur des Fbank dès que le RSB descend en dessous de 20 dB, ce qui est le cas pour une grande partie de nos enregistrements (RSB moyen de 22,2 dB pour l'ensemble de test, voir tableau 4.7). Cette dégradation est encore une fois due à la corrélation des paramètres : le bruit augmente le niveau de corrélation et modifie la distribution des spectres de parole, qui deviennent difficilement modélisables par des ensembles de gaussiennes. Utiliser des alignements générés par un GMM-HMM entraîné sur des MFCC permet d'obtenir un bien meilleur score PER (32,3%), laissant penser que seul le GMM-HMM est affecté par cette redondance d'information.

Parmi les paramètres issus du TS, les PLP et Fbank obtiennent un score très proche ou égal à celui des paramètres de référence, les MFCC. Ce résultat diffère de [Shivakumar 2014], où un GMM-HMM entraîné sur des PLP obtient un WER 5% plus haut que lorsqu'il est entraîné sur des MFCC. Les Fbank, dont la seule différence avec les MFCC réside dans l'application d'une transformée en cosinus discrète visant à réduire la redondance de l'information, sont donc aussi adaptés que les MFCC au modèle TDNNF-HMM, qui, contrairement au GMM-HMM, semble gérer cette redondance. Les RASTA-PLP obtiennent un score significativement moins bon que celui des PLP : le filtrage RASTA, censé améliorer la robustesse des paramètres à l'environnement sonore, est donc néfaste sur nos enregistrements de parole d'enfants. Cet effet est potentiellement dû à une suppression excessive de composantes fréquentielles, détériorant la qualité du spectre de parole d'enfants. Les coefficients Gammatones, qui diffèrent des MFCC uniquement par le type de filtrage, se révèlent beaucoup moins efficaces.

Des combinaisons de paramètres pourraient être envisagées afin d'exploiter les avantages de chaque paramètre [Zolnay 2005, Schluter 2007]. Pour avoir une idée de la complémentarité des différents paramètres, nous calculons le PER-Oracle de chaque combinaison de modèles (MFCC, Fbank, PLP, RASTA-PLP et Gammatone), ainsi que leurs pourcentages de différence dans leurs réponses correctes et incorrectes, dans le tableau 4.10. Les paramètres les plus complémentaires sont les Fbank et Gammatones : leurs réponses correctes diffèrent à 19,9%, et leurs erreurs à 32,0%. Ensemble, ils obtiennent un PER-Oracle de 28,5%, ce qui correspond à une amélioration relative de 12,0% et 31,1% par rapport aux modèles Fbank et Gammatone séparément. Le meilleur PER-Oracle est de 26,2% et est obtenu par la combinaison des modèles MFCC et PLP, ce qui correspond à une amélioration relative de 16,0% par rapport à chacun des modèles individuels. Aux vues des grandes différences dans leurs techniques d'extraction (figures 3.1 et 3.2 du chapitre précédent), la combinaison de ces paramètres en entrée du système de reconnaissance de phonèmes pourrait potentiellement améliorer la performance. Une étude plus poussée serait cependant nécessaire pour déterminer la façon de les combiner afin d'éviter une trop grande redondance des informations.

Table 4.10 – PER-Oracle (%) entre les différents paramètres TS

Paramètres	MFCC	Fbank	PLP	RASTA-PLP	Gammatones
MFCC	32,1				
Fbank	27,3	32,3			
RASTA-PLP	26,8	26,9	26,6	35,6	
Gammatones	28,3	28,5	28,6	29,5	40,7

4.3.2.2 Paramètres extraits avec des modèles auto-supervisés (EAS)

Les paramètres EAS se révèlent également moins performants que les MFCC, avec des scores PER de 1,7% à 2,8% plus élevés. Plusieurs pistes pourraient conduire à de meilleurs résultats via ces techniques prometteuses. Tout d'abord les modèles Wav2vec et PASE+ utilisés pour extraire ces paramètres ont été entraînés sur de la parole en langue anglaise, et non française [Schneider 2019, Ravanelli 2020]. Certaines caractéristiques du français, et notamment certains phonèmes comme les voyelles nasales, ne se retrouvent pas dans la parole anglaise. Malgré leur objectif d'extraction de paramètres agnostiques de la tâche, la performance de ces modèles pourrait être variable en fonction de la langue, en particulier pour une tâche de reconnaissance de la parole. De la même façon, les modèles ont été entraînés uniquement sur de la parole d'adultes, et pourraient ainsi négliger certaines caractéristiques acoustiques des voix d'enfant.

Suite à nos différentes expériences, les MFCC semblent rester les meilleurs paramètres pour un modèle hybride DNN-HMM sur notre tâche de reconnaissance de phonèmes dans la parole d'enfants. Tous nos modèles TDNNF-HMM seront donc entraînés sur des MFCC de dimension 40, avec des fenêtre de 25 ms et un décalage de 10 ms.

4.3.3 Évaluation de la méthode d'apprentissage par transfert

Nous étudions à présent l'efficacité de l'apprentissage par transfert sur un modèle TDNNF-HMM, étant données nos 13,2 heures de parole d'enfants. Nous utilisons un modèle entraîné sur le corpus Common Voice de parole d'adultes, qui obtient un PER de 23,5% sur le jeu de test correspondant. Ce modèle obtient cependant un PER de 49,7% sur le jeu de test d'enfants, c'est-à-dire deux fois plus élevé. Cela montre bien qu'un modèle TDNNF-HMM entraîné pour la RAP d'adultes ne convient pas à la parole d'enfants et insiste sur la nécessité d'adapter ce modèle par TL. Six variantes de TL seront évaluées, correspondant aux combinaisons de deux points d'action : le modèle utilisé pour fournir les alignements servant à l'entraînement du modèle TL et la stratégie d'application du TL. Le premier point agit sur la qualité des alignements : nous possédons un GMM-HMM entraîné sur des voix d'enfants, type de modèle habituellement utilisé pour générer des alignements, ainsi qu'un TDNNF-HMM enfant, servant de référence à cette expérience et *a priori* plus performant que le GMM-HMM. Le second point comprend trois stratégies, détaillées dans la section 3.2.1 du chapitre précédent :

— *Ré-init* : ré-initialisation de la couche de sortie du modèle source, puis ré-entraînement

du modèle complet ;

— *Gelée* : gel de la couche de sortie du modèle source, puis ré-entraînement du reste du modèle ;

— *Complet* : ré-entraînement du modèle complet directement.

Le tableau 4.11 présente les scores PER obtenus avec notre référence TDNNF-HMM enfant et les six modèles TL à évaluer.

Table 4.11 – PER (%) obtenus avec différentes stratégies de TL sur le modèle TDNNF-HMM

Nom du modèle	Modèle d'alignement	Stratégie de TL	PER
Enfant	GMM-HMM enfant	-	32,6
TL 1a	GMM-HMM enfant	Ré-init	32,9
TL 1b	GMM-HMM enfant	Gelée	32,6
TL 1c	GMM-HMM enfant	Complet	32,1
TL 2a	TDNNF-HMM enfant	Ré-init	30,7
TL 2b	TDNNF-HMM enfant	Gelée	31,6
TL 2c	TDNNF-HMM enfant	Complet	**29,6**

Nous observons tout d'abord que le TL est efficace, quelle que soit la variante d'application, avec des améliorations relatives allant de 1,8% à 11,6%. Sans surprise, les alignements générés par le TDNNF-HMM sont de meilleure qualité que ceux du GMM-HMM : les modèles 1* obtiennent de meilleurs scores que les modèles 2*. La stratégie *Complet* semble la plus efficace, quels que soient les alignements utilisés. La mauvaise performance de la stratégie *Gelée* suggère que la couche de sortie du TDNNF modélise en partie des caractéristiques acoustiques qui doivent être adaptées à la parole d'enfants. Celle de la stratégie *Ré-init* permet d'ajouter que cette couche de sortie contient également des informations décisionnelles acquises sur une grande quantité de données d'adultes : perdues lors de la ré-initialisation, elles ne peuvent pas être retrouvées avec 13 heures de parole d'enfants.

Le meilleur modèle TDNNF-HMM TL s'avère être le modèle 2c, utilisant des alignements générés par le TDNNF-HMM enfant, et suivant la stratégie de TL *Complet*, consistant à ré-entraîner le modèle entier. Nous conserverons donc cette configuration d'apprentissage par transfert pour nos modèles TL dans la suite de ce manuscrit. Le score PER de ce modèle (29,6%) sera cependant amené à évoluer légèrement dans les chapitres suivant, puisque le corpus *Lalil-o-riginel* sera remplacé par le corpus *Lalil-o-fficiel*, contenant une proportion égale de mots isolés et de phrases courtes pour mieux correspondre à l'exercice de lecture orale de la plateforme.

4.3.4 Évaluation de la méthode VTLN

L'objectif de cette étude est d'améliorer la performance du TL en adaptant les voix d'adultes aux voix d'enfants par application de VTLN. Rapprocher les fréquences de la parole

d'adultes de celles d'enfant pourrait permettre au TL de mettre à profit les 13,2 heures de données cibles pour apprendre d'autres caractéristiques spécifiques de la parole d'enfants, et ainsi améliorer la reconnaissance.

Les auteurs de [Lee 1996] conseillent le choix d'un facteur α compris entre 0,88 et 1,12 (valeurs qui représentent la gamme des longueurs de conduit vocal des adultes, mais pas celle des enfants). Afin d'étendre la gamme de fréquences des adultes, nous appliquons des coefficients $\alpha = 1{,}2$ et $\alpha = 1{,}3$ pour les locutrices et locuteurs, respectivement. Cette configuration a donné les meilleurs résultats dans une étude préliminaire sur notre tout premier jeu d'entraînement, contenant environ trois heures de parole d'enfants. Nous utilisons la fonctionnalité VTLN de Kaldi, qui fixe par défaut les fréquences f_{min} et f_{max} (voir équation 3.7) à 60 et 7800 Hz. Les coefficients sont appliqués sur la parole d'adultes avant l'entraînement du modèle source, puis le modèle TL est entraîné de façon habituelle sur des données d'enfant sans normalisation de la longueur du conduit vocal. Le tableau 4.12 présente les scores PER obtenus sur de la parole d'enfants par les modèles adultes avec et sans VTLN, et les modèles TL qui en découlent.

Table 4.12 – PER (%) obtenus sur des modèles TDNNF-HMM entraînés sur parole d'adultes avec et sans VTLN, ainsi que sur les modèles TL qui en découlent

Modèle	VTLN	PER
Adulte	Non	51,8
Adulte	Oui	53,6
TL	Non	29,6
TL	Oui	31,3

Nous voyons que la VTLN ne fournit pas les bénéfices espérés pour la reconnaissance de parole d'enfants : une dégradation relative de 3,5% est observée sur les modèles adultes, et de 5,6% sur les modèles TL. Ces résultats sont surprenants, car ils contredisent ceux de [Gray 2014], où le contexte est relativement similaire : les auteurs adaptent un modèle GMM-HMM adulte avec des données de parole d'enfants sur lesquelles a été appliqué du VTLN (procédure inverse de la nôtre). Ils rapportent que cette technique a amélioré de 5% relatifs le WER sur la parole d'enfants. Il est cependant important de noter qu'ils disposent de beaucoup plus de données d'adultes (plusieurs milliers d'heures contre 150 heures) et d'enfants (200 heures contre 13 heures), et que leur corpus contient de la parole d'enfants jusqu'à 12 ans (donc plus âgés que les nôtres), ce qui pourrait expliquer la différence d'efficacité. Ces dégradations pourraient également être dues au choix d'un coefficient α fixe pour chaque genre de locuteur·rice·s, dont les longueurs de conduit vocal sont variables. Une sélection du coefficient VTLN par locuteur·rice aurait potentiellement permis de meilleures performances, comme dans [Serizel 2014b, Shivakumar 2014]. Nos résultats sont cependant en accord avec ceux de [Yeung 2018], qui établit l'efficacité de la VTLN uniquement pour les enfants de plus de 11 ans, ayant un effet négatif sur les jeunes enfants.

4.3.5 Évaluation de la méthode d'augmentation de données avec bruit

Nous étudions enfin l'efficacité de l'augmentation de données avec bruit de brouhaha sur le modèle TDNNF-HMM avec TL. L'objectif est double : améliorer la reconnaissance globale de par l'augmentation de données, et améliorer la robustesse du modèle aux bruits de salles de classe présents dans le corpus *Lalil-o-riginel*.

Les enregistrements de bruit utilisés proviennent de deux jeux de données : le corpus public DEMAND [Thiemann 2013] et le corpus privé de bruit *Lali-noise*. Dans le premier, nous avons sélectionné 16 enregistrements, d'une durée de cinq minutes chacun, de bruit de brouhaha provenant de personnes adultes discutant dans une cafétéria, situation qui se rapproche d'un environnement de salles de classe. Le bruit contenu dans ces enregistrements est peu diversifié et assez diffus. Le second a été constitué d'enregistrements obtenus par la plateforme Lalilo ayant été manuellement étiquetés comme « bruités ». Ce corpus est représentatif des environnements de salles de classe et inclut des enfants qui parlent, des enseignant·e·s qui font cours, mais également des bruits de microphones, des claquements de portes, des raclements de chaises et autres bruits typiques des salles de classe. Ces enregistrements contiennent donc une grande diversité de bruits, parfois de très courte durée.

Chaque enregistrement du jeu d'entraînement est augmenté à différents niveaux de RSB, ce qui nous permet d'imiter une salle de classe plus ou moins bruyante. Nous suivons les recommandations de Gibson et al [Gibson 2018] en créant quatre versions de chaque enregistrement, avec des RSB de 2, 5, 10, et 15 dB, comme montré sur la figure 4.2.

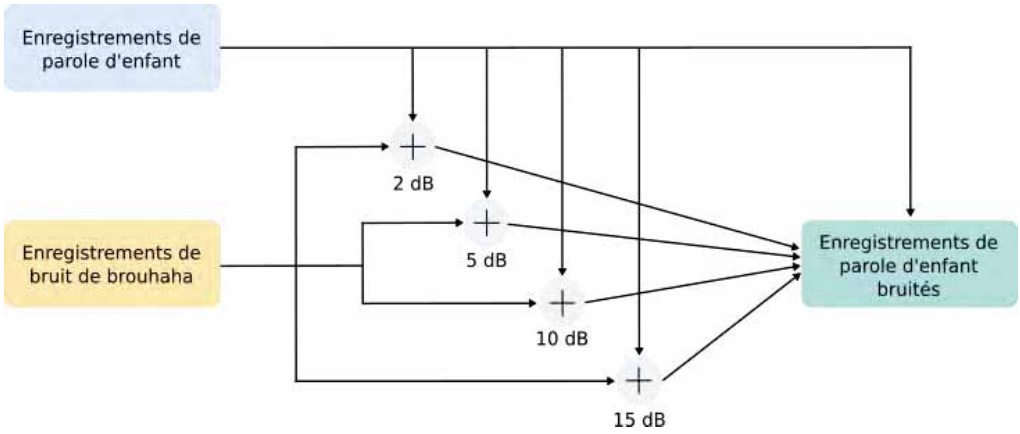


FIGURE 4.2 – Procédure d'augmentation de données avec bruit de brouhaha

Notre méthode d'augmentation d'un enregistrement d'entraînement consiste en plusieurs étapes :

— Alignement forcé avec un modèle GMM-HMM afin de localiser les segments avec et sans parole de l'enfant cible ;

— Calcul du RSB initial de l'enregistrement :

$$\mathrm{RSB}_{\mathrm{dB}} = 10 \times log \frac{P_{\mathrm{parole}}}{P_{\mathrm{bruit}}} \tag{4.2}$$

où P_{parole} et P_{bruit} désignent respectivement les puissances moyennes sur les segments avec et sans parole de l'enfant cible. Cette méthode de calcul du RSB est légèrement biaisée, car dépendante de la qualité de l'alignement forcé, qui, elle, dépend du niveau de bruit présent sur l'enregistrement ;

— Pour chaque RSB cible, uniquement si celui-ci est plus faible que le RSB initial de l'enregistrement de parole :

  ⋆ Choix aléatoire d'un enregistrement de bruit, qui est ramené à la taille de l'enregistrement de parole. Si le signal de parole est plus long, l'enregistrement est concaténé plusieurs fois jusqu'à atteindre la taille désirée. Si le signal de parole est plus court, un segment est extrait aléatoirement dans le signal de bruit ;

  ⋆ Fusion pondérée (à partir de l'équation 4.2) des signaux de parole et de bruit afin d'obtenir le RSB final désiré ;

  ⋆ Normalisation du pic à 3 dB pour éviter la saturation.

Les modèles source sont entraînés sur de la parole d'adultes non augmentée, puis le TL est appliqué en ré-entraînant le modèle avec des données de parole d'enfants augmentées avec du bruit de brouhaha. Deux modèles TL sont entraînés, l'un avec du bruit DEMAND et l'autre avec du bruit *Lali-noise*, afin d'étudier l'impact du choix de bruit en fonction de l'application. Les résultats sont présentés dans le tableau 4.13 : nous observons que les deux modèles avec augmentation dépassent le modèle TL de référence, démontrant l'efficacité de cette méthode d'augmentation de données. Le modèle TL Lali-noise-aug obtient le meilleur score PER (27,7%), qui correspond à une amélioration relative de 6,4% par rapport au modèle TL sans augmentation. Ce résultat confirme qu'utiliser des bruits identiques à ceux présents sur les enregistrements de test est à privilégier.

Table 4.13 – PER (%) obtenus avec augmentation de données avec du bruit de brouhaha, sur des modèles TDNNF-HMM avec TL

Modèle	PER
Sans augmentation	29,6
DEMAND-aug	28,7
Lali-noise-aug	**27,7**

Outre l'avantage d'augmenter la quantité de données d'entraînement, cette technique a pour objectif de rendre la reconnaissance plus robuste au bruit présent sur les enregistrements. Afin de vérifier l'accomplissement de cet objectif, nous classons les enregistrements du corpus de test *Lalil-o-riginel* en trois catégories de niveaux de bruit, mesurés avec un RSB selon la même procédure qu'en section 3.2.2 :

— « Bruité » pour des enregistrements avec un RSB < 10 dB ;

— « Moyennement bruité » pour des enregistrements avec un RSB entre 10 et 25 dB ;

— « Propre » pour des enregistrements avec un RSB > 25 dB.

La figure 4.3 présente les scores PER des modèles TL sans augmentation et avec augmentation DEMAND et *Lali-noise* sur ces intervalles de RSB.

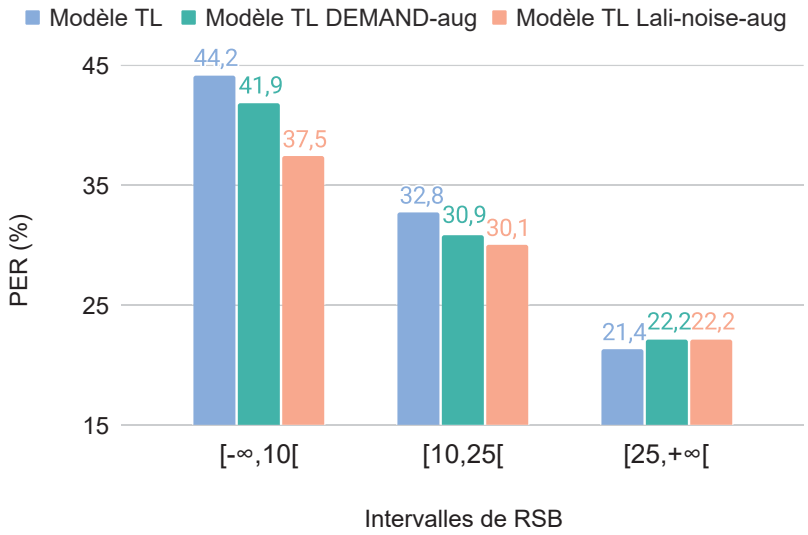


FIGURE 4.3 – PER (%) obtenus avec le modèle TDNNF-HMM, avec et sans augmentation de bruit sur différents intervalles de RSB

La première observation n'est pas surprenante : la qualité de la reconnaissance se dégrade de façon draconienne avec le niveau de bruit. Heureusement, nous voyons une très nette amélioration du PER avec les deux augmentations de données sur les enregistrements bruités, et notamment avec le modèle TL *Lali-noise*-aug qui apporte une amélioration relative de 15,2% par rapport au modèle TL. Sur les enregistrements moyennement bruités, l'amélioration est plus légère, mais toujours significative. La contrepartie à ces progrès est une légère dégradation de la performance sur les enregistrements non bruités.

4.4 Bilan

Dans ce chapitre, notre processus d'acquisition des données de parole d'enfants a été décrit : création du contenu avec l'aide des experts pédagogiques de Lalilo, collectes auprès de nos familles, via la plateforme Lalilo ou en me rendant directement dans des écoles, sélection des données et annotation. La procédure d'annotation est particulièrement complexe et a nécessité un travail rigoureux : les enfants AL peuvent se montrer très créatif·ive·s dans leurs erreurs de lecture, et il est difficile de gérer toutes les possibilités. Après l'annotation de nos données, une analyse détaillée sur un échantillon représentatif expose les erreurs de lecture les plus fréquentes faites par les enfants AL : répétitions, substitutions de mots ou de phonèmes...

Les expériences sur un premier corpus de parole d'enfants, le corpus *Lalil-o-riginel*, et liées aux techniques présentées dans le chapitre précédent, ont été détaillées et analysées. Ces analyses ont permis d'établir la référence de notre système de reconnaissance de phonèmes sur parole d'enfants : un modèle hybride TDNNF-HMM, entraîné sur des MFCC à l'aide d'apprentissage par transfert. Le modèle TDNNF-HMM a largement surpassé son aîné le

TDNN-HMM, avec un gain relatif de 11,2%. Aucun des paramètres acoustiques évalués, fondés sur du traitement du signal ou extraits par modèle auto-supervisé, n'a obtenu de meilleur score PER que les MFCC. La stratégie d'apprentissage par transfert sélectionnée consiste à ré-entraîner toutes les couches du modèle adulte source avec les données de parole d'enfants, car les caractéristiques complexes de la parole d'enfants très jeunes sont modélisées à tous les niveaux du réseau de neurones. Étonnamment, appliquer de la VTLN sur les données d'adulte pour étendre leur gamme de fréquences de façon à les rapprocher acoustiquement de celles des enfants n'a pas amélioré la reconnaissance sur parole d'enfants. Cette méthode n'est donc pas retenue dans notre système de référence pour la suite de nos travaux. Une technique d'augmentation de données par ajout de bruit de brouhaha a apporté une amélioration de la performance globale, et plus particulièrement de la robustesse du système au bruit de salle de classe caractérisant nos enregistrements. Le type de bruit le plus performant est celui du corpus *Lali-noise*, constitué d'enregistrements de salles de classe contenant de la parole d'enfants superposée. Lors de la comparaison de notre modèle de référence TDNNF-HMM aux modèles *end-to-end* dans le chapitre 6, cette augmentation ne sera appliquée sur aucun modèle pour faciliter l'interprétation des résultats. Nous la rétablirons cependant dans le chapitre 7, afin d'améliorer la robustesse au bruit de notre modèle final.

Troisième partie

Modélisation acoustique *end-to-end* et améliorations

Chapitre 5

Méthodes et architectures de nos modèles *end-to-end*

Les approches hybrides ont permis d'utiliser de nouvelles structures de réseaux de neurones pour la modélisation acoustique, comme les réseaux de neurones profonds ou récurrents, en les couplant à des HMM, qui fournissent les étapes de pré-segmentation des données et de post-traitement des probabilités de sorties. Cependant, le processus d'entraînement des modèles DNN-HMM est complexe : il faut générer les alignements de chaque énoncé en amont, et utiliser des lattices pour inférer une séquence de symboles. De plus, ces systèmes hybrides sont composés de plusieurs modules (modèles acoustique, de langage et de prononciation) entraînés séparément, ce qui peut causer des décalages de comportement entre ces modules.

Les architectures *end-to-end* ont été créées dans le but de simplifier et d'unifier le processus d'entraînement des systèmes de RAP, en réunissant tous les modules en un seul réseau de neurones. Ces architectures peuvent s'appuyer sur différentes méthodes, qui, bien que déjà universellement utilisées pour la parole d'adultes, n'ont que peu été étudiées pour la parole d'enfant. Cela motive une étude poussée de leurs effets sur la performance de nos systèmes de reconnaissance automatique de parole d'enfant. Ce chapitre présente notre sélection de méthodes et architectures *end-to-end* pour la reconnaissance automatique de phonèmes appliquée à la parole d'enfants AL.

Sommaire

5.1 Méthodes choisies pour la transcription en phonèmes de parole d'enfant

Pour une séquence d'entrée de paramètres de parole $X = (x_1,...,x_T)$ et la séquence de sortie de phonème correspondante $Y = (y_1,...,y_L)$, avec T le nombre de trames dans l'enregistrement de parole, et L la longueur de la séquence de phonèmes, l'objectif d'un modèle acoustique est de déterminer la probabilité d'avoir la séquence Y sachant X. L'ensemble des symboles de sortie est nommé S et contient les 33 phonèmes de notre vocabulaire définis en section 3.3.3. Cette probabilité est définie selon l'équation (5.1) comme le produit de la probabilité conditionnelle de chaque symbole de sortie y_i, $i = 1...L$, sachant les précédents symboles $y_{<i}$ et l'entrée X.

$$P(Y|X) = \prod_{i=1}^{L} P(y_i|X, y_{<i}) \tag{5.1}$$

Plusieurs méthodes, que nous présentons dans les sections suivantes, peuvent être utilisées dans les modèles acoustiques, seules ou combinées entre elles, avec pour objectif l'extraction d'information pour la prédiction ou le calcul de cette probabilité $P(Y|X)$.

5.1.1 Réseaux de neurones récurrents

Les réseaux de neurones récurrents « simples » (*Recurrent Neural Networks*, RNN) [Goodfellow 2016] permettent une modélisation séquentielle de l'information acoustique à différents niveaux d'abstraction. Le terme « récurrent » vient de la structure particulière en cycles, qui prend chaque élément d'entrée et sort un vecteur à une position t, dépendant des positions précédentes $1...t-1$. Cette « mémoire » est extrêmement adaptée à la modélisation de signaux temporels, et notamment aux signaux de parole qui contiennent différents niveaux d'information. La lecture orale d'apprenant·e·s lecteur·rice·s, en particulier, contient une grande quantité d'informations très complexes, du niveau très précis du phone et de l'articulation entre phones à des niveaux plus larges pour la formation de mots et de phrases, l'intonation et l'expressivité. Un réseau récurrent pourrait donc s'avérer plus efficace qu'un simple DNN pour l'extraction de ces informations complexes. La difficulté que pourraient rencontrer les RNN face à cette parole spécifique est liée à la gestion de la mémoire : une parole lente, avec des phonèmes étendus dans le temps et possiblement des erreurs de fluence comme des répétitions de mots pourraient saturer la mémoire de ces réseaux et les contraindre à oublier des informations pertinentes pour la tâche de reconnaissance de phonèmes.

Un RNN « simple » est composé de trois transformations linéaires U, V et W, définies selon le système d'équations (5.2), où $X = (x_1,...,x_T)$ et $Y = (y_1,...,y_T)$ sont les séquences d'entrée et de sortie, dans le cas simplifié où elles ont la même taille, et $H = (h_1,...,h_T)$ est

l'état caché du réseau. f est une fonction d'activation (qui varie selon le type de RNN). Les matrices U et W correspondent respectivement aux transformations linéaires liées à l'entrée x_t et à l'historique h_t à chaque position t. La matrice V incorpore l'information contextuelle contenue dans h_t dans le calcul du vecteur de sortie y_t.

$$\begin{aligned} h_t &= \tanh(Ux_t + Wh_{t-1}) \\ y_t &= f(Vh_t) \end{aligned} \tag{5.2}$$

La figure 5.1 représente un cycle de réseau récurrent pour la prédiction de la séquence Y à partir de la séquence X, en versions enroulée et déroulée. Un RNN peut également être bi-directionnel, comme sur la figure 5.2 : en couplant un réseau allant de gauche à droite et un second réseau allant de droite à gauche, et en sommant par exemple leurs vecteurs de sortie $H = (h_1,...,h_T)$ et $H' = (h'_1,...,h'_T)$ pour obtenir Y, le réseau bénéficie d'information contextuelle gauche et droite à chaque position.

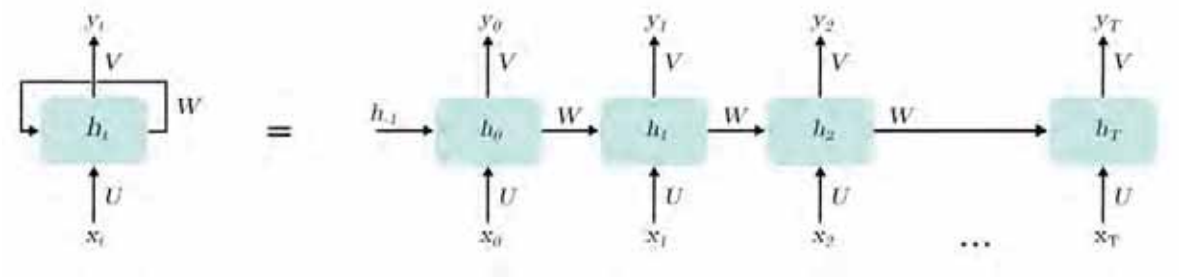


FIGURE 5.1 – Fonctionnement d'un RNN, versions enroulée (à gauche) et déroulée (à droite)

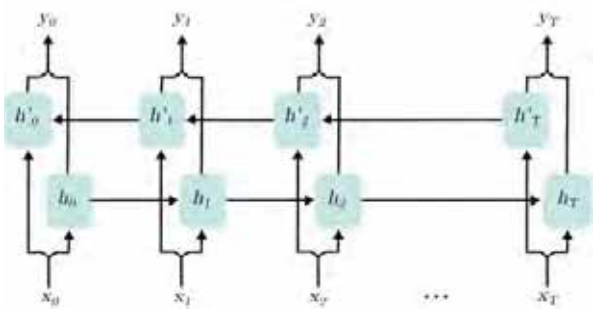


FIGURE 5.2 – Fonctionnement d'un RNN bi-directionnel, version déroulée

La structure du RNN, destinée à mémoriser des dépendances à court-terme et à long-terme à la fois, peut causer des difficultés à l'apprentissage, notamment d'explosion du gradient (un gradient trop grand est amplifié à chaque position t, jusqu'à exploser) et de disparition du gradient (au contraire, un gradient trop petit se dissipe au fur et à mesure jusqu'à ne plus agir sur les paramètres). Un mécanisme de portes permettant de contrôler et filtrer l'information à long terme offre une solution à ces problèmes [Hochreiter 1997]. Parmi les RNN utilisant ce mécanisme, les plus répandus sont les réseaux récurrents à mémoire court et long terme (*Long-Short Term Memory*, LSTM) [Hochreiter 1997] et les réseaux récurrents à portes (*Gated Recurrent Unit*, GRU) [Cho 2014].

Réseaux LSTM [Hochreiter 1997]

Les réseaux LSTM sont composés de cellules, et transmettent à chaque position, outre l'état caché h_t (voir équation (5.2)), l'état de la cellule c_t, qui sert de variable mémoire. Chaque cellule comporte trois portes (entrée, sortie et oubli). La porte d'entrée contrôle la mise à jour de l'information dans la cellule, celle de sortie contrôle le transfert de l'état du réseau à la position suivante, et la porte d'oubli contrôle la dissipation de l'information dans la cellule.

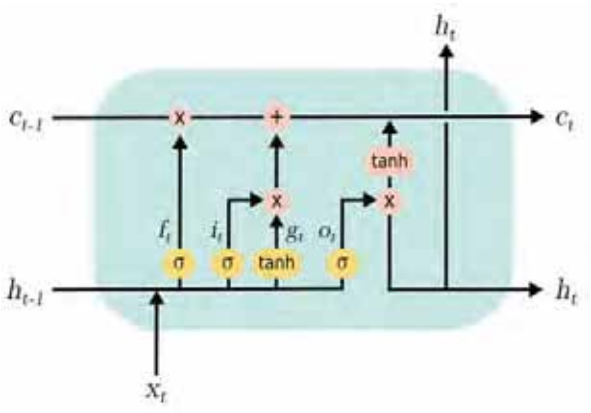


FIGURE 5.3 – Cellule de réseau LSTM

La figure 5.3 représente une cellule de réseau LSTM, où x_t est le vecteur d'entrée, h_t l'état caché et c_t l'état mémoire à la position t. Les variables f_t, i_t, et o_t représentent respectivement les fonctions d'activation des portes d'oubli, d'entrée et de sortie, qui sont définies dans l'ensemble d'équations 5.3. La variable g_t représente le vecteur mémoire candidat, qui transporte l'information à rajouter à la variable mémoire c_t. Par rapport à l'expression générale d'un RNN (équation (5.2)), un vecteur biais b est introduit, qui est mis à jour durant l'apprentissage de la même façon que les matrices de poids U et W. La fonction sigmoïde est désignée par σ, et la fonction tangente hyperbolique par *tanh*.

$$\begin{aligned} i_t &= \sigma(U_i x_t + W_i h_{t-1} + b_i) \\ f_t &= \sigma(U_f x_t + W_f h_{t-1} + b_f) \\ o_t &= \sigma(U_o x_t + W_o h_{t-1} + b_o) \\ g_t &= \tanh(U_g x_t + W_g h_{t-1} + b_g) \end{aligned} \tag{5.3}$$

L'état de la cellule c_t est mis à jour selon l'équation (5.4) grâce à une somme pondérée entre la mémoire de la position précédente c_{t-1} contrôlée par la porte d'oubli f_t et les nouvelles informations contenues dans g_t et contrôlées par la porte d'entrée i_t. Le vecteur de sortie h_t, est lui calculé à partir de l'état de la cellule c_t, pondéré par l'activation de la porte de sortie o_t.

$$\begin{aligned} c_t &= i_t \cdot g_t + f_t \cdot c_{t-1} \\ h_t &= o_t \cdot \tanh(c_t) \end{aligned} \tag{5.4}$$

Réseaux GRU [Cho 2014]

Les réseaux GRU proposent une modélisation similaire aux réseaux LSTM dans leur utilisation du mécanisme de portes, mais plus compacte car contenant uniquement deux portes (de remise à zéro et de mise à jour), et ne stockant et ne transmettant pas de variable mémoire. La porte de remise à zéro a le même rôle que la porte d'oubli du LSTM, et celle de mise à jour combine les portes d'entrée et de sortie du LSTM : elle détermine quelles parties d'information antérieure doivent être communiquées aux étapes postérieures.

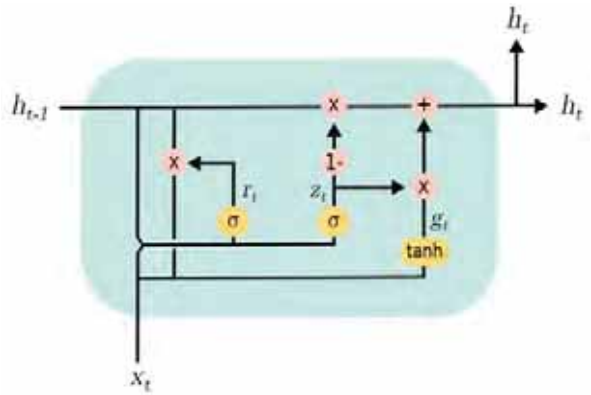


FIGURE 5.4 – Cellule de réseau GRU

La figure 5.4 représente une cellule de réseau GRU, où x_t et h_t sont les vecteurs d'entrée et de sortie de la cellule. Les variables r_t et z_t représentent les fonctions d'activation des portes de remise à zéro et de mise à jour, et sont définies par les équations (5.5), où $\odot$ désigne le produit matriciel. La variable g_t, de la même façon que pour le LSTM, transporte l'information de mise à jour, à la différence qu'elle prend ici en compte l'activation de la porte de remise à zéro r_t. La sortie de la cellule h_t est calculée selon l'équation (5.6) grâce à une pondération entre l'information nouvelle g_t et ancienne h_{t-1} contrôlée par l'activation de la porte de mise à jour z_t.

$$\begin{aligned} r_t &= \sigma(U_e x_t + W_r h_{t-1} + b_r) \\ z_t &= \sigma(U_z x_t + W_z h_{t-1} + b_z) \\ g_t &= \tanh(U_g x_t + r_t \odot W_g h_{t-1}) + b_g) \end{aligned} \tag{5.5}$$

$$h_t = z_t \cdot g_t + (1 - z_t) \cdot h_{t-1} \tag{5.6}$$

Le plus petit nombre de paramètres du réseau GRU par rapport au réseau LSTM lui permet un entraînement plus rapide. Cependant, des études suggèrent que les LSTM sont plus adaptés à la modélisation de la parole grâce à une meilleure généralisation sur des données inconnues [Weiss 2018]. Les réseaux LSTM pourraient donc être mieux adaptés à la grande variabilité intra- et inter-locuteur de la parole d'enfant.

5.1.2 La fonction CTC

La méthode *Connectionist Temporal Classification* (CTC) [Graves 2006] outrepasse la nécessité d'un HMM, en alignant automatiquement les séquences d'entrée et de sortie. Outre la simplification des processus d'entraînement, le CTC répond à une limite des HMM : la définition de ces modèles suppose que la probabilité de se trouver dans un certain état à un temps t ne dépend que de l'état au temps $t-1$, et que des observations successives sont indépendantes, ce qui n'est pas forcément le cas pour les signaux de parole, dont les trames sont liées sur des intervalles de temps plus ou moins grands [Chakraborty 2016]. Cette dépendance est de plus amplifiée dans le cas de la parole d'enfant : une étude en langue anglaise [Lee 1999] montre que les phonèmes prononcés par des enfants de 5 à 8 ans ont une durée significativement plus longue que pour des adultes ou des enfants plus âgés. Considérant les similarités entre les langues anglaise et française, nous pouvons supposer que ces résultats sont généralisables au français. Additionnellement, les débutant·e·s en lecture lisent plus lentement et ont tendance à étendre la durée des phones lors de leur déchiffrage. Cela motive l'utilisation de la fonction CTC, qui permet d'aligner des symboles de façon monotone en prenant en compte l'interdépendance entre les trames audio, pour la modélisation de la parole de jeunes enfants.

