Beyond words : leveraging language models for incremental and context-aware text-to-speech synthesis

Brooke Stephenson

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : EEATS - Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)
Spécialité : Signal Image Parole Télécoms
Unité de recherche : Grenoble Images Parole Signal Automatique

Au delà des mots: utilisation des modèles de langage pour une synthèse vocale incrementale et adaptable au contexte linguistique

Beyond words: leveraging language models for incremental and context-aware text-to-speech synthesis

Présentée par :

Brooke STEPHENSON

Direction de thèse :

Thomas HUEBER Directeur de recherche, CNRS Délégation Alpes	Directeur de thèse
Laurent BESACIER Ingénieur HDR, NAVER LABS Europe	Co-directeur de thèse
Laurent GIRIN Professeur des Universités, Grenoble INP	Co-encadrant de thèse

Rapporteurs :

Damien LOLIVE
PROFESSEUR DES UNIVERSITES, UNIVERSITE DE RENNES
Mireia FARRUS CABECERAN
ASSOCIATE PROFESSOR, Université de Barcelone

Thèse soutenue publiquement le **26 septembre 2023**, devant le jury composé de :

Thomas HUEBER DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directeur de thèse
Damien LOLIVE PROFESSEUR DES UNIVERSITES, UNIVERSITE DE RENNES	Rapporteur
Mireia FARRUS CABECERAN ASSOCIATE PROFESSOR, Université de Barcelone	Rapporteure
Olivier KRAIF PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Président
Joakim GUSTAFSON FULL PROFESSOR, KTH Royal Institute of Technology	Examinateur
Laurent BESACIER INGENIEUR HDR, NAVER LABS EUROPE	Co-directeur de thèse

Invités :

Laurent Girin
PROFESSEUR DES UNIVERSITES, Grenoble-INP

Résumé —

Cette thèse vise à améliorer les systèmes de synthèse vocale à partir du texte en ciblant deux axes, la réactivité et la qualité. En effet, les systèmes actuels présentent un délai important car l'utilisateur doit entrer une phrase complète avant que cette dernière ne puisse être synthétisée. Lorsque utilisé comme voix de substitution par une personne présentant un trouble de la communication, ces systèmes ne permettent donc pas une interaction fluide. De plus, les systèmes actuels exploitent exclusivement le texte de la phrase à synthétiser en ignorant le contexte linguistique général associé (par exemple les phrases précédentes). Dans cette thèse, nous proposons d'utiliser les modèles de langage basés sur des architectures de type Transformer pour (1) prédire le texte futur, à partir du texte déjà saisi, et ainsi débuter la synthèse d'un mot juste après sa saisie - on parlera de synthèse incrémentale, et (2) encoder le contexte linguistique général associé à la phrase à synthétiser pour améliorer la qualité prosodique de la synthèse - on parlera de synthèse adaptée au contexte.

Dans une première étude, nous étudions l'évolution des représentations internes d'un système TTS neuronal lorsque ce dernier synthétise un mot avec une connaissance seulement partielle des k mots à venir (le *lookahead*). Une analyse statistique (de type forêts aléatoires) est utilisée pour determiner quels sont les descripteurs linguistiques qui influent sur la stabilité de ces représentations internes. Enfin, nous complétons ces mesures objectives par un ensemble de tests perceptifs visant à quantifier la qualité prosodique en fonction du contexte linguistique considéré. Ces évaluations montrent que les systèmes TTS actuels exploitent un horizon d'environ 2 mots et que la stabilité des représentations internes associé à un mot dépend fortement de sa longueur.

Notre seconde contribution porte sur l'intégration, à un système TTS neuronal, d'un modèle de langage autoregressif, basé sur une architecture de type *Transformer* (tel que GPT) afin de prédire, au fur et à mesure de la saisie du texte, les mots suivants les plus probables. Les évaluations objectives et perceptives menées montrent que cette approche permet un bon compromis entre réactivité et naturel de la synthèse, mais reste très dépendante de la qualité de la prédiction du texte.

Notre troisième contribution porte sur l'amélioration générale de la prosodie d'un système TTS et plus spécifiquement sur la prédiction de la focalisation contrastive d'une part, et d'autre part sur la segmentation d'un texte en cours de saisie en groupe de souffle. Il s'agit de tâches particulièrement difficiles car elles nécessitent l'extraction d'informations au niveau sémantique. Nous proposons d'utiliser les modèle de language pour capturer ces informations en exploitant un contexte linguistique plus large que la phrase à synthétiser. Plus spécifiquement, nous adaptons un modèle de type BERT pour qu'il prédise directement des caractéristiques acoustiques associées à la focalisation contrastive. Pour évaluer cette approche, nous avons constitué un corpus spécifique présentant de nombreuses occurences de focus contrastifs sur des pronoms personnels. Enfin, nous proposons d'utiliser les modèles autoregressifs (GPT) pour décomposer de facon incrémental un texte en cours de saisie, ce qui permet de réaliser un compromis entre le naturel et la réactivité de la synthèse vocale.

Abstract — Text-to-speech (TTS) technology has the potential to enable real-time communication for applications such as automatic interpreters or assistive technologies for the speech impaired. However, current TTS models are not optimized for such use cases because they require full-sentence inputs, leading to delays between conversation turns. Furthermore, these models are unaware of the surrounding context and are thus unable to adapt their prosody to suit the current situation. These limitations impede engagement and understanding. In this thesis, we aim to improve the suitability of TTS for interactive applications by addressing two main challenges. Firstly, we focus on reducing the time required to initiate speech synthesis while at the same time maintaining natural prosody. Secondly, we explore the prediction of appropriate prosodic features for a given linguistic context. Language models (LMs), known for their effectiveness in natural language processing tasks, are employed as a primary tool for investigation for both of these challenges.

We begin by investigating the importance of degrees of lookahead (i.e., future words) for a vanilla, full-sentence TTS model. We do this by measuring the distance between the final internal representation of a word (i.e., when the full sentence is known) and the intermediate representations at each degree of lookahead. We also compare the prosodic quality of the outputs with a subjective test. Finally, we use random forest analysis to study which factors contribute the most to the stability of the internal representations (i.e., to determine whether the representation is likely to change or not). These tests show that word representations are shaped mostly by the next two words of lookahead and that word length is the largest predictor of stability.

We then investigate the use of pseudo-future text (generated by a language model) to enhance incremental text-to-speech (iTTS) synthesis. By leveraging linguistic clues present in the already provided text, language models anticipate the future context, filling in missing information for prosody modelling purposes. The objective and perceptual evaluations carried out show that this approach offers a good compromise between responsiveness and naturalness of synthesis, but remains highly dependent on the quality of text prediction.

Finally, we address the challenge of producing contextually appropriate speech. We identify an aspect of prosody modelling, contrastive focus on personal pronouns, which can be particularly challenging due to the high-level discursive knowledge which is often required for correct prediction. We evaluate the contribution pretrained LMs can make to this task compared to less linguistically sophisticated baselines. We also compare prediction accuracy with different amounts of context and test the control of prominence in the speech output. We go on to evaluate the use of LMs to guide speech segmentation for high input latency applications. We compare LM-informed methods with simpler count-based methods using subjective tests and a sentence verification test.

Acknowledgments

When I began this thesis on predicting uncertain futures, I was myself uncertain of what the future would bring. Lacking formal training in mathematics and engineering, I was unsure about whether I could contribute meaningful work to the technical field of text-to-speech synthesis. If I was able to produce something of value, it was in large part due to the encouragement and support of a network of people to whom I am eternally grateful.

First and foremost, I would like to thank my dream team of advisors: Thomas, LB and LG. I could not have asked for a better group of mentors. Despite your very busy schedules, you were always available when I needed you. You pushed me to make my work better and your insights into how to shape and clarify my work were invaluable. Not only are you brilliant researchers, but you are some of the kindest people I've ever met.

I would also like to thank the members of the jury. To Damien and Olivier, who provided their perspectives and feedback throughout the entire PhD process as part of my CSI, and to Mireia and Joakim, whose thoughtful questions shone a light on new reflections for my work.

During my PhD, I was fortunate to be a part of not one, but two teams of brilliant PhD students. Whether it was on trips abroad to conferences or simply picnics at Parc Paul Mistral, the time spent in your company made this experience truly enjoyable. I hope the friendships we developed continue long into the future.

My thanks also go out to my parents who sent their love and support from afar.

And last but certainly not least, I would like to thank my partner Dave, without whom I would have lost my mind during the last stage of the thesis writing process. You supported me in so many ways: not only did take on additional household chores so I would have time to work, but you also contributed your immense artistic talents to making some of the most beautiful slides I've ever seen. I am forever thankful.

Contents

List of Figures

List of Tables

Table of abbreviations and acronyms

ASR	*Automatic speech recognition*
BPE	*Byte-pair encoding*
f0	*fundamental frequency*
GAN	*Generative adversarial network*
iTTS	*incremental text-to-speech*
JND	*Just-noticeable difference*
LM	*Language model*
LSTM	*Long short-term memory*
MOS	*Mean opinion score*
MSE	*Mean squared error*
MT	*Machine translation*
MLM	*Masked language modelling*
MUSHRA	*Multiple stimuli with hidden reference and anchor*
NLU	*Natural language understanding*
NLP	*Natural language processing*
POS	*Part of speech*
PP	*Prepositional phrase*
seq2seq	*Sequence-to-sequence*
SVO	*Subject-Verb-Object*
ToBI	*Tones and break indices*
TTS	*Text-to-speech*
VP	*Verb phrase*

Introduction

Text-to-speech for real-time communication Text-to-speech (TTS) is a technology that can give voice to the voiceless, bridge language divides and allow humans to converse with machines. However, current TTS models are not optimized for real-time communication: synthesis is performed at the sentence level which can cause long delays between conversation turns, stilting the interaction. Nor are they good at adapting their output to suit the needs of an evolving linguistic context or the communicative intentions of the user. The incongruence between the speech and the context causes the listener to expend extra mental energy to understand the speaker, further degrading engagement.

In this thesis, our overarching goal is to make TTS more suitable for interactive applications. This includes the sub-goals of (1) reducing the time it takes to start outputting speech while maintaining natural prosody and (2) predicting appropriate prosodic features for a given linguistic context. We investigate these issues primarily through the use of language models (LMs). The recent progress in LM technology has substantially enhanced our capability to model linguistic phenomena, and the application of these models has already proven useful for a wide range of natural language processing (NLP) tasks. We test whether pseudo-future text (generated by a language model) can improve incremental text-to-speech (iTTS) synthesis by filling in missing contextual information that has not yet been provided by the user, and we evaluate the ability of language model-encoded linguistic representations to improve prosody modelling.

Incremental text-to-speech (iTTS) iTTS synthesis is the process of generating speech from an incomplete/evolving text input. The current state-of-the-art paradigm for TTS training (full-sentence TTS) involves sequence-to-sequence (seq2seq) modelling of utterance-level phoneme sequences to their corresponding Mel-spectrograms; inputting an incomplete sequence into a model trained in this fashion will result in a loss in quality due to missing contextual information. Adapting TTS for online processing could take several forms: these could include the use of lookahead (i.e., waiting for some additional words) or training the model to ignore future context (by training on truncated inputs). Both these options require a trade-off with regards to the quality/latency of the system; lookahead reintroduces latency (although perhaps not as extreme as waiting until the end of the sentence) and ignoring right-context means optimizing for a generic future where all context-dependent prosodic phenomena have to be neutralized to fit with any possible future. Another possibility, which has the potential to both reduce latency and keep quality high, is to predict what is coming next using the linguistic clues present in the already-provided text. Since language models (specifically causal ones) are skilled at anticipating what comes next, using this tool could prove useful for iTTS. We evaluate this method in the thesis.

iTTS applications can be divided into two categories: those where the input stream of text content is roughly equivalent to the production rate of natural speech (e.g., simultaneous

translation, dialogue systems) and those where the input stream is considerably slower than natural speech (e.g., Augmentative and Alternative Communication (AAC) applications used to assist the speech impaired). Both types could benefit from reduced latency with predicted futures, in the former case to continuously try to replicate natural speech as closely as possible, but for the latter type, we may want to alternate between speeding up synthesis with a predicted future and tolerating some latency in order to output more natural chunks of speech (as opposed to a one-word-at-a-time output). For this use case, we investigate using the linguistic knowledge contained within language models to guide the segmentation of the speech stream.

Contextually appropriate speech The quality of full-sentence TTS systems has increased dramatically in recent years, and in small doses, their output is basically indistinguishable from human speech. In longer form however (i.e., when synthesizing multiple sentences), the weaknesses of TTS become more apparent: it may sound like a person speaking, but not like a person thinking and embodied in the current context. Humans naturally emphasize the parts of their message that are new or informative for the current discourse. They also add nuances of meaning by shifting emphasis to different words in an utterance. Current TTS systems are incapable of adapting their output to suit the current context because (1) they are trained on single sentences, hence they have no awareness of the surrounding context (the lack of contextual awareness is only exacerbated in iTTS) and (2) they are trained to output the most likely prosodic pattern for a given input (for English, this usually means placing emphasis on the last lexical word of an utterance); their training objective does not support the learning of marked/non-canonical patterns. In order to produce speech that is appropriate for a given context and for the speaker's intentions, a TTS model must have access to higher-level knowledge about meaning and discourse than that inferable from the phoneme sequence of a single sentence alone.

A recent trend in TTS is to incorporate language model embeddings into the TTS model as a way of exploiting additional contextual/linguistic information for prosody modelling. These language models, which are trained on massive corpora containing significantly more words than those typically used to train TTS models, are able to learn representations of various linguistic phenomena. Previous TTS studies incorporating language models have seen improvements in mean opinion scores of speech samples, however the exact contribution of the additional input is not well understood. In this work, we probe whether pretrained language models are able to provide high-level discourse knowledge to a prosody predictor or if it simply adds lower-level information about syntax and distributional semantics. We also examine whether extended contexts can help improve the prediction of prosodic prominence.

The contextual appropriateness of speech (or lack thereof) can have consequences on the cognitive load imposed on the listener. Evidence suggests that listeners do not expend equal amounts of energy on all aspects of the speech signal, but rather use prosodic and other linguistic cues to focus their decoding efforts and structure the discourse. These cues include the relative prominence of words in an utterance and the segmentation of speech into information units. As we would like to synthesize speech that is easy to process, we experiment

with a testing paradigm, sentence verification, that aims to assess the mental effort required to understand an utterance; we attempt to move beyond the most common TTS evaluation methodology which is to gather subjective ratings of single-sentence utterances.

Controlling TTS In previous TTS paradigms (e.g., HMM), prosody modelling was divided into two phases: (1) a front-end NLP analysis which would predict discrete prosodic features (e.g., the phoneme sequence, pitch accent placement and the position of phrase breaks) and (2) an acoustic model that predicts continuous features (e.g., f0, duration). With the introduction of seq2seq TTS architectures, these two feature sets were inferred jointly. This eliminated some error propagation issues and resulted in overall more natural speech, but came at the expense of interpretability and feature control. In this work, we reintroduce a two-stage process in order to regain the control that is necessary for making speech contextually appropriate. We use LMs to predict prosodic feature tags (for prominence and boundaries) and then we use these tags to control the output of a TTS model. This method of control has been shown to be successful in previous works which have tested control of content words (which are frequently prominent in training corpora). We test the limits of the control afforded by this technique by testing more marked prominence structures: prominence on personal pronouns.

Organization and contributions In this thesis, we approach the topics of iTTS and the application of LMs to this paradigm from several angles: evolving neural representations, future word prediction, and prosody modelling. The thesis is organized into five chapters. The first two are dedicated to existing research on communication and linguistic context and its effects on prosody. In the next three, we present our contributions to the field.

1) In the first chapter, we review the literature on characteristics of human communication that should be considered when building a TTS system for interactive purposes. Specifically, we look at (1) the way interaction and backchanneling shapes the quality of a conversation and (2) the foregrounding and backgrounding of elements of the speech signal to facilitate processing. We also present theoretical background on prosody and information structure (i.e., the techniques used by speakers to relate their speech to the common ground shared with their interlocutor(s)). We finish by examining the weaknesses of current TTS models and the methods used to evaluate them.

2) In the second chapter, we look at the ways context affects prosody at the different levels of the linguistic hierarchy. We then review the literature on LM probes which explore the types of linguistic knowledge learnt by these models during training.

3) In the third chapter, we review previous and concurrent work in iTTS before presenting our first contribution, which was one of the first works on neural iTTS. In this work, we study the importance of different degrees of lookahead (i.e., future words) on speech synthesis. To do this, we measure the distance between the final internal representation of a TTS model (i.e., when the full sentence is known) and the intermediate representations at each degree of lookahead. We also evaluate the audio outputs at each stage with a subjective test. We use random forest analysis to study which factors contribute the most to the stability of the

internal representations. The results of this analysis could be used to build an adaptable latency mechanism which only tolerates additional latency when the current word is likely to change significantly.

4) In the fourth chapter, we test the use of language models to predict future context. The aim of this work is to reduce latency by replacing ground-truth lookahead with pseudo-lookahead. We test whether this method is able to improve prosody prediction using both objective and subjective measures. We compare several different synthesis conditions: (1) full-sentence, (2) no lookahead, (3) language model generated lookahead, (4) randomly generated lookahead (a control) and (5) ground-truth lookahead.

5) In the fifth chapter, we use LMs to predict and control prosody. We identify an aspect of prosody modelling, contrastive focus on personal pronouns, which can be particularly challenging due to the high-level discursive knowledge which is often required for correct prediction. We evaluate the contribution pretrained language models can make to this task compared to less linguistically sophisticated baselines. We also compare prediction accuracy with different amounts of context (incremental, full-sentence and extended context). Furthermore, we conduct a perceptive test to gauge the amount of control our controllable TTS system has over infrequently prominent words like personal pronouns. We go on to evaluate the use of language models to guide speech segmentation for high input latency applications. We compare language model informed methods with simpler count-based methods using subjective tests and a sentence verification test.

Research context and publications This thesis was funded by the Multidisciplinary Institute in Artificial Intelligence (MIAI) and began in January 2020. The work was conducted within the CRISSP team at GIPSA-lab (Grenoble) and the GETALP team at LIG (Grenoble). The following works were published as part of this thesis:

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber (2020). "What the future brings: Investigating the impact of lookahead for incremental neural TTS." in: *Proceedings of Interspeech*. Shanghai, China, pp. 215–219

Brooke Stephenson, Thomas Hueber, Laurent Girin, and Laurent Besacier (2021). "Alternate endings: Improving prosody for incremental neural TTS with predicted future text input." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 3865–3869

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber (2022). "BERT, can HE predict contrastive focus? Predicting and controlling prominence in neural TTS using a language model." In: *Proceedings of Interspeech*. Incheon, South Korea, pp. 3383–3387

Audio samples Audio samples used for the research in this work can be found at:

`https://bstephen99.github.io/iTTS/thesisHome.html`

Chapter 1

How humans communicate and the limits of current TTS

Contents

In this chapter, our goal is to review some of the characteristics of human speech which facilitate communication, namely the cobuilding of message through incremental feedback and the foregrounding of important information. We also present some theoretical background in prosody and information structure in order to better understand our objectives in this thesis. And finally, in light of these communication characteristics, we examine the ways in which synthetic speech, despite its recent advancements, is not currently well suited to interactive applications.

1.1 Speech is a duet, not a series of solos (Clark 1996)

Early psychological models of the production and comprehension of speech in conversation (Miller 1951) influenced by information theory (Shannon and Weaver 1949) presented these processes as a series of consecutive of actions: a speaker would first encode a message into its

linguistic form and transmit it to the listener. The listener would then decode the linguistic signal to understand the message and then take the turn of the speaker and formulate a new message to continue the discussion. More recent developments in this field have come to see the act of conversation as a more collaborative activity, as evidenced by phenomenon such as split utterances where the listener is able to coherently complete the speaker's sentence (Purver et al. 2009). A listener is able to accomplish this because they are not a passive vessel but rather an active agent who is incrementally predicting what the speaker is going to say next (Pickering and Garrod 2007; Pickering and Garrod 2013). The listener is also active in that they are continuously communicating their level of understanding of what the speaker has already said.

The listener's ability to backchannel (Yngve 1970) has a direct effect on their understanding. Schober and Clark 1989 compared the speech understanding of addressees and overhearers. In their experiment, a speaker and addressee were in direct communication while trying to complete a card ordering game; an overhearer was given a recording of the exchange between the speaker and addressee and they were asked to complete the same task. The overhearer had access to the same linguistic input as the addressees, but their inability to provide feedback to the speaker and to negotiate meaning resulted in worse performance on the task.

In addition to communicating their own understanding, the listener is able to help shape the production of the speaker through backchanneling. Bavelas et al. 2000 studied productions by speakers narrating a story under conditions where the listener was either attentive or distracted. In the attentive condition, speakers received normal backchanneling from the listener. In the distracted condition, less feedback was provided and the quality of story suffered as a result. The type of backchanneling also has an influence on the course the speaker will take. Tolins and Tree 2014 compared generic (e.g., *uh huh*, *oh*) and specific (i.e., context sensitive commentary, e.g., *wow*) listener responses and found that generic backchanneling encouraged the speaker to provide discourse new information whereas context specific backchanneling would cause the speaker to elaborate on given events. Furthermore, as mentioned above, listeners are constantly making predictions about upcoming speech and they sometimes vocalize these predictions as a form of backchannel to communicate they have been paying attention. Even if an offered prediction was not precisely the word the speaker had in mind, the speaker will often adopt the listener's words. This has the effect of *grounding* conversation (Clark and Marshall 1981), i.e., establishing a common ground between the speaker and listener's knowledge.

It is the speaker's job to monitor the addressee's understanding and to adapt their speech if there has been a breakdown in communication. This alteration can take place part way through an ongoing utterance if the speaker suspects the addressee has not identified one of their referents (Clark and Krych 2004). Speakers can also elicit feedback from the listener through the use of prosodic cues (Buschmeier and Kopp 2014).

Other studies have demonstrated that the quality of speech is altered if listener feedback is delayed. Krauss et al. 1977 had participants communicate over an audio channel with a one second delay. The communication becomes a lot less efficient/more redundant (i.e., speakers used considerably more words) when compared to communication without a delay.

Similarly, Vartabedian 1966 found participants accomplished a shared task 28% slower while teleconferencing with delay versus no delay.

Active communication is most effective when there is a timely back and forth between participants. This allows a speaker to react to and a listener to express their comprehension needs in real time. Going beyond comprehension, reactivity in conversation can help build social connections as the participants work together to build meaning. Reactivity is therefore an important element in a communicative TTS system.

1.2 Focus on what is important, minimize the rest

As noted by Winkler 2005, the auditory processing of speech has many parallels with the visual processing of a visual scene: we cannot pay attention to everything at once, we focus on some elements while relegating others to the background. If we cannot make a clear distinction between what should be focused and what should be ignored, like in Figure 1.1, then we have trouble deciphering the message. If we take the time to examine the picture, we can see a dog in the woods, but the interpretation is not automatic the way it would be if the background and focus were clearly distinguished. The same is true in audio processing, we have to work overtime if what is uninformative is not reduced. This is something that humans do naturally, but TTS systems struggle with.

Figure 1.1: Heterogeneous background. Auditory comprehension is similar to the visual system in that we must make a distinction between the background and foreground. When this distinction is not made, more mental effort must be exerted to comprehend. Image from Winkler 2005, originally from Goldstein 1996.

1.2.1 Speech production: emphasize the unpredictable

Language is a complex and noisy system that humans use to communicate information. A key strategy humans employ to optimize the transfer of their message is to take into account the

predictability of the words they use. Lieberman 1963 studied the relationship between word redundancy (i.e., how predictable a word is in its context) and its acoustic realization. Words present in idiomatic or cliché expressions ("A stitch in time saves nine") were compared with matched words in less predictable sentences ("The number that you will hear is nine"). The words were recorded in context, excised from the sentence and then study participants were asked to identify the words presented in isolation. The words excised from the less predictable sentences were more easily recognized.

Several subsequent works have found similar effects of predictability on acoustic reduction (Jescheniak and Levelt 1994; Jurafsky et al. 2008; Arnon and Snider 2010). To explain this phenomenon, Aylett and Turk 2004 propose the **Smooth signal redundancy hypothesis** which stipulates an inverse relationship between language redundancy and acoustic redundancy (duration). In other words, speakers make an effort to evenly distribute information throughout the speech signal to maximize the chance their interlocutor will understand what they are trying to say: frequent, predictable words are accorded less time than infrequent, less predictable words. This allows for reduced articulation effort as well as robustness to information loss over a potentially noisy channel of communication.

1.2.2 Speech perception: Good enough processing

Processing speech is a complicated task that requires the rapid decoding of sounds to interpret meaning at several structural levels (phonological, syntactic, semantic and discursive). This involves recognizing the individual sounds that make up the words, identifying the words themselves, understanding the grammar and structure of the sentence, and interpreting the meaning of the sentence in context. Unlike when reading, where it is possible to return to an earlier section, speech/conversation keeps moving forward. In order to handle the stream of information, humans have to be shown where to focus their processing efforts; for the rest, shallow representations will usually suffice.

Ferreira and Lowder 2016 call this differentiated system of attention *Good enough processing*. This theory is built on several experiments that show humans often use partial/underspecified linguistic representations, which can sometimes lead to misunderstandings, but in most circumstances allow them to understand the gist of what their interlocutor is saying. For example, when asked *How many animals of each kind did Moses take on the Ark?*, most people respond "two" even though it was actually Noah who brought animals onto an ark (Erickson and Mattson 1981).[1] Another example (Ferreira et al. 2001) comes from a garden-path study looking at sentences like *While Anna bathed the baby played in the crib.* When asked comprehension questions, subjects scored poorly on questions that dealt with the initial misreading of the sentence (e.g., *Did Anna bathe the baby?*), indicating they had not properly updated their representation of the sentence when disambiguating syntactic cues became available.

A possible explanation for the shallow treatment in these experiments is the form in which

[1] For this illusion to work, the incorrect word must share semantic traits with the correct word (the effect goes away if you substitute *Moses* with *Nixon*).

the poorly processed sections appear. In the Moses illusion, *Moses* is not the focused element of the question and *bathe* is part of an initial subordinate clause which is often used to present given information (Ferreira and Lowder 2016). Support for this idea comes from an experiment by Sanford et al. 2006 designed to test focus and depth of processing. In this experiment, participants were played an audio recording twice; on the second listening, the audio was either identical to the first, or one word was changed. The target word was either narrowly focused (as in (1)a) or part of a broad focus (as in (1)b). Results showed participants were more likely to notice the change in the narrow focus condition, where it was more important to the discourse.

(1) a. Narrow focus:
They wanted to know which money had been stolen.
The money from the WALLET/PURSE had gone missing.
Thefts in the area were becoming all too common.
b. Broad focus:
They wanted to find out what had happened.
The money from the wallet/purse had gone missing.
Thefts in the area were becoming all too common.

Other research shows a prioritized treatment of prosodically prominent words. Stressed words are processed faster than unstressed words. This has been tested in phoneme monitoring experiments where subjects must push a button when they detect a target phoneme: Shields et al. 1974 found reaction times were faster for stressed words. There is also evidence that the preceding context allows listeners to anticipate upcoming focus. Cutler and Foss 1977 spliced the same recorded word into two intonational contexts; one leading to a stressed word and one to an unstressed one. Despite the identical acoustic content at the target, reaction times were faster for the stressed condition. Nooteboom 1987 posed the question why all words are not accented if it aids processing. Their experiments revealed that there is an interaction between the new/giveness status of a word and its accentuation. Subjects were faster to identify a given referent when it was unaccented.

1.3 Prosody

Words in spoken language are made more or less prominent than others by manipulating prosodic features. Prosody refers to the rhythmic, intonational and phrasing patterns used in speech to convey both linguistic and paralinguistic meaning. These meanings are expressed through pitch, energy, duration and voice quality variations. A widely adopted annotation system to describe the features of prosody is ToBI (Tones and Break Indices) (Silverman et al. 1992 based on the work of Pierrehumbert 1980). This is a symbolic system that envisions the prosodic contour as a series of targets which include pitch accents, phrase accents and boundary tones.

Pitch accents are used to highlight important words and they are (usually) aligned with stressed syllables which are lexically determined in English (although they can shift to focus an informative morpheme, e.g., *It is STRESSED, not **UN**stressed*). ToBI defines pitch accents by their targets in the pitch range: high (H^*) and low (L^*). There are also more complex pitch accents that combine multiple targets, e.g., L^*+H, L+H^* (* shows the target that aligns with the stressed syllable). These accents have been associated with different discursive functions, for example H^* accents have been associated with discourse new words and L+H^* with contrastively focused words.

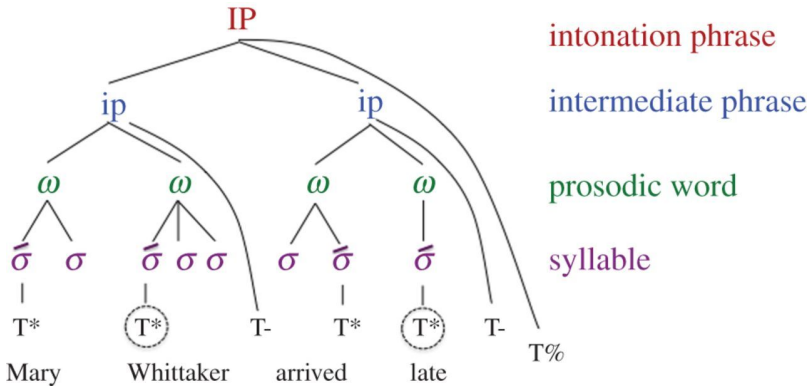


Figure 1.2: The prosodic hierarchy (Image from Krivokapić 2014). T represents tonal targets. Ts followed by * designate pitch accents, the circled ones show nuclear accents. T- indicates phrase accents and T% boundary tones.

Phrase accents and boundary tones are used to elucidate the hierarchical organization of utterances. The domains in this hierarchy include the syllable, the prosodic word, the intermediate phrase and the intonational phrase (See Figure 1.2). Phrase accents are used to segment intermediate phrases and boundary tones segment intonational phrases. These are marked using pitch targets similar to pitch accents (H and L). However, rather than associating with stressed syllables, phrase accents define the tonal movement from the last stressed syllable until the end of the phrase, and boundary tones are limited to the edge of intonational phrases. Both these boundary markers are associated with domain-final lengthening and they can signal the type of discourse unit being used (e.g., a declarative or interrogative sentence). Furthermore, each intermediate phrase must contain at least one pitch accent, the most prominent of which is referred to as the *nuclear accent*.

ToBI is a symbolic abstraction of a noisy acoustic signal that is subject to speaker and dialect variations and that is simultaneously trying to represent several streams of meaning (e.g., speaker intention and affect, information structure, etc.). As such, the clean categories defined by the system are not always easy to identify. Furthermore, the assignment of pitch accents to specific discourse functions has not always been supported by empirical evidence (Katz and Selkirk 2011; Chodroff and Cole 2019). Chodroff and Cole 2019, for example, found that the relationship between given, new and contrastive status and their associated pitch accents is only probabilistic; a contrastive accent, for example, is more likely to be marked with a L+H^* accent, but other accents can do the same job.

There is also evidence that speakers and listeners are very sensitive to longer-range context in their judgements of prominence. So while the L+H* accent is typically associated with a higher degree of prominence than the L* accent, these roles can be reversed depending on how frequently they are used. Kakouros et al. 2018 tested perceptions of prominence by having subjects listen to five minutes of audio with a biased distribution of either rising or falling intonation on the sentence final word: when asked to evaluate subsequent sentences, the underrepresented/less likely prosodic trajectory was rated as more prominent.

Nonetheless, the underlying truth that speech highlights important words and organizes itself hierarchically is general accepted and ToBI is commonly used by the research community to discuss such phenomena.

1.4 Information structure

The themes of predictability, prominence and givenness seen in the previous sections are all related to the subject of **Information structure** (Halliday 1967). Information structure reflects the presumed common ground between speakers (Stalnaker 2002). A speaker will make syntactic and prosodic choices to package information in such a way that it can be easily understood by their interlocutor (Chafe 1976). This includes structuring sentences into **topic** and **comment** to aid coherence, drawing attention to certain words to highlight their informativeness in the discourse (**focus** or **contrastive focus**) and de-emphasizing elements that are assumed to be understood (givenness).

Coherent discourse requires consideration be paid to how a sentence fits into the larger context. For example, the two sentences in (2) contain the same information content/semantic truth values, but the focus assignment in each only makes them appropriate in certain settings. If the preceding question is *Who did Mary give the letter to?* then **a** would be felicitous but **b** would not, and the opposite would be true if the questions were *What did Mary give to Kim?*.

(2) a. Mary gave KIM the letter.
b. Mary gave Kim the LETTER.

Sentences are organized into a topic and comment (or theme and rheme) structure. Topic refers to what the sentence is about and comment updates knowledge about the topic. Topic is strongly associated with the grammatical subject in English, but these two roles do not have to be taken by the same entity, as demonstrated by sentences with topic marking phrases (e.g., ***As for the dog***, *Bill left it at home.*).

There are different schools of thought on the notion of focus. Some define it in terms of **newness/informativeness** (Halliday 1967) or what is asserted (Lambrecht 1994) and others in terms of the evocation of **alternatives** (Rooth 1992; Krifka 2008). Halliday 1967 describes the focus of the utterance as the part which is "not recoverable from the preceding discourse". Krifka 2008 defines it as that which "indicates the presence of alternatives that are relevant

for the interpretation of linguistic expression."

Newness accounts divide focus into **presentational focus**, which satisfies information requirements and is less marked prosodically and **contrastive focus**, which makes salient alternatives in the discourse (converging with the alternatives account). For the alternatives account, these two types of focus are simply a difference in the saliency of pragmatically appropriate alternatives; elements in presentational focus have many possible alternatives (an open set) whereas the alternatives are more restricted in the contrastive case (a closed set) (Kiss 1998).

Topics can also take on a contrastive meaning, in which case they are referred to as contrastive topics. Elements that are neither focused or contrastively focused are marked as given.

1.4.1 Common ground

The common ground refers to a shared space of knowledge where there is a set of propositions and referents that are known by both conversation participants (Stalnaker 2002). Krifka 2008 also includes communicative interests and goals in this space. Elements in the common ground can be built from a shared physical environment, mutual past experience, world knowledge or from previous linguistic context. Previously unmentioned elements can enter the common ground without explicitly being evoked if they are inferable from existing referents (e.g., we can accommodate the word *seat belt* if we have been discussing a car).

The common ground is constantly being updated. When a speaker wants to add something new to the discourse, they must be conscious of what is already in the common ground and structure their expression so the new contribution can be connected to what is already there. And referents already in the shared space, must be marked as such, so the listener knows where to look for them. The status of referents are differentiated with different types of referring expressions (e.g., nouns vs. pronouns, definite vs. indefinite articles, etc.) and also, importantly for our purposes, prosodically.

1.4.2 Devices for information structure expression

Languages differ in the way they express information structure. English, French and Italian are all Subject-Verb-Object (SVO) languages whose unmarked prosodic accent pattern is to place the sentence accent at the end of the sentence. However, when the information structural properties require a marked structure each of these languages uses a different strategy. To illustrate these different strategies, we will take a look at an example from Lambrecht 1994 ((3)). The imagined scenario is one in which a woman is struggling to get her many shopping bags onto a crowded bus. She looks apologetically at the other passengers and says:

(3) a. My CAR broke down.

b. Mi si è rotta la MACCHINA.
to-me itself is broken the car.
c. J'ai ma VOITURE qui est en PANNE.
I have my car that is in breakdown

This is an example of what Lambrecht calls an "event-reporting" sentence which introduces a new discourse referent and is distinguished from more common "topic-comment" sentences with a marked focus structure. English, which has very strict word-order rules, uses prosodic means to indicate the focus, moving it from the sentence final position. Italian, which has looser word-order rules, inverts the subject (*la macchina*) and the verb phrase (*è rotta* in order to the maintain the prosodic prominence/the indicator of focus in its natural position. French, which has a strong aversion to placing focus on the subject of a sentence, will often resort to syntactic devices to express information structure. In this case, the simple proposition *Ma voiture est en panne* is broken into two *J'ai ma voiture* and *qui est en panne.* The first proposition (*I have my car* does not add anything to the discourse; it serves merely to move the subject to the proposition final position where it can be focused).

English also has syntactic means to mark information structure (e.g., cleft sentences - *It is my car that is broken.*), however it is most commonly indicated prosodically. Because information structure is not explicitly written in the text, learning context appropriateness is particularly challenging for English TTS.

1.4.3 Interaction with prosody

According to Halliday 2015, there are three ways to manipulate the interpretation of an utterance in terms of prosody and information structure. The first is **tonality** which concerns prosodic phrasing (i.e. the hierarchical grouping of words, which are signaled by weak or strong boundaries between words). Leonarduzzi and Herment 2013 contend that non-canonical phrasing (which typically involves breaking an intonational phrase into subphrases) is a strategy to highlight the informativeness of elements in each of the subphrases. The second is **tonicity** which describes the placement of the nuclear accent. The nuclear accent is usually placed on the last lexical word in an intonational phrase. If it is placed elsewhere, this signals a marked prosody that can point to a narrowly focused word. The third is **tone** which describes the type of pitch accent.

1.4.3.1 Focus projection

The focus of a proposition and prosodic markings of focus are not the same thing. The examples in (4) (simplified from Selkirk 1995) illustrate this fact: The word **BATS** receives the nuclear accent (represented by capital letters) in all five utterances, but the focus in the information structural sense is dependent on the discourse (i.e., the corresponding question

below each sentence), ranging from narrow focus on a single constituent in (a) to the full utterance in (e).

(4) a. Mary bought a book about [BATS]$_{\text{FOC}}$.
(What did Mary buy a book about?)
b. Mary bought a book [about BATS]$_{\text{FOC}}$.
(What kind of book did Mary buy?)
c. Mary bought [a book about BATS]$_{\text{FOC}}$.
(What did Mary buy?)
d. Mary [bought a book about BATS]$_{\text{FOC}}$.
(What did Mary do?)
e. [Mary bought a book about BATS]$_{\text{FOC}}$.
(What's been happening?)

While the distribution of prominent words in an utterance does not directly point to the underlying focus structure, it does constrain the interpretation. Selkirk 1984 explains the phenomenon of focus projection through interactions with syntax. A pitch accented word can licence the spread of focus to its head or its internal arguments. On the contrary, an accented word cannot licence focus onto an adjoining adjunct. For example, in the sentence *He only smoked in the tent.*, both *smoked* and *tent* must be accented if they are both in the focus domain (Gussenhoven 1983, cited in Selkirk 1995). If only *tent* is accented, this does no spread to the VP.

1.4.3.2 Degree of prominence

Focused words will be longer and louder than non-focused words, however the degree of prominence will be influenced by the *type* of focus. In the course-grained focus/contrastive focus distinction, contrastive focus is more prominent. More fine-grained focus categories have also been proposed which differ in terms of the number and the saliency of the members in its alternative set. The focus types listed below are ranked from least to most prominent (examples from Féry 2013):

(5) a. **Broad information focus** What is happening? Tom is going to VIENNA.
b. **Informational narrow focus** Who is going to Vienna? TOM is going to Vienna.
c. **Exhaustive/identificational interpretation of a narrow focus** Which of your sons is going to Vienna? It is TOM who is going to Vienna.
d. **Association-with-focus (particles)** Are both Alain and Tom going to Vienna? Only TOM is going to Vienna.
e. **Contrastive focus: parallelism, right-node-raising, selection** Where are your sons going to? TOM is going to VIENNA, and ALAIN to BERLIN.
f. **Contrastive focus: correction** Is Alain going to Vienna? No, TOM is going to

Vienna/No, it is TOM who is going to Vienna.

Zimmermann 2008 proposes an alternative explanation for relative prominence, not in terms of focus type, but rather with respect to the speaker's beliefs about the listener's expectations. If a speaker thinks their contribution will be unexpected by the listener (i.e., unlikely to enter the common ground), they will add extra emphasis. And the more surprising a contribution, the more emphasis it will receive.

Calhoun 2009 theorizes that the perceived level of prominence of a word is related to how salient that word is relative to how salient it is expected to be. So words that fall naturally in a prominent position (e.g., sentence final/default nuclear accent position, where we expect sentence stress to be) will not necessarily be perceived as particularly prominent unless they are overly exaggerated. Conversely, a function word, that is usually expected to be reduced, only has to be somewhat more salient than expected (i.e., not excessively emphasized) to attract a contrastive focus interpretation.

1.5 Where text-to-speech is lacking

In this section, we will look at some of the shortcomings of both current TTS and the methods used to evaluate it. Using traditional evaluation metrics, state-of-the-art TTS systems are almost on par with human speech. However, these high scores are a reflection of the limited evaluation techniques more than actual human communication parity. The standard practices in TTS evaluation include measures of intelligibility (i.e., how well a listener can identify the segments in the speech signal) and naturalness (i.e., how human-like is the speech). Naturalness is often evaluated using a Mean Opinion Score (MOS) test. Here, single isolated sentences are evaluated on a five point scale. Many criticisms have been levied at this way of testing, due to its failure to capture important elements of a system's success. These elements include contextual appropriateness and variability. When these elements are lacking, increased listening effort is imposed on the listener. Developing evaluations that measure this effort could help move TTS forward.

1.5.1 Contextual appropriateness

Contextual appropriateness in TTS can refer both to how adapted the system is to its use case (the contextual framework) and to how adapted the prosody is to the current linguistic context. We will only briefly touch on these topics here, and develop them more fully in the next two chapters.

Contextual framework. Evaluating speech in a vacuum does not necessarily reflect how well the system will be perceived when put to use. Wagner et al. 2019 illustrate this point by comparing potential style clashes, like a dramatic poetry recitation versus a telephone-based

inquiry system; the speech can sound good but be completely inappropriate. The opposite is also true: Baumann and Schlangen 2013 found users of an interactive application preferred adaptable speech with lower acoustic quality to higher quality but more rigid speech.

Linguistic context TTS models are typically trained using single, isolated sentences. With such limited context, it is not possible for a model to adapt itself to the current discursive or environmental context when it is applied to an interactive application. What's more, the single-sentence paradigm is not suited to incremental synthesis. If we simply adopt a full-sentence model to produce speech one word at a time, there is serious degradation in the prosodic quality. Furthermore, the contextual appropriateness difficulties observed in (single) full sentence TTS is exacerbated in an incremental setting, where not only is the previous context unknown, but so is the future context.

1.5.2 Variability

Monotony in speech, be it human or synthetic, can be painful to listen to. While a certain amount of regularity helps make speech decipherable, a complete absence of variation reduces a listener's ability to prioritize what is important in the message. Monotony in TTS stems from its training regime which is at odds the way humans employ expectation for communication. Humans will often subvert expectation (i.e. use a less probable pattern to mark their intended meaning (Calhoun 2007; Kakouros and Räsänen 2016; Kakouros et al. 2018)) but neural networks are only trained to replicate the most likely intonational pattern. Training with limited context aggravates this problem because longer-range patterns in variation are not accessible to the model.

In addition to repeated intonational patterns, synthetic speech often suffers from flattened expression (*average prosody*). Again, due to the training objectives (to reduce the corpus-wide mean squared error which encourages conservative/close-to-the-mean predictions), the predicted intonational contours tend to be flatter than those found in natural speech. This too can affect comprehension as fundamental frequency has been shown to aid understanding in adverse conditions. Laures and Bunton 2003 compared the intelligibility of both natural speech and speech with a flattened f0 contour in noise and found the number of transcription errors was significantly higher in the flattened f0 condition.

Recent efforts to introduce more variability involve the use of auto-encoders to learn a latent prosodic space and then the use of either random sampling or linguistic features to condition speech generation (Kenter et al. 2019; Tyagi et al. 2020; Hodari et al. 2021)

1.5.3 Listening effort

There have been some experiments that go beyond the standard MOS and try to find evidence of cognitive effects from TTS. These studies have looked for differences in comprehension,

processing speed and memory when compared to natural speech.

Direct effects on comprehension have not been easy to find (Wester et al. 2016; Pisoni and Hunnicutt 1980; Boogaart and Silverman 1992). While TTS is often lacking the prosodic features that facilitate the processing of discourse and information structure, this information can usually be recovered from the linguistic content alone (i.e., the sequence of words). It does however place an additional burden on the listener who must actively work to reconstruct a map of the structure of discourse and meaning, whereas in natural speech, the speaker provides clear signposts, easing the cognitive effort.

More online measures have been used to try and quantify the increased listening effort. These include physiological response tests (e.g., pupillometry (Simantiraki et al. 2018; Govender and King 2018b)) and tasks that measure reaction times in decision tasks, such as phoneme monitoring or sentence verification where subjects decide whether a sentence is true or false (Nix et al. 1993, Pisoni et al. 1987). These tests demonstrate a cognitive penalty for synthetic speech.