Pour l'entraînement d'un système de reconnaissance de phonèmes, la CTC aligne les trames de paroles $X = (x_1,...,x_T)$ et la séquence de phonèmes correspondante $Y = (y_1,..,y_L)$ à condition que $L \leq T$, où T et L sont respectivement le nombre de trames audio dans l'énoncé, et la longueur de la séquence de phonèmes. La différence de longueur entre T et L est comblée par la répétition de symboles sur plusieurs trames, ainsi que par l'utilisation d'un symbole « vide », représenté par « - », dont la fonction d'activation correspond à la probabilité de n'observer aucun symbole. Ce symbole « vide » sert de plus à discriminer et séparer les symboles en phonèmes (*Z-o-o* donne *Zoo* mais *Z-oo* donne *Zo*). Pendant l'alignement, le système apprend un jeu de chemins possibles $\pi = (\pi_1,...,\pi_T)$, qui sont des séquences de phonèmes de longueur T. En effet, de par la différence entre L et T, une séquence de phonèmes donnée Y peut être obtenue par différents chemins π contenus dans un ensemble θ_Y : par exemple, la séquence *Zoo* peut être obtenue par un chemin *ZZ-o-oo* ou un autre *Zooo-oo*. La probabilité d'un chemin $\pi \in \theta_Y$ est le produit des probabilités $p(\pi_t|X)$ de chacun de ses éléments π_t au temps t, et la probabilité totale d'obtenir une séquence Y est la somme des probabilités de tous ses chemins potentiels :

$$P(Y|X) = \sum_{\pi \in \theta_Y} \prod_{t=1}^{T} p(\pi_t|X) \tag{5.7}$$

Lors de l'entraînement, le système effectue une passe en avant pour chaque séquence acoustique X du jeu de données, calculant les probabilités $p(s|x)$ d'obtenir chaque symbole s (ici, les phonèmes et le symbole vide) pour chaque trame d'entrée x. Chaque phonème du texte de référence est intercalé d'un symbole vide afin de créer la séquence à aligner S : par exemple si le mot est « *Zoo* », la séquence à aligner est « *-Z-o-o-* ». Un graphe est ensuite construit à partir de la séquence à aligner S, comme sur la figure 5.5. Chaque ligne représente un symbole

de la séquence S, de longueur $L' = 2L + 1$, et chaque colonne représente le temps $t = 1...T$. Les chemins possibles pour l'obtention de cette séquence sont contraints à être monotones, c'est-à-dire qu'ils commencent en haut à gauche du graphe, et terminent en bas à droite. Les flèches bleues et vertes correspondent respectivement à rester sur le même symbole et passer au symbole suivant. Seul le symbole vide peut être ignoré, en début et en fin de séquence, ainsi qu'entre deux symboles différents (flèches oranges).

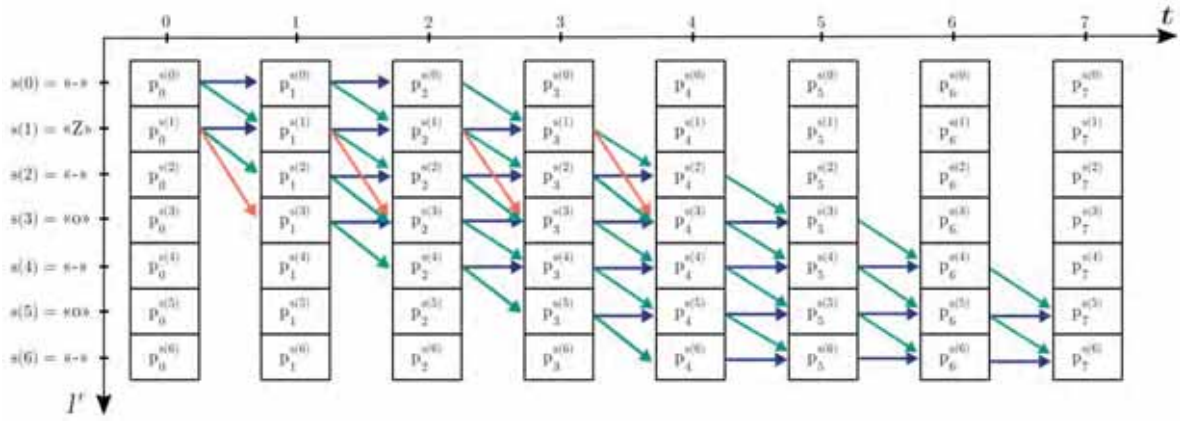


FIGURE 5.5 – Graphe d'alignement CTC pour le mot « *Zoo* »

L'algorithme suit le chemin donné, et calcule la probabilité $P(\pi_t = s_{l'}|S,X)$ de voir un certain symbole $s_{l'}$ ($l' = 1...L'$ désignant l'index courant dans la séquence S) à chaque temps t, étant données la séquence à aligner $S = S_0...S_{L'-1}$ et la séquence d'entrée $X = X_0...X_{T-1}$. Cette probabilité est notée $p_t^{s(l')}$ par souci de simplicité sur la figure 5.5. L'algorithme sépare ainsi la somme de l'équation (5.7) sur des chemins $\pi_{1,...,T} \in \theta_{Y_{1,...,T}}$ de longueur T en une somme itérative de chemins $\pi_{1,...,t} \in \theta_{Y_{1,...,t}}$ de longueur t. Les itérations sont ensuite calculées avec une propagation avant-arrière dynamique, ce qui rend possible le calcul simultané d'un grand nombre de chemins potentiels. Pour plus de détails sur l'algorithme CTC, voir l'annexe D.

5.1.3 Les mécanismes d'attention

Les mécanismes d'attention permettent de rechercher l'information nécessaire à la prédiction d'un symbole dans un vecteur source de taille variable, en localisant automatiquement les portions du vecteur dignes d'intérêt. Ces mécanismes sont très souvent utilisés dans des architectures *seq2seq* (*sequence-to-sequence*), où ils font le lien entre l'encodeur et le décodeur en cherchant l'information acoustique pertinente pour prédire chaque phonème et donc calculer la probabilité $\mathrm{P}(Y|X)$ d'avoir une séquence de phonèmes Y en fonction du vecteur acoustique X, selon l'équation (5.1). À la différence de la CTC, les mécanismes d'attention ne supposent pas l'indépendance conditionnelle entre les symboles de sortie : cela autorise une plus grande liberté mais peut également causer des confusions plus fréquentes. Ils peuvent également être inclus dans des couches d'auto-attention dans un objectif d'extraction d'information (acoustique ou textuelle). Appliqués à la parole d'enfant, les mécanismes d'attention ont l'avantage de la flexibilité, qui leur permet de considérer toutes les trames audio sans contraintes, et si besoin

de partager l'attention entre différentes trames dans le cas de phones anormalement longs. Cependant, une grande flexibilité pourrait, en présence de parole d'AL composée de nombreux évènements acoustiques et linguistiques atypiques, causer une perte de concentration vis à vis de l'information pertinente pour la tâche de reconnaissance de phonèmes.

L'attention fonctionne grâce à trois composants : les vecteurs clef, valeur, et requête. Dans le cas d'un usage de l'attention pour lier l'encodeur et le décodeur, les deux premiers sont des projections linéaires de la sortie de l'encodeur, représentant l'information acoustique, et le troisième est généré par le décodeur à partir de l'information textuelle. Certains modèles utilisent des modules d'auto-attention, contenus dans l'encodeur ou le décodeur : les trois vecteurs sont alors trois projections linéaires de la même couche de l'encodeur ou du décodeur, et permettent d'extraire directement l'information acoustique ou textuelle.

Pour chaque séquence de phonèmes (phonèmes de référence en phase d'apprentissage, et phonèmes prédits en phase d'inférence) ou de trames audio, le mécanisme calcule l'énergie entre la requête et la clef afin de localiser la portion de la clef qui correspond le mieux à la requête, puis applique l'énergie obtenue au vecteur valeur. Il existe plusieurs types d'attention, la plus simple étant l'attention à produit scalaire (PS), définie dans l'équation (5.8), où Q, K, V correspondent respectivement à la requête (*Query*), la clef (*Key*) et la valeur (*Value*).

$$\mathrm{A}_{\mathrm{PS}}(Q, K, V) = \mathrm{softmax}(QK^T)V \tag{5.8}$$

Une pondération par $1/\sqrt{d_k}$ est parfois ajoutée (voir équation (5.9), où PSP signifie produit scalaire pondéré), notamment dans les modules d'auto-attention, afin d'empêcher le produit scalaire d'exploser lorsque d_k est grand, ce qui engendrerait de très petits gradients avec la fonction softmax [Vaswani 2017]. Le facteur d_k est généralement égal à la dimension des vecteurs Q, K et V.

$$\mathrm{A}_{\mathrm{PSP}}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.9}$$

5.2 Méthodes d'inférence

L'objectif du modèle acoustique est de calculer pour chaque énoncé la probabilité $P(Y|X)$ d'obtenir une certaine séquence Y en fonction de la séquence audio X, selon l'équation (5.1). Lors de la phase d'entraînement, la séquence Y de référence est connue. Lors de la phase d'inférence, cependant, nous ne disposons pas de la vérité terrain : il faut donc utiliser un algorithme de recherche dont le rôle est de trouver la séquence Y la plus probable étant donnée la séquence audio X. Ces algorithmes cherchent à maximiser la probabilité $P(Y|X)$ en recherchant les symboles y_i les plus adaptés. Les deux algorithmes les plus utilisés pour les architectures *end-to-end* sont la recherche gloutonne (*Greedy search*) et la recherche en faisceau (*Beam search*).

5.2.1 *Greedy search*, ou recherche gloutonne

Cette première méthode consiste à choisir à chaque pas i le symbole y_i ayant la probabilité conditionnelle $p(y_i|X)$ la plus élevée parmi tous les symboles, ce qui revient à prendre le maximum local à chaque pas. Ainsi, l'algorithme fournit une solution optimale locale qui se rapproche de (et parfois atteint) la solution optimale globale, en un temps de décodage minimal. Cependant, cet algorithme peut manquer la meilleure séquence Y en éliminant un symbole y_i sans considérer les choix futurs qui pourraient en découler. Notamment, dans le cas de la reconnaissance de la parole où les trames audio sont intrinsèquement liées les unes aux autres, et où des évènements infimes peuvent troubler le modèle acoustique, la recherche gloutonne échoue souvent à trouver la meilleure séquence de phonèmes.

5.2.2 *Beam search*, ou recherche en faisceau

La recherche en faisceau, plus complexe, définit un arbre de décision en conservant k chemins et en calculant à chaque pas la probabilité totale des chemins courants, pour ne garder que les plus probables. Le pseudo-algorithme du *beam search* est présenté ci-dessous.

Algorithm 1 Pseudo-algorithme *beam search*

Initialisation : $\Omega_0 \leftarrow \{<\text{sos}>\}, \hat{\Omega} \leftarrow \emptyset$
1 : **for** $l = 1...L_{\max}$ **do**
2 : $\quad\Omega_l \leftarrow \emptyset$
3 : $\quad$**while** $\Omega_{l-1} \neq \emptyset$ **do**
4 : $\quad\quad g \leftarrow$ premier élément de Ω_{l-1}
5 : $\quad\quad$Suppression du premier élément de Ω_{l-1}
6 : $\quad\quad$**for each** $s \in S \cup \{<\text{eos}>\}$ **do**
7 : $\quad\quad\quad h \leftarrow g \cdot s$
8 : $\quad\quad\quad$Calcul de la probabilité totale $P(Y = h|X)$ du chemin $h = y_1,...y_i$
9 : $\quad\quad\quad$**if** $s = <\text{eos}>$ **then**
10 : $\quad\quad\quad\quad$Ajout de h dans $\hat{\Omega}$
11 : $\quad\quad\quad$**else**
12 : $\quad\quad\quad\quad$Ajout de h dans Ω_l
13 : $\quad\quad\quad\quad$**if** $|\Omega_l| > k$ **then**
14 : $\quad\quad\quad\quad\quad$Sélection des k meilleurs chemins de Ω_l et suppression des autres
15 : $\quad\quad\quad\quad$**end if**
16 : $\quad\quad\quad$**end if**
17 : $\quad\quad$**end for**
18 : $\quad$**end while**
19 : **end for**
20 : **return** $\arg\max_{h \in \hat{\Omega}} \alpha(h,X)$

Ω_l et $\hat{\Omega}$ représentent respectivement le set d'hypothèses partielles de longueur l et le set d'hypothèses complètes de la recherche. Le premier est initialisé avec le symbole *start-of-sequence* <sos>, le second correspond à l'ensemble vide. Aux lignes 1 à 8, pour chaque pas

$l = 1,...,L_{\max}$, avec $L_{\max}$ le nombre de symboles maximum à prédire, l'algorithme prend une séquence g contenue dans Ω_{l-1}, ajoute un symbole s pour former le chemin h et calcule la probabilité totale $P(Y = h|X)$ de ce chemin. Cette opération est faite pour tous les symboles contenus dans le vocabulaire S ainsi que pour le symbole *end-of-sequence* <eos>. Si le symbole s marque la fin de la séquence, le chemin est ajouté à l'ensemble des hypothèses complètes $\hat{\Omega}$, sinon, à celui des hypothèses partielles Ω_l du pas courant l. Une sélection est ensuite faite à la ligne 14 pour ne garder que les k chemins appartenant à Ω_l ayant la meilleure probabilité totale. L'algorithme se poursuit jusqu'à atteindre le nombre de symboles maximal $L_{\max}$, puis retourne la séquence Y avec la plus haute probabilité totale $P(Y|X)$. Cette valeur $L_{\max}$ est un hyper-paramètre fixé par l'utilisateur : il faut qu'elle soit suffisamment grande pour éviter des hypothèses partielles, mais la réduire permet d'optimiser le temps de décodage. La taille du faisceau k (*beam size*) est également un hyper-paramètre : plus elle est grande, plus l'algorithme parcourt de chemins et donc a de chances de trouver la séquence optimale, mais plus le temps de décodage est long. Si $k = 1$, cela équivaut à une recherche gloutonne (voir section précédente).

Tous nos modèles *end-to-end* utilisent lors de la phase d'inférence une recherche en faisceau, avec une taille de faisceau $k = 5$, et une longueur maximale de séquence $L_{\max}$ de 30 phonèmes pour les enregistrements de mots, et de 130 phonèmes pour les enregistrements de phrases. Dans notre cadre d'application sur la parole d'enfants AL, ces longueurs maximales ont été sur-estimées volontairement, afin de rendre possible la détection d'éventuelles répétitions et autres erreurs de lecture ou de non-maîtrise de la langue.

5.3 Modèles acoustiques *end-to-end* mises en place pour la parole d'enfant

Les architectures *end-to-end* sont des architectures plus ou moins complexes, soit fondées sur l'utilisation d'un encodeur et d'une fonction CTC, soit sur une structure *seq2seq* avec un encodeur et un décodeur reliés optionnellement par un mécanisme d'attention. Ainsi, elles peuvent combiner de façon diverse les trois méthodes proposées dans la section précédente : les RNN, la fonction CTC et les mécanismes d'attention. La figure 5.6 présente uniquement les architectures *end-to-end* explorées dans cette étude, choisies afin d'évaluer les différentes combinaisons de méthodes pour l'efficacité de la reconnaissance automatique de la parole d'enfants AL. LAS est l'acronyme de *Listen, Attend and Spell*, architecture *end-to-end* présentée dans [Chan 2016]. Chaque architecture est présentée en détail dans les sections suivantes.

Les systèmes *end-to-end* ont été développés dans un objectif de simplification maximale du processus d'entraînement des modèles. Afin de se soustraire à une étape antérieure de phonétisation et une étape postérieure de formation de mots, nécessitant un dictionnaire de prononciations, la plupart des modèles *end-to-end* sont entraînés à reconnaître des caractères alphabétiques, obtenus directement à partir des transcriptions au niveau du mot. Cependant, cette approche semble moins adaptée que la reconnaissance de phonèmes pour notre application de RAP pour l'apprentissage de la lecture.

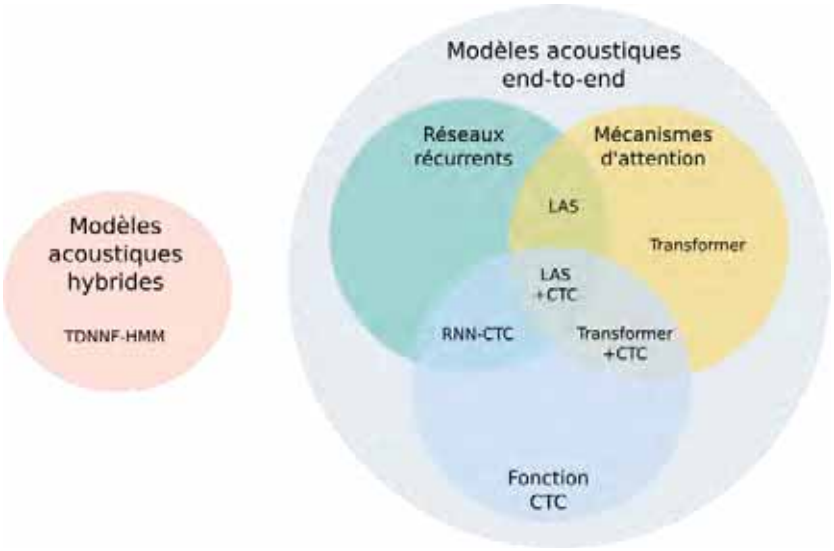


FIGURE 5.6 – Présentation des méthodes utilisées dans notre travail, en fonction des architectures de modèle acoustique

Tout d'abord, l'évaluation de la lecture à voix haute correspond par définition à l'évaluation de la prononciation des mots lus par l'enfant, et reconnaître la séquence de caractères menant à cette prononciation plutôt que directement la séquence de phonèmes a peu de sens. Deuxièmement, il est difficile d'annoter certaines erreurs de lecture au niveau du caractère, notamment quand le mot prononcé n'existe pas dans la langue française. Troisièmement, la langue française contient un grand nombre de mots homophones, c'est-à-dire qui se prononcent de façon identique malgré une orthographe parfois différente (1200 d'après [Bertrand 1990]), ainsi que de mots homographes (environ 500 d'après [Dégardin 1857]), c'est-à-dire des mots de sens différents qui s'écrivent de façon identique, et qui peuvent être homophones ou non. Un grand nombre de mots contiennent de plus des lettres muettes, dont la présence ou l'absence forme parfois des couples de mots homophones. Ces mots spéciaux ont toujours été sources d'erreur pour les systèmes de reconnaissance au niveau du mot, qu'ils utilisent un dictionnaire de prononciation ou une séparation en caractères. Ils constituent également des défis pour les apprentis lecteurs, causant de nombreuses erreurs de lecture :

— Les homophones sont souvent causés par une même prononciation pour différentes lettres ou combinaisons de lettres : par exemple le son [ɛ] est le même dans les mots « mer », « maire » et « mère » aux graphies différentes. Les graphies « è » et « ai » sont naturellement plus difficiles à lire pour l'enfant, qui peut confondre avec la graphie « é » et prononcer le phonème [e], ou ne pas connaître la combinaison des deux voyelles et les lire séparément ([ai]). Les homophones sont également formés par la présence de lettres muettes, notamment en fin de mot et sur l'accord de verbes : « il préside » et « ils président » ou encore « la cour de récréation », « le cours de français » et « elle court ». Deux erreurs classiques sont de lire la lettre muette ou de la combiner avec d'autres lettres pour former un son (« ils président » donnerait [ils pʁezidã]) ;

— Les homographes non homophones sont également sources d'erreurs car ils nécessitent une bonne compréhension du contexte pour être prononcés correctement : par exemple,

le mot « président » ne sera pas lu de la même façon dans les phrases « ils président la session » ([pʁezid]) et « le président a parlé » ([pʁezidã]).

Un modèle acoustique au niveau du caractère pour la langue française se retrouverait donc confronté à l'existence de lettre muettes et à de nombreuses orthographes différentes pour des mots homophones, ce qui augmenterait artificiellement le taux d'erreur du modèle. Le traitement de ces mots spéciaux demanderait l'utilisation d'un modèle de langage pour choisir la suite de caractères la plus adaptée au contexte. Cela pourrait dégrader les performances de nos modèles sur la parole d'AL en gommant des erreurs de lecture non existantes dans la langue française ou peu probables en fonction du contexte. Pour notre application spécifique, il est donc naturel d'apprendre au modèle acoustique à reconnaître des phonèmes et non des caractères ou des mots. Les homophones seront ainsi traités de façon identique, et de potentielles erreurs de lecture seront détectées comme de mauvaises prononciations. Le traitement au niveau phonétique des homographes permettra ainsi de distinguer des homographes non homophones, ainsi qu'une bonne d'une mauvaise prononciation en fonction du contexte.

Enfin, notre modèle de référence étant entraîné pour la reconnaissance automatique de phonèmes, utiliser la même tâche permet une comparaison plus aisée. Pour ces raisons, tous les modèles acoustiques *end-to-end* étudiés sont entraînés pour la reconnaissance automatique de phonèmes.

5.3.1 RNN-CTC

La première, et la plus simple, architecture *end-to-end* de notre étude est un modèle combinant RNN et CTC, ce qui lui vaut son nom : RNN-CTC. Comme nous pouvons le voir sur la figure 5.7, le modèle est constitué d'un simple encodeur fondé sur des réseaux de neurones récurrents, dont le rôle est d'extraire l'information acoustique pour la prédiction de parole, et d'un module CTC, qui aligne les séquences et génère la probabilité d'obtenir la séquence de sortie, selon l'équation (5.7). Les RNN constituant l'encodeur sont des réseaux GRU bidirectionnels (Bi-GRU), choisis par souci de simplicité et de rapidité d'entraînement par rapport aux réseaux LSTM.

Les paramètres d'entrée sont des MFCC de dimension 40, avec une normalisation CMVN, comme dans [Bayerl 2019]. Afin d'améliorer la concentration de l'information et d'augmenter la vitesse de traitement, la résolution temporelle est divisée par deux en combinant les paires de trames d'entrée consécutives. L'encodeur est composé de couches Bi-GRU : une couche d'entrée avec 2x80 neurones, puis quatre couches cachées de taille 2x160. Enfin, une couche linéaire de sortie de dimension 34 permet d'obtenir des séquences composées des 33 phonèmes du français et du symbole vide du CTC. Nous appliquons un taux de *dropout* de 10% sur les sorties de chaque couche cachée. La couche de sortie utilise comme fonction d'activation la fonction logarithmique softmax. Une recherche par quadrillage (*grid search*) a été effectuée pour l'optimisation des hyper-paramètres.

L'optimiseur Adam [Kingma 2017] a été utilisé pour l'entraînement, qui a été fait sur 100 epochs, avec un mécanisme d'arrêt lié à la valeur de coût sur le jeu de validation. Un

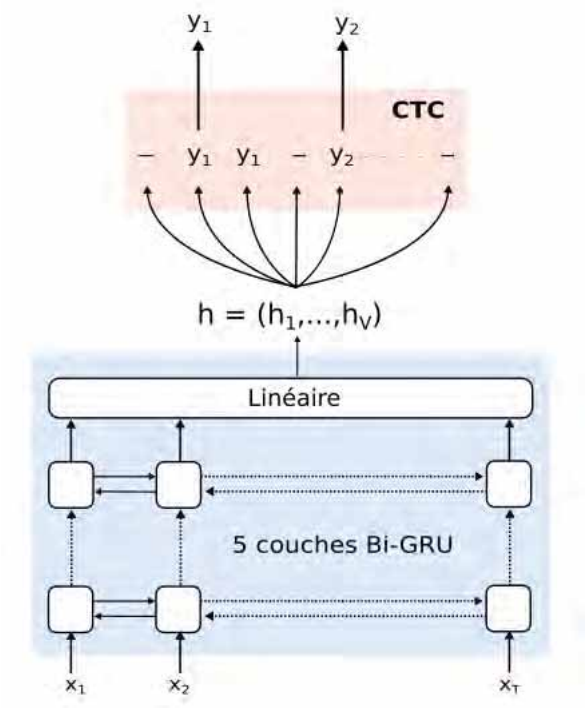


FIGURE 5.7 – Architecture du modèle RNN-CTC

planificateur de taux d'apprentissage a été utilisé, divisant par 10 le taux après deux epochs successifs sans amélioration de la perte sur le jeu de validation. Concernant l'implémentation des modèles en Pytorch, j'ai repris celle du modèle de reconnaissance de caractères dénommé CTC-GRU+TR2 dans [Heba 2019] et l'ai adaptée pour notre application : ajout de l'étape de phonétisation du texte et remplacement du modèle de langage tri-gramme par un uni-gramme. Des expériences ont été effectuées pour choisir les hyper-paramètres optimaux pour nos jeux de données (taille des couches, taux d'apprentissage, taille de batch...). Les modèles adulte, enfant et Transfer Learning (TL) ont respectivement été entraînés avec des batchs de taille 100, 50 et 50, et un taux d'apprentissage de 9e-5, 1e-4 et 1e-4. Leur temps moyen d'entraînement sur un seul GPU GTX 2018 Ti a été de 24,3, 4 et 4 heures, respectivement. Chaque modèle contient 2,1 millions de paramètres.

5.3.2 *Listen, Attend and Spell*

Le second modèle utilisé est un modèle *seq2seq*. D'abord présentés pour des tâches de traduction automatique [Sutskever 2014], où une séquence de mots dans un certain langage doit être traduite en une séquence de mots dans un autre langage, les architectures *seq2seq* visent à se passer des HMM pour faire la prédiction de séquences avec des tailles variables d'entrée et de sortie. Cette structure est généralisable à d'autres applications, comme la génération de légendes d'images [Vinyals 2015b, Xu 2015], la modélisation de conversation [Vinyals 2015a], et bien sûr la modélisation acoustique pour la RAP. En comparaison avec les autres tâches, la RAP se caractérise par de très longues séquences d'entrée (les trames de parole) et des

séquences de sortie relativement courtes (phonèmes ou caractères). Cette caractéristique est d'autant plus vraie pour les jeunes enfants, dont les réalisations acoustiques sont étendues par rapport à celles d'adultes [Lee 1999].

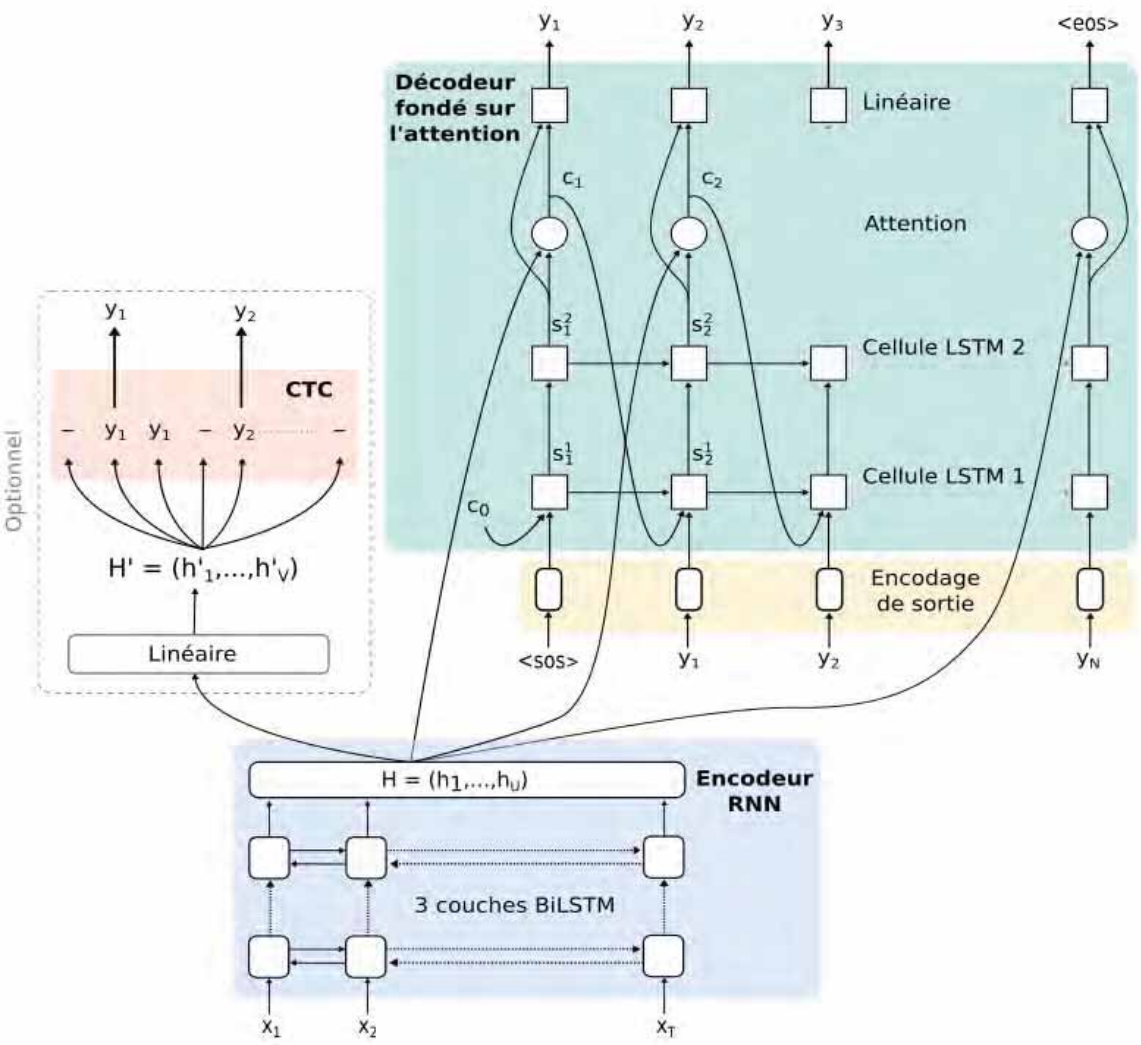


FIGURE 5.8 – Architecture des modèles LAS et LAS+CTC

Les modèles *seq2seq* sont généralement composés d'un encodeur qui extrait l'information audio et la concentre sous la forme d'un vecteur de taille variable, et d'un décodeur qui prend ce vecteur et génère une séquence de symboles (phonèmes, caractères ou autres unités de parole). Le décodeur prend également en entrée la séquence de symboles de référence, et utilise les $i-1$ symboles de référence pour prédire le symbole i. Cela a pour effet l'apprentissage d'un modèle de langage « implicite » par le décodeur, qui peut ou non être favorable à la prédiction, selon si les données vues en apprentissage concordent ou non avec les données de test. L'utilisation d'un mécanisme d'attention [Bahdanau 2014], qui localise les parties du vecteur en sortie d'encodeur qui contiennent de l'information intéressante pour prédire les phonèmes de sortie, améliore grandement les performances, et notamment pour la RAP, car

elle permet de pallier à la grande différence de longueur entre les séquences d'entrée et de sortie. Ce mécanisme génère à chaque étape du décodeur un vecteur d'attention, qui transmet de l'information contextuelle de l'encodeur au décodeur. Les encodeur et décodeur du système contiennent chacun des couches RNN, ce qui nous permet d'étudier les RNN en conjonction avec les mécanismes d'attention, plutôt qu'avec la fonction CTC dans l'architecture précédente.

Le modèle que nous utilisons est semblable à une architecture *Listen, Attend and Spell* (LAS), introduite récemment par [Chan 2016], qui combine des modules RNN et un mécanisme d'attention. Notre architecture est présentée sur la figure 5.8, où le module CTC est en option (voir section 5.3.3). Notre architecture diffère légèrement de l'originale, notamment par l'absence de la structure pyramidale des couches de l'encodeur, dont l'effet s'est avéré trop drastique pour nos enregistrements courts, concentrant l'information jusqu'à la perte d'éléments pertinents pour la reconnaissance. Nous référerons néanmoins à nos modèles grâce au nom « LAS ».

Les paramètres d'entrées de l'encodeur sont des filterbank de dimension 40, ainsi que leurs dérivées premières et secondes, suivant [Watanabe 2017]. L'encodeur contient trois couches LSTM bidirectionnelles (Bi-LSTM), qui génèrent une matrice $H = (h_1, ..., h_U)$, où chaque h est un vecteur de dimension 256, et U correspond au nombre (variable) de trames audio. Un taux de *dropout* de 20% est appliqué à chaque couche de l'encodeur. Le décodeur prend en entrée, à chaque étape i, un vecteur imbriqué de dimension 512 représentant le symbole de sortie y_i. Ce vecteur passe à travers une couche LSTM, qui génère un état $s_i^{(1)}$ selon l'équation (5.10). Une seconde couche LSTM génère ensuite un état $s_i^{(2)}$ à partir de la sortie de la couche précédente $s_i^{(1)}$ et de l'état précédent de la couche courante $s_{i-1}^{(2)}$.

$$\begin{aligned} s_i^{(1)} &= \text{Bi-LSTM}(s_{i-1}^{(1)}, y_{i-1}, c_{i-1}) \\ s_i^{(2)} &= \text{Bi-LSTM}(s_i^{(1)}, s_{i-1}^{(2)}) \end{aligned} \tag{5.10}$$

Le composant c_i est le vecteur contextuel de dimension 512 défini par l'équation (5.11), généré par un mécanisme d'attention en produit scalaire (équation (5.8)). La sortie de l'encodeur H sert à la fois de clef (K) et de valeur (V), et l'état courant $s_i^{(2)}$ correspond au vecteur requête (Q).

$$c_i = \text{A}_{\text{PS}}(s_i^{(2)}, H, H) \tag{5.11}$$

Enfin, un MLP composé de deux couches linéaires de taille 512 séparées par une couche d'activation tangente hyperbolique, fourni la probabilité $p(y_i|X, y_{<i})$ de l'équation (5.1). Pour cela, il prend en entrée l'état du décodeur $s_i^{(2)}$ et le contexte d'attention c_i :

$$p(y_i|X, y_{<i}) = \text{MLP}(s_i^{(2)}, c_i) \tag{5.12}$$

Nos modèles LAS sont entraînés avec l'optimiseur Adam, une régularisation l_2 au taux de 1e-5, et un taux d'apprentissage de 1e-4. Ce dernier est divisé par deux après chaque epoch où

la valeur de coût sur le jeu de validation ne s'améliore pas. Les modèles sont entraînés avec de l'échantillonnage programmé (*scheduled sampling*) [Bengio 2015] : cette méthode consiste à fournir au décodeur soit le phonème de référence, soit le dernier phonème prédit par le modèle. Elle a pour objectif de limiter l'apprentissage d'un modèle de langage implicite par le décodeur afin de le rendre plus robuste à des erreurs potentielles durant l'inférence. Elle permet en outre au modèle de converger plus rapidement. Le choix d'utiliser le phonème prédit plutôt que celui de référence se fait avec une probabilité fixe de 10%. J'ai implémenté personnellement un modèle LAS en Pytorch, mais les performances obtenues étaient légèrement en dessous de celles obtenues avec le code de Kaituo Xu [1]. J'ai donc repris ce modèle et l'ai adapté pour notre application. Tous les modèles sont entraînés sur 50 epochs, et uniquement le modèle avec la meilleure valeur de coût sur la validation est conservé. Les modèles adulte, enfant et TL ont nécessité en moyenne 50,7, 3,9 et 3,9 heures d'entraînement, respectivement, sur un unique GPU GTX 2080 Ti. Les modèles contiennent chacun 7,6 millions de paramètres.

5.3.3 *Listen, Attend and Spell* + CTC

Le mécanisme d'attention simple utilisé dans les modèles LAS est souvent trop flexible pour une tâche de RAP, car il autorise la génération d'alignements non monotones, alors que les entrées et sorties d'un système de RAP sont généralement monotones [Watanabe 2017]. Cette flexibilité peut être problématique pour la lecture de jeunes élèves dans le cas d'erreurs de fluence avec des répétitions de phonèmes, qui peuvent troubler l'attention. Une autre caractéristique de la tâche de RAP est la grande variabilité de longueur des séquences d'entrée et de sortie, ce qui rend les alignements parfois difficiles à générer pour les mécanismes d'attention. Cette variabilité est d'autant plus importante pour des apprenant·e·s lecteur·rice·s aux niveaux de lecture hétérogènes, dont la parole est lente et contient des erreurs de fluence.

La fonction CTC, par définition, contraint les alignements à être monotones, ce qui a motivé les auteurs de [Watanabe 2017] à combiner ces deux paradigmes, engendrant un système hybride CTC/attention, schématisé sur la figure 5.8. Ces modèles utilisent une méthode d'apprentissage multi-objectifs qui combine la fonction de coût initiale du LAS (une fonction d'entropie croisée, dénotée CE) et la fonction de coût CTC selon :

$$\text{loss} = \lambda \times \text{loss}_{\text{CTC}} + (1 - \lambda) \times \text{loss}_{\text{CE}} \tag{5.13}$$

où λ est un hyper-paramètre. Les modèles LAS+CTC sont en capacité de générer deux séquences de sorties différentes, une par l'encodeur via la fonction CTC, et l'autre par le décodeur via le mécanisme d'attention.

Les modèles LAS+CTC ont la même architecture et les mêmes hyper-paramètres que les modèles LAS décrits dans la section précédente. J'ai adapté le code du modèle LAS simple pour y ajouter l'entraînement multi-objectifs CE+CTC et obtenir le modèle LAS+CTC. Les paramètres audio en entrée sont également identiques : des filterbanks de dimension 40 et leurs premières et secondes dérivées, comme dans [Watanabe 2017]. La méthode d'échantillonnage

1. https://github.com/kaituoxu/Listen-Attend-Spell

programmé est également appliquée au taux de 10%. Le coût CTC est combiné avec le coût du décodeur suivant l'équation (5.13) avec $\lambda = 0{,}2$. D'autres valeurs plus grandes de λ (0,5, 0,8) ont donné de moins bons résultats, ce qui corrobore les résultats de [Watanabe 2017].

Le LAS+CTC a la particularité d'avoir deux sorties distinctes, comme nous pouvons le voir sur la figure 5.8 : une prédiction est faite par le décodeur grâce aux mécanismes d'attention, et une autre est faite par l'encodeur via la fonction CTC. Fondées sur des paradigmes différents, ces sorties émettent des prédictions différentes. Les performances de l'une ou de l'autre peuvent dépendre de différents paramètres affectant plus ou moins la fonction CTC ou les mécanismes d'attention : longueur de l'énoncé, rareté/difficulté des mots à lire, présence d'erreurs de lecture, niveau de bruit... Afin de tirer parti des capacités de l'une et de l'autre sans avoir à choisir au cas par cas (impossible en pratique), le décodage joint CTC/attention a été proposé dans [Hori 2017a]. Les auteurs obtiennent des gains significatifs avec cette méthode de décodage par rapport à l'utilisation d'un algorithme *beam search* sur l'unique sortie du décodeur.

Le décodage joint CTC/attention consiste à inclure les scores obtenus par les deux sorties distinctes pour calculer la probabilité totale d'une hypothèse dans l'algorithme *beam search*. Néanmoins, la simple addition de ces scores n'est pas possible, car le décodeur émet des probabilités au niveau des symboles de sortie, alors que l'encodeur émet des probabilités au niveau des trames audio. La probabilité $P_{\text{CTC}}(h|X)$ de l'hypothèse h étant donné la séquence d'entrée X est définie comme le logarithme de la probabilité de préfixe CTC [Graves 2008] :

$$P_{\text{CTC}}(h|X) = \log p_{\text{ctc}}(h,...|X) \tag{5.14}$$

Cette probabilité de préfixe représente la probabilité cumulative de toutes les séquences de symboles contenant en préfixe l'hypothèse partielle h et est calculée selon :

$$p_{\text{ctc}}(h,...|X) = \sum_{\nu \in (S \cup \{<\text{eos}>\})} p_{\text{ctc}}(h \cdot \nu|X) \tag{5.15}$$

où ν représente toutes les séquences de symboles possibles sauf la séquence vide. La probabilité CTC est combinée à la probabilité attention dans l'algorithme 1 à la ligne 8 selon :

$$P(Y = h|X) = \eta P_{\text{CTC}}(h|X) + (1 - \eta) P_{\text{Att}}(h|X) \tag{5.16}$$

où η est un hyper-paramètre de pondération, souvent pris égal au λ de l'équation 5.13. La fonction de calcul de la probabilité de préfixe CTC est décrite par l'algorithme 2 (tiré de [Hori 2017a]). J'ai récupéré la fonction Pytorch effectuant ce calcul dans l'outil SpeechBrain [Ravanelli 2021], récemment mis à disposition de la communauté scientifique, et l'ai adapté à notre implémentation de LAS+CTC. L'adaptation a notamment requis d'ajouter le symbole « vide » nécessaire à la fonction CTC dans le vocabulaire phonétique commun. Soient $\gamma_t^{(n)}(h)$ et $\gamma_t^{(v)}(h)$ les probabilités de l'hypothèse h sur les trames $t = 1,...,T$ dans les différents cas où tous les chemins CTC menant à cette hypothèse finissent par un symbole non-vide (exposant (n)) et vide (exposant (v)), respectivement. À l'initialisation de l'algorithme 1

s'ajoutent les probabilités initiales $\gamma_t^{(n)}()$ et $\gamma_t^{(v)}()$ pour $t = 1,...,T$ selon :

$$\begin{aligned}\gamma_t^{(n)}(<\text{sos}>) &= 0 \\ \gamma_0^{(v)}(<\text{sos}>) &= 1 \\ \gamma_t^{(v)}(<\text{sos}>) &= \prod_{\tau=1}^{t} \gamma_{\tau-1}^{(v)}(<\text{sos}>) p(z_\tau = <\text{->}|X)\end{aligned} \tag{5.17}$$

où <-> représente le symbole « vide ». Étudions maintenant l'algorithme 2. L'hypothèse h est d'abord décomposée en son dernier symbole s et le reste de la séquence de symboles g. Si le symbole s est <eos>, marquant la complétion de la séquence h, les probabilités à la dernière trame audio T sont renvoyées. Sinon, l'hypothèse est partielle et les probabilités $\gamma_t^{(n)}(\text{g})$ et $\gamma_t^{(v)}(\text{g})$, ainsi que la probabilité de préfixe $\Psi = p_{\text{ctc}}(h,...|X)$, sont calculées récursivement (initialisation aux lignes 5-7, et récurrence aux lignes 8-12). À chaque étape, nous rappelons que les probabilités $\gamma_t^{(n)}(\text{g})$ et $\gamma_t^{(v)}(\text{g})$ ont déjà été calculées à l'itération précédente de l'algorithme 1 puisque g est un préfixe de h.

Algorithm 2 Calcul de la probabilité de préfixe CTC

function $p_{\text{ctc}}(h,...|X)$
1 : $g, s \leftarrow h$
2 : **if** $s = <\text{eos}>$ **then**
3 : **return** $\{\gamma_T^{(n)}(g) + \gamma_T^{(v)}(g)\}$
4 : **else**
5 : $\gamma_1^{(n)}(h) \leftarrow \begin{cases} \text{p}(\text{z}_t = s|X) & \text{si g} = <\text{sos}> \\ 0 & \text{sinon} \end{cases}$
6 : $\gamma_1^{(v)}(h) \leftarrow 0$
7 : $\Psi \leftarrow \gamma_1^{(n)}(h)$
8 : **for** $t = 2...T$ **do**
9 : $\Phi \leftarrow \gamma_{t-1}^{(v)}(g) + \begin{cases} 0 & \text{si le dernier symbole de } g \text{ est } s \\ \gamma_{t-1}^{(n)}(g) & \text{sinon} \end{cases}$
10 : $\gamma_t^{(n)}(h) \leftarrow (\gamma_{t-1}^{(n)}(h) + \Phi) p(z_t = s|X)$
11 : $\gamma_t^{(v)}(h) \leftarrow (\gamma_{t-1}^{(v)}(h) + \gamma_{t-1}^{(n)}(h)) p(z_t = <\text{->}|X)$
12 : $\Psi \leftarrow \Psi + \Phi \cdot p(z_t = s|X)$
13 : **end for**
14 : **end if**
15 : **return** Ψ

Dans le chapitre suivant, où nous présenterons les performances des différentes architectures appliquées à la reconnaissance automatique de parole d'enfant, nous distinguerons les trois sources de prédiction du modèle LAS+CTC :

— la sortie de l'encodeur via la fonction CTC, dénommée « LAS+CTC enc »,

— la sortie du décodeur via les mécanismes d'attention, dénommée « LAS+CTC dec »,

— la sortie jointe « LAS+CTC joint ».

5.3.4 Transformer

Présenté par [Vaswani 2017] et adapté à la reconnaissance automatique de la parole par [Dong 2018], le modèle Transformer suit une architecture encodeur-décodeur *seq2seq*. Cependant, il se fonde uniquement sur des mécanismes d'attentions, abandonnant les réseaux de neurones récurrents habituels des modèles *seq2seq*. La récurrence, essentielle pour extraire l'information de position des trames audio, est remplacée par des encodages positionnels qui sont concaténés aux encodages d'entrée initiaux, ainsi que par des modules d'auto-attention à plusieurs têtes parallèles et des réseaux de neurones à propagation avant (*Feed-forward neural network*, FFNN) tenant compte de la position. Supprimer la nécessité de réseaux de neurones récurrents permet de calculer les dépendances entre chaque paire de positions en même temps, plutôt qu'une par une. Cela engendre une plus grande vitesse d'entraînement et plus de parallélisation possible que pour les modèles LAS vus précédemment.

Le choix de cette architecture repose, outre le fait qu'elle ait démontré d'excellentes performances sur des tâches de reconnaissance de parole d'adulte [Karita 2019b], sur la possibilité d'étudier un modèle reposant uniquement sur des mécanismes d'attention, et ainsi d'analyser le comportement des modules d'attention et d'auto-attention pour la prédiction de phonèmes et l'extraction d'informations issues de signaux de parole d'enfant. Comme mentionné dans la section 5.1.3, les mécanismes d'attention pourraient être particulièrement adaptés à la parole d'enfants AL pour leur capacité à gérer une grande complexité acoustique et linguistique.

La figure 5.9 présente l'architecture du Transformer, avec un encodeur et un décodeur, et en option le module CTC pour de l'entraînement multi-objectifs (voir section 5.3.5). L'encodeur prend en entrée la séquence audio $X = (x_1,...,x_T)$, où chaque x est un vecteur de paramètres filterbank de dimension 80. Les entrées sont tout d'abord traitées par une couche linéaire et une couche de normalisation de taille $d_{\text{model}} = 256$. Des encodages positionnels extraient des informations sur la position relative de chaque élément de la séquence (ici, les trames audio) et remplacent les récurrences ou convolutions utilisées dans les modèles classiques. Ces encodages ont la même dimension (256) que les vecteurs d'entrée pour faciliter la somme entre les deux vecteurs. Comme dans la littérature [Vaswani 2017, Dong 2018, Karita 2019b], nous utilisons des encodages positionnels sinusoïdaux, définis par :

$$\begin{aligned} \text{EncPos}_{(pos,2i)} &= \sin(pos/10000^{2i/d_{\text{model}}}) \\ \text{EncPos}_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{\text{model}}}) \end{aligned} \tag{5.18}$$

où pos est la position de l'unité (trame audio ou symbole) dans la séquence, et i est la i-ème dimension de l'encodage positionnel, avec $0 \leq i < d_{\text{model}}$.

L'encodeur contient six couches, chacune étant composée de deux sous-couches : un module d'auto-attention à plusieurs têtes, et un réseau de neurones à propagation avant tenant compte de la position. Chacun est suivi par une couche de normalisation avec une connexion résiduelle [He 2016]. La première sous-couche contient un mécanisme d'attention multi-têtes à produit scalaire pondéré [Vaswani 2017], qui relie les différentes positions des éléments

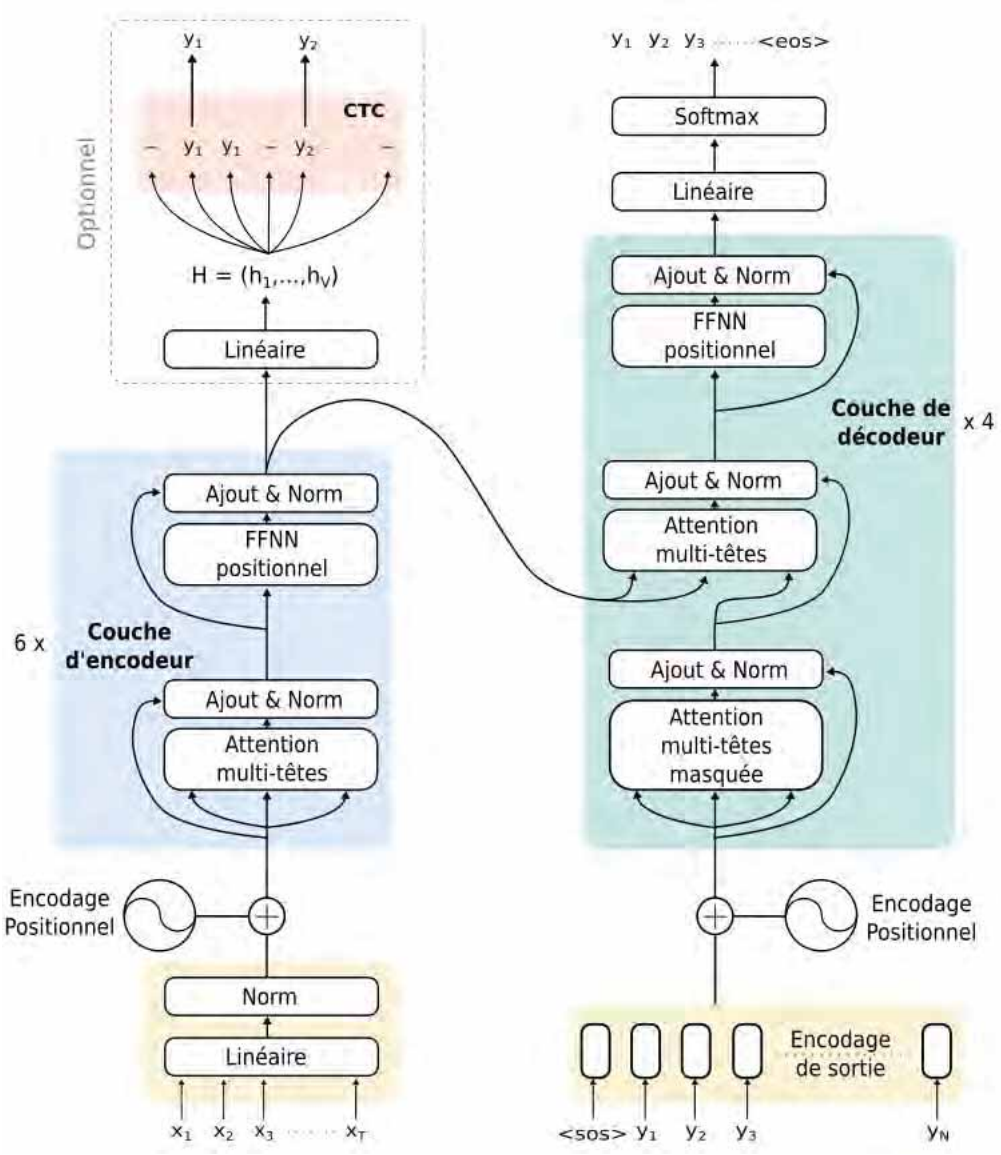


FIGURE 5.9 – Architecture des modèles Transformer et Transformer+CTC

d'entrée pour créer des représentations contextuelles. Chacune des $i \in \{1..h, h = 4\}$ têtes suit l'équation (5.19), où Q_i, K_i et V_i représentent respectivement les i-èmes projections linéaires de la requête (Q), de la clef (K) et de la valeur (V) du mécanisme d'attention en produit scalaire pondéré ($\mathrm{A}_{\mathrm{PSP}}$, voir équation (5.9)). Le facteur de pondération est égal à la dimension de Q, K, et V ($d_k = 256$). Finalement, la sortie de ce module est une projection linéaire de la concaténation de ces quatre têtes d'attention. La seconde sous-couche est un FFNN positionnel, constitué de deux couches linéaires séparées par une activation ReLU [Nair 2010], et appliqué à chaque position séparément.

$$\mathrm{head}_i = \mathrm{A}_{\mathrm{PSP}}(Q_i, K_i, V_i) \tag{5.19}$$

Les séquences d'entrée du décodeur sont les phonèmes de référence durant la phase d'entraînement, et les phonèmes prédits précédemment durant la phase d'inférence. Ces séquences sont décalées d'une position par l'insertion du symbole <sos> (*start of sequence*), puis sont représentés sous forme d'embeddings de dimension 256. Comme pour les entrées de l'encodeur, des encodages positionnels sont créés et ajoutés aux vecteurs imbriqués d'entrée du décodeur. Comme tout modèle *end-to-end* avec une structure encodeur-décodeur, le Transformer apprend un modèle de langage implicite de par l'utilisation des phonèmes de référence fournis en entrée du décodeur durant l'entraînement. Contrairement aux modèles LAS qui décodent phonème par phonème, le décodeur du Transformer utilise directement la séquence entière, ce qui permet de paralléliser et de décoder plus rapidement.

Cette caractéristique empêche cependant d'utiliser la technique classique d'échantillonnage programmé [Bengio 2015], où le dernier phonème prédit séquentiellement est fourni pour prédire le phonème suivant, dans le but d'améliorer la robustesse aux potentielles erreurs d'inférence. Une expérience a été menée sur ce sujet, suivant la méthode d'échantillonnage programmé adaptée au Transformer présentée dans [Mihaylova 2019]. La méthode consiste à effectuer deux passages dans le décodeur, le premier permettant d'obtenir les prédictions du décodeur à partir de la séquence de phonème de référence, et le second de nouvelles prédictions à partir d'un mélange de phonèmes de référence et de phonèmes prédits lors de la première étape. Le choix entre phonèmes de référence et prédits a été testé selon deux méthodes (par énoncé, ou par phonème), avec un taux d'introduction de phonèmes prédits fixe ou croissant au fur et à mesure de l'apprentissage. Ces expériences n'ont pas amené d'améliorations significatives des performances du Transformer.

Les quatre couches du décodeur contiennent chacune trois sous-couches. La première est un module d'auto-attention identique à celui de l'encodeur, à la différence qu'un masque lui est appliqué. Ce masque, combiné avec le décalage de 1 de la séquence de phonèmes d'entrée du décodeur, garantit que la prédiction d'un phonème en position i ne dépende que des phonèmes aux positions précédentes $\leq i - 1$, les phonèmes suivants n'étant pas connus durant l'inférence. La seconde sous-couche est également un module d'attention multi-têtes, qui fonctionne comme le module d'attention du LAS : le bloc d'encodage génère les clefs et les valeurs, tandis que la requête est fournie par la première sous-couche du décodeur. La troisième sous-couche est un FFNN positionnel, identique à la deuxième sous-couche de l'encodeur. La sortie du décodeur est traitée par un simple MLP composé d'une projection linéaire et d'une opération $softmax$,

qui produit la probabilité $p(y_i|X, y_{<i})$ de l'équation (5.1).

Les modèles Transformer sont entraînés avec l'optimiseur Adam, avec $\beta_1 = 0{,}9$, $\beta_2 = 0{,}98$ et $\epsilon = 1e-9$. Le même planificateur de taux d'apprentissage que le Transformer original [Vaswani 2017], défini dans l'équation (5.20) est utilisé, où N est le nombre de pas d'échauffement et est égal à 4000, et n est le pas courant.

$$\mathrm{lr} = d_{\mathrm{model}}^{-0.5} \cdot \min(\mathrm{n}^{-0.5}, \mathrm{n} \cdot \mathrm{N}^{-1.5}) \tag{5.20}$$

Pour l'implémentation du Transformer, je me suis également appuyée sur un code de Kaituo Xu [2], que j'ai adapté pour notre application. Les modèles Transformer sont entraînés sur 100 epochs, et le meilleur modèle sur la validation est conservé. Les modèles adulte, enfant et TL nécessitent en moyenne 30, 3 et 3 heures d'entraînement sur un unique GPU GTX 2080 Ti. Chaque modèle contient 14,3 millions de paramètres.

5.3.5 Transformer + CTC

Pour les mêmes raisons que le modèle LAS+CTC, utiliser la fonction CTC permet d'améliorer les performances du Transformer pour la reconnaissance automatique de la parole [Karita 2019a]. Lors de l'entraînement, les fonctions de coût CTC et entropie croisée sont combinées de la même façon que pour le LAS+CTC, suivant l'équation (5.13). Lors de l'inférence, le décodage joint est également utilisé : dans [Moritz 2020], les auteurs ont comparé les performances des prédictions de l'encodeur (via CTC), du décodeur (via l'attention) et du décodage joint CTC/attention, et ont observé des améliorations significatives avec le décodage joint par rapport aux deux sorties distinctes.

Les modèles Transformer+CTC utilisent les mêmes paramètres d'entrée, des filterbank de dimension 80, que les modèles Transformer et que les systèmes de la littérature [Karita 2019a, Karita 2019b]. La procédure d'entraînement est identique également à celle décrite en section 5.3.4. Le code des modèles Transformer+CTC est adapté de celui des modèles Transformer simples. Comme dans [Karita 2019a], les meilleures performances sont données par une combinaison des objectifs avec $\lambda = 0{,}3$. Les modèles adulte et enfant ont nécessité 100 epochs d'apprentissage, alors que les modèles TL ont convergé rapidement en 10 epochs, donnant lieu à un temps d'apprentissage très court (moins d'une heure).

Similairement aux modèles LAS+CTC, nous étudierons au chapitre 6 les résultats de :

— la sortie de l'encodeur, dénommée « Transformer+CTC enc »,

— la sortie du décodeur, dénommée « Transformer+CTC dec »,

— la sortie jointe « Transformer+CTC joint ».

2. https://github.com/kaituoxu/Speech-Transformer

5.4 Bilan

Ce chapitre était dédié à la présentation des méthodes et architectures choisies pour la transcription phonétique automatique de la parole d'enfants. Nous avons tout d'abord cherché à établir quelles techniques pouvaient être pertinentes pour notre tâche : nous nous sommes concentré·e·s sur les principales méthodes existant dans la littérature, présentant un intérêt marqué pour la modélisation acoustique et textuelle ou la prédiction d'une séquence de phonèmes à partir des trames audio. Nous avons ainsi décrit dans une première partie trois méthodes : les RNN, la fonction CTC et les mécanismes d'attention. Les RNN sont particulièrement adaptés à la modélisation de signaux temporels complexes, comme ceux de parole d'enfants qui contiennent une forte variabilité acoustique et phonologique, mais pourraient rencontrer des problèmes de gestion de mémoire sur des lectures très lentes. La fonction CTC promet une modélisation efficace prenant en compte l'interdépendance entre les trames audio, ce qui est particulièrement important pour de la parole de jeunes lecteur·rice·s, composée de longs segments phonétiques. Enfin, les mécanismes d'attention démontrent une grande flexibilité qui pourrait s'avérer pertinente pour la modélisation de la parole d'enfant, mais présentent le risque d'ignorer certains évènements atypiques, comme des erreurs de lecture.

Nous avons rapidement présenté les techniques classiques d'inférences, puis les différents modèles acoustiques *end-to-end* choisis : chaque architecture utilise une ou plusieurs méthodes de modélisation décrites en première section, ce qui nous permettra d'en étudier l'efficience pour notre application spécifique. Chacune a été implémentée par mes soins en Pytorch, en m'appuyant sur des tutoriels ou des codes existants, et en ajoutant les modules et fonctions nécessaires à l'obtention des modèles présentés dans la littérature. Alors que les modèles *end-to-end* agissent généralement au niveau du caractère, supprimant ainsi l'étape de phonétisation, nous continuerons de faire de la reconnaissance de phonèmes, qui est plus en phase avec notre application où nous devons évaluer si les correspondances graphème-phonème sont bien établies par les apprenant·e·s lecteur·rice·s. Nous avons tout d'abord présenté un simple modèle RNN-CTC composé uniquement d'un encodeur, puis des architectures *seq2seq* à l'état de l'art : les modèles LAS et Transformer. Le premier s'appuie sur des RNN et un mécanisme d'attention reliant l'encodeur et le décodeur, tandis que le second remplace les RNN par des mécanismes d'auto-attention. Des versions améliorées de ces deux architectures, dénommées LAS+CTC et Transformer+CTC, sont obtenues avec un entraînement multi-objectifs CE+CTC et un décodage hybride CTC/attention. Nous étudierons dans le chapitre 6 les résultats de toutes ces architectures, et les compareront à notre système hybride de référence TDNNF-HMM. Nous décomposerons notamment les résultats obtenus par le LAS+CTC et le Transformer+CTC en fonction de trois sorties : celle de l'encodeur via la fonction CTC, celle du décodeur via l'attention, et celle combinée via le décodage hybride CTC/attention.

Chapitre 6

Modélisation acoustique *end-to-end* : résultats et analyses

Le chapitre précédent exposait différentes méthodes de modélisation acoustique nous semblant pertinentes pour la reconnaissance automatique de parole d'enfants AL et présentait de façon détaillée les architectures *end-to-end* proposées en conséquence. Ce chapitre est consacré à l'étude de ces modèles acoustiques et de leurs performances pour notre tâche spécifique.

Nous présentons dans un premier temps le corpus *Lalil-o-fficiel* sur lequel les expériences sont réalisées, puis les résultats de nos expérimentations. Nous nous interrogeons sur la pertinence d'utiliser l'apprentissage par transfert pour entraîner des modèles acoustiques *end-to-end* avec une quantité d'enregistrements de parole d'enfants réduite, et mesurons son efficacité, variable selon l'architecture étudiée. Enfin, nous projetons notre étude dans le monde réel de Lalilo et analysons les performances de nos modèles en fonction de la tâche de lecture proposée aux enfants et de la qualité de leur lecture.

Sommaire

6.1 Présentation du corpus *Lalil-o-fficiel*

En 2020, l'exercice de lecture à voix haute de la plateforme Lalilo a évolué pour proposer un plus grand choix de contenus, avec des mots isolés et des phrases courtes. Le corpus *Lalil-o-riginel* ne contenant quasiment que des mots isolés, il a donc été rééquilibré afin de correspondre à cette nouvelle utilisation des systèmes de reconnaissance vocale. Le corpus obtenu, dénommé *Lalil-o-fficiel*, et présenté dans le tableau 6.1, sera utilisé pour toutes les expériences ci-après.

La deuxième phase d'annotation, effectuée par des transcripteur·rice·s en été 2020, a concerné des enregistrements de phrases, et nous avons pu ajouter suffisamment de phrases pour avoir une quantité égale de mots isolés et de mots venant de phrases. Le corpus *Lalil-o-riginel* a donc été complété avec 1317 phrases correspondant à environ 7 300 mots puis scindé en deux parties : *Lalil-o-fficiel* « Train » et *Lalil-o-fficiel* « Valid ». Le premier contient ainsi 13 heures de parole d'enfants, pour un total de 17 000 énoncés (15600 mots isolés, 1200 phrases courtes et 200 histoires). Le jeu « Valid » contient 500 enregistrements de mots isolés et 97 de phrases, et sera utilisé pour choisir le meilleur modèle lors de nos expérimentations.

Le jeu de test de mots isolés « Test M » est resté identique, et un jeu de test additionnel, « Test P », a été créé à partir de 391 enregistrements de phrases courtes annotées au niveau du phonème. Ces deux jeux de test seront utilisés combinés (« Test ») ou séparés dans les expériences des chapitres suivants.

Table 6.1 – Information sur le corpus *Lalil-o-fficiel*. « Test M », « Test P » et « Test » désignent respectivement les jeux de test contenant des mots isolés, des phrases et les deux.

Jeu	Train	Valid	Test	Test M	Test P
Durée (h)	13,0	0,41	1,33	0,85	0,48
Nombre de locuteur·rice·s	3014	459	687	425	262
Temps moyen (s)					
Par énoncé	2,3	2,5	2,6	2,6	2,6
Par locuteur	15,2	3,2	6,9	7,0	6,7
Nombre d'énoncés	17k	0,6k	1,7k	1172	391
Nombre de mots	31k	1,0k	2,4k	1172	2280
Nombre de phonèmes	100k	3,2k	10k	4107	5860
% de mots avec erreur(s)	0	0	28,1	46,7	17,1
RSB (dB)					
Moyenne	21,0	20,6	20,6	22,2	22,8
Std	13,0	12,6	12,6	13,2	12,1

Comme le corpus précédent *Lalil-o-riginel*, ce nouveau corpus *Lalil-o-fficiel* contient un très grand nombre de locuteurs dans chaque jeu de données, ce qui induit un faible temps de parole par locuteur. La durée moyenne des énoncés reste faible malgré l'ajout de phrases : l'algorithme de distribution des exercices de la plateforme Lalilo s'adaptant au niveau de chaque élève, la lecture de phrases est proposée à des élèves au niveau de lecture plus avancé,

qui lisent donc généralement plus vite. Les phrases sont de plus assez courtes (5 à 8 mots) car adaptées à des élèves de CP. Nous observons également que le temps moyen d'un énoncé du Test M est égal à celui d'un énoncé du Test P, alors que ce dernier contient plus de mots à lire : par le même effet de l'algorithme adaptatif, les mots isolés sont proposés aux enfants au niveau de lecture plus bas, qui lisent donc généralement plus lentement et font plus d'erreurs de fluence (répétitions, faux départs...).

Comme précisé dans le chapitre 4, l'annotation de toutes nos données au niveau phonétique n'était pas envisageable. La plupart des enregistrements sont donc uniquement classés comme « correctement lus » ou non. Les premiers peuvent être inclus dans les jeux d'entraînement et de validation, avec comme transcription la phrase proposée à l'enfant. Dans le tableau 6.1, les pourcentages de mots corrects de ces deux ensembles sont donc 100%. Afin de pouvoir évaluer la robustesse de nos systèmes de RAP sur de la lecture de jeunes apprentis lecteurs, qualité indispensable pour notre application, le jeu de test contient des énoncés annotés manuellement au niveau phonétique et incluant des erreurs de lecture. Test M et Test P contiennent respectivement 53,3% et 82,9% de mots corrects. Cette différence tient de nouveau à l'utilisation de l'algorithme adaptatif : des lecteurs plus avancés se voient proposer des phrases plutôt que des mots, et font peu d'erreurs de lecture, alors qu'à l'inverse, des lecteurs débutants doivent lire plus souvent des mots isolés et font plus d'erreurs de déchiffrage.

6.2 Comparaison des modèles hybrides et *end-to-end*

Cette section présente les performances des différentes architectures *end-to-end* présentées au chapitre précédent. Chaque architecture a servi à l'entraînement de trois modèles :

- Un modèle adulte, entraîné sur le corpus Common Voice (~150h, voir tableau 4.1) ;
- Un modèle enfant, entraîné sur le corpus *Lalil-o-fficiel* (13h, voir tableau 6.1) ;
- Un modèle TL, entraîné avec apprentissage par transfert à partir du modèle source adulte et des données *Lalil-o-fficiel*.

Les objectifs sont ainsi de démontrer la pertinence de l'apprentissage par transfert pour l'adaptation adulte-enfant de modèles acoustiques *end-to-end*, et de comparer les performances des divers modèles *end-to-end* à celles du modèle hybride de référence, le TDNNF-HMM. La variété des approches *end-to-end* nous permettra d'inférer sur l'effet des différentes méthodes de modélisation acoustique pour la parole d'enfants.

6.2.1 Comparaison sur la parole d'adultes

Une première étude des performances des modèles adultes sur de la parole d'adultes nous permet de valider les architectures *end-to-end* proposées sur une tâche de reconnaissance de phonèmes avec une quantité de données suffisante. Le tableau 6.2 contient les scores de PER obtenus avec les différents modèles adultes, lorsque l'inférence est faite sur le jeu de test Common Voice (première colonne).

TABLE 6.2 – PER (%) obtenu sur la parole d'adultes (Common Voice) puis d'enfants (*Lalilo-fficiel*) avec les différents modèles entraînés sur la parole d'adultes uniquement (Common Voice)

Entraînement Test	Common Voice Common Voice	Common Voice *Lalil-o-fficiel*
TDNNF-HMM *(référence)*	23,5	**45,9**
RNN-CTC	16,1	59,4
LAS	12,6	77,9
LAS+CTC enc	15,5	62,8
LAS+CTC dec	11,2	70,5
LAS+CTC joint	10,5	66,0
Transformer	**7,4**	76,6
Transformer+CTC enc	12,4	59,2
Transformer+CTC dec	9,1	67,7
Transformer+CTC joint	8,2	64,2

Nous observons tout d'abord que la performance des modèles suit leur apparition chronologique dans la littérature : la référence TDNNF-HMM obtient le PER le plus haut, 23,5%, tandis que le modèle Transformer, très récent, obtient le meilleur score de 7,4%. Outre l'avantage de leur simplicité d'entraînement, tous les modèles *end-to-end* obtiennent de meilleures performances que le modèle hybride de référence, avec des améliorations relatives allant de 29% à 68,5%. Les structures *seq2seq* avec encodeur et décodeur liés par un mécanisme d'attention (modèles LAS et Transformer) semblent plus efficaces que la combinaison d'un simple encodeur RNN et de la fonction d'alignement CTC du RNN-CTC. L'utilisation de modules d'auto-attention en remplacement des RNN permet au Transformer de diviser par deux le score du LAS, suggérant une meilleure capacité d'extraction d'information des signaux de parole.

L'utilisation de la fonction CTC pour un entraînement multi-objectifs améliore les performances du LAS+CTC dec par rapport au LAS, ce qui est probablement dû à l'apport d'information complémentaire lors de la phase d'alignement. L'inférence via la sortie CTC du modèle LAS+CTC est beaucoup moins précise (+4,8% absolus) que via la sortie du décodeur, fondée sur l'attention. Cette observation classe les mécanismes d'attention en tête pour le calcul des probabilités P(Y|X) (voir équation 5.1). Combiner les deux sorties par décodage joint montre que leurs hypothèses sont complémentaires : le PER est réduit de 5,0% par rapport à la sortie de l'encodeur, et de 0,7% par rapport à la sortie du décodeur.

Le potentiel de l'entraînement multi-objectifs démontré sur le LAS n'a toutefois pas été observé pour le Transformer : le modèle Transformer+CTC, quelque soit la sortie utilisée, obtient des scores supérieurs à celui du Transformer. De la même façon que pour le LAS, les sorties encodeur et décodeur du Transformer obtiennent des scores très différents, avec un avantage significatif pour la sortie du décodeur, et le décodage joint permet une amélioration de 4,2% et 0,9% par rapport aux sorties encodeur et décodeur.

6.2.2 Comparaison sur la parole d'enfants

Après avoir validé l'efficacité des différentes architectures pour la tâche conventionnelle de reconnaissance de parole d'adultes avec une quantité suffisante de données, nous étudions leurs performances sur de la parole d'enfants, avec les différentes stratégies d'entraînement présentées en début de section. Les résultats sur le jeu de test *Lalil-o-fficiel* (mots isolés et phrases confondus) sont présentés, pour les modèles adulte, enfant et TL respectivement, dans les tableaux 6.2, 6.3.

6.2.2.1 Modèles adultes

La différence entre les scores obtenus par les modèles adultes sur la parole d'adultes et la parole d'enfants (tableau 6.2, première et deuxième colonne) est drastique : les modèles perdent en moyenne 52 points de PER. Le score du TDNNF-HMM est seulement doublé entre la parole d'adultes et d'enfants, alors que les scores des modèles *end-to-end* sont multipliés par des facteurs allant de 3,8 (pour le LAS+CTC enc) à 10,2 (pour le Transformer). Ainsi, alors que le modèle de référence TDNNF-HMM était sans hésitation le moins efficace sur la parole d'adultes, il s'avère être le meilleur sur la parole d'enfants, surpassant largement tous les modèles *end-to-end*. Ces écarts énormes entre parole d'adultes et d'enfants peuvent être expliqués par les importantes différences acoustiques et prosodiques entre la parole d'adultes et d'enfants.

Nous pouvons observer que les scores des modèles LAS(+CTC) et Transformer(+CTC) sont très proches, suggérant que les RNN autant que les modules d'auto-attention ont des difficultés à extraire des informations pertinentes dans ces données inhabituelles, difficultés probablement majoritairement causées par les différences acoustiques.

Par la comparaison des modèles LAS et Transformer avec leurs alternatives LAS+CTC et Transformer+CTC, nous voyons que l'entraînement multi-objectifs CE+CTC aide légèrement les modèles à s'adapter à des caractéristiques inconnues. Nous voyons également que parmi les deux sorties indépendantes de ces modèles, la sortie de l'encodeur fondée sur la fonction CTC donne de meilleures performances que celle du décodeur fondée sur l'attention. Enfin, nous observons que le modèle RNN-CTC obtient un PER significativement plus bas que les modèles *seq2seq* utilisant des mécanismes d'attention pour l'alignement des données. Ces observations suggèrent que les modèles à base d'attention nécessitent des conditions d'entraînement et de test identiques, alors que ceux à base de CTC s'adaptent mieux à des conditions de test différentes. Les sorties jointes des modèles LAS+CTC et Transformer+CTC obtiennent des scores à mi-chemin entre ceux des sorties décodeur et encodeur : cela s'explique par le poids plus important donné aux hypothèses du décodeur, qui, de moins bonne qualité que celles du décodeur, dégradent le score final.

Les données de *Lalil-o-fficiel* se détachent également des données de Common Voice par la longueur des énoncés : alors que les adultes ne lisent que des phrases relativement longues, les enfants se voient proposer des mots isolés ou des phrases courtes. Cette caractéristique

des enregistrements pourrait avoir un impact plus ou moins important sur les méthodes de modélisation acoustiques utilisées. Nous pouvons supposer, grâce aux observations précédentes sur la CTC et l'attention, que la différence de longueur d'énoncés a un impact négatif plus fort sur les mécanismes d'attention que sur la fonction CTC. Cela rejoint les observations faites sur un modèle LAS dans [Chan 2016] : les mécanismes d'attention sont sensibles à la longueur des énoncés, et le WER augmente significativement lorsque le nombre de mots est inférieur à 5.

6.2.2.2 Modèles enfants

TABLE 6.3 – PER (%) obtenu sur la parole d'enfants (*Lalil-o-fficiel*) avec les différents modèles entraînés sur la parole d'enfants uniquement

Entraînement Test	*Lalil-o-fficiel* *Lalil-o-fficiel*
TDNNF-HMM *(référence)*	**32,5**
RNN-CTC	61,4
LAS	52,5
LAS+CTC enc	88,1
LAS+CTC dec	72,1
LAS+CTC joint	65,6
Transformer	69,7
Transformer+CTC enc	61,5
Transformer+CTC dec	62,4
Transformer+CTC joint	48,7

Nous observons premièrement dans le tableau 6.3 que le modèle de référence TDNNF-HMM obtient le meilleur score de tous les modèles enfants, avec une différence absolue de 20% de PER avec le second meilleur modèle (LAS). Les approches hybrides semblent ainsi réussir à modéliser les phonèmes contenus dans la parole d'enfants à partir d'une quantité très limitée de données. Au contraire, les modèles *end-to-end* obtiennent, sans exception, des scores supérieurs à 50% : les méthodes utilisées pour les architectures *end-to-end* sont plus flexibles que de simples DNN et HMM, et nécessitent par conséquent une plus grande quantité de données pour modéliser un signal complexe. Ce phénomène est exacerbé par la grande variabilité présente dans la parole de jeunes enfants.