TTS can affect memory because it becomes more difficult to encode information when more effort is being exerted to decipher it. Paris et al. 2000 found subjects recalled fewer words when asked to reconstruct a sentence that had been presented with a synthetic voice than a natural one. Wolters et al. 2014 studied the recall of medications after a message from either a human or a synthetic voice. Recall was the same for items the subjects were already familiar with, but when asked to recall new information, human-presented messages were remembered better. Interestingly, the inclusion of some human like features (e.g., breath sounds) can help improve recall for synthetic speech (Whalen et al. 1995; Elmers et al. 2021).

One investigation into the cognitive load imposed by TTS uncovered some surprising results. Govender and King 2018a conducted a dual task experiment where subjects had to shadow (i.e., repeat the words) sentences read by either a human or a synthetic voice and then answer other questions in parallel. Unexpectedly, subjects performed better on the high quality synthetic speech than on the human speech. Most people would agree that human speech is easier/more enjoyable to listen to, and so we suggest that this result perhaps reflects the selected task more than the listening effort imposed. To repeat a sentence (in this case, semantically nonesense sentences), you must correctly identify the phonemes being spoken. This is likely easier when each word is clearly enunciated. Since TTS is not great at predicting contextually appropriate prosody, it tends to hedge its bets and accent all content words, which perhaps facilitated this task. To understand a discourse in real time though, you must be able to ignore the parts of the message that are unimportant. As we have seen previously, this is done by reducing articulation/duration on the redundant parts of the message.

There are of course application of TTS where evenly articulated speech is appropriate, for example the announcement of schedule changes in a noisy train station. However for longer-form communication, emphasizing the important and de-emphasizing the unimportant should help with processing. The difficulty for TTS is deciphering between the two.

1.6 Conclusion

In this section, we have seen two fundamental traits of interactive communication: (1) the incremental negotiation of meaning and (2) the highlighting of important information on the part of the speaker and the differential processing based on these cues by the listener. The use of current TTS models in interactive applications fall short on these two fronts. TTS models trained in a single-sentence paradigm require full-sentence inputs and this will cause response delays degrading the quality of the interaction. Moreover, they do not posses sufficient information to properly predict the elements in the text input that should be foregrounded and those that should be backgrounded. In this thesis, we try to improve these aspects by adapting TTS to an incremental setting and by predicting prominence features. We further try to reduce listening effort by predicting appropriate prosodic boundaries.

Chapter 2

Why context matters

Contents

In this chapter, we will investigate the importance of context for prosodic expression and processing. In the first section, we provide an overview of the ways context influences the speech signal at all levels of the linguistic hierarchy. In the second section, we will look at a potential tool for adding contextual knowledge to a TTS system, Transformer language models (LMs).

2.1 How context shapes the speech signal

A single string of phonemes can be said in a number of different ways. Even controlling for context, there is no such thing as a unique "correct" prosody. This one-to-many nature of speech makes TTS a particularly challenging problem. That said, speech is not a completely random process and there definitely is such a thing as wrong prosody that violates a native speaker's grammatical and pragmatic expectations and can contribute to misunderstanding. Listener judgements of naturalness and appropriateness will be influenced by the level of adherence to these expectations and TTS systems should strive to meet them.

The contextual factors reviewed in this section will not be exhaustive, there are of course a large number of different ways context can impact speech. But we do hope to provide an overview of the linguistic research into speech regularities, so that we can better understand what exactly we are trying to predict when we enrich a TTS/iTTS system with additional LM-produced information.

2.1.1 Phonological effects

The realization of phonemes is dependent on their context in a myriad of ways. Here we will focus on co-articulation phenomena (i.e., modifications that are made to facilitate the articulation of two or more speech sounds when they are produced sequentially). These changes can occur within words and across word boundaries. They are a common occurrence in connected speech and can become exaggerated as speech rate increases. The co-articulation processes can be summarized into three main types: (1) sound changes (**assimilation** and **allophonic variation**), (2) sound deletion (**elision**) and (3) sound addition (**intrusion**) (descriptions below based on Setter 2015).

Assimilation is a process where one phoneme takes on the characteristics of its neighbour(s). Both past and future phonemes can assert an influence on the current phoneme; these are referred to as progressive and regressive assimilation respectively. The voicing, the manner and/or the point of articulation of a segment can all be affected by assimilation; in the regressive case, the speaker anticipates upcoming articulatory features and as they prepare to say the next phoneme, the current phoneme adopts these features to smooth the transition. The reverse is true in the progressive case.

A similar process to assimilation is allophonic variation, although here the phonemes do not change into other phonemes, but rather into another version of the same phoneme (i.e., an allophone). For example, aspirated /p/ will become unaspirated and sound more like a /b/ when preceded by an /s/ in English (*speech* →*sbeech*).[1].

Elision is a process where phonemes are deleted from a word. Elision is particularly prevalent in sequences of words that are frequently said together (e.g. must be →mus be). It can also occur within words, as in the reduction of the four syllable word *interesting* (in.te.res.ting) to three syllables (int.res.ting).

Sometimes phonemes that do not exist in a word when it is pronounced in isolation appear out of nowhere when the word is in specific contexts. In non-rhotic varieties of English, where /r/ only exists in prevocalic positions, an /r/ will insert itself between two words if the second word begins with a vowel (e.g., *law*[r] *and order*). This is known as *intrusive R* (or *linking R* if the r exists in the spelling, but is usually not pronounced). It is used to avoid adjacent vowels (Broadbent 1991).

[1]This is only true if the /s/ is in the same word as the unvoiced consonant (and not in an easily separable morpheme, an /s/ in the preceding word (e.g., Miss [p^h]iggy).

2.1.2 Metrical effects

Language is characterized by rhythmic patterns; there is a natural tendency to see alternating strong and weak syllables. This is known as the *Principle of Rhythmic Alternation* (Selkirk 1984). These beats are important for the parsing of the speech stream. In English, stress is an encoded feature of lexical items, however, when two strong syllables are adjacent in a prosodic phrase, speakers make adjustments to remove the stress clash. They do this by either removing or shifting one of the stresses (Hayes 1984; Selkirk 1984). For example, the lexical stress for the word *fourteen*, pronounced on the second syllable when spoken in isolation, is moved to the first syllable when it is followed by a word that has stress on its first syllable (e.g., *women*). See more examples of stress shift adapted from Hayes 1984 in (1) (primary stress is marked with capital letters and an acute accent; secondary stress with a grave accent):

(1) a. fourTÉEN → fòurteen WÓmen
b. MissisSÍPpi → Mìssissipi LÉGislature
c. seventy-SÉVen→ sèventy-seven SÉALS

Quené and Port 2002 attribute the occurrence of stress shift to more global sentence considerations with their *Equal spacing constraint* which states: "Prominent vowel onsets are attracted to periodically spaced temporal locations." The implication of this being that both the left and right context play a role in determining stress shift, as in (2) where the stress on *ideal* moves from the first syllable in (a) to the second in (b) due to the contrastively focused *their* that immediately precedes it.

(2) a. [John and Ben have been searching for an acceptable partner, but] they will NÉVer FIND their ÍDeal partners.
b. [Other people have found their perfect partners, but] John and Ben will NÉVer find THÉIR idEAL partners.

2.1.3 Prosodic domain effects

Where a phoneme falls within the prosodic hierarchy will affect its expression. Domain-final lengthening is one of the essential clues for decoding the prosodic structure of an utterance. Wightman et al. 1992 measured normalized foot durations at pre-boundary positions and observed a linear relationship with annotator perceived boundary strength. In other words, a foot preceding an intonational phrase boundary is longer than one preceding an intermediate phrase boundary, which is in turn longer than a prosodic word boundary.

Domain boundaries also have an effect on the articulatory strength of the phonemes that border these divisions. Fougeron and Keating 1997 detected differences in linguopalatal contact for consonants and vowels in domain initial, medial and final positions. Consonants exhibit more linguopalatal contact in initial position and vowels exhibit less in domain-final

position. This has the effect of making the phonemes in these positions more pronounced (i.e., distinguishable from other phonemes). The articulation of heads of prosodic domains (e.g., the nuclear accent in an intermediate phrase) are also differentiated from non-heads (Beckman and Edwards 1994). The duration will be longer and it will have a larger and faster opening movement.

Another durational effect has been observed for syllables and feet known as anticipatory compensation or compensatory shortening (Klatt 1973; Fowler 1981; Munhall et al. 1992). The more segmental content within these units, the more compressed they become. For example, if a syllable consists of a single vowel, it will be longer than if the syllable contains a coda with one consonant, and shorter still if the coda has two consonants. The same applies to feet, where the stressed syllable will be shorter depending on the number of unstressed syllables that follow it (e.g., stick > sticky > stickiness).

2.1.4 Lexical and frequency effects

Knowing the lexical identity of a word will influence its pronunciation. For example, homographs like *bass* will contain different phonological content depending on whether one is discussing the fish or the musical instrument. Most homographs in English can be distinguished by their POS category (e.g., *The bandage was wound around the wound* (Noun vs. Verb)) and thus can be easily distinguished by their syntactic environment. Generally speaking, if the word in a homographic pair is a noun, the lexical stress will fall on the first syllable and if it is a verb, it will fall on the second.

How frequently a word is used will have an impact on how expected it is and therefore on how quickly it is pronounced. Baker and Bradlow 2009 studied the effects of both general word frequency and contextual predictability (using second mentions) and found both factors led to shorter durations even when controlling for speech style (clear/hyper-articulate vs. plain).

Whether or not a word belongs to a larger lexicalized expression will also impact its pronunciation and stress patterns. When the elements of a compound structure are used frequently enough, they start to fuse together. Morrill 2012 found intensity, duration and pitch differences could differentiate between lexicalized compounds and equivalent phrases (e.g., *greenhouse*/*green house*). The stress pattern a compound adopts (i.e., relative prominence on the first or second word) is largely dependent on the identity of its constituents parts (Plag 2010). For example, compounds whose second word is *street* (*Máin Street, Óxford Street*) will have primary stress on the first word and compounds with *avenue* (*Fifth Ávenue, Madison Ávenue*) will have it on the second. The same type of rigidity can be seen in larger idiomatic expressions as well (Ashby 2006). See (3).

(3) a. She has eyes in the back of her HEAD.
    b. She has eyes in the BACK of her head.

The standard idiomatic reading of the sentence is (a). If someone says (b), it comes across as

odd, even though non-idiomatic sentences with similar structures can readily shift the nuclear accent to *back* (e.g., *She has a shovel in the BACK of the house.*) This gives a contrastive reading (*back* vs. *front*), but the same should be true for (3), since the implicit contrast is also with *front*. The lack of flexibility is evidence that we may encode prosodic features as part of the mental lexicon.

2.1.5 Syntactic effects

Syntactic structure, while not the same as prosodic structure, has been shown to influence the prosodic form and judgments of prosodic appropriateness (e.g., Pynte 1998). Syntactic complexity also impacts prosody, specifically pause durations between constituents, due to the difference in time it takes to plan the production of simple versus complex structures (Ferreira 1991).

We know that syntactic and prosodic structures are not the same, because multiple prosodic phrasings are judged as acceptable for the same syntactic structure (See (6)). The ends of large syntactic constituents (clauses) usually align with large prosodic boundaries, but smaller constituents are less predictable.

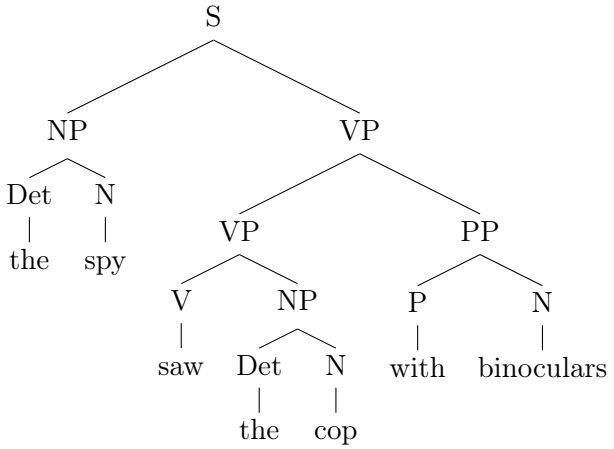
There are some syntactic constructions that require separate phrasing (in written text these are usually set off by commas). These include parentheticals, appositions, and non-restrictive relative clauses. Price et al. 1991 found that parentheticals and appositions could be reliably differentiated prosodically from non-parenthetical/non-appositions containing identical phoneme sequences (e.g., *Mary knows many languages*(,) *you know*) thanks to boundary strength. Other potentially ambiguous constructions are not always made clear in a written text. We will look at some of these in the following subsections.

2.1.5.1 Syntactic attachment

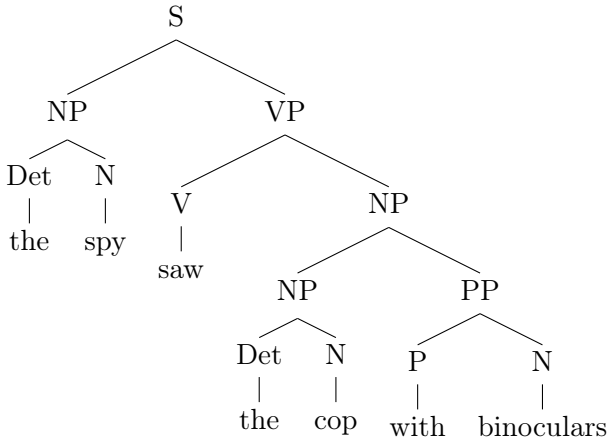
There are several types of syntactic attachment ambiguities whose underlying structure can affect prosody. These include prepositional phrase ((4)), relative clause ((5)a), and coordinate structure ambiguities ((5)b). These possible points of confusion arise because the grammar allows for the legal attachment to two positions in the syntactic tree, like in (4) where the PP *with binoculars* can modify either the verb *saw* or the noun *cop* (as can be seen in the two alternative trees (4) a and b).

(4) The spy saw the cop with binoculars.

a.



b.



(5) a. The propeller of the plane which the mechanic of the plane ...
b. The old men and women stayed at home.

Disambiguating these structures when processing speech will be guided by contextual, semantic and lexical valence factors (e.g., Trueswell et al. 1993; Crain and Steedman 2010), but prosody has also been shown to have an influence. The relative strength of prosodic boundaries impacts attachment preference (Carlson et al. 2001; Clifton et al. 2002): if the prosodic boundary immediately preceding the ambiguous constituent is larger than the one separating the candidate attachment sites (e.g., *saw* and *the cop* in (4)) then high attachment is preferred and if it is not, then low attachment is. Equally sized boundaries can function as a neutral prosody permitting both interpretations.

Schafer 1997 studied attachment preferences for multiple phrasing variations (e.g., (6)). This research found support for the prosodic visibility hypothesis: words are more prominent candidates for attachment if they are being processed within the same prosodic phrase as the unresolved ambiguous node. So the preferred reading of (a) is attachment to *the bus driver* and for (b), it is to *the rider*. (6)d is truly ambiguous with no preferred interpretation. (6)c shows a bias towards *the bus driver*, stemming from non-prosodic factors, as both NPs are made available within in the current prosodic phrase.

(6) a. The bus driver angered the rider / with a mean look.
b. The bus driver angered / the rider with a mean look.
c. The bus driver angered the rider with a mean look.
d. The bus driver / angered / the rider / with a mean look.

The preferred interpretation of attachment sites is also swayed by prominence features. A pitch accent can act as an attractor when there are different candidate sites. Schafer et al. 1996 tested sentences like those in (5)a, where *propeller* or *plane* are attachment site candidates and found the more prosodically prominent noun was selected by listeners. Schafer formulated the *focus attraction hypothesis* that states that listeners assume more information (e.g., a relative clause) will be conveyed regarding important/focused elements than about unimportant/unfocused elements.

Duration cues play a significant role in disambiguating the intended meaning of coordination ambiguities as in (5)b, where it is unclear if only the *men* are *old* or if both the *men and women* are *old*. Lehiste et al. 1975 found that manipulating the duration of the conjuncts could flip the interpretation.

2.1.5.2 Garden-path sentences

Syntactic ambiguities can exist within the global structure of a sentence (as we have just seen), but they can also exist at specific local positions; while processing a sentence linearly, we are led to one interpretation, but when we reach a syntactic cue that does not conform with this initial reading, we must reanalyze the sentence to obtain an acceptable structure. These are know as **garden-path sentences**.

Grillo et al. 2018 conducted a production study to assess the prosody of garden-path sentences and found that ambiguities that arise while reading are differentiated in spoken language. Speakers read two versions of sentences that contained identical sequences of words except for the presence/absence of a coordinating conjunction which alters the syntactic structure (e.g., *The radio reported that the owners offered tempting food (and) gulped it down.*). The sentences with relative clause structure were read faster from the noun head (*owners*) until the point of the critical disambiguating word.

As well as timing differences, garden-path sentences are likely to be differentiated by prosodic phrasing (Kjelgaard and Speer 1999; Nagel et al. 1996), like in phrasal verb/non-phrasal (e.g., (7) a and b), NP/reduced complement clauses (e.g., (7) c and d) and early/late subordinate clause closures (e.g., (7) e and f) ambiguities.

(7) a. He checked the guests in / at the hotel.
b. He checked the guests / in the morning. (e.g., for their vaccination passes)
c. The company owner promised the wage increase to the workers.
d. The company owner promised / the wage increase would be substantial.
e. When Roger leaves / the house is dark.

f. When Roger leaves the house / it's dark.

2.1.6 Semantic effects

Speakers use prosodic cues to distinguish between different semantic interpretations. These cues can have an effect on the processing of semantic roles and the scope of negation and even the truth conditional status of an utterance.

2.1.6.1 Semantic roles

Different types of intransitive verbs have been associated with different nuclear stress patterns. Whether or not the subject *boy* in (8) a and b (Irwin 2011) will have nuclear stress, will depend on the type of verb that follows and the type of argument it takes. In a broad focus condition, unergative verbs (i.e., verbs that have an agent as subject) present with a verb-accent pattern (as in (8)a). Unaccusative verbs (i.e., verbs whose subject is not an agent but a theme[2]) present with a noun-accent pattern (as in (8)b).

(8) a. A boy JUMPED. (**Verb-accent pattern**)
b. A BOY fell. (**Noun-accent pattern**)
c. Allison ate the cake / with a large fork. (**Instrument**)
d. Allison ate / the cake with the chocolate ganache. (**Modifier**)

The semantic role of a constituent will also provide clues to the constituency structure of an utterance, which we have seen can influence prosodic phrasing. For example, the PPs in (8) c and d, which are both introduced by the preposition *with* have different semantic roles: *with a large fork* is an instrument attached to the verb, whereas *with the chocolate ganache* is a modifier attached to the noun *cake*.

2.1.6.2 Truth-conditional effects

The focus sensitive particle *only* is known to interact with truth-conditional value of a proposition. Take for example the sentences in (9) (Rooth 1992). If the situation is such that Mary introduced Bill and Tom to Sue and made no other introductions, then the truth-value of (9)a is false and that of (9)b is true. The placement of pitch accents in these types of constructions can therefore have a drastic effect on the interpretation of the meaning.

(9) a. Mary only introduced $[\text{Bill}]_{\text{F}}$ to Sue.
b. Mary only introduced Bill to $[\text{Sue}]_{\text{F}}$.

[2] A theme is a participant that undergoes a change or experiences an action, as opposed to an agent who causes an action.

2.1.6.3 Negation scope ambiguities

In potentially ambiguous sentences involving the scope of negation (i.e., the amount of the sentence that is being negated), prosody is a critical for differentiation. The two possible readings of (10), the narrow reading (a) where William does not drink can be contrasted with the broad reading (b) where William does in fact drink. These two renditions are distinguished by the number of prosodic phrases and boundary tone contours (Hirschberg and Avesani 1997); the narrow reading is split into two phrases and ends in a low tone; the broad reading is a single intonational phrase ending with high tone.

(10) William doesn't drink because he's unhappy.
   a. Narrow scope: William does not drink and the reason for him not drinking is his unhappiness.
   b. Broad scope: William does drink, but the reason is not his unhappiness.

2.1.6.4 Reference resolution

Coreference resolution is the task of linking entities in the discourse to other mentions of the same entity or a derived version (e.g. the house -> the roof (of the house)) also present in the text. Prosody has been shown to influence the interpretation of pronouns in ambiguous cases. For example, in (11)a, the expected referent of *he* is *John*. However, if the referent is in fact *David*, then the speaker will signal this prosodically by making the pronoun more prominent ((11)b). Different explanations have been proposed for this change in interpretation. Explanations based on Centering theory (Grosz et al. 1995), which posits that pronouns usually refer to the most prominent entity in the discourse[3] and that subjects are more prominent than objects, explain pronoun accenting as a method of overriding the default assumption and signalling a shift in topic (Kameyama 1999). Venditti et al. 2002 explain shifts in reference as a side effect of the interpretation of the discourse relation between segments (Occasion vs. Resemblance). Jasinskaja et al. 2007 point to the need for a contrasting alternative in the discourse to licence pronoun focus.

(11) a. John hit David and then he hit George. (he = John)
   b. John hit David and then HE hit George. (HE = David)

2.1.7 Discourse effects

Prosody provides important clues that allow listeners to interpret how the elements of the speaker's discourse fit together. This happens at both the global and local levels. At the global level, speakers signal changes in topic (shifts between major branches in the discourse

[3]Less prominent entities can also be presented as pronouns, but only if the current *center* is also a pronoun.

tree) and at the local level, they indicate the type of discourse relations between adjacent units as well as the hierarchical structure between these smaller units.

Topic change is marked by high pitch onsets and low pitch closes (Yule 1980; Grosz and Hirschberg 1992; Smith 2004). Utterances between the opening and closing of topic units gradually decline. In addition to changes in pitch, other acoustic correlates of discourse structure have been observed. The pauses between discourse segments increase with their distance on a discourse tree, as well as increased energy in the elements following a move to the next tree branch (Tyler 2013). The structure traced by prosody can also influence coreference resolution, as it can override recency bias and encourage the listener to look for referents in a larger linear span, in a segment that dominates the current one (Grosz and Sidner 1986; Khosla et al. 2021).

Phrase accents and boundary tones play an important role in communicating how an utterance should be interpreted with respect to the surrounding utterances (Pierrehumbert and Hirschberg 1990). A high boundary tone signals that the current utterance should be understood with respect to the future utterances. This is typical of question intonation. A falling tone signals the unit is complete, as in typical declarative intonation.

Efforts to classify discourse relations based on prosodic features with machine learning have had mixed results (Kleinhans et al. 2017; Murray et al. 2006), with some relations being easier to identify than others. Tyler 2014 studied human differentiation of ambiguous discourse relations, like the example in (12), where the second and third sentences could either be in a coordinate or a subordinate structure with the first sentence. In the coordinate case, three separate events are described. In the subordinate case, the history class is the main event and the next two sentences elaborate on that event. The results showed that rising or falling pitch at the end of the first sentence could bias the interpretation.

(12) I sat in on a history class. I read about housing prices. And I watched a cool documentary.

To discourse effects, we can also include the influence of the discourse mode and the relationship between discourse participants. Formal and informal speaking styles differ (Sityaev et al. 2007), as do read speech and spontaneous speech (Howell and Kadi-Hanifi 1991). Participants in a conversation also adapt to the style of their conversation partner in a phenomenon known as entrainment (Edlund et al. 2009; Michalsky et al. 2018). Furthermore, speakers use prosody to convey their propositional attitudes. This includes sentiments such as uncertainty, incredulity and surprise (Bolinger 1982). They are expressed using different tunes (i.e., combinations of pitch accents, phrase accent and boundary tones).

2.1.8 Information structure effects

When interlocutors participate in a conversation, they must keep track of what knowledge is present in the common ground. And every new contribution or modification to that common

ground must be packaged in such a way that the conversational partner can easily interpret how the new information fits into the shared representation. This can be accomplished through prosodic, syntactic or lexical means; here we focus on prosody.

Dahan et al. 2002 studied the online processing of pitch accents as a marker of new/given status. They displayed items on a computer screen which shared primary syllables (e.g., *candle*/*candy* along with other distractor objects, and they gave subjects instructions to move objects around. In situations where one of the target items had previously been mentioned, the use of a pitch accent on the initial syllable caused participants to focus their gaze on the theretofore unmentioned object and an unaccented version of the initial syllable had the opposite effect, functioning as an anaphoric marker. In a second experiment, they controlled the semantic role (theme or goal) of a target item in the first instruction (e.g., *Put the candle below the triangle* or *Put the necklace below the candle*) and then prosodically focused the target word in the second instruction. In the goal condition, the accented word was interpreted as referring to the previously mentioned word, as opposed to a new item that had not been part of the discussion up to that point. This demonstrates that the new/given distinction is not so clear cut; a previously known entity can be highlighted when it is entering into attentional focus. See Watson et al. 2008 for similar work on established contrastive sets.

An entity that has previously been part of attentional focus can also lose its saliency if it has not been part of the discussion for awhile. If this entity is to return to focus, it will likely be pitch accented.

Other factors such as the scope of focus (as we saw in the previous chapter) and contrastive focus will influence prosody. We will treat the latter topic in Chapter 5.

2.2 What do Transformer language models know about linguistic context?

In this section, we explore a tool that could be used to anticipate future context and to enhance the contextual understanding of known text: the Transformer language model.

2.2.1 What are Transformer language models?

Transformer LMs (e.g., Devlin et al. 2019) are the current state of the art in language modelling. These models offer considerable improvements over earlier language modelling techniques such as n-grams (Shannon and Weaver 1949; Jelinek 1976), or even earlier neural network architectures (Bengio et al. 2000; Mikolov et al. 2013) because they can successfully adapt their representations to fit the sentential context. So for example, when you have a polysemous word (e.g., bank, which can either refer to a financial institution or the side of a river), the model is able to differentiate between the different meanings by attending to the other words in the sentence. This is done through the use of a self-attention mechanism; this

allows a model to learn the relevant relationships between the input tokens and to weigh the influence of all tokens in the shaping of an individual token's representation.

A further innovation of Transformers is the use of multiple self-attention heads. By duplicating self-attention units on the same data, each head can develop a specialization in a specific linguistic task. For example, one head can learn the relationship between a direct object and its verb, while another focuses on the relationship between a determiner and its head noun (Clark et al. 2019).

2.2.1.1 Training and architecture

These models are trained in a self-supervised manner. Using a large corpus of text as input data, the models are trained to recreate the sequences in the original texts. In so doing, they are able to learn linguistic features without any specialized teaching. Transformer LMs come in a variety of designs. These can be classified into three types: (1) Encoders (e.g., BERT (Devlin et al. 2019), Electra (Clark et al. 2020)) (2) Decoders (e.g., GPT-2 (Radford et al. 2019), Transformer-XL (Dai et al. 2019)) and (3) Encoder-Decoder (e.g., Bart (Lewis et al. 2020), T5 (Raffel et al. 2020)) (See Figure 2.1).

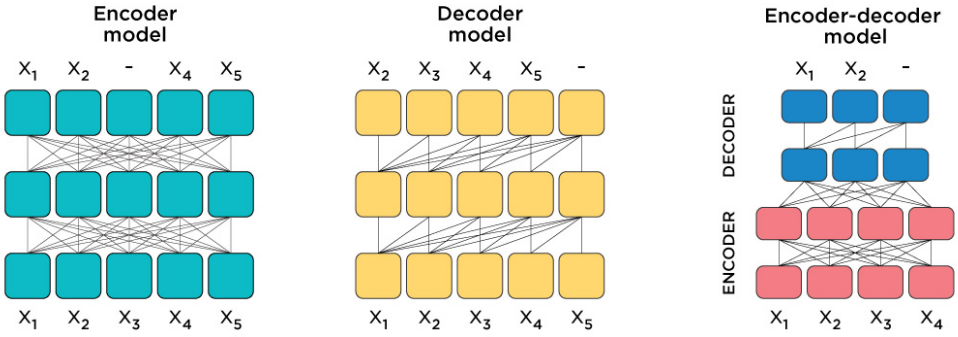


Figure 2.1: Types of Transformer language models.

These models differ in their architecture and their training regime. Many variations on training LMs have been proposed in the literature. Here we will look at three illustrative examples: GPT-2, BERT and BART. GPT-2, a decoder model, is a traditional LM in that it is trained to predict the next word when given a sequence of past words. Decoder models mask out all future tokens in their input when training. BERT is an encoder model that is trained using masked language modelling (MLM): random tokens in the input sequence are masked out and they must be reconstructed using the unmasked words as context clues. BERT is further trained with a next sequence prediction task (i.e., it must decide if a given sentence is the true next sequence or a random one). BERT is bidirectional and so it has the advantage of being able to see all the (unmasked) tokens in the sequence. While GPT-2 has a more limited context, its left to right processing makes it more conducive to text generation. BART is an encoder-decoder that combines the attributes of BERT (bidirectionality) and GPT-2 (text

generation abilities). It is trained with MLM, with both individual words and spans of words masked out. Furthermore, BART's training includes the random shuffling of input tokens. The decoder must reproduce the original sequence in the correct order.

To deal with potential out-of-vocabulary words, the common practice of using subword tokens has been adopted in NLP. Various techniques, such as WordPiece (Kudo and Richardson 2018), Byte-pair encoding (BPE) (Sennrich et al. 2016) and Unigram (Kudo 2018) have been proposed. These methods either break apart or build up words from n-gram units until a desired vocabulary size can be achieved, with special considerations for keeping common words as single tokens. Subword tokenization is not quite the same as decomposing words based on morphology; the subwords are derived from frequency features in the training corpus as opposed to semantically relevant morphemes. It is however easy to implement, it reduces the number of computations that must be performed with character based models and its use has resulted in impressive results.

All Transformer LMs take word embeddings as input. These are dense vector representations that are learnt when training the models. The embeddings learn the features of distributional semantics (Harris 1954); that is, they learn that words used in similar contexts have similar meanings. The words/embeddings position themselves within the vector space to reflect similarities and differences in meaning. The specific dimensions of meaning that are taken into consideration will depend on the model's training objective and training data (Vulić et al. 2020; Wei et al. 2021).

The self-attention layers use a query (Q), key (K), value (V) system (represented mathematically in Equation 2.1). The name originates from a filing system metaphor where the keys represent the file labels and the values represent the contents of the files. Each token in the input acts as a query that is compared to the keys of all tokens in the sentence. This can be seen as the model searching the file labels for relevant information. When there is a high similarity between the query and the key, greater attention is paid to the corresponding value from that file. In practice, the query-key similarities are used to build a similarity matrix (W_{att}) for the input sentence, and then, the output vectors of the self-attention layer are given by the weighted sum (derived from the similarity matrix) of the values. A common pattern seen in attention weights is all attention focused on punctuation or separator tokens. This is interpreted as the model returning a null search result for a given linguistic function (Clark et al. 2019).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V = W_{att}V \tag{2.1}$$

Transformers do not naturally encode the linear order of text since they are permutation invariant. To allow the models to learn important linguistic features such as word order (*Dog bites man* vs. *Man bites dog*) and subordination, Transformer LMs usually employ positional embeddings. These embeddings encode distance measures so the model can see which words are close together (and hence are more likely to impact each other's representations).

2.2.2 Techniques for exploring language models' knowledge

Transformer LMs have contributed to improved performance on a number of NLP tasks. However, what exactly their contribution is is not always clear due to the black box nature of neural networks. To help explore the kind of knowledge these models encode, several techniques have been developed. These include probing, perturbed masking, curated test samples and studying continuations.

Probing (e.g., Jawahar et al. 2019; Klafka and Ettinger 2020) is typically done by extracting the hidden representations from the primary model and using those representations to train a simpler secondary model on a task that requires specific linguistic knowledge such as syntactic or semantic structure. If the simpler model is able to perform well on the task, this is a good indication that the primary model has successfully encoded the information of interest.

One common approach to probing is to extract the hidden representations from the primary model at different layers and evaluate the performance of the secondary model on the task using each layer's representations separately. This can reveal whether certain layers of the primary model are more specialized for encoding particular linguistic properties.

An alternative exploration technique called **perturbed masking** was proposed by Wu et al. 2020. This method does not involve a secondary model, eliminating the possibility that the additional parameters introduced by the secondary model are responsible for some of the prediction accuracy. Instead, perturbed masking measures inter-word correlations by measuring the distance between the hidden representation of a single masked word and the representation of that same word when a second word has also been masked. The resulting impact matrices can be used to induce the dependency and constituency trees implicit in the Transformer.

Curated test samples (e.g., Ettinger 2020) can be used to evaluate performance on specific linguistic task. These studies usually employ surprisal (i.e., the negative log probability) as a measure of the model's representation (i.e., the model should not be surprised to see human expected words in a given slot and should be surprised to see the unexpected). Studying **continuations** made by the models is another possibility (e.g., Aina and Linzen 2021). By generating future text from a prompt, this provides clues as to how the model interprets the prompt.

2.2.3 Transformer language model's linguistic representations

In this subsection, we will look at what Transformer LMs know and don't know about linguistic context. We caution that the studies reviewed here apply to BERT and similar sized Transformer models. The rapid revolution that is currently taking place in language modeling, with larger and larger models overcoming the limitations of smaller models, is very promising for future applications. However, for the moment, the integration of GPT-3/4-like models

(Brown et al. 2020), with billions of parameters, is impractical for a portable iTTS system.

2.2.3.1 Representations at different layers

Probes into Transformer LMs show that they build different linguistic representations at successive layers in their architecture and the representations at these layers will depend on the training objectives of the model. Voita et al. 2019 studied the evolution of embeddings from MLMs and causal LMs using mutual information; embeddings in causal models start out building a representation of the past and then gradually forget that information to focus on projecting into the future. MLM embeddings also develop a contextual understanding at early layers and then forget their own identity in middle layers before rebuilding it at the last layers.

Probes conducted by Jawahar et al. 2019 show that BERT learns phrasal information at the lower layers, it learns syntactic information in the middle layers and it learns semantic information at the higher levels. This was tested using the BERT CLS tag (a special token that is used to learn sentence representations) at each layer and a suite of sentence probes developed by Conneau et al. 2018. These tests evaluate both simple surface traits (number of words in the sentence and presence of a specific word) and more complex syntactic and semantic features: sensitivity to word order, knowledge of tree depth, the sequence of high level constituents (e.g., NP VP), tense of the main clause, the subject and object number, sensitivity to random substitutions (e.g., *...I wanted to know if it was real or a spoonful* (orig: *ploy*) and to coordinate clause inversions (e.g., *They might be only memories, but I can still feel each one. →I can still feel each one, but they might be only memories*).

Klafka and Ettinger 2020 found that BERT more evenly distributed information about semantic and syntactic traits of the sentence tokens than GPT. For example, the animacy of the direct object can be predicted from all individual word embeddings in BERT, but only from the direct object embedding itself in GPT.[4]

2.2.3.2 Word sense disambiguation

As previously mentioned, Transformer LMs are quite good at learning distributional semantics and word sense disambiguation. This has been tested using k-means clustering of word senses (Wiedemann et al. 2020, Chawla et al. 2021). Models (of roughly equivalent size) trained with MLM outperform causal models on this task. All LMs do however show some limitations when it comes to subtleties of word composition; they lean heavily on sentence word content for predictions and do not necessarily pay attention to some word order distinctions.

Shwartz and Dagan 2019 looked at meaning shift and implicit meaning in Transformer representations. Meaning shift refers to changes in meaning that result from lexical compo-

[4]This study only looked at the final layer in each of the models; the results may be different at other layers since to successfully predict the next word/direct object, GPT would have to learn the types of arguments typically associated with the verb.

sition, like in multiword expressions (e.g., *carry* →*carry on*). Implicit meaning refers to the implied relationships that result from compounding (e.g., *olive oil* →made *of* olives; baby oil →made *for* babies). Transformer models were able to successfully distinguish differences from the first category, but struggled with the second. In similar work, Yu and Ettinger 2020 compared LM phrase representations to human paraphrase similarity scores. On global datasets, the models correlated well with human judgments but when controlling for word overlap (e.g., adult female = female adult; law school ≠ school law), performance dropped significantly.

2.2.3.3 Syntax

Syntactic probes show that LMs have a fairly solid grasp on the grammatical structure of language. Clark et al. 2019 found certain attention heads were specialized in attending to syntactic relationships (e.g., focusing on the direct objects of verbs or the noun head of determiners). Furthermore, POS and constituent tagging probes achieve high scores (Tenney et al. 2019; Hewitt and Manning 2019).

While Transformer LMs often perform well on syntactic tests, that is not to say that they process language in the same way as humans. For example, their ability to develop symbolic rules that can be applied to unseen word combinations is not as robust as in humans. Goldberg 2019 studied BERT's subject-verb agreement preferences on syntactically correct but semantically nonsense sentences (e.g., colorless green ideas **sleep/sleeps** furiously). BERT preferred the correct verb form for approximate 85% of the tested sentences. This is impressive, but not on par with humans. Wei et al. 2021 found that BERT could generalize subject-verb agreement to lexical items that had not been paired in the training set, but the application of this rule was dependent on both the relative frequency of the competing verb forms in the corpus and the absolute frequency of the verb (i.e., a minimum number of samples must be seen before verb rules can be mapped to a lexical item).

Aina and Linzen 2021 looked at the syntactic representations held by causal LMs (GPT-2 and an LSTM model) when they encounter ambiguous syntactic junctures in a sentence (e.g., NP/S ambiguities→*The scientist proved the theory* ... a) *through two experiments* (NP) b) *was correct* (S); NP/Zero-complement ambiguities→*Even though the band left the party* ... a) *I stayed* (NP) b) *went on for another hour* (Zero-complement). They generated multiple continuations from the locus of ambiguity and measured the portion of responses that conformed with one or the other syntactic reading. These tests showed that GPT-2 is able to hold multiple structural representations at once for NP/S ambiguities, generating both types of continuations. For NP/Zero-complement, GPT-2 had a strong preference for NP interpretation, but was (usually) able to adjust its representation when disambiguating clues became available after the locus.[5]

[5]While not something we tested in this current work, LM's ability to maintain representations of alternative syntactic representations at locally ambiguous junctures could be exploited for iTTS.

2.2.3.4 Common sense and pragmatics

An area where BERT and its ilk are far from human capacity is in pragmatics and common sense knowledge. Ettinger 2020 analyzed BERT using a number of psycholinguistic tests originally designed to study linguistic processing in humans. These evaluations probe BERT's "reasoning" skills by looking at probabilities of pragmatically likely/unlikely cloze completions when given a specific context or when a predicate has been negated. The first type of test requires inferences to be made regarding the context and the pragmatic relationship between sequential sentences (e.g., *He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ...* →*lipstick/mascara/bracelet*). BERT did assign high probability to the correct completions at least half of the time, but inappropriate completions were also assigned a high probability. As for negation, BERT showed clear weaknesses, failing to modify its predictions to reflect the change in polarity. So likely continuations of *A robin is a* and *A robin is not a* were essentially the same (e.g., *bird*, *robin*).

More evidence that BERT relies on heuristics for its decision making come from McCoy et al. 2019. They tested BERT on a natural language inference task, where machine models have to decide whether one sentence entails another (e.g., *The banker near the judge saw the actor* entails *The banker saw the actor*, but does not entail *The judge saw the actor*). BERT performed very poorly on a dataset that excluded examples that could be predicted based on shallow features such as lexical, subsequence or constituent overlap.

2.2.3.5 Coreference and information structure

The same pattern seen for syntax and semantic probes also applies to coreference and information status: Transformer LMs have globally good results, but falter on difficult cases. Sorodoc et al. 2020 looked at referential representations in Transformer-XL (Dai et al. 2019). They probed the model to see if it could identify previous mentions of pronominal anaphora. The probe yielded very high results. The model learnt that nouns and other pronouns are likely antecedent and that the pronouns should agree in gender and number. The model performed less well on challenging examples where there was a mismatch in number features (e.g., *the audience* →*they*) or when there were distractors (i.e., intervening nouns/pronouns that share features of the target pronoun). Loáiciga et al. 2022 extended this work to look at new/given status in entity representations in general. This probe also resulted in high accuracy.

Apart from information status, information structure representations (topic and focus) in LMs has received very little attention, with the exception of Fujihara et al. 2022. Fujihara et al. looked at topicalization decisions in Japanese LMs; Japanese is a topic-prominent language where the topic is marked with the particle (*wa*).[6] The authors prepared a dataset of sentences where the initial NP could be a topic (marked with *wa*) or a subject (marked with the subject particle *ga*). On a subset of samples, human preferences for topic or subject changed when

[6]This type of discourse element should be more apparent for a Japanese mono-modal LM than it would be for one trained on English where topicalization is usually marked prosodically.

broader context was made available; the LM preferences on the other hand stayed the same with both limited and extended context.

2.2.3.6 Discourse relations

Shi and Demberg 2019 applied BERT to the task of implicit discourse relation classification. This task cannot use explicit discourse markers (e.g., *therefore*) to intuit the relationship between sentences; it must rely on the semantic content of the two propositions and learn the typical reasons and consequences for different types of events. A fine-tuned BERT improved results by a wide-margin over previous techniques, a success that the authors attribute to BERT's next sentence training objective. Even so, accuracy scores were only slightly over 50 percent for the 11-way classification task.

2.3 Discussion and conclusion

In this chapter, we have seen that context can affect speech and speech perception in a variety of ways. If we look at these factors from a TTS/iTTS perspective, we can see that some of these contextual constraints can be inferred from text alone with a fairly local context. Others will require at least a full phrase or sentence and possibly knowledge of the semantic content of the message. Others still will require knowledge of the broader discursive context (beyond the sentence).

Phonological environments can be accounted for with only a few phonemes lookahead. Rhythmic clashes could possibly be avoided by looking at the next (couple of) word(s). Garden-path sentences can be differentiated at some point in the current sentence, and usually quickly after the point of ambiguity. Incrementally derived knowledge of the semantic characteristics of words could help in predicting their syntactic attachment and their relative prominence.

Other factors like global attachment ambiguity or discursive context can only be confidently inferred (statistical biases notwithstanding) if a larger context is made available. Prosodic boundary prediction is complicated in the incremental setting, both due to structural ambiguities which may have not yet been resolved and because global phrase length considerations is a factor in determining breaking points.

To what extent LMs can help resolve ambiguities, provide plausible future context and aid prosody predictions is the major theme of this thesis. Research probing LMs shows that the amount of information gleaned from text alone is very impressive. But given the fact they lack embodied world experience, there are limitations to how much they can know about the intricacies of human experience and pragmatic reasoning. Nonetheless, LMs do have a large potential to improve TTS compared to the simple mapping of unenhanced phoneme sequences to speech.

In the next chapter, we put LMs aside and investigate how vanilla TTS models make use of pure (not LM-enriched) textual context. In Chapter 4, we return to LMs and use them to predict future context for iTTS. In Chapter 5, we investigate the use of LM-enriched text as well as additional previous context for prominence and boundary prediction.

Chapter 3

Incremental text-to-speech

Contents

In this chapter, we will discuss incremental text-to-speech (iTTS). We will begin by presenting the standard (non-incremental) text-to-speech pipeline. We will then survey the relevant issues that arise when building an iTTS model and review the research that has been carried out on these topics. And finally, we will present our first contribution to the field of iTTS, the work presented at Interspeech 2020 *What the future brings: Investigating the impact of lookahead for incremental neural TTS*. When we started this work, neural iTTS had received almost no attention from the research community (with the exception of Yanagita et al. 2019) and so we posed the question: how much future context is necessary in a seq2seq paradigm?

3.1 Text-to-speech synthesis

Text-to-speech, be it incremental or not, is a process that is divided into three steps (See Figure 3.1):

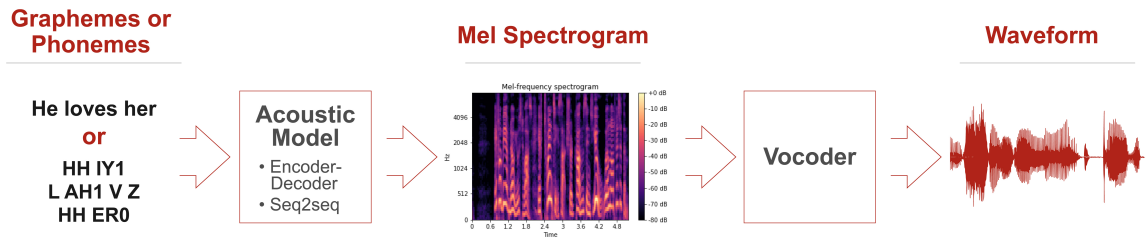


Figure 3.1: Neural text-to-speech pipeline. Grapheme or phoneme sequences are converted into Mel-spectrograms with an acoustic model. The Mel-spectrograms are then converted into waveforms with a vocoder.

1. The front-end which consists in text normalization (i.e., transforming ambiguous forms such as numbers and abbreviations into an orthographic representation) and grapheme-to-phoneme (G2P) conversion.

2. The acoustic model that converts the phoneme sequence into a compact representation of the speech signal (usually a Mel-spectrogram).