Les résultats des modèles *end-to-end* sont très élevés et il est difficile d'en tirer des conclusions. Certaines architectures (LAS, Transformer, Transformer+CTC) obtiennent de meilleures performances lorsqu'elles sont entraînées sur de la parole d'enfants que sur de la parole d'adultes, mais nous observons également l'inverse pour d'autres (RNN-CTC, LAS+CTC). L'influence de l'entraînement multi-objectifs CE+CTC est positive pour le Transformer, mais négative pour le LAS. Enfin, la différence de scores entre les sorties de l'encodeur et du décodeur est très faible pour le Transformer+CTC (<1%) mais très importante pour le LAS+CTC (13,3%). Ces résultats peu cohérents peuvent s'expliquer par le fait du manque de données, qui, pour des modèles nécessitant une grande quantité de données d'entraînement, peut entraîner des

résultats aléatoires.

Nous observons néanmoins que les scores des sorties jointes des modèles LAS+CTC et Transformer+CTC sont significativement meilleurs que ceux de chaque sortie indépendamment : le décodage joint réussit à extraire le meilleur des deux sorties. Cela peut s'expliquer par des valeurs de probabilité P(y|X) très proches entre différents phonèmes, de par l'incertitude du modèle : si chaque sortie attribue au phonème à détecter une probabilité seulement légèrement supérieure aux autres phonèmes, la bonne prédiction peut se perdre dans le décodage *beam search*. En revanche, la combinaison des deux probabilités (encodeur et décodeur) pour ce phonème à détecter, chacune légèrement supérieure à celles attribuées aux autres phonèmes, renforce la probabilité de le prédire correctement.

6.2.2.3 Modèles TL

TABLE 6.4 – PER (%) obtenu sur la parole d'enfants (*Lalil-o-fficiel*) avec les différents modèles entraînés par TL

Entraînement Test	TL *Lalil-o-fficiel*
TDNNF-HMM *(référence)*	30,1
RNN-CTC	32,4
LAS	33,9
LAS+CTC enc	32,2
LAS+CTC dec	31,5
LAS+CTC joint	27,7
Transformer	28,5
Transformer+CTC enc	29,2
Transformer+CTC dec	28,3
Transformer+CTC joint	**25,0**

Au premier coup d'oeil, nous voyons dans le tableau 6.4 que les modèles entraînés avec apprentissage par transfert sont plus efficaces que les modèles entraînés seulement sur la parole d'enfants. L'amélioration relative est déjà significative pour le TDNNF-HMM (7,3% relatifs), mais dépasse tout entendement pour les modèles *end-to-end*, qui, lorsqu'ils étaient entraînés sur les 13 heures de parole d'enfants, obtenaient tous un score au-delà de 50%. Ces résultats montrent qu'il est bénéfique d'utiliser l'apprentissage par transfert lorsque nous disposons d'une quantité de données limitée, et que cette méthode est particulièrement pertinente pour les modèles *end-to-end*, qui nécessitent de grandes quantités de données d'entraînement.

Les modèles TL performent considérablement mieux que les modèles adultes, ce qui montre l'effet positif de l'apprentissage par transfert, même avec seulement 13 heures de données de parole d'enfants. Le gain moyen relatif entre les scores des modèles adultes (tableau 6.2, deuxième colonne) et des modèles TL (tableau 6.4) est de 53,1%.

Les modèles *seq2seq* avec attention (LAS et Transformers) bénéficient le plus de l'apprentissage par transfert : l'amélioration la plus spectaculaire est celle du Transformer, avec une réduction relative du PER de 62,8%. Les modèles qui n'utilisent pas d'attention, c'est-à-dire le TDNNF-HMM et le RNN-CTC en bénéficient dans une moindre mesure (34,4% et 45,5% d'amélioration relative, respectivement). Nous observons que les scores des modèles TL sont inclus dans un intervalle (28%-34%) beaucoup plus réduit que les modèles adultes (45%-78%). La convergence de toutes les architectures vers des scores autour de 30% pourrait signifier que l'efficacité du TL atteint un palier. L'étude de Shivakumar et al. montre que le WER obtenu avec un modèle hybride DNN-HMM entraîné avec TL pour la reconnaissance de parole d'enfants décroît de façon exponentielle avec l'augmentation de la quantité de données d'enfants, et ne converge toujours pas avec 100 heures de parole d'enfants [Shivakumar 2020]. Ces résultats suggèrent que ce palier pourrait être dépassé avec un jeu d'entraînement de parole d'enfants plus conséquent.

Grâce au TL, les modèles *seq2seq* à base d'attention, dont les mécanismes entraînés sur les phrases d'adulte n'étaient pas capables de s'adapter aux diverses longueurs d'énoncés et aux caractéristiques acoustiques et prosodiques des enfants, atteignent des performances comparables aux modèles sans attention. L'adaptation des mécanismes d'attention et d'auto-attention aux caractéristiques spécifiques de la parole d'enfants permet aux modèles Transformer(+CTC) de surpasser le modèle de référence TDNNF-HMM, avec des améliorations relatives du PER allant jusqu'à 16,9%. Les architectures utilisant des RNN (RNN-CTC, LAS) sont moins performantes que le TDNNF-HMM et les Transformers, ce qui suggère que les RNN nécessitent plus de données que le TDNNF ou les modules d'auto-attention pour une bonne pertinence dans l'extraction d'information sur de la parole d'enfants. Ces résultats font écho au classement établi par [Karita 2019b] entre les architectures hybrides DNN-HMM, *seq2seq* fondés sur des RNN de type LAS, et Transformers : les systèmes DNN-HMM et Transformer sont, sur la grande majorité des corpus testés, plus performants que les systèmes de type LAS.

Si nous nous concentrons sur l'influence de la CTC, nous faisons les mêmes observations pour les architectures LAS et Transformer mais dans des mesures différentes. La CTC se montre extrêmement utile pour l'entraînement du modèle LAS+CTC, améliorant de 7,1% le score du LAS+CTC dec relativement au LAS. L'amélioration est légère (0,7%) entre le Transformer et le Transformer+CTC dec. De la même façon que pour les modèles adultes sur parole d'adultes (tableau 6.2, première colonne), nous voyons que les sorties encodeur des modèles sont invariablement moins précises que les sorties décodeur. La combinaison des deux sorties est néanmoins ici très efficace, démontrant la complémentarité de la CTC et des mécanismes d'attention : le LAS+CTC joint obtient ainsi un PER de 27,7%, améliorant le score du modèle LAS de 18,3% relatifs. Le Transformer+CTC joint est le modèle le plus performant, avec un PER de 25,0%, soit une amélioration relative de 12,3% par rapport au Transformer, et de 16,9% avec le modèle de référence TDNNF-HMM. Cette amélioration par rapport au modèle de référence est significative, avec un $p = 1{,}2.10^{-4}$ au test de Wilcoxon [Wilcoxon 1945].

6.3 Analyses détaillées des performances pour notre application Lalilo

Le domaine d'application de l'exercice de lecture orale de Lalilo est loin d'être idéal pour une tâche de reconnaissance vocale : variabilité des supports de lecture et des niveaux des élèves, vitesses de lecture variables inter-locuteurs, présence d'erreurs de déchiffrage et de fluence... Cette section a pour objectif d'évaluer les différents modèles entraînés avec TL en fonction de différents facteurs pouvant heurter la précision de la reconnaissance.

6.3.1 Application aux tâches de lecture de Lalilo : mots isolés et phrases

Les enfants se voient offrir différentes tâches de lecture lors de leur apprentissage, en fonction de leur niveau de lecture ou de la compétence qu'ils doivent améliorer. Nous nous concentrons ici sur les deux tâches de lecture actuellement proposées dans l'exercice de lecture orale de Lalilo : les mots isolés, et les phrases. Sachant que les composants de chaque architecture de modélisation acoustique ont des caractéristiques aux effets potentiellement différents en fonction de la longueur des énoncés à reconnaître, nous nous intéressons à leurs comportements sur ces deux tâches de lecture. Les figures 6.1 et 6.2 affichent les performances (en termes de PER) des modèles sur les jeux de test de parole d'enfant, pour les tâches de lecture de mots isolés et de phrases, respectivement. Les types de modèles y sont regroupés, et suivent le code couleur suivant :

— gris pour le modèle hybride de référence TDNNF-HMM ;

— coloré pour les modèles *end-to-end* : vert pour le RNN-CTC, nuances de bleu pour les modèles de type LAS, et nuances de rouge-jaune pour les modèles de type Transformer.

6.3.1.1 Lecture de mots isolés

Les scores PER sur les mots isolés (Test M) sont significativement plus hauts que sur les phrases, suggérant que cette tâche est plus ardue pour les modèles acoustiques. Cette difficulté est probablement amenée par la courte durée des énoncés, privant les modèles d'un contexte suffisant, ainsi que par la grande proportion d'erreurs de lecture (46,7% des mots contiennent au moins une erreur, voir tableau 6.1).

La figure 6.1 ne dévoile pas de différence flagrante entre les différentes architectures *end-to-end*, ce qui suggère que les RNN comme les mécanismes d'auto-attention subissent le manque de contexte des énoncés très courts pour l'extraction d'information acoustique et textuelle. Le modèle TDNNF-HMM, en revanche, obtient des scores relativement proches sur les mots isolés et sur les phrases (2,8% de différence absolue), et atteint le meilleur PER sur les mots isolés : 31,7%. Nous pouvons en conclure que la structure du TDNNF-HMM est adaptée à l'extraction d'information sur des longueurs d'énoncés variables, probablement grâce à son système de délais temporels prenant en compte différentes largeurs de contextes pour différentes couches.

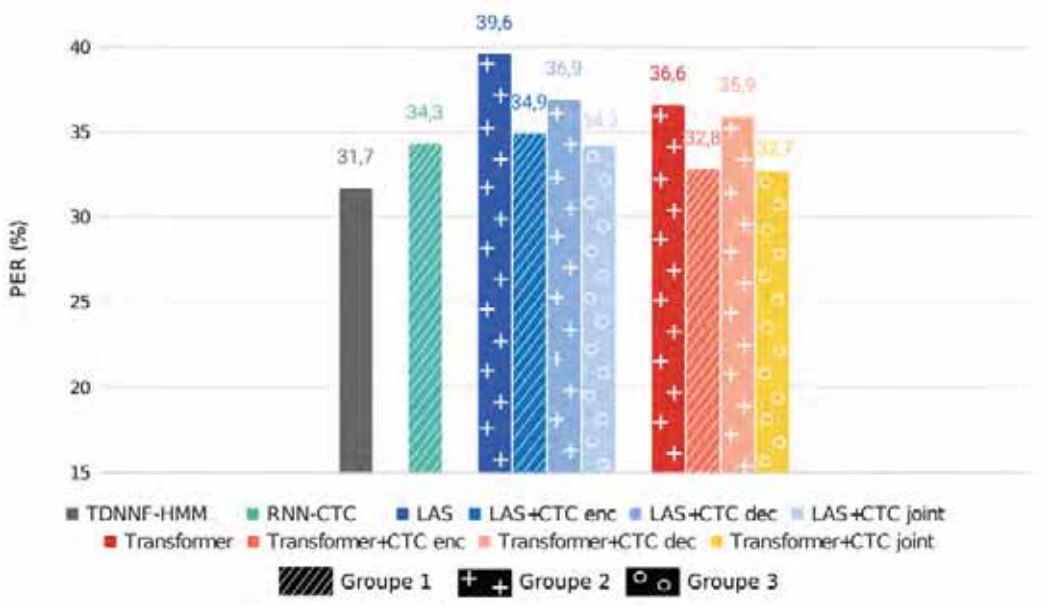


FIGURE 6.1 – PER (%) de tous les modèles sur le jeu Test M contenant des mots isolés

En étudiant l'influence de la CTC et de l'attention, nous pouvons définir plusieurs groupes de modèles *end-to-end*. Le premier groupe comprend les modèles indiqués par des traits diagonaux blancs sur la figure 6.1. Ces trois modèles (RNN-CTC, LAS+CTC enc et Transformer+CTC enc) ont en commun leur utilisation de la CTC comme sortie, et obtiennent des PER entre 32,8% et 34,9%. Nous pouvons inférer que la CTC devance les mécanismes d'attention lorsque les énoncés sont trop courts, et réussit à aligner les phonèmes correctement malgré un contexte très étroit.

Nous voyons ensuite un second groupe, qui contient les quatre modèles indiqués par des croix blanches, utilisant une sortie inférée par le décodeur via l'attention : LAS, LAS+CTC dec, Transformer et Transformer+CTC dec. L'entraînement multi-objectifs CE+CTC a ici un effet mitigé sur la sortie décodeur des modèles LAS+CTC et Transformer+CTC. Alors qu'il permet une amélioration sur le premier (6,8% d'amélioration relative entre le LAS et le LAS+CTC dec), il dégrade légèrement les performances du deuxième (-1,6% relatifs). Les quatre modèles utilisant un mécanisme d'attention liant l'encodeur et le décodeur, et effectuant la prédiction avec cette attention, obtiennent les scores les plus hauts, supérieurs à 36,6%. Cela montre que ces mécanismes n'arrivent pas à décoder sur des séquences audio très courtes, probablement à cause du manque de contexte. Ce résultat fait écho à [Chan 2016], où les auteurs montrent que le WER du LAS augmente significativement lorsque le nombre de mots contenus dans un énoncé descend en dessous de 3.

Le troisième et dernier groupe est constitué des modèles LAS+CTC et Transformer+CTC avec décodage hybride CTC/attention, qui sont indiqués à l'aide de rond blancs sur la figure 6.1. Nous distinguons pour ces deux modèles la même amélioration très légère par rapport à la sortie de l'encodeur (0,7% et 0,1% d'amélioration absolue entre LAS+CTC enc et joint, et Transformer+CTC enc et joint). Cela montre que les mécanismes d'attention ne modélisent que

peu d'informations complémentaires à celles modélisées par la fonction CTC, probablement à cause du manque de contexte dû à la petite taille des énoncés. Le Transformer+CTC notamment atteint un score de 32,7%, dont la différence avec le score du modèle de référence TDNNF-HMM (31,7%) n'est pas significative (Wilcoxon, $p = 0{,}27$).

6.3.1.2 Lecture de phrases courtes

Sur les phrases (Test P), la référence TDNNF-HMM se place en septième position. Les modèles Transformer(+CTC) obtiennent des scores significativement meilleurs que les autres modèles, dépassant le TDNNF-HMM de 10,0% à 32,2% relatifs. Le meilleur score, 19,6% est obtenu par le modèle Transformer+CTC joint (voir figure 6.2). La différence de score avec le modèle de référence est significative : $p = 1{,}2.10^{-20}$ au test de Wilcoxon.

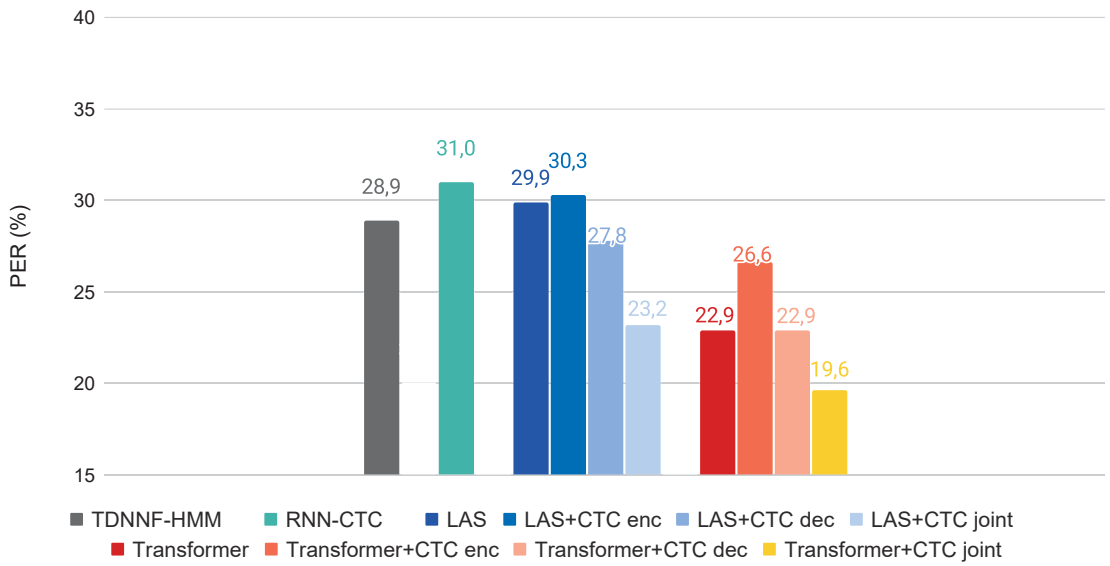


FIGURE 6.2 – PER (%) de tous les modèles sur le jeu Test P contenant des phrases

Nous voyons que les architectures s'appuyant sur des réseaux RNN[5] offrent de moins bonnes performances que les architectures Transformer, fondées uniquement sur de l'attention. Cette tendance était déjà observée sur les mots et phrases confondus (tableau 6.4), mais n'apparaît pas dans le cas des mots isolés de la figure 6.1 : cela suggère que les énoncés longs peuvent heurter la précision des RNN. Nous pourrions supposer que la gestion de la mémoire est plus difficile pour des énoncés longs.

De la même façon que sur les mots isolés, les scores sur les phrases démontrent une tendance claire d'amélioration par l'entraînement multi-objectifs CE+CTC avec décodage joint : le LAS+CTC joint améliore de 22,4% relatif le score du LAS, et le Transformer+CTC joint de 14,4% celui du Transformer. La sortie décodeur apporte des résultats mitigés, avec une amélioration pour le LAS+CTC dec par rapport au LAS, mais pas pour le Transformer+CTC

dec par rapport au Transformer. En revanche, l'utilisation de la sortie encodeur via la CTC détériore grandement les performances en comparaison avec la sortie décodeur, dans une mesure bien plus importante que celle observée sur les mots et phrases confondus. La fonction CTC pourrait ainsi être en difficulté pour aligner des énoncés plus longs, alors que les mécanismes d'attention profiteraient au contraire d'un large contexte pour améliorer leurs prédictions. Les importantes améliorations apportées par les modèles avec entraînement CE+CTC et décodage joint suggèrent cependant que les informations extraites par la fonction CTC et les mécanismes d'attention sont complémentaires sur des enregistrements de phrases.

6.3.1.3 Quel modèle pour quelle tâche ?

Les résultats et classements des modèles, différents en fonction de la longueur du texte à reconnaître, pourraient poser des difficultés pour le choix d'un modèle à intégrer dans la plateforme Lalilo. Dans le cadre de l'exercice de lecture à voix haute de Lalilo, néanmoins, nous connaissons le type de contenu que l'enfant doit lire (mot isolé ou phrase), et sommes en mesure de contrôler le modèle utilisé pour inférer la séquence de phonèmes lus. Il serait donc en théorie possible d'utiliser deux modèles différents, un pour chaque tâche. Mais qu'en est-il en pratique ? Utiliser le meilleur modèle pour chaque tâche, c'est-à-dire le TDNNF-HMM pour les mots et le Transformer+CTC joint pour les phrases, serait très coûteux en pratique, pour les raisons suivantes :

— Les deux modèles sont lourds (7,6 et 14,1 millions de paramètres, respectivement), ce qui peut ralentir la plateforme et poser des problèmes d'intégration ;
— L'entraînement et l'optimisation de deux modèles sont coûteux en ressources de calcul et en temps ;
— Les deux modèles sont implémentés avec des boîtes à outils différentes : Kaldi pour le TDNNF-HMM, et Pytorch pour le Transformer+CTC, ce qui complique l'intégration et la maintenance ;
— L'inférence du TDNNF est faite avec un décodage de lattices WFST, alors que celle du Transformer+CTC est une recherche par faisceau, ce qui augmente la quantité de code à développer et à maintenir.

En revanche, nous pouvons voir sur la figure 6.1 que le modèle Transformer+CTC joint obtient un score intéressant de 32,7%, ce qui le classe à la deuxième position pour les mots isolés, avec une différence absolue de 1,0% avec le score du TDNNF-HMM. Ce résultat motive la possibilité d'utiliser uniquement le Transformer+CTC joint, exploitant le décodage hybride pour obtenir de bonnes performances, quelque soit la taille de l'énoncé. Nous utiliserons ainsi dans la suite de ce chapitre le terme simplifié « Transformer+CTC » pour désigner le Transformer+CTC joint.

6.3.2 Influence des particularités de la lecture d'AL

Créé pour la reconnaissance automatique de la parole d'enfants AL, notre système rencontre inévitablement des vitesses de lecture hétérogènes et de nombreuses erreurs de lecture. Pour

remplir sa mission, il doit pouvoir reconnaître avec précision les phonèmes lus par l'enfant, indépendamment du niveau de lecture de l'enfant. Les lectures très lentes ou contenant des erreurs de lectures peuvent cependant rendre difficile cette tâche. C'est pourquoi nous cherchons dans cette section à évaluer la performance du Transformer+CTC lorsqu'il est confronté à ces évènements. De la même façon que dans la section précédente, nous comparons les performances du Transformer+CTC à celles du TDNNF-HMM et du Transformer.

6.3.2.1 Influence de la vitesse de lecture

Imitant les méthodes des professeurs des écoles, nous mesurons la vitesse de lecture grâce à un nombre de mots corrects par minute (*Word Correct Per Minute*, WCPM). Habituellement, cette métrique est calculée en prenant le nombre de mots correctement lus dans un énoncé d'exactement une minute. Nos enregistrements étant d'une durée inférieure à une minute, nous divisons le nombre de mots par le temps de l'énoncé, ce qui cause une légère sur-estimation de la métrique, puisque l'enfant n'a pas besoin de respirer entre les phrases, de changer de ligne ou d'avaler sa salive. Le déroulement de l'exercice de Lalilo permet de plus à l'enfant de lire la phrase dans sa tête avant d'enregistrer sa lecture orale, ce qui implique encore une fois une sur-estimation de notre WCPM. Afin d'éviter des biais trop importants, nous ne calculerons le WCPM que sur les phrases : par conséquent tous les résultats de cette analyse seront présentés sur le jeu Test P.

Nous définissons des intervalles de WCPM qui correspondent aux valeurs attendues en fin de chaque classe de l'école élémentaire (du CP au CM2). Ces valeurs sont définies par le programme de lecture du Ministère de l'Éducation Nationale [1]. Les valeurs de WCPM attendues en fin de CP, CE1, CE2, CM1 et CM2 sont respectivement 50, 70, 90, 110 et 120. Le tableau 6.5 présente les intervalles de WCPM définis en fonction de ces normes : par exemple, un enregistrement est classé dans la catégorie « CP » s'il a un WCPM inférieur ou égal à 50, qui est le niveau attendu pour un élève de CP. Le nombre d'énoncés du Test P correspondant à chaque catégorie est également renseigné dans le tableau 6.5 : comme précisé dans le paragraphe précédent, notre calcul du WCPM est légèrement sur-estimé, ce qui fait que les classes CM1 et CM2 sont sur-représentées alors que les enregistrements proviennent d'enfants en classes de CP à CE2. Pour la même raison, le jeu de données Test P a un WCPM moyen très haut de 96,7 avec une déviation standard de 47,0.

Table 6.5 – Intervalles de WCPM correspondant au niveau attendu dans chaque classe de l'école élémentaire, et le nombre d'énoncés correspondant à chaque catégorie

Classe	CP	CE1	CE2	CM1	CM2
Intervalle de WCPM	[0,50]	]50,70]	]70,90]	]90,110]	]110,+∞[
Nombre d'énoncés	57	51	48	78	119

La figure 6.3 affiche les valeurs de PER obtenues par le TDNNF-HMM, le Transformer

1. https://colibris.link/national-standards-WCPM

et le Transformer+CTC, en fonction de la catégorie de WCPM de chaque enregistrement. Nous pouvons voir que les trois modèles sont en difficulté sur les enregistrements de « CP », c'est-à-dire sur la lecture très lente, mais que leurs performances s'améliorent considérablement lorsque la vitesse de lecture augmente. Ce phénomène peut s'expliquer par la faible vitesse de lecture, avec des hésitations intra- et inter-mots et des extensions de phones, qui peut rendre la reconnaissance laborieuse, notamment pour les modules d'(auto-)attention des modèles Transformer et Transformer+CTC. De plus, le nombre WCPM n'inclut que les mots correctement lus par l'enfant : un enregistrement avec un faible WCPM comporte donc probablement plus d'erreurs de lecture qu'un enregistrement avec un WCPM haut, ce qui augmente la difficulté de la tâche de RAP. Parmi les erreurs de lecture se trouvent les erreurs de fluence, notamment les répétitions, qui augmentent la durée de lecture sans être comptées dans le nombre de mots corrects, et donc diminuent drastiquement le WCPM. Une solution pourrait être d'incorporer une plus grande quantité de lectures lentes dans le jeu d'entraînement, qui actuellement présente un WCPM moyen de 114,3 (±40,8). De la même façon, incorporer dans le jeu d'entraînement des enregistrements à WCPM faible de par la présence d'erreurs de lecture pourrait être bénéfique. Nous pourrions également imaginer des techniques d'augmentation de données spécifiques pour créer des enregistrements à WCPM faible en créant synthétiquement des hésitations, des extensions de phones ou d'autres erreurs de fluence.

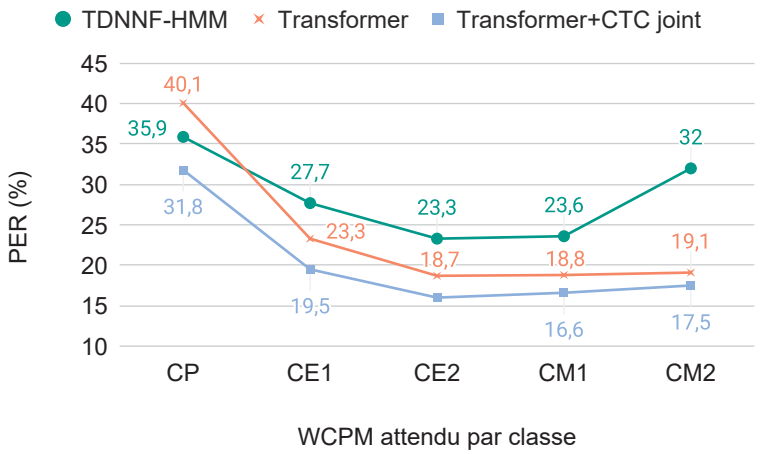


FIGURE 6.3 – PER (%) des modèles TDNNF-HMM, Transformer et Transformer+CTC en fonction du WCPM, affichés selon les niveaux attendus dans chaque classe de l'école élémentaire

Nous observons cependant sur la catégorie « CP » une très nette amélioration de la reconnaissance du Transformer+CTC par rapport au Transformer, ce qui lui permet de surpasser le TDNNF-HMM. Cette amélioration est possiblement due à une meilleure appréhension des erreurs de fluence comme les répétitions : les mécanismes d'attention, trop flexibles, ont tendance à fusionner plusieurs occurrences d'une séquence de phonème en une seule, ignorant ainsi les mots répétés. La fonction CTC contraint l'attention à être monotone, ce qui limite l'oubli des répétitions et donc améliore le PER.

De l'autre côté du graphique, nous voyons que la performance du TDNNF-HMM se dégrade fortement dans la catégorie « CM2 », c'est-à-dire pour les lectures très rapides avec un WCPM

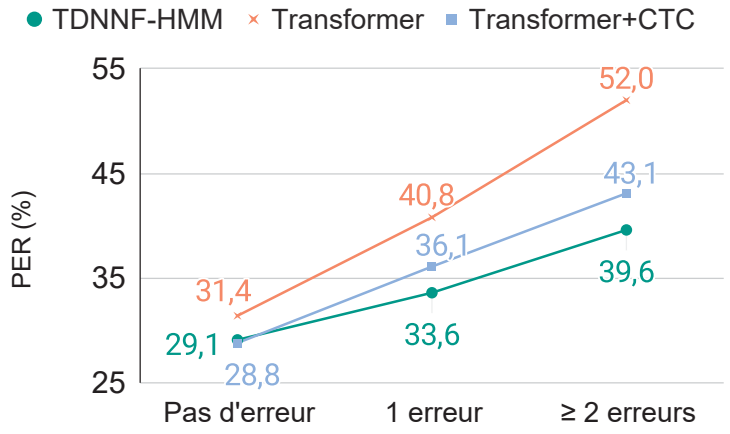
supérieur à 110. Cela pourrait être dû au sous-échantillonnage (de facteur 3) opéré dans les modèles hybrides dits « chaînes » [Povey 2016], qui ignorerait une proportion trop importante de trames acoustiques pour les phones de très courte durée. Le score des modèles Transformer et Transformer+CTC sur cette catégorie augmente seulement légèrement, ce qui suggère une bonne adaptabilité des mécanismes d'attention aux lectures rapides.

6.3.2.2 Influence de la présence d'erreurs de lecture

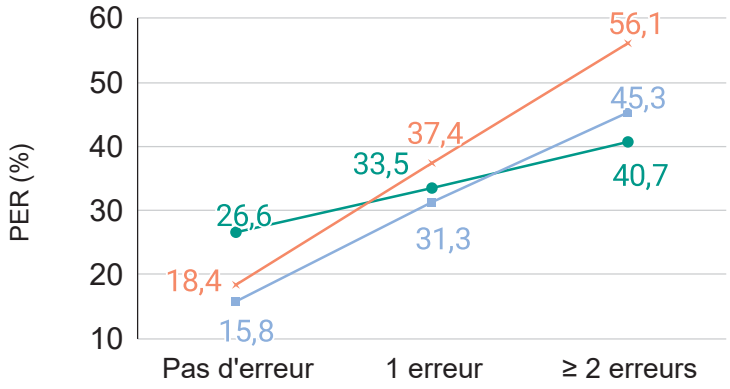
Nous souhaitons maintenant évaluer l'influence de la présence d'erreurs de lecture sur la précision de notre modèle Transformer+CTC. Nous nous comparons sur ces aspects au modèle de référence TDNNF-HMM, ainsi qu'au Transformer simple afin d'analyser l'impact de l'utilisation de la fonction CTC. Les niveaux de PER sur les deux tâches de lecture (mots isolés et phrases) étant très différents, nous analysons chaque tâche séparément. Les figures 6.4a et 6.4b présentent les scores PER des modèles TDNNF-HMM, Transformer et Transformer+CTC pour les mots qui contiennent 0, 1 et $\geq$2 erreurs, pour les mots isolés et les phrases, respectivement. Nous nous intéressons ici aux erreurs de déchiffrage, et non aux erreurs de fluence. Une « erreur » correspond donc ici à la substitution, insertion ou suppression d'un phonème dans le mot. Pour cette raison, les mots répétés sont inclus dans les catégories en fonction du nombre d'erreurs au niveau du phonème qu'ils contiennent : par exemple, si un·e enfant qui doit lire le mot « pouce » lit [pus pus] alors les deux mots seront inclus dans la catégorie « Pas d'erreur », mais s'il·elle lit [puk pus] alors le premier mot correspond à la catégorie « 1 erreur ».

Une première observation, commune aux deux tâches de lecture, est que les performances des modèles se dégradent drastiquement avec l'apparition d'erreurs de déchiffrage. Le TDNNF-HMM semble le moins affecté, avec des dégradations absolues entre 0 et $\geq$2 erreurs de 11,8% et 10,5% sur les phrases et mots respectivement. Le Transformer est le plus affecté, avec 33,2% et 20,6%, contre 25,6% et 14,5% pour le Transformer+CTC.

La mauvaise performance du Transformer en présence d'erreurs de lecture peut être expliquée par l'influence du modèle de langage implicite appris par le décodeur durant l'entraînement à travers les séquences de phonèmes de référence. Or, les séquences de phonèmes vues à l'entraînement ne contiennent pas d'erreurs de lecture : durant l'inférence, ce modèle de langage implicite aura ainsi tendance à gommer les erreurs de lecture, et notamment celles résultant en des mots non-existants. Cet effet impacte donc la précision sur les mots contenant au moins une erreur de lecture. L'utilisation de la fonction CTC a un effet positif sur le Transformer : nous voyons que les performances du Transformer+CTC sont meilleures que celles du Transformer, sur les mots et les phrases, en présence d'erreurs ou non. Nous voyons de plus que l'amélioration apportée croît avec le nombre d'erreurs. Les mots contenant au moins une erreur de lecture sont probablement mieux reconnus grâce au décodage joint attention/CTC, qui permet de diminuer l'effet du modèle de langage implicite du décodeur par le contrepoids de l'hypothèse prédite par la CTC. Le TDNNF-HMM est moins affecté par la présence d'erreurs de lecture, ce qui est vraisemblablement dû à notre utilisation d'un modèle de langage uni-gramme au niveau du phonème, limitant ainsi la tendance à gommer



(a) Mots isolés



(b) Phrases

FIGURE 6.4 – PER (%) des modèles TDNNF-HMM, Transformer et Transformer+CTC pour chaque mot en fonction du nombre d'erreurs de lecture contenues dans ce mot, provenant d'énoncés (a) de mots isolés (Test M) (b) de phrases (Test P)

les erreurs de lecture.

Sur les mots isolés (figure 6.4a), tâche sur laquelle les modèles Transformer ont montré de moins bonnes performances globales que le TDNNF-HMM, nous voyons que le Transformer+CTC dépasse à peine le score du TDNNF-HMM sur les mots correctement lus. Malgré les améliorations substantielles apportées au Transformer par l'utilisation de la fonction CTC, le TDNNF-HMM conserve son avance quant à la reconnaissance des mots contenant des erreurs de lecture.

Sur les phrases (figure 6.4b), en revanche, nous observons que les modèles Transformer sont significativement meilleurs que le TDNNF-HMM sur les mots ne contenant pas d'erreurs. L'utilisation de la fonction CTC, apportant une amélioration absolue du PER de 6,1%, permet de dépasser la performance du TDNNF sur les erreurs légères. En revanche, le TDNNF-HMM reste imbattable sur les mots contenant plus de deux erreurs.

Nous souhaitons maintenant analyser plus en détail l'impact de la fonction CTC sur deux aspects : la trop grande flexibilité des mécanismes d'attention, et l'influence du modèle de langage implicite appris par le décodeur. Nous prenons un enregistrement contenant plusieurs erreurs de lecture typiques. Alors que l'enfant devait lire « entre le pouce et le majeur, il y a l'index », il·elle lit les trois premiers mots « entre le pouce » en oubliant le phonème [ʁ] du mot « entre », puis se rend compte de son erreur et recommence au début de la phrase. Le dernier mot de la phrase contient également plusieurs erreurs : le [ɛ] du mot « index » est lu [e], et les phonèmes [k s] sont supprimés. Les transcriptions phonétiques de ce que l'enfant devait lire et ce qu'elle a vraiment lu se trouvent aux premières lignes du tableau 6.6.

TABLE 6.6 – Étude d'un enregistrement exemple, où l'enfant lit un énoncé avec plusieurs mots répétés et un mot contenant une erreur de lecture. Les hypothèses émises par le Transformer, et les trois sorties du Transformer+CTC (T+CTC) sont indiqués en phonétique.

Zone d'intérêt	1	2	3
Affiché à l'enfant	entre le pouce ɑ̃ t ʁ ə l ə p u s	et le majeur il y a e l ə m a ʒ ə ʁ i l i a	l'index l ɛ̃ d ɛ k s
Lu par l'enfant	ɑ̃ t l ə p u s ɑ̃ t ʁ ə l ə p u s	e l ə m a ʒ ə ʁ i l i a	l ɛ̃ d e
Transformer	v ɑ̃ t ʁ ə l ə p u s	e l ə m a ʒ ə ʁ i l i a	l ɛ̃ d ɛ k s
T+CTC enc	p ɑ̃ k - ə p u s ɑ̃ t ʁ ə l ə p u s	e e n ə m a ʒ o - - l i a	l ɛ̃ d ɛ
T+CTC dec	p ɑ̃ p l ə p u s ɑ̃ t ʁ ə l ə p u s	e l ə m a ʒ ə ʁ i l i a	l ɛ̃ d ɛ k s y ʁ l ɑ̃ d ʁ
T+CTC joint	p ɑ̃ k l ə p u s ɑ̃ t ʁ ə l ə p u s	e l ə m a ʒ o - - l i a	l ɛ̃ d ɛ

La figure 6.5 présente les transcriptions phonétiques du texte affiché, de ce qui a effectivement été lu par l'enfant, ainsi que les prédictions phonétiques faites par le modèle Transformer et la sortie décodeur du modèle Transformer+CTC. Les poids d'attention extraits des modules d'attention liant l'encodeur et le décodeur sont également affichés. Cette figure permet d'étudier l'effet de la fonction CTC pendant l'apprentissage uniquement, écartant l'effet du décodage hybride CTC/attention.

La première zone d'intérêt de cet exemple est le début de l'enregistrement : le Transformer+CTC dec réussit à prédire les phonèmes des mots répétés (malgré un [p] inséré et un [t] substitué par un [p], en rouge sur la figure 6.5), alors que le Transformer ignore complètement les mots répétés. Les poids d'attention montrent très clairement que le Transformer confond les trames audio correspondant aux deux occurrences de la séquence « entre le pouce » et considère une seule occurrence : ce phénomène est souligné par deux diagonales parallèles rouges sur la figure 6.5. Sur les poids d'attention du Transformer+CTC en revanche, une seule diagonale est visible, ce qui signifie que l'attention considère bien les deux occurrences comme séparées. Cette observation illustre que la fonction CTC contraint effectivement les mécanismes d'attention à la monotonie lors de l'entraînement du modèle Transformer+CTC.