3. A vocoder that converts the compact representation into a waveform.

These steps are not fixed: G2P is not obligatory, as TTS models are capable of learning mappings directly from grapheme representations, however the use of a phonemic representation usually results in fewer pronunciation errors.[1] Some newer models bypass the second step and predict a waveform directly from the phoneme input sequence (e.g., Weiss et al. 2021; Donahue et al. 2021; Kim et al. 2021), however these models are difficult to train and most do not quite reach the quality of cascaded systems. In what follows, we will provide an overview of the techniques used for the three stages in modern TTS, with an emphasis on the models used in this thesis. For a more thorough summary of the current state-of-the-art, we recommend Tan et al. 2021.

3.1.1 Front-end

In previous TTS frameworks (HMM-based (Tokuda et al. 2000) and concatenative (Hunt and Black 1996)), the front-end consisted of in-depth textual analysis to predict linguistic features relevant for acoustic feature prediction. These linguistic features included attributes such as quin-phone (the identity of the current phone as well as that of the two preceding and two following ones), part of speech (POS) tags, lexical stress, ToBI features (pitch accents and boundary tones) and positional features (e.g., the position of the current phrase in the utterance). Neural network-based models typically use a much reduced text analysis unit that serves to normalize and disambiguate non-standard text, like abbreviations (e.g., *Dr.*→*Doctor*

[1] Jia et al. 2021 and Kastner et al. 2019 found that training with both graphemes and phonemes could offer more flexibility and reduce pronunciation errors.

or *Drive*), and convert graphemes to phonemes. These tasks are often treated as seq2seq conversions (e.g., Zhang et al. 2019; Yao and Zweig 2015) in contemporary models.

Neural systems let the acoustic model learn which other features from the text are most pertinent directly from the input sequence. This method has resulted in more natural speech, in part because the tools used for linguistic analysis were sometimes inaccurate and this caused a mismatch between the extracted features and the ground-truth audio used for training. The improvements resulting from end-to-end acoustic training (using only text/phoneme inputs) did however decrease the level of control of the system: a given sequence of phonemes will always be pronounced the same way, irrespective of the contextual factors that should influence its expression (See Chapter 2). Recent work (e.g., Kenter et al. 2020; Zou et al. 2021; Talman et al. 2019; Xiao et al. 2020; Hodari et al. 2021), including our own (Chapter 5), attempt to reintroduce linguistic analysis by leveraging language models. The accuracy improvements in language modelling/NLP technology in just the last five years has been seismic and so reintroducing a linguistic analysis stage could help improve contextual appropriateness.

3.1.2 Acoustic models

Speech synthesis is a seq2seq task where the length of the input sequence is not equal to the length of the output sequence. Target waveforms are usually sampled at 22050 samples per second; in normal speech, 10 to 15 phonemes are spoken per second (Levelt 1993); an approximately 1:2205/1:1470 ratio between input (phonemes) and output (speech samples). Using an intermediate feature has been shown to be an effective way to bridge the gap between these two disparate representations and the most commonly used one in neural TTS is the Mel-spectrogram. This is a representation of the frequency content of a signal and its evolution over time. The Mel-scale takes into consideration the way humans perceive differences in sounds, which is not linear: differences between pitches at the lower range are not perceived in the same way as those in the higher range. While Mel-spectrograms are still the standard, some recent research (Lim et al. 2021; Siuzdak et al. 2022) has tried incorporating representations from self-supervised audio models such as Wav2Vec 2.0 (Baevski et al. 2020) and these tests show promising results.

Just like Transformer LMs (Section 2.2.1), TTS acoustic models come in autoregressive (Tacotron (Wang et al. 2017), Tacotron2 (Shen et al. 2018), Deep Voice 3 (Ping et al. 2018), Transformer TTS (Li et al. 2019)) and non-autoregressive varieties (Fastspeech (Ren et al. 2019), Fastspeech 2 (Ren et al. 2021), Glow-TTS (Kim et al. 2020), BVAE-TTS (Lee et al. 2021). We will take a closer look at two of these models, Tacotron 2 (Figure 3.2) and Fastspeech 2 (Figure 3.3), which are illustrative of their respective categories (and the models used for the research in this thesis).

Both Tacotron 2 and Fastspeech 2 are encoder-decoder models that predict Mel-spectrogram frames. Tacotron 2's encoder consists of a series of convolutional layers and a bi-directional LSTM and Fastspeech 2's encoder uses a series of Transformer layers. Despite the difference in architecture, these encoders perform the same task: they build a global representation of

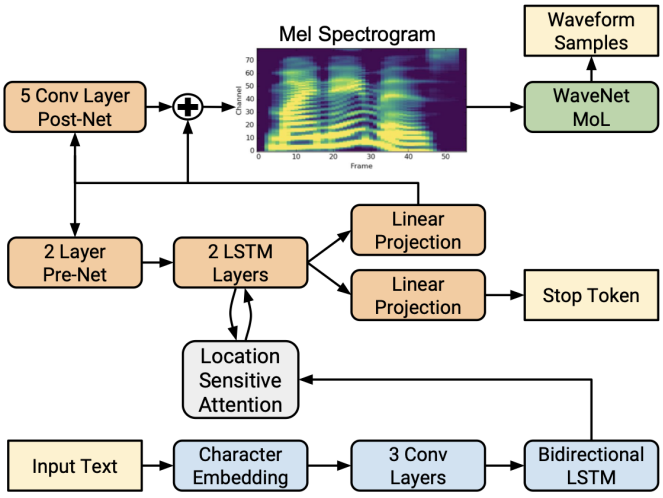


Figure 3.2: Tacotron 2 pipeline. Image from Shen et al. 2018.

the input sequence. However they do this in slightly different ways: Tacotron 2's convolutional layers extract local features and these are then consolidated by processing the data sequentially (both in the forwards and backwards directions) with an LSTM (Long short-term memory) layer; FastSpeech 2's transformer layers have access to the complete phoneme sequence when building its representation. The self-attention modules capture relevant features from the entire sequence in parallel, without the need for sequential processing, which speeds up inference.

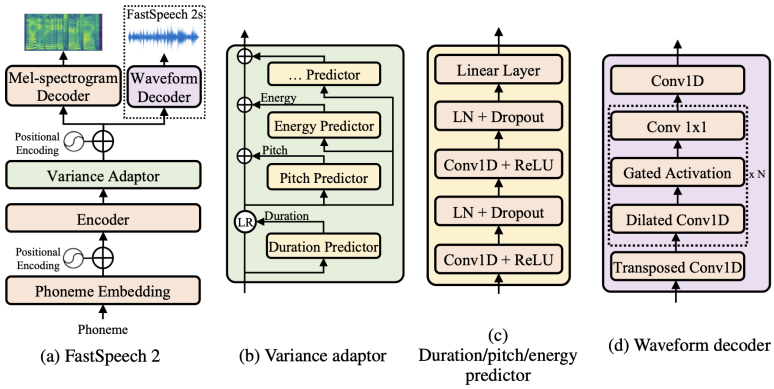


Figure 3.3: Fastspeech 2 pipeline. Image from Ren et al. 2021.

The two models also differ in how they learn alignments between the input and output sequences. To account for the size difference between these two forms, the models need to learn alignments between the phonemes and their corresponding representations within the mel-spectrogram frames. Tacotron 2 uses an attention mechanism that learns transition probabilities based on a context vector (i.e., a representation of the input tailored to the current decoding step), the previously predicted frames and previous context alignment vectors. Fastspeech 2, on the other hand, incorporates an explicit phoneme duration model and then

upsamples the phone embeddings to match the target spectrogram. Alignments for training Fastspeech 2 are either obtained from the attention weights of a teacher model (e.g., Tacotron 2) or through forced-alignment.

Tacotron 2 is trained using teacher forcing; i.e., when predicting the next spectrogram frame in the sequence, the model has access to the ground truth previous frame. This helps the model learn coherent sequences of speech, but it can lead to issues at inference time when successive predictions can drift away from the distribution of natural speech and confuse the attention mechanism. This can result in slurred speech or skipped or repeated sounds. Some methods have been proposed to mitigate this issue including **double feed training** (Shechtman and Sorin 2019) where both the predicted and the ground truth previous frames are passed to the decoder at training time (the two are concatenated together) and **scheduled sampling** (Bengio et al. 2015) where the model is trained on either the ground truth previous frame or the predicted one based on a coin flip using a defined distribution probability.

A further difficulty with training attention-based acoustic models is learning the alignments between the encoder outputs and the decoder steps. To overcome this issue, Tachibana et al. 2018 introduced a guided attention loss which encourages the attention matrix to be diagonal. As a result of this training, the decoder progresses monotonically through the input, but for our purposes in iTTS, it is important to remember that the model does still depend on the full context due to the bidirectional LSTM layer in its encoder.

Training Fastspeech 2 is faster and easier than Tacotron 2 because all the predictions are made in parallel. It is also more robust to unseen data; it does not suffer from the same slurred speech problem. The first Fastspeech model, which also employed a duration prediction + upsampling method, did suffer from less expressivity than Tacotron 2, and that is why Fastspeech 2 introduced additional variance predictors (pitch and energy) to mitigate this issue.

In this Chapter (Section 3.3), we use Tacotron 2 for our experiment, as Fastspeech 2 had not yet been released in 2020. But in later chapters, we switch to Fastspeech 2 because of its speed and robustness.

3.1.3 Vocoders

The current state-of-the-art in vocoding is neural models. WaveNet (Oord et al. 2016) was the first major entry in this paradigm. It is an autoregressive model that uses dilated convolution layers to increase the receptive field and capture long-range dependencies. WaveNet produces very high quality speech, however its inference time is very slow. To make vocoding more efficient, modifications to the autoregressive system have been proposed (e.g., WaveRNN (Kalchbrenner et al. 2018)), and alternative solutions based on generative and source-filter techniques have been developed. Generative techniques include Generative Adversarial networks (GANs), (e.g., Parallel WaveGAN (Yamamoto et al. 2020), HiFi-GAN (Kong et al. 2020)), Flow-based models (e.g., WaveGlow (Prenger et al. 2019)), and Diffusion models (e.g., DiffWave (Kong et al. 2021)). A source-filter model, LPCNet (Valin and Skoglund 2019), uses

a neural network to predict the excitation of the speech signal, but uses a less computational intensive linear prediction method to model the spectral envelope.

In our work, we have experimented with both flow-based (WaveGlow) and GAN-based models (Parallel WaveGAN and Hifi-GAN). These models were selected because of their quality (e.g., acoustic feature reconstruction abilities (Perrotin et al. 2021)) and their inference speed (AlBadawy et al. 2022). While these models are not the absolute simplest/fastest models available, we favoured models with high synthesis quality since our primary concern is with creating natural prosody and we did not want evaluations to be effected by distortions in the vocoding process.

3.2 What considerations for iTTS

Text-to-speech systems, discussed in the previous section, have made great strides with the introduction of seq2seq neural models, combined with end-to-end trainable architectures. These models are able to learn a direct mapping between phonemes and spectrogram or waveform outputs without the need for feature engineering and they produce very natural sounding speech. However, most of these neural TTS systems are designed to work at the sentence level, i.e., the synthetic speech signal is generated after the user has typed a complete sentence. When processing a given word, the system can thus rely on its full linguistic context (i.e. both past and future words) to build its internal representation.

Despite its ability to generate high-quality speech, this synthesis paradigm is not ideal for several applications. For example, when used as a substitute voice by people with severe communication disorders (Augmentative and alternative communication (AAC)) or integrated in a dialog system (e.g. personal assistant, simultaneous speech interpretation, etc.), the system's need to wait until the end of a sentence introduces a latency which might be disruptive to conversational flow and system interactivity.

Incremental TTS (iTTS, sometimes called low-latency or online TTS) aims to address these issues by synthesizing speech on-the-fly, that is by outputting audio chunks as soon as a new word (or a few of them) become available. This task is particularly challenging since producing speech without relying on the full linguistic context can result in both segmental (phonological) and supra-segmental (prosodic) errors (Le Maguer et al. 2013).

There are several factors that need to be considered when building and evaluating an iTTS system. In this section, we will examine these aspects and review the related work that has been carried out in the field.

3.2.1 Unit of processing and text entry interface

The incremental unit of processing (or granularity) has to be determined for an iTTS system (e.g., the word, the phrase), particularly for AAC purposes where the rate of typed input

can be very slow. Both the interface units and synthesis units need to be selected. Some existing systems go as granular as the phoneme (the Synthé 5 and 6 models[2], HandiVoice and Finger Foniks (described in Glennen and DeCoste 1997) or the syllable level (Leblatphone[3]). Synthesis at this rate (i.e. every time a new phoneme/syllable is entered) can be quite difficult for the listener to comprehend, as word boundaries are not well demarcated, and coarticulation and syllabification (in the case of phoneme units) cannot be properly modelled. The full speech sequence must therefore be repeated when the sentence is complete. But this technique can be used to maintain the line of communication, in the same manner as a filled pause in direct human-to-human communication (Ball 1975). Synthesis can be delayed until larger (but still incremental) units are made available, but machine learning techniques will be required to group the phonemes into syllables/words (Bartlett et al. 2009).

An interface that contains syllables or morphemes as input has the potential to speed up text input since fewer keystrokes are required for each word, but it will require the user to learn a new spelling system, whereas a standard alphabetic keyboard may be more intuitive to use for an experienced typist. Selecting the most appropriate interface will of course depend on the user.

Synthesizing speech at the word level does provide units that are easy to delineate, but this may still not provide sufficient input to the TTS model: Saeki et al. 2021a and Yanagita et al. 2019 found that synthesized units smaller than two words were unintelligible. Moreover, processing one-word-at-a-time may not be sufficient to build natural intonational contours. The phrase level corresponds to more natural thought groupings, but determining phrase boundaries, generally but especially in an incremental setting, is not a straightforward task. Some systems (Proloquo4Text[4], Predictable[5]) offer the possibility to store and reuse full sentences, which is very useful for commonly used utterances; but keyboard interfaces are still necessary for more granular and flexible inputs.

In this thesis, we consider increments at the word level (the current chapter, chapters 4 and 5) and at the phrase level (Chapter 5).

3.2.2 Amount of context

Deciding when to trigger synthesis is a fundamental question for iTTS. Is speech synthesized when the current incremental unit is made available or is it delayed to provide the system with some lookahead (i.e., future/right context)? And if lookahead is permitted, does this involve a fixed lookahead policy (e.g., always waiting for the next word/next two words/etc.) or an adaptable one where the delay can be modulated based on the ambiguity of the current context.

[2]http://www.synthe-aria.com

[3]https://hacavie.fr/aides-techniques/essais-d-aides-techniques/articles/machine-a-parler-portative-leblaphone-fabriquee-par-la-sas-leblat/

[4]https://www.assistiveware.com/products/proloquo4text

[5]https://therapy-box.co.uk/predictable

3.2.2.1 Fixed lookahead

In the context of HMM-based TTS synthesis, Astrinaki et al. 2012 evaluated a limited training context (the previous, current and next syllable) to obtain phoneme labels, as well as a limited phoneme label context to obtain vocoder parameters (the current and next phoneme). Subjective and objective evaluations showed only a slight degradation from full context, however the naturalness of even full-context TTS was not very high at this time. Baumann and Schlangen 2012a examined the effects of lookahead on the prosody of the current chunk depending on when the TTS system gained knowledge of the subsequent chunk. The words near the beginning of current chunk were not hugely affected by a lack of knowledge about the future chunk, but if integration was delayed for too long, the level of distortion (pitch and duration) for words near the end of the chunk would rise rapidly. This is because the later words did not form a continuation intonation (preparing to lead into the next chunk), but rather an utterance final intonation. The authors conclude that one chunk of lookahead is sufficient and that the incorporation of that chunk can wait until after the first word of the current chunk has been processed; this allows for a balance between the delivery of speech and processing time (i.e., the critical prosodic updates from the future chunk can be processed while the first word is being spoken).

In the context of speech-to-speech translation using neural TTS, Ma et al. 2020 used a wait-k policy (inspired by the prefix-to-prefix framework introduced for translation in Ma et al. 2019a), which consists in having access to a future context of k input tokens while generating speech output (*Prefix* in this type of system denotes the combined previous and current context). The system must wait for an initial k tokens to be entered, but it then outputs one new token at each subsequent timestep. The authors found that one word of lookahead, for both the Mel-spectrogram prediction and the waveform prediction, gave the best results, however the vocoder lookahead only provided very minor improvements.

3.2.2.2 Adaptable lookahead

A fixed lookahead policy has the advantage of being easy to determine. But while an adaptable policy does require some additional computation, adaptability is the better option with regards to the latency/quality trade-off, because a delay is only tolerated for input that is ambiguous and requires more context to disambiguate. Pouget et al. 2016, in the context of HMM synthesis, proposed an adaptive decoding policy based on the online estimation of the stability of the linguistic features: the synthesis of a given word is delayed if its part-of-speech (POS) is likely to change when additional (future) words are added. While also contributing to synthesis quality, this method improved chunking properties (i.e., deciding on natural breaks to output speech). We extend this work in Chapter 5.

Mohan et al. 2020 investigated the use of reinforcement learning to establish the optimal *read-speak* policy where *read* refers to the encoding of an additional character and *speak* refers to the decoding of the current queue of encoded characters. Their model was able to balance latency and quality and it successfully navigated some ambiguous pronunciation issues, for

example waiting for more information when the character sequence *secret* was encountered, as with more information the pronunciation could change to secretary (/siːkrət/ →/sɛkrəˈtɛri/).

3.2.2.3 History

Considering the maximum amount of history to retain is a further design question. Including all past context from an on-going conversation will soon become unwieldy. Limiting context history to the prefix for the current sentence is more computationally feasible. Martos et al. 2021 tried setting a six word limit for previous context, however they hypothesize that this limited context may be partially responsible for the degraded speech quality they observed.

It is possible to summarize the past context in the form of a context embedding, as done in Saeki et al. 2021a. As far as we know, most of the works on iTTS that incorporate previous context reprocess the past as the context window moves forward. However, if there is some overlap between past contexts (e.g., when a context window moves one step forward, all but the previous "current" word were part of the past context of the previous timestep), there is a potential for computational savings by updating a past context embedding from its previous state instead of recalculating everything from scratch, similar to the average embedding layer proposed for incremental machine translation by Zhang et al. 2020.

3.2.3 Latency

Every stage of the iTTS process has the potential to introduce latencies. These latencies stem from either the time collecting input data, the time processing the data, or from a backlog of previously synthesized audio. Processing speed will depend on the size, architecture and complexity of the model, as well as the strength of the hardware the system is deployed on.

3.2.3.1 Input latency

The latency caused by the input stream will vary greatly depending on the intended application. In the case of automatic interpreters, the rate of input will depend on the automatic speech recognition (ASR) system transcribing the incoming speech and the machine translation (MT) unit that converts the source language into the target language. Dialogue systems will similarly depend on ASR, on natural language understanding (NLU) and on text generation. In AACs the input will depend on the typing speed of the user, which can vary greatly based on the input mode (i.e., manual typing or eye gaze typing) and the user's motor control abilities (Koester and Arthanat 2018).

ASR should not be a major cause of delay, since several current models are able to transcribe speech in real time (See Addlesee et al. 2020 for an evaluation of ASR systems judged on incremental criteria). MT and NLU can similarly be processed quickly, however these applications also have latency/performance tradeoffs and there are stability issues to consider (i.e., is

the translation/meaning likely to change when more input is made available?) (Arivazhagan et al. 2019; DeVault et al. 2011). Input latencies for AAC can be very large but they can be partially reduced through the implementation of predictive text (Judge and Landeryou 2007). Predictive text can either provide next word options for the user to choose from or eliminate input keys based on statistical probabilities of future letters.

3.2.3.2 Computational latency

Input length Inference speed will be affected by the size of the input units. Ma et al. 2020 evaluated the computational latency for TTS in a neural speech-to-speech translation setting. The latency for full sentence synthesis grew in a linear manner with the size of the sentence, whereas their wait-k policy gave a fixed latency. Using wait-k, they were able to achieve a positive time balance (i.e., the next word could be synthesized in the time it took to play the current word).

Architectural design and complexity The architectural choice of model can be a potential cause of latency. Autoregressive TTS models can be slow to produce the speech sequence since you must wait for one frame to be inferred before you can infer the next. Ellinas et al. 2020 propose methods to accelerate this process with their lightweight neural TTS model, designed to be more agile and operable on CPUs. Their modified Tacotron (a combination of both Tacotron 1 and 2), uses a simplified attention mechanism and its decoder infers several spectrogram frames at each timestep; they found that predicting 5 frames/timestep (240ms increments of speech) was comparable in quality to 2 frames/timestep while reducing latency by almost half. Increasing the prediction rate to 10 frames/timestep resulted in a dramatic drop in quality. Non-autoregressive models that predict all Mel-spectrogram frames in parallel could also be used to speed up inference.

Other possibilities for reducing the complexity of TTS models (and in so doing, decrease latency) include techniques such as pruning (Lam et al. 2022) to eliminate redundant or unnecessary model parameters, neural architecture search (Luo et al. 2021) to find the most efficient design, and compression strategies (e.g., quantization, low-rank matrix approximation (Koc et al. 2021) and knowledge distillation (Wang et al. 2021b) to simplify computations. Making the model as efficient as possible is paramount if the iTTS system is to be used on a portable device with limited battery power.

3.2.3.3 Backlog

Liu et al. 2022, Zheng et al. 2020 and Fukuda et al. 2021 call attention to the issue of backlog as a cause of latency. This is an issue that can arise when one audio unit has been synthesized, but the previous unit has not finished playing. This can occur for instance in speech-to-speech translation when the target speech is significantly slower than the source speech. To eliminate this problem, Liu et al. 2022 propose duration scaling (i.e., speeding up the audio) and Zheng

et al. 2020 propose *self-adaptive translation* which modifies the length of the translated text to adapt to the speaking rate of the source speaker. The latter method has the added advantage of reducing unnatural pauses when the speaking rate decreases. A further cause for backlog comes from the asynchronous deployment of modules in a system's pipeline.

3.2.4 Predicting future context

Language is not a random string of words. It has recognizable syntactic structure and collocational/phraseological patterns that make predicting the future text (or at least something similar) possible. Furthermore, language is constrained by context. For example, in a discussion about cooking, the probability of hearing the words *broil, bake* and *steam* is much higher than hearing *abacus, tractor* and *constellation*. Using these constraints has the potential to fill in some of the missing information for an iTTS model.

In Chapter 4, we explore the use of language models for future word prediction. In a contemporary work, Saeki et al. 2021a also investigated the use of "pseudo-text". They (1) predicted five words into the future using GPT-2, (2) passed both the past context and predicted future to a context encoder and (3) conditioned the synthesis of two-word chunks on this context embedding. They further propose fine-tuning the context encoder so that the representations of the pseudo and ground truth futures are closer together. They report positive results from this method, however it is unclear whether they would have achieved similar results using a random future context, as this condition was not evaluated in their listening test.[6] In a follow up paper, Saeki et al. 2021b sped up inference time by training an LSTM model to replace GPT-2; through knowledge distillation, the LSTM learns to predict the context embedding directly from the past context and the current words, as opposed to first predicting the future text and then encoding that representation.

3.2.4.1 Quality of predictions

Liu et al. 2022 also conditioned an iTTS model on pseudo-lookahead as part of their speech translation model. They improved future word prediction by using the speech translation unit, which is conditioned on the source speech signal, to generate their predictions. This resulted in major gains in accuracy (70%+ accuracy versus less than 20% when using a separate language model).

The quality of predictions depends on the prompt given to the language model. With too little context, the predictions are not grounded and hence are likely to simply be the most common words in English (or whichever language it was trained on). Saeki et al. 2021a tested

[6] Saeki et al. 2021a do report the fine-tuned context embeddings give better results than the non-finetuned version, i.e., context embeddings closer to the ground truth result in more natural audio, and the random future embeddings were farther away than the GPT-2 generated ones. The details of the random text generation are not reported; it is possible that the randomly generated text does not match the distribution of function words in natural speech and this could cause the large gap between pseudo and random, since function words are an important clue for phrase boundary prediction in end-to-end models.

this experimentally[7] and confirm that predictions at the beginning of the sentence (after the first word) are worse than those at the end. This issue could likely be overcome by including context from previous sentences (Mikolov and Zweig 2012; Tiedemann and Scherrer 2017), as sentence initial words are usually more predictable because they serve to link the current discourse segment to what has come before (Ferreira and Chantavarin 2018).

Future prediction will never be entirely accurate and incorrect predictions could have a negative effect on speech quality. Skantze and Hjalmarsson 2013 propose *speech plans* that consider multiple dialogue paths at a time. By doing the same, an iTTS model could prepare multiple intonational contours and be ready to deploy the correct one as soon as disambiguating information becomes available.

3.2.4.2 How far into the future?

How much future text to predict is an important question. In the work described in the next chapter, we chose to limit prediction to the next word, so we could study the effects of language model predicted text to random next word generation. Predicting longer stretches of text could be beneficial, as it provides more context to the model, but with each subsequent prediction you risk greater divergence from the ground truth text. As we shall see, predicting beyond a certain point may have little impact on the features of the current word and could therefore add unnecessary computational steps.

3.2.5 Revision/Disfluencies, disruptions and repetition

In an incremental system, it is inevitable that mistakes will be made. This is also true in human speech where disfluencies such as false starts, repetitions and revisions are common occurrences (2 to 26 disfluencies for every 100 words according to Faure 1980). If an error is detected as more data becomes available (i.e., as we get more input text), it may be beneficial for cognitive processing to signal this revised understanding/prosodic representation to the listener. For such a system to function properly, it would be necessary to define/threshold prosodic changes that alter the understanding of a sentence; in other words, we would expect there to be minor changes to the prosodic predictions with each unit of additional context, but we would only want to revise the speech if the new predictions alter the meaning. This would not be straightforward, as shifts in meanings are gradient and dependent on context. Consideration would also have to be paid to the stability of the revisions.

The insertion of disfluencies may be advantageous when the speaker/typist wants to revise the content of their message. In an AAC, we could imagine a restart triggered by the deletion of previously typed words. Other disfluencies could be integrated in order to assist turn-management. For this purpose, Betz et al. 2015 tested the use of filled and silent pauses, word fragments and word lengthening in a dialogue system. Users rated the silent pauses and

[7] The prediction quality was measured using the cosine difference between ground truth context encoder embeddings and pseudo-lookahead context embeddings.

word lengthening positively, but not the other disfluencies; the authors attribute this to a lack of variability in these synthetic features. Skantze and Hjalmarsson 2013 and Baumann and Schlangen 2013 used filled pauses to bide time while a dialogue system waited for crucial information to be processed.

Buschmeier et al. 2012 tested a feature for dialogue systems that allows the system to adapt to environmental factors, such as loud noises. If speech is disrupted, the model stops the ongoing utterance and rephrases the parts that went unsaid. Human raters judged this behaviour as significantly more human-like than a second condition where the model simply paused during the interruption and then continued the already generated utterance. The authors advocate a *just-in-time* processing strategy, where generation is held off until right before it is needed, so that the model can remain as reactive as possible to the communicative context. Similar work was conducted in the context of an in-car dialogue system (Kousidis et al. 2014) where for safety, the system must sometimes pause as to not distract the driver from events that require their full attention.

Also in the context of dialogue systems, Yu et al. 2015 investigated reactivity to the listener's attention. Sentence restarts have been linked with speaker coordination (i.e., if the speaker does not have the attention of the listener at the beginning of their utterance, they will employ pauses and/or restarts as a means to coordinate a grammatical phrase with the listener's shifting gaze), as in the following example (from Goodwin 1981 as cited in Yu et al. 2015):

(1) She - she's reaching the p- she's at the* point I'm

The dotted line shows the listener's moving gaze, * indicates the point where the gaze meets and the solid line shows a period of mutual gaze. Their implemented attention aligning procedure was not entirely successful, as the pausing behaviour was frequently interpreted as the end of utterances/floor releases. This problem could potentially be overcome with modifications to the speech synthesis unit, which could convey desired continuation.

Wester et al. 2017 looked into adaptive systems that respond to user interruptions. They tested a model that would react with increasing levels of annoyance in the synthetic voice upon repeated interruptions by the user. A focus group had mixed reactions to this behaviour, some members were amused by the sassiness of the system, but generally the public felt an artificial agent should remain cooperative. Baumann 2014a investigated methods for aligning synthetic speech with a robot's pointing gestures. Speech tempo was modulated on-the-fly to ensure deictic expressions (e.g., ***this*** *button*) were coordinated with the gesture. Alignments were achieved but at the cost of speech quality degradation.

3.2.6 Model and training modifications

Modifying models or their training regime is another option for improving incremental predictions. These methods take direct account of the demands of on-line processing and prepare

the models to deal with uncertainty or simply to process incoming data as soon as it becomes available.

The *Inpro_iSS* was one of the first speech synthesis systems built specifically for incremental processing (Baumann and Schlangen 2012b). It was part of an incremental processing toolkit for dialogue systems *InproTK* (Baumann et al. 2010; Baumann and Schlangen 2012c). This toolkit handles incremental units (IUs) at different levels of linguistic abstraction (e.g., phoneme, word, pragmatic plan) and keeps track of the dependencies between these units. As the system's state changes, connected units in the network are updated. Unspoken dependent units can thus be modified to suit the current context.

Other early (HMM) iTTS systems developed strategies to deal with uncertainty. (Baumann and Schlangen 2012a; Pouget et al. 2015; Pouget et al. 2016). In this paradigm, models are trained on a set of explicit linguistic features (e.g. number of syllables in the next word) and Baumann and Schlangen 2012a and Pouget et al. 2015 developed coping mechanisms to handle missing features when making predictions for iTTS: unknown future context information is replaced with the most common values for these features at inference time in Baumann and Schlangen 2012a, whereas uncertainty on those features is explicitly integrated at training time by Pouget et al. 2015.

Training TTS models directly on incremental data (i.e. randomly selecting a truncated sequence that is shorter than the full sequence) has the potential to increase TTS quality in incremental mode. Liu et al. 2022 and Ellinas et al. 2020 supplement the full input training data with prefixes. When training on prefixes/sentence fragments, it is important to make the model aware of the position in the sentence as the prosody of words/phrases at the beginning, middle and end of an utterance are very different (e.g., declination (Ladd 1984), boundary tones (Pierrehumbert 1980)). Yanagita et al. 2019 tested an approach which consists in (1) marking three subunits within the training sentences using *start*, *middle* and *end* tags, (2) training a Tacotron 2 TTS model with these tags so it learns intrasentential boundary characteristics, and (3) synthesizing sentences by inputting chunks of length n words (up to half a sentence) with the appropriate *middle* or *end* tag. In subjective tests, it was found that three-word units were indistinguishable from the full sentence speech.

3.3 What the future brings: the effect of lookahead in neural TTS

In this section, we present our own investigation into the topics presented in Section 3.2.2: lookahead and adaptability. This work was presented at Interspeech 2020.

How many future words do you need to know to predict a natural sounding, coherent utterance with a neural seq2seq model? Human speech production is incremental in nature (Levelt 1989); full speech utterances are not usually completely planned out when the utterance begins. A certain amount of lookahead is certainly necessary if one is to satisfy the contextual constraints discussed in the previous chapter. And there is evidence of anticipatory speech

errors (e.g., *Fill the gas up with car* Target: *Fill the car up with gas* (Dell and Reich 1981)) that show that there is some advanced planning. For humans, it has been hypothesized that abstract conceptual plans of the upcoming utterance are sketched out and that specific words/phonemes are slotted into place as they become activated.

Of course an artificial neural network does not function in the same way as a human brain. Where a human is the author of their own conceptual thoughts and this is what shapes their speech, an artificial neural network is relying solely on statistical regularities (that it encountered in the training data) to predict an adequate prosodic form. Now the question is, do these models take advantage of long-range features or do contextual effects remain fairly local? And is the degree of dependence on the future contingent on the specific context?

The goals of the work presented here are twofold: (1) to evaluate the amount of necessary future context for neural TTS models and (2) to pave the way toward an adaptive decoding policy for neural iTTS. Similarly to the HMM-based iTTS system described in Pouget et al. 2016, where synthesis was delayed based on the stability of POS tag predictions, the envisioned neural iTTS is expected to modulate the lookahead (and thus the latency) by the uncertainty on some features due to the lack of future context.

Unlike in the HMM-framework, where models are trained on explicit linguistic features, neural networks learn which features are most important during training. The gain in naturalness provided by end-to-end models is accompanied by reduced interpretability. Because of the black box nature of the models, studying the importance of missing features is a challenging task. To address this, we analyse the evolution of the encoder representations of a neural TTS (Tacotron 2) when words are incrementally added (i.e. when generating speech output for token n, the system only has access to $n+k$ tokens from the text sequence, k being the lookahead parameter). We also investigate which text features are the most influential on this evolution towards the final encoder representation. Finally, we evaluate the effects of the lookahead at the perceptual level using a MUSHRA listening test.

The heart of our investigation here is the influence of incrementality on the acoustic model, as it is responsible for predicting the prosody of the input text. We do however perform some initial evaluations to verify that other components of the TTS pipeline, G2P and the vocoder, are not causing major disturbances. We use the variable k to refer to lookahead (i.e., future context) for both the study on G2P and the acoustic model, however lookahead is defined slightly differently in the two cases. For the acoustic model a trailing space character will affect the output whereas for G2P, it does not, and so space characters are counted as tokens in the acoustic model. See Section 3.3.4.2, for a formal description of the encoding policy.

3.3.1 Models

For these experiments, we use a sequence-to-sequence TTS model, Tacotron 2 (Shen et al. 2018, see Section 3.1.2) (this architecture was state-of-the-art when this study was conducted). We

use an implementation developed by NVIDIA[8]. The TTS model was pre-trained on the LJ Speechset (Ito 2017), a collection of non-fiction books read by a single, female, American speaker. The corpus contains 24 hours of audio recordings. The sampling rate of the audio clips is 22050Hz. The Mel-spectrogram frames have 80 bands and they were computed with short time Fourier transform (STFT) of hann window size 1024 samples (46ms) and a hop size of 256 samples (11ms). For vocoding, we use a flow-based model, WaveGlow (Prenger et al. 2019) also developed and pretrained by Nvidia.[9]

3.3.2 Incremental grapheme-to-phoneme conversion

To test the effects of the incremental mode on G2P conversion, we evaluate the changes from $k = 0$ and $k = 1$ (i.e., a lookahead of no/one word(s)) to the full context on the tokens in the LJSpeech corpus using the G2P algorithm (Park 2019). G2P performs text normalization, differentiates between heteronyms using POS labels, uses the CMU Pronouncing Dictionary (Rudnicky 2015) for unambiguous words, and uses a neural network to predict the pronunciation of out-of-vocabulary words.

We find that only 67 of the 221,097 tokens in the corpus change from $k = 0$ to the full context. This number drops to 13 when G2P is given one word of lookahead. Out of the 67 phoneme shifts, 29 involve a change in voicing (most of these changes are for the word *used* and this change does not affect the token's identity but would have an impact on naturalness) and 38 involve a vowel shift and/or a stress shift (e.g. the representation of *presents* changes from a noun to a verb: *P R EH1 Z AH0 N T S* →*P R IY0 Z EH1 N T S*.

Based on these results, G2P conversion in an incremental setting does not appear to be a major issue for iTTS in English: only 0.03% of the tokens we tested required additional input. English may however be a particularly easy language for this task, since the majority of homographs can be differentiated by their POS and the separation between words is clearly demarcated. Additional modification may be necessary for languages with greater numbers of homographs like Arabic (Azmi et al. 2022) and for languages that also require word segmentation as part of their front-end processing like Chinese (Chang et al. 2008) or Thai (Yamasaki 2022).

3.3.3 Incremental vocoding

To test the quality of incremental, neurally-vocoded audio, we compared distortions between ground-truth utterances and vocoded utterances synthesized with different quantities of Mel-spectrogram frames at each timestep. In other words, we fed the vocoder (WaveGlow) f Mel-spectrogram frames at a time, concatenated the resulting waveforms together, and then compared the root mean squared error (averaged over the frames of each utterance) of the newly computed Mel-spectrograms with the spectrogram from the ground-truth audio. 50

[8] `https://github.com/NVIDIA/tacotron2`
[9] https://github.com/NVIDIA/waveglow

utterances from the LJSpeech corpus, vocoded with chunks of $f = \{1, 2, \ldots, 25\}$ and $f = full$ were used for this evaluation.

While the distortion between full context and incremental inputs does not become statistically insignificant (as measured by a paired t-test) until $f = 25$ (p-value $= 0.17$), the differences in RMSE (Figure 3.4) from $f = 4$ onwards are fairly minor: a Cohen's d test (Cohen 2013) indicates that the mean difference is less than half the standard deviation of the pooled groups $f = 4$ and $f = full$ (Cohen's d $= 0.47$). Furthermore, the distortions are usually isolated to the points where segments were concatenated together. The segments do not necessarily connect smoothly and this results in audible artifacts. We found that using a simple cross-fade could eliminate this issue. Since very few words are less than 4 Mel-spectrograms frames in length (only 3 in our test corpus), we conclude that incremental vocoding does not degrade the quality of speech (provided a cross-fade procedure is implemented).[10]

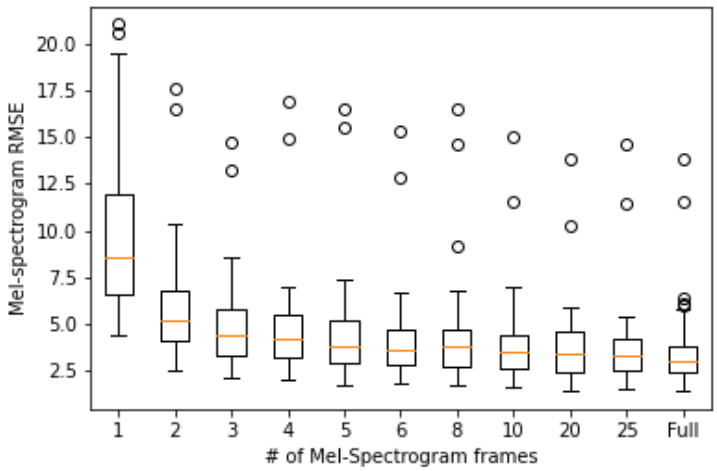


Figure 3.4: Mel-spectrogram RMSE for different value of f (i.e., the number of frames fed to the WaveGlow at each timestep).

3.3.4 Incremental acoustic modelling

3.3.4.1 Test corpus

For our analysis, the test sentences used as input sequences are taken from the LibriTTS corpus (Zen et al. 2019). We filter 1,000 utterances with sentence length ranging from 5 to 42 words. We follow the procedure outlined on Sketch Engine[11] (Kilgarriff et al. 2004; Kilgarriff et al. 2014; Kilgarriff 2001) to verify that word distribution is similar to that of larger general corpora (Brown and BNC corpora). This is done by (1) extracting the 5000 most common

[10] Ma et al. 2020 propose an alternative strategy for handling this problem: they include additional buffer frames around each chunk to be synthesized.

[11] https://www.sketchengine.eu

words in each corpus and then aggregating the two lists (removing duplicates), (2) calculating the keyness score for each word (a measure of the relative significance of a word between corpora (see Gries 2016 for more information)) and (3) averaging the 500 largest keyness scores. The difference measures, using the BNC as the reference, between Brown, our corpus and a specialized Covid-19 corpus (Wang et al. 2020) are 1.59, 2.24 and 4.0 respectively. The small divergence between Brown and our corpus can be attributed to the smaller size of our corpus which contains $34,768$ tokens and $4,085$ types.

3.3.4.2 Incremental encoding policy

We consider an input sequence of tokens, where each token can be either a word, a space or a punctuation mark. We define an iTTS system with the following simple policy (similar to Ma et al. 2020): the encoder starts by reading k input tokens (k is the lookahead parameter) and then it alternates between generating speech output and reading the next token until the complete input token sequence is consumed. Formally, we use the following notations and definitions (see Table 3.1 for an example of the listed items):

- N is the length of the input sequence (in number of tokens);
- $\mathbf{x}_n$ is the token at position n (the "current" token); $\mathbf{x}_{1:N}$ is the complete sequence of input tokens; $\mathbf{x}_{1:n}$ is the subsequence of input tokens from position 1 to position n;
- $\mathbf{y}_n$ is the speech output segment corresponding to token $\mathbf{x}_n$;
- $c(n,k) = \min(n+k, N)$ is the number of input tokens read when generating $\mathbf{y}_n$ (recall that k is the lookahead parameter); $\mathbf{z}_n^{n,k}$ is the corresponding encoder output.[12] In other words, $\mathbf{y}_n$ is obtained after reading the partial sequence of input tokens $\mathbf{x}_{1:c(n,k)}$; $\mathbf{z}_{1:c(n,k)}^{n,k}$ is the sequence of encoder representations obtained so far;

Conventional offline encoding (using the full sequence of input tokens $\mathbf{x}_{1:N}$ at each position n) is also processed for comparison, and $\mathbf{z}_{1:N}^{\text{full}}$ denotes the corresponding encoded sequence.

3.3.4.3 From character to word representations

In the Tacotron 2 model, input sequences are encoded at the character level. However, in our study, we consider an iTTS decoding policy at the word level; this is because token breaks are a natural trigger for synthesis or evaluation in a practical iTTS system. Consequently, we need to go from character representation to word representation. We start from the encoder's bidirectional LSTM network: forward and backward layers each provide a 256-dimensional vector for each input character. For each new token $\mathbf{x}_n$, we extract the output of the forward layer corresponding to the last character of $\mathbf{x}_n$. The forward layer continues up to the last

[12] n is used two times in $\mathbf{z}_n^{n,k}$ since we will see that the value at other positions, e.g. $\mathbf{z}_{n-1}^{n,k}$, also depends on n and k.

Table 3.1: Incremental inputs (for different lookahead k) for sentence "The dog is in the yard." to generate $\mathbf{x}_3$ (the word "dog.").

n	k	$c(n,k)$	Input at $c(n,k)$	$\mathbf{x}_n$
3	0	3	The_dog	dog
3	1	4	The_dog_	dog
3	2	5	The_dog_is	dog
3	...	...	...	dog
3	8	11	The_dog_is_in_the_yard	dog
3	9	$N = 12$	The_dog_is_in_the_yard.	dog

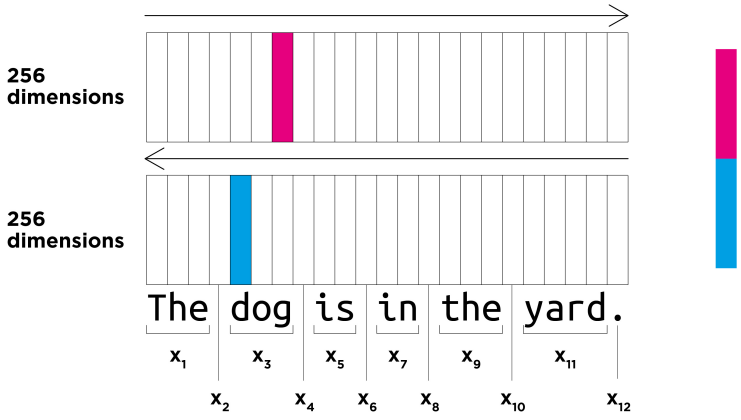


Figure 3.5: Illustration of the word embedding extraction procedure for the word dog. Embeddings are extracted from the LSTM layer of the Tacotron 2 encoder. The embedding corresponding to the first character is taken from the backward layer and the last character from the forward layer. These two embeddings are concatenated together to represent the word.

character of token $\mathbf{x}_{c(n,k)}$. Then the backward layer goes from the last character of token $\mathbf{x}_{c(n,k)}$ to the first character of token $\mathbf{x}_1$. We extract the output of the backward layer corresponding to the first character of $\mathbf{x}_n$. Both vectors are concatenated to get a 512-dimensional vector representation $\mathbf{z}_n^{n,k}$ of $\mathbf{x}_n$. See Figure 3.5 for an illustration of the process for a full sentence input. Note that the input sequence is re-encoded for each new token (i.e., for each increment of n), leading to new values for the sequence $\mathbf{z}_{1:n-1}^{n,k}$. Of course, this sequence also depends on k, which is the purpose of this study. In other words, when encoding the sequence $\mathbf{z}_{1:c_n^k}$ with different values of n (and of course of k), we obtain different values for each vector $\mathbf{z}$ of the sequence, which is not apparent from the notation. In contrast, there is only one single value for the sequence $\mathbf{z}_{1:N}^{\text{full}}$. We keep this notation for simplicity of presentation.

3.3.4.4 Incremental decoding

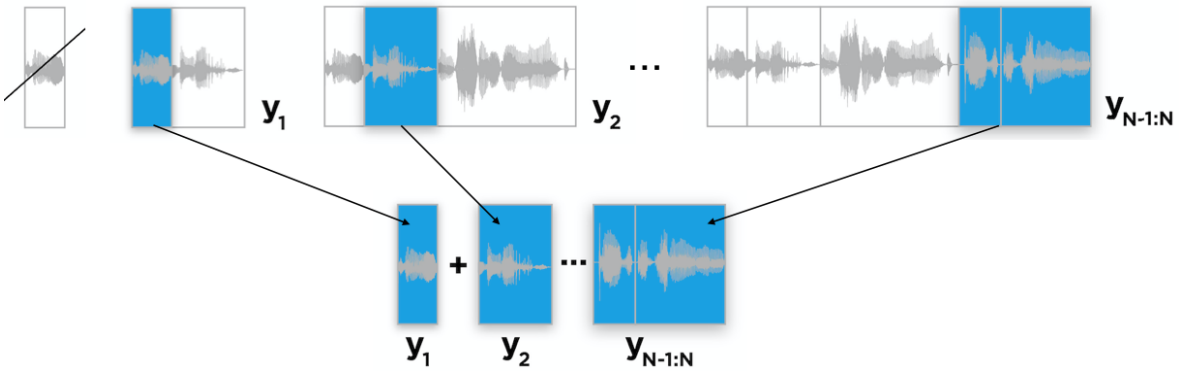


Figure 3.6: Illustration of the incremental speech waveform generation process for lookahead parameter $k = 1$. The available context for the current word (including lookahead) is synthesized at each timestep, but only the waveform corresponding to the current word is retained. The excised waveforms are concatenated together to obtain the full utterance.

We build the iTTS decoder output as follows. For a given value of k, and for the current token $\mathbf{x}_n$, we first produce the speech waveform corresponding to the encoded sequence $\mathbf{z}_{1:n}^{n,k}$. Then, using the Munich Automatic Segmentation system (Kisler et al. 2017) (an automatic speech recognition and forced alignment tool which employs an HMM and Viterbi decoding to find the best alignment between the text and audio), we select the portion $\mathbf{y}_n$ of the waveform corresponding to $\mathbf{x}_n$. Finally we concatenate this speech segment $\mathbf{y}_n$ to the speech segment resulting from the processing of previous tokens, that we can denote as $\mathbf{y}_{1:n-1}$. In short, we simply update the generated speech waveform as $\mathbf{y}_{1:n} = [\mathbf{y}_{1:n-1}\ \mathbf{y}_n]$. For example, for $k = 2$, we extract the speech waveform segment $\mathbf{y}_1$ corresponding to token $\mathbf{x}_1$ from the signal generated from $\mathbf{z}_{1:3}^{1,2}$; then we extract the speech waveform segment $\mathbf{y}_2$ corresponding to token $\mathbf{x}_2$ from the signal generated from $\mathbf{z}_{1:4}^{1,2}$; we concatenate $\mathbf{y}_1$ and $\mathbf{y}_2$, and we continue this process until the end of the input sequence is read. This process is illustrated in Fig. 3.6 for $k = 1$. Segment concatenation is done with a 5-ms cross-fade, a simple and efficient way to prevent audible artefacts in our experiments. Note that the overall encoding and decoding process simulates an effective k-lookahead iTTS system that generates a new speech segment $\mathbf{y}_n$ when entering the new input token $\mathbf{x}_{c(n,k)}$. Sound examples obtained with this procedure are available online.[13]

3.3.4.5 Analyzing the impact of lookahead on encoder representation

Our first goal is to analyze the impact of the lookahead parameter k on the representation of a given token $\mathbf{x}_n$ at the encoder level. Given the two encoder representations of $\mathbf{x}_n$ ($\mathbf{z}_n^{n,k}$ in incremental mode and $\mathbf{z}_n^{\text{full}}$ in offline mode), we compute the cosine distance between them with Equation 3.1.

[13] `https://shorturl.at/emyPV`

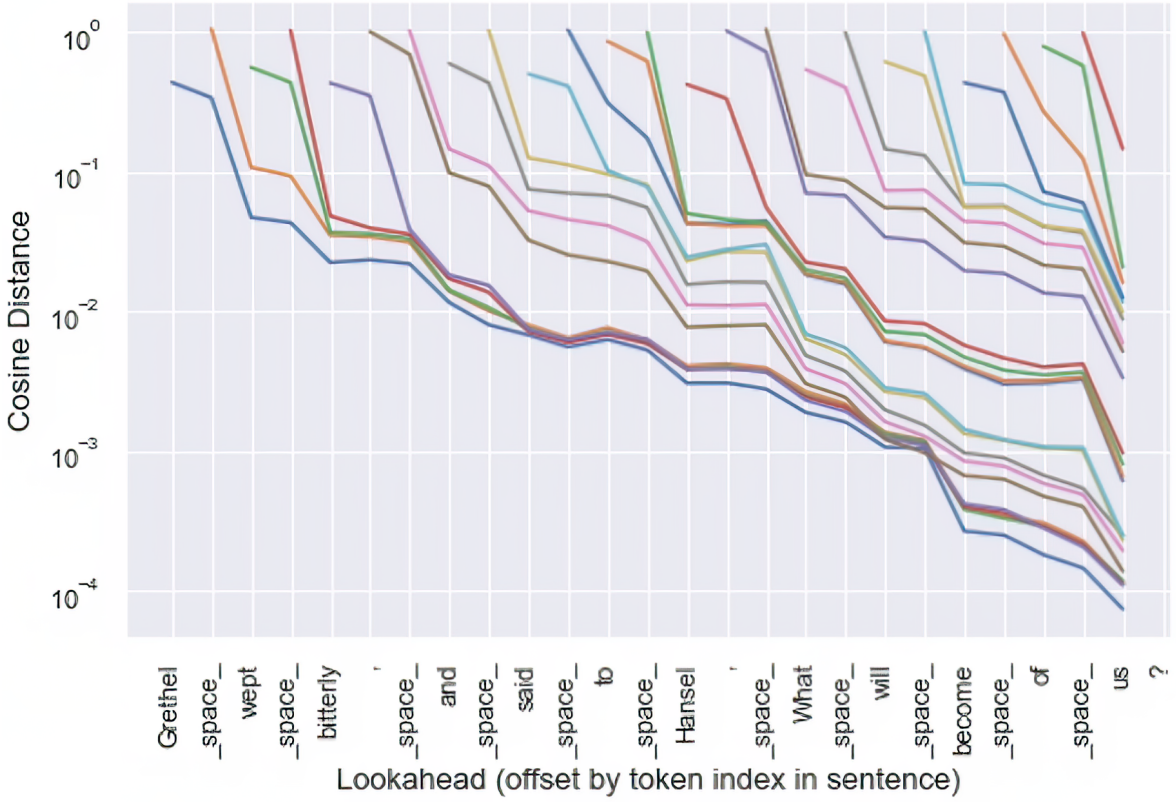


Figure 3.7: Change in token representations over time. Each colored line represents a token from the utterance "Grethel wept bitterly, and said to Hansel, What will become of us?" The height of the colored line shows the distance between encoder outputs $\mathbf{z}_n^{n,k}$ (incremental decoding) and $\mathbf{z}_n^{\text{full}}$ (offline decoding) at different values of k. The vertical grid line where $\mathbf{x}_n$ (the colored line) first appears represents $k = 0$ for that token; the next vertical grid line to the right represents $k = 1$ for $\mathbf{x}_n$ and $k = 0$ for $\mathbf{x}_{n+1}$, etc.

$$d(n,k) = 1 - \frac{\mathbf{z}_n^{n,k}.\mathbf{z}_n^{\text{full}}}{||\mathbf{z}_n^{n,k}||.||\mathbf{z}_n^{\text{full}}||} \tag{3.1}$$

We then average this distance for all tokens of our corpus or all tokens of a given syntactic category.

Our analysis consists in investigating which token features could best explain the observed variance in our data (i.e., why are some tokens relatively far from their final representation while others are close at the same value of k?). We did this using random forest (RF) regressors (Pedregosa et al. 2011) which optimize cosine distance predictions and can provide information about which input features contribute the most towards these predictions.

We investigate a range of different features in our model including: the frequency of the current word in the training corpus, the word length and POS of the current/next/previous word, the relative position of the current word in the full sentence, as well as correlates of

its position within smaller phrases (i.e., distance from the next punctuation mark, distance to the parent phrase end in a constituency tree.[14]) Our selected features and their statistical significance are summarized in Table 3.2.

The RFs were fit using 100 estimators, mean squared error measures and bootstrapping. We followed the following procedure to determine which features are the most significant: (1) we add a column of randomly distributed values to our data set; (2) we fit an initial RF and eliminate all variables with a Gini importance[15] lower than the random feature; and (3) we fit a new RF using only the remaining features and then calculate the permutation feature importance (i.e. the drop in R^2 that results from swapping columns in the dataset) (Altmann et al. 2010).

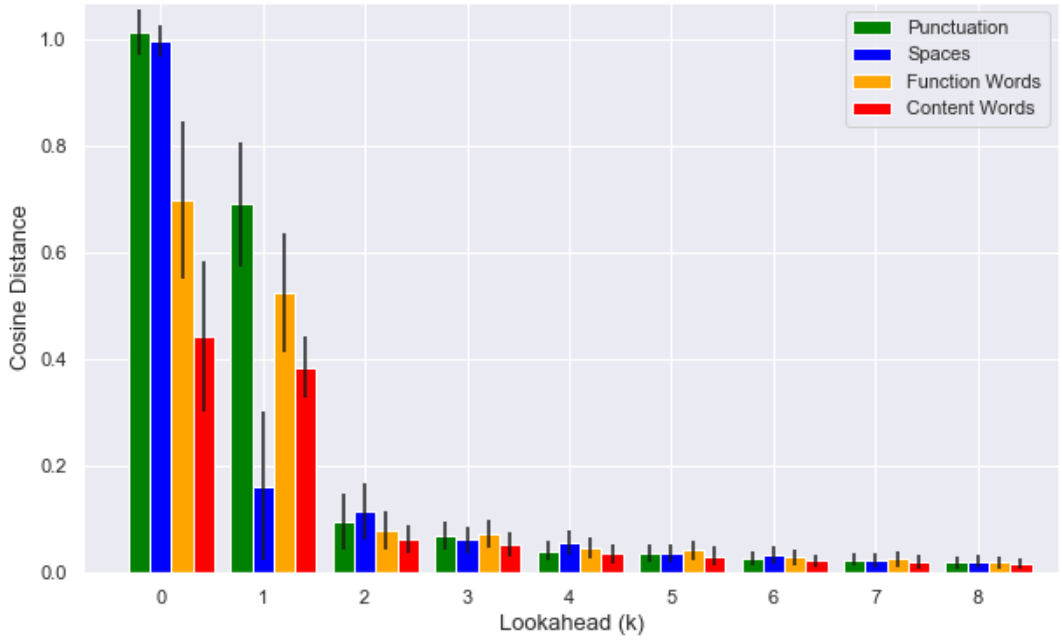


Figure 3.8: Distance $d(n, k)$ between encoder representations $\mathbf{z}_n^{\text{full}}$ (offline) and $\mathbf{z}_n^{n,k}$ (incremental decoding) averaged over all tokens of a given category (punctuation, space, function word, content word), for lookahead parameter $k = 0$ to 8. Error bars represent standard deviation.

3.3.4.6 Analyzing the effect of lookahead on decoder output

We evaluate the perceptual impact of the lookahead k using a MUSHRA listening test (ITU-R 2015). To that purpose, we selected 20 sentences and generated each at multiple values of k, namely $k = 1, 2, 4, 6$. $k = 1$ corresponds to a lookahead of one space (or one punctuation

[14] We use the Benepar constituency parser (Kitaev and Klein 2018) (available at https://spacy.io/).

[15] The Gini importance score (Breiman 2001) is a measure used to evaluate the importance of features in a decision tree. It is calculated based on how much the node impurity (i.e., the degree to which data points belong to different classes at the node) decreases when a feature is used to split the data. This measure has a bias towards features that have many split points (Strobl et al. 2007) and that is why it should not be used alone to judge importance.

mark). It was chosen as the baseline and should be considered as the low-range anchor for the test. In general, $k = 2$ represents a 1-word lookahead, $k = 4$ a 2-word lookahead and $k = 6$ a 3-word lookahead, although other permutations occasionally occur (e.g. a space followed by an open parenthesis). Note that $k = 0$ was not selected because the output signal was deemed too unintelligible to warrant evaluation. The reference stimuli were generated with the offline TTS mode, and were used both as reference and as the hidden high-range anchor. 21 participants, all native English speakers, were asked to assess the similarity between the reference and each of the stimuli obtained with the incremental decoding policy (plus the high-range anchor) on a 0-100 scale (100 means that sample and reference are identical). The MUSHRA test was done online, using the Web Audio Evaluation Tool (Jillings et al. 2016). 3 participants were excluded from analysis because they did not give high similarity ratings for the reference and the hidden high-range anchor (which were identical). Statistical significance between different experimental conditions (different values of k and incremental vs. offline synthesis) were assessed using paired t-tests.

3.3.4.7 Results and discussion

Encoder representations Figure 3.7 displays an example of distance (in log-scale) between encoder representations in incremental *versus* offline modes for a given sentence and all possible values of k. While the global trend is a movement towards the final (offline) representation as k increases ($d(n, k)$ decreases with k), we also observe some cases where an increment of the context leads to a representation that is farther away from its offline counterpart (see for instance the comma after the word "Hansel"). One possible explanation for this might be that Tacotron interprets the input as the end of an intonational or intermediate phrase and when further input is received, it reassesses the token representation.

We further notice that the addition of future context effects groups of words in a similar fashion. For example, the clause *Grethel wept bitterly*, syncs up after the first few timesteps, and every additional word pushes or pulls the group away from or towards its final representation to a more or less equal degree. This may be the result of the model trying to accommodate rhythmic constraints that are affected by the length of the units within the global structure (Zvonik and Cummins 2003; Krivokapic 2010).

Figure 3.8 also shows the distance $d(n, k)$ for $k = 0$ to 8, but this time averaged over all tokens of the $1,000$ test sentences and for the different token categories: punctuation, space, function word and content word. As in Figure 3.7, increasing the lookahead consistently reduces the distance between the encoder outputs in incremental and in offline mode, on average. Importantly, the most significant decrease is observed between $k = 1$ and $k = 2$, that is, when considering a lookahead of one space and one word (in addition to the current word). A slower decrease toward the final representation is observed for $k \geq 2$. This is consistent with Figure 3.7. A series of paired t-tests on $d(n, k)$ (averaged over all test sentences and all token categories) reveals a tiny but systematically significant difference between pairs of consecutive lookaheads (e.g., $k = 3$ vs. $k = 4$, $k = 7$ vs. $k = 8$) up to the end of the sentence. This might show that, on average, each new token considered in future context contributes slightly but

significantly to the evolution of the encoder representation.

We also observe that, while representations of content words are more stable to context variation, those of punctuation, spaces and function words are further away from their final representation in offline mode when not enough context is given ($k < 2$). This makes sense for function words that are usually unstressed and often become cliticized with their surrounding content words in continuous speech, unless they are at the end of a prosodic phrase (Selkirk 2008). Since function words would initially be interpreted as phrase final, but then be reinterpreted with future context, we would expect to see a large shift in the representation. Similarly for space characters, which sometimes correspond to pauses in the speech stream, but are more often empty symbols that represent coarticulatory features between two consecutive words (exploration of the vector space shows that space vectors preceded and followed by the same phoneme pairings form clusters). The lack of stability in punctuation marks is slightly more surprising since they typically mark the end of a clause or a sentence. Clause endings in human speech possess phrase final features (phrase accents and boundary tones) that indicate how one clause/sentence relates to the next and are thus highly dependent on future context. However synthetic speech has not mastered this skill; it nonetheless relies heavily on future context to shape its representation.

A more fine grain analysis of the factors that impact $d(n, k)$ is provided by the results of the RF analysis, which are summarized in Table 3.2. For $k = 0$, the length of $\mathbf{x}_n$ is the most effective predictor of cosine distance, and for $k = 2$ the lengths of $\mathbf{x}_{n+1}$ and $\mathbf{x}_{n+2}$ (i.e. the future tokens that the encoder sees when encoding $\mathbf{x}_n$) are the most effective predictors. For instance, at $k = 2$, our model correctly predicts that the token "to" in Sentence A below (lookahead = *space* + "be") is farther away from its final representation than "to" in Sentence B (lookahead = *space* + "Kitty"). The cosine distances are 0.135 and 0.057 respectively. An alternative possibility for this difference, that "to" in Sentence A is farther from its parent constituency phrase end than Sentence B, is not considered significant by our RF model.

A) *I suppose, he said, I ought* ***to*** *be glad of that.*

B) *And the Captain of course concluded (after having been introduced* ***to*** *Kitty) that Mrs Norman was a widow.*

Perceptual evaluation of the decoder output Results of the MUSHRA listening test are presented in Figure 3.9. First, statistical analyses show significant differences for all pairs of considered lookahead ($k = 1$ vs. $k = 2$, $k = 2$ vs. $k = 4$, and $k = 4$ vs. $k = 6$). This confirms at a perceptual level the tendency observed on the evolution of the encoder representation (see Section 3.3.4.7): each additional lookahead brings the incremental synthesis closer to the offline one. The degree of change to the decoder outputs is however much larger than the corresponding change in the encoder outputs, which suggests there may be other factors than context influencing the audio quality. We also found a significant difference between $k = 6$ and $k = N$ (offline mode), i.e. with a lookahead of typically 3 words. This is in contradiction

Table 3.2: Influence of text features on the distance $d(n, k)$ estimated by RF regression for $k = 0$ and 2 (NS=not significant; * = weak effect; ** = medium effect; *** = strong effect).

		Permutation Feature Importance	
Feature	**Definition**	$k = 0$	$k = 2$
Token Length	# of characters in x_n	***	**
POS	Part of speech of x_n	NS	NS
Frequency in Training	# of instances of x_n in LJ Speechset	*	NS
Relative Position	Token's relative position in input sequence = n/N	*	*
Penultimate	Does $n = N - 1$?	*	NS
Followed by Punctuation	Is x_{n+1} a punctuation mark?	NS	NS
Distance to Punctuation	# of tokens before next punctuation mark	*	*
Distance to Parent Phrase End	# of tokens to the end of parent constituent group of x_n	NS	NS
POSPrev + m	Part of speech of token x_{n-m}	NS	NS
POSNext + m	Part of speech of token x_{n+m}	NS	m=1 *
Word Length of Prev + m	# of characters in x_{n-m}	$m = 1$ * $m = 2$ *	NS
Word Length of Next + m	# of characters in x_{n+m}	$m = 2$ *	$m = 1$: *** $m = 2$: *** $m = 4$: *

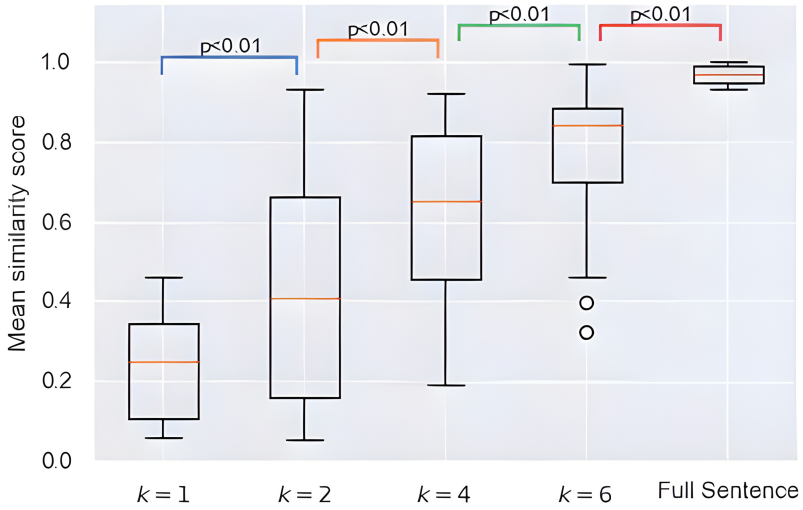


Figure 3.9: Perceptual evaluation of the impact of lookahead parameter k using MUSHRA listening test.

with Yanagita et al. 2019 who did not report any difference between incremental and offline synthesis for such lookahead. Possible explanations for this include (1) the use of a different evaluation paradigm (MUSHRA vs. MOS in Yanagita et al. 2019), (2) duration distortions caused by the concatenation of speech segments or (3) by the fact that contrary to Yanagita et al. 2019, we did not retrain Tacotron 2 on shorter linguistic units. We observed increased instability of the attention mechanism when it was presented with truncated inputs, and so retraining on smaller units is probably advisable for such models (in the following chapters, we use a Transformer-based TTS model which does not have the same issue).

3.4 Conclusion

This study presents several experiments which probe the impact of future context in a neural TTS system, based on a sequence-to-sequence model, both in terms of encoder representation and perceptual effect. Reported experimental results allow us to draw the contours of an adaptive decoding policy for an incremental neural TTS, which modulates the lookahead (and thus the overall latency) by potential change in the internal representations. Shorter words are more dependent on future context than longer ones. Therefore, in a practical iTTS, if the lookahead buffer is fed a short word, it may be preferable to delay its synthesis because the internal representation associated with it is likely to change when additional tokens become available. Also, it may be more useful to define the lookahead parameter in terms of future syllables rather than words.

In addition, perceptual evaluation shows that the dynamics between encoder and decoder are such that even if the encoder representation of an individual token changes slightly, the length of the encoder representation sequence will influence the way in which the decoder treats that token. By examining the attention weights the decoder uses when making predictions,

we see that focus is placed on the current character's immediate surroundings (one or two characters ahead and behind), but focus is also placed on the end of the encoded sequence (the Tacotron 2 model we used in this experiment was not trained with a guided attention loss (See Section 3.1.2) and thus was not encouraged to learn diagonal attention weights). We can conjecture that the decoder is regulating the duration of each segment with respect to sequence length. The decoding phase would most likely be more stable if trained with a guided attention loss. Now that the importance of future context has been assessed, in the next chapter we will move on to context extension through prediction of future tokens using contextualized language models.

Chapter 4

Predicting future text

Contents

In this chapter, we present the work from our Interspeech 2021 paper *Alternate Endings: Improving prosody for Incremental Neural TTS with predicted future text input* (Stephenson et al. 2021).

Inferring the prosody of a word in text-to-speech synthesis requires information about its surrounding context. In incremental text-to-speech synthesis, where the synthesizer produces an output before it has access to the complete input, the full context is often unknown, which can result in a loss of naturalness. In this work, we investigate whether the use of predicted future text from a Transformer language model can attenuate this loss in a neural iTTS system. We compare several test conditions of next future word: (a) unknown (zero-word), (b) language model predicted, (c) randomly predicted and (d) ground-truth. We measure the

prosodic features (pitch, energy and duration) and find that predicted text provides significant improvements over a zero-word lookahead, but only slight gains over random-word lookahead. We confirm these results with a perceptive test.

Our analyses show that the additional syntactic context provided by the LM-generated text does not improve prosody predictions. To further investigate the syntactic sensitivity of the TTS model used for this experiment (Fastspeech 2), we analyze the prosody feature predictions for (1) garden-path sentences and (2) a single prefix with multiple LM-generated sentence completions controlled for syntactic construction. The garden-path sentences, in line with the results from the previous chapter, show that standard TTS models make shallow use of context beyond the very local environment. The multiple sentence completions show prosodic distinctions based on frame building function words (e.g., *who*, *and*).

4.1 Introduction

In incremental text-to-speech synthesis (iTTS), the system starts to output chunks of synthetic audio before the full text input is known (Baumann and Schlangen 2012a; Baumann 2014c; Pouget et al. 2015; Pouget et al. 2016). The missing input information often hinders the ability to produce a natural sounding speech sequence, mostly because prosodic features that will be determined by the future context (i.e., the remaining words in the sentence) have not yet been specified. Fortunately, the future input is not completely random; human language is characterized by several lexical and syntactic patterns, which can be statistically learnt and then predicted to a certain extent. Recent advances in language modelling, namely the use of transformer models such as BERT (Devlin et al. 2019) and GPT-2 (Radford et al. 2019) (and more recently GPT-3 and 4 (Brown et al. 2020; OpenAI 2023)) give us accurate representations of the probability distribution of future words. If this information can be mobilized to fill in the missing data for an iTTS system, it may be possible to retain naturalness while minimizing latency.

4.2 Related work

4.2.1 Human prediction

Language is a balance between the predictable and the unpredictable; we need some structure/predictability to facilitate the transfer of information, but if everything is predictable, there is no point in actually speaking, because our interlocutor can already anticipate what we are going to say. Anticipating upcoming semantic and syntactic content is a natural part of human language processing, although the degree to which exact words can be predicted is highly dependent on the context (Figure 4.1); the level of predictability fluctuates from word to word, with varying degrees of constraints coming from structural, semantic, collocational, and topical factors as well as the broader situational context. Levels of predictability have

been shown to affect the duration of spoken words (e.g., Jurafsky et al. 2008), reading times (e.g., Roland et al. 2012) and the types of contributions offered by conversation partners when the other partner hesitates (Howes et al. 2012).

A standard measure for predictability is the cloze score (Taylor 1953). This is similar to the language modelling task used for GPT-2. Participants have access to the past words in a passage and they must try to reconstruct the next word using contextual clues. The percentage of responses that match the original word show the word's predictability. Luke and Christianson 2016 conducted a large scale cloze test and found that only a small percentage of words have a high cloze probability: 5% of content words and 19% of function words. Many of the words in this evaluation had a more probable competitor, but reading time tests showed no processing penalty for the incorrect guesses. These results provide evidence for graded prediction in human language processing, where many words are activated based on their contextual probability. Contrary to specific lexical predictability, semantic and morphosyntactic features were highly predictable. Human predicted words matched the POS of the actual word 70% of the time; predicted nouns and verbs shared the same morphological characteristics (number and tense) in over 72% of cases.

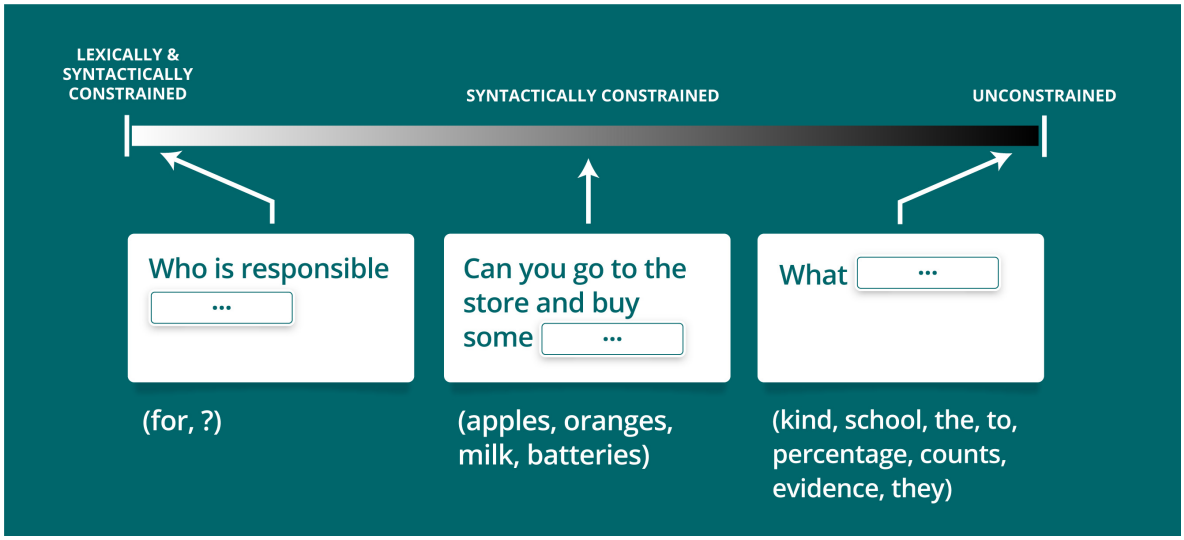


Figure 4.1: Constraints affecting the predictability of future words.

4.2.2 Predictive text

Predictive text, which is related to our proposed method, is commonly used in AAC and text-messaging applications. These applications offer word options for the user to select from when entering their message. The options are based on word frequency, language model word probabilities, or context-based vocabulary builders. Efforts to improve word prediction have included training or fine-tuning LMs on an individual user's data (Lee et al. 2017), on the current topic of conversation (Trnka et al. 2006) and text style (Li et al. 2020). The incorporation of the conversational partner's speech through ASR has also been tested (Adhikary

et al. 2019; Wisenburn and Higginbotham 2008), as has the use of text external context such as location (Epp et al. 2012; Kane et al. 2012) and objects in the immediate environment (Kane and Morris 2017). Adaptation to specific task types can further improve prediction as measured by number of keystrokes required from the user (Higginbotham et al. 2009).

The work presented in this chapter differs from predictive text in two ways: (1) Our system does not require the user to select from potential words. Rather it runs in the background, sampling future words that will not be part of the audio output, but simply used as additional input to an iTTS system. (2) We process input at the word level, whereas predictive text systems continually update or filter their predictions with each additional character (e.g., Schadle 2004). Permitting slightly more latency so an iTTS user can enter the first character of the next word could have dramatic effects on prediction accuracy, however we leave this for future work.

4.2.3 Pseudo-lookahead for iTTS and other neural model applications

When part of the input for a statistical model is missing, as is the case for iTTS, accommodations must be made to replace or compensate for the unknown. This can be done with either concrete replacement values obtained from imputation techniques or with descriptive statistics (e.g., the mean) which can help to infer average behaviour.

In the HMM-based paradigm of TTS, explicit features derived from the linguistic content were fed to the models in order to predict the acoustic signal. An intermediate step in this design included decision trees that would learn to cluster different contexts that were used to predict prosodic features. Baumann 2014b studied the effects of missing future features by replacing decision tree split criteria with default values and evaluating the degradation in speech quality: while cepstral and aperiodicity features could be estimated fairly accurately with just a local context, prosodic features (f0 and duration) were found to be more dependent on longer range context. Pouget et al. 2015 explicitly specified unknown features in the context clustering process and found improvements over a default value strategy.

Recent research in iTTS (contemporary and subsequent works to the research presented in this chapter) has focused on end-to-end neural models. While these models create more natural speech, they are also more difficult to analyze because the relevant features for the task are learnt during training and are subsequently not easily human interpretable. In this framework, both predicted future text and more abstract future-context representations have been tested.

Saeki et al. 2021a used a language model to predict the next five-word sequence for an iTTS system (pseudo-lookahead) and then used a context encoder to learn a more abstract representation of that future. The context encoder was first trained on ground-truth future and past textual contexts. Then the context encoder was fine-tuned to bring the ground-truth and pseudo-lookahead embeddings closer together. In a follow-up study (Saeki et al. 2021b), they trained a language model to predict the context embeddings directly, without a separate context encoder, to further reduce latency. Saeki et al.'s work, while similar in concept to our

proposed method, differs in the amount of future context predicted and the level of abstraction of the future context. We will address these differences throughout this chapter.

As part of a simultaneous speech-to-speech translation system, Liu et al. 2022 mobilized the translation unit to generate future text for the TTS unit; the translation decoder would predict the translation up to the current input state and then simply predict one additional word. Since the translation was conditioned on the source language audio, it had more contextual information available compared to a standard language model. This increased prediction accuracy considerably.

Other works in natural language processing have also experimented with predicted future text to adapt to an incremental setting. Zheng et al. 2019 hallucinated future words to balance the latency/quality trade-off in simultaneous translation and Madureira and Schlangen 2020 tested "prophecies" to improve the performance of a full-sentence language model (BERT) on incremental sequence tagging and classification tasks.

4.3 Proposed Method

Figure 4.2: Utilizing language model predictions to improve incremental TTS quality while keeping limited latency.

In the present work, we propose an iTTS system that incorporates a language model to predict future lookahead. Our approach (described in Figure 4.2) takes all available context from the input stream and predicts one word into the future. The current context and pseudo-future are then passed to a TTS model that synthesizes the audio and all but the current word is discarded. The current word is vocalized and this procedure repeats for each successive word. For example, if the first word entered by a user is *What*, we use *What* as a prompt for the language model and sample a possible next word (*are*). We then synthesize *What are*, but only vocalize the word *What*.

We chose a limited lookahead so that the effects of correct and incorrect predictions could be studied. We assume that a large number of sampled words will not match the ground-truth future text and so we want to evaluate how LM-generated text, that is likely to share syntactic traits with the ground-truth, compares with randomly generated future words. A larger lookahead complicates matters because the combinatorics that increase with each additional word (e.g., the n^{th} future word/POS may match, but $n+1$ word may not or vice versa) must be accounted for, but the distribution of these conditions can be heavily skewed. Furthermore, the results from the last chapter show only a limited amount of lookahead was actually responsible

for bringing word representations the majority of the way to their final representation in a TTS model.

We evaluate our system by contrasting different future word contexts: (a) unknown, (b) language model predicted, (c) randomly predicted (a control group) and (d) ground-truth (see Table 4.1 for examples). Differences are measured at the TTS encoder level and from the generated speech signal through a listening test.

Table 4.1: Examples of input sequences with unknown, ground-truth, predicted and random future context. In each sequence, the word in bold is the word which is synthesized from the sequence.

Input Type	**Lookahead**	**Input Sequences**
Ground-truth	Full sentence, $k = N - n$	**Do you think that you could manage, Tidy?**
Unknown (future)	$k = 0$ word	$\mathbf{s}^{\text{GT}}_{1:n+0}$ = **Do**, Do **you**, Do you **think**, ...
Ground-truth	$k = 1$ word	$\mathbf{s}^{\text{GT}}_{1:n+1}$ = **Do** you, Do **you** think, Do you **think** that, ...
GPT-2 prediction	$k = 1$ word	$\mathbf{s}^{\text{Pred}}_{1:n+1}$ = **Do** they, Do **you** agree, Do you **think** this, ...
Random	$k = 1$ word	$\mathbf{s}^{\text{Rand}}_{1:n+1}$ = **Do** dance, Do **you** until, Do you **think** art, ...

4.3.1 Language model feature prediction and sampling techniques

By sampling pseudo-future text, our aim is to predict textual features that will improve prosody modelling. Future features that proved helpful before end-to-end models took over TTS included the POS, the number of syllables and the stress pattern of the next word, as well as its accent and boundary status.

Ideally, to match all of these features, we would like to predict the exact next word. However, we have seen from human cloze tests that the number of highly constrained contexts where the exact next word can be predicted is fairly limited. And comparing LM's exact next word prediction capabilities to humans', we can expect even worse performance, since LMs have the disadvantages of restricted world/common-sense knowledge and imperfect syntactic and semantic representations (see Section 2.2.3). Moreover, their predictions are conditioned on a limited context (one/a few sentence(s)) which further hinders performance. In their pseudo-lookahead experiment, Madureira and Schlangen 2020 compared ground-truth sentences with GPT-2 predicted continuations until the end of the sentence using BLEU scores (Papineni et al. 2002) and found the results were very low (0.004).

Indeed, correctly projecting the exact future text is quite unlikely. If we take an example sentence from our test corpus *Besides, he's not the sort of person to complain* (Figure 4.3), and we use GPT-2 to predict the next word at each point in the sentence we see that only two of the words are correctly predicted when sampling the most probable next word (*not* and *of*). Figure 4.3 shows the top three most likely next words for each point in the sentence and their corresponding probabilities according to the language model. At some points, GPT-2 is fairly

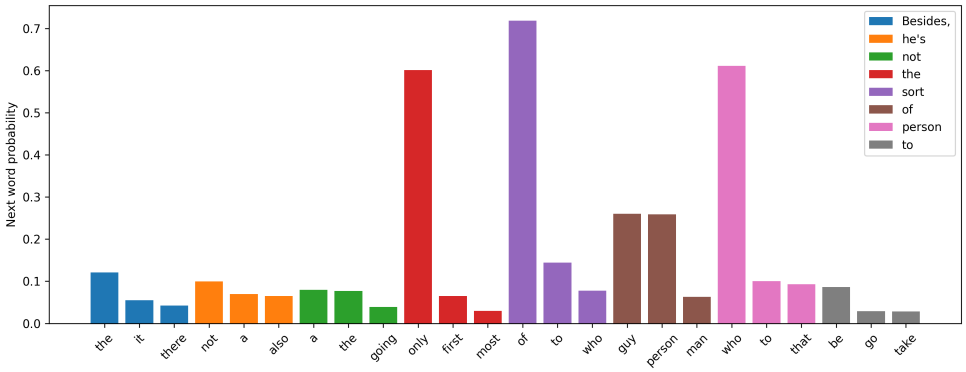


Figure 4.3: Probability of future words for the sentence *Besides, he's not the sort of person to complain.* The words in the legend show the point in the sentence when the prediction is made and the corresponding bars show the probability of the top three most likely next words given by GPT-2.

confident about its predictions (at *the*, *sort* and *person*), but only one of these predictions ends up matching our sentence.

Nonetheless, the predicted words are plausible continuations and many match the POS of the words in our sentence. If words belonging to the same syntactic category can be predicted, this could have potential benefits for the prediction of natural prosodic contours. Previous studies in TTS have demonstrated that including syntactic information can be useful for predicting prosodic feature assignment (Hirschberg and Rambow 2001; Fitzpatrick and Bachenko 1989; Liu et al. 2021). Some of these works entail deep syntactic parsing, however even surface features like POS tags can assist with both structural and prominence prediction. Sanders and Taylor 1995 found a trigram POS model could successfully predict phrase boundaries, and even yielded better results than some more complicated models which considered additional features. And Hirschberg and Litman 1993 achieved high pitch accent prediction accuracy (77%) by basing predictions on fine-grained POS categories.

The quality/naturalness of the generated words will depend on the sampling technique used to select them. Our limited lookahead reduces the impact of the sampling strategy, however we briefly outline these options below to discuss their potential influence on our experiment and related works.

1. Greedy Search: The most probable next word is selected at each timestep.

2. Beam Search: The top n most probably next words are selected, and branches for each of these words are created with the top m words to follow them. The most likely next sequence is chosen based on the combined probability of the words in each branch.

3. Top-k sampling: The next word is sampled from the distribution of the top k most probable next words (Fan et al. 2018) (not to be confused with the variable k used to

denote lookahead in iTTS). This eliminates low-probability words from being selected, but an appropriate number k is difficult to set for continuous generation, since it could include inappropriate words for highly constrained conditions and exclude appropriate words for less constrained conditions.

4. Top-p/Nucleus sampling (Holtzman et al. 2020): Sampling is performed over the words in the top-p set, where p is cumulative probability. This allows for more adaptable sampling (overcoming the rigid parameters of top-k sampling).

5. Locally typical sampling (Meister et al. 2023): This sampling method models the expected information at a given point in the generation. It was developed to emulate the way humans communicate, which is not to always say what is most likely since words that are less likely carry greater information content. Locally typical sampling attempts to keep the rate of information transfer constant by conditioning the generation of the current word on the information value of the previous words in the sequence.

Using greedy sampling is the most straightforward method for getting the most likely next words, however if one is sampling repeatedly into the future, this can result in a dull, repetitive text that does not match the normal sequence distribution in terms of word lengths, because more common/probable words are often shorter. Locally typical sampling is a proposal that post dates our work from this chapter, however it has the potential to provide better phrase length approximations because it encourages less likely/longer words to be sampled at regular intervals (this remains to be verified). In this work, we do notice a short word bias in the predicted text, and we control for this in our random sampling method (details below).

Saeki et al. 2021a used top-k sampling to project five words into the future and they evaluated different values of k for their similarity to the ground-truth (as measured by the cosine distance between vectors from a context encoder). They found that $k = 1$ (i.e, greedy search) gave the best results. In this work, since we are interested in teasing apart the effects of language model predicted text from random future text, we sample multiple future words for the same $k = 1$ position (here k refers to lookahead). This allows us to compare the prosodic feature predictions from multiple future contexts. We are only projecting one word into the future, so we do not employ beam search.

4.4 Method

4.4.1 Definitions

For each token in our corpus, we prepare different sequences which are used as input to the TTS model, FastSpeech 2 (Ren et al. 2021).

- $\mathbf{x}_{1:n} = x_1, x_2, ..., x_n$ is the sequence of tokens up to n. In the proposed iTTS system, the tokenization policy is to split the sentence on space characters, and then synthesis is triggered when a space character is encountered.

- k is the lookahead parameter (number of future tokens available when synthesizing token x_n).
- $\mathbf{s}_{1:n+k} = \{x_1, x_2, ..., x_n, \hat{x}_{n+1}, ..., \hat{x}_{n+k}\} = \{\mathbf{x}_{1:n}, \hat{\mathbf{x}}_{n+1:n+k}\}$ is the sequence used for the synthesis of token x_n, where for the ground-truth condition (*GT*) $\hat{\mathbf{x}}_{n+1:n+k} = \mathbf{x}_{n+1:n+k}$, for the prediction condition (*Pred*) $\hat{\mathbf{x}}_{n+1:n+k}$ is given by the language model, and for the random condition (*Rand*) $\hat{\mathbf{x}}_{n+1:n+k}$ is random. The random token generation is described in Section 4.5.1.
- $\mathbf{s}_{1:n}$ is the input prompt used to generate language model predictions.
- Near the end of the sequence, we replace $n + k$ with $\min(n + k, N)$ where N is the length of the full utterance.

4.4.2 Models

4.4.2.1 Language model used for prediction

We use the GPT-2 language model for our study. This is an auto-regressive model trained to predict the next word given a sequence of past words (causal language modeling task), based on a Transformer architecture. The auto-regressive architecture is well suited to incremental prediction. The original GPT-2 (Radford et al. 2019) is large (1.5B parameters) and since our intended use requires fast predictions, we opted to use a smaller version of GPT-2, called "distilled GPT-2" (Wolf et al. 2020).[1] This model has been trained to produce the same output probability distribution as the original GPT-2 but using fewer layers/parameters. It is a six-layered model that uses twelve attention heads and a hidden layer size of 768 dimensions.

4.4.2.2 TTS model

For TTS we select a fast and high-quality end-to-end model: FastSpeech 2. The implementation we use (Hayashi et al. 2020),[2] trained on the LJ Speech Dataset (Ito 2017), takes characters as input and converts them to phonemes. Phoneme embeddings are passed through several self-attention layers before the model makes duration, pitch and energy predictions for each phoneme. These feature predictions and the latent phoneme representations are then passed to the decoder (more self-attention layers) which produces a Mel-spectrogram.[3] The Mel-spectrogram is then input into a Parallel WaveGAN vocoder (Yamamoto et al. 2020) (trained on full-sentence inputs) for waveform generation. This model is well suited to iTTS because (1) it is fast which is desirable when the objective is to reduce latency (the speed is achieved by predicting all Mel-spectrogram frames in parallel), and (2) it makes explicit duration predictions for each phoneme, which makes it possible to segment words and only synthesize the word(s) of interest.

[1] https://huggingface.co/distilgpt2
[2] https://github.com/espnet/espnet
[3] For implementation details, see https://tinyurl.com/s7p38hcr

4.4.3 Incremental synthesis (iTTS)

We implement an incremental synthesis procedure where each token x_n is synthesized from the input sequence $\mathbf{s}_{1:n+k}$. Mel-spectrogram frames corresponding to individual tokens are identified using the internal duration predictions made by FastSpeech 2. Successive word-level Mel-spectrograms are input into the Parallel WaveGAN vocoder on a word-by-word basis. Resulting waveforms are concatenated together using a 1-ms crossfade to eliminate glitches (synthetic audio samples are available at `https://bstephen99.github.io/iTTS/interspeech2021/interspeech2021.html`).

4.5 Experiments

4.5.1 Corpus and predictions

The English corpus we use for analysis consists of 1,000 sentences from LibriTTS (Zen et al. 2019). Sentence length ranges from 5 to 42 words, with a total of 16,965 tokens and 62,556 phonemes.

For each token x_n in the corpus, we sampled five GPT-2 and five random next word predictions ($\hat{x}_{n+1}$). The GPT-2 predictions are constrained to the 30 most likely next words (top-30 sampling strategy). The random words were selected from a list of 1,266 of the most common words in English (Speer et al. 2018). Importantly, we force GPT-2 predictions and random predictions to have comparable lengths in term of characters/phonemes because (1) GPT-2 tends to predict shorter words because they are more frequent, (2) in our previous study (Stephenson et al. 2020), we found that longer future words have more influence on the current token's internal representation (in a seq2seq model) than shorter ones, (3) otherwise, our results may be biased by the fact that the random condition simply has more future context. To control for word length in the random condition, we (1) took the word-length distribution of GPT-2 predictions, (2) randomly sampled a word-length category from this distribution (e.g., 2-4 characters), (3) limited our most-common list to only words in this category and (4) randomly sampled a word from this list using a uniform distribution.

GPT-2 uses byte pair encoding (BPE) which breaks words down into subword units to better handle out-of-vocabulary tokens. As such, some of its predictions extend the final prompt word rather than predicting a new token (e.g., previous → previously). To avoid such distortions to our input text, we sample until the first character in the predicted text is a space. This also prevents erroneous punctuation marks from being predicted.