Regardons les séquences de phonèmes prédites par les trois sorties du Transformer+CTC (tableau 6.6), afin d'étudier l'effet de la fonction CTC dans le décodage hybride CTC/attention. Nous voyons que la prédiction jointe est effectivement un mélange des prédictions des sorties

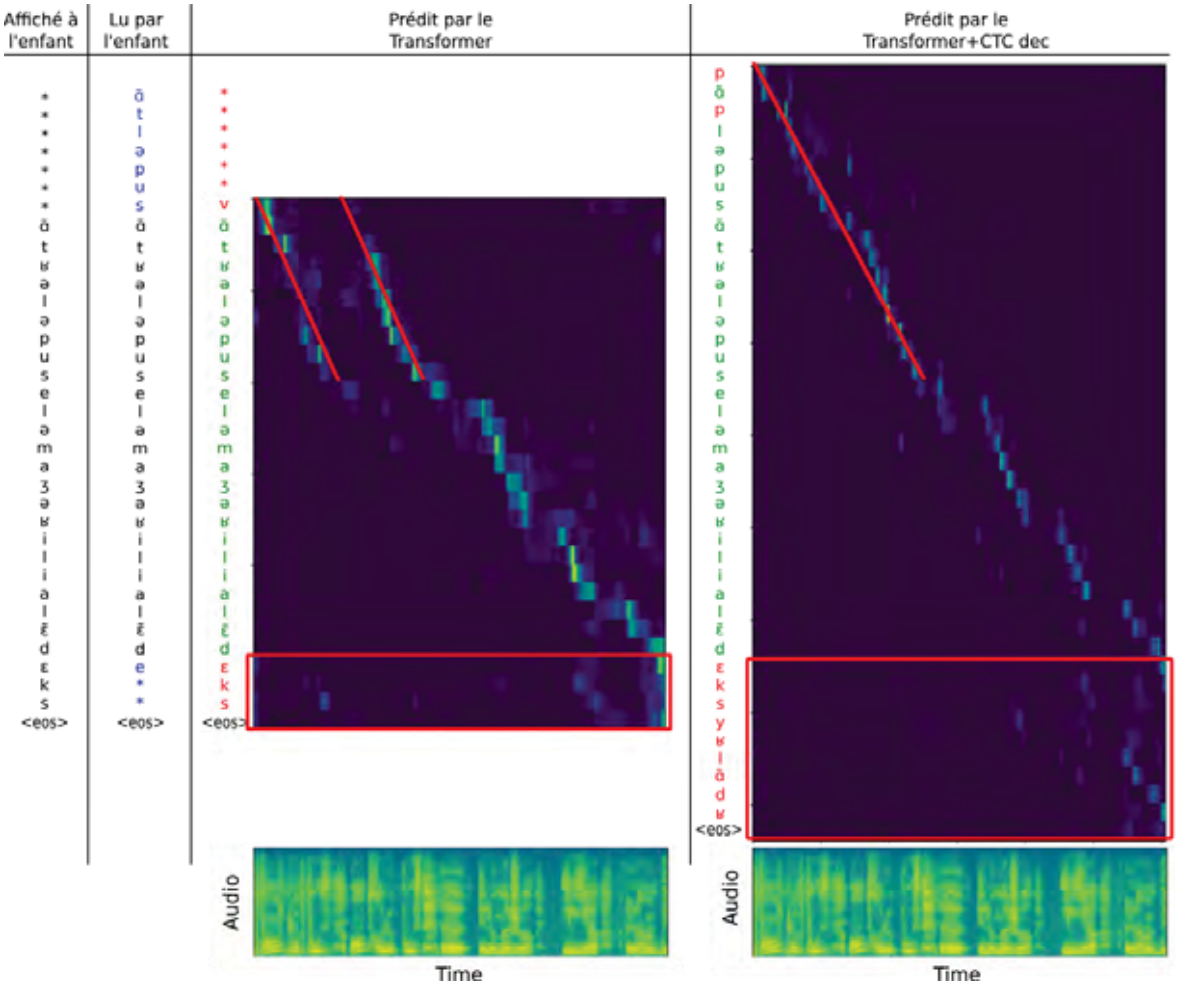
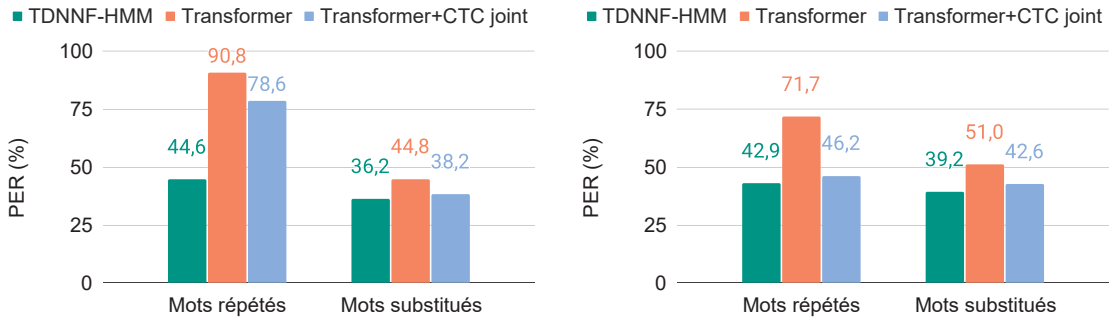


FIGURE 6.5 – Texte affiché à l'enfant, texte lu par l'enfant et transcriptions obtenues avec les modèles Transformer et Transformer+CTC (sortie du décodeur fondée sur l'attention) sur un enregistrement exemple. Les erreurs de lecture de l'enfant (répétitions, substitution et suppressions de phonèmes) sont en bleu ; les prédictions correctes et incorrectes dans les transcriptions phonétiques des modèles sont en vert et rouge, respectivement. Sont affichés également les poids d'attention extraits du module d'attention liant l'encodeur et le décodeur des modèles, ainsi que les spectrogrammes de l'enregistrement.

encodeur (utilisant la fonction CTC) et décodeur (utilisant l'attention). Le mélange est parfois heureux, parfois malheureux, et quelques erreurs de transcription de la CTC persistent dans la prédiction jointe (par exemple sur le mot « majeur » dans la zone n°2). Sur la zone d'intérêt n°1, la prédiction jointe est meilleure que celle de la sortie « enc », et équivalente à celle de la sortie « dec ».

Un bon support pour vérifier l'effet de cette flexibilité sur la parole d'apprenant·e·lecteur·rice·s sont les disfluences, puisque nous supposons que les mécanismes d'attention ont tendance à les ignorer. Nous affichons ainsi sur les figures 6.6a et 6.6b les performances des modèles TDNNF-HMM, Transformer et Transformer+CTC joint sur les mots répétés, provenant d'énoncés de mots isolés et de phrases, respectivement. Les scores PER du Transformer sont catastrophiques sur les mots répétés (90,8% et 71,7% sur les mots isolés et les phrases, respectivement) en comparaison avec ses scores obtenus sur les mots correctement lus (31,4% et 26,6%, voir figure 6.4). L'utilisation de la fonction CTC améliore effectivement les performances, de façon drastique sur les phrases (-35,6% relatifs) mais plus modérée sur les mots isolés (-13,4% relatifs). Le TDNNF-HMM reste cependant significativement meilleur. Ces observations semblent confirmer que les mécanismes d'attention sont trop flexibles pour une correcte détection de disfluences comme les répétitions de mots, et que la fonction CTC contraint effectivement ces mécanismes à être plus monotones.



(a) Mots isolés

(b) Phrases

FIGURE 6.6 – PER (%) des modèles TDNNF-HMM, Transformer et Transformer+CTC pour chaque mot répété ou substitué, provenant d'énoncés (a) de mots isolés (Test M) (b) de phrases (Test P)

Revenons à l'exemple du tableau 6.6, et plus précisément à la zone d'intérêt n°3. L'enfant y émet un mot non existant [ɛ̃ d e] à la place du mot « index » ([ɛ̃ d ɛ k s]). Le Transformer ne détecte pas l'erreur de lecture et prédit le mot existant, alors que l'enregistrement ne contient pas les phonèmes [k s]. Nous voyons sur la figure 6.5 que les poids d'attention du Transformer à l'intérieur du rectangle rouge sont d'ailleurs éparpillés sur les trames audio de l'enregistrement, montrant la confusion du module d'attention. Nous supposons que ce phénomène est dû au modèle de langage implicite que le décodeur Transformer apprend à partir des transcriptions phonétiques de référence. En effet, la quantité de contenu de Lalilo n'étant pas infinie, la phrase « entre le pouce et le majeur il y a l'index » est présente trois fois dans l'apprentissage, lue correctement par d'autres élèves. Le modèle a ainsi appris la suite de

phonèmes par cœur et l'a ensuite prédite lors de l'inférence malgré la présence d'une erreur de lecture. La prédiction du Transformer+CTC dec n'est malheureusement pas meilleure : non seulement le mot existant « index » est prédit alors qu'il n'a pas été prononcé, mais le modèle prédit des phonèmes additionnels, qui peuvent être assimilés à des mots existants vus pendant l'apprentissage également ([s y ʁ] donnant le mot très courant « sur », [ã d ʁ] comme la fin de plusieurs verbes : prendre, apprendre...). Nous voyons de plus que les poids d'attention sont dispersés dans la zone délimitée par le rectangle rouge, même s'ils le sont moins que ceux du Transformer. Ces observations suggèrent que malgré l'entraînement multi-objectifs avec la fonction CTC, l'influence du modèle de langage implicite est encore importante.

Étudions également l'effet du décodage hybride CTC/attention en comparant les transcriptions phonétiques du tableau 6.6 sur cette zone n°3. Nous voyons que le décodage joint a su favoriser sur le mot « index » la prédiction d'une séquence de phonèmes qui dévoile une erreur de lecture : [ɛ̃dɛ] et non [ɛ̃dɛks]. Même si la séquence prédite n'est pas exactement la bonne, il est primordial pour le module de détection d'erreurs du système global de Lalilo (figure 1.12) que la reconnaissance de phonèmes ne prédise pas un mot correct alors qu'une erreur est présente.

De la même façon que pour les mots répétés, nous étudions les mots substitués sur les figures 6.6a et 6.6b : cela nous permet de vérifier l'influence supposée du modèle de langage implicite appris par le décodeur du Transformer sur les substitutions de mots. Il apparaît que le Transformer est significativement moins performant sur les mots substitués que sur les mots correctement lus, et que ce phénomène est plus important sur les phrases (dégradation absolue de 32,6%) que sur les mots isolés (13,4%). En comparaison, les dégradations subies par le TDNNF-HMM sont plus légères : -12,6% sur les phrases et -7,1% sur les mots. Nous supposons ce phénomène dû au modèle de langage implicite, dont l'effet est plus notable sur les phrases de par le contexte plus large dont bénéficie le modèle. L'utilisation de la fonction CTC semble limiter cet effet, mais le Transformer+CTC reste derrière le TDNNF-HMM sur les mots substitués.

6.4 Bilan

Ce chapitre avait pour objectif l'évaluation des différentes méthodes et architectures, présentées au chapitre précédent, pour la reconnaissance de parole d'enfants AL. Nous avons dans un premier temps présenté le corpus *Lalil-o-fficiel* utilisé pour les expériences, qui a été complété avec des phrases courtes suite à l'évolution de l'exercice de lecture orale de Lalilo. Nous comparons ensuite les différents modèles, entraînés sur de la parole d'adultes (~150 heures), d'enfants (~13 heures), ou les deux successivement grâce à un apprentissage par transfert. Nos résultats démontrent que cette dernière option donne les meilleurs résultats sur un jeu de test de parole d'enfants (~1,5 heures), étant donné notre quantité de données et l'âge de nos locuteur·rice·s. Nous nous concentrons ensuite sur l'application Lalilo et ses spécificités. Une comparaison des modèles est faite en fonction de la tâche de lecture proposée par l'exercice de lecture orale (mots isolés ou phrases). Nous analysons enfin les performances

de nos meilleurs modèles en fonction des vitesses de lecture, très hétérogènes, car liées au niveau de lecture de l'enfant, et de la présence de disfluences et d'erreurs de lecture, fréquentes dans la parole d'AL.

Nos différentes analyses ont fait ressortir les avantages et inconvénients des différents paradigmes présentés au chapitre précédent : réseaux de neurones récurrents, fonction CTC et mécanismes d'attention. Nous observons que les mécanismes d'auto-attention sont plus performants que les RNN pour l'extraction d'information acoustique ou textuelle. Les mécanismes d'attention subissent cependant le manque de contexte sur des énoncés très courts, alors que les réseaux TDNNF, RNN et la fonction CTC sont moins affectés. Les mécanismes d'attention font en outre preuve d'une très grande flexibilité : elle octroie de bonnes performances sur des lectures correctes mais implique une fâcheuse tendance à ignorer des disfluences comme les répétitions de mots, évènements courants dans la parole d'AL. Cette flexibilité peut être canalisée par un entraînement multi-objectifs CE+CTC, qui impose une contrainte de monotonie aux mécanismes d'attention. Nous observons également que les décodeurs des modèles *seq2seq* apprennent un modèle de langage implicite à partir des séquences de phonèmes vues en entraînement : ce modèle est bénéfique pour la reconnaissance de mots correctement lus, mais tend à gommer de potentielles erreurs de lectures lorsque le mot substitué ne lui est pas familier. Cet effet peut également être réfréné par l'entraînement multi-objectifs CE+CTC. Enfin, nous montrons que le décodage joint CTC/attention est très performant, notamment sur les phrases, où les informations apportées par la CTC et l'attention sont complémentaires. Ces résultats ont été communiqués dans un article de la revue internationale *Speech Communication* [Gelin 2021a].

Ce chapitre nous a ainsi permis de déterminer le meilleur modèle pour notre application, le Transformer+CTC avec décodage joint CTC/attention, obtenant un PER de 25,0%. Il surpasse ainsi largement le modèle de référence hybride TDNNF-HMM (PER 30,1%). Bien qu'il reste très légèrement moins performant que le TDNNF-HMM sur la tâche de lecture de mots isolés (32,7% *vs.* 31,7%), il est drastiquement meilleur que ce dernier sur la tâche de lecture de phrases (19,6% *vs.* 28,9%), et le contexte réel d'application nous pousse à faire le choix d'un modèle unique. Nous observons de plus qu'il est plus robuste à des vitesses de lecture hétérogènes, grâce à la fonction CTC qui améliore fortement les performances du Transformer sur les lectures lentes. Notre analyse de comportement en présence d'erreurs de lecture nous montre premièrement que le Transformer+CTC est significativement meilleur que le TDNNF-HMM sur les mots ne contenant pas d'erreurs. Deuxièmement, nous observons que malgré l'apport important de la fonction CTC quant aux mots erronés, le Transformer+CTC reste légèrement moins performant que le TDNNF-HMM sur ces mots. Des améliorations, que nous tenterons d'apporter dans le chapitre 7, sont donc nécessaires pour envisager l'utilisation du Transformer+CTC dans l'exercice de lecture orale de Lalilo, puisqu'il est primordial que le système de reconnaissance de phonème soit capable d'identifier correctement les mots contenant des erreurs de lecture.

Chapitre 7

Amélioration de la robustesse aux erreurs de lecture et au bruit

Le chapitre précédent nous a permis d'évaluer différentes architectures *end-to-end* sur notre tâche de reconnaissance de phonèmes sur parole d'enfant, et de choisir la meilleure pour la suite de nos études : le Transformer+CTC avec décodage hybride CTC/attention. Nous avons étudié ses performances quant aux spécificités de notre application, liées à la parole d'apprenant·e·s lecteur·rice·s. Nous avons observé, en particulier, des performances dégradées lorsque le mot à reconnaître contient une ou plusieurs erreurs de lecture. Bien que présent également pour le TDNNF-HMM, ce phénomène est accentué dans le cas du Transformer de par la grande flexibilité des mécanismes d'attention et l'effet du modèle de langage implicite appris par le décodeur. L'utilisation de la fonction CTC améliore significativement les performances en présence d'erreurs de lecture, grâce à ses capacités à contraindre les mécanismes d'attention à la monotonie et à diminuer l'influence du modèle de langage implicite. Cependant, le Transformer+CTC reste très impacté. Cela est sans aucun doute lié au fait que les données d'entraînement ne contiennent pas d'erreurs de lecture, coûteuses à annoter.

Ce chapitre présente ainsi des techniques d'augmentation de données visant à améliorer la robustesse du modèle Transformer+CTC à la parole d'AL, ainsi qu'au bruit de salle de classe. Nous commençons par appliquer l'augmentation de données par ajout de bruit de brouhaha qui a apporté des améliorations significatives sur le TDNNF-HMM au chapitre 4. Nous introduisons ensuite une technique novatrice d'augmentation de données par simulation d'erreurs de lecture. Nous décrivons la procédure utilisée pour générer ces erreurs synthétiques, puis évaluons son efficacité sur les mots contenant des erreurs de lecture, tout en vérifiant que cela ne dégrade pas la reconnaissance des mots correctement lus. Enfin, nous combinons les deux augmentations pour étudier leur complémentarité.

Sommaire

7.1 Augmentation par ajout de bruit de brouhaha

Les scores de PER, vus au chapitre précédent dans le tableau 6.4, peuvent sembler encore élevés : il faut cependant garder en tête que nos données, de par leur enregistrement en salle de classe, peuvent contenir de hauts niveaux de bruits de brouhaha, typique des environnements où évoluent des enfants. Les RSB moyens pour les ensembles d'entraînement, de validation et de test sont respectivement 20,9 dB, 20,1 dB et 20,6 dB, avec des écarts types de 13,0, 12,6 et 12,6 dB (tableau 6.1). Certains enregistrements dévoilent même un RSB proche de 0 dB.

Dans cette section, nous appliquons la méthode d'augmentation de données par ajout de bruit de brouhaha présentée au chapitre 3 et utilisée sur le modèle de référence au chapitre 4, section 4.3.5. Cette méthode a montré de fortes améliorations de la robustesse du modèle TDNNF-HMM au bruit typiquement présent sur des enregistrements réalisés en salle de classe (gain relatif de 6,4% sur tous les fichiers, et de 15,2% sur les fichiers très bruités). Nous l'appliquons ici à notre modèle Transformer+CTC, ce qui nous permet d'étudier son efficacité sur une architecture *end-to-end*

De la même façon qu'au chapitre 4, nous entraînons deux modèles TL en appliquant de l'augmentation de bruit sur les données de parole d'enfants. Nous utilisons deux sources de bruit : le corpus DEMAND [Thiemann 2013] de bruit de brouhaha adulte, et le corpus de bruit de salles de classe de *Lali-noise*. Le tableau 7.1 présente les scores PER obtenus sur le jeu de test Lalilo avec et sans augmentation de bruit. Nous observons que les deux augmentations apportent une amélioration du PER, notamment sur les phrases, où l'amélioration est significative. La plus efficace, sur les mots et phrases combinés, semble être le bruit de salles de classe *Lali-noise*, obtenant un PER de 23,3%.

Table 7.1 – PER (%) obtenus par le Transformer+CTC TL, avec et sans augmentation de données par ajout de bruit de brouhaha, sur les mots isolés (Test M), les phrases (Test P) et les deux combinés (Test M+P)

Modèle	Test M	Test P	Test M+P
Sans augmentation	32,7	19,6	25,0
DEMAND-aug	33,0	17,2	23,7
Lali-noise-aug	32,2	17,1	**23,3**

La figure 7.1 présente les performances de ces modèles en fonction du niveau de bruit grâce à trois intervalles de RSB : RSB $\leq$ 10 dB (très bruité), 10 dB $<$ RSB $\leq$ 20 dB (moyennement bruité) et RSB $>$ 20 dB (non bruité). Nous observons tout d'abord que, comme pour le TDNNF-HMM, la présence de bruit a un fort impact sur les performances du Transformer+CTC : une perte absolue d'environ 10% de PER se remarque entre chaque niveau de bruit, pour le modèle sans augmentation. Étonnamment, l'augmentation de bruit avec le corpus DEMAND n'a pas d'effet significatif sur les enregistrements très bruités, mais améliore de 7,2% relatifs le PER sur les enregistrements propres. L'augmentation de bruit *Lali-noise* offre au contraire des améliorations de 10,0% et 9,0% relatifs sur les enregistrements très bruités et moyennement

bruités, respectivement, sans dégrader les performances sur les enregistrements propres.

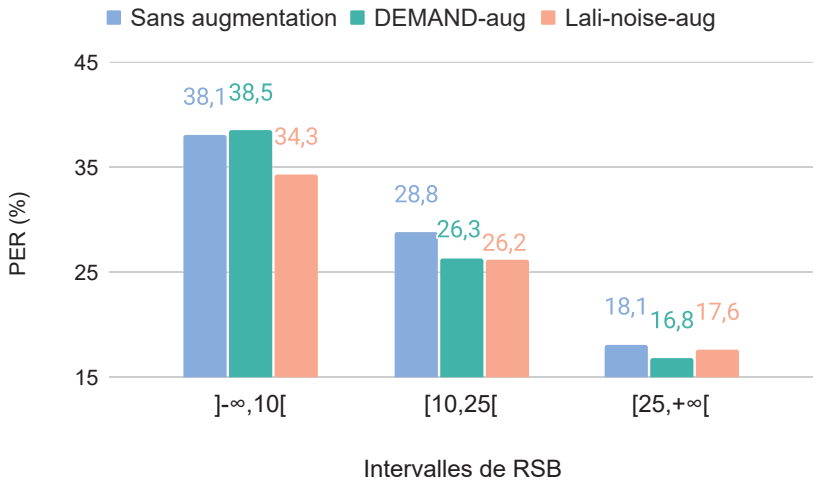


FIGURE 7.1 – PER (%) obtenus avec le modèle Transformer+CTC, avec et sans augmentation de bruit sur différents intervalles de RSB

L'ensemble de ces résultats montre que l'augmentation de données par ajout de bruit de brouhaha fonctionne sur notre modèle Transformer+CTC. La comparaison entre les deux sources de bruit suggère qu'un bruit correspondant à l'environnement sonore réel est plus adapté. Dans notre cas, où le bruit présent sur nos enregistrements est de la parole superposée d'enfants dans une salle de classe, il semble important d'augmenter les données avec du bruit de brouhaha provenant d'enfants du même âge, et non d'adulte.

7.2 Augmentation par simulation d'erreurs de lecture

7.2.1 Méthodes de simulation d'erreurs

Nous avons observé dans le chapitre 4 que les données reçues via la plateforme Lalilo contiennent des erreurs de lecture. Or, pour les raisons détaillées dans ce même chapitre, notre jeu d'entraînement ne contient que des enregistrements de lecture correcte : notre modèle n'est donc pas préparé à gérer les erreurs de lecture des enfants. En effet, la section 6.3.2.2 du chapitre précédent montrait que les performances des modèles étaient fortement dégradées (multiplication du PER par deux) en présence d'erreurs de lecture. Nous montrons ici comment créer des erreurs de lecture synthétiques à intégrer dans le jeu d'entraînement, dans l'optique d'améliorer la robustesse des modèles à la parole d'apprenant·e lecteur·rice. Nous présumons que l'augmentation par simulation d'erreurs agira sur deux leviers, identifiés au chapitre précédent :

— apprendre aux mécanismes d'attention à distinguer les mots répétés malgré leur grande flexibilité, en leur présentant des exemples de répétitions de mots durant l'entraînement ;

— limiter la tendance du modèle de langage implicite du décodeur à gommer les erreurs de lecture de l'enfant, en lui faisant apprendre des séquences de phonèmes plus diverses contenant des substitutions de mots et de phonèmes.

La table 7.2 présente les erreurs de lecture synthétiques créées par cette méthode : (a) répétitions de mots, (b) substitutions de mots et (c) substitutions de phonèmes. Nous présentons en détail la procédure de simulation de chacune de ces erreurs dans les sections suivantes.

TABLE 7.2 – Simulation d'erreurs de lecture sur la phrase : « il roule à vélo » [il ʁul a velo]

(a) Répétition de mot(s)

Type d'opération	Description
Répétition par motif	Un motif de mots est répété *Exemple :* [il ʁul a il ʁul a velo]
Répétition individuelle	Un ou plusieurs mots sont répétés individuellement *Exemple :* [il ʁul ʁul a velo velo]

(b) Substitution de mot

Type d'opération	Description
Substitution de voyelle	Une voyelle est substituée par une autre *Exemple :* [il ʁɑl a velo]
Substitution de consonne	Une consonne est substituée par une autre *Exemple :* [il bul a velo]
Inversion de phonèmes	Un mot à deux phonèmes est inversé *Exemple :* [li ʁul a velo]
Arrêt dans un mot	Seul le début du mot est lu. La coupure peut se faire au niveau des phonèmes ou des graphèmes *Exemple :* [il ʁu a velo] *ou* [il ʁo a velo]

(c) Substitution de phonème (création d'un mot inexistant)

Type d'opération	Description
Substitution de voyelle	Une voyelle est substituée par une autre *Example :* [il ʁyl a velo]
Substitution de consonne	Une consonne est substituée par une autre de la même famille (fricatives, plosives, liquides) *Example :* [il lul a velo]

Une étape préalable à toute simulation d'erreur est l'alignement forcé des données d'entraînement afin d'obtenir les frontières des segments d'intérêt. L'alignement peut être fait au niveau du mot ou au niveau du phonème. Deux modèles différents sont utilisés pour cela : un GMM-HMM entraîné à la reconnaissance de mots, et un TDNNF-HMM entraîné à la reconnaissance de phonèmes. Le GMM-HMM correspond au modèle utilisé pour le module d'alignement forcé du système global d'évaluation de lecture orale (figure 1.12). Le TDNNF-HMM est le

modèle de référence présenté dans le chapitre 3 et évalué dans le chapitre 6.

7.2.1.1 Répétition de mot(s)

Nous avons vu au chapitre 4 que les répétitions constituent une proportion non négligeable des erreurs de lecture rencontrées (3,5%, tableau 4.6), et que 80% des mots répétés le sont entièrement et sans erreur phonétique. Cela motive la création de répétitions de mots entiers correctement lus à partir des enregistrements contenus dans le jeu d'entraînement : nous ne jugeons pas nécessaire, dans un premier temps, de créer des mots répétés avec erreur de lecture, marginaux dans la distribution d'erreurs et plus complexes à générer artificiellement.

L'analyse des erreurs a mis en lumière deux types de répétitions (table 7.2(a)) : les répétitions par motif et les répétitions individuelles. Comme observé dans le chapitre 4, le premier type correspond plutôt à des élèves de niveau avancé, qui commencent tout juste à gagner en confiance, à comprendre le sens de ce qu'il·elle·s lisent et à faire attention à leur fluidité de lecture. Nous retrouvons le second type de répétitions plutôt chez les élèves débutant·e·s en lecture, qui butent sur les mots, ne comprennent pas le sens des mots qu'il·elle·s lisent et donc sont moins en capacité de s'auto-corriger.

Les données d'entraînement sont alignées au niveau du mot en amont, permettant l'obtention des frontières de chaque mot dans le signal audio. Pour augmenter une phrase contenant K mots avec des répétitions de mots, le type de répétition est choisi aléatoirement, puis, en fonction de ce choix, l'une des deux procédures suivantes est appliquée :

Répétition par motif

1. Choix aléatoire d'un mot de la phrase à la position $i \leq K$
2. Extraction des segments audio pour tous les mots de position $\leq i$
3. Concaténation des segments audio extraits au début du signal originel

Répétition individuelle

1. Choix aléatoire de $k = 1...K$ mots de la phrase
2. Extraction des segments audio correspondant à ces k mots
3. Insertion des segments audio dans le signal original en amont du segment correspondant au mot à répéter

7.2.1.2 Substitution de mot

Les substitutions de mots sont également très présentes dans nos données : il est courant qu'un enfant remplace un mot par un autre mot, que ce soit un mot ressemblant, une conjugaison différente du verbe, un mot du même champ lexical... Il est difficile de générer automatiquement des substitutions de mots par des mots complètement différents, donc nous nous concentrons ici sur la substitution de mots ressemblants. Nous établissons pour cela quatre types d'opérations possibles au niveau du phonème pour créer des mots ressemblants au mot à substituer (tableau 7.2(b)) : substitution de voyelle, substitution de consonne, inversion de phonèmes (seulement pour les mots contenant deux phonèmes) et faux départ.

Les substitutions de voyelle et de consonne sont uniquement des substitutions de phonèmes, et non de graphèmes. Les faux départs peuvent néanmoins être fondés sur les graphèmes ou sur les phonèmes : par exemple pour le mot « roule », un faux départ fondé sur les phonèmes serait [ʁu] et un faux départ fondé sur les graphèmes serait [ʁo].

À partir de ces quatre opérations, et pour chaque mot présent dans le jeu d'apprentissage, une liste de mots de substitution possibles est dressée. Par exemple, la liste du mot « roule » contiendra huit mots formés par substitution de voyelles (« râle, rôle, roula, roulé, roulait, rouleau, roulons, roulant »), sept mots par substitution de consonnes (« boule, moule, poule, foule, cool, route, rousse, rouge ») et de deux faux départs, dont un fondé sur les graphèmes (« rot ») et un fondé sur les phonèmes (« roux/roue »). La liste pour le mot « il » sera constituée d'un mot formé par substitution de voyelle (« elle/aile »), d'une inversion de phonèmes (« lit/lis ») et d'un faux départ (« y »). Cette liste ne peut contenir que des mots existants et présents dans le jeu de données d'entraînement. Les homophones sont regroupés, comme dans nos exemples ci-dessus, puisque seule la prononciation phonétique compte.

Les substitutions de mots sont plus complexes à simuler que les répétitions : les phrases lues par les enfants étant très courtes, elles ne contiennent quasiment jamais deux mots ressemblants, et il n'est donc pas possible de créer des substitutions de mots à partir d'un seul enregistrement. Une solution serait d'utiliser les enregistrements d'un·e même locuteur·rice et de chercher quelles substitutions sont possibles à partir de ces enregistrements. Cependant, notre jeu d'entraînement contient 3014 locuteur·rice·s pour 13 heures de parole, correspondant à 15,2 secondes de parole par locuteur·rice. Cela limite fortement les possibilités de substitutions. Dans notre méthode, nous utilisons donc toutes les données d'entraînement pour générer des substitutions de mots, sans distinction de locuteur·rice. Nous avons cependant limité le nombre de mots substitués à un par enregistrement pour éviter de trop importants décalages d'identité vocale ou d'environnement sonore. Nous avons également appliqué une normalisation d'énergie sur les segments de substitution pour adoucir les transitions.

Les étapes de simulation de substitutions de mots sont les suivantes :

1. Choix aléatoire d'un mot à substituer dans la phrase
2. Choix aléatoire d'un mot de substitution parmi la liste de substitutions possibles pour ce mot
3. Choix aléatoire d'un enregistrement du jeu d'entraînement contenant le mot de substitution, et extraction du segment audio correspondant à ce mot
4. Remplacement du segment à substituer par le segment de substitution

7.2.1.3 Substitution de phonème

Les substitutions de phonèmes sont les plus délicates à effectuer. L'objectif est de substituer un phonème contenu dans un mot par un autre phonème contenu dans l'enregistrement, afin d'avoir la possibilité de créer des mots non existants dans la langue française (voir tableau 7.2(c)). Cela nécessite donc un alignement très précis afin que l'enregistrement corresponde effectivement à la transcription phonétique générée. Nous utilisons pour cela

le TDNNF-HMM évalué au chapitre 6, qui génère de meilleurs alignements qu'un modèle GMM-HMM. La substitution d'un phonème se fait avec un phonème de la même famille afin de garder une cohérence linguistique. Les familles sont définies ainsi :

— Voyelles (orales et nasales confondues) : [a, ɑ, e, ɛ, ə, i, œ, o, ɔ, u, y, ã, ɛ̃, œ̃, ɔ̃]

— Consonnes plosives : [p, t, k, b, d, ɡ]

— Consonnes fricatives : [f, s, ʃ, v, z, ʒ]

— Consonnes liquides et nasales : [l, ʁ, m, n, ŋ, ɲ]

— Semi-voyelles : [j, w, ɥ]

À cause du nombre limité d'occurrences de semi-voyelles dans notre jeu d'entraînement, nous bloquons les substitutions de ces dernières afin d'éviter le sur-apprentissage sur ces quelques segments. Nous limitons également l'augmentation à une seule substitution de phonème par enregistrement.

Les étapes pour l'augmentation d'un enregistrement avec une substitution de phonème sont les mêmes que pour les substitutions de mots, à la différence que nous opérons au niveau du phonème, et que le phonème de substitution est extrait obligatoirement du même enregistrement. Si, pour le phonème à substituer choisi, aucun phonème de la même famille n'est présent dans la phrase, nous réitérons le processus jusqu'à trouver une substitution possible.

7.2.2 Expériences et évaluation globale

Les enregistrements sur lesquels appliquer l'augmentation sont sélectionnés. Les histoires sont mises de côté pour le moment, par crainte que de mauvais alignements altèrent la qualité de l'augmentation. En effet, non seulement les longs enregistrements tendent à être plus difficile à aligner pour le système de RAP [Katsamanis 2011], mais nous avons observé que les enfants AL font plus d'erreurs de fluence sur des textes longs, ce qui dégrade encore la qualité des alignements. Les données à augmenter sont présentées dans le tableau 7.3 : 1221 phrases courtes contenant en moyenne 5,8 mots, et environ 15 000 mots isolés. Chaque enregistrement de phrase est augmenté trois fois, avec chaque type d'augmentation (répétition de mots, substitution de mot, substitution de phonème). Les enregistrements de mot isolé ne sont augmentés que deux fois, car il est inutile de faire une substitution de mot sur un mot isolé. Nous avons fait le choix de ne pas combiner les différents types d'erreurs synthétiques, pour ne pas compliquer inutilement la procédure d'augmentation. Cela implique que les mots répétés créés sont tous correctement lus, ce qui est en phase avec les observations faites au chapitre 4 : la grande majorité des mots répétés ne contiennent pas d'erreur de lecture.

Nous nous appuyons sur les observations faites au chapitre 4 pour déterminer les proportions d'erreurs synthétiques à ajouter au jeu d'entraînement. Des sous-ensembles de données pour chaque type d'augmentation sont créés aléatoirement, puis regroupés pour correspondre à des proportions d'erreurs entre 1 et 10% (27 combinaisons testées). Ces divers sous-ensembles, concaténés au jeu d'entraînement originel, sont ensuite utilisés pour entraîner un modèle Transformer+CTC.

TABLE 7.3 – Description des données disponibles augmentées avec la méthode ErrSyn

Type de contenu	Nombre de données initiales	Répétition de mot(s)	Substitution de mot	Substitution de phonème	Nombre de données finales
Mot isolé	15k	1	0	1	45k
Phrase	1221	1	1	1	4884

Le corpus de validation est également augmenté. De cette façon, il correspond mieux au contenu de l'ensemble de test, ce qui limite un potentiel décalage de performance entre les jeux de validation et de test. Dans le cadre de cette étude, cela nous permet de choisir un modèle en fonction de ses capacités à correctement transcrire phonétiquement des mots contenant des erreurs de lecture.

Le modèle avec le meilleur PER sur cet ensemble de validation est choisi. Il a été entraîné sur 6,0% de répétitions de mot(s) (dont 3,5% sur des phrases et 2,5% sur des mots isolés), 1,5% de substitutions de mot (uniquement sur des phrases) et 3,5% de substitutions de phonème (dont 2% sur des phrases et 1,5% sur des mots isolés).

Les résultats de ce modèle, baptisé ErrSyn-aug pour « Erreurs synthétiques », sont présentés dans la table 7.4, en comparaison avec le même modèle non augmenté. Nous voyons que l'augmentation de données améliore le PER à la fois sur les mots isolés (Test M) et sur les phrases (Test P). Le modèle Transformer+CTC aug obtient un PER global de 23,9%, soit une amélioration relative de 4,4% par rapport au modèle sans augmentation.

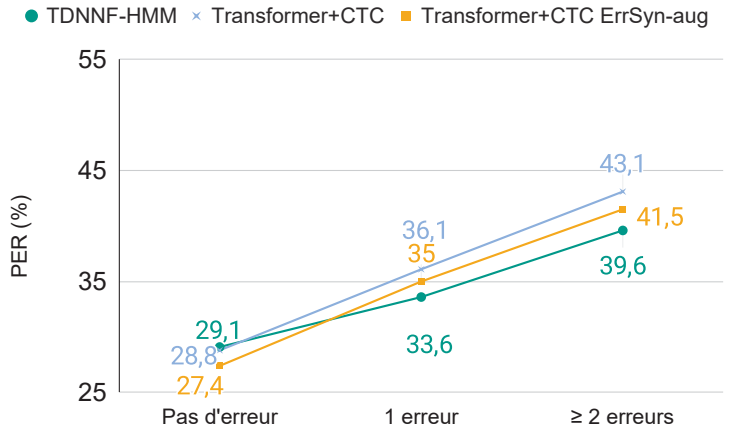
TABLE 7.4 – PER (%) obtenus par le Transformer+CTC TL, avec et sans augmentation par simulation d'erreurs de lecture, sur les mots isolés (Test M), les phrases (Test P) et les deux combinés (Test M+P)

Jeu de test	Test M	Test P	Test M+P
Sans augmentation	32,7	19,6	25,0
ErrSyn-aug	31,2	18,7	**23,9**

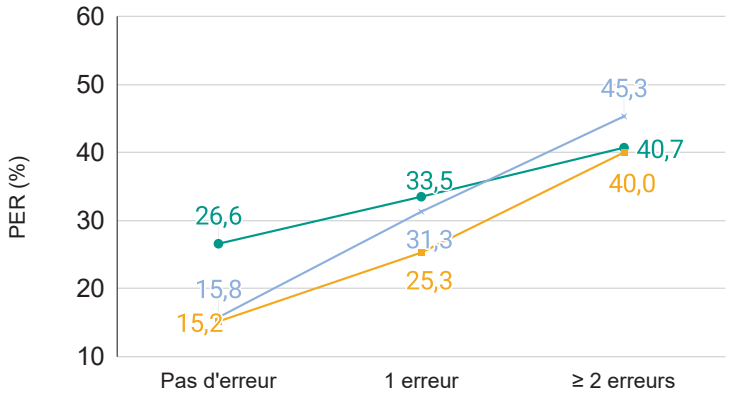
7.2.3 Évaluation de la robustesse aux erreurs de lecture

De la même façon que dans le chapitre précédent (section 6.3.2.2), nous observons l'influence de la présence d'erreurs de lecture sur la précision de nos modèles, et additionnellement évaluons l'impact de l'augmentation de données par simulation d'erreurs synthétiques sur cette précision. Pour cela, nous reprenons les figures 6.4a et 6.4b du chapitre précédent en rajoutant le modèle Transformer+CTC ErrSyn-aug, et obtenons les figures 7.2a et 7.2b. Nous rappelons que sur ces graphiques, seules sont comptées les erreurs au niveau du phonème dans un mot : les mots répétés sont considérés en fonction du nombre d'erreurs phonétiques (substitution, insertion ou suppression) qu'ils contiennent, et peuvent donc appartenir aux catégories 0, 1 ou ≥ 2 erreurs. De la même façon, les figures 7.3a et 7.3b reprennent les figures 6.6a et 6.6b du chapitre

précédent en affichant les scores des modèles sur les mots répétés uniquement, puis sur les mots substitués uniquement.