We compare the exact word and word category (POS) prediction rates between the *Pred* and *Rand* conditions and the *Ref* sentences. As expected, the *Pred* exact-word matches are quite low (6.8%), however they are a lot more frequent than the Rand condition (0.09%). To obtain POS tags, we use the Spacy POS tagger.[4] This is not an incremental tagger and so the

[4]https://spacy.io/

tagging error rate will be higher than under full-sentence conditions. However comparisons are made with incrementally tagged *Ref* sentences, so the errors should often be aligned. For example, the word *that* in sentence final position is tagged as a pronoun, even though with more context it could be revealed to be a determiner (e.g., *that cat*); if the *Pred* and *Ref* $k = 1$ words are both *that*, this will be counted as a match, even though the tag is incorrect. The confusion matrices for *Pred* and *Rand* are shown in Figure 4.4. The overall POS accuracy for the *Pred* condition is 43.5% and 18.0% for *Rand*.

4.5.2 Metrics

4.5.2.1 FastSpeech 2 representations

We aim at evaluating the prosody obtained in the different test conditions: no context ($k = 0$), ground-truth context (GT), predicted context (*Pred*), random context (*Rand*). For this aim, we compare the pitch, duration and energy values produced in those conditions with the values produced in the *reference* condition (*Ref*) where the full context (full sentence input) is used. In the present study, we concentrate on the case $k = 1$ (one-word lookahead).

As for duration and energy, they are first computed at the phoneme level, using the FastSpeech 2 internal predictions (see Figure 4.5 for a plot of duration values from an example sentence). A phoneme duration is defined as (the log of) the number of Mel-spectrogram frames of that phoneme. The energy is the squared magnitude of the short-time Fourier transform (STFT), averaged across all frequency bins and across the duration of the phoneme. Then the mean absolute error (MAE) is computed by averaging the absolute value of the difference of duration values obtained in each test condition and in the reference condition across all phonemes of the dataset, and the same for the energy feature. The results are reported in Table 4.2.

Pitch is evaluated at the sentence level.[5] We first align the Mel-spectrograms obtained in the test and reference conditions with Dynamic Time Warping using the Librosa library (McFee et al. 2015). Then we extract the pitch curves from the concatenated audio (see Section 4.4.3) using Praat/Parselmouth (Boersma and Weenink 2018; Jadoul et al. 2018) and we compute the MAE in cents between the aligned f_0 trajectories:

$$MAE = \frac{1200}{T} \sum_{t=1}^{T} \left| \log_2 \left(\frac{f_0^{Test}(t)}{f_0^{Ref}(t)} \right) \right|. \tag{4.1}$$

Then the sentence-level MAEs are averaged across all sentences of the dataset. The results are reported in Table 4.3.

[5] We did not evaluate error in the internal FastSpeech 2 pitch predictions because we observed a few extreme prediction values which did not materialize in the resultant audio. We do however use these values in our follow up analyses after verifying the absence of outlier values.

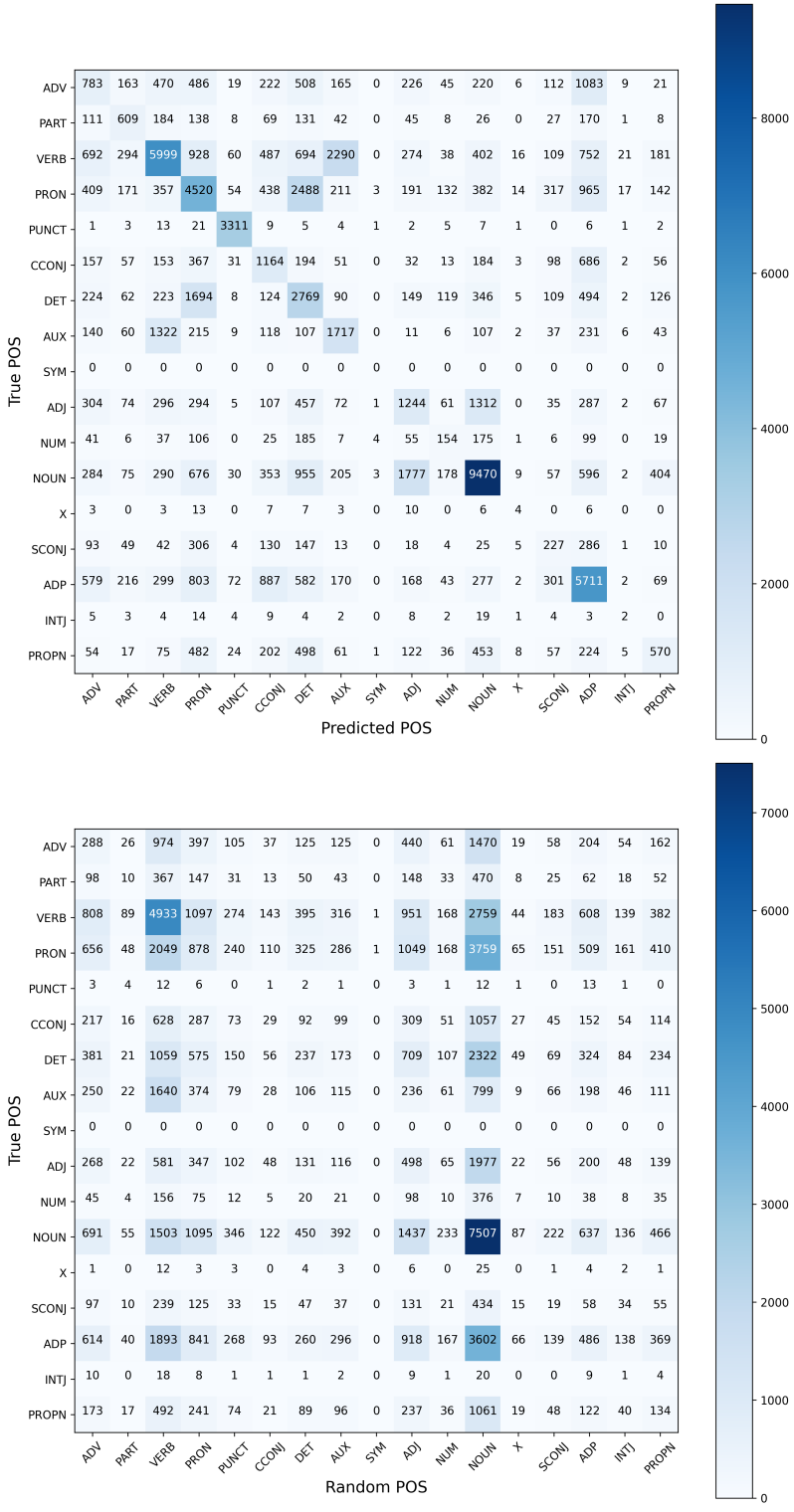


Figure 4.4: Confusion matrices for predicted POS categories. The upper matrix shows the GPT-2 predicted categories and the lower matrix the randomly selected future words.

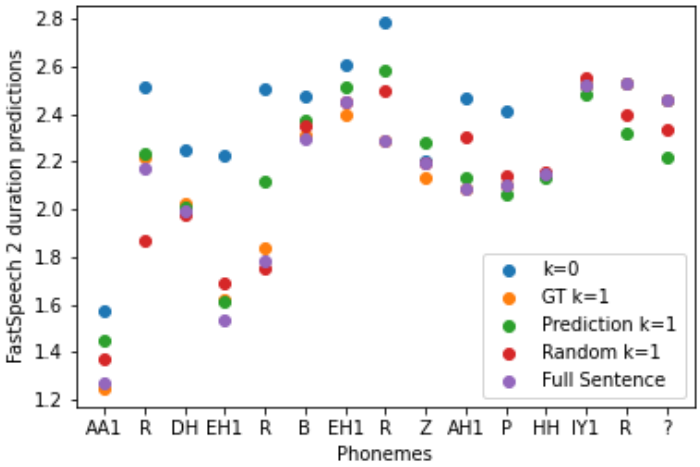


Figure 4.5: Duration predictions from FastSpeech 2 (in number of Mel-spectrogram frames on a log scale) for each phoneme in the sentence "Are there bears up here?" and for the different tested prediction conditions.

Table 4.2: MAE (and standard deviation across phonemes) between duration (resp. energy) obtained with full context and with limited context. *unit = number of Mel-spectrogram frames on a log scale; **arbitrary unit: signal is digital, normalized and averaged.

Input type	**# phonemes**	**Duration***	**Energy****
$k = 0$	62,556	0.262 ± 0.297	0.301 ± 0.364
GT $k = 1$	62,556	0.077 ± 0.133	0.176 ± 0.241
Pred $k = 1$	5×62,556	0.135 ± 0.198	0.247 ± 0.296
Rand $k = 1$	5×62,556	0.147 ± 0.208	0.260 ± 0.304
Correct pred.	38,274	0.086 ± 0.132	0.187 ± 0.239
Incorrect pred.	274,506	0.142 ± 0.205	0.255 ± 0.301

Table 4.3: MAE between the pitch curves obtained with the full context and with limited context.

Input type	**# sentences**	**Pitch MAE (Cents)**
$k = 0$	1,000	203.56 ± 45.50
GT $k = 1$	1,000	88.57 ± 26.33
Pred $k = 1$	5×1,000	120.03 ± 29.34
Rand $k = 1$	5×1,000	123.03 ± 30.27

4.5.2.2 Perceptive test

Finally, we evaluate the global quality using 40 native English speaking evaluators[6] and a MUSHRA test (ITU-R 2015). We selected 20 sentences from our corpus and for each sentence, we presented the listeners with a reference audio clip (generated with the full-sentence context) and then asked them to assign a similarity score to five test clips: the hidden reference (identical to the reference and used as the MUSHRA high anchor), $k = 0$ (used as the low anchor), Ground-Truth $k = 1$, GPT-2 prediction $k = 1$ and random prediction $k = 1$. We then compare the distributions of the similarity scores. The responses from four of the participants were removed because these listeners consistently failed to assign a high similarity score to the high anchor. See Figure 4.7 for results.

4.6 Results and discussion

For all metrics, with regards to the mean, we see a clear ranking in the similarity to the full-sentence reference: $k = 0$ is farthest away, GT $k = 1$ is the closest and *Pred* and *Rand* are in between, the former being slightly closer to full context than the latter. Statistical tests (t-test for pitch, duration and energy measures and Wilcoxon for the listening test) confirmed that *Pred* and *Rand* do not belong to the same distribution (p-value < 0.05) and that *Pred* is better by a small but significant margin.

We notice that duration predictions for $k = 0$ are almost always longer than the other conditions (Figure 4.5). And as in Baumann and Schlangen 2012a, we observe pitch drops for $k = 0$ words. This is because all words are interpreted as the end of a sentence (as they are the final word in the FastSpeech 2 input, hence sentence final characteristics are predicted by the model). Both the prediction and the random conditions reduce this effect thanks to the additional padding words.

4.6.1 Correct vs. incorrect predictions

When we separate the correct from the incorrect GPT-2 next word predictions (see Table 4.2), we see that the MAE for the incorrect predictions is almost identical to the MAE for the random condition. This suggests that the improved syntactical accuracy gained from the GPT-2 predictions (recall the POS of the predicted token matches that of the GT next token 43.5% of the time vs. 18.0% for random)[7] does not translate into improved prosodic features.

Since we only see improvement when the exact next word is predicted, it is clear that the minor difference between GPT-2 and random is explainable by the low exact-word prediction

[6] Anonymous participants were recruited using Prolific (www.prolific.co). They were compensated at a rate slightly above the UK minimum wage.

[7] The accuracy/meaningfulness of the POS tags for the random sequences is questionable due to their often nonsensical nature. However, it is highly probable that the GTP-2 syntax is better than chance.

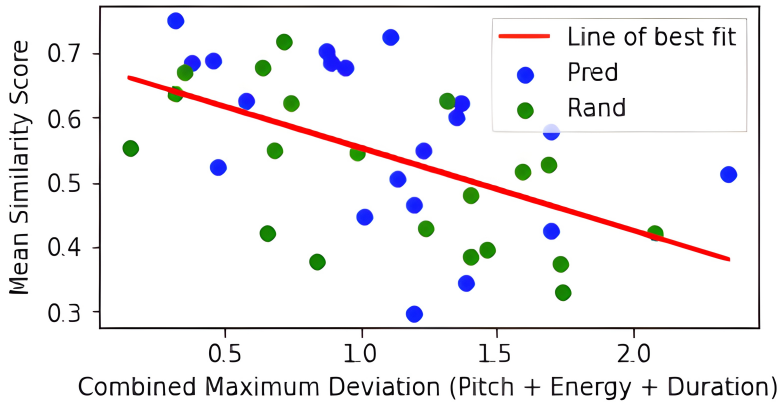


Figure 4.6: Each point represents a sentence (synthesized under the *Pred* or *Rand* condition) from the MUSHRA test. The x-axis shows the scaled and combined (pitch, energy, duration) maximum deviation values (deviation from the full-context value) for the phonemes in the sentence. The y-axis shows the mean similarity score for (*Pred*, *Rand*) sentences to their full context counterpart, given by the MUSHRA participants. The Pearson correlation coefficient is equal to -0.53.

rate. We observe that 76% of the GPT-2 sequences have a prediction rate lower than 10%, and 97% have a rate lower than 21%. It is likely that as language models continue to improve (Brown et al. 2020), we will see greater gains in naturalness from the proposed method (improvements in semantic modelling will narrow the range of word choice, resulting in more frequent exact-word predictions). However, these potential advances will have a fairly low ceiling if we consider human prediction abilities as the upper limit. Further prediction gains could be achieved if the language model was fine-tuned on the traits of a specific author (Delasalles et al. 2019); this would be an advisable step in the use case of assistive technologies for the speech impaired.

4.6.2 Context sensitivity

Previous studies investigating the impact of lookahead have shown the contrast between $k = 0$ context and different degrees of ground-truth lookahead. The setup of the present study allows us to investigate where the choice of future context modifies the output the most (i.e., where do prosodic features remain stable irrespective of the future context and where do they vary dependent on the future context). To this purpose, we calculated the range of phoneme duration and pitch feature values predicted by the TTS model in all test conditions except $k = 0$. More precisely, from the 12 predicted and ground-truth conditions ($5 \times$ *Pred*, $5 \times$ *Rand*, *GT* $k = 1$ and full context)[8], we take the max and min values from this set and calculate the difference. This analysis shows that a large portion of phonemes in the corpus alter only slightly when provided with different next word contexts. The pitch range does not exceed 300 cents for approximately 75% of our samples, which falls below the Just Noticeable Difference

[8]The predicted word and the *GT* next word are sometimes the same.

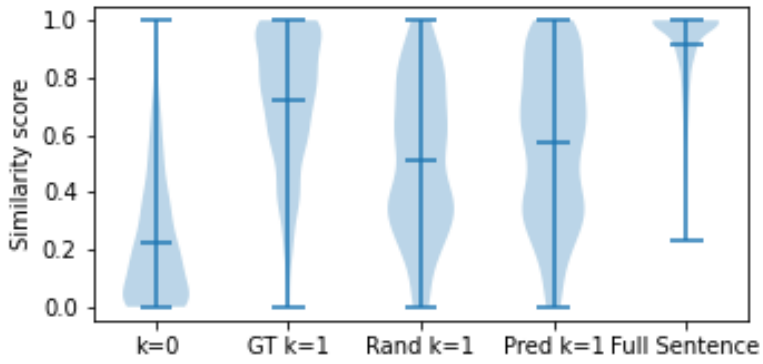


Figure 4.7: Violin plots of the distribution of similarity scores between signals generated with full context and signals generated with limited context for the 20 sentences in the MUSHRA test. The middle bars show the mean value.

(JND) threshold for pitch distance found by Hart 1981. The duration range is limited to a single spectrogram frame (11.75ms) for 40% of phonemes, which, depending on the length of the phoneme, may be imperceptible to the average listener (Quené 2007 found a JND of 5%).

We do however see some wide range values in the corpus which explain the large standard deviations in Table 4.2 and the significant variability of the *Pred* and *Rand* scores in the MUSHRA test (Figure 4.6: the maximum deviation values in a sentence show strong correlation with the mean MUSHRA similarity scores). By examining the corpus, we notice discernible patterns in the locations of large context sensitivity. With respect to pitch, we see large variation when there is a mismatch between predicted and ground-truth punctuation at the end of the next word or when there is a reporting verb (e.g., *said*, *exclaimed*) rather than the beginning of a new sentence following a punctuation mark. With respect to duration, the largest variance occurs at the beginning of sentences, at punctuation marks and in function words, especially in the coordinating conjunction *and*.

4.6.3 Full-sentence context sensitivity

Given the unexpected result that the syntactic context provided by GPT-2 does not improve prosody, we conduct some small-scale additional tests to (1) probe Fastspeech 2's use of syntax and (2) see if the syntactic context is more impactful in the full-sentence/longer context condition (since Saeki et al. 2021a report positive results with five-word lookahead). In the previous chapter, we saw that Tacotron 2 uses limited future context, however this does not necessarily hold for Transformer models that have access to the full input sequence at every timestep. Moreover, we only looked at global trends with a single expanding lookahead. Here, we look at the same fixed prefix with multiple sentence continuations to see if there is any evidence that Fastspeech 2 makes use of syntax.

We expect a TTS model would be able to capture some features of syntax, given the

common morphological features that distinguish different parts of speech (e.g., the *ed* past tense morpheme on verbs). And the presence of common function words like prepositions and determiners should assist the model to learn grouping like prepositional and determiner phrases. But whether these features lead to local or global structural representations that influence prosody prediction is an open question.

4.6.3.1 Garden-path sentences

In this probe, we look at garden-path sentences (i.e., sentences that trigger reinterpretation when a certain word in the sentence is encountered) and their alternative versions (i.e., the version of the sentence that does not require reinterpretation) (see 2.1.5.2). Sentence (1) below is an example: When the coordinating conjunction *and* is present, the word *loaned* is interpreted as a main verb for the subject NP *the large corporations*. When the coordinating conjunction is absent, *loaned* is interpreted as part of a reduced relative clause (RC).

(1) The press reported that the large corporations loaned money at low interest rates **(and)** kept accurate records of their expenses.

Grillo et al. 2018 found that garden-path sentences differ in prosody from their alternative version. In their experiment, the reduced clause reading was pronounced significantly faster than in the alternative condition, beginning from the head noun (*the large corporations* through to the critical word *kept*). There was also a marginally significant effect on pitch reset. This result makes intuitive sense since reduced RCs are usually used to convey given/presupposed information and are therefore spoken faster (this association between speed and reduced RCs goes beyond garden-path sentences).

These sentence variations are an interesting testing ground for context sensitivity in TTS: unlike many other contextual effects (see Chapter 2) that require a deeper understanding of the meaning of the text to select appropriate prosody[9], this difference is apparent from the surface syntactic form. The presence of a second verb with no coordinating conjunction forces a particular structural interpretation (at least for humans).

Here we are interested in seeing if the TTS model is taking global structural features into account, which should result in prosodic differences at a distance before the critical disambiguating word (distal effects), or if the divergent contexts (i.e., garden-path or not) only affect the immediately adjacent words. To do this, we look at the duration features (predicted by Fastspeech 2) for twenty garden-path sentences and their non-garden-path counterpart. We compare the identical phoneme sequence prior to the critical word and measure the difference in the duration predictions.[10]

[9] We assume the model has no understanding of semantics.

[10] Garden-path sentences are often read incorrectly by humans processing the sentence sequentially, and so we might expect the same from an iTTS system, but in this case, we use the full context, which Fastspeech 2 processes in parallel and so the difference should be apparent to the model.

In Figure 4.8, we see the distribution of changes in phoneme duration between two sets of phonemes: (1) phonemes from the word adjacent to the critical word and (2) phonemes at a greater distance in the past context. These results show that changes to the distal phonemes are very minor, typically a shift less than 0.1 Mel-spectrogram frames on a log scale. We observe larger changes for the critical word adjacent phonemes. Figure 4.9 shows the duration predictions for an example prefix under both conditions. The distal phoneme predictions are essentially identical. We do not find evidence that global structural considerations influence Fastspeech 2.

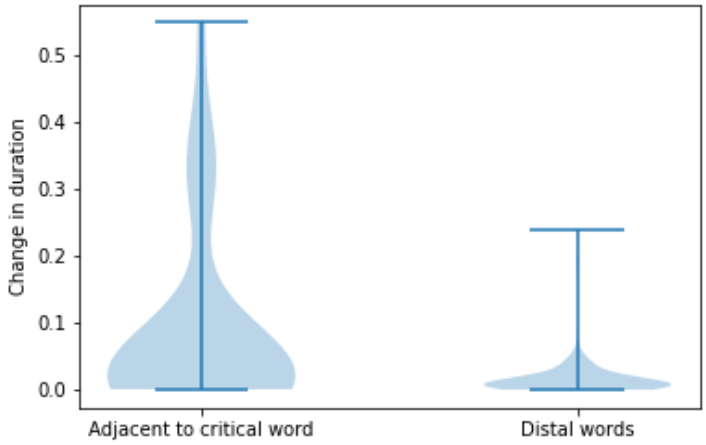


Figure 4.8: Changes in phoneme duration predictions for 20 garden-path sentences. The phonemes are separated into two groups: (1) the words immediately adjacent to the critical word and (2) the words at a farther distance from the critical word (distal words).

4.6.3.2 Multiple continuations

We perform a second experiment to investigate the impact of syntactic context at the local level. We take a sample sentence prefix (*The child*) and we vary the next word syntactic content. Forty utterance continuations were generated using ChatGPT (OpenAI n.d.)[11] for different next word syntactic contexts: preposition, relative pronouns, verb, the coordinating conjunction *and*, noun, adverb, determiner and participle. This resulted in 320 sentence continuations, each of which was generated until an end-of-sentence punctuation mark (see Table 4.4 for example sentences). The continuations vary in the lexical content and phrase lengths. The full sentences were then synthesized with Fastspeech 2 and the pitch prediction for each phoneme in the sentence prefix were extracted.

[11]Replicating the prediction experiments of this chapter with ChatGPT would likely result in improved accuracy. We do not however test this here; we simply use the tool to generate multiple continuations with controlled POS of the next word. Our purpose being to evaluate the use of structure by the TTS model and not prediction accuracy.

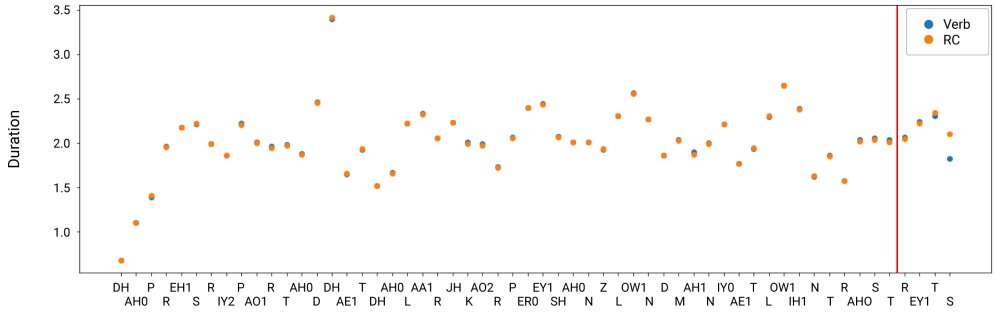


Figure 4.9: Fastspeech 2 predicted duration values for the phonemes preceeding the critical word *and* or *kept* in the sentence *The press reported that the large corporations loaned money at low interest rates* ***(and) kept*** *accurate records of their expenses.*. The blue points show the duration predictions when *loaned* is interpreted as the main verb and the orange points when the *loaned* is part of a reduced relative clause. The red line indicates the beginning of the word immediately adjacent to the critical word.

The predicted pitch values are displayed in Figure 4.10. Most of the syntactic contexts result in similar contours, with two notable exceptions: *and* and relative pronouns (*that* and *who*). We cannot of course draw any broad conclusions from this one prefix evaluation, but these results suggest that the model is quite sensitive to frequent function words that serve to structure the utterance (we also saw large context sensitivity with the word *and* in the limited-context setting). For our pseudo-lookahead use case, if one of these words is predicted incorrectly, or inversely not predicted where it should be, then we will see large difference between the ground-truth and the predicted text. The other syntactic contexts seen here have a lot more overlap in their predicted values, and so, we would not necessarily expect two noun future contexts to result in closer predictions than a noun and a verb for instance.

4.7 Conclusion and perspectives

The results from all metrics for our pseudo-lookahead evaluation show that the language model predicted text does improve prosody when compared to the $k = 0$ condition. This difference can be attributed to medial/final position distinctions, as phrase-final prosodic features are predicted for $k = 0$ words. Slight improvements over the random-text condition are also observed. These improvements are attributed to cases where the exact next word was predicted correctly, rather than to the pseudo-syntactic context provided by the language model.

Language model predictions are often incorrect and context mismatches can occasionally cause major distortions compared to the full-context prosody. To improve our model, we could implement a wait policy that delays synthesis when a context-sensitive word is encountered, similar to Pouget et al. 2016) (where predicted POS tag stability is used to guide output) or the proposed strategy in Chapter 3 (where stability features like word length are used to

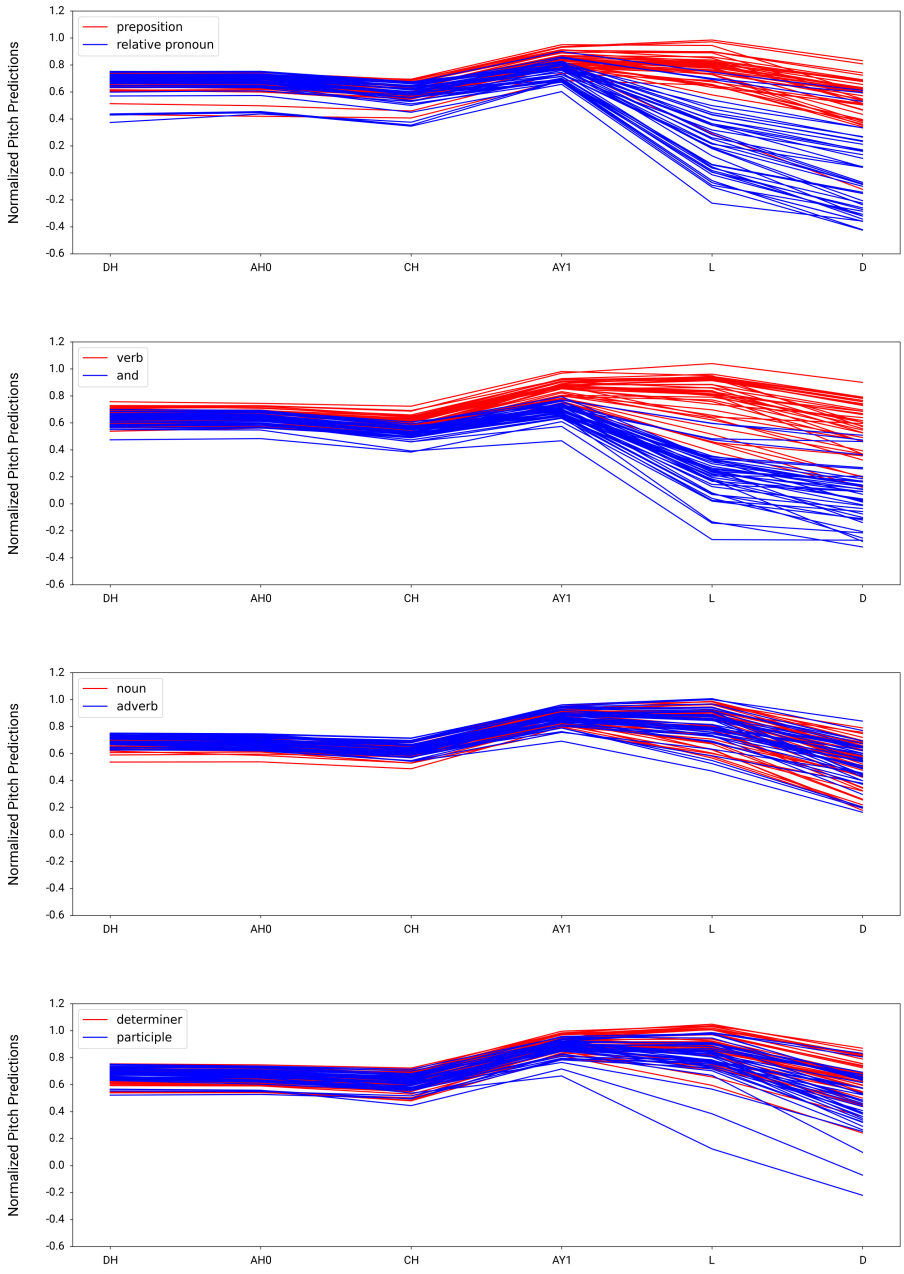


Figure 4.10: Pitch predictions from Fastspeech 2 for the prefix *The child* with different next word syntactic conditions (preposition, relative pronoun, verb, the coordinating conjunction *and*, noun, adverb, determiner (*the*) and participle). The Fastspeech 2 pitch predictor is trained on log f0 values normalized to zero mean and unit variance for each utterance.

Table 4.4: Example sentences for next word syntactic context evaluation.

Next word syntactic context	**Example Sentences**
preposition	The child in the park played on the swings. The child on the soccer team scored the winning goal.
relative pronoun	The child who hit Tommy is crying. The child that was picking flowers was admiring their beauty.
verb	The child listened attentively as their teacher told a story. The child has been diagnosed with a form of epilepsy.
coordinating conjunction (and)	The child and the dog played in the park. The child and the sibling argued over who gets to go first.
noun	The child athlete excelled in sports. The child dancer gracefully performed on stage.
adverb	The child eagerly waited for his turn. The child sadly remembered the loss of a loved one.
determiner	The child the teacher was talking about is very intelligent. The child the park ranger was leading on a hike was having a great time.
participle	The child studying hard for their exams was determined to do well. The child exhausted by the long hike collapsed on the ground.

predict how close the word is to its final representation).

Another option is to use more abstract future context representations similar to Saeki et al. 2021a. These abstractions could potentially help the TTS model learn more neutral solutions for context-sensitive words, smoothing over jarring mismatches between predictions and the ground-truth. However these neutral solutions would most likely lead to overall flatter prosody, which is at odds with our objectives in this thesis. We have to question what these abstractions can learn: if different futures that diverge in contextual features such as word choice, number of syllables, POS, and position in the prosodic phrase/utterance are encouraged to resemble one another in a generic future embedding, what exactly do they represent? Using concrete predicted words can encourage more bold/less generic choices from the TTS model, but unfortunately these choices may be at odds with the actual future content.

In this chapter, we have also seen further evidence that TTS models trained exclusively on phonemes employ a rather shallow use of syntax/context. Prediction decisions appear to be dominated by the lexical/phonological features of the immediate local context with little influence from the underlying structural features. As far as the local context is concerned, Fastspeech 2 does appear to be somewhat sensitive to the syntactic context, in particular when function words that mark coordination or subordination are involved.

In the next chapter, we will explore the use of language models to predict prosodic features directly, as opposed to predicting future text and then relying on the TTS system to model appropriate prosody. It is hypothesized that these representations, which encode syntactic and semantic features, will make richer use of context when predicting prosody.

Chapter 5

Predicting and controlling prosody with language models

Contents

In the previous chapters, we saw that vanilla neural TTS models make very shallow use of context when predicting prosodic features: inferred prosodic values/representations are mostly influenced by the immediate next word (when the previous context is held constant) and we saw some evidence that larger structural factors are barely taken into consideration (e.g., with garden-path sentences). What's more, single sentence vanilla models can neither make use of semantic factors in the local context, nor discursive factors in the wider context.

These observations led us to believe that the "success" of limited lookahead iTTS, as judged by the difference from full-sentence TTS, was more a reflection of the shortcomings of these full-sentence end-to-end models than it was of the true value of further lookahead. And so, in the remainder of our work, we experimented with prosody prediction using more contextually-informed models. More precisely, instead of using LMs to predict future context for a vanilla TTS model, we pivoted to using LMs to predict prosodic features directly. These predictions were then used to condition TTS synthesis; this division into prediction and synthesis allows for more controllable/context-adaptable TTS. We also test conditioning on extended contexts (i.e., previous utterances) with the aim of increasing contextual appropriateness.

We build on previous research that has seen positive effects of conditioning TTS on Transformer LMs and extended context. We apply these techniques to iTTS models and we provide new contributions by (1) investigating how the accuracy of linguistically-aware prosody prediction changes in different context conditions (incremental, full-sentence and/or extended context), (2) investigating the type of information provided by the LM and (3) testing context-informed adaptation strategies for segmenting speech in large input latency iTTS application. We explore these themes through two information structural properties: prominence, with special attention on contrastive focus, and the segmentation of information units (boundary prediction).

Prosodic prominence is a speech feature that would seemingly benefit from increased future context (beyond the very limited lookahead our results from Chapters 3 would suggest). It is much more difficult to predict appropriate prominence assignment if future disambiguating information is not available (e.g., *I wore the RED hat, not the BLUE one* vs. *I wore the red HAT, not the red SCARF*). Previous context should likewise provide important information, and so we compare predictions when our model does and does not have access to previous sentences. In addition to evaluating the contribution of the amount of context, a further aim in this chapter is to evaluate the type of contextual information provided by the LMs. And so we selected a particularly challenging task that cannot rely on simple heuristics (such as POS or lexical identity) to make correct predictions: the prediction of contrastively focused personal pronouns, which often require discursive and pragmatic knowledge to predict correctly.

Prosodic boundaries are similarly a feature whose prediction could be aided by full-sentence context, since the global syntactic structure and phrase length have been shown to affect boundary placement and strength (see Section 2.1.5.1). Evaluating the effects of lookahead for this task would be informative for continuous input iTTS applications like speech-to-speech translation. Here however, our goal is not to evaluate prediction accuracy with regards to normal-speed ground-truth speech samples, but rather to adapt the synthesized speech to an incremental context where the input latency is quite high (AACs[1]). In this context, there is a need to find a balance between reactivity and the digestibility of the speech stream. Synthesizing each word as it becomes available is the most reactive form that could be adopted, however this is not a natural form for speech to take, as language is typically presented in informationally-motivated chunks. In this work, we compare methods for segmenting speech that are either based on LM-predicted features (POS tags or boundary

[1] Augmentative and Alternative Communication

strength values derived from the audio signal) or count-based methods (one/two-word(s)-at-a-time). We evaluate the quality of these models with traditional subjective tests (a modified MUSHRA and an AB test) and we experiment with a sentence verification test designed to evaluate cognitive load.

To train our adaptable synthesis models for both the prominence and boundary tasks, we adopt an automatic prosody annotation tool based on continuous wavelet transforms. Previous studies have reported successful control of content words when conditioning on these annotations. We conduct a perceptive test to evaluate the control provided over short function words, which have relatively few prominent examples in TTS corpora, but whose emphasis can have a large effect on intended meaning.

To summarize, our objectives in this chapter are to evaluate the ability of LMs to predict and control prosody with limited and extended contexts. We investigate:

1. prominence prediction in different context settings (incremental, full-sentence, extended context).
2. the knowledge provided by LMs. We do this by assessing contrastive focus prediction on a corpus of difficult examples (that cannot be predicted correctly using simple heuristics).
3. the amount of prosodic control that can be achieved on a class of function words (personal pronouns).
4. user preferences for segmentation in a high-input latency context.

This chapter is organized as follows: we begin by reviewing related work on the topics of prosodic representation and control and on context-aware TTS. In Section 5.2, we present our work on prominence and contrastive focus (an extended version of Stephenson et al. 2022 which was presented at Interspeech 2022)[2]. Section 5.3 presents our work evaluating segmentation techniques and we finish with a conclusion.

5.1 Related works

5.1.1 Prosodic representations and control

Adapting TTS to suit a given context requires symbolic or latent representations of prosodic features in order to control speech output (which can be controlled at different levels of granularity, from phonemes to the utterance). Techniques to detect, represent and control prosodic features in neural TTS fall into two main categories: unsupervised and supervised methods.

[2]Here we present updated results after manually correcting transcription and forced alignment errors in the corpus. We see slightly better results, but the overall trends remain the same as in the original paper, with the exception of the randomly initialized BERT model, which shows some improvement with extended context.

5.1.1.1 Unsupervised methods

Exemplar-based methods In exemplar-based methods (e.g., Skerry-Ryan et al. 2018; Wang et al. 2018), the prosodic characteristics of reference speech recordings are transferred to new utterances. This usually involves passing a Mel-spectrogram to a reference encoder unit, which extracts a representation of the prosodic or stylistic features of the speech. This representation is later combined with text representations, and the TTS model uses these combined features to generate a speech sample. These models are not always entirely successful at disentangling form from content and so adversarial training and style losses have been proposed to further separate these features (e.g., Ma et al. 2019b; Wang et al. 2021c).

Variational autoencoders (VAEs) VAEs can be used to learn prosody/style representations in a latent acoustic space (e.g., Hsu et al. 2017; Hono et al. 2020; Sun et al. 2020). To achieve interpretable control over the audio output of these models, effort must be made to disentangle the features of interest. And while labels are not required at the onset to learn these spaces, human interpretation of the discovered space is required. For example, Sun et al. 2020 proposed a hierarchical latent variable model trained with conditional VAEs where lower-level prosody is conditioned on higher-level prosody latent representations (phone, word, utterance). They used a training schedule to separate prosodic attributes among latent dimensions (i.e., they progressively added trainable dimensions) and after training they measured the levels of disentanglement by modifying individual dimensions in the latent space while holding the other layers constant and measuring differences in variance between different prosodic attributes. Clustering methods are another option for interpreting the latent space (e.g., Ellinas et al. 2023).

Because the prosodic features learned by unsupervised models are difficult to disentangle and interpret, work on categorized prosodic event detection (i.e., the presence or absence of pitch accents and prosodic boundaries) has mostly been carried out using supervised methods.

5.1.1.2 Supervised methods

In supervised methods, the TTS model is trained on human or automatically generated labels:

Human labelling Human labelling is likely to be the most accurate reflection of perceived prominence and boundaries, however obtaining such labels is a time consuming process that requires expert knowledge or a large number of annotators to achieve reliable results (e.g., Rapid Prosody Transcription (Cole et al. 2017)). As a result, the size of human annotated corpora is usually quite small (Calhoun et al. 2010; Ostendorf et al. 1995), which is not ideal for neural network training. Some speech corpora have been designed specifically to provide variations in prominence placement (e.g., Latif et al. 2021; Strom et al. 2007) but the reduced linguistic complexity (i.e., simple SVO or repeated template sentences) and the disassociation from a meaningful communicative context can oversimplify and/or over-exaggerate the

expression of prominence, making (detection or synthesis) models trained on these datasets less transferable to more naturalistic speech.

Automatic labelling Automatically generated labels may not be as accurate as human labels, however they make it possible to treat a much larger quantity of data. These labels can be based on acoustic or linguistic features, or a combination of the two.

Prominence. Good overall performance in *prominence* prediction can be achieved with text-only features (lexical, syntactic, information status properties). Models trained on these features however do not generally perform well on marked/non-standard prosodic realizations (an issue we address in this chapter). Previous works addressing general accent placement (i.e., marked and unmarked structures combined) include Altenberg 1987 (as cited in Ross and Ostendorf 1996) who came at the issue from a POS perspective, using fine-grained categories to identify those that are likely to be prominent (e.g., "wh" adverbs, ordinals and quantifying pronouns). Hirschberg and Litman 1993 predicted labels from POS, surface order and given/new status, Pan and McKeown 1999 used word informativeness (as measured by n-gram probabilities) and Nenkova and Jurafsky 2007 used accent-ratio, a probability measure of an individual lexical item being accented.

Acoustic-based models for prominence detection are better able to capture deviations from standard prosodic pattern. These models either use measurable prosodic values as input or learn relevant features directly from a neural network. Rosenberg 2010 developed an automatic ToBI labelling system, using pitch, energy, duration, spectral tilt and contour slopes as features and traditional machine learning techniques for identification and classification (logistic regression, SVM). Nielsen et al. 2020 evaluated different amounts of context with either a CNN or a CNN + bidirectional LSTM architecture and found the full utterance input with bidirectional encoding yielded better results than single or three token windows. They also compared text-only (using GloVe embeddings (Pennington et al. 2014)), speech-only and combined models. They found that text-only models performed no better than an all-content-word-accented baseline but that text features in combined models could provide a slight improvement over speech-only ones (see Ananthakrishnan and Narayanan 2008; Chen et al. 2004 for similar work).

Boundaries. Studies looking at *boundary* detection and prediction have likewise used text and acoustic features. Some elaborate methods to predict boundaries based on the syntactic structure of the utterance have been developed (Cooper and Paccia-Cooper 1980; Gee and Grosjean 1983; Ferreira 1988). For example, Cooper and Paccia-Cooper 1980 devised an algorithm that counts and weighs dominating nodes on the left and right of each potential boundary site in a syntactic tree, excluding minor categories and non-terminal nodes on the left side. Watson and Gibson 2004 proposed a simpler method based on the length of the most recently completed syntactic constituent, as well as that of the upcoming syntactic constituent. In order to execute these algorithms, the syntactic structure of the complete utterance, or at least the distance to the end of the next syntactic constituent, are important factors, making their implementation in the incremental setting difficult.

Acoustic models have focused on pause, breath, pitch and duration features. Wightman and Ostendorf 1991 trained classifiers to detect silence and breath features which could identify major prosodic boundaries. For smaller boundaries, they tested phone duration features, such as the normalized lengths of the rhymes of word final syllables. These were compared with the duration of onset consonants, which in the case of boundaries should not lengthen as much as the rhyme, whereas in the case of pitch accents they should undergo a comparable lengthening. The model was not entirely successful at fine-grained boundary type prediction, but fairly good at a binary boundary/no boundary classification. Wang and Narayanan 2004 studied pitch breaks, pitch resets and break intervals to differentiate between fluent and disfluent boundaries in spontaneous speech and achieved 75% accuracy with no other additional linguistic information. Biron et al. 2021 used speech rate discontinuities (which reflect pre-boundary lengthening and post-boundary acceleration) to detect breaks.

Continuous Wavelet Transforms. Suni et al. 2017 proposed a signal processing-based, hierarchical method for extracting prosodic events (both prominence and boundaries). This is the technique we adopt for our work because it mimics human multiscale perceptive processes and allows for the automatic extraction of word-level prosody tags. The technique uses continuous wavelet transforms (CWTs) to analyse the speech signal at several timescales which correspond to different levels in the prosodic hierarchy (phones →syllables →words →phrases). By detecting changes at multiple-levels, a richer representation of prosody is extracted. Values for the strength of word-level prominence or prosodic break (i.e., boundaries) can be obtained by: (1) combining normalized f0, energy and duration into a composite signal, (2) performing the CWT, (3) establishing lines of maximum or minimum amplitude connecting the various timescales (maximum for prominence (the black lines in Figure 5.1) and minimum for boundaries (the white lines)) and then (4) calculating a weighted sum of the points in this line (attributing greater weight to the higher levels).

5.1.2 Context-aware TTS

5.1.2.1 Transformer language models

Several recent works have investigated the use of Transformer LMs in TTS models. These experiments either condition synthesis directly on extracted embeddings or divide the pipeline into prosodic feature/embedding prediction and context-conditioned synthesis (as we do in this chapter).

Direct conditioning. Hayashi et al. 2019 tested the integration of BERT-derived representations of subword tokens and of sentence-level representations into a Tacotron 2 model. MOS scores showed small but significant improvements with both models, with larger gains from the more fine-grained subword embeddings. Kenter et al. 2020 incorporated BERT into an RNN TTS model and found it could help with the pronunciation of complex compound noun grouping which can be determined based on semantic content (e.g., (diet (cat food)) not ((diet cat) food)). Their experiments showed smaller BERT models performed better than

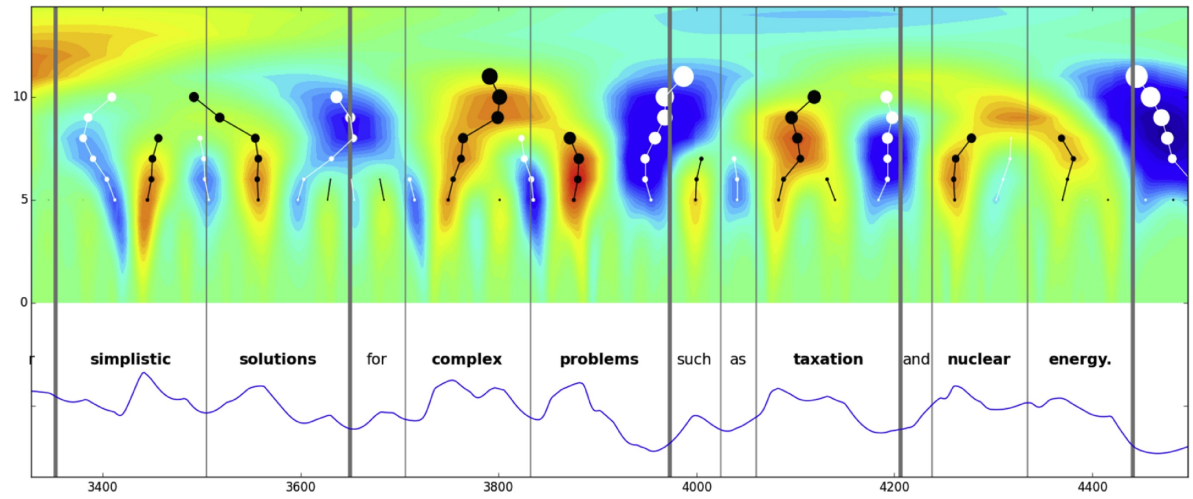


Figure 5.1: The lower panel shows the combined f0, energy and duration signal. The upper panel shows the scalogram resulting from CWT analysis. The black lines show the lines of maximum amplitude (LOMA) and the white lines show the lines of minimum amplitude (LomA). Image from Suni et al. 2017.

larger ones and fine-tuning on prosodic data was crucial.