(a) Mots isolés



(b) Phrases

FIGURE 7.2 – PER (%) des modèles TDNNF-HMM, Transformer+CTC et Transformer+CTC ErrSyn-aug pour chaque mot en fonction du nombre d'erreurs de lecture contenues dans ce mot, provenant d'énoncés (a) de mots isolés (Test M) (b) de phrases (Test P)

Étudions d'abord les mots isolés : nous observons sur la figure 7.2a que l'impact de l'augmentation ErrSyn sur les mots contenant des erreurs de lecture est modéré et presque constant selon le nombre d'erreurs. En fait, l'amélioration est même légèrement plus importante sur les mots ne contenant pas d'erreurs, ce qui est au premier abord inattendu. En effet, il ne faut pas oublier que la plupart des mots répétés sont lus correctement (par exemple, « pouce » est lu [pus pus]) : la proportion de mots répétés sans erreur est d'environ 80% dans notre échantillon étudié au chapitre 4, et d'environ 71% dans notre corpus Test M. La majeure partie des mots répétés sont donc inclus dans les mots ne contenant pas d'erreur. Ainsi, une partie de l'amélioration observée sur la catégorie « Pas d'erreur » peut être due à une amélioration de la reconnaissance sur les mots répétés. Cette hypothèse peut être vérifiée en observant la figure 7.3a : nous y voyons une nette amélioration (15,8% relatifs) sur les mots

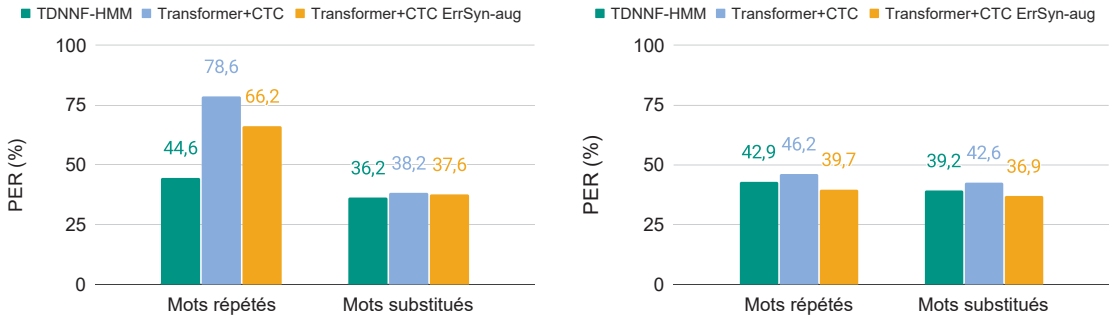


FIGURE 7.3 – PER (%) des modèles TDNNF-HMM, Transformer+CTC et Transformer+CTC ErrSyn-aug pour chaque mot répété ou substitué, provenant d'énoncés (a) de mots isolés (Test M) (b) de phrases (Test P)

répétés entre le modèle Transformer+CTC et le modèle Transformer+CTC ErrSyn-aug. Une étude plus poussée où nous séparons les mots répétés selon les catégories de la figure 7.2 nous apprend que l'amélioration apportée par l'augmentation ErrSyn est plus marquée sur les mots répétés ne contenant pas d'erreur phonétique. Ce résultat n'est pas étonnant puisque notre procédure d'augmentation consiste à créer uniquement des mots répétés correctement lus. La figure 7.3a nous montre également que l'augmentation ErrSyn n'apporte pas d'amélioration significative sur les mots substitués, ce qui corrobore les résultats modérés observés pour les mots contenant au moins une erreur de lecture sur la figure 7.2a. Cela peut également être le signe que l'influence du modèle de langage implicite n'est pas très forte sur les mots isolés puisque le contexte est étroit, et donc l'augmentation par ajout de substitutions de mots, supposée apporter de la diversité à ce modèle de langage, n'a que peu d'impact sur les mots isolés. Sur le corpus Test M, le TDNNF-HMM reste meilleur que le Transformer+CTC ErrSyn-aug sur les mots répétés, les mots substitués, et plus généralement les mots contenant au moins une erreur de lecture au niveau du phonème. Des expériences ont été menées avec de plus grandes quantités de mots isolés contenant des erreurs synthétiques dans le corpus d'entraînement, mais n'ont pas été davantage concluantes.

Sur les phrases, nous voyons grâce à la figure 7.2b que l'augmentation ErrSyn a été extrêmement efficace sur les mots erronés. Les performances du Transformer+CTC ErrSyn-aug surpassent de 19,2% et 11,7% relatifs celles du Transformer+CTC sur les mots à une et plus de deux erreurs, respectivement. Nous voyons également que l'augmentation a permis au Transformer+CTC ErrSyn-aug de supprimer l'écart résiduel avec le TDNNF-HMM sur les mots avec plus de deux erreurs. Le Transformer+CTC ErrSyn-aug obtient ainsi les meilleurs scores PER, peu importe le nombre d'erreurs de lecture contenues dans le mot.

La figure 7.3b nous montre que l'augmentation ErrSyn a un effet significatif sur les mots répétés (-14,1% relatifs), ce qui permet au Transformer+CTC ErrSyn-aug de dépasser le score du TDNNF-HMM de 7,5% relatifs. Ce résultat porte à croire que les mécanismes d'attention sont en effet mieux entraînés à distinguer les répétitions de séquences de phonèmes. Contrairement à ce qui a été observé sur les mots isolés, l'augmentation apporte une amélioration significative

(13,4% relatifs) sur les mots substitués, et surpasse une nouvelle fois le TDNNF-HMM de 5,9% relatifs. Ce phénomène peut être expliqué par plusieurs facteurs : la proportion de substitutions de mots et de phonèmes (1,5% + 2%) sur les phrases est plus élevée que sur les mots isolés (1,5%). De plus, nous n'avons simulé sur les mots isolés que des substitutions de phonèmes : ce type d'erreurs synthétiques est plus difficile à générer, car nécessitant des alignements très précis, ce qui peut en dégrader la qualité et les rendre moins efficaces que les substitutions de mots. Enfin, le modèle de langage implicite appris par le décodeur a sûrement plus d'impact sur les phrases que sur les mots, grâce à l'exploitation d'un contexte plus large. L'amélioration observée sur les mots substitués peut donc être le signe que l'augmentation ErrSyn diversifie en effet les séquences apprises par le modèle de langage et lui permet de mieux reconnaître les mots substitués.

Ces résultats montrent que notre augmentation novatrice par simulation d'erreurs de lecture a effectivement amélioré la robustesse de notre modèle aux erreurs de lecture, et lui a permis de surpasser le modèle TDNNF-HMM sur une grande partie des données. Les améliorations significatives sur les mots répétés et substitués montrent que l'augmentation ErrSyn a effectivement agi sur les deux leviers identifiés : les mécanismes d'attention sont plus habitués à rencontrer des répétitions de mots et les distinguent ainsi mieux, et le modèle de langage implicite du décodeur a appris des séquences de phonèmes plus diversifiées et peut ainsi mieux reconnaître les substitutions effectuées par les enfants AL.

7.3 Augmentations combinées

Nous avons vu aux sections précédentes que les deux techniques d'augmentation proposées sont efficaces sur le modèle Transformer+CTC : l'augmentation par ajout de bruit de salles de classe *Lali-noise* réduit le PER de 6,8% relatifs, et l'augmentation par simulation d'erreurs de lecture de 4,4%. Ces deux méthodes permettent de plus d'améliorer la robustesse du modèle, d'une part à l'environnement de salle de classe, et d'autre part à la présence d'erreurs de lecture. Cette section est donc consacrée à l'évaluation d'un modèle sur lequel sont appliquées ces deux méthodes de façon combinée.

Pour conserver la même proportion d'erreurs de lecture dans le corpus d'apprentissage, nous appliquons l'augmentation par ajout de bruit sur les enregistrements contenant des erreurs de lectures synthétiques. Le modèle est ainsi entraîné sur les enregistrements originaux, les enregistrements originaux bruités, les enregistrements contenant des erreurs de lecture synthétiques, et ces derniers enregistrements auxquels du bruit a été ajouté. Puisqu'il a été vu à la section 7.1 que le bruit du corpus DEMAND était beaucoup moins efficace que le bruit de salle de classe du corpus *Lali-noise*, nous utilisons uniquement ce dernier.

Le tableau 7.5 reporte les résultats des sections précédentes, obtenus grâce à chaque augmentation appliquée séparément, auxquels sont ajoutés les scores PER obtenus par le modèle entraîné avec les deux augmentations combinées. Nous notons une amélioration significative (Wilcoxon, $p = 0{,}03$) par rapport au modèle initial sans augmentation (15,2% relatifs sur Test M+P). Le PER est également réduit par rapport aux modèles entraînés avec

l'augmentation par ajout de bruit (9,0% relatifs) et par simulation d'erreurs (11,3% relatifs) séparément, ce qui montre la complémentarité des deux augmentations. Nous observons par ailleurs que l'amélioration est plus importante lorsque l'augmentation ErrSyn est appliquée au modèle *Lali-noise*-aug (-2,1%) qu'au modèle sans augmentation (-1,1%), alors que la proportion d'erreurs synthétiques est la même. Ce phénomène peut s'expliquer par le fait qu'un enregistrement original et un enregistrement ErrSyn seront augmentés avec différents enregistrements de bruit, ce qui accroît la diversité du corpus.

TABLE 7.5 – PER (%) obtenus par le Transformer+CTC, avec augmentation par ajout de bruit *Lali-noise* et augmentation par simulation d'erreurs de lecture ErrSyn sur les mots isolés (Test M), les phrases (Test P) et les deux combinés (Test)

Jeu de test	Test M	Test P	Test
Sans augmentation	32,7	19,6	25,0
Lali-noise-aug	32,2	17,1	23,3
ErrSyn-aug	31,2	18,7	23,9
Lali-noise-aug + ErrSyn-aug	**29,9**	**15,1**	**21,2**

7.4 Bilan

Ce dernier chapitre portait sur deux méthodes d'augmentation de données : ne possédant que 13 heures de données de parole d'enfants, ces méthodes ont pour objectif d'améliorer les performances globales du modèle Transformer+CTC, mais également ses performances sur des évènements spécifiques à la parole d'enfants AL en environnement de salle de classe.

Nous avons d'abord présenté les résultats obtenus grâce à l'augmentation par ajout de bruit de brouhaha : ayant démontré son efficacité pour l'amélioration de la robustesse aux bruits de salle de classe du TDNNF-HMM au chapitre 4, nous l'avons appliqué ici au Transformer+CTC. Nos résultats indiquent des améliorations grâce aux deux types de bruit (bruit de brouhaha d'adultes du corpus DEMAND, et bruit de salle de classe avec parole superposée d'enfants du corpus *Lali-noise*), mais plus importantes avec le second type. L'augmentation Lali-noise est en outre la seule à apporter une amélioration sur les enregistrements très bruités (RSB $\leq$ 10 dB), permettant au modèle de s'entraîner à reconnaître la parole dans un environnement proche de la réalité.

Nous avons ensuite introduit une méthode novatrice d'augmentation de données par simulation d'erreurs de lecture, visant à améliorer la robustesse du modèle en présence de ces dernières, courantes dans la parole d'apprenant·e·s lecteur·rice·s. Nous avons présenté les procédures de création de répétitions de mots, de substitutions de mots et de phonèmes, et l'entraînement de notre modèle sur un corpus auquel sont ajoutées ces erreurs de lecture. Nos résultats montrent que cette méthode apporte une réduction globale du PER, et permet une meilleure reconnaissance des mots contenant des erreurs. Grâce à des analyses poussées, nous suggérons que cette augmentation agit effectivement sur la capacité des mécanismes

d'attention à distinguer les mots répétés, ainsi que sur la diversification des séquences de phonèmes apprises par le modèle de langage implicite du décodeur, lui permettant de mieux détecter de potentielles substitutions, insertions ou suppressions de phonèmes. Nous avons montré que cette méthode permet au Transformer+CTC de rattraper les performances du TDNNF-HMM sur les mots contenant des erreurs de lecture, tout en restant largement meilleur sur les mots correctement lus, et le rend ainsi éligible à une utilisation dans la plateforme Lalilo. L'utilité de cette méthode d'augmentation pourrait s'étendre au-delà de notre application : détection de troubles de langage, apprentissage d'une seconde langue, évaluation de l'expression orale...

Enfin, nous avons montré que ces deux techniques d'augmentation sont complémentaires : un Transformer+CTC entraîné sur des données combinant ajout de bruit et simulation d'erreurs de lecture obtient de meilleurs scores à la fois sur les mots isolés (8,6% d'amélioration relative par rapport au Transformer+CTC originel) et les phrases (23,0% relatifs). Sur les deux jeux de test combinés, ce modèle apporte une réduction relative du PER de 15,2% par rapport au Transformer+CTC originel, atteignant un PER global de 21,2%.

Conclusion

Cette thèse CIFRE avait pour objectif, via une collaboration entre l'entreprise Lalilo et l'IRIT, l'amélioration du système de détection d'erreurs de lecture derrière l'exercice de lecture orale de leur plateforme. Cet exercice permet aux apprenant·e·s lecteur·rice·s de s'entraîner à lire oralement, activité clef pour maîtriser la lecture, et de recevoir un retour adapté sur la qualité de leur lecture. Le système de détection d'erreurs s'appuie principalement sur un module fondamental de reconnaissance automatique de phonèmes, dont l'exactitude des transcriptions automatiques fournies influe fortement sur la bonne classification des lectures correctes et incorrectes. Une grande précision de ce module de reconnaissance de phonèmes est donc primordiale pour générer un retour juste, et ainsi favoriser la progression de l'enfant.

Ces travaux de thèse ont ainsi été portés par cet objectif premier, en choisissant les chemins susceptibles d'apporter le plus de valeur, autant du côté scientifique qu'applicatif. Des progrès énormes ayant été réalisés dans le domaine de la RAP d'adultes au cours de ces dix dernières années, nos recherches ont principalement porté sur l'implémentation et l'évaluation d'architectures de modèles acoustiques, d'abord de type hybride puis *end-to-end*, pour la parole de jeunes enfants. Nous avons également cherché à améliorer la robustesse de nos modèles quant aux spécificités de notre domaine d'application. Nous présentons ici la synthèse de nos travaux, et notamment de nos découvertes et contributions principales pour la reconnaissance automatique de la parole d'enfants apprenant·e·s lecteur·rice·s en salle de classe.

Synthèse des travaux

Ce manuscrit était divisé en trois parties. Dans la première, nous avons posé le contexte de la thèse et identifié les défis liés à notre problématique. La grande complexité de la parole d'enfants représente notre principal défi. Cette parole est constituée de gammes de fréquences décalées et élargies par rapport à la parole d'adultes, ainsi que d'importantes variabilités acoustiques intra- et inter-locuteur·rice·s. La co-articulation des jeunes enfants est également moins précise et plus variable que celle des adultes. Ces caractéristiques rendent la parole d'enfants plus difficile à distinguer que celle d'adultes, même pour une oreille humaine. Cela implique qu'une grande quantité de données est nécessaire pour en modéliser correctement la complexité, ce qui représente un second grand défi puisqu'il n'existe pas à notre connaissance de corpus public de parole d'enfants en langue française. La parole d'enfants peut également présenter une qualité linguistique et prosodique dégradée, due à un mauvais contrôle moteur et de la langue française. Cette qualité est empirée dans notre cas : nous étudions spécifiquement de la lecture orale d'enfants débutants en lecture, qui contient de nombreuses erreurs de fluence et de déchiffrage. Enfin, le bruit présent en salle de classe, noyant la parole de l'enfant cible dans le brouhaha de ses camarades, ajoute une difficulté supplémentaire. Cette première partie a permis d'éclairer notre compréhension des différences entre parole d'adultes et d'enfants

et des particularités de la parole d'AL, et d'observer leur impact sur la RAP via d'autres études, apportant ainsi un début de réponse aux questions 1 et 6 posées en introduction de ce manuscrit.

La seconde partie était dédiée à l'étude de l'approche de modélisation acoustique hybride DNN-HMM, visant à l'établissement d'un modèle qui nous servira de référence pour la suite de nos travaux. Nous y avons également présenté le travail conséquent qu'a représenté la collecte de données de parole d'enfants durant la première année de cette thèse. Ce projet a inclus la création du contenu de lecture orale, la collecte en personne dans les écoles, la création de la procédure d'annotation et l'analyse détaillée des données récoltées visant à en comprendre les spécificités. Notre premier corpus, dénommé *Lalil-o-riginel* a été décrit. Nous avons ensuite étudié différents paramètres acoustiques, extraits à l'aide de techniques de traitement du signal ou de modèles auto-supervisés pré-entraînés. Aucun n'a surpassé les paramètres MFCC, classiquement utilisés pour la RAP d'adultes, que nous avons donc adoptés dans la suite de cette partie. Sur notre corpus *Lalil-o-fficiel*, le modèle hybride TDNNF-HMM s'est montré significativement plus efficace que son aîné le TDNN-HMM (11,2% d'amélioration relative du PER), le qualifiant comme modèle de référence. L'adaptation d'un modèle adulte à l'aide d'apprentissage par transfert (TL) a permis d'améliorer encore de 9,2% relatifs ses performances. Nous avons effectivement observé une forte disparité de performance d'un modèle adulte, lorsqu'appliqué sur la parole d'adultes puis d'enfants. Nous avons montré que la meilleure stratégie de TL, étant donnés l'âge de nos locuteur·rice·s et notre quantité de données, était d'adapter l'entièreté du réseau TDNNF, concluant ainsi que les caractéristiques complexes de la parole de jeunes enfants sont modélisées par toutes les couches du réseau. Nous avons pu enfin améliorer la robustesse du modèle TDNNF-HMM au bruit de brouhaha typique des salles de classe, grâce à une augmentation des données d'entraînement par ajout de bruit (gain relatif de 6,4%). Parmi les deux types de bruit, un bruit de brouhaha d'adulte et un bruit de brouhaha d'enfants enregistré en salle de classe, le second a démontré être le plus adapté, confirmant l'importance de faire correspondre les données d'entraînement à celles de test.

Cette seconde partie apporte ainsi des réponses aux questions 1, 2, 3 et 5 posées en introduction. Ces travaux ont fait l'objet de deux publications :

— un article à la conférence nationale des Journées d'Étude de la Parole [Gelin 2020b], qui présente les résultats du TL sur un modèle TDNN-HMM entraîné sur un petit corpus de 3 heures de parole, antérieurement à notre collecte de données ;

— un poster au *Speech in Noise workshop* [Gelin 2020a], qui confirme l'utilité du TL sur un modèle TDNNF-HMM avec une quantité plus importante de données (corpus *Lalil-o-riginel*), puis montre que l'augmentation par ajout de bruit réduit l'impact du bruit de brouhaha sur nos modèles. Ce poster nous a valu l'obtention du prix « Colin Cherry Award 2020 » récompensant le meilleur poster [1].

Suite à l'apparition de nouvelles architectures *end-to-end* révélant de très bonnes performances sur des tâches de RAP d'adultes, nous nous sommes concentré·e·s sur leur application à la parole d'enfants. La troisième partie de ces travaux portait ainsi sur l'étude de ces modèles.

1. https://2020.speech-in-noise.eu/

Nous avons tout d'abord sélectionné soigneusement les méthodes de modélisation semblant les plus pertinentes pour notre parole spécifique. Nous avons ensuite choisi plusieurs architectures *end-to-end* combinant ces méthodes de diverses façons : un modèle RNN-CTC, un modèle LAS fondé sur des RNN et un mécanisme d'attention, et son évolution le LAS+CTC, et un modèle Transformer fondé uniquement sur de l'attention, pouvant également évoluer en Transformer+CTC. Nous entraînons nos modèles via apprentissage par transfert et comparons leurs performances afin d'identifier des tendances liées à chaque méthode. Le Transformer+CTC s'est avéré être le plus performant, avec un PER de 25,0% sur le corpus *Lalil-o-fficiel*, dépassant ainsi largement le modèle de référence TDNNF-HMM (PER 30,1%). Cela nous mène à plusieurs conclusions : (1) les mécanismes d'auto-attention offrent une meilleure pertinence de l'information acoustique ou textuelle extraite que les couches RNN (2) la fonction CTC et les mécanismes d'attention fonctionnent très bien de façon combinée. En effet, l'entraînement multi-objectifs CE+CTC réduit la flexibilité de l'attention, l'aidant à ne pas manquer des évènements comme les répétitions. Le décodage hybride CTC/attention permet au modèle de s'appuyer sur la sortie CTC en cas d'énoncés très courts, où l'attention manque de contexte. Il limite de plus l'effet du modèle de langage implicite appris par le décodeur du Transformer+CTC, lui permettant ainsi de mieux détecter les erreurs de lecture. Malgré cela, les performances du Transformer+CTC restent fortement dégradées par la présence d'erreurs de lecture, alors que le modèle de référence TDNNF-HMM est moins affecté. Nous introduisons alors une méthode novatrice d'augmentation de données par simulation d'erreurs de lecture afin de le rendre plus robuste. Cette méthode procure une amélioration globale significative (PER 23,9%), et remplit son rôle en apportant de fortes améliorations sur les mots erronés, répétés ou substitués. Notre méthode d'augmentation par simulation d'erreurs est de plus complémentaire avec l'augmentation par ajout de bruit présentée dans la partie 2. Le Transformer+CTC bénéficiant des deux augmentations atteint un PER de 21,2%, ce qui représente une réduction relative de 15,2% par rapport au Transformer+CTC initial.

Cette troisième et dernière partie a visé principalement à répondre aux questions 3, 4, 5 et 6, et a mené à nos deux principales contributions :

— un article dans la revue internationale *Speech Communication* [Gelin 2021a], présentant la comparaison de nos modèles hybride et *end-to-end*. Nous y analysons l'efficacité du TL en fonction de l'architecture et des méthodes utilisées, étudions les forces et faiblesses de chaque modèle et montrons que le modèle Transformer+CTC surpasse les autres, y compris le modèle hybride de référence. Nous y évaluons enfin l'impact des spécificités de la parole d'AL ;

— un article publié lors de la conférence internationale Interspeech [Gelin 2021b], introduisant notre technique novatrice d'augmentation de données par simulation d'erreurs de lecture et mesurant son impact sur la reconnaissance de mots erronés.

Par ailleurs, j'ai assisté mes collègues dans l'utilisation de mon modèle TDNNF-HMM entraîné sur parole d'adultes afin de mesurer automatiquement l'intelligibilité de patients atteints d'un cancer de la gorge ou de la bouche [Balaguer 2021]. J'ai également aidé à la supervision de deux stagiaires à Lalilo. Le premier, dont le travail sur la classification d'enregistrements ne contenant pas de parole d'enfants ou étant trop bruités a été publié à la conférence internationale SLaTE [Hajji 2019]. Le second a fourni l'étude sur les paramètres

acoustiques présentée aux chapitres 3 et 4. Enfin, j'ai participé tout au long de cette thèse au développement de l'exercice de lecture orale de Lalilo, et notamment de l'algorithme de détection d'erreurs de lecture s'appuyant sur la reconnaissance de phonèmes, présenté en format *Show & Tell* à la conférence internationale Interspeech [Hembise 2021].

Perspectives

Bien évidemment, à l'issue de cette thèse, un certain nombre de perspectives peuvent être identifiées. Ci-dessous, nous présentons d'abord plusieurs perspectives scientifiques, rangées par ordre de priorité, puis une perspective applicative visant à exploiter les résultats de cette thèse dans la plateforme Lalilo.

Utilisation d'attention localisée Nous avons vu au chapitre 6 que le mécanisme d'attention liant l'encodeur et le décodeur du Transformer+CTC est trop flexible pour correctement détecter les erreurs de fluence comme les répétitions. Le module CTC aide à limiter cette flexibilité, mais nous observons toujours une dégradation significative des performances sur les mots répétés. Nous envisageons d'explorer des variantes de l'attention globale visant à en contraindre la flexibilité, comme l'attention localisée de [Chorowski 2015] ou l'attention locale et monotone de [Tjandra 2017]. La première est implémentée dans l'outil SpeechBrain [Ravanelli 2021], ce qui nous permettrait de l'évaluer à coût réduit en récupérant le module correspondant et en l'adaptant à notre implémentation du Transformer+CTC.

Adaptation des modèles auto-supervisés pour l'extraction de paramètres L'étude des modèles Wav2vec et PASE+ n'a pas été concluante au chapitre 4, les paramètres extraits n'ayant pas surpassé les MFCC. Ces modèles sont cependant pré-entraînés sur de la parole d'adultes en langue anglaise, ce qui peut expliquer leur mauvaise performance pour la parole d'enfants en langue française. Il pourrait être envisagé d'entraîner notre propre modèle PASE+, puisqu'il ne nécessite que 50 heures de parole (alors que Wav2vec est entraîné sur 960 heures). Ces 50 heures pourraient être réunies en annotant des données supplémentaires récoltées sur la plateforme Lalilo. Cependant, la procédure d'entraînement de PASE+ est lourde et complexe. Une solution serait d'adapter les modèles existants : utilisés jusque là comme simples extracteurs de paramètres, ils peuvent être ré-entraînés avec des données cibles selon le même principe que l'apprentissage par transfert, ce qui affinerait la pertinence des paramètres pour la parole d'enfants en langue française. La seconde version du modèle Wav2vec, nommée Wav2vec 2.0 [Baevski 2020], promet notamment une plus grande facilité d'adaptation à des domaines avec peu de données. L'utilisation du modèle multilingue XLSR-53 [Conneau 2020], dérivé de Wav2vec 2.0, pourrait offrir l'avantage d'un unique modèle pour les différentes langues proposées par Lalilo. Enfin, nous pourrions greffer le modèle auto-supervisé choisi en entrée du Transformer+CTC de façon à entraîner conjointement les deux modèles avec la parole d'enfants : les paramètres extraits seraient ainsi adaptés spécialement au Transformer+CTC.

Exploration de nouvelles architectures *end-to-end* Au regard du nombre élevé de propositions de méthodes et architectures démontrant des performances à l'état de l'art, notre comparaison des modèles *end-to-end* n'est pas exhaustive et mériterait d'être étendue au-delà de cette thèse. De futurs axes de recherche pourraient inclure le RNN-Transducer (RNN-T) [Graves 2012]. Ce modèle *end-to-end* est composé d'une structure encodeur-décodeur et constitué entièrement de réseaux récurrents. L'alignement des trames audio d'entrée et des phonèmes de sortie se fait actuellement par l'intermédiaire d'un réseau à propagation avant baptisé *joiner*. Cette méthode d'alignement répond à plusieurs inconvénients de la fonction CTC et des mécanismes d'attention. Alors que la fonction CTC ne modélise pas les inter-dépendances entre les phonèmes de sortie, le RNN-T modélise conjointement ces dépendances linguistiques, importantes au traitement de la parole, ainsi que les dépendances entre audio et transcription phonétique. De plus, contrairement aux mécanismes d'attention, le RNN-T aligne les trames audio d'entrée et les phonèmes de sortie de façon monotone, ce qui peut être mieux adapté aux signaux de parole d'enfants contenant des erreurs de fluence. Ce modèle a uniquement été utilisé, à notre connaissance, pour de la RAP d'adultes : il a surpassé un modèle de type RNN-CTC pour une tâche de reconnaissance de phonèmes dans [Graves 2012], et concurrence des modèles de type LAS ou Transformer dans des tâches de RAP [Prabhavalkar 2017, Battenberg 2017, Li 2020]. D'un autre côté, nous avons constaté dans notre étude que les mécanismes d'auto-attention étaient plus pertinents que les RNN pour extraire l'information acoustique ou textuelle. Nous pourrions alors imaginer un modèle se situant à l'embranchement entre un Transformer et un RNN-T.

Une autre direction serait d'étudier l'utilisation de réseaux de neurones convolutifs (*Convolutional Neural Network*) à deux dimensions (2D-CNN). Contrairement aux CNN à une dimension, comme le TDNNF, qui exploitent des représentations classiques telles que les MFCC, les 2D-CNN construisent une image à partir des paramètres acoustiques, avec comme axes les domaines temporels et fréquentiels, ce qui leur permet d'extraire des informations discriminantes dans les deux domaines. Il a été observé dans [Abdel-Hamid 2014] qu'un modèle CNN-HMM pouvait modéliser de légères variations de fréquences dues aux différences entre locuteurs. Cette capacité pourrait améliorer l'efficacité de l'apprentissage par transfert adulte-enfant, par une meilleure adaptation des caractéristiques fréquentielles de la parole d'enfants. Des modules CNN pourraient être utilisés, en remplacement ou complément, dans plusieurs architectures *end-to-end* étudiées dans cette thèse, comme dans l'encodeur du modèle LAS+CTC [Hori 2017b] ou en complément de modules d'auto-attention type Transformer dans le modèle Conformer [Gulati 2020].

Exploitation du modèle Transformer+CTC dans la plateforme Lalilo Comme décrit au chapitre 1, le système de reconnaissance automatique de phonèmes constitue pour l'instant seulement une brique du système global de détection d'erreurs de lecture. L'actuel modèle utilisé par la plateforme Lalilo étant un TDNNF-HMM, largement surpassé par le Transformer+CTC, nous voudrions implanter ce dernier à la place. Or, l'algorithme de décision du système global (voir figure 1.12) s'appuie notamment sur des mesures de *Goodness of Pronunciation* [Witt 2000] calculées en comparant les sorties des modules de reconnaissance de phonèmes et d'alignement forcé au niveau de la trame audio. Le Transformer+CTC, lui,

génère des probabilités au niveau des symboles de sortie et non au niveau de la trame, ce qui le rend à première vue incompatible avec le système existant. Une première solution serait de convertir les poids d'attentions liant l'information acoustique aux symboles inférés en matrice de probabilités au niveau de la trame audio, ce qui permettrait de simplement remplacer le TDNNF-HMM par le Transformer+CTC. Une seconde solution, impliquant un projet de plus grande ampleur, s'inspire du domaine de la détection et du diagnostic de mauvaises prononciations (*Mispronunciation detection and diagnosis*, MDD) appliqué aux apprenant·e·s d'une seconde langue [Li 2014b, Leung 2019]. Nous pourrions remplacer le système global par le seul Transformer+CTC, entraîné pour une tâche similaire à la MDD, mais pour des erreurs de lecture. Les informations liées au texte que l'enfant doit lire pourraient être incorporées dans le modèle via des mécanismes d'attention, comme dans [Feng 2020]. Nous pourrions également mettre à parti les métadonnées obtenues via la plateforme Lalilo sur chaque enfant (classe, niveau de lecture, leçons maîtrisées ou non, erreurs fréquentes...), toujours via des mécanismes d'attention, pour perfectionner la précision du Transformer+CTC.

Il faudra ensuite évaluer si le nouveau système ainsi créé permet d'obtenir une amélioration dans la qualité des retours fournis aux enfants, et ainsi homologuer son utilisation. Nous utiliserons pour cela des métriques classiques d'évaluation de systèmes de classification : rappel, précision, score F1, taux de faux négatifs et de faux positifs. Il est important pour l'application de minimiser le taux de faux positifs (un mot correctement lu est détecté comme erroné) car cela génère de la frustration et trouble l'apprentissage de l'enfant. Enfin, la validation optimale quant à l'utilité applicative de notre travail nécessitera une étude d'efficacité (par essais contrôlés randomisés dans les salles de classe ou par test A/B intégré à la plateforme) portant spécifiquement sur l'exercice de lecture orale. L'objectif de cette étude est d'apporter une réponse à la question suivante : les améliorations apportées par ces travaux de thèse sur les modèles acoustiques pour la reconnaissance de phonèmes dans la parole d'enfants sont-elles à l'origine d'une meilleure progression des enfants dans l'apprentissage de la lecture ?

Annexe A

Quelques histoires pour la première collecte auprès des familles

Voici des exemples d'histoires lues par les enfants participants à la première collecte, effectuée auprès des familles des employés de Lalilo en 2018. Les histoires étant relativement longues, seules les premières pages sont affichées sur les figures A.1, A.2 et A.3.

Figure A.1 – Histoire de niveau 1 : « Attention à l'araignée ! » de Kanchan Bannerjee

Figure A.2 – Histoire de niveau 2 : « Tyranno le terrible » de Hans Wilhelm

Qu'il est bon de se blottir sous les couvertures le soir !

Après une journée bien riche d'apprentissages, de rires et de jeux, Victorien apprécie son doux lit moelleux. Cette impression de velours, de chaleur et de confort est accentuée par le bon bain chaud qu'il vient de prendre.

- Tu peux encore lire quelques minutes, lui dit maman, après quoi il faudra éteindre les lumières et « dodo ». Bonne nuit mon chéri !

Elle l'embrasse et quitte la pièce à petits pas.

Seul dans sa chambre, Victorien savoure ce moment de solitude et de silence quelquefois interrompu par le doux frottement des draps et des pages du livre qu'il est en train de lire.

FIGURE A.3 – Histoire de niveau 3 : « Super Héros » de Odysseus

Annexe B

Documents officiels pour la collecte de parole d'enfant en école

---

La collecte a été organisée avec des écoles volontaires de l'agglomération toulousaine, et avec l'aide de Charlotte SICRE, chargée de valorisation scientifique à l'IRIT. Les documents officiels suivants étaient nécessaires, et sont ajoutés aux pages suivantes :

— *« Autorisation d'enregistrement de la voix »* : à faire signer par l'enfant et son·sa représentant·e légal·e ;

— *« Animation d'ateliers de lecture à voix haute avec enregistrement audio, en vue de l'amélioration d'un système de reconnaissance vocale à application pédagogique »* : explication du projet de recherche et du déroulement de la collecte, à destination des enseignant·e·s et directeur·ice·s d'école ;

— *« Animation d'ateliers de lecture à voix haute avec enregistrement audio, pour un projet de recherche - Notice d'information aux parents »* : notice d'explication de la procédure pour les enfants, statuant les droits de leur enfant.

Autorisation d'enregistrement de la voix
Projet de recherche

La présente demande est destinée à recueillir le consentement et les autorisations nécessaires dans le cadre du projet spécifié ci-dessous, étant entendu que les objectifs de ce projet ont été préalablement expliqués aux élèves et à leurs responsables légaux. Une notice d'information est par ailleurs remise aux représentants légaux des élèves.

1. Désignation du projet de recherche

Votre enfant est invité à prendre part à un projet de recherche intitulé «Évaluation automatique multidimensionnelle de la lecture à voix haute d'enfants apprenants lecteurs ». Ce projet de recherche entre dans le cadre de la réalisation d'une thèse de doctorat par Mme Lucile Gelin (Doctorante) sous la responsabilité de M. Julien Pinquier (MCF UPS) et de M. Thomas Pellegrini (MCF UPS), en collaboration entre l'IRIT et la société Lalilo.

L'enregistrement audio de votre enfant et son exploitation seront réalisés par :
L'Université Toulouse III – Paul Sabatier, agissant au nom et pour le compte de l'Institut de Recherche en Informatique de Toulouse (UMR 5505), 118 Route de Narbonne 31062 Toulouse cedex 9,
ET
Lalilo, Société par actions simplifiée, 136 Avenue du Général de Gaulle 92130 Issy-les-Moulineaux.

2. Modes d'exploitation envisagés

Les enregistrements audios de votre enfant seront réalisés sur un ordinateur portable protégé par un mot de passe à accès unique par la doctorante Lucile Gelin. Ils seront, à la fin de chaque séance d'enregistrement, stockés sur deux serveurs informatiques, un de l'IRIT et un de Lalilo, qui seront tous deux cryptés et à accès restreint protégé par un mot de passe.

Seuls, l'IRIT (équipe SAMoVA - Structuration, Analyse et Modélisation de documents Vidéo et Audio) et la société Lalilo auront accès à l'enregistrement audio de votre enfant.
Votre accord à participer implique que vous acceptez que l'IRIT et Lalilo puisse :

- Enregistrer la voix de votre enfant
- Utiliser cet enregistrement dans le cadre de la réalisation du projet de recherche décrit au paragraphe 1
- Diffuser les résultats du projet de recherche sous forme d'articles dans des publications scientifiques, d'essai, de thèse, ou dans le cadre de conférences et communications scientifiques à la condition qu'aucune information permettant d'identifier votre enfant ne soit divulguée publiquement à moins d'un consentement explicite de votre part et de l'accord de votre enfant.

Les enregistrements audios collectés dans le cadre de ce projet seront conservés 5 ans après la fin du projet de recherche.

3. Informations de l'élève très jeune ou non lecteur

Rappel : l'article 16 de la Convention internationale des Droits de l'Enfant consacre le droit au respect de sa vie privée, ce qui implique notamment le respect de son droit à l'image. Lorsque l'enfant est trop jeune pour exprimer son consentement de façon autonome et éclairée (compréhension des enjeux et des conséquences), il importe de lui fournir les explications adaptées à son âge et de s'assurer autant qu'il est possible, compte tenu de son âge et de sa compréhension, de son adhésion au projet.