More specialized uses of language-models for TTS have included contextualizing multilingual text representations for code-switching TTS (Zhou et al. 2020) and word-boundary prediction in Mandarin (Xiao et al. 2020).

Two-stage models. Conscious that directly conditioning prosodic prediction on context embeddings may have the effect of entwining different layers of speech representation (segmental and suprasegmental), Hodari et al. 2021 used a word-level prosody reference encoder with a high bottleneck to encourage the learning of prosodic representations (separate from phonemic ones). They then trained a BERT model to predict the prosody embeddings at inference time. Zou et al. 2021 use a similar two-stage method, but instead of data-learnt prosody representations they used human-annotated ToBI labels to train the speech synthesis model and then had an LM (ELECTRA (Clark et al. 2020)) predict the labels. Yoon et al. 2022 use GPT-3 (Brown et al. 2020) to predict an appropriate emotion category and its strength from text.

Similar to Hodari et al. 2021 and Zou et al. 2021, Talman et al. 2019 used BERT to predict word-level prosodic features. They however used CWT-derived prominence values as their prediction target. We build upon this work because (1) (as previously mentioned) we can extract the labels automatically and (2) predicting discretized prosodic values is easier to interpret than vector representations which may encode other aspects of prosody in addition to the features we want to study (i.e., prominence and boundaries).

5.1.2.2 Discourse-aware/Extended context TTS

Other efforts to increase contextual appropriateness have included modelling discourse phenomena (i.e., discourse relations, speech acts, information structure), taking the position in the discourse into account and incorporating a larger textual and/or acoustic context. Some of these works also incorporate Transformer LMs.

Discourse phenomena Hu et al. 2016 modified speech synthesis to reflect the pause patterns that correlate with the discursive hierarchy and duration differences that are associated with degrees of discourse centrality (i.e. nucleus vs. satellite units). Syrdal and Kim 2008; Syrdal et al. 2010 clustered the acoustic characteristics of different speech acts and then used a symbolic representation of the groupings to condition a TTS model. Guo et al. 2021 proposed a conversation context encoder that used BERT sentence embeddings of the past and current sentence(s) to replace broad speech act categories with more fine-grained contextual representations. Domínguez et al. 2022 tested a hierarchical thematicity model to improve information structure expression in complex sentences.

Position in speech paragraph Farrús et al. 2016 studied sentence-level prosodic variations within multi-sentence discourses. In Peiró-Lilja and Farrús 2018, they applied the significant start/middle/end sentence features found in the original study to speech synthesis and found that these characteristics, such as the gradual decline in pitch range and speech rate variations, improved subjective evaluations.

Extended context Cong et al. 2021 conditioned synthesis on global style tokens (i.e., prosody embeddings (Wang et al. 2018)) from previous speech segments to model entrainment features in conversation speech. Makarov et al. 2022 trained multi-sentence TTS models, while also conditioning on LM-representations, and observed improved pausing and pacing in long form synthesis. Gallegos et al. 2021 compared acoustic and linguistic representations of past context at different levels of granularity (word and utterance levels). They used representation derived from either pretrained models (a Mel-spectrogram image encoder for acoustic context and an LM for textual context) or from a context encoder unit learned jointly with a TTS model; attention was employed to learn relevant features for the current sentence. Results showed the pretrained models were more effective at extracting important features for the acoustic condition, but both pretrained and jointly learned encoders were comparable for the text condition. Furthermore, the different types of contextual input were complementary, with combined utterance-level acoustic and word-level text representations providing the largest gains in a subjective test. In the current study, we limit our investigation to text features of previous utterances. In assistive technology applications, previous utterances by the user will be synthetic and so the information provided by past acoustic features would be limited. On the other hand, incorporating the acoustic features of the conversation partner's contributions has the potential to improve predictions, but this is not a topic we explore in this thesis.

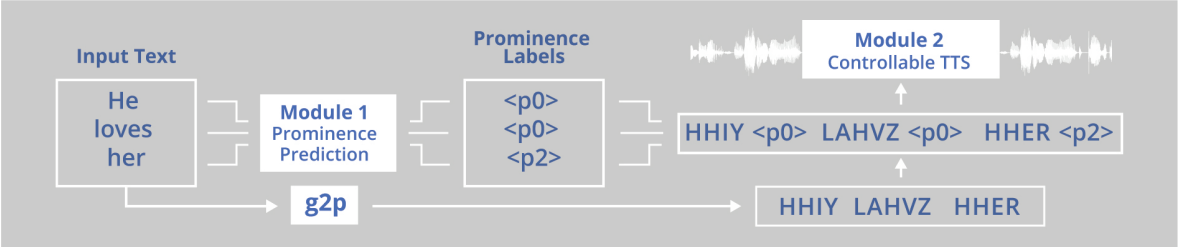


Figure 5.2: TTS overview: The system is split into two modules. The first uses an LM to predict prominence labels. The second controls the prominence in the synthetic speech in accordance with the predicted labels.

5.2 Prominence

Contrastive focus and high-level linguistic context. "*HE loves her*", "*he LOVES her*", and "*he loves HER*" all have the same textual content, but three distinct communicative goals. Indeed, such *contrastive focus* is used by speakers to evoke alternative sets in the discourse (Rooth 1992). This can be utilized to make explicit intended discourse relations between clauses/paragraphs/sections, to highlight a fact that the listener may find surprising, or to express a specific semantic or pragmatic meaning. The prediction of contrastive focus placement therefore often requires high-level linguistic understanding. Current vanilla neural text-to-speech (TTS) synthesis systems lack this understanding and will always pronounce the above sentences in the same way, irrespective of the context. In this section, we investigate methods to predict the placement of contrastive focus (and other forms of strong prosodic prominence) and to control it in a TTS system. Figure 5.2 illustrates our overall approach which addresses both *predicting* word-level prominence labels (Module 1) and *controlling* the realization of prominence in synthetic speech using the prominence labels (Module 2).

One way to insert sophisticated linguistic information may be through the use of contextualized word embeddings. While other works have explored the use of Transformer LMs to predict prosodic and stylistic features (Section 5.1.2.1), it has not been fully explored how much the encoded word representations actually imbue high-level knowledge. In other words, do they provide information about the content and context of the message or do they only provide/reinforce low-level linguistic features such as the likelihood of lexical prominence, parts of speech and position in the sentence? For prominence/pitch accent prediction, a fairly high baseline can be achieved using word majority/accent-ratio alone (i.e., if a lexical item is usually prominent in the training set, it is likely to be prominent in the test set) (Nenkova et al. 2007). Moreover, Hirschberg and Litman 1993 found that in a binary pitch accent prediction task, using broad word category distinctions (open/content or closed/function) could achieve 68% accuracy; more fine-grained division of the closed class category brought that number up to 77%.

In this work, we probe LMs, in the present case BERT (Devlin et al. 2019) and GPT-2 (Radford et al. 2019), by choosing a testing ground that cannot rely on simple heuristics to achieve good results: the prediction of contrastive focus on personal pronouns. Personal

pronouns, at least in the corpora typically used to train TTS systems, are majority non-prominent.[3] Therefore, instances that are contrastive cannot be predicted solely based on their lexical identity or word class affiliation.[4]

Incremental and extended-context conditions. In addition, to these difficult examples, we evaluate corpus-wide prediction of prominent words from an incremental (using GPT-2), a full-sentence (using BERT) and an extended context perspective (using both LMs). As we did in Chapter 3, we look at different degrees of lookahead (k), but this time looking at the prediction of a specific prosodic feature (i.e., prominence).

Controlling prominence. Assuming we are able to predict good candidates for contrastive focus, it remains to be verified if we can *control* the contrastive prominence on pronouns in a neural TTS system. The previous study by Suni et al. 2020, on which we base this work, report successful control over other word categories using a CWT-conditioned model (although this was not evaluated with a perceptual test). Assuming these results are reliable, prominence on personal pronouns, and other function words, may still present a particular problem for control, due to the sparsity of examples in the training data (Latif et al. 2021 found that a model trained on subject and object focus samples did not generalize to verb focus). Even if prominence on certain function words is not as common as on other types of words, their control is still important for contextual appropriateness and the expression of subtle nuances in meaning. We evaluate control (which can be thought of as the degree of disentanglement between word-level prominence and global acoustic considerations) that is attained using CWT-prominence labels. We do this with a perceptive ranking test (Section 5.2.7).

Naturalness and user preferences. In terms of naturalness and user preferences, previous efforts to control contrastive focus, or emphatic expression, have seen mixed results. Pitrelli and Eide 2003, using ToBI labels to insert contrastive focus in TTS, found a 0.4 increase in MOS scores for appropriately focused words. Strom et al. 2007 also saw improved listener ratings when synthesizing with contrastive focus in a unit-selection model. Li et al. 2012, however, saw negative results: they tested listener preferences for emphasis on (1) a member of a lexical contrast pair, (2) a word other than those in the contrast pair and (3) no words (a neutral, no emphasis sentence). Their results showed subjects preferred contrast on contrastive element over non-contrastive elements, but they preferred no emphasis overall due to distortions in the audio caused by the increased emphasis.

For the moment, we do not evaluate naturalness or user preference. It is possible the CWT-control causes a small degradation in terms of overall smoothness compared to a vanilla model, but regardless, if appropriate words are emphasized, this could increase global discourse understanding.[5] Properly weighing these two criteria for an interactive TTS application did

[3] We conflate the notions of prominence and contrastive focus for personal pronouns; when they are prominent, they typically possess a contrastive meaning.

[4] We also considered including prepositions and determiners in the corpus, similarly words that can bear contrastive focus, but are usually unstressed. However we found too few examples of these types in our corpus.

[5] Research in second language acquisition has shown contrastive focus to be a high-value pronunciation feature (i.e., its proper usage contributes more to understanding than other types of pronunciation errors.

not seem feasible in the time we had available: Evaluations that can be done relatively quickly, where listeners rate short audio clips divorced from a natural communicative context, are likely to overvalue naturalness and undervalue expressiveness of discourse dependent phenomena. So for now, we only evaluate perception of prominence.[6]

5.2.1 Contrastive focus

What is contrastive focus? An influential framework for understanding contrast was developed by Rooth (Rooth 1992) in his theory of *Alternative Semantics*. The theory makes a distinction between the ordinary semantic value of a sentence and its focus semantic value. The focus semantic value is "a set of alternatives from which the ordinary semantic value is drawn, or a set of propositions which potentially contrast with the ordinary semantic value." For example, the focus semantic value of [[Mary]$_F$ likes Sue] is the set of propositions that fit the structure *x likes Sue* (e.g., {*Mary likes Sue*, *John likes Sue*, *The teacher likes Sue*, etc.}) and the focus semantic value of [Mary likes [Sue]$_F$] is the set {*Mary likes Sue*, *Mary likes John*, *Mary likes cheese*, etc.}. From the focus semantic value, a contextually appropriate alternative set is constructed using pragmatic constraints (i.e., discourse, lexical and world knowledge). If members of the alternative set are salient in the discourse (either implicitly or explicitly), contrastive relationships can be inferred.

Two constituents, A and B, are said to be symmetrically contrastive if A belongs to the focus semantic value of B. (e.g., *An **American** farmer met a **Canadian** farmer.*). Wagner 2006 places a further constraint on constrastiveness: that suitable alternatives must form a partition. So *high-end convertible* and *cheap convertible* are non-overlapping partitions that will evoke a contrastive relationship whereas *red convertible*, which could potentially overlap with *high-end* or *cheap*, does not.

We saw in Chapter 1 that it is a subject of debate whether or not there is a difference between focus and contrastive focus. Whether contrast is a distinct category from focus, or simply at the strong end of a focus gradient scale, does not concern us here. What we are interested in is predicting all instances of contrast (or focus) that result in increased prosodic prominence. Our prepared corpus includes instances of contrast belonging to both the topic and focus domains.

5.2.1.1 Where does contrast occur?

While ultimately the decision to mark contrastive focus with prosodic emphasis is the choice of the speaker, the probability of it occurring does increase within certain discourse relations,

Levis and Levis 2018 studied the effects of dedicated instruction in contrastive focus prosody and found it could significantly improve comprehensibility ratings of students even while other fluency measures remained the same.

[6] Additional losses to the TTS system (e.g., an adversarial loss that discriminates between natural and synthetic speech) could potentially reduce any degradation in naturalness.

near lexical items marking contrast, in the presence of salient alternative sets and in certain syntactic constructions (some examples in (1) and (2) adapted from Umbach 2004 and Repp 2016).

Repp 2016 describes four discourse relations that have been studied in relation to contrastive prosody: similar, oppose, correction and Q-A pairs. The first two relations are defined in terms the notion of the current (usually implicit) question under discussion (QUD): two segments expressing a similar relation make the same contribution to the QUD (e.g., (1)a) and the segments forming an oppose relation make contrasting contributions (e.g., (1)b - where there is a violation of expectations). In a corrective relationship, one of the segments rejects the other (e.g., (1)c and d). In Q-A pairs, a question evokes an alternative set and the answer selects one of these alternatives. Contrastive discourse relations are often expressed using contrastive syntactic forms like those in (2).

(1) a. John was mowing the lawn. Pete was too.
b. John is tall, but he's no good at basketball.
c. Last week, John went to London.
[No,] He went to PARIS
d. John didn't invite SUE but MARY.
e. Who ate the cookie?
JANE ate the cookie.
f. John only saw SUE at the dinner party.

(2) a. Parallelism - *JOHN went to PARIS and/but MARY went to LONDON.*
b. Ellipsis - *I don't want cake, but PETER does.*
c. Right-node rising - *Susan BUYS and Mary EATS the cookies*
d. Topicalization - *JILL, they like. JOHN, they don't.*

Repp 2016 also defines contrasting constituents in terms of explicitness and number of alternatives in the discourse. An **explicit alternative** (*John put an apple in his new bowl, then he put a BANANA.*), an **explicit alternative set** (*John bought a banana and an apple. He put the BANANA in his new bowl.* or an **implicit alternative set**, evoked by semantic relations such as *kind of* (*John was choosing fruit for his new bowl. He picked a BANANA.*) can all elicit contrast.

The presence of certain lexical items are also good indicator of contrast. For instance, focus particles (e.g., *only* (e.g., (1)f), *even*, *just*), and other words marking sets or parts of an entity or event (e.g., *the first/last/next, turn, side*). Discourse markers can also explicitly mark a contrast (e.g., *but, however*).

5.2.1.2 Is contrastive focus predictable?

Related works The prediction of contrastive prominence is a difficult task due the combination of lexical, syntactic, semantic, discursive and pragmatic factors that can all contribute to its occurrence. A further complication is the fact that prosodically marking contrast appears to be optional in some cases. Previous works in text-only contrastive focus prediction (Badino and Clark 2008; Badino et al. 2009; Li et al. 2012) simplified the task by limiting samples to those where the contrast was explicit, apparent from a single sentence context and expressed through word pairs (as opposed to contrasting phrases/sentences). These works trained binary classifiers that labelled word pairs (belonging to the same POS category) as contrastive or not. They used features such as explicit lexical contrast markers (e.g., *rather than*, *instead of*), dependency relations (e.g., are both words in a subject relationship to their verb?), measures of parallelism (the edit distance between two clauses), morphological similarity/difference measures (e.g., *formal*/*informal*), semantic relationships extracted from ontologies (e.g., hypernyms, antonyms, member-of, etc.), accent ratio and word identity. Even with the simplified task, these models struggle to identify contrast, with recall scores below 31%. Other works interested in contrast detection for natural language understanding include acoustic correlates of prominence among their feature set (Zhang et al. 2006; Nenkova and Jurafsky 2007). This improves detection rate, however these features are not available for TTS.

We are interested in improving upon these more evident forms of contrast but also in predicting contrasts arising from discursive saliency. Discourse induced cases are equally important for producing contextually appropriate speech, though they do present a significant challenge for prediction due to the higher-level knowledge required. Modern LMs should be able to provide some information about the discourse relations (Section 2.2.3.6) and provide more contextually refined semantic representations (Section 2.2.3.2), but whether they encode an understanding of discursively salient contrastive sets is yet to be verified. One probe into BERT's representation of contrast ((Lei et al. 2021) for a lexical cohesion identification task), saw some evidence that contrast understanding relies on some local, basic heuristics. They tested BERT's ability to distinguish between lexical items being used to communicate a contrast relation from identical word pairs that just happen to occur together in the same neighbourhood. For example, the words *positive* and *negative* are lexical opposites that often express contrast, but in the passage *The reviewers are rather positive about this paper. They are nominating it for Best Paper for the discovery of a negative finding that dispels the conventional wisdom.*, they do not. BERT's accuracy on this task was around 70%; the authors found performance was weaker for lexical contrast pairs that occurred in different sentences and performance was also best for pairs that were in similar lexical contexts (e.g., *lived in the darkness* and *live in the light*). This suggests that classification was dependent on surface features and perhaps less so on the larger discourse.

Degrees of difficulty In this work, we chose to study prominent personal pronouns because these marked prosodic forms cannot be predicted from the basic heuristics of POS category or lexical identity alone (critical features in previous prominence prediction models). These

	Limited context	Extended context
1	Very odd! And **YOU** only the second.	"Yes, ma'am, all." "All! What, all five out at once?
2	I was sometimes quite provoked, but then I recollected my dear Elizabeth and Jane, and for **THEIR** sakes had patience with her.	I talked to her repeatedly in the most serious manner, representing to her all the wickedness of what she had done, and all the unhappiness she had brought on her family. If she heard me, it was by good luck, for I am sure she did not listen.
3	Even if **HE** could form such a design against a young woman of Lydia's connections, which is not likely, can I suppose her so lost to every thing?"	Her pale face and impetuous manner made him start, and before he could recover himself to speak, **SHE**, in whose mind every idea was superseded by Lydia's situation, hastily exclaimed, "I beg your pardon, but I must leave you.
4	**I** shall not be able to keep you–and so I warn you.	"I am sure I do not know who is to maintain you when your father is dead.
5	I have nothing to say against **HIM**; he is a most interesting young man;	"Seriously, I would have you be on your guard. Do not involve yourself or endeavour to involve him in an affection which the want of fortune would make so very imprudent.

Table 5.1: Example sentences from personal pronoun corpus.

challenging examples must rely on different sources of information to be predicted correctly (which could potentially be provided by LMs). However, even within this set of difficult examples, there are different degrees of prediction difficulty. Some samples possess surface markers of contrast in the local environment, while others gain their contrastiveness from more distant discursive content.

In this section, we will examine a sampling of the sentences from our collected corpus to illustrate the degrees of predictability (more details on the corpus preparation are provided in the next section). The samples and their extended contexts are provided in Table 5.1.[7]

A number of examples contain lexical items that are commonly used to communicate

[7]We attempt to evaluate predictability through two lenses: basic lexical/syntactic heuristics and discourse understanding. The neural networks we train may however use other heuristics, not immediately evident from a human perspective, when making their predictions.

contrastive perspectives or considerations, for example the word *sakes* in Sentence 2. These are typically paired with a possessive determiners (e.g., *their*). Other examples from our corpus include *turn*, *part*, *side*, *account*, and *opinion*. While the pronoun is not predictable from its own lexical identity, it can be inferred through its syntactic or semantic relationship with a neighbouring word or simply through proximity.

Common phraseology marking contrastive topics can also be used for prediction. *And x* (*And you* in Sentence 1) is commonly used to signal a topic shift/contrast with the previous topic. *As for x* and *Speaking of x* are other common examples. Sentence 1 possesses additional contrastive cues which should make it an easy sample for prediction. These include the elliptical structure and the alternative set markers *only* and *the second*. The extended context further reinforces the saliency of the set (*all five*).

The contrast in Sentence 3 requires higher-level understanding. There is an overt discourse marker signaling a contrastive discourse relation (*even if*), however this does not automatically entail that *HE* will be the focus of the contrast. The pronoun *her* later in the sentence makes this contrast clear, but this appears at a sizeable distance from the first pronoun[8]; making the connection between the two requires comprehension of the individual discourse units and their relation to one another.

Other instances are even more subtle, as they do not have an overt comparison, like in Sentence 5 where the implied comparison is between two aspects of *HIM*; his character and his fortune. Making this connection is likely to require world knowledge about the alternative set of desirable traits of a romantic partner.

Other sentences, such as Sentence 4, are highly dependent on the extended context: In isolation, the relevant alternative set (i.e., *those who may maintain you*), is not made clear and so *I* would seem an unlikely candidate for contrast.

5.2.2 Prepared datasets

5.2.2.1 Corpus selection and preprocessing

The selected audiobooks for our corpus are literary texts sourced from Librivox[9] and from the Blizzard Challenge 2013 dataset.[10] Criteria for book selection included open-source status, the availability of multiple recordings with different speakers (minimum 3 so that we could examine the likelihood of contrastive focus being employed), audio quality and the subject matter (we favoured books dealing with interpersonal relationships as they are more likely to contain contrastively focused pronouns). Five novels (x 3 speakers) were selected for the

[8]Contrasts that are only made explicit at a later stage are often marked typographically by the author (e.g., italics) to facilitate comprehension for the reader who is reading linearly. Our corpus is stripped of typographical differences and so predictions must be made on text content alone.

[9]`https://librivox.org`

[10]`https://www.synsig.org/index.php/Blizzard_Challenge_2013`

training set (41,593 utterances, approx. 66 hours of audio/speaker[11] and one novel (x 3 speakers) was selected for testing (6838 utterances, approx. 11 hours of audio per speaker). The corresponding transcriptions were obtained from Project Gutenberg.[12] Transcripts were split into chapters and then sentences using *Chapterize* (Reeve 2016) and *SpaCy* (Honnibal et al. 2020). The audio files were segmented into utterances using the Aeneas library[13] and phoneme alignment was obtained using the Montreal Forced Aligner McAuliffe et al. 2017. The average sentence length is 16.8 words.

5.2.2.2 Prosodic feature extraction

The audio files were analyzed with the continuous wavelet transform (CWT) method, described in Section 5.1.1.2 and Suni et al. 2017, as implemented in the Wavelet Prosody Toolkit.[14,15] The continuous values obtained from this procedure are quantized into three categories: <p2> (strong prominence), <p1> (intermediate prominence) and <p0> (no prominence).[16] A three-way quantization was selected to reflect an unaccented, accented, and emphasized distinction (Badino et al. 2012 showed that a binary distinction between accented/non-accented words had no effect on perceived TTS quality, but a three-way distinction that included an emphatic/contrastive level, did improve subjective scores.)

As has been shown in previous studies (e.g., Duběda and Mády 2010) the tendency for prominent, nuclear accents to appear at the ends of intonational phrases is also prevalent in our training and test corpora. Figure 5.3 shows the percentage of <p0>, <p1> and <p2> tagged words grouped by train/test set and sentence position (we use punctuation as a proxy for the end of prosodic utterances). This bias in the data is another feature that naive (i.e., discourse unaware) prominence prediction models could exploit to achieve a decent baseline.

5.2.3 Contrastive personal pronoun subcorpus

With the processed data from the previous section, we searched for utterances containing <p2> labelled personal pronouns (strong prominence) in the test set. With manual verification, we collected positive examples of contrastive focus on pronouns. We randomly selected an equal number of negative samples where the pronoun was tagged as <p0> for all three speakers; these samples were also manually checked.

We then enlisted three native English speakers to validate 200 pronouns (x 3 speakers) from the collected samples (100 positive samples and 100 negative samples, just over 20% of the full pronoun corpus). Before evaluating the corpus, validators underwent a short training

[11] Two of the five book sets are single speaker. The third set is made up of multiple speakers.
[12] `https://www.gutenberg.org`
[13] `https://github.com/readbeyond/aeneas`
[14] `https://github.com/asuni/wavelet_prosody_toolkit`
[15] Complete implementation settings are available at `https://tinyurl.com/4rt3sa2f`
[16] Dataset available at https://doi.org/10.5281/zenodo.6646827.

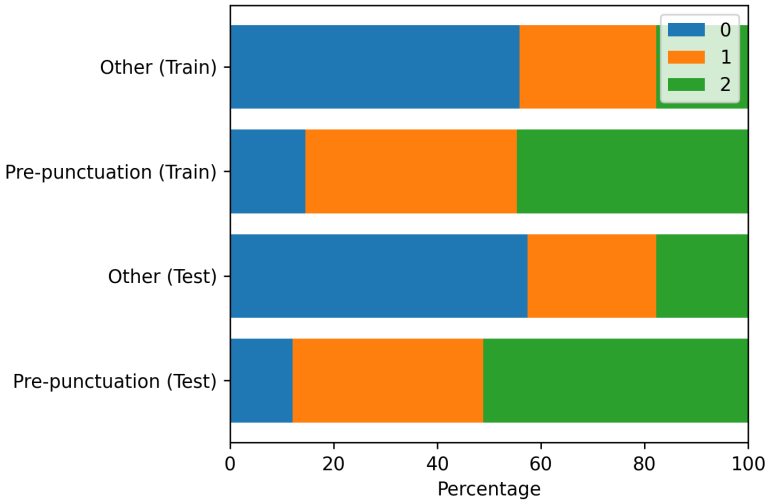


Figure 5.3: Percentage of prominence tags (0, 1, 2) for the training and test sets grouped by words immediately preceding punctuation marks and other.

session to ensure they understood what was meant by contrastive focus. They were told they needed to answer the following question: "Does the speaker use prosody (changes in pitch, duration and energy) to evoke alternatives/highlight a contrast between entities?" They were also presented with a series of example sentences similar to those in Table 5.1 to further clarify the task. In the evaluation phase, validators were presented with the audio clips (for all three speakers) and a transcription of the text with the pronoun of interest highlighted in red. They were asked to assign a value of 1 if they deemed the speaker had used prosody to convey contrastive focus and 0 if they did not. Cohen-kappa scores (Cohen 1960) were used to evaluate inter-annotator agreement between the three raters. These scores range from 0.85–0.90; this shows strong to almost perfect agreement.

For evaluation purposes, we sorted the positive samples into two groups: (1) those where the majority of speakers (at least 2 out of 3) used contrastive focus (**Pronoun maj.**). This group contains 393 pronouns; 310 of which are prosodically contrasted by all three speakers; (2) those where only 1 out of 3 speakers used contrastive focus (**Pronoun min.**). This group contains 100 pronouns. All contrastive pronouns come from 406 utterances as several utterances contain multiple examples.

In this study, our intention is to find challenging examples for prominence prediction (i.e., words that are not frequently prominent). Subjective pronouns (e.g., I, we), objective pronouns (e.g., me, us) and possessive determiners (e.g., my, our) all fit this requirement, but possessive pronouns (e.g., mine, ours) are more often prominent than not. We decided to include these "easy" words in the corpus because they may be of interest for future work on contrastive focus. However, they do not present a particular challenge to prominence prediction.

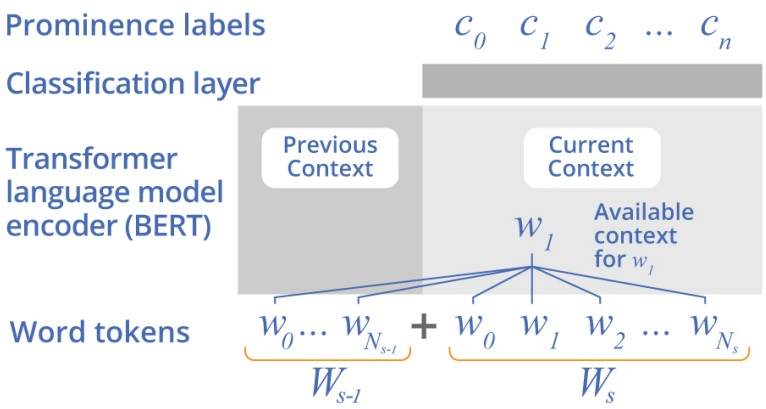


Figure 5.4: Module 1: Predicting prominence. The figure represents the bidirectional encoding that is used by BERT. In the case of GPT-2, only previous words and the lookahead are encoded.

5.2.4 Predicting prominence

Prediction task. For the sequence of words $\{w_0, w_1, ..., w_N\}$ our objective is to predict a sequence of prominence labels $\{c_0, c_1, ..., c_N\}$, where c_n is either <p0>, <p1> or <p2>. Since knowledge about the previous context is sometimes essential for determining whether a word should be contrastive or not, we experiment with models conditioned on different degrees of past context. We note $W_s = \{w_1, \dots, w_{N_s}\}$ the sequence of N_s words in the current sentence s which we want to synthesize. The LM is given either the current sentence only, i.e., W_s, or both the previous and the current sentences, i.e., $\{W_{s-1}, W_s\}$, or both the two previous sentences and the current one, i.e., $\{W_{s-2}, W_{s-1}, W_s\}$. See Figure 5.4 for an illustration of the model's components.

Only the encoded representation of the current sentence is passed to the classification layer. Since "current sentences" for one training sample also serve as "previous sentences" in other training sample, we did not want the model to become overly familiar with specific sentences (i.e., memorize the prominence features), while ignoring the influence of past context. So when a sentence is part of the past, the model is not being optimized to predict its prominence features, but rather to maximize the contribution of the past to the prediction of the current sentence's prominence features.

5.2.5 Models and linguistic knowledge.

We evaluate three methods for prominence prediction with increasing access to linguistic knowledge:

- Our baseline is a simple word majority method: word statistics from the training corpus are computed; we count how often a lexical item belongs to each of the three prominence

categories and the majority category is used for all predictions in the test set.

- The second method involves the use of the BERT architecture, but instead of using weights pretrained on a masked language modelling task, we randomly initialize the model and train it to predict the prominence labels for each word in the input sequence. This model can presumably learn the same word statistics available to the word majority method and additionally the self-attention layers and positional embeddings provide the model with information about the surrounding lexical items and the positional context of each word. We expect this model will be able to learn canonical patterns of English prosody (i.e. that prominent, nuclear accents are typically found at the end of an intonational phrase) even if the semantic knowledge about the content of the sentences will be non-existent or at best, very naive.

- The third method involves finetuning a pretrained LM on the prominence features. From the beginning of training, this model has access to the syntactic and semantic representations learned from training on large amounts of textual data; during the finetuning process it must find the optimal way to use this information for the prosody/prominence prediction task. We use both a bidirectional model (BERT[17]) and a unidirectional model (GPT-2[18]). Fine-tuning is performed on all layers using a learning rate of = 5e−5. A softmax layer is added to both models for prominence label prediction.

Lookahead for incremental prominence prediction To investigate the effects of lookahead on prediction accuracy when using GPT-2 as an encoder, we modify the model architecture slightly (the process is illustrated in Figure 5.5): after encoding the words in the input text with GPT-2, we duplicate the resulting vector sequence, once for every value of m in the set $\{1, 2, 3, \ldots, k-1, k\}$ (where k is the total lookahead). For each duplication, we remove the first m vectors from the beginning of the sequence and add m zero vectors to the end of the sequence. Then, to replicate the backwards pass of a bidirectional LSTM, we feed the sequence w_{n+k}, w_{n+k-1}, ..., w_n into a unidirectional LSTM and take the last hidden state. It is important to pass the sequence in backwards, so that the last hidden state corresponds to the current word (word identity is an essential feature for predicting prominence and if the sequence is fed in forward, the LSTM must learn to recall the identity of the word k-words back). The hidden state is then passed to a linear layer which predicts the value of prominence.[19] We test models for values of $k = 0, 1, 2, 3, 4, 9, 17$ (up to the average sentence length).

Table 5.2: Classification results for the prominence prediction task for the <p2> (high prominence) category. Recall (R), Precision (P) and F1 are reported for the full test set (all POS categories combined). Models that use the same architecture and lookahead are colour-coordinated in the *Current sentence*, *+1 Previous sentence* and *+2 Previous sentences tables*.

Model	**Current sentence**				
	F1	**R**	**P**	**#<p2>**	**# Tokens**
Word majority	0.458	0.394	0.546	79,793	365,330
Randomly initialized BERT	0.553	0.497	0.625	79,793	365,330
Fine-tuned BERT-uncased	0.603	0.566	0.646	79,793	365,330
Fine-tuned GPT-2 k=0	0.448	0.361	0.591	79,793	365,330
Fine-tuned GPT-2 k=1	0.582	0.554	0.614	79,793	365,330
Fine-tuned GPT-2 k=2	0.595	0.581	0.611	79,793	365,330
Fine-tuned GPT-2 k=3	0.598	0.576	0.622	79,793	365,330
Fine-tuned GPT-2 k=4	0.600	0.589	0.622	79,793	365,330
Fine-tuned GPT-2 k=9	0.602	0.574	0.633	79,793	365,330
Fine-tuned GPT-2 k=17	0.602	0.566	0.643	79,793	365,330

Model	**+1 Previous sentence**				
	F1	**R**	**P**	**#<p2>**	**# Tokens**
Randomly initialized BERT	0.572	0.567	0.578	79,793	365,330
Fine-tuned BERT-uncased	0.595	0.543	0.658	79,793	365,330
Fine-tuned GPT-2 k=0	0.560	0.553	0.568	79,793	365,330
Fine-tuned GPT-2 k=1	0.589	0.576	0.603	79,793	365,330
Fine-tuned GPT-2 k=2	0.595	0.580	0.612	79,793	365,330
Fine-tuned GPT-2 k=3	0.594	0.571	0.620	79,793	365,330
Fine-tuned GPT-2 k=4	0.597	0.572	0.625	79,793	365,330

Model	**+2 Previous sentences**				
	F1	**R**	**P**	**#<p2>**	**# Tokens**
Randomly initialized BERT	0.505	0.427	0.620	79,793	365,330
Fine-tuned BERT-uncased	0.599	0.555	0.650	79,793	365,330
Fine-tuned GPT-2 k=0	0.561	0.552	0.570	79,793	365,330
Fine-tuned GPT-2 k=1	0.586	0.562	0.612	79,793	365,330
Fine-tuned GPT-2 k=2	0.586	0.557	0.617	79,793	365,330
Fine-tuned GPT-2 k=3	0.587	0.559	0.617	79,793	365,330
Fine-tuned GPT-2 k=4	0.594	0.569	0.621	79,793	365,330

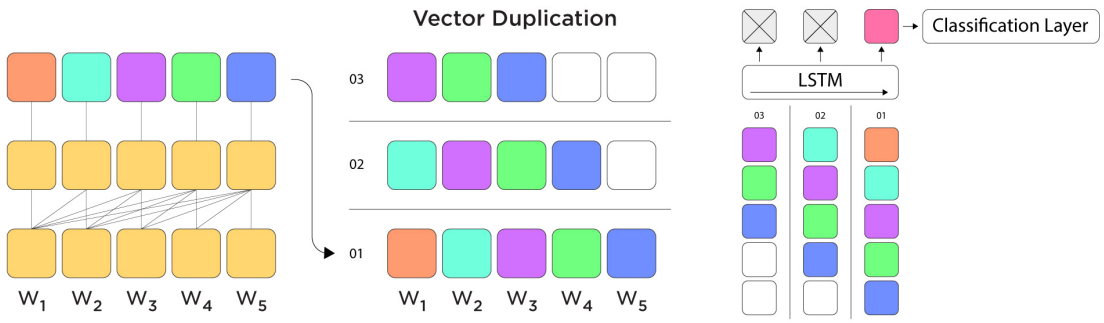


Figure 5.5: Incremental prominence prediction with lookahead. Words are first encoded with GPT-2 (left panel). The word embeddings are then duplicated for each degree of lookahead (middle panel). Finally, each lookahead sequence + current word embedding are fed to an LSTM. The output of the LSTM is used to predict the level of prominence (right panel).

5.2.6 Results

All POS categories The results for $< p2 >$ classification for the full dataset (all POS categories) are shown in Table 5.2. The fine-tuned BERT with no previous context achieves the highest F1 score, with GPT-2 k=9 and k=17 (also with no previous context) close behind. The weakest scoring models are the two baselines (Word Majority and Randomly initialized BERT) and the GPT-2 k=0 with no previous context. To test the statistical significance of LMs (vs. baselines), extended context and lookahead for prominence prediction, we use 10-fold cross validated paired t-tests with Holm-Bonferroni correction. We compare both global F1 scores (weighted for the three prominence categories), as well as the F1 scores for the strong prominence ($< p2 >$) category alone.

Baselines: The BERT and GPT-2 models all outperform the baselines in the current sentence condition, with the exception of the GPT-2 k=0 models. **Extended context:** Global F1 scores show no statistical significance between the pretrained BERT models with different degrees of past context, but for the $k = 0$ incremental model improves when adding the previous sentence. Especially at the beginning of a sentence, this model has very little context available on which to base its predictions; expanding the window of text appears to be beneficial. More surprisingly, the other incremental models suffer from previous context with degraded global scores for k=1,2,3,4 in the +1 and +2 *Previous* conditions compared to the current sentence condition (there is no difference between the +1 and +2 conditions), although the k=1 +1 *Previous* is more accurate for the $< p2 >$ category specifically.

[17]https://huggingface.co/bert-base-uncased
[18]https://huggingface.co/gpt2
[19]In this experiment, lookahead is counted in the number of tokens which may not always correspond to complete words (due to byte-pair encoding). This decision was made to simplify model training which includes batching and requires sequences of equal length. Furthermore, when evaluating prediction accuracy, only the final token in a word made up of multiple tokens is considered.

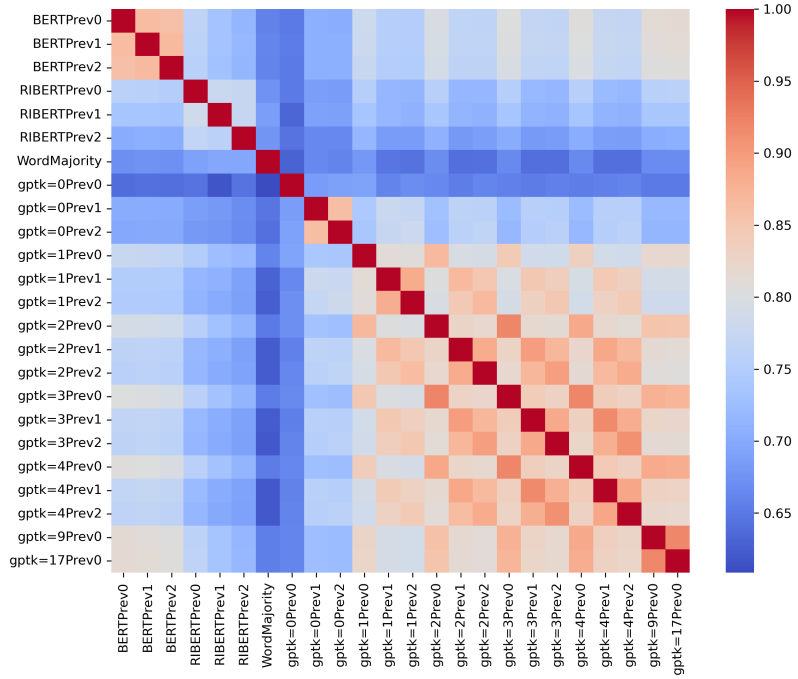


Figure 5.6: Jaccard similarity scores between tested prominence prediction models on the full test set for all levels of prominence (<p0>, <p1>, <p2>). High scores (red) reflect high similarity in prediction behaviour between the two models.

Lookahead: As for lookahead, we again see improvement when passing from k=0 to k=1, and also from k=1 to k=2 (both globally and for $< p2 >$). There is no statistical difference between k=2, 3 and 4, but there is one between k=2 and k=9 and 17 for $< p2 >$ prediction. We also see a significant difference between k=3 and 4 and k=17 global F1 scores. In summary, prediction accuracy gradually increases with increased lookahead.

We further compare model similarity with Jaccard scores (Jaccard 1908) calculated on the test set. The results are displayed in Figure 5.6. The prediction behaviour of the incremental models gradually moves closer to the bidirectional model as the amount of lookahead increases, however even at $k = 17$ when F1 scores are not statistically different between BERT and GPT-2, the two types of models exhibit different prediction behaviour, perhaps reflecting differences in the LM-encodings.

Personal pronouns The recall scores for the full pronoun subset are shown in Table 5.3. Analyzing these results, we notice that performance on the contrastive pronoun sets is significantly lower than the full dataset (Recall with fine-tuned BERT is 0.318 for the pronoun majority group and 0.566 for all POS). Furthermore, the tokens in the pronoun minority group are rarely predicted to be prominent (highest recall score= 0.14 by the GPT-2 $k = 4$ model).

We again compare significance between our various models using 10-fold cross validated paired t-tests with Holm-Bonferroni correction. Here we compare global F1 scores for the

combined *Pronoun majority*, *Pronoun minority* and *Pronoun negative* sets. The pretrained BERT and GPT-2 class of models, with the exception of GPT-2 $k = 0$ models, result in higher prediction accuracy than the *Word majority* and *Randomly initialized* baselines. There is however no significant difference between the LM-informed models on this subset, again with the exception of GPT-2 $k = 0$ and also the fine-tuned, +2 *Previous* BERT model which have a worse performance than the other models.

The fact that we see a large jump in accuracy from $k = 0$ to $k = 1$ models is likely an indication that the gains achieved by using LMs is not a result of the contribution of discursive knowledge (as knowledge of discourse is not dramatically different between these two conditions). The gains are more likely attributable to syntactic and semantic features in the local context.

The predictions made by the various models are not always interpretable, but we do see some trends in the strategies employed by models with different levels of linguistic knowledge: The word majority method correctly predicts possessives (e.g., mine, yours); the randomly initialized BERT learns structurally prominent positions; it often correctly predicts prominence at the ends of prosodic phrases (immediately preceding punctuation marks). The poor performance of the GPT-2 $k = 0$, current sentence model can in part be explained by this missing punctuation information. And the LM-informed predictors make more advanced use of syntax, for example pronouns followed by gerunds are predicted to be prominent (e.g., *Till she recollected that HIS being the intimate friend ...*).

So while the LMs provide some improvement over baseline models, prediction accuracy is still fairly low on these pronoun examples. We can imagine several possible causes for this, beyond the inherent difficulty of the task. It may be that we have an insufficient number of samples of contrastively focused pronouns to train the model to recognize focus patterns; there is an average of 7893.6 <p2> labelled personal pronouns/speaker in the training set, and given the complexity of the task, this may not be enough. The LMs, during pretraining, may have learnt a representation of contrast, but the finetuning dataset may be too small to connect this representation to the new multimodal setting. Alternatively, learnt representations may not be sophisticated enough to encode the higher level linguistic information required for the discursive element of this task. The recent research trend in language modeling is to scale models bigger and bigger and this increase in size results in better quality on tasks such as text generation. Replicating this experiment with larger LMs (e.g. GPT3 or 4) is a perspective of this work. Finally, as can be expected with any automatic annotation method, there is some noise in the data: we did find examples for which prominence was questionable, predominately at phrase boundaries, where boundary tone features may be mistaken for strong prominence (i.e., words are tagged <p2> because of a sharp rise in f0). Hence, human intervention may still be necessary for better fine-grained annotation/control of prosodic data.