Mon enfant, Nom : Prénom : Classe : Age :	❏ A été informé des objectifs de ce projet, ❏ sait qui pourra voir, entendre l'enregistrement, ❏ a compris et dit qu'il était d'accord pour qu'on enregistre sa voix.

4. Autorisation parentale

Vu le Code civil, en particulier son article 9, sur le respect de la vie privée,
Vu le Code de la propriété intellectuelle,
Vu le consentement préalablement exprimé par la personne mineure ci-avant,

La présente autorisation est soumise à votre signature, pour la fixation sur support audiovisuel de la voix de votre enfant mineur dont l'identité est donnée au paragraphe 3, ci–avant, dans le cadre du projet désigné au paragraphe 1 et pour les modes d'exploitation désignés au paragraphe 2.

Cet enregistrement de la voix du mineur que vous représentez sera réalisé sous l'autorité de:

L'Université Toulouse III – Paul Sabatier, agissant au nom et pour le compte de l'Institut de Recherche en Informatique de Toulouse (UMR 5505), 118 Route de Narbonne 31062 Toulouse cedex 9
Et
Lalilo Société par actions simplifiée, 136 Avenue du Général de Gaulle 92130 Issy-les-Moulineaux.

Ci-après désigné par les « bénéficiaires »
L'enregistrement aura lieu aux dates/moments et lieux indiqués ci-après.
Date(s) d'enregistrement : ……
Lieu(x) d'enregistrement : ……
Les bénéficiaires de l'enregistrement exerceront l'intégralité des droits d'exploitation attachés à cet enregistrement. L'enregistrement demeurera leur propriété. Les bénéficiaires de l'autorisation, s'interdisent expressément de céder les présentes autorisations à un tiers.

Ils s'interdisent également de procéder à une exploitation illicite, ou non prévue ci-avant, de l'enregistrement de la voix du mineur susceptible de porter atteinte à sa dignité, sa réputation ou à sa vie privée et toute autre exploitation préjudiciable selon les lois et règlements en vigueur.
Dans le contexte défini dans le paragraphe 1 et 2, l'enregistrement ne pourra donner lieu à aucune rémunération ou contrepartie sous quelque forme que ce soit. Cette acceptation expresse est définitive et exclut toute demande de rémunération ultérieure.

Je soussigné(e) *(prénom, nom)* ……..... **déclare être le représentant légal du mineur désigné au paragraphe 3.**

Je reconnais être entièrement investi de mes droits civils à son égard. Je reconnais expressément que le mineur que je représente n'est lié par aucun contrat exclusif pour l'utilisation de sa voix, voire de son nom.

Je reconnais avoir pris connaissance des informations ci-dessus concernant le mineur que je représente et donne mon accord pour la fixation de sa voix, dans le cadre exclusif du projet exposé au paragraphe 1 et tel qu'il y a consenti au paragraphe 3 : ☐ OUI ☐ NON

Fait en autant d'originaux que de signataires.

Fait à : **Le (date) :**	
Signature du responsable scientifique	**Signature du représentant légal du mineur**

Animation d'ateliers de lecture à voix haute avec enregistrement audio, en vue de l'amélioration d'un système de reconnaissance vocale à application pédagogique

Les enfants sont de véritables machines à apprendre ! Toutefois, l'école manque d'outils et de méthodes et les professeurs manquent de temps pour pouvoir s'adapter à chaque enfant selon son propre rythme. Aujourd'hui, ¼ des élèves arrivent en 6ème avec des difficultés majeures pour lire et bien comprendre. Lalilo conçoit un assistant pédagogique dont l'objectif est d'aider le professeur à personnaliser son enseignement en lui permettant de faire travailler en autonomie un groupe d'enfants sur Lalilo et en lui fournissant des outils de suivi et de préparation de classe. Lalilo se présente sous la forme d'une plateforme web, où chaque enfant travaille sur des exercices de lecture adaptés à son niveau et ses difficultés, et où l'enseignant a accès à un tableau de bord lui donnant un suivi du niveau des élèves.

En collaboration avec l'Institut de Recherche en Informatique de Toulouse (IRIT), Lalilo conçoit un système intelligent capable de détecter automatiquement les erreurs de lecture d'enfant apprenants lecteurs, ainsi que d'évaluer leur niveau de fluence. Cette technologie est d'ores et déjà intégrée dans Lalilo, sous la forme d'un exercice de lecture à voix haute. Pour que ce système soit encore plus efficace et fiable, nous avons besoin d'un grand nombre d'enregistrements de voix d'enfants. C'est pourquoi nous avons besoin de vous !

Nous cherchons donc des professeurs nous autorisant à venir animer des ateliers de lecture à voix haute. Ces ateliers feront l'objet d'un enregistrement audio de quelques minutes. Les visites peuvent être prévues pendant les horaires scolaires ou extra-scolaires (temps d'étude, activités organisées par l'école pendant les vacances scolaires..).

Quelle tranche d'âge ?
Notre système étant destiné aux enfants apprenants lecteurs, nous cherchons des enfants de CP-CE1-CE2, capables de déchiffrer des syllabes, mots ou phrases.

Qui effectue la collecte ?
Une petite équipe, formée de 2 personnes, travaille sur la reconnaissance vocale à Lalilo, bénéficiant des conseils de chercheurs de l'IRIT. Lucile Gelin, qui démarrera prochainement une thèse sur ce sujet, s'occupera de venir animer les ateliers.

Comment se passera la collecte ?
Les ateliers se dérouleront un enfant à la fois, si possible dans une salle à part, afin que l'enfant se sente bien et pour éviter le bruit, qui pourrait gâcher l'enregistrement. Selon leur niveau de lecture, les enfants devront lire à voix haute sur un support papier des syllabes, des mots, des phrases ou de petites histoires. L'enregistrement s'effectuera grâce à un micro-casque relié à un ordinateur. Les enregistrements obtenus seront **entièrement anonymisés** et utilisés dans le cadre des recherches scientifiques de Lalilo et l'IRIT.

Contact ? Lucile Gelin, 07 82 39 34 29, lucile.gelin@irit.fr

Animation d'ateliers de lecture à voix haute avec enregistrement audio, pour un projet de recherche

Notice d'information aux parents

Identification du projet et des responsables scientifiques

Projet de recherche : Evaluation automatique multidimensionnelle de la lecture à voix haute d'enfants apprenants lecteurs

Directeurs de thèse : Julien PINQUIER et Thomas PELLEGRINI, Maîtres de Conférences, Université Toulouse III - Paul Sabatier - Institut de Recherche en Informatique de Toulouse (IRIT) - 118 Route de Narbonne 31062 Toulouse Cedex 9 - Equipe SAMoVA (Synthèse, Analyse et Modélisation de documents Audio et Vidéo)

Contact : julien.pinquier@irit.fr - 05 61 55 74 34
thomas.pellegrini@irit.fr - 05 61 55 68 86

Doctorante : Lucile GELIN, Doctorante, employée par la société Lalilo (136 Avenue du Général de Gaulle 92130 Issy-les-Moulineaux) pour la réalisation du projet de recherche en collaboration avec l'IRIT dans le cadre du dispositif CIFRE (Conventions Industrielles de Formation par la REcherche)

Objectif général du projet

Votre enfant est invité à prendre part à un projet de recherche intitulé « Evaluation automatique multidimensionnelle de la lecture à voix haute d'enfants apprenants lecteurs »

Ce projet de recherche entre dans le cadre de la réalisation d'une thèse de doctorat par Mme Lucile Gelin (Doctorante) sous la responsabilité de M. Julien Pinquier et M. Thomas Pellegrini (MCF UPS) en collaboration entre l'IRIT et la société Lalilo.

Afin de remédier à la proportion croissante d'enfants arrivant en 6ème avec des difficultés majeures en lecture, Lalilo conçoit un assistant pédagogique dont l'objectif est d'aider le professeur à personnaliser son enseignement en lui permettant de faire travailler en autonomie un groupe d'enfants sur Lalilo et en lui fournissant des outils de suivi et de préparation de classe. Lalilo se présente sous la forme d'une plateforme web, où chaque enfant travaille sur des exercices de lecture adaptés à son niveau et ses difficultés, et où l'enseignant a accès à un tableau de bord lui donnant un suivi du niveau des élèves.

L'IRIT et Lalilo collaborent dans l'objectif d'améliorer l'assistant pédagogique Lalilo en le rendant capable de détecter automatiquement les erreurs de lecture d'enfants apprenants lecteurs. Pour que ce système soit encore plus efficace et fiable, nous avons besoin d'un grand nombre d'enregistrements de voix d'enfants.

La contribution de votre enfant favorisera l'avancement des connaissances dans le domaine de la reconnaissance automatique de la parole.

Le/la directeur(rice) de l'école, ainsi que le/la professeur(e) de votre enfant, ont été prévenus et ont donné leur accord.

Tâches demandées à votre enfant

Avec votre autorisation et l'accord de votre enfant, ce dernier sera invité lors d'un entretien individuel à lire à voix haute des syllabes, des mots, des phrases ou de petites histoires, selon son niveau de lecture. Cette session de lecture sera enregistrée grâce à un micro-casque relié à un ordinateur. La session durera environ 10 minutes, et se déroulera dans une salle de l'Ecole Jules Ferry, sous la surveillance du professeur de votre enfant.

Avantages et risques d'inconfort

Il n'y a pas de risque associé à la participation de votre enfant à ce projet. Les activités proposées à votre enfant sont similaires à celles qui rencontre dans une journée de classe ordinaire. Néanmoins, soyez assuré que les responsables scientifiques resteront attentifs à toute manifestation d'inconfort chez votre enfant durant sa participation.

Anonymat et confidentialité

Il est entendu que seuls, l'IRIT (équipe SAMoVA) et Lalilo auront accès à l'enregistrement audio de votre enfant.

L'IRIT et Lalilo peuvent être amenés à diffuser les résultats du projet de recherche sous forme d'articles dans des publications scientifiques, d'essai, de thèse, ou dans le cadre de coférences et communications scientifiques. Cependant, aucune information permettant d'identifier votre enfant ne sera divulguée publiquement, à moins d'un consentement explicite de votre part et l'accord de votre enfant.

Les enregistrements audio seront conservés sur un serveur sécurisé de l'IRIT. Ils seront détruits 5 ans après la fin du projet de recherche.

Participation volontaire

La participation de votre enfant à ce projet est volontaire. Cela signifie que même si vous consentez aujourd'hui à ce que votre enfant participe à cette recherche, il demeure entièrement libre de ne pas participer ou de mettre fin à sa participation en tout temps sans justification ni pénalité. Vous pouvez également retirer votre enfant du projet en tout temps.

Aspect réglementaire (mention obligatoire)

La voix de votre enfant constitue une donnée personnelle protégée par le Règlement européen sur la protection des données.
Ces données personnelles seront traitées par l'Institut de Recherche en Informatique de Toulouse – IRIT – UMR5505 (118 Route de Narbonne 31062 Toulouse cedex 9) et la société LALILO (136 Avenue du Général de Gaulle 92130 Issy-les-Moulineaux).
Ces données sont collectées dans le seul objectif de réaliser le projet de recherche intitulé « valuation automatique multidimensionnelle de la lecture à voix haute d'enfants apprenants lecteurs».

La base légale du traitement est le consentement (cf. article 6.1.a) du Règlement européen sur la protection des données.
Les destinataires des données collectées sont l'IRIT et LALILO.
Ces données seront conservées 5 ans après la fin du projet de recherche.

Vous pouvez accéder aux enregistrements audio de votre enfant ou demander leur effacement. Vous disposez également d'un droit d'opposition, d'un droit de rectification et d'un droit à la limitation du traitement de vos données (cf. cnil.fr pour plus d'informations sur vos droits).

Pour exercer ces droits ou pour toute question sur le traitement de vos données dans ce dispositif, vous pouvez contacter :
M Julien Pinquier - M Thomas Pellegrini
Institut de Recherche en Informatique de Toulouse (IRIT)
118 Route de Narbonne 31062 Toulouse cedex 9
julien.pinquier@irit.fr - thomas.pellegrini@irit.fr
05 61 55 74 34 - 05 61 55 68 86

Si vous estimez, après nous avoir contactés, que vos droits Informatique et Libertés ne sont pas respectés ou que le dispositif de contrôle d'accès n'est pas conforme aux règles de protection des données, vous pouvez adresser une réclamation en ligne à la CNIL ou par voie postale.

Des questions sur le projet ou sur vos droits ?

Vous pouvez contacter le Directeur de thèse pour des questions additionnelles sur le déroulement du projet. Vous pouvez également discuter avec le Directeur de thèse des conditions dans lesquelles se déroulera la participation de votre enfant et de ses droits en tant que participant de recherche.

Votre collaboration et celle de votre enfant sont importantes à la réalisation de ce projet, et nous tenons à vous en remercier.

Annexe C

Format des annotations en base de données

Toutes les opérations de gestion et préparation de données sont effectuées à partir des données en base, exploitables par un fichier json contenant tous les enregistrements et les informations nécessaires pour chaque mot à partir des transcriptions.

Par exemple, pour un enregistrement où l'enfant doit lire « *Le chat est gris* » mais où il lit « *Le ch...chat est gros* » et fait un commentaire à la fin « *pfff* », le json serait de la forme suivante :


```
{
    "id": 839,
    "random_speaker_id": "db56zeoi47",
    "path": "chemin/vers/enregistrement.wav",
    "lang": "fr",
    "text_type": "sentence",
    "text": "le chat est gris",
    "corpus": "TEST",
    "annotations": [
        {"id": 4689, "label": "OK", "word_id": 0, "word": "le", "word_pos": 0},
        {"id": 4690, "label": "FS", "word_id": 1, "word": "chat", "word_pos": 1, "value":["SH"]},
        {"id": 4691, "label": "OK", "word_id": 1, "word": "chat", "word_pos": 2},
        {"id": 4692, "label": "OK", "word_id": 2, "word": "est", "word_pos": 3},
        {"id": 4693, "label": "MISPRON_1", "word_id": 3, "word": "gris", "word_pos": 4, "value":["G", "R", "o"]}
        {"id": 4694, "label": "COMMENT", "word_id": None, "word_pos": 5, "value":["P", "F"]}
    ]
}
```


Figure C.1 – Exemple de fichier json contenant les informations pour un enregistrement où l'enfant lit « *Le ch...chat est gros pff* »

Chaque enregistrement possède un numéro d'identification, un chemin, un langage (anglais ou français), et un numéro d'identification du locuteur, anonymisé pour répondre aux normes RGPD. Le *text_type* indique s'il contient un mot isolé ou une phrase. Le texte que l'enfant doit lire est indiqué dans la clé *text*. La clé *annotations* contient une liste d'annotations pour chaque mot prononcé par l'enfant dans l'enregistrement :

— Le premier mot « *Le* » est correctement lu (*label* = OK), c'est le premier mot du texte à lire (*word_id* = 0) et le premier mot lu par l'enfant (*word_pos* = 0). Puisque le mot est bien lu, la transcription phonétique n'est pas gardée en mémoire, mais sera générée automatiquement au besoin grâce à un dictionnaire de prononciation ;

— Le second mot du texte de référence « *chat* » est lu deux fois par l'enfant :

  - ⋆ La première fois, correspondant à l'annotation ligne 11, il fait un faux départ (*label* = FS) en ne lisant que le premier phonème (*value* = ["SH"]). C'est le deuxième mot du texte à lire (*word_id* = 1) et le deuxième mot lu par l'enfant (*word_pos* = 1) ;
  - ⋆ La deuxième fois (ligne 12), il lit le mot correctement (*label* = OK). C'est toujours le deuxième mot du texte à lire (*word_id* = 1) mais le troisième mot lu par l'enfant (*word_pos* = 2) ;
- Le mot « *est* » est correctement lu (*label* = OK), c'est le troisième mot du texte à lire (*word_id* = 2) mais le quatrième mot lu par l'enfant (*word_pos* = 3) ;
- Le mot « *gris* » est mal lu, l'enfant confond les phonèmes [i] et [o] en lisant le mot « *gros* » (*label* = MISPRON_1). La transcription phonétique est précisée (*value* = [G, R, o]) ;
- Le dernier mot prononcé par l'enfant est le commentaire « *pff* » (*label* = COMMENT), mot qui n'existe pas dans la langue française, il n'a donc ni clé *word* ni indice dans le texte de référence (*word_id* = None) mais a bien une position dans la phrase lue par l'enfant (*word_pos* = 5) et une transcription phonétique associée (*value* = [P, F]).

Annexe D

Algorithme progressif-rétroprogressif CTC

L'algorithme CTC permet de calculer la probabilité d'obtenir une séquence Y selon l'équation (5.7), en décomposant la somme sur des chemins $\pi_{1,...,T} \in \theta_{Y_{1,...,T}}$ de longueur T en une somme itérative de chemins $\pi_{1,...,t} \in \theta_{Y_{1,...,t}}$ de longueur t. Il cherche donc à calculer la probabilité $P(\pi_t = s_{n'}|S,X)$ de voir un certain symbole $s_{n'}$ ($n' = 1...N'$ désignant l'index courant dans la séquence) à chaque temps t, étant données la séquence à aligner $S = S_0...S_{N'-1}$ et la séquence d'entrée $X = X_0...X_{T-1}$.

Cette probabilité peut être exprimée grâce au théorème de dérivation des fonctions composées, suivant l'équation (D.1). Le terme $P(S|X)$ représente la probabilité d'avoir la séquence S, étant donnée la séquence d'entrée, qui est constante lors de la phase d'alignement et peut être négligée.

$$P(\pi_t = s_{n'}|S,X) = P(\pi_t = s_{n'},S|X)P(S|X) \propto P(\pi_t = s_{n'},S|X) \tag{D.1}$$

Sans le symbole «-»

Pour calculer la probabilité $P(\pi_t = s_{n'}|S,X)$, nous ignorons dans un premier temps la présence du symbole vide. Elle peut alors être décomposée selon l'équation (D.2), où $[s_{n'+}]$ indique que π_{t+1} pourrait être soit $s_{n'}$ soit $s_{n'+1}$, et $[s_{n'-}]$ que π_{t-1} pourrait être soit $s_{n'}$ soit $s_{n'-1}$. La suppression de la séquence à aligner S à la deuxième ligne est possible, car les séquences $\pi_0...\pi_{T-1}$ sont contraintes à être dérivées de S.

$$\begin{aligned} & P(\pi_t = s_{n'},S|X) \\ &= \sum_{\pi_0...\pi_{t-1} \to s_1...[s_{n'-}]} \; \sum_{\pi_{t+1}...\pi_{T-1} \to [s_{n'+}]...s_{N'}} P(\pi_0...\pi_{t-1}, \pi_t = s_{n'}, \pi_{t+1}...\pi_{T-1}, S|X) \\ &= \sum_{\pi_0...\pi_{t-1} \to s_1...[s_{n'-}]} \; \sum_{\pi_{t+1}...\pi_{T-1} \to [s_{n'+}]...s_{N'}} P(\pi_0...\pi_{t-1}, \pi_t = s_{n'}, \pi_{t+1}...\pi_{T-1}|X) \end{aligned} \tag{D.2}$$

À l'aide de la loi de Bayes, nous obtenons l'équation (D.3) :

$$P(\pi_t = s_{n'},S|X) = \sum_{\pi_0...\pi_{t-1}\rightarrow s_1...[s_{n'-}]} \sum_{\pi_{t+1}...\pi_{T-1}\rightarrow[s_{n'+}]...s_{N'}} [P(\pi_0...\pi_{t-1},\pi_t = s_{n'}|X)P(\pi_{t+1}...\pi_{T-1}|\pi_0...\pi_{t-1},\pi_t = s_{n'},X)] \tag{D.3}$$

Pour une utilisation de la CTC avec un réseau RNN sans retour de la sortie, nous pouvons supposer que les probabilités d'observer chaque symbole sont conditionnellement indépendantes les unes des autres sachant la séquence d'entrée :

$$P(\pi_{t+1}...\pi_{T-1}|\pi_0...\pi_{t-1},X) = P(\pi_{t+1}...\pi_{T-1}|X) \tag{D.4}$$

Cette hypothèse permet d'obtenir l'équation (D.5), puis par simple factorisation, l'équation (D.6).

$$P(\pi_t = s_{n'},S|X) = \sum_{\pi_0...\pi t-1\rightarrow s_1...[s_{n'-}]} \sum_{\pi_{t+1}...\pi_{T-1}\rightarrow[s_{n'+}]...s_{N'}} [P(\pi_0...\pi_{t-1},\pi_t = s_{n'}|X)P(\pi_{t+1}...\pi_{T-1}|\pi_t = s_{n'},X)] \tag{D.5}$$

$$\begin{aligned} &P(\pi_t = s_{n'},S|X) \\ &= \left[\sum_{\pi_0...\pi t-1\rightarrow s_1...[s_{n'-}]} P(\pi_0...\pi_{t-1},\pi_t = s_{n'}|X)\right]\left[\sum_{\pi_{t+1}...\pi_{T-1}\rightarrow[s_{n'+}]...s_{N'}} P(\pi_{t+1}...\pi_{T-1}|\pi_t = s_{n'},X)\right] \\ &= \alpha(t,n')\beta(t,n') \end{aligned} \tag{D.6}$$

Le premier terme de l'équation (D.6), allant de 0 à $t-1$ est la probabilité *progressive* $\alpha(t,n')$. Le second terme, allant de $t+1$ à $T-1$, est la probabilité *rétroprogressive* $\beta(t,n')$. Les calculs de ces deux termes (équations (D.7) et (D.8)) mènent à des expressions récursives, qui permettent à l'algorithme une grande efficacité. Les sous-algorithmes récursifs sont présentés à la suite, où nous avons simplifié les termes $P(\pi_t = s_{n'}|X)$ en $p_t^{s(n')}$, $P(\pi_{t+1} = s_{n'}|X)$ en $p_{t+1}^{s(n')}$ et $P(\pi_{t+1} = s_{n'+1}|X)$ en $p_{t+1}^{s(n'+1)}$.

$$
\begin{aligned}
\alpha(t,n') &= \sum_{\pi_0 \dots \pi_{t-1} \rightarrow s_1 \dots [s_{n'}-]} P(\pi_0 \dots \pi_{t-1}, \pi_t = s_{n'}|X) \\
&= \sum_{\pi_0 \dots \pi_{t-1} \rightarrow s_1 \dots [s_{n'}-]} P(\pi_0 \dots \pi_{t-1}|X) P(\pi_t = s_{n'}|(\pi_0 \dots \pi_{t-1}, X) \\
&= \sum_{\pi_0 \dots \pi_{t-1} \rightarrow s_1 \dots [s_{n'}-]} P(\pi_0 \dots \pi_{t-1}|X) P(\pi_t = s_{n'}|X) \\
&= \left(\sum_{\pi_0 \dots \pi_{t-2} \rightarrow s_1 \dots [s_{n'}-]} P(\pi_0 \dots \pi_{t-2}, \pi_{t-1} = s_{n'}|X) \right. \\
&\quad \left. + \sum_{\pi_0 \dots \pi_{t-2} \rightarrow s_1 \dots [s_{(n'-1)}-]} P(\pi_0 \dots \pi_{t-2}, \pi_{t-1} = s_{n'-1}|X) \right) \times P(\pi_t = s_{n'}|X) \\
&= \left(\alpha(t-1,n') + \alpha(t-1,n'-1)\right) P(\pi_t = s_{n'}|X)
\end{aligned}
\tag{D.7}
$$

$$
\begin{aligned}
\beta(t,n') &= \sum_{\pi_{t+1} \dots \pi_{T-1} \rightarrow [s_{n'}+] \dots s_{N'}} P(\pi_{t+1} \dots \pi_{T-1}|X) \\
&= \sum_{\pi_{t+2} \dots \pi_{T-1} \rightarrow [s_{n'}+] \dots s_{N'}} P(\pi_{t+1} = s_{n'}, \pi_{t+2} \dots \pi_{T-1}|X) \\
&+ \sum_{\pi_{t+2} \dots \pi_{T-1} \rightarrow [\pi_{(n'+1)+}] \dots s_{N'}} P(\pi_{t+1} = s_{n'+1}, \pi_{t+2} \dots \pi_{T-1}|X) \\
&= P(\pi_{t+1} = s_{n'}|X) \sum_{\pi_{t+2} \dots \pi_{T-1} \rightarrow [s_{n'}+] \dots s_{N'}} P(\pi_{t+2} \dots \pi_{T-1}|\pi_{t+1} = s_{n'}, X) \\
&+ P(\pi_{t+1} = s_{n'+1}|X) \sum_{\pi_{t+2} \dots \pi_{T-1} \rightarrow [s_{(n'+1)+}] \dots s_{N'}} P(\pi_{t+2} \dots \pi_{T-1}|\pi_{t+1} = s_{n'+1}, X) \\
&= P(\pi_{t+1} = s_{n'}|X) \sum_{\pi_{t+2} \dots \pi_{T-1} \rightarrow [s_{n'}+] \dots s_{N'}} P(\pi_{t+2} \dots \pi_{T-1}|X) \\
&+ P(\pi_{t+1} = s_{n'+1}|X) \sum_{\pi_{t+2} \dots \pi_{T-1} \rightarrow [s_{(n'+1)+}] \dots s_{N'}} P(\pi_{t+2} \dots \pi_{T-1}|X) \\
&= P(\pi_{t+1} = s_{n'}|X) \beta(t+1,n') + P(\pi_{t+1} = s_{n'+1}|X) \beta(t+1,n'+1)
\end{aligned}
\tag{D.8}
$$

Algorithm 3 Sous-algorithme récursif pour le calcul de $\alpha(t,n')$

Initialisation : $\alpha(0,1) = p_0^{s(1)}, \alpha(0,n') = 0, n' > 1$

for $t = 1...T-1$ **do**
  $\alpha(t,1) = \alpha(t-1,1)p_t^{s(1)}$
  for $n' = 2...R$ **do**
    $\alpha(t,n') = (\alpha(t-1,n') + \alpha(t-1,n'-1))p_t^{s(n')}$
  end for
end for

Algorithm 4 Sous-algorithme récursif pour le calcul de $\beta(t,n')$

Initialisation : $\beta(T-1,N') = 1, \beta(T-1,n') = 0, n' < N'$

for $t = T-1...0$ **do**
  $\beta(t,N') = \beta(t+1,N')p_{t+1}^{s(N')}$
  for $n' = N'-1...1$ **do**
    $\beta(t,n') = p_{t+1}^{s(n')}\beta(t+1,n') + p_{t+1}^{s(n'+1)}\beta(t+1,n'+1)$
  end for
end for

Une fois les termes $\alpha(t,n')$ et $\beta(t,n')$ calculés, nous pouvons reprendre l'équation (D.1) :

$$\begin{aligned} P(\pi_t = s_{n'}|S,X) &= \frac{P(\pi_t = s_{n'}, S|X)}{\sum_{s'_{n'}} P(\pi_t = s'_{n'}, S|X)} \\ &= \frac{\alpha(t,n')\beta(t,n')}{\sum_{n''} \alpha(t,n'')\beta(t,n'')} \end{aligned} \tag{D.9}$$

Avec le symbole «-»

L'ajout du symbole vide altère légèrement les sous-algorithmes 3 et 4 et donne les sous-algorithmes 5 et 6. Puisque le premier symbole peut être un symbole cible ou le symbole vide, deux valeurs sont initialisées et des termes sont ajoutés aux expressions récursives.

Algorithm 5 Sous-algorithme récursif pour le calcul de $\alpha(t,n')$ avec le symbole vide

Initialisation : $\alpha(0,0) = p_0^{s(0)}, \alpha(0,1) = p_0^{s(1)}, \alpha(0,n') = 0, n' > 1$

for $t = 1...T-1$ **do**
  $\alpha(t,1) = \alpha(t-1,1)p_t^{s(1)}$
  for $n' = 2...R$ **do**
    if $s(n')$ ="-" or $s(n') = s(n'-2)$ **then**
      $\alpha(t,n') = (\alpha(t-1,n') + \alpha(t-1,n'-1))p_t^{s(n')}$
    else
      $\alpha(t,n') = (\alpha(t-1,n') + \alpha(t-1,n'-1) + \alpha(t-1,n'-2))p_t^{s(n')}$
    end if
  end for
end for

Algorithm 6 Sous-algorithme récursif pour le calcul de $\beta(t,n')$, avec le symbole vide

Initialisation : $\beta(T-1,N') = 1, \beta(T-1,2N'+1) = \beta(T-1,2N') = \beta(T-1,n') = 0, n' < 2N'$

for $t = T-1...0$ **do**
  $\beta(t,N') = \beta(t+1,N')p_{t+1}^{s(N')}$
  for $n' = N'-1...1$ **do**
    if $s(n')$ ="-" or $s(n') = s(n'+2)$ **then**
      $\beta(t,n') = (p_{t+1}^{s(n')}\beta(t+1,n') + p_{t+1}^{s(n'+1)}\beta(t+1,n'+1)$
    else
      $\beta(t,n') = (p_{t+1}^{s(n')}\beta(t+1,n') + p_{t+1}^{s(n'+1)}\beta(t+1,n'+1) + p_{t+1}^{s(n'+2)}\beta(t+1,n'+2)$
    end if
  end for
end for

Bibliographie

[Abad 2020] A. Abad, P. Bell, Andrea Carmantini et S. Renals. *Cross Lingual Transfer Learning for Zero-Resource Domain Adaptation.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6909–6913, 2020.

[Abdel-Hamid 2014] O. Abdel-Hamid, A-R. Mohamed, H. Jiang, L. Deng, G. Penn et D. Yu. *Convolutional Neural Networks for Speech Recognition.* IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), vol. 22, no. 10, page 1533–1545, 2014.

[Abka 2015] A. F. Abka et H. F. Pardede. *Speech recognition features : Comparison studies on robustness against environmental distortions.* In Proc. of the International Conference on Computer, Control, Informatics and its Applications (IC3INA), pages 114–119, 2015.

[Airaksinen 2019] M. Airaksinen, L. Juvela, P. Alku et O. Räsänen. *Data Augmentation Strategies for Neural Network F0 Estimation.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6485–6489, 2019.

[Alexander 2008] M. P. Alexander et A. E. Hillis. *Chapter 14 - Aphasia.* In Neuropsychology and Behavioral Neurology, volume 88 of *Handbook of Clinical Neurology*, pages 287–309. Elsevier, 2008.

[Alwan 2007] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman et S. Wang. *A System for Technology Based Assessment of Language and Literacy in Young Children : the Role of Multiple Information Sources.* In IEEE Workshop on Multimedia Signal Processing, pages 26–30, 2007.

[Amodei 2016] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng et et al G. Chen. *Deep speech 2 : End-to-end speech recognition in English and Mandarin.* In Proc. of the International Conference on Machine learning (ICML), volume 48, pages 173–182, 2016.

[Anastasakos 1997] T. Anastasakos, J. McDonough et J. Makhoul. *Speaker Adaptive Training : a Maximum Likelihood Approach to Speaker Normalization.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 2, pages 1043–1046, 1997.

[A.Varga 1993] A.Varga et Herman J.M. Steeneken. *Assessment for automatic speech recognition : II. NOISEX-92 : A database and an experiment to study the effect of additive noise on speech recognition systems.* Speech Communication, vol. 12, no. 3, pages 247–251, 1993.

[Baevski 2020] A. Baevski, H. Zhou, A. Mohamed et M. Auli. *Wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations.* ArXiv preprint :2006.11477, 2020.

[Bahdanau 2014] D. Bahdanau, K. Cho et Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate.* ArXiv preprint :1409.0473, 2014.

[Bahl 1991] L. R. Bahl, P. V. de Soutza, P. S. Gopalakrishnan, D. Nahamoo et M. A. Picheny. *Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees.* In Proceedings of the Workshop on Speech and Natural Language, HLT '91, page 264–269, USA, 1991. Association for Computational Linguistics.

[Baker 1975] J. Baker. *The DRAGON system–An overview.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23, no. 1, pages 24–29, 1975.

[Balaguer 2021] Mathieu Balaguer, Lucile Gelin, Virginie Woisard, Jérôme Farinas et Julien Pinquier. *Mesure de l'intelligibilité après cancer oral ou oropharyngé par un système de reconnaissance automatique de la parole.* In 1ère Journée Scientifique d'Orthophonie, Congrès en ligne, France, Octobre 2021. SURO Société Universitaire de Recherche en Orthophonie.

[Barker 2017] J. Barker, R. Marxer, E. Vincent et S. Watanabe. *The third 'CHiME' Speech Separation and Recognition Challenge : Analysis and outcomes.* Computer Speech & Language, vol. 46, pages 605–626, 2017.

[Barker 2018] J. Barker, S. Watanabe, E. Vincent et J. Trmal. *The fifth 'CHiME' Speech Separation and Recognition Challenge : Dataset, task and baselines.* In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyberabad, 2018.

[Battenberg 2017] E. Battenberg, J. Chen, R. Child, A. Coates, Y. Gaur, Y. Li, H. Liu, S. Satheesh, D. Seetapun, A. Sriram et Z. Zhu. *Exploring Neural Transducers for End-to-End Speech Recognition.* ArXiv preprint :1707.07413, 2017.

[Baum 1967] Leonard E. Baum et J. A. Eagon. *An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology.* Bulletin of the American Mathematical Society, vol. 73, no. 3, pages 360 – 363, 1967.

[Bayerl 2019] S. P. Bayerl et K. Riedhammer. *A Comparison of Hybrid and End-to-End Models for Syllable Recognition.* In Text, Speech, and Dialogue, pages 352–360, 2019.

[Beaume 1987] E. Beaume. *La lecture à voix haute.* Les actes de lecture, vol. 18, 1987.

[Bengio 2015] S. Bengio, O. Vinyals, N. Jaitly et N. Shazeer. *Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks.* In Proc. of the International Conference on Neural Information Processing Systems (NIPS) - Volume 1, page 1171–1179. MIT Press, 2015.

[Bertrand 1990] J. Bertrand. Dictionnaire des homonymes. Nathan, 1990.

[Bolaños 2011] D. Bolaños, R. Cole, W. Ward, E. Borts et E. Svirsky. *FLORA : Fluent Oral Reading Assessment of Children's Speech.* IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), vol. 7, no. 4, page 16, 2011.

[Bourlard 1990] H. Bourlard et N. Morgan. *A Continuous Speech Recognition System Embedding MLP into HMM.* In D. Touretzky, editeur, Advances in Neural Information Processing Systems, volume 2. Morgan-Kaufmann, 1990.

[Bradbury 2017] J. Bradbury, S. Merity, C. Xiong et R. Socher. *Quasi-Recurrent Neural Networks.* In Proc. of the International Conference on Learning Representations, ICLR, 2017.

[Braun 2017] S. Braun, D. Neil et S-C. Liu. *A curriculum learning method for improved noise robustness in automatic speech recognition*. In 2017 25th European Signal Processing Conference (EUSIPCO), pages 548–552, 2017.

[Chakraborty 2016] C. Chakraborty et P. Talukdar. *Issues and Limitations of HMM in Speech Processing : A Survey*. International Journal of Computer Applications, vol. 141, pages 13–17, 2016.

[Chan 2016] W. Chan, N. Jaitly, Q. Le et O. Vinyals. *Listen, Attend and Spell : A Neural Network for Large Vocabulary Conversational Speech Recognition*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4960–4964, 2016.

[Chen 2020] G. Chen, X. Na, Y. Wang, Z. Yan, J Zhang, S. Ma et Y. Wang. *Data Augmentation For Children's Speech Recognition – The "Ethiopian" System For The SLT 2021 Children Speech Recognition Challenge*. ArXiv preprint :2011.04547, 2020.

[Chiu 2018] C-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski et M. Bacchiani. *State-of-the-art Speech Recognition With Sequence-to-Sequence Models*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4774–4778, 2018.

[Cho 2014] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk et Y. Bengio. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. ArXiv preprint :1406.1078, 2014.

[Cho 2018] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe et T. Hori. *Multilingual Sequence-to-Sequence Speech Recognition : Architecture, Transfer Learning, and Language Modeling*. In IEEE Spoken Language Technology Workshop (SLT), pages 521–527, 2018.

[Chorowski 2014] J. Chorowski, D. Bahdanau, K. Cho et Y. Bengio. *End-to-end Continuous Speech Recognition using Attention-based Recurrent NN : First results*. In Proc. of the International Conference on Neural Information Processing Systems (NIPS) : Workshop on Deep Learning, pages 1–10, 2014.

[Chorowski 2015] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho et Y. Bengio. *Attention-Based Models for Speech Recognition*. In Proc. of the International Conference on Neural Information Processing Systems (NIPS), page 577–585. MIT Press, 2015.

[Claus 2013] F. Claus, H. Gamboa-Rosales, R. Petrick, H-U. Hain et R. Hoffmann. *A Survey about Databases of Children's Speech*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Lyon, 2013.

[Cohen 2011] W. Cohen et C. Anderson. *Identification of Phonological Processes in Preschool Children's Single-word Productions*. International journal of language & communication disorders, vol. 46, pages 481–488, 07 2011.

[Cole 2006a] R. Cole, P. Hosom et Pellom B. *University of Colorado Prompted and Read Children's Speech Corpus*. Technical Report TR-CSLR-2006-02, 2006.

[Cole 2006b] R. Cole et B. Pellom. *University of Colorado Read and Summarized Story Corpus*. Technical Report TR-CSLR-2006-03, 2006.

[Cole 2019] R. Cole, W. Ward et S. Pradhan. *My Science Tutor and the MyST Corpus*, 02 2019.

[Conneau 2020] A. Conneau, A. Baevski, R. Collobert, A. Mohamed et M. Auli. *Unsupervised Cross-lingual Representation Learning for Speech Recognition.* ArXiv preprint :2006.13979, 2020.

[Davis 1980] S. W. Davis et P. Mermelstein. *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Se.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980.

[Dehak 2011] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel et P. Ouellet. *Front-End Factor Analysis for Speaker Verification.* IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), vol. 19, no. 4, pages 788–798, 2011.

[Detey 2016] S. Detey, L. Fontan et T. Pellegrini. *Traitement de la prononciation en langue étrangère : approches didactiques, méthodes automatiques et enjeux pour l'apprentissage.* Revue TAL, vol. 57, no. 3, pages 15–39, 2016.

[Dong 2018] L. Dong, S. Xu et B. Xu. *Speech-Transformer : A No-Recurrence Sequence-to-Sequence Model for Speech Recognition.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5884–5888, 2018.

[Duan 2020] R. Duan, T. Kawahara, M. Dantsuji et H. Nanjo. *Cross-Lingual Transfer Learning of Non-Native Acoustic Modeling for Pronunciation Error Detection and Diagnosis.* IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), vol. 28, pages 391–401, 2020.

[Duchateau 2009] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquière, W. Verhelst et H. Van hamme. *Developing a Reading Tutor : Design and Evaluation of Dedicated Speech Recognition and Synthesis Modules.* Speech Communication, vol. 51, no. 10, pages 985–994, 2009. Spoken Language Technology for Education.

[Dégardin 1857] F. Dégardin. Les homonymes et les homographes de la langue française. Vve Maire-Nyon, 1857.

[Eguchi 1969] S Eguchi et I. J. Hirsh. Development of speech sounds in children, volume 257. Acta Oto-Laryngol, 1969.

[Elenius 2005] D. Elenius et M. Blomberg. *Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 year old Children.* In Proc. Interspeech 2005, pages 2749–2752, 2005.

[Ellis 2001] D.P.W. Ellis, R. Singh et S. Sivadas. *Tandem Acoustic Modeling in Large-Vocabulary Recognition.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 517–520, 02 2001.

[Ellis 2009] J. B. Ellis et P. R. Ramig. *A Handbook on Stuttering.* Journal of Fluency Disorders, vol. 34, no. 4, pages 295–299, 2009.

[Eskenazi 1996] M. Eskenazi. *Kids : a Database of Children's Speech.* The Journal of the Acoustical Society of America, vol. 100, no. 4, 1996.

[Eskenazi 1997] M. Eskenazi et J. Mostow. *The CMU KIDS Speech Corpus (LDC97S63)*, 1997. Linguistic Data Consortium, University of Pennsylvania.

[Feng 2020] Y. Feng, G. Fu, Q. Chen et K. Chen. *SED-MDD : Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3492–3496, 2020.

[Fringi 2015] E. Fringi, J. Fain Lehman et M. J. Russell. *Evidence of Phonological Processes in Automatic Recognition of Children's Speech*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, pages 1621–1624, 2015.

[Gales 1998] M.J.F. Gales. *Maximum Likelihood Linear Transformations for HMM-based Speech Recognition*. Computer Speech & Language, vol. 12, no. 2, pages 75–98, 1998.

[Gales 2008] M. Gales et S. Young. *The Application of Hidden Markov Models in Speech Recognition*. Foundations and Trends® in Signal Processing, vol. 1, no. 3, pages 195–304, 2008.

[Garofolo 1993a] J. Garofolo, D. Graff, D. Paul et D. S. Pallett. *CSR-I (WSJ0) Complete LDC93S6A*, 1993. Web Download. Linguistic Data Consortium.

[Garofolo 1993b] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus et D. Pallett. *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*. NASA STI/Recon Technical Report N, vol. 93, page 27403, 01 1993.

[Gelin 2020a] Lucile Gelin, Morgane Daniel, Thomas Pellegrini et Julien Pinquier. *Babble Noise Augmentation for Phone Recognition applied to Children Reading Aloud in a Classroom Environment*. Speech in Noise Workshop (SPiN 2020), 2020. Poster.

[Gelin 2020b] Lucile Gelin, Morgane Daniel, Thomas Pellegrini et Julien Pinquier. *Reconnaissance de phones fondée sur du Transfer Learning pour des enfants apprenants lecteurs en environnement de classe*. In Conférence conjointe Journées d'Études sur la Parole (JEP), Traitement Automatique des Langues Naturelles (TALN), et Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), volume 1, pages 253–261, Nancy, France, 2020. ATALA.

[Gelin 2021a] Lucile Gelin, Morgane Daniel, Julien Pinquier et Thomas Pellegrini. *End-to-end Acoustic Modelling for Phone Recognition of Young Readers*. Speech Communication, vol. 134, pages 71–84, 2021.

[Gelin 2021b] Lucile Gelin, Thomas Pellegrini, Julien Pinquier et Morgane Daniel. *Simulating Reading Mistakes for Child Speech Transformer-Based Phone Recognition*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, pages 3860–3864, 2021.

[Gerosa 2006a] M. Gerosa, D. Giuliani et S. Narayanan. *Acoustic Analysis and Automatic Recognition of Spontaneous Children's Speech*. In International Conference on Spoken Language Processing, volume 4, pages 1886–1889, 2006.

[Gerosa 2006b] M. Gerosa, S. Lee, D. Giuliani et S. Narayanan. *Analyzing Children's Speech : An Acoustic Study of Consonants and Consonant-Vowel Transition*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages I – I, 06 2006.

[Gibson 2018] M. Gibson, C. Plahl, P. Zhan et G. Cook. *Multi-condition Deep Neural Network Training*. Studientexte zur Sprachkommunikation : Elektronische Sprachsignalverarbeitung, pages 77–84, 2018.

[Godde 2017] E. Godde, G. Bailly, D. Escudero, M-L. Bosse, M. Bianco et C. E. Vilain. *Improving Fluency of Young Readers : introducing a Karaoke to learn how to breath during a Reading-while-Listening task*. In ISCA Workshop on Speech and Language Technology in Education (SLaTE), pages 127–131, 2017.

[Godde 2019] E. Godde, G. Bailly et M-L. Bosse. *Un Karaoké pour Entraîner Prosodie et Compréhension en Lecture*. In Environnements Informatiques pour l'Apprentissage Humain (EIAH), Paris, France, 2019.

[Godfrey 1993] J. J. Godfrey et E. Holliman. *Switchboard-1 Release 2 LDC97S62*, 1993. Web Download. Linguistic Data Consortium.

[Goodfellow 2016] I. Goodfellow, Y. Bengio et A. Courville. Deep learning. MIT Press, 2016.

[Graves 2006] A. Graves, S. Fernández, F. Gomez et J. Schmidhuber. *Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks*. In Proc. of the International Conference on Machine learning (ICML), pages 369–376, 2006.

[Graves 2008] A. Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. PhD thesis, Technische Universität München, 2008.

[Graves 2012] A. Graves. *Sequence Transduction with Recurrent Neural Networks*. ArXiv preprint :1211.3711, 2012.

[Graves 2013] A. Graves, A. Mohamed et G. Hinton. *Speech Recognition with Deep Recurrent Neural Networks*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6645–6649, 2013.

[Graves 2014a] A. Graves. *Generating Sequences With Recurrent Neural Networks*. ArXiv preprint :1308.0850, 2014.

[Graves 2014b] A. Graves et N. Jaitly. *Towards End-to-End Speech Recognition with Recurrent Neural Networks*. In Proc. of the International Conference on Machine learning (ICML), volume 32, pages 1764–1772, 2014.

[Gray 2014] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner et N. Bodenstab. *Child automatic speech recognition for US English : child interaction with living-room-electronic-devices*. In Proc. of the International Workshop on Child Computer Interaction (WOCCI), pages 21–26, 2014.

[Gulati 2020] A. Gulati, J. Qin, C-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu*et al. Conformer : Convolution-augmented Transformer for Speech Recognition*. ArXiv preprint :2005.08100, 2020.

[Gutmann 2010] M. Gutmann et A. Hyvärinen. *Noise-contrastive estimation : A new estimation principle for unnormalized statistical models*. In Proc. of the International Conference on Artificial Intelligence and Statistics, pages 297–304, 2010.

[Hagen 2003] A. Hagen, B. Pellom et R. Cole. *Children's Speech Recognition with application to Interactive Books and Tutors*. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 186 – 191, 01 2003.

[Hajji 2019] Mohamed El Hajji, Morgane Daniel et Lucile Gelin. *Transfer Learning based Audio Classification for a Noisy and Speechless Recordings Detection Task, in a classroom*

context. In ISCA Workshop on Speech and Language Technology in Education (SLaTE), pages 109–113, Graz, Austria, Septembre 2019. ISCA.

[He 2016] K. He, X. Zhang, S. Ren et J. Sun. *Deep Residual Learning for Image Recognition*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[Heba 2019] A. Heba, T. Pellegrini, J-P. Lorré et R. André-Obrecht. *Char+CV-CTC : Combining Graphemes and Consonant/Vowel Units for CTC-Based ASR Using Multitask Learning*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, pages 1611–1615, 2019.

[Hembise 2021] Corentin Hembise, Lucile Gelin et Morgane Daniel. *Lalilo : a Reading Assistant for Children featuring Speech Recognition-based Reading Mistake Detection*. In Annual Conference of the International Speech Communication Association (INTERSPEECH), Show & Tell contribution, Brno, Czech Republic, Août 2021.

[Hermansky 1991] H. Hermansky et L.A. Cox. *Perceptual Linear Predictive (PLP) Analysis-Resynthesis Technique*. In IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, pages 37–38, 1991.

[Hermansky 1992] H. Hermansky, N. Morgan, A. Bayya et P. Kohn. *RASTA-PLP speech analysis technique*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 121 – 124 vol.1, 04 1992.

[Hermansky 1999] H. Hermansky et S. Sharma. *Temporal patterns (TRAPs) in ASR of noisy speech*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 289–292, 1999.

[Hinton 1983] G. E. Hinton et T. J. Sejnowski. *Optimal Perceptual Inference*. In Proc. of the IEEE conference on Computer Vision and Pattern Recognition, volume 448. Citeseer, 1983.

[Hinton 2006] G. E. Hinton, S. Osindero et Y-W. Teh. *A Fast Learning Algorithm for Deep Belief Nets*. Neural Comput., vol. 18, no. 7, page 1527–1554, 2006.

[Hinton 2012] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever et R. R. Salakhutdinov. *Improving Neural Networks by preventing Co-Adaptation of Feature Detectors*. ArXiv preprint :1207.0580, 2012.

[Hirano 1981] M. Hirano, S. Kurita et T. Nakashima. *The Structure of the Vocal Folds*. Vocal Fold Physiology, pages 33—-41, 1981.

[Hochreiter 1997] S. Hochreiter et J. Schmidhuber. *Long Short-term Memory*. Neural computation, vol. 9, pages 1735–80, 12 1997.

[Hori 2017a] T. Hori, S. Watanabe et J. Hershey. *Joint CTC/attention decoding for end-to-end speech recognition*. In Proc. of the Annual Meeting of the Association for Computational Linguistics, pages 518–529, 2017.

[Hori 2017b] T. Hori, S. Watanabe, Y. Zhang et W. Chan. *Advances in Joint CTC-Attention Based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Sweden, pages 949–953, 2017.

[Jelinek 1976] F. Jelinek. *Continuous Speech Recognition by Statistical Methods.* Proceedings of the IEEE, vol. 64, no. 4, pages 532–556, 1976.

[Jiang 2018] H. Jiang, B. Kim, M. Guan et M. Gupta. *To Trust Or Not To Trust A Classifier.* In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi et R. Garnett, editeurs, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.

[Joshi 2013] V. Joshi, N.V. Prasad et S. Umesh. *Modified cepstral mean normalization - Transforming to utterance specific non-zero mean.* In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Lyon, pages 881–885, 01 2013.

[Juang 1985] B.-H. Juang. *Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains.* AT & T Technical Journal, vol. 64, no. 6, pages 1235–1249, 1985.

[Kanters 2009] S. Kanters, C. Cucchiarini et H. Strik. *The Goodness of Pronunciation Algorithm : a detailed Performance Study.* In ISCA Workshop on Speech and Language Technology in Education (SLaTE), 2009.

[Karita 2019a] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa et T. Nakatani. *Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration.* In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, pages 1408–1412, 2019.

[Karita 2019b] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin et R. Yamamoto. *A Comparative Study on Transformer vs RNN in Speech Applications.* IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), no. April 2020, pages 449–456, 2019.

[Katsamanis 2011] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein et S. Narayanan. *SailAlign : Robust long speech-text alignment.* In Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research, 01 2011.

[Kazemzadeh 2005] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Andersen, S. Narayanan et A. Alwan. *TBALL Data Collection : the making of a Young Children's Speech Corpus.* In Proc. of the European Conference on Speech Communication and Technology (Eurospeech), pages 1581–1584, 09 2005.

[Kenny 2007] P. Kenny, G. Boulianne, P. Ouellet et P. Dumouchel. *Joint Factor Analysis Versus Eigenchannels in Speaker Recognition.* IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pages 1435–1447, 2007.

[Kent 1976] R. D. Kent. *Anatomical and Neuromuscular Maturation of the Speech Mechanism : Evidence from Acoustic Study.* Journal of Speech and Hearing Research, vol. 19, no. 3, pages 421–445, 1976.

[Kent 1980] R. D. Kent et L. L. Forner. *Speech Segment Durations in Sentence Recitations by Children and Adults.* Journal of Phonetics, vol. 8, no. 2, pages 157–168, 1980.

[Kent 1992] R. D Kent et C. Read. The acoustic analysis of speech. Singular Publising Group, 1992.

[Kim 2011] Y-S. Kim, R. Wagner et E. Foster. *Relations Among Oral Reading Fluency, Silent Reading Fluency, and Reading Comprehension : A Latent Variable Study of First-Grade Readers.* Scientific Studies of Reading, vol. 15, pages 338–362, 07 2011.

[Kingma 2017] D. P. Kingma et J. Ba. *Adam : A Method for Stochastic Optimization.* ArXiv preprint :1412.6980, 2017.

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

[Lecun 1995] Y. Lecun et Y. Bengio. Convolutional networks for images, speech, and time-series. MIT Press, 1995.

[Lee 1996] L. Lee et R.C. Rose. *Speaker Normalization using Efficient Frequency Warping Procedures.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 353–356 vol. 1, 1996.

[Lee 1997] S. Lee, A. Potamianos et S. Narayanan. *Analysis of Children's Speech : Duration, Pitch and Formants.* In Proc. of the European Conference on Speech Communication and Technology (Eurospeech), pages 473–476, 1997.

[Lee 1999] S. Lee, A. Potamianos et S. S. Yegna Narayanan. *Acoustics of Children's Speech : Developmental Changes of Temporal and Spectral Parameters.* The Journal of the Acoustical Society of America, vol. 105, no. 3, pages 1455–1468, 1999.

[Leung 2019] W-K. Leung, X. Liu et H. Meng. *CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8132–8136, 2019.

[Levenshtein 1966] V. I. Levenshtein. *Binary Codes capable of Correcting Deletions, Insertions and Reversals.* Soviet Physics Doklady, vol. 10, no. 8, pages 707–710, 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[Levinson 1983] S. E. Levinson, L. R. Rabiner et M. M. Sondhi. *An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition.* The Bell System Technical Journal, vol. 62, no. 4, pages 1035–1074, 1983.

[Levy 1995] B.A. Levy, J. Campsal, J. Browne, D. Cooper, C. Waterhouse et C. Wilson. *Reading fluency : Episodic integration across texts.* Journal of Experimental Psychology : Learning, Memory & Cognition, vol. 21, pages 1169–1185, 1995.

[Li 2001] Q. Li et M. J. Russell. *Why is automatic recognition of children's speech difficult ?* In Proc. of the European Conference on Speech Communication and Technology (Eurospeech), pages 2671–2674, 2001.

[Li 2014a] J. Li, L. Deng, Y. Gong et R. Haeb-Umbach. *An Overview of Noise-Robust Automatic Speech Recognition.* IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), vol. 22, no. 4, pages 745–777, 2014.

[Li 2014b] K. Li et H. Meng. *Mispronunciation Detection and Diagnosis in L2 English Speech using Multi-distribution Deep Neural Networks*. In International Symposium on Chinese Spoken Language Processing, pages 255–259, 2014.

[Li 2020] J. Li, Y. Wu, Y. Gaur, C. Wang, R. Zhao et S. Liu. *On the Comparison of Popular End-to-End Models for Large Scale Speech Recognition*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, pages 1–5, 2020.

[Liao 2015] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays et M. Bacchiani. *Large Vocabulary Automatic Speech Recognition for Children*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, pages 1611–1615, 2015.

[Lippmann 1990] Richard P. Lippmann. *Review of Neural Networks for Speech Recognition*. In A. Waibel et K-F. Lee, editeurs, Readings in Speech Recognition, pages 374–392. Morgan Kaufmann, San Francisco, 1990.

[Lu 2015] L. Lu, X. Zhang, K. Cho et S. Renals. *A Study of the Recurrent Neural Network Encoder-Decoder for Large Vocabulary Speech Recognition*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, volume 2015-January, pages 3249–3253, 2015.

[Lust 2006] B. Lust. Child language : Acquisition and growth. Cambridge University Press, 01 2006.

[Matassoni 2018] M. Matassoni, R. Gretter, D. Falavigna et D. Giuliani. *Non-Native Children Speech Recognition Through Transfer Learning*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6229–6233, 2018.

[Metallinou 2014] A. Metallinou et J. Cheng. *Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Singapore, pages 1468–1472, 2014.

[Mihaylova 2019] T. Mihaylova et A. F. T. Martins. *Scheduled Sampling for Transformers*. In Proc. of the Annual Meeting of the Association for Computational Linguistics : Student Research Workshop, pages 351–356. Association for Computational Linguistics, Juillet 2019.

[Mohamed 2011] A-R. Mohamed, G. E. Dahl et G. Hinton. *Acoustic Modeling Using Deep Belief Networks*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pages 14–22, 2011.

[Mohri 2008] M. Mohri, F. Pereira et M. Riley. Speech recognition with weighted finite-state transducers, pages 559–584. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[Morgan 1990] N. Morgan et H. Bourlard. *Continuous Speech Recognition using Multilayer Perceptrons with Hidden Markov Models*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 413–416, 1990.

[Moritz 2020] N. Moritz, T. Hori et J. Le Roux. *Streaming automatic speech recognition with the transformer model*. ArXiv preprint :2001.02674, 2020.

[Mostow 2001] J. Mostow et G. Aist. *Evaluating Tutors that Listen : An Overview of Project LISTEN*. In Smart machines in education : The coming revolution in educational technology., pages 169–234. The MIT Press, 2001.

[Mugitani 2012] R. Mugitani et S. Hiroya. *Development of Vocal Tract and Acoustic Features in Children*. The Journal of the Acoustical Society of Japan, vol. 68, no. 5, pages 234–240, 2012.

[Nair 2010] V. Nair et G. E. Hinton. *Rectified Linear Units Improve Restricted Boltzmann Machines*. In Proc. of the International Conference on Machine learning (ICML), ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress.

[Ng 2020] S-I. Ng, W. Liu, Z. Peng, S. Feng, H-P. Huang, O. Scharenborg et T. Lee. *The CUHK-TUDELFT System for The SLT 2021 Children Speech Recognition Challenge*. ArXiv preprint :2011.06239, 2020.

[Panayotov 2015] V. Panayotov, G. Chen, D. Povey et S. Khudanpur. *Librispeech : An ASR corpus based on public domain audio books*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210, 2015.

[Pascual 2019] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte et Y. Bengio. *Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, pages 161–165, 2019.

[Patterson 1992] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang et M. Allerhand. *Complex Sounds and Auditory Images*. In Auditory Physiology and Perception, 1992.

[Pelanek 2016] R. Pelanek. *Applications of the Elo Rating System in Adaptive Educational Systems*. Computers & Education, 2016.

[Pellom 2001] B. Pellom. *Sonic : The University of Colorado Continuous Speech Recognizer*. Technical Report TR-CSLR-2001-01, 2001.

[Potamianos 1997] A. Potamianos, S. Narayanan et S. Lee. *Automatic Speech Recognition for Children*. In Proc. of the European Conference on Speech Communication and Technology (Eurospeech), 01 1997.

[Potamianos 1998] A. Potamianos et S. Narayanan. *Spoken Dialog Systems for Children*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 197–200 vol.1, 1998.

[Potamianos 2003] A. Potamianos et S. Narayanan. *Robust Recognition of Children's Speech*. IEEE Transactions on Speech and Audio Processing, vol. 11, no. November 2003, pages 603–616, 2003.

[Potamianos 2007] A. Potamianos et S. Narayanan. *A Review of the Acoustic and Linguistic Properties of Children's Speech*. In IEEE Workshop on Multimedia Signal Processing, pages 22–25, 2007.

[Povey 2011] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer et K. Vesely. *The Kaldi Speech Recognition Toolkit*. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 1–4, 2011.

[Povey 2016] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang et S. Khudanpur. *Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), San Francisco, pages 2751–2755, 2016.

[Povey 2018] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi et S. Khudanpur. *Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyberabad, pages 3743–3747, 2018.

[Prabhavalkar 2017] R. Prabhavalkar, K. Rao, T. Sainath, B. Li, L. M. Johnson et N. Jaitly. *A Comparison of Sequence-to-Sequence Models for Speech Recognition*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, 2017.

[Proença 2018] J. D. L. Proença. *Automatic Assessment of Reading Ability of Children*. PhD thesis, University of Coimbra, 2018.

[Qian 2016] Y. Qian, X. Wang, K. Evanini et D. Suendermann-Oeft. *Improving DNN-Based Automatic Recognition of Non-native Children Speech with Adult Speech*. In Proc. of the International Workshop on Child Computer Interaction (WOCCI), pages 40–44, 2016.

[Rabiner 1983] L. R. Rabiner, S. E. Levinson et M. M. Sondhi. *On the application of Vector Quantization and Hidden Markov Models to Speaker-independent, Isolated Word Recognition*. The Bell System Technical Journal, vol. 62, no. 4, pages 1075–1105, 1983.

[Rabiner 1986] L. Rabiner et B. Juang. *An introduction to Hidden Markov Models*. IEEE ASSP Magazine, vol. 3, no. 1, pages 4–16, 1986.

[Rajnoha 2009] J. Rajnoha. *Multi-Condition Training for Unknown Environment Adaptation in Robust ASR Under Real Conditions*. Acta Polytechnica, vol. 49, 01 2009.

[Ravanelli 2018] M. Ravanelli et Y. Bengio. *Speaker recognition from raw waveform with sincnet*. In IEEE Spoken Language Technology Workshop (SLT), pages 1021–1028. IEEE, 2018.

[Ravanelli 2019] M. Ravanelli et Y. Bengio. *Learning Speaker Representations with Mutual Information*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, pages 1153–1157, 2019.

[Ravanelli 2020] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal et Y. Bengio. *Multi-Task Self-Supervised Learning for Robust Speech Recognition*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 01 2020.

[Ravanelli 2021] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J-C. Chou, S-L. Yeh, S-W. Fu, C-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori et Y. Bengio. *SpeechBrain : A General-Purpose Speech Toolkit*, 2021. arXiv :2106.04624.

[Ravishankar 1996] M. K. Ravishankar. *Efficient Algorithms for Speech Recognition*. PhD thesis, Carnegie Mellon University, 1996.

[Roehrig 2008] A. Roehrig, Y. Petscher, S. Nettles, R. Hudson et J. Torgesen. *Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes*. Journal of school psychology, vol. 46, pages 343–66, 07 2008.

[Ros-Dupont 1999] M. Ros-Dupont. La lecture à haute voix : du CP au CM2. Bordas, 1999.

[Rumelhart 1986] D. E. Rumelhart, G. E. Hinton et R. J. Williams. *Learning representations by back-propagating errors*. Nature, vol. 323, pages 533–536, 1986.

[Sabatini 2019] J. Sabatini, Z. Wang et T. O'Reilly. *Relating Reading Comprehension to Oral Reading Performance in the NAEP Fourth-Grade Special Study of Oral Reading*. Reading Research Quarterly, vol. 54, pages 253–271, 03 2019.

[Saon 2013] G. Saon, H. Soltau, D. Nahamoo et M. Picheny. *Speaker Adaptation of Neural Network Acoustic Models using i-vectors*. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 55–59, 2013.

[Schluter 2007] R. Schluter, I. Bezrukov, H. Wagner et H. Ney. *Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 4, pages 649–652, 2007.

[Schneider 2019] S. Schneider, A. Baevski, R. Collobert et M. Auli. *Wav2vec : Unsupervised Pre-Training for Speech Recognition*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, pages 3465–3469, 2019.

[Schroeder 1977] M.R. Schroeder. *Recognition of Complex Acoustic Signals*. In T.H. Bullock, editeur, Life Sciences Research Report, volume 5, pages 323–328, 1977.

[Schwartz 1985] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner et J. Makhoul. *Context-dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech*. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 10, pages 1205–1208, 1985.

[Schwartz 2009] P. Schwartz. *Phoneme Recognition based on Long Temporal Context*. PhD thesis, Brno University of Technology, 2009.

[Senior 2014] A. Senior et I. Lopez-Moreno. *Improving DNN Speaker Independence with I-vector Inputs*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 225–229, 2014.

[Senior 2015] A. Senior, H. Sak, F. de Chaumont Quitry, T. Sainath et K. Rao. *Acoustic modelling with CD-CTC-SMBR LSTM RNNS*. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 604–609, 2015.

[Sereno 1985] J.A. Sereno, S.R. Baum, G.C. Marean et P Lieberman. *Acoustic Analyses and Perceptual Data on Anticipatory Labial Coarticulation in Adults and Children*. The Journal of the Acoustical Society of America, vol. 77, 1985.

[Serizel 2014a] R. Serizel et D. Giuliani. *Deep Neural Network Adaptation for Children's and Adults' Speech Recognition*. In Proc. of the Italian Computational Linguistics Conference (CLiC-it), pages 137–140, 2014.

[Serizel 2014b] R. Serizel et D. Giuliani. *Vocal Tract Length Normalisation Approaches to DNN-based Children's and Adults' Speech Recognition*. In IEEE Spoken Language Technology Workshop (SLT), pages 135–140, 2014.

[Shahnawazuddin 2016] S. Shahnawazuddin, A. Dey et R. Sinha. *Pitch-Adaptive Front-End Features for Robust Children's ASR*. In Proc. Interspeech 2016, pages 3459–3463, 2016.

[Shivakumar 2014] P. G. Shivakumar, A. Potamianos, S. Lee et S. S. Narayanan. *Improving Speech Recognition for Children using Acoustic Adaptation and Pronunciation Modeling*. In Proc. of the International Workshop on Child Computer Interaction (WOCCI), 2014.

[Shivakumar 2020] P. G. Shivakumar et P. Georgiou. *Transfer Learning from Adult to Children for Speech Recognition : Evaluation, Analysis and Recommendations*. Computer Speech & Language, vol. 63, page 101077, 2020.

[Shivakumar 2021] P. G. Shivakumar et S. Narayanan. *End-to-End Neural Systems for Automatic Children Speech Recognition : An Empirical Study*. ArXiv preprint :2102.09918, 2021.

[Shobaki 2000] K. Shobaki, J-P. Hosom et R. Cole. *The OGI Kids' Speech Corpus and Recognizers*. In Proc. of the International Conference on Spoken Language Processing (ICSLP), pages 564–567, 2000.

[Shobaki 2007] K. Shobaki, J-P. Hosom et R. Cole. *CSLU : Kid's Speech Version 1.1 LDC2007S18*, 2007. Web Download. Philadelphia : Linguistic Data Consortium.

[Smith 1978] B. Smith. *Temporal aspects of English Speech Production : A Developmental Perspective*. Journal of Phonetics, vol. 6, no. 1, pages 37–67, 1978.

[Stoddard 1993] K. Stoddard, G. Valcante, P. Sindelar, L. O'Shea et B. Algozzine. *Increasing Reading Rate and Comprehension : The Effects of Repeated Readings, Sentence Segmentation, and Intonation Training*. Reading Research and Instruction, vol. 32, no. 4, pages 53–65, 1993.

[Sutskever 2014] I. Sutskever, O. Vinyals et Q. V. Le. *Sequence to Sequence Learning with Neural Networks*. In Proc. of the International Conference on Neural Information Processing Systems (NIPS), page 3104–3112, Cambridge, MA, USA, 2014.

[Tepperman 2007] J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan et S. Narayanan. *A Bayesian Network Classifier for Word-level Reading Assessment*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Antwerp, pages 2185–2188, 08 2007.

[Thiemann 2013] J. Thiemann, N. Ito et E. Vincent. *DEMAND : a collection of multi-channel recordings of acoustic noise in diverse environments*, 2013.

[Ting 2004] H. N. Ting et J. Yunus. *Speaker-independent Malay Vowel Recognition of Children using Multi-layer Perceptron*. In IEEE Region 10 Conference TENCON, volume A, pages 68–71 Vol. 1, 2004.

[Tjandra 2017] A. Tjandra, S. Sakti et S. Nakamura. *Local Monotonic Attention Mechanism for End-to-End Speech and Language Processing*. ArXiv preprint :1705.08091, 2017.

[Tong 2017a] R. Tong, L. Wang et B. Ma. *Transfer Learning for Children's Speech Recognition*. Proc. of the International Conference on Asian Language Processing (IALP), pages 36–39, 2017.

[Tong 2017b] S. Tong, P. Garner et H. Bourlard. *Multilingual Training and Cross-lingual Adaptation on CTC-based Acoustic Model*. Speech Communication, vol. 104, 2017.

[Vaswani 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, A. N. Jones L.and Gomez, L. Kaiser et I. Polosukhin. *Attention is All You Need*. In Proc. of the International Conference on Neural Information Processing Systems (NIPS), page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Verma 2015] P Verma et P Das. *i-Vectors in Speech Processing Applications : A Survey*. International Journal of Speech Technology, vol. 18, pages 529–546, 12 2015.

[Veselý 2013] K. Veselý, A. Ghoshal, L. Burget et D. Povey. *Sequence-discriminative training of deep neural networks*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Lyon, pages 2345–2349, 2013.

[Viikki 1998] O. Viikki et K. Laurila. *Cepstral domain segmental feature vector normalization for noise robust speech recognition*. Speech Communication, vol. 25, no. 1, pages 133–147, 1998.

[Vinyals 2015a] O. Vinyals et Q. Le. *A Neural Conversational Model*. Proc. of the International Conference on Machine learning (ICML) : Deep Learning Workshop, 06 2015.

[Vinyals 2015b] O. Vinyals, A. Toshev, S. Bengio et D. Erhan. *Show and tell : A neural image caption generator*. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3156–3164, 2015.

[Vygotsky] L.S. Vygotsky. Mind in society : Development of higher psychological processes. Harvard University Press.

[Waibel 1989] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano et K. J. Lang. *Phoneme Recognition using Time-Delay Neural Networks*. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pages 328–339, 1989.

[Watanabe 2017] S. Watanabe, T. Hori, S. Kim, J. R. Hershey et T. Hayashi. *Hybrid CTC/Attention Architecture for End-to-End Speech Recognition*. IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pages 1240–1253, 2017.

[Weiss 2018] G. Weiss, Y. Goldberg et E. Yahav. *On the Practical Computational Power of Finite Precision RNNs for Language Recognition*. ArXiv preprint :1805.04908, 2018.

[Wilcoxon 1945] F. Wilcoxon. *Individual Comparisons by Ranking Methods*. Biometrics Bulletin, vol. 1, no. 6, pages 80–83, 1945.

[Witt 2000] S.M Witt et S.J Young. *Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning*. Speech Communication, vol. 30, no. 2, pages 95–108, 2000.

[Wu 2019] F. Wu, P. Garcia, D. Povey et S. Khudanpur. *Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, pages 1–5, 2019.

[Xu 2015] K. Xu, J. L. Ba, R. Kiros, K. Cho, Aaron Courville, R. Salakhutdinov, R. S. Zemel et Y. Bengio. *Show, Attend and Tell : Neural Image Caption Generation with Visual Attention*. In Proc. of the International Conference on Machine learning (ICML), page 2048–2057. JMLR.org, 2015.

[Yeung 2018] G. Yeung et A. Alwan. *On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children*. In Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, pages 1661–1665, 2018.

[Yong 2011] B. F. Yong et H. N. Ting. *Speaker-Independent Vowel Recognition for Malay Children Using Time-Delay Neural Network*. In 5th Kuala Lumpur International Conference on Biomedical Engineering 2011, pages 565–568. Springer Berlin Heidelberg, 2011.

[Yu 2021] F. Yu, Z. Yao, X. Wang, K. An, L. Xie, Z. Ou, B. Liu, X. Li et G. Miao. *The SLT 2021 Children Speech Recognition Challenge : Open datasets, rules and baselines*. In IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, Janvier 2021.

[Yuliani 2017] A. Yuliani, R. Sustika, R. Yuwana et H. Pardede. *Feature transformations for robust speech recognition in reverberant conditions*. In Proc. of the International Conference on Computer, Control, Informatics and its Applications (IC3INA), pages 57–62, 10 2017.

[Zharkova 2015] N. Zharkova, W. Hardcastle, F. Gibbon et R. Lickley. *Development of Lingual Motor Control in Children and Adolescents*. In Proc. of the International Congress of Phonetic Sciences (ICPhS), 2015.

[Zolnay 2005] A. Zolnay, R. Schluter et H. Ney. *Acoustic feature combination for robust speech recognition*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 457–460, 2005.

Résumé

À travers ces travaux de thèse, nous cherchons à perfectionner les transcriptions phonétiques de lectures orales d'enfants apprenant·e·s lecteur·rice·s réalisées en environnement scolaire. Ces transcriptions automatiques forment la base d'un système de détection d'erreurs utilisé dans un exercice de lecture orale de la plateforme pédagogique Lalilo. Une bonne précision est primordiale pour fournir un retour adapté à l'enfant, et ainsi favoriser son apprentissage.

Une première partie présente les principaux défis de notre tâche. La reconnaissance automatique de la parole d'enfants est plus ardue que celle d'adultes, en raison de ses très grandes variabilités acoustique et prosodique. La rareté des données disponibles, notamment en français, nous oblige de plus à redoubler d'inventivité pour en modéliser correctement la variabilité. Enfin, de fréquentes occurrences d'erreurs de fluence et de déchiffrage, ainsi que la présence de bruit de brouhaha typique des salles de classe, constituent des difficultés supplémentaires.

Nous construisons dans une seconde partie un modèle acoustique hybride TDNNF-HMM, qui deviendra notre modèle de référence. Son entraînement via un apprentissage par transfert permet de pallier au manque de données et d'atteindre un PER de 30,1%. Nous étudions différents paramètres acoustiques et méthodes de normalisation, visant à maximiser la performance de notre modèle. Une technique d'augmentation de données par ajout de bruit, visant à améliorer la robustesse du modèle aux bruits de salle de classe, apporte également une amélioration relative du PER de 6,4%.

Dans notre dernière partie, nous explorons les architectures récentes *end-to-end* fondées sur des réseaux RNN, des modules CTC et des mécanismes d'attention. Notre travail est l'un des premiers à appliquer des architectures *end-to-end* sur de la parole d'enfants, et à analyser leurs forces et faiblesses quant aux spécificités de la lecture orale d'apprenant·e·s lecteur·rice·s. Notre système Transformer+CTC fournit les meilleurs résultats (PER de 25,0%) grâce à la pertinence des informations acoustiques et textuelles extraites par ses mécanismes d'auto-attention et à la complémentarité des modules CTC et d'attention. Notre système est ensuite enrichi de techniques d'augmentation de données. Nous introduisons notamment une méthode novatrice de simulation d'erreurs de lecture, afin d'entraîner le modèle à mieux les détecter. Celle-ci s'avère complémentaire à l'augmentation par ajout de bruit étudiée en deuxième partie. Ces deux techniques permettent alors au Transformer+CTC de surpasser largement le modèle hybride de référence, avec un PER de 21,2%, et d'améliorer la qualité de ses transcriptions sur de la lecture incorrecte ou en présence de bruit de brouhaha.

Mots clés : reconnaissance automatique de phones, parole d'enfant, modélisation acoustique *end-to-end*, peu de données, erreurs de lecture, bruit de brouhaha

Abstract

In this PhD thesis, we aim at perfecting the phonetic transcriptions of oral readings of children learning to read, recorded in a classroom environment. These automatic transcriptions power a reading mistakes detection system used in the reading aloud exercise of the Lalilo pedagogical platform. Good accuracy is essential to provide appropriate feedback to the child, thus promoting his·her learning.

A first section presents the main challenges of our task. The automatic recognition of children's speech is more difficult than adults' speech, due to its very high acoustic and prosodic variability. The scarcity of available data, especially in French, requires us to be more inventive as to correctly model its variability. Finally, frequent occurrences of fluency and decoding mistakes, as well as the presence of classroom babble noise, constitute additional difficulties.

In a second section, we build a hybrid TDNNF-HMM acoustic model, which will become our baseline model. Using transfer learning allows to overcome the lack of data and achieve a PER of 30.1%. We study different acoustic parameters and normalization methods, aiming at maximizing our model's performance. Data augmentation by adding noise with the objective of improving the model's robustness to classroom babble noise further improves the PER by 6.4% relative.

In our final section, we explore recent end-to-end architectures based on RNNs, CTC modules and attention mechanisms. Our work is one of the first to apply end-to-end architectures to child speech and to analyze their strengths and weaknesses with respect to the specificities of oral reading by children learning to read. Our Transformer+CTC system provides the best results (25.0% PER) thanks to the relevance of the acoustic and textual information extracted by its self-attention mechanisms and the complementarity of the CTC and attention modules. Our system is then enhanced with data augmentation techniques. In particular, we introduce an innovative method of simulating reading mistakes, that seeks to train the model to better detect them. It reveals complementary to the noise data augmentation previously studied. These two techniques then allow the Transformer+CTC to greatly outperform the hybrid reference model, with a PER of 21.2%, and to improve the quality of its transcriptions over misreadings or classroom babble noise.

Keywords : automatic phone recognition, child speech, end-to-end acoustic modelling, low data, reading mistakes, babble noise