Table 5.3: Results for pronoun subset on the prominence prediction task. Recall is reported for the two manually verified subsets of contrastively focused personal pronouns and the non-contrastive pronoun subset (Neg: 493 samples). Maj: group of 393 samples where majority of speakers used contrastive focus. Min: group of 100 samples where only 1 out of 3 speakers used contrastive focus. (Cat. = Category)

Model	**Data**	**Cat.**	**R**	**R**	**R**
Context			**Current**	**+1 Previous**	**+2 Previous**
Word	Maj	<p2>	0.079		
Majority	Min	<p2>	0.000		
	Neg	<p0>	1.000		
Randomly	Maj	<p2>	0.186	0.216	0.087
Initialized	Min	<p2>	0.110	0.082	0.020
BERT	Neg	<p0>	0.990	0.976	0.947
Fine-	Maj	<p2>	0.318	0.305	0.277
tuned	Min	<p2>	0.090	0.080	0.070
BERT	Neg	<p0>	0.984	0.986	0.986
GPT-2	Maj	<p2>	0.099	0.255	0.262
k=0	Min	<p2>	0.090	0.090	0.100
	Neg	<p0>	0.935	0.931	0.943
GPT-2	Maj	<p2>	0.303	0.318	0.326
k=1	Min	<p2>	0.170	0.130	0.140
	Neg	<p0>	0.972	0.968	0.960
GPT-2	Maj	<p2>	0.310	0.328	0.316
k=2	Min	<p2>	0.090	0.090	0.120
	Neg	<p0>	0.980	0.974	0.966
GPT-2	Maj	<p2>	0.310	0.338	0.328
k=3	Min	<p2>	0.130	0.120	0.130
	Neg	<p0>	0.984	0.972	0.964
GPT-2	Maj	<p2>	0.328	0.331	0.313
k=4	Min	<p2>	0.140	0.140	0.140
	Neg	<p0>	0.982	0.968	0.968
GPT-2	Maj	<p2>	0.310		
k=9	Min	<p2>	0.120		
	Neg	<p0>	0.978		
GPT-2	Maj	<p2>	0.298		
k=17	Min	<p2>	0.120		
	Neg	<p0>	0.974		

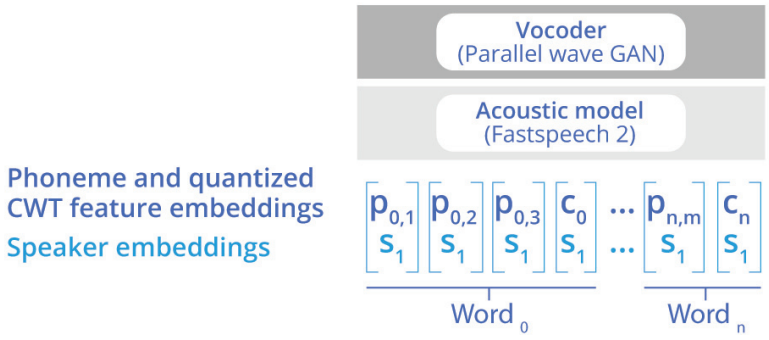


Figure 5.7: Module 2: Controlling prominence. The TTS model is conditioned on sequences of phoneme embeddings (p) and prominence labels (c), concatenated with speaker embeddings (s).

5.2.7 Controlling prominence

5.2.7.1 Controllable TTS model

To synthesize speech with controllable prominence, we follow the method proposed in Suni et al. 2020 where the TTS is conditioned on prominence labels. Our implementation differs in that we use FastSpeech 2 (Ren et al. 2021) (as implemented by Wang et al. 2021a) instead of DC-TTS and we refrain from using boundary tags as we are primarily interested in prominence control for this experiment. FastSpeech 2 is a non-autoregressive, transformer encoder-decoder model that makes explicit duration, f0 and energy predictions between the encoder and decoder. The input to our model is a sequence of phonemes and prominence labels. Each word w_n in the utterance is converted into a sequence of phonemes $\{p_{n,0}, ..., p_{n,m}\}$ and this phoneme sequence is followed by the prominence label c_n for w_n (phonemes and prominence labels are converted into embeddings in the first layer of the model). The output of the modified FastSpeech 2 model is a Mel-spectrogram, which is converted into a waveform using a Parallel WaveGAN vocoder (Yamamoto et al. 2020).[20]

To train and test the TTS model, we use the data described in Section 5.2.2, but only for a single speaker (Blizzard Challenge 2013; this speaker has read all 6 books). While the training corpus is read by a single speaker, this speaker portrays several different characters with different accents and pitch ranges. This tends to introduce fuzziness into the synthetic speech. To help the model learn these characteristics and improve quality, we added a speaker embedding to the TTS input. To obtain this embedding, we (1) encoded each utterance in the training set with a pretrained speaker identification model (ECAPA-TDNN Desplanques et al. 2020 available at Ravanelli et al. 2021), (2) used k-means clustering on these embeddings to obtain 30 different 'speakers' and (3) used these speaker labels as an additional input to FastSpeech 2; the speaker embeddings are concatenated with the phoneme and prominence label embeddings.

[20] https://github.com/kan-bayashi/ParallelWaveGAN

5.2.7.2 Listening test

To test the controllability of our TTS model in terms of prominence, we conducted an ordinal ranking listening test using the Web Audio Evaluation Tool (Jillings et al. 2016), following this procedure: (1) we randomly sampled 100 utterances containing pronouns from our full test set (not solely from the contrastive subset); (2) from this selection, we took the first 10 utterances containing a subjective pronoun, the first 10 with an objective pronoun and the first 10 with a possessive determiner (for a total of 30 utterances); (3) using our pretrained TTS system, we synthesized three versions of each utterance, changing only the prominence label for the relevant pronoun ($c \in \{$<p0>,<p1>,<p2>$\}$). The prominence labels for all other words in the sentence were kept constant with the ground truth values (extracted from the original audio with the CWT method); (4) 30 native English evaluators, recruited on Prolific,[21] were presented with the three versions of the synthesized utterances (in random presentation order) and a transcript of the audio with the pronoun of interest in uppercase letters. Participants were asked to rank the prominence of the pronoun by dragging and dropping the movable audio clips so that they were arranged from most prominent to least; (5) clips ranked as most prominent are assigned a score of 1; clips ranked as second most prominent are assigned a score of 0.5 and clips ranked as least prominent are given a score of 0. Sample audio files are available at `https://bstephen99.github.io/iTTS/interspeech2022/interspeech2022.html`.

Results. The results of the listening test are shown in Figure 5.8. We see the median values align with the prosody labels used (median: <p0>= 0, <p1>= 0.5, <p2>= 1.0). We do however see a wide distribution in the responses. This, and the examination of the ratings for individual utterances, indicates that this method works, but not consistently (i.e., there are some utterances for which the evaluators could not detect a difference). The only pronoun category for which we see a fairly clear distinction between <p0>, <p1>, and <p2> in perceived prominence is for possessive determiners. This may be because there is more natural variation within the training corpus for this category. Or, it may be due to the labelling errors at phrase boundaries discussed in the previous section: the sampled subjective and objective pronouns were found more often at the beginning and end of phrases than the possessive determiners. More work is thus needed to disentangle the global prosodic representations from that of individual words, but this separation is difficult because it may result in less natural utterances.

5.2.8 Summary and perspectives

In this section, we investigated the task of prominence prediction for all POS categories and for a subset of contrastively focused personal pronouns. We also tested the prosodic control of personal pronouns in TTS.

Prediction: we compared models with varying degrees of access to linguistic knowledge, past context and future context on a word-level prominence label prediction task. We found

[21] `https://www.prolific.co`

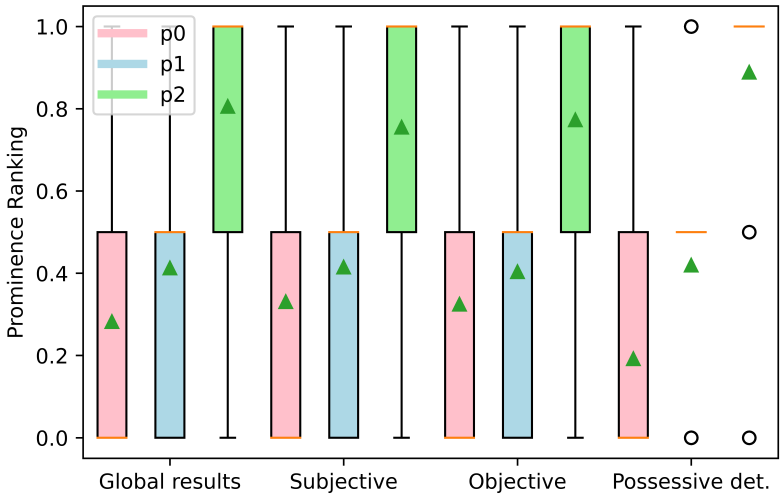


Figure 5.8: Results from the prominence ordinal ranking listening test. Audio ranked as most prominent was assigned a score of 1, second most prominent 0.5 and least prominent 0. Orange lines show the median and green triangles show the mean.

that fine-tuned LMs could improve prominence prediction over baselines, but it is unlikely the models of the size we tested are contributing deep understanding of the discourse to the prediction task. In the future, investigating the use of word representations from larger LMs with more sophisticated linguistic understanding would likely yield better results. Anecdotally, when asking ChatGPT to explain the contrast present in Sentence 5 of our corpus examples (*I have nothing to say against HIM* + extended context), it returns a correct, coherent response, demonstrating an "understanding" beyond that possible by BERT and GPT-2: (*In the given passage, the word "HIM" is being contrasted with the warning given by the speaker to be on guard and not involve oneself with someone who lacks fortune, as it would be imprudent. The speaker is acknowledging that there is nothing negative to say about this "interesting young man", but is still advising caution in pursuing an affectionate relationship with him due to financial considerations.*).

Control: with a perceptive test, we evaluated the control of prominence on pronouns in a TTS model conditioned on prominence labels. The results show the model is able to provide some control but the performance is not consistent over all samples. It may be possible to achieve better control of these marked prosodic patterns if we can take into consideration expected prominence values for unmarked prosodic utterances when labelling the data. In the present setup, for the majority of utterances, the labels we obtain through CWT will align with the expected values for a given sentence/syntactic position (i.e., will not diverge from the common patterns that would be learnt naturally (without the labels) by a vanilla TTS model). Due to this overlap, the model may struggle to learn/disentangle the contribution that the prominence labels provide to the input. By incorporating a vanilla model into the labelling process, the role of the labels could be made more explicit and possibly improve control (e.g., if the ground-truth value is more/less prominent than the vanilla prediction, a +1/-1 label could be used to condition a new controllable TTS model).

Natural variation: we used multiple spoken versions of the same written text to see the agreement among speakers on the use of contrastive focus. But we must keep in mind that while the textual context remains the same, the interpretation of the text can vary. For example, we infer that one of the speakers interprets some of the characters in the novel to be passive aggressive and they convey this through the frequent use of contrastive focus on 'I' (e.g., **I** am going to Gretna Green (intended meaning: and **YOU** are not). Removing the contrastive focus here is not wrong, but gives a very different impression of character/situation/relationship. This illustrates the difficulty of the prediction task and therefore, depending on the intended usage of the TTS system, it might be fruitful to explore other sources of input to the prominence prediction model (e.g., the source speech in a speech-to-speech translation system) in order to be as faithful to the intended meaning as possible.

5.3 Boundaries

Incremental text-to-speech requires decisions about how to divide up the speech signal to output chunks of speech incrementally. Humans naturally package words into meaningful units/phrases to make their message digestible in normal continuous speech. We hypothesize that listeners will prefer to hear natural groupings of words in the incremental setting and we test different methods for obtaining these groupings. We look at both user preferences and the cognitive load effects of these different segmentation styles.

The iTTS method we propose (see Figure 5.9), consists of two modules (similar to the prominence prediction pipeline in the last section): the first module processes the text stream as each new word becomes available; it is an LM fine-tuned to either predict (a) simplified POS tags which can be used to make boundary decisions (the *Chinks 'n Chunks* method (Liberman and Church 1992)) or (b) CWT boundary features (Suni et al. 2017). If a boundary is not predicted after the current word, the word is added to a buffer. If a boundary is predicted, then the current word and all the words stored on the buffer are sent to Module 2. Module 2 is a neural TTS conditioned on boundary features that can control pre-boundary lengthening at the end of the synthesized unit. In addition to the LM-informed chunking strategies, we also test more reactive, less computationally expensive count-based methods (i.e., one/two-word(s)-at-a-time).

5.3.1 Speech segmentation

Adaptation for large input latency conditions Few previous works have investigated the size/selection of output units for iTTS. Yanagita et al. 2019 found that synthesizing two or three word chunks at a time was equal in quality to full-sentence quality. This work however assumed that the input latency would be minor and that the output speech could play more or less continuously; so while synthesis was done in chunks incrementally, the audio was evaluated as a connected whole (which may be realistic for speech-to-speech translation or dialogue systems). What we are interested in is evaluating the case where the input latency can

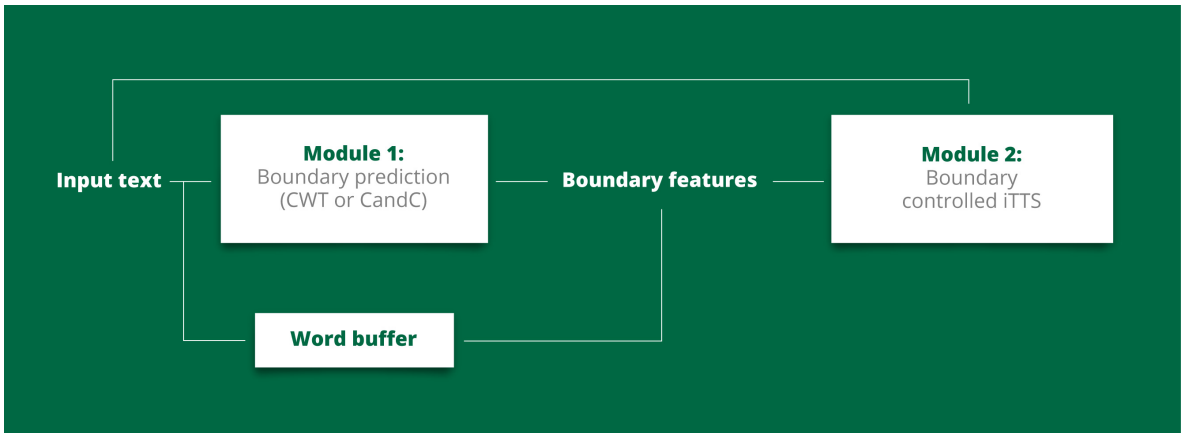


Figure 5.9: Boundary controlled speech synthesis: in this pipeline, the first module is responsible for predicting whether or not there will be a boundary after the current word. The second module is responsible for speech synthesis. It is conditioned on boundary tags.

potentially be very large and where there will be gaps between speech segments: Alternative and assistive technologies (AACs). These are applications whose users may suffer from physical or cognitive disabilities which make text entry particularly challenging. The work closest to that presented here is Pouget et al. 2016, where the stability of POS predictions was used to guide segmentation. In other words, if the predicted POS tag for the current word was unlikely to change with additional input, synthesis was triggered. A subjective test showed this strategy was preferred over a random one, but it did not rate as high as human-determined segments, showing there is room for improvement with richer syntactic and semantic representations guiding segmentation choices.

In this work, we make the assumption that the positions of prosodic boundaries are suitable locations to break up the speech stream. To find these boundaries in our training corpus, we extract CWT-derived boundary values (Section 5.1.1.2) from the audio signal; these features were shown to be successful at predicting human-annotated boundaries (85.7% accuracy) on the Boston News Corpus (Ostendorf et al. 1995; Suni et al. 2017). We use these values to train both a prediction model (for one of our tested conditions) and a controllable TTS model (used to synthesize speech for all our tested conditions). The controllable TTS can be used to modify the prosody of predicted segments so that they exhibit pre-boundary characteristics, indicating the end of a chunk, rather than having abrupt interruptions in normal continuous speech.

At a normal speech rate, there is a prosodic hierarchy of major and minor phrases, each representing a different degree of juncture. We use the boundaries of both these domains (by setting a low threshold on the CWT values) as segmentation points in order to keep the model as reactive as possible while still trying to output meaningful/cohesive units. At synthesis time, minor boundaries are converted into major boundaries (i.e., characterized by stronger lengthening and a subsequent pause), much in the same way that prosodic phrasing changes at slower speech rates; at slow speeds, the number and strength of boundaries increases (Trouvain and Grice 1999; Fougeron and Jun 1998).

5.3.2 Cognitive processing of speech chunks

5.3.2.1 Chunks

Information processing units. The basis for our hypothesis that people will prefer to hear and find it easier to process chunks is based on the fact that these seem to be a natural unit of information processing/storage in the human brain. For example, Johnson 1965 studied transitional error probabilities (i.e., in a recall task, the probability of not recalling the next word in a sentence given the current word was recalled correctly). This study showed that transitional errors between phrases was higher than those within phrases, suggesting that phrases were learnt/stored as a unit. Similarly, sequences of words are better recalled the closer they are to the standard phrasal patterns of the language of study, even if the sentences are nonsensical (Miller and Selfridge 1950).

The study of event related potentials (ERP) (Steinhauer et al. 1999; Steinhauer and Friederici 2001; Pannekamp et al. 2005) also show the cognitive reality of phrasing. Approximately 500ms after hearing a phrase boundary, there is a positive deflection in the brain which has been termed a closure positive shift (CPS). The nature (e.g., size and onset) of the CPS can be affected by both acoustic and syntactic features. This effect is not simply a reaction to pausing; the brain appears to distinguish between appropriate and inappropriate chunking. A study by Anurova et al. 2022 shows that pauses at the ends of phrases induce different reactions from those that appear mid-phrase, with mid-phrase breaks being interpreted as interruptions. This stresses the importance of finding appropriate chunks for iTTS, as inappropriate ones could give the sense of being constantly interrupted.

Unit size. Chunks are not fixed in length and several appropriate parses are possible for the same sentence (even when syntactic and semantic factors are held constant). These variations can reflect information structure differences or simply speech rate, as prosodic boundaries (as mentioned previously) become more frequent in slowed down speech. In this section, we look at the extreme case where the speech is so slow each word can potentially be presented as its own chunk. The comprehension of speech is usually facilitated by slower rates (e.g., Griffiths 1992), but if breaks are as frequent as after every single word, does this impose a cognitive penalty? Does the listener have to work harder to reconstruct the syntactic and semantic structure from singleton units? And is a similar penalty incurred if the grouping are not linguistically motivated (i.e., using random groupings or simple count-based strategies)? We test the hypothesis that the chunking of speech into larger, language-model informed units will help the listener better encode and understand speech. We do this with a sentence validation task: We present the subjects with a short passage which first provides some context and then follows it up with a sentence that is either coherent or incoherent with respect to the context (the coherence hinges on the final word of the passage). To complete this task successfully, the subject must understand and recall the earlier parts of the passage to compare them with the final proposition. We measure accuracy and reaction time on this task.

Some support for the hypothesis that single word units may degrade comprehension comes

from research on speech rate and online vs. offline processing of specific syntactic phenomenon, such as pronoun resolution ((Love et al. 2009; Nicol and Swinney 1989)), verb phrase ellipsis interpretation (Callahan et al. 2012) and gap filling (Love et al. 2008). For example, humans automatically resolve pronoun coreference online when the relationship with the antecedent is syntactically bound (e.g., local/non-local constraints on interpretation: *Sue asked Mary to help her/herself.*). When however the relationship is not constrained, more offline semantic and pragmatic processes must be utilized. At slower rates, it has been shown that automatic processes become degraded and more cognitively demanding strategies have to compensate (Love et al. 2009). It also becomes more difficult to resolve the syntactic roles of noun phrases in non-canonical sentences (e.g., object-relative clauses like *The man that the boy pushes...*) at slower rates (Love et al. 2008). While all the conditions we study here will be doled out at a slow pace (potentially affecting syntactic processing), it is hypothesized these effects will be exacerbated, and may extend to other syntactic dependency relationships, in the one-word-at-a-time condition.

5.3.2.2 Non-standard speech/incongruous prosody

Several works in psycholinguistics have demonstrated that the cognitive processing of speech can be hampered by the presentation of speech in a degraded or unnatural manner. For example, noise (Tun 1998), unusual intonational patterns (Braun et al. 2011), prosody/context mismatch (Nooteboom and Kruyt 1987) and accented (McLaughlin and Engen 2020) or synthetic speech (Govender and King 2018a) can all result in a higher cognitive load for the listener. We extend this work to the realm of speech segmentation.

Related works on cognitive processing and speech segmentation/phrasing have investigated incongruous phrasing that does not match the syntactic content of the sentences being investigated:

By looking at event related potentials, Pauker et al. 2011 and Bögels et al. 2013 studied the online processing effects of natural, missing and superfluous prosodic boundaries in early/late closure garden path sentences (e.g., *Whenever John walks the dog is chasing him*/ *Whenever John walks the dog the kids are chasing him*). Processing was easiest when a boundary separated the two clauses in the stimuli (the natural condition). For the distorted conditions, no boundary between the clauses was easier to recover from than the superfluous boundary condition.

Sanderman and Collier 1997 studied the effects of prosodic boundary placement on reaction times in a hybrid sentence verification/question answering task. Participants were presented with sentences, like those in (3), in one of three prosodic conditions relative to a contextualizing question: (1) a congruent phrasing structure (2) an incongruent phrasing structure, (3) a neutral phrasing. Participants were asked to press a button as soon as they could answer the question; reaction times were significantly slower when there was a mismatch between context and phrasing.

(3) a. How did I reserve a room?
I reserved / a room in the hotel / with a fax machine.
b. Which facility did the hotel have?
I reserved / a room / in the hotel with a fax machine.

Acceptable/unacceptable phrasing Phrasing differences like those in (3) represent alternative meanings. Meaning change however is not the only factor that can influence judgments of phrase acceptability. For example, *right annexation* (Taglicht 1998), where an argument is split from its head, is dispreferred by listeners (e.g., (4)a from Selkirk 2000), although the degree of unacceptability can be modulated by other concerns, such as the expectation that long utterances will be broken up into smaller units (Hwang and Steinhauer 2011). Information structure considerations also play a role in phrasing. Halliday 1967 posits that positioning a theme in a separate information unit builds the expectation that that theme will hold over the entire next information unit. So *John // saw the play and liked it* would be considered acceptable, but *John // saw the play and Mary went to the concert* would be unacceptable due to change in topic (*Mary*) part way through the unit. Focus also interacts with phrasing. The undesirable break between the verb and object in (4)a is rendered acceptable when the verb is in focus ((4)b). These general principles for well-formedness do not necessarily hold at slow speech rates, where there may be more flexibility. Our subjective experiments compare the level of acceptability produced by a range of boundary prediction methods at a slow rate of output.

(4) a. *She loaned // her rollerblades to Robin.
b. She LOANED // her rollerblades // to Robin.

5.3.3 Experiments

We seek to evaluate user preferences for different segmentation strategies and also the cognitive load they impose. We adapt boundary prediction models that have been proposed in full-sentence TTS and we test them in the incremental setting. We also compare these methods with more reactive ones (i.e., where the speech is generated as soon as it becomes available/or with a one word delay). Our research questions are the following: (1) Do listeners prefer to wait for longer (LM/linguistically-informed chunks) or do they prefer shorter wait times in a large input latency application? (2) Which boundary prediction method (a heuristic one or data driven one) provides more acceptable chunks according to subjective evaluation? (3) Does short/unnatural chunking incur a cognitive penalty for the listener?

We evaluate the first question with a modified MUSHRA test (with no anchor and no reference, since there is no ground-truth rendition with which to compare). Participants listen to multiple versions of the same passage (differing in segmentation) and they rate each sample, relative to the others, on a 100 point scale. Our second question is explored with a forced choice AB test; participants listen to two versions of the same utterance and select the version they prefer.

We selected a sentence verification task to study our third research question. This testing modality has previously been used to evaluate online processing effects of synthetic speech in Pisoni et al. 1987. Sentence/passage verification was chosen over other online measures, like word or phoneme monitoring tasks (e.g., Nix et al. 1993) where subjects listen for a specific word/phoneme in the stimuli and react once the target has been detected). Sentence verification was preferred because it requires the listener to contend with the meaning of the words and not simply to listen for specific sounds. It also requires them to recall earlier parts of the context to make their decisions on the coherence of the passage. Since in our iTTS application we would like listeners to be able to understand and remember elements of the ongoing (and slowly progressing) discourse, this seemed an appropriate measure. The requirement to understand the passage for successful task completion also provides a complementary measure to that provided by the MUSHRA and AB tests, where stylistic properties of the speech may have a stronger influence on listener choice than whether or not they have understood what is being communicated.

We devised twenty short passage pairs (with coherent and incoherent versions), where the final word determines the coherence of the sentence. The passages vary in terms of length (12 - 30 words, median length = 18.5) and syntactic structures (see 5.4 for an example passage and Appendix A for full list of passages). We use only the coherent versions of the passages for the MUSHRA and AB tests and both versions for sentence verification. The passages are relatively long compared to samples used in most synthetic speech subjective evaluations. This was deemed necessary to increase the segmentation differences between conditions and to provide sufficient context to compare the final sentence against.

Our proposed iTTS pipeline divides synthesis into two parts (see Figure 5.9). Module 1 predicts boundaries and Module 2 synthesizes chunks, imposing boundary features at the end of each chunk. All stimuli, for all our tested conditions, are synthesized with the same CWT-boundary conditioned model (Module 2). Module 1 varies according to the tested conditions. All models are described below.

5.3.4 Conditions and prediction models

5.3.4.1 Count-based: One/Two-word(s)-at-a-time

These segmentation strategies are the simplest to implement. Synthesis can be triggered after every word or after every other word. One-word-at-a-time (*One*) will incur the least wait time between audio segments and will thus be the most reactive method. Two-words-at-a-time (*Two*) requires a slightly longer wait period, but eliminates the isolated presentation of short function words which can be unnatural. This policy does not take into consideration the syntactic structuring or the semantic content of the input text and so the word groupings may be unnatural.

Conditions	Example Segmentation
One-word-at-a-time	Jimmy / and / Molly / got / married / the / week / before / last. / At / the / reception, / Jimmy / introduced / Molly / to / his / brother. /
Two-words-at-a-time	Jimmy and / Molly got / married the / week before / last. At / the reception, / Jimmy introduced / Molly to / his brother. /
Chinks 'n Chunks	Jimmy / and Molly / got married / the week before last. / At the reception / Jimmy / introduced Molly / to his brother. /
Continuous wavelet transform (CWT)	Jimmy and Molly / got married / the week before last. / At the reception / Jimmy introduced / Molly / to his brother.

Table 5.4: Examples of the speech segmentation resulting from the different evaluated techniques.

5.3.4.2 Language model guided

These segmentation strategies rely on the predictions made by an autoregressive LM (GPT-2), fine-tuned to predict boundary relevant features.

Chinks 'n Chunks (*CandC*) This method, originally developed in a non-incremental setting by Liberman and Church 1992, relies on the categorization of words into two groups: chinks which roughly correspond to function words and chunks which roughly correspond to content words (with some exceptions described below). This approach predicts boundaries where there is shift from a sequence of chunks to a sequence chinks.

To implement the *CandC* method in incremental mode, we fine-tune a GPT-2[22] model to predict a simplified set of POS tags (i.e., chinks and chunks). To train the model we use the English Universal Dependencies corpus (Zeman et al. 2017) which contains human annotated POS labels which we collapse into the two categories. Nouns, proper nouns, adjectives, adverbs and numerals are classified as **chunks**. Auxiliary verbs, determiners, coordinating and subordinating conjunctions are classified as chinks. Verbs and pronouns are split, with finite verbs and most pronouns counted amongst the chinks and non-finite verbs and objective pronouns with the chunks. The finite verb distinction reflects the fact that tensed verbs are more commonly grouped prosodically with their objects than with their subjects. Objective pronouns are differentiated because they typically end a functional grouping rather than start one. We adapt the original chinks and chunks method to split prepositions (chinks) from particles (chunks). This keeps phrasal verb components together as a coherent group (e.g., *I have filled in / the form.*). The incremental CandC tagger achieves 97.4% accuracy on the English Universal Dependencies test set (for comparison we also fine-tuned a BERT/full-sentence

[22] https://huggingface.co/gpt2

model on this data and it achieves 98.8% accuracy).

Provided accurate POS tags are available, the *CandC* method is easy to implement, but its decision making strategy will be shallow as it does not take into consideration the semantic content of the message. With ambiguous sentences like *The friends you praise sometimes deserve it* (where *sometimes* could modify *praise* or *deserve*), the algorithm will always group the ambiguous adverb with the first verb; given the ambiguity, this is as good a choice as any. But when presented with an unambiguous sentence *The friends you praise undoubtedly deserve it*, the adverb will still be grouped with the first verb, even though collocation statistics and semantic meaning should more naturally group it with the second.

Features derived from continuous wavelet transforms (*CWT*) This method utilizes boundary features extracted from the speech signal using a continuous wavelet transform (Suni et al. 2017). Similar to our *CandC* model, we implement the *CWT* method by fine-tuning GPT-2, but instead of POS tags, we predict CWT values (continuous values resulting from lines of minimum amplitude for each word in the sentence) with a regression training objective. We use the Wavelet Prosody Toolkit[23] to extract these boundary features. For configuring the CWT feature extraction, we used the parameter set originally proposed by Suni et al. 2017 and available here: `https://tinyurl.com/3ep8w529`.

Our aim is to find a segmentation method that is linguistically motivated but also one that provides relatively small chunks in order to meet the incremental/reactive concerns of our application. To select a threshold for splitting segments with CWT LomA values (the values in our corpus range from 0 to 1.44), we try to balance the number of long chunks (which would impose an unacceptable wait-time) with the number of single word 'chunks'. There are of course some natural single word chunks (e.g., sentential adverbs like *Hopefully,* - we find 2379 of them with the *CandC* segmentations), however, single word phrasing is not that common and so we select a threshold of 0.4 to balance the two extremes. Figure 5.10 shows the segment-length distribution in number of words for *CandC* and *CWT* (at thresholds 0.2, 0.4 and 1.0 for *CWT*) for our test corpus. The longest chunk in the test sentences used for evaluation is seven words long (*so as not to wake her family*), a series of mostly short words which does not inflict a typing delay beyond that of two or three long ones.

5.3.5 Speech synthesis

5.3.5.1 Models and training data

Speech synthesis Module 2 is used to synthesize the samples for all four test conditions. This controllable TTS unit is similar to that described in Section 5.2.7. We use a Fastspeech 2 model that is conditioned on CWT features, in this case boundary values (calculated from LomA). We quantize these values into ten different categories to train the model to learn a range of boundary strengths ($< b0 >$ for no boundary through to $< b9 >$ for a strong/utterance

[23] `https://github.com/asuni/wavelet_prosody_toolkit`

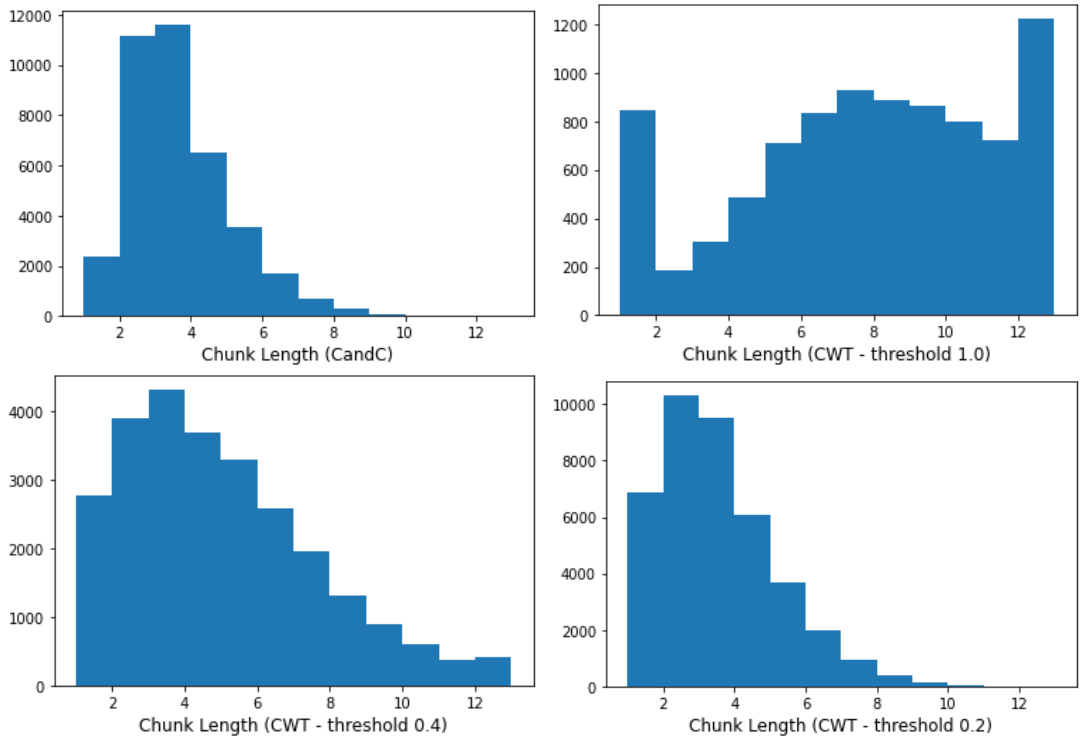


Figure 5.10: Distribution of chunk lengths from test corpus with segmentation methods (*CandC* and *CWT*). Chunk length is counted in the number of words.

final boundary). Each word in the input is followed by one of these tags. Ground-truth values are used at training time, and we use either language-model inferred values or count-based boundary assignment at inference. At inference, we limit our tags to three categories: $< b0 >$ tags are placed after words within chunks, $< b5 >$, a mid-range boundary, after words at the ends of chunks, and $< b9 >$ at the ends of utterances. The model is trained from scratch using the FairSeq S^2 (Wang et al. 2021a) implementation of Fastspeech 2.

In early models trained on boundary features, we noticed there were occasional distortions to the phoneme content (e.g., /g/ would change into /d/). In order to remedy this issue, we added an additional training objective to the standard Fastspeech 2: we encourage the model to keep phoneme representations distinct by reconstructing them at the output of the encoder. Encoder outputs are passed to a linear layer, the output of which is compared to the original phoneme embedding from the input using a mean squared error loss.

In addition to phoneme and boundary embeddings, we also include a speech rate value and speaker embeddings as input to the model. Speech rate is measured as the average phoneme duration over the utterance, measured in spectrogram frames. We concatenate this value to the phoneme embeddings. Ground-truth values are used for training and we use slightly slower than average speech rate at inference (12 frames/phoneme). Speaker embeddings (described in Section 5.2.7) are input to the decoder (as opposed to the encoder) in an effort to separate the modelling of prosodic phrasing from that of speaker characteristics; the model sees multiple

versions of the same utterance during training with variations in phrasing.

We use a HifiGan vocoder[24] (Kong et al. 2020) to convert Mel-spectrograms into waveforms. We train the model on time-aligned ground-truth waveforms from our speech corpus and synthetic Mel-Spectrograms inferred from the Fastspeech 2 model. The model was trained for 50,000 epochs.

Training data The training data and preprocessing are the same as those employed for the prominence task (the general corpus in Section 5.2.2) except in this case lines of minimum amplitude are used to extract boundary values. We also add some additional training data to improve the synthesis of the *One* condition. In connected speech, function words are usually pronounced in a weakened form (e.g., eɪ →ə) and they combine with more prominent lexical items to form prosodic words. Synthesizing normally reduced words independently gave poor results. To improve the quality, we added a recording originally intended for language learners, that contained basic phrases spoken one word at a time (Kendra's Language School 2021; Kendra's Language School 2022). This was an additional two hours of audio data.

Timing and lookahead To determine the timing between speech segments, we employ an average typing speed of 3 characters per second. Segments are vocalized when all characters in the segment + the next word of lookahead have been typed, or once previous backlogged segments have finished playing. Koester and Arthanat 2018 found that the average text entry rate for people with physical disabilities was 15.4 words per minute (1.2 characters per second). Our chosen average speed is significantly faster than this. This choice was made for practical testing purposes: otherwise the test samples would be prohibitively long. The selected rate allows for clear pauses between (most) segments allowing us to test disconnected speech, while at the same time being short enough to fit into a twenty minute evaluation period.

Due to a concern that study participants might prematurely think that the audio had ended when long pauses were encountered, we added faint typing noise (simulating the use case) to play continuously in the background of each of the stimuli. Samples are available at `https://bstephen99.github.io/iTTS/boundaries2023/boundaries2023.html`.

A $k = 1$ lookahead was selected to keep the reactivity high while also satisfying the constraints of our tested methods: the *CandC* method requires the next tag to be known to make a boundary decision.

[24] https://github.com/jik876/hifi-gan

5.3.6 Subjective evaluation

5.3.6.1 Method

To evaluate subjective opinion on the four segmentation strategies, we conduct two separate evaluations: (1) a modified MUSHRA test and (2) an AB forced choice test. The decision to do two tests was taken due to the length of the audio samples and concern for listener fatigue: In the MUSHRA test, we evaluated all four conditions while restricting the amount of audio quantity to approximately twenty minutes (10 audio clips $\times$ 4 version). In the AB test, we only looked at the two LM-guided methods, while doubling the number of audio clips (20 clips $\times$ 2 version). There is no ground-truth reference for these stimuli and thus this element is not included in the test.

Both tests were conducted online with the Web Audio Evaluation Tool (Jillings et al. 2016). Participants were recruited using Prolific (www.prolific.co); they were required to be native English speakers with no reported hearing issues. 26 subjects completed the MUSHRA test and a different set of 27 subjects completed the AB test. One of the participants from the AB test was removed because they finished the test too quickly. Participants in both tests were informed that they would be evaluating a TTS system for people who are speech impaired and they were asked to focus their evaluation on the word groupings.

5.3.6.2 Results and discussion

The results for the MUSHRA test are presented in Figure 5.11. We see there is a clear preference for longer chunks over shorter ones with both *CandC* and *CWT* receiving higher scores than *One* and *Two*. We use a Wilcoxon signed-rank test with Holm-Bonferroni correction to compare each pair of conditions. All pairs are statistically significant with the exception of *CandC* and *CWT* (p-value = 0.188). To control for rater differences in the use of the 100 point scale, we also compare the results of raw scores converted into rank order and again conduct Wilcoxon signed-rank test with Holm-Bonferroni correction. These significance tests align with the first set of results.

In addition to the LM-chunking preference, the *Two* condition is preferred over *One*, suggesting that larger chunks are preferred over shorter wait times, even when the groupings are random. However, we cannot know for certain whether this is a universal preference for longer chunks or whether the high quality chunks (*Two* will generate both acceptable and unacceptable chunks by chance), weighed more heavily on the ratings than the longer but more awkwardly segmented chunks.

The results for the AB test, for each of the tested passages, are displayed in Figure 5.12. We use a binomial test to compare user preferences globally and for individual samples. Globally, preferences are almost perfectly split with 49.6% for *CWT* and 50.4% for *CandC* (two-sided binomial p-value = 0.97). This is not to say that evaluators are insensitive to chunking differences, with four individual test items reaching statistical significance (Passages 1, 7, 13

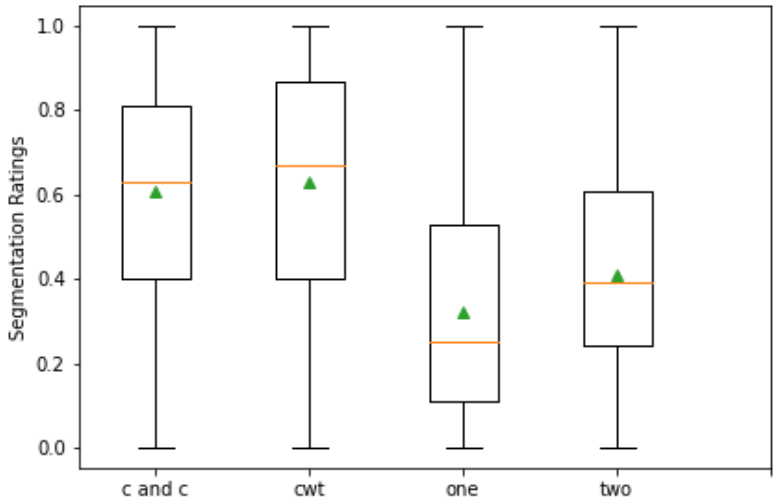


Figure 5.11: Mushra Results for segmentation strategies. The yellow bars show the median values and the green triangles show the mean scores.

and 18 : p-value= 0.04); but the preferred samples are evenly split across the two conditions. While the results of this test do not allow us to select a preferred method for chunk selection, qualitative analysis of the user preferences on these samples could lead to strategies for model improvements/areas of future research.

Compound nouns The *CandC* version is favoured for Passage 1. This is likely due to splits in the *CWT* version that separate compound nouns (*Mrs. // Jones* and *kindergarten // teacher*); the CWT tends to place boundaries after longer words. Separating arguments from their heads is known to be dispreferred in normal speed speech, however in our samples, this rule does not automatically determine the preferred version. In Passages 2 and 20, where the *CWT* and *CandC* scores are evenly split, direct objects are/are not separated from their verbs (e.g., *to pay // his bills*/*to pay his bills*). So while an argument-head rule may be relaxed at slow speeds, compound nouns may be a special case, where segmentation preference is more rigid because the two words are treated as a single unit.

Tagging errors For Sentence 13, the CWT parse is preferred, likely caused by a tagging error from the CandC tagger: *as* in the adverbial "Try as he might", is mislabelled as a chunk, causing the expression to be split in two *Try as // he might*. This error is possibly due to the infrequency of this type of syntactic construction (i.e., the base form of the verb followed by a conjunction) and could be overcome with more training data.

Short segments and timing *CandC* is preferred for Sentence 7; this may be because CWT splits a short, objective pronoun from its verb (*The dog chases // the cat // with the black eyes everyday. // He never catches // **it** // because // it is too fast*). There is a counter example

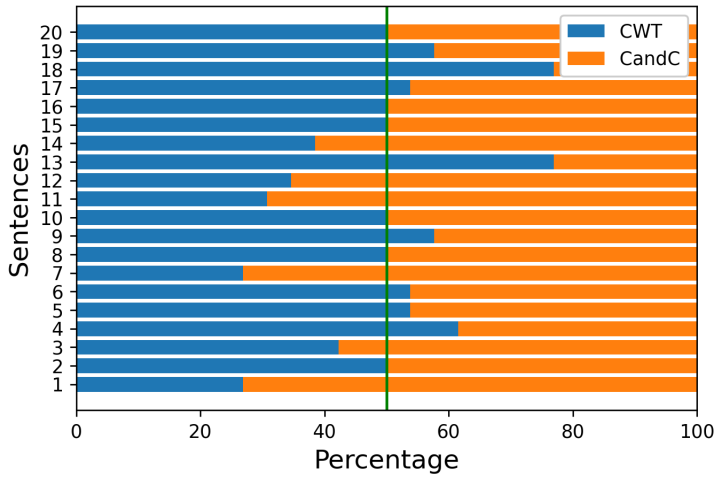


Figure 5.12: AB results for the 20 test sentences comparing segmentation strategies (Incremental Chinks 'n Chunks (*CandC*) and Incremental CWT-derived boundary feature prediction (*CWT*)).

for the averseness to short isolated words within our test sentences ((5) below, Sentence 19 in Figure 5.12 and the Appendix)): the CWT version is slightly preferred despite splitting the word *down* from the rest of its clause. This preference is however likely attributable to a timing issue with the *CandC* version; The conjunction *so* is played immediately after the first sentence finishes, leaving a long pause before the remainder of the sentence it is associated with. It is possible this chunk would be considered acceptable (as conjunctions making discursive links are commonly followed by a pause) if it were played closer in time to the second sentence. A more sophisticated trigger for synthesis (beyond playing a chunk as soon as it is available) is perhaps required to keep distinct linguistic units separate for the listener (e.g., requiring a minimum pause after sentence completion).

There are two differences between the segmentations of Sentence 18, where *CWT* is preferred: *CWT - John was late*, *CandC - John // was late* and *CWT - as the light turned green*, *CandC - as the light // turned green*. This could again be an issue of isolating short words (here *John* or there may be an aversion to splitting adverbial clauses. More controlled experiments are required to separate these two potential factors.

(5) a. ***CandC***
The house // remained on the market // for years // despite being in excellent condition. // **So (PAUSE)** // the sellers // decided to bring // the price down. //

b. ***CWT***
The house // remained // on the market // for years // despite // being in excellent // condition. **(PAUSE)** // So the sellers // decided // to bring the price // down. //

5.3.7 Sentence validation

5.3.7.1 Method

Test materials We use the same synthetic speech samples that were used for the subjective tests and also include the incoherent versions for each passage. We remove the background typing sounds and add babble noise to increase the baseline cognitive load required to complete the task. Babble noise, which results from the addition of a large number of individual speech signals, was selected to reflect a potential use-case environment for an AAC user (i.e., a cocktail party/crowded area). The speech signal $s(n)$ and the noise signal $b(n)$ are first normalized by their respective standard deviation to have a unitary power (i.e., a unitary variance). Then the noise is assigned a weighting factor α and summed to the speech signal to obtain the noisy speech signal $x(n)$:

$$x(n) = s(n) + \alpha \cdot b(n), \tag{5.1}$$

The factor α is set to control the signal-to-noise ratio R (in dB) of the resulting noisy speech signal as follows:

$$\alpha = \sqrt{10^{\frac{R}{10}}}. \tag{5.2}$$

Finally, $x(n)$ is normalized to have a unitary power.

The segmentation and coherence conditions were counterbalanced. Eight versions of the test were prepared. Four random sequences including ten coherent and ten incoherent conditions were generated for the first four tests; the inverse of these sequences were used for the second four. The ordering of segmentation conditions were determined with a latin square process[25] Four different condition sequences are repeated and combined with the eight coherence sequences to achieve even coverage of our stimuli. Following Nix et al. 1993 and Sanderman and Collier 1997, we maintain the same presentation order for the passages to reduce noise induced by participants answering rapidly and incorrectly at the beginning of the test and more accurately but slower towards the end.

32 native English speaking participants, with no reported hearing issues were recruited on Prolific. The test was implemented in Psycho Py (Peirce et al. 2019). Participants were first told they would be testing different delivery systems for people who are speech impaired and that they had to judge the coherence of passages as quickly as possible. In the training phase, they were presented with a text example of an incoherent sentence, its coherent form and an explanation for the incoherence. They then were presented with the same sentence in audio form, along with other practice sentences and were asked to press the right button on the keyboard if the sentence was coherent and the left button if it was incoherent. They then moved on to the test phase where they heard and evaluated each of the 20 test passages only once. The presentation of each new passage was triggered by the participant. Reaction times are measured from the onset of the target/final word of the passage. Participants had a maximum of 3 seconds to respond true or false.

[25] We rotate through the rows of a 4 x 4 latin square to obtain five blocks of four condition sequences (for our 20 stimuli). Four such sequences are produced by starting from a different row each time.

5.3.7.2 Results

We evaluate both the response accuracy and reaction times. Mixed effects models were used for statistical analysis (using the lme4 package (Bates et al. 2015)). The fixed effects used are the coherence and segmentation conditions (2x4 models) and the random effects are the passages and participants. Mixed effects models allow us to account for factors such as the length, syntactic complexity and predictability of individual stimuli and the reactivity/speed of different participants.

We begin by fitting generalized linear models to predict the binary response variable (Correct/Incorrect response), one for segmentation and coherence conditions and one for segmentation alone. We conduct an analysis of deviance comparison which reveals a significant interaction between coherence and segmentation ($\chi^2(4) = 16.733$, $p = 0.002178$). Figure 5.13 shows the response accuracy for the 640 collected responses, grouped by coherence and segmentation conditions.

We then fit a generalized linear mixed effects model to predict the response variable with random intercepts for stimuli and participant. We generate pairwise comparisons of the estimated marginal means between levels in the model (with single-step p-value adjustment) and this shows a statistically significant difference between *One* and *CandC* with an odds ratio estimate of 1.29 (SE = 0.38, z = 3.38, p = 0.01).

To study reaction time, we use only correct responses (460 in total). Here we fit a linear mixed-effects model to predict the log reaction time. Pairwise comparisons of the four segmentation methods (with Tukey adjustment (Tukey 1949)) reveal no significant differences.

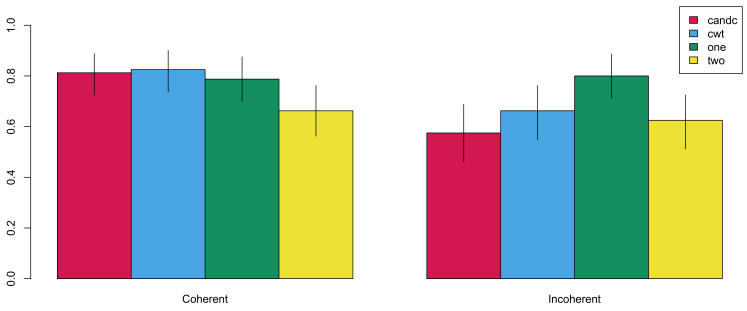


Figure 5.13: Proportion of correct responses on the sentence verification task for the four test conditions (Incremental Chinks 'n Chunks (*CandC*), Incremental CWT-derived boundary feature prediction (*CWT*), One-word-at-a-time (*One*), Two-words-at-a-time (*Two*). Error bars represent the 95% confidence interval.

5.3.8 Discussion and perspectives

The results of these tests do not support the hypothesis that chunked speech can facilitate cognitive processing. The response accuracy results, with more correct responses for *One* than *CandC*, actually support the inverse conclusion, but since we do not observe a difference between the other longer chunk condition (*CWT*) and *One*, this is unlikely to be the direct effect of the segmentation strategy. These results (or lack thereof) could stem from different causes:

(1) the sentence verification task may not be suitable for measuring online processing in this slowed down conditions. The final decision on the coherence of the passage must be delayed until the final word is spoken, however the resolution of earlier sentence structures could be carried out in an offline manner during the extended pauses and participants could be making predictions about upcoming content, reducing the processing effort when the final words are actually spoken. Other online testing paradigms (e.g., phoneme monitoring) may be required for more sensitive measurements.

(2) Other factors such as rhythmic regularity may influence accuracy and reaction time more than the effects of segmentation strategy. In this study, we applied a policy where speech was vocalized either as soon as it was synthesized or after the backlog of previous words had been vocalized. In the case of backlogs due to long words, there was sometimes an odd rhythmic pattern, where there was a long period of silence followed by two back-to-back chunks. Humans are very sensitive to rhythmic patterns in speech and English speakers have a preference for evenly spaced stress intervals. It is possible these irregularities had an affect on the reaction times. In future work, experimenting with constraints that attempt to evenly distribute chunks over a given time period could be fruitful.

(3) Very short segments may in fact be the most effective way to process speech, despite a stylistic aversion by listeners. Further testing that places iTTS users and their conversation partners in a real communicative context will be necessary to obtain more definitive results regarding segmentation preferences and processing.

Further investigation into segmentation preferences In our subjective evaluations, participants were asked to make global assessments of speech segmentations. We propose some possible explanations for the observed preferences, however to gain a more detailed understand of acceptable chunking, more fine-grained observations are required. In future work, we could imagine continuous audio assessment, as in Hansen and Kollmeier 1999, where evaluators must press a button every time they hear something disagreeable in the audio; this could allow us to pinpoint exactly which chunks led to reduced scores.

Thresholding and chunking policies We applied a threshold for CWT boundaries that gave regular chunks below the clause level, as one of our goals is to make TTS more incremental. It is possibly that adjusting the threshold to a slightly higher value could yield more favourable results, as it could eliminate some of the one-word chunks that the current threshold yielded. Alternatively, we could implement a minimum chunk threshold which would override single word chunks.

5.4 Conclusion

In this chapter, we explored prosodic event prediction with LMs for iTTS. We evaluated global prominence prediction results as well as a subcorpus of contrastively focused personal pronouns, designed to test the degree of linguistic knowledge provided by LMs for this task. We also investigated context-enhanced boundary prediction as a method for segmenting speech output for iTTS.

We compared prominence prediction accuracy under different context conditions (incremental, full-sentence, extended context). Previous context (one or two previous sentences) did not improve performance on full-sentence models or models with at least one-word lookahead, but did have a positive impact on an incremental model with no lookahead ($k = 0$). Lookahead was shown to provide improvement in prominence prediction on the full corpus (all POS), with gradual increases in F1 scores as more future context is added, but did not significantly improve prediction on a set of difficult samples. Prediction models which use pretrained LMs do improve prediction over baselines (except $k = 0$), but we see little evidence that they are contributing discursive knowledge.

Subjective evaluation of speech segmentation strategies showed a clear preference for LM-derived segmentation over more basic count-based methods. There was however no preference between the simplified POS-sequence method *CandC*) and a model fine-tuned on acoustic features (*CWT*). We conducted a further evaluation on the cognitive load imposed by the different segmentation strategies. Here we did not see definitive results, possibly due to offline processing which can take place in this slow output speech style.

Conclusion and perspectives

In this thesis, we have explored the topic of incremental text-to-speech synthesis and investigated the application of language models to this synthesis paradigm. We began by examining the impact of lookahead on a vanilla text-to-speech model and found that the first couple of future words had the largest impact on the internal representation of each word. We then sought to replace this lookahead with language model generated text. This method did improve prosodic prediction over a random next word control condition, but only when the exact next word was predicted correctly, and not when syntactic/POS predictions aligned with the ground-truth future context. Further analyses into the prosodic predictions made in different structural contexts reinforced our earlier finding that vanilla models only make use of very local/shallow context.

This limitation of end-to-end models trained on single sentence phoneme/Mel-spectrogram pairs, along with their inability to incorporate semantic or discursive factors into prosodic predictions, motivated us to shift from using language models to infer future text and to instead use them to predict prosodic features directly. We looked at both prominence and boundary prediction using pretrained language models which are known to encode syntactic and semantic features. For prominence prediction, we found that the use of language models and increased lookahead (and extended context in a minimum lookahead condition) could improve prediction accuracy on corpus-wide prediction, but on a set of difficult samples that require more understanding of discourse, we only saw minor improvements with additional context, suggesting we may need to apply more sophisticated language models to this task. As for boundaries, our objective was to guide the segmentation of speech for AACs where the input latency can be very high. We compared different methods for predicting appropriate chunks and found that language model informed segmentation was preferred over simpler count-based methods. However, we found no preference between the two test LM-based models (one rule-based and one data-driven).

There are several future avenues for continued research in incremental and contextualized text-to-speech synthesis. We outline a few below:

Model and data scale The language capabilities of massive language models like GPT-4 have demonstrated the benefits of size for NLP tasks. Some initial exploratory probes into contrast understanding by ChatGPT indicate these tools also encode discourse information useful for prosody prediction. Their increased size does however present problems for incorporating them into portable communication devices for the moment, but assuming continued engineering advancements, this could be feasible in the future. We would also expect improvements in prosody modelling with increased audio data. We investigated cases of marked prosodic structures (i.e., emphasis on personal pronouns), that are (by definition) less common in the training data. Collecting a larger training corpus to increase the number of samples of this type would encourage a neural network to learn the triggers for these distinct patterns,

rather than ignore them as outliers.

Latency and the effects on interactivity and engagement The subjective evaluations done in this thesis were all conducted online (i.e., over the internet) with pre-synthesized recordings. To truly evaluate an iTTS for interactive applications, we need to conduct tests that put this system into real use in real time in order to gauge its strengths and weaknesses. Betz et al. 2018 found that speech evaluated alone in a MOS test and that same speech evaluated as part of an interactive dialogue system received very different scores. It is possible we could see a similar reversal when testing chunking methods in real time. Our evaluations showed a clear preference for longer chunks, but when this system is being used to interact, will the periods of silence these longer chunks impose be tolerable? We also saw that longer range lookahead could be beneficial for prominence prediction, which in turn should facilitate discourse comprehension, but again, how long are listeners willing to wait? While trade-offs are inevitable for iTTS, we would like to find the system that best optimizes speech quality, listener comprehension and conversational engagement. Evaluation metrics from human-robot interaction research could serve to assess engagement: these include physiological measures such as heart rate and skin conductivity and interaction measures such as nodding and gaze (see Anzalone et al. 2015 and the references therein).

Other correlates of prosody for human-in-the-loop applications In this thesis, we concentrated on the use of text representations to infer prosodic features. As text-to-speech is a one-to-many problem, using this information alone we will never replicate the intentions of the speaker exactly. In future work, it would be interesting to explore other potential sources of prosodic signaling For example, there is some evidence that prosodic boundaries can be inferred from the duration of keystrokes (Fuchs and Krivokapić 2016). It remains to be seen whether other features like prominence could also be inferred from duration or from the force of keystrokes. Facial expressions have been shown to correlate with prosodic prominence for spoken language (e.g., Swerts and Krahmer 2008); we may find similar cues when a person is typing. In the case of speech-to-speech translation, the source speech will be a rich source of prosodic information. Some efforts have been made to transfer pauses from the source to target speech (Tam et al. 2022) as well as emphasis (Do et al. 2017). Systems that allow for direct user control over prosody are another possibility, but this could increase latency as the user would be burdened with trying to input both the content and the form of their message.

Contextual appropriateness Our focus in rendering TTS contextually appropriate was on information structural properties, but there are of course other criteria for appropriateness. These include emotional characteristics and speech acts. These features could also be inferred by modelling the ongoing discourse or from facial expressions. Bertero et al. 2016, for example, has conducted work on recognizing user's emotions in real time to increase the empathy level of a dialogue system.

Further work is also required to match the speech style with the use case. The models used

in this thesis were trained on read speech because there is a large quantity of clean recordings with reliable transcriptions. To gain a more conversational feel, training on spontaneous speech will be necessary. Traits specific to spontaneous speech (e.g., hesitations/filled pauses and repetitions) could also be utilized to make the gaps in iTTS more natural.

Appendix A

Test sentences

---

a - CandC method; b - CWT method

1. (a) Mrs. Jones // is a beloved kindergarten teacher // The parents // appreciate how patient // she is with the children // and her students // think she is very **nice/mean** //
   (b) Mrs. // Jones // is a beloved // kindergarten // teacher // The parents appreciate // how patient // she is with the children // and her students // think she is very **nice/mean** //
2. (a) Jimmy // and Molly // got married // the week before last // At the reception // Jimmy // introduced Molly // to his **brother/bride** //
   (b) Jimmy and Molly // got married // the week before last // At the reception // Jimmy introduced // Molly // to his **brother/bride** //
3. (a) Richard // is trying // to lose // some weight // He goes for a jog // every morning // Furthermore // he eats nothing but **salad/cakes** //
   (b) Richard // is trying to lose // some weight // He goes for a jog every morning // Furthermore // he eats nothing // but **salad/cakes** //
4. (a) Seth // finished washing // the dishes // He then // put them // in the **cupboard/dishwasher** //
   (b) Seth finished washing // the dishes // He then put them in the **cupboard/dishwasher** //
5. (a) Bill // was out // in the rain // for hours // He lost his car keys // in the park // When he finally // got home // he was very cold // and his coat // was **wet/dry** //
   (b) Bill was out in the rain // for hours // He lost // his car keys // in the park // When he finally // got home // he was very cold // and his coat // was **wet/dry** //
6. (a) Julia // was late getting back // from work // So // as not to wake her family // she walked on her **tiptoes/elbows** //
   (b) Julia // was late getting // back // from work // So as not to wake her family // she walked on her **tiptoes/elbows** //

7. (a) The dog chases // the cat // with the black eyes everyday // He never // catches it // because it is too **fast/slow** //

   (b) The dog chases // the cat // with the black eyes everyday // He never catches // it // because // it is too **fast/slow** //

8. (a) Laurence // is a skilled surgeon // He has performed // over six hundred surgeries // and his success rate // is very **high/low** //

   (b) Laurence // is a skilled // surgeon // He has performed over // six hundred surgeries // and his success rate // is very **high/low** //

9. (a) It was 11 o'clock // in the evening // Harry // heard a noise // in the yard // He wanted to see // what it was so // he turned the back light **on/off** //

   (b) It was 11 o'clock // in the evening // Harry heard a noise // in the yard // He wanted // to see what it was so // he turned // the back light // **on/off** //

10. (a) Martha // broke her hip // a month ago // due to osteoporosis // This was surprising // because she is very **young/old** //

    (b) Martha broke her hip // a month // ago // due // to osteoporosis // This was surprising // because she is very **young/old** //

11. (a) Record temperatures // are expected // this summer // in the Southern hemisphere // Even // in Antarctica // it will be unusually **hot/cold** //

    (b) Record temperatures // are expected // this summer in the Southern hemisphere // Even in Antarctica // it will be unusually // **hot/cold** //

12. (a) The baby // took a bad spill // Lucky // for him // the floor // of his playpen // was **soft/hard** //

    (b) The baby // took // a bad // spill // Lucky // for him // the floor of his playpen // was **soft/hard** //

13. (a) Try as // he might // he could not reach // the light switch // He was just too **short/tall** //

    (b) Try as he might // he could not reach // the light switch // He was just // too **short/tall** //

14. (a) It was taking ages // to chop // the vegetables // for the soup // The knife // was **blunt/sharp** //

    (b) It was taking ages // to chop the vegetables // for the soup // The knife // was **blunt/sharp** //

15. (a) Victoria // was accused // of murdering her husband // To prove her innocence // she had to hire // a good **lawyer/gardener** //

    (b) Victoria // was accused // of murdering // her husband // To prove her innocence // she had to hire a good **lawyer/gardener** //

16. (a) You may not cross // this fence // the land // on the other side // is **private/public** //

    (b) You may not cross // this fence // the land on the other side // is **private/public** //

17. (a) To be sure // you do not miss // your flight // make sure // you arrive at the airport extra **early/late** //

    (b) To be sure you do not // miss // your flight // make sure you arrive // at the airport // extra **early/late** //

18. (a) John // was late // for work // He waited impatiently // in his car // for the traffic // to start moving // Finally // the traffic // sped forward // as the light // turned **green/red** //

    (b) John was late // for work // He waited impatiently // in his car // for the traffic // to start moving // Finally // the traffic // sped forward // as the light turned **green/red** //

19. (a) The house // remained on the market // for years // despite being in excellent condition // So // the sellers // decided to bring // the price **down/up** //

    (b) The house // remained // on the market // for years // despite // being in excellent // condition // So the sellers // decided // to bring the price // **down/up** //

20. (a) Philip // could not afford // to pay // his bills // But now // that he has won // the lottery // he is no longer **poor/rich** //

    (b) Philip // could not afford // to pay his bills // But now that he has won // the lottery // he is no longer **poor/rich** //

Appendix B

French summary

Introduction

La synthèse vocale à partir du texte (Text-to-speech) est une technologie qui permet de donner la voix à ceux qui n'en ont pas, de combler les barrières linguistiques et de permettre aux humains de dialoguer avec les machines. Cependant, les systèmes TTS actuels présentent deux défauts majeurs. D'une part, ils fonctionnent classiquement à l'échelle de la phrase, c'est-à-dire que la synthèse sonore n'est déclenchée qu'après la saisie d'une phrase complète. Ce mode de fonctionnement ne permet pas une interaction fluide lorsque la synthèse vocale est utilisée par exemple comme une voix de substitution, l'interlocuteur devant attendre la saisie complète de la phrase à vocaliser avant de pouvoir la percevoir. D'autre part, en fonctionnant principalement à l'échelle de la phrase, les systèmes TTS actuels ne tiennent assez peu compte du contexte linguistique, fourni par exemple par les phrases précédentes du dialogue, notamment pour la prédiction d'indices prosodiques fins.

Dans cette thèse, notre objectif global est d'améliorer à la fois la qualité et l'interactivité des systèmes de synthèse TTS. Cela inclut les sous-objectifs suivants : (1) réduire le temps nécessaire pour commencer la production de la parole tout en maintenant une prosodie naturelle (paradigme de synthèse vocale incrémentale ou iTTS) et (2) prédire certaines caractéristiques prosodiques d'une phrase à synthétiser en exploitant son contexte. Nous abordons ces problématiques à travers l'utilisation de modèles de langage (LM). Les LM récents basés sur l'apprentissage profond (et plus précisément sur l'architecture *Transformer*) semblent capables de capturer des relations complexes entre différents niveaux linguistiques (morphologique, syntaxique, sémantique). Ils occupent aujourd'hui une place centrale dans de nombreux systèmes de traitement automatique des langues (TAL). Dans cette thèse, nous utilisons les LM d'une part pour encoder, sous une forme utilisable par un TTS, le contexte du dialogue, et d'autre part

pour prédire, au fur et à mesure de la saisie du texte, des informations contextuelles manquantes (par exemple le mot suivant le plus probable ou sa fonction grammaticale) ainsi que certains indices prosodiques comme le focus ou les limites de syntagmes (groupes de souffle).

Ce manuscrit de thèse est divisé en cinq chapitres dont nous résumons ci-après le contenu.

Chapitre 1

Au chapitre 1, nous discutons l'importance de l'interaction dans la communication parlée et montrons que les systèmes TTS actuels, lorsque utilisés par exemple comme voix de substitution, ne permettent pas une interaction fluide. Nous commençons par une revue de la littérature qui montre que la qualité d'une interaction orale entre deux interlocuteurs se dégrade en cas de retard ou de suppression des réactions du destinataire. Ces réactions servent à indiquer le niveau de compréhension de l'interlocuteur, à orienter les contributions futures et à construire un ensemble de connaissances partagées. Lorsque la parole est livrée phrase par phrase, comme c'est le cas dans la plupart des systèmes TTS actuels, le destinataire ne peut pas réagir avant d'avoir perçu l'intégralité de la phrase synthétisée. Un système capable de délivrer une parole de synthèse au fur et à mesure de sa saisie devrait donc permettre un échange plus fluide, s'approchant ainsi d'une interaction naturelle. Nous examinons ensuite la littérature sur la production et la perception de la parole. Du côté de la production, la littérature montre que les humains fournissent plus d'efforts sur les éléments moins prévisibles/plus importants pour le discours. Du côté de la perception, il a été démontré que les humains optimisent leurs efforts de décodage d'un stimulus auditif de parole en fonction de sa nouveauté.

Nous présentons ensuite quelques notions sur la prosodie et la structure de l'information. La prosodie fait référence aux variations de hauteur, d'énergie, de durée et de qualité de la voix utilisées pour attirer l'attention, structurer le message, et pour transmettre émotion et attitudes sociales. La structure de l'information désigne les méthodes utilisées par les locuteurs pour ajouter une information au "socle commun" qu'il partage avec leur interlocuteur, ce qui, pour l'anglais par exemple, est souvent fait à l'aide de patrons prosodiques spécifiques.

Nous terminons le chapitre par une discussion sur les lacunes à la fois des systèmes TTS actuels et des mesures utilisées pour les évaluer. La mesure la plus courante est le test MOS (i.e., score moyen d'opinion), où les participants donnent des évaluations subjectives de phrases individuelles isolées. Nous notons que ce type de test ne capture pas des éléments tels que l'adéquation d'une parole synthétique à un contexte de communication (discours). Nous nous intéressons donc à d'autres paradigmes de test, comme par exemple le paradigme de vérification des phrases, où les destinataires doivent juger de la véracité d'une déclaration sur la base d'un contexte. Nous expérimentons notamment ce type de test dans le chapitre 5.

Chapitre 2

Au chapitre 2, nous explorons les façons dont le contexte affecte le signal de la parole, et ce pour différents niveaux linguistiques (phonologique, syntaxique, sémantique, discursif). Si certaines informations peuvent être déduites facilement à partir du contexte proche (par exemple à l'échelle du syntagme), d'autres nécessitent une compréhension plus approfondie du contenu du message et/ou une prise en compte d'un empan temporel plus large (par exemple à l'échelle d'un paragraphe).

Nous étudions ensuite l'utilisation d'un modèle de langage basé sur l'architecture *Transformer* pour capturer ses informations contextuelles et les transformer en descripteurs prosodiques utilisables par un synthétiseur TTS. Différentes études ont montré que ces modèles apprennent des représentations qui encodent relativement bien la syntaxe et la sémantique. Cependant, d'autres études ont montré certaines limitations comme par exemple l'incapacité des modèles à généraliser des règles grammaticales pour des énoncés dont le vocabulaire est mal représenté dans leur corpus d'entraînement. De plus, les modèles de langage actuels ne semblent pas être en mesure d'encoder différemment des expressions qui font l'objet d'un chevauchement lexical (par exemple, "law school" et "school law"). Ainsi, bien qu'ils soient capables de produire un texte semblable à celui des humains, les modèles actuels semblent le faire en raisonnant de manière a priori différente. Dans cette thèse, l'une des questions que nous posons est de savoir si les modèles de langage sont capables de fournir à un système TTS des informations sur le contexte du discours, informations nécessaires à une bonne définition du contenu prosodique de la parole de synthèse.

Chapitre 3

Au chapitre 3, nous nous intéressons à la synthèse vocale incrémentale ou iTTS. Après une présentation des travaux précédents dans ce domaine de recherche relativement récent, nous présentons notre première contribution dans ce domaine, qui porte sur l'étude du comportement d'un système TTS neuronal (Tacotron2), préentrainé de façon standard - c'est-à-dire en exploitant à l'apprentissage la phrase complète - lorsqu'il est utilisé en mode incrémental - c'est-à-dire lorsqu'il synthétise un mot avec une connaissance seulement partielle des k mots à venir (le *lookahead*). Nous avons étudié, d'une part, l'évolution des représentations internes du modèle en sortie de l'encodeur, et d'autre part, la qualité du signal de parole synthétique, évaluée à l'aide d'un test perceptif. Les résultats montrent qu'en moyenne, la variation de la représentation interne associée à un mot diminue de 88% si on l'on considère le mot suivant, et de 94% si l'on considère 2 mots suivants.

Nous présentons ensuite une série d'expériences complémentaires basées sur une analyse de type forêt aléatoire, montrant que la longueur du mot à synthétiser est le meilleur prédicteur de la stabilité de la représentation interne associée à ce mot. Ce résultat permet de définir une première stratégie pour la synthèse vocale incrémentale, à savoir retarder davantage la synthèse de mots courts dont la représentation interne risque de changer avec la saisie des mots suivants.

Chapitre 4

Notre seconde contribution porte sur l'intégration, à un système TTS neuronal, d'un modèle de langage autoregressif, basé sur une architecture de type *Transformer* (tel que GPT) afin de prédire, au fur et à mesure de la saisie du texte, les mots suivants les plus probables. Nous faisons l'hypothèse que même si le modèle n'est pas capable de prédire exactement le ou les mots suivants, il prédira un texte dont la structure syntaxique sera cohérente et utile pour la prédiction du contenu prosodique de la parole de synthèse.

Pour évaluer l'efficacité de cette méthode, nous comparons plusieurs conditions : (1) synthèse TTS non-incrémentale à partir d'une phrase complète (utilisée comme référence), (2) synthèse TTS en utilisant uniquement le contexte gauche (c'est-à-dire les mots précédemment saisis), (3) synthèse TTS en utilisant unique-

ment le contexte gauche et un contexte droit prédit à l'aide d'un modèle de langage (méthode proposée), (4) synthèse TTS en utilisant uniquement le contexte gauche et un contexte droit constitué de mots choisis aléatoirement et (5) synthèse TTS en utilisant uniquement le contexte gauche et un contexte-droit limité (ce qui revient à introduire une latence de quelques mots dans la synthèse). Nous utilisons à la fois des mesures objectives et subjectives pour l'évaluation. Les mesures objectives évaluent la différence en termes de fréquence fondamentale, de durée segmentale et d'intensité entre une synthèse non-incrémentale et les différentes conditions de synthèses incrémentales. Pour l'évaluation subjective des différentes conditions, nous avons utilisé un test d'écoute de type MUSHRA.

Les évaluations objectives et perceptives menées montrent que l'approche par prédiction d'un contexte futur à l'aide d'un modèle de langage de type GPT permet un bon compromis entre réactivité et naturel de la synthèse, mais reste très dépendante de la qualité de la prédiction du texte.

Chapitre 5

Dans le chapitre 5, nous étudions la prédiction et le contrôle de la prosodie à l'aide de modèles de langage. Nous examinons à la fois la prédiction de l'emphase et de la limite des syntagmes (groupes de souffle). Pour l'emphase, nous étudions plus spécifiquement le focus contrastif sur les pronoms personnels, qui peut être particulièrement difficile en raison de la nécessite d'accéder à des connaissances de haut niveau sur la structure du discours.

Tout d'abord, nous avons constitué un corpus dédié contenant de nombreuses occurrences de pronoms pour lesquels on observe un focus contrastif. Pour quantifier le niveau de focus, nous utilisons la technique basée sur une transformée en ondelette d'un signal composite créé à partir de la courbe de fréquence fondamentale et des variations d'intensité du signal de parole. Cette approche permet d'extraire 3 niveaux de proéminences à partir du signal audio de parole, notés respectivement p1, p2 et p3.

Des modèles de langage pré-entrainés causaux (de type GPT) et non-causaux (de type BERT) sont ensuite adaptés pour une tache de prédiction du niveau de focus (p1,p2,p3) à partir du texte de la phrase à synthétiser, mais aussi des phrases précédentes. Les expériences montrent de bonnes performances globales

lorsqu'on considère l'ensemble des mots faisant l'objet d'un focus contrastif, mais une performance moindre (et peu d'amélioration par rapport à des techniques plus simples) si l'on considère uniquement les pronoms personnels. Ces résultats tendent à démontrer l'incapacité des modèles testés à capturer la structure fine du discours.

Nous étudions ensuite la possibilité de contrôler directement le niveau de focus dans la synthèse TTS. Un synthétiseur TTS de type *Fastspeech2* est entraîné à partir de données audio pour lesquelles une séquence d'étiquettes de focus (p1, p2, p3) est ajoutée au texte. Une évaluation perceptive montre une assez bonne corrélation entre les degrés de focus souhaités et perçus.

Nous terminons ce chapitre en évaluant l'utilisation de modèles de langage pour prédire les limites de syntagmes. Nous nous intéressons notamment à la segmentation, en ligne, d'une parole synthétique en groupes de mot (groupe de souffle), pour des applications en suppléance vocale. Nous comparons notre méthode basée sur les modèles de langage à des méthodes plus simples, telles que par exemple la synthèse d'un ou deux mot à la fois, ou bien l'approche par règle *chink-chunk*. Des tests subjectifs basés d'une part sur des tests de préférence et d'autre part sur le paradigme de vérification de phrase mentionné précédemment, ont montré une nette préférence par les sujets des segmentations réalisées par les approches basées LM et les approches par règles, en comparaison avec les heuristiques plus simples, mais cependant sans différence significative entre ces dernières.

Bibliography

Addlesee, Angus et al. (2020). "A Comprehensive Evaluation of Incremental Speech Recognition and Diarization for Conversational AI." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online), pp. 3492–3503 (cit. on p. 47).

Adhikary, Jiban et al. (2019). "Investigating Speech Recognition for Improving Predictive AAC." In: *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*. Minneapolis, Minnesota, pp. 37–43 (cit. on p. 69).

Aina, Laura and Tal Linzen (2021). "The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty Through Generation." In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Punta Cana, Dominican Republic, pp. 42–57 (cit. on pp. 32, 34).

AlBadawy, Ehab A et al. (2022). "Vocbench: A Neural Vocoder Benchmark for Speech Synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Singapore, pp. 881–885 (cit. on p. 44).

Altenberg, B (1987). *Prosodic Patterns in Spoken English: Studies in the Correlation Between Prosody and Grammar for Text-to-speech Conversion*. Lund University Press, p. 409 (cit. on p. 93).

Altmann, André et al. (2010). "Permutation importance: a corrected feature importance measure." In: *Bioinformatics* 26.10, pp. 1340–1347 (cit. on p. 60).

Ananthakrishnan, Sankaranarayanan and Shrikanth S. Narayanan (2008). "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence." In: *IEEE Transactions on Audio, Speech and Language Processing* 16 (1), pp. 216–228 (cit. on p. 93).

Anurova, Irina et al. (2022). "Event-related responses reflect chunk boundaries in natural speech." In: *NeuroImage* 255, p. 119203 (cit. on p. 118).

Anzalone, Salvatore M. et al. (2015). "Evaluating the Engagement with Social Robots." In: *International Journal of Social Robotics* 7 (4), pp. 465–478 (cit. on p. 134).

Arivazhagan, Naveen et al. (2019). "Monotonic Infinite Lookback Attention for Simultaneous Machine Translation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 1313–1323 (cit. on p. 48).

Arnon, Inbal and Neal Snider (2010). "More than words: Frequency effects for multi-word phrases." In: *Journal of Memory and Language* 62 (1), pp. 67–82 (cit. on p. 8).

Ashby, Michael (2006). "Prosody and idioms in English." In: *Journal of Pragmatics* 38 (10), pp. 1580–1597 (cit. on p. 22).

Astrinaki, Maria et al. (2012). "Reactive and continuous control of HMM-based speech synthesis." In: *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*. Miami, USA, pp. 252–257 (cit. on p. 46).

Aylett, Matthew and Alice Turk (2004). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech." In: *Language and Speech* 47 (1), pp. 31–56 (cit. on p. 8).

Azmi, Aqil M. et al. (2022). "Light Diacritic Restoration to Disambiguate Homographs in Modern Arabic Texts." In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 21 (3), 60:1–60:14 (cit. on p. 54).

Badino, Leonardo and Robert A.J. Clark (2008). "Automatic labeling of contrastive word pairs from spontaneous English." In: *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*. Goa, India, pp. 101–104 (cit. on p. 101).

Badino, Leonardo et al. (2009). "Identification of contrast and its emphatic realization in HMM based speech synthesis." In: *Proceedings of Interspeech*. Brighton, UK, pp. 520–523 (cit. on p. 101).

Badino, Leonardo et al. (2012). "Towards hierarchical prosodic prominence generation in TTS synthesis." In: *Proceedings of Interspeech*. Portland, USA, pp. 2398–2401 (cit. on p. 104).

Baevski, Alexei et al. (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Online, pp. 12449–12460 (cit. on p. 41).

Baker, Rachel E. and Ann R. Bradlow (2009). "Variability in word duration as a function of probability, speech style, and prosody." In: *Language and Speech* 52 (4), 391–413 (cit. on p. 22).

Ball, Peter (1975). "Listeners' Responses to Filled Pauses in Relation to Floor Apportionment." In: *British Journal of Social and Clinical Psychology* 14 (4), 423–424 (cit. on p. 45).

Bartlett, Susan et al. (2009). "On the Syllabification of Phonemes." In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado, pp. 308–316 (cit. on p. 45).

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4." In: *Journal of Statistical Software* 67.1, pp. 1–48 (cit. on p. 130).

Baumann, Timo (2014a). "Coordinating Speech Delivery to Gesture Progress for Deictic Expressions with Incremental Speech Synthesis." In: *Proceedings of the Workshop on Timing in Human Robot Interaction*. Bielefeld, Germany (cit. on p. 51).

— (2014b). "Decision tree usage for incremental parametric speech synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Florence, Italy, pp. 3819–3823 (cit. on p. 70).

— (2014c). "Partial representations improve the prosody of incremental speech synthesis." In: *Proceedings of Interspeech*. Singapore, pp. 2932–2936 (cit. on p. 68).

Baumann, Timo and David Schlangen (2012a). "Evaluating prosodic processing for incremental speech synthesis." In: *Proceedings of Interspeech*. Portland, USA, pp. 438–441 (cit. on pp. 46, 52, 68, 80).

— (2012b). "INPRO_iSS: A Component for Just-In-Time Incremental Speech Synthesis." In: *Proceedings of the ACL 2012 System Demonstrations*. Jeju Island, Korea, pp. 103–108 (cit. on p. 52).

— (2012c). "The InproTK 2012 release." In: *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*. Montréal, Canada, pp. 29–32 (cit. on p. 52).

— (2013). "Interactional adequacy as a factor in the perception of synthesized speech." In: *Proceedings of 8th ISCA Workshop on Speech Synthesis (SSW 8)*. Barcelona, Spain, pp. 223–227 (cit. on pp. 16, 51).

Baumann, Timo et al. (2010). "InproTK in Action: Open-Source Software for Building German-Speaking Incremental Spoken Dialogue Systems." In: *Electronic Speech Signal Processing 2010*, pp. 204–211 (cit. on p. 52).

Bavelas, Janet B. et al. (2000). "Listeners as co-narrators." In: *Journal of Personality and Social Psychology* 79 (6), pp. 941–952 (cit. on p. 6).

Beckman, Mary E and Jan Edwards (1994). "Articulatory evidence for differentiating stress categories." In: *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*. Ed. by Patricia Keating. Cambridge University Press, pp. 7–33 (cit. on p. 22).

Bengio, Samy et al. (2015). "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks." In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Montreal, Quebec, Canada, pp. 1171–1179 (cit. on p. 43).

Bengio, Yoshua et al. (2000). "A Neural Probabilistic Language Model." In: *Advances in Neural Information Processing Systems 13: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Denver, CO, USA, pp. 932–938 (cit. on p. 29).

Bertero, Dario et al. (2016). "Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, pp. 1042–1047 (cit. on p. 134).

Betz, Simon et al. (2015). "Micro-structure of disfluencies: Basics for conversational speech synthesis." In: *Proceedings of Interspeech*. Dresden, Germany, pp. 2222–2226 (cit. on p. 50).

Betz, Simon et al. (2018). "Interactive hesitation synthesis: Modelling and evaluation." In: *Multimodal Technologies and Interaction* 2 (1), p. 9 (cit. on p. 134).

Biron, Tirza et al. (2021). "Automatic detection of prosodic boundaries in spontaneous speech." In: *PLoS ONE* 16 (5 May), e0250969 (cit. on p. 94).

Boersma, Paul and David Weenink (2018). *Praat: doing phonetics by computer [Computer software]*. Version 6.0.37, retrieved 3 February 2018 `http://www.praat.org/` (cit. on p. 77).

Bolinger, Dwight (1982). "Intonation and Its Parts." In: *Language* 58 (3), pp. 505–533 (cit. on p. 28).

Boogaart, T. and Kim Silverman (1992). "Evaluating the overall comprehensibility of speech synthesizers." In: *Proceedings of 2nd International Conference on Spoken Language Processing (ICSLP 1992)*. Banff, Alberta, Canada, pp. 1207–1210 (cit. on p. 17).

Braun, Bettina et al. (2011). "An unfamiliar intonation contour slows down online speech comprehension." In: *Language and Cognitive Processes* 26, pp. 350 –375 (cit. on p. 119).

Breiman, Leo (2001). "Random Forests." In: *Machine Learning* 45 (1), pp. 5–32 (cit. on p. 60).

Broadbent, Judith (1991). "Linking and intrusive r in English." In: *UCL Working Papers in Linguistics* 3, pp. 281–302 (cit. on p. 20).

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Online, pp. 1877–1901 (cit. on pp. 33, 68, 81, 95).

Buschmeier, Hendrik and Stefan Kopp (2014). "When to elicit feedback in dialogue: Towards a model based on the information needs of speakers." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 71–80 (cit. on p. 6).

Buschmeier, Hendrik et al. (2012). "Combining Incremental Language Generation and Incremental Speech Synthesis for Adaptive Information Presentation." In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Seoul, South Korea, pp. 295–303 (cit. on p. 51).

Bögels, Sara et al. (2013). "Processing consequences of superfluous and missing prosodic breaks in auditory sentence comprehension." In: *Neuropsychologia* 51 (13), pp. 2715–2728 (cit. on p. 119).

Calhoun, Sasha (2007). "Predicting focus through prominence structure." In: *Proceedings of Interspeech*. Antwerp, Belgium, pp. 622–625 (cit. on p. 16).

— (2009). "What makes a word contrastive? Prosodic, semantic and pragmatic perspectives." In: *Studies in Pragmatics* 8, pp. 53–78 (cit. on p. 15).

Calhoun, Sasha et al. (2010). "The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue." In: *Language Resources and Evaluation* 44 (4), pp. 387–419 (cit. on p. 92).

Callahan, Sarah M. et al. (2012). "The processing and interpretation of verb phrase ellipsis constructions by children at normal and slowed speech rates." In: *Journal of Speech, Language, and Hearing Research* 55 (3), pp. 710–725 (cit. on p. 119).

Carlson, Katy et al. (2001). "Prosodic Boundaries in Adjunct Attachment." In: *Journal of Memory and Language* 45 (1), pp. 58–81 (cit. on p. 24).

Chafe, Wallace (1976). "Givenness, Contrastiveness, Definiteness, Subject, Topics, and Point of View." In: *Subject and Topic*. Ed. by C. Li. New York: Academic Press, pp. 25–55 (cit. on p. 11).

Chang, Pi-Chuan et al. (2008). "Optimizing Chinese Word Segmentation for Machine Translation Performance." In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio, pp. 224–232 (cit. on p. 54).

Chawla, Avi et al. (2021). "A Comparative Study of Transformers on Word Sense Disambiguation." In: *Communications in Computer and Information Science*, pp. 748–756 (cit. on p. 33).

Chen, Ken et al. (2004). "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Montreal, Canada, pp. 509–512 (cit. on p. 93).

Chodroff, Eleanor Rosalie and Jennifer Cole (2019). "The phonological and phonetic encoding of information status in American English nuclear accents." In: *Proceedings of the 19th International Congress of Phonetic Sciences*, p. 10 (cit. on p. 10).

Clark, Herbert H (1996). *Using language*. Cambridge university press (cit. on p. 5).

Clark, Herbert H. and Meredyth A. Krych (2004). "Speaking while monitoring addressees for understanding." In: *Journal of Memory and Language* 50 (1), pp. 62–81 (cit. on p. 6).

Clark, Herbert H and Catherine R Marshall (1981). "Definite reference and mutual knowledge." In: *Elements of discourse understanding*. Ed. by A Joshi et al. Cambridge University Press, pp. 10–63 (cit. on p. 6).

Clark, Kevin et al. (2019). "What Does BERT Look at? An Analysis of BERT's Attention." In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy, pp. 276–286 (cit. on pp. 30, 31, 34).

Clark, Kevin et al. (2020). "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." In: *International Conference on Learning Representations*. Addis Ababa, Ethiopia (cit. on pp. 30, 95).

Clifton, Charles et al. (2002). "Informative prosodic boundaries." In: *Language and Speech* 45 (2), pp. 87–114 (cit. on p. 24).

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1, pp. 37–46 (cit. on p. 105).

— (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge (cit. on p. 55).

Cole, Jennifer et al. (2017). "Crowd-sourcing prosodic annotation." In: *Computer Speech and Language* 45, pp. 300–325 (cit. on p. 92).

Cong, Jian et al. (2021). "Controllable context-aware conversational speech synthesis." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 4658–4662 (cit. on p. 96).

Conneau, Alexis et al. (2018). "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 2126–2136 (cit. on p. 33).

Cooper, William E. and Jeanne Paccia-Cooper (1980). *Syntax and Speech*. Harvard University Press (cit. on p. 93).

Crain, Stephen and Mark Steedman (2010). "On not being led up the garden path: the use of context by the psychological syntax processor." In: *Natural Language Parsing*. Cambridge University Press (cit. on p. 24).

Cutler, Anne and Donald J. Foss (1977). "On the role of sentence stress in sentence processing." In: *Language and Speech* 20 (1), pp. 1–10 (cit. on p. 9).

Dahan, Delphine et al. (2002). "Accent and reference resolution in spoken-language comprehension." In: *Journal of Memory and Language* 47 (2), pp. 292–314 (cit. on p. 29).

Dai, Zihang et al. (2019). "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 2978–2988 (cit. on pp. 30, 35).

Delasalles, Edouard et al. (Nov. 2019). "Learning Dynamic Author Representations with Temporal Language Models." In: *2019 IEEE International Conference on Data Mining (ICDM)*. Beijing, China: IEEE, pp. 120–129 (cit. on p. 81).

Dell, Gary S. and Peter A. Reich (1981). "Stages in sentence production: An analysis of speech error data." In: *Journal of Verbal Learning and Verbal Behavior* 20 (6), pp. 611–629 (cit. on p. 53).

Desplanques, Brecht et al. (2020). "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification." In: *Proceedings of Interspeech*. Shanghai, China (Online), pp. 3560–3564 (cit. on p. 113).

DeVault, David et al. (2011). "Incremental interpretation and prediction of utterance meaning for interactive dialogue." In: *Dialogue and Discourse* 2 (1), pp. 143–170 (cit. on p. 48).

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, pp. 4171–4186 (cit. on pp. 29, 30, 68, 97).

Do, Quoc Truong et al. (2017). "Preserving word-level emphasis in speech-to-speech translation." In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 25 (3), pp. 544–556 (cit. on p. 134).

Domínguez, Mónica et al. (2022). "The Information Structure-prosody interface in text-to-speech technologies. An empirical perspective." In: *Corpus Linguistics and Linguistic Theory* 18 (2), pp. 419–445 (cit. on p. 96).

Donahue, Jeff et al. (2021). "End-to-end Adversarial Text-to-Speech." In: *International Conference on Learning Representations*. Online (cit. on p. 40).

Duběda, Tomáš and Katalin Mády (2010). "Nucleus position within the intonation phrase: A typological study of English, Czech and Hungarian." In: *Proceedings of Interspeech*. Makuhari, Japan, pp. 126–129 (cit. on p. 104).

Edlund, Jens et al. (2009). "Pause and gap length in face-to-face interaction." In: *Proceedings of Interspeech*. Brighton, UK, pp. 2779–2782 (cit. on p. 28).

Ellinas, Nikolaos et al. (2020). "High quality streaming speech synthesis with low, sentence-length-independent latency." In: *Proceedings of Interspeech*. Shanghai, China, pp. 2022–2026 (cit. on pp. 48, 52).

Ellinas, Nikolaos et al. (2023). "Controllable speech synthesis by learning discrete phoneme-level prosodic representations." In: *Speech Communication* 146, pp. 22–31 (cit. on p. 92).

Elmers, Mikey et al. (2021). "Take a breath: Respiratory sounds improve recollection in synthetic speech." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 3196–3200 (cit. on p. 17).

Epp, Carrie Demmans et al. (2012). "Towards providing just-in-time vocabulary support for Assistive and Augmentative Communication." In: *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*. New York, NY, USA, pp. 33–36 (cit. on p. 70).

Erickson, Thomas D. and Mark E. Mattson (1981). "From words to meaning: A semantic illusion." In: *Journal of Verbal Learning and Verbal Behavior* 20 (5), pp. 540–551 (cit. on p. 8).

Ettinger, Allyson (2020). "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models." In: *Transactions of the Association for Computational Linguistics* 8, pp. 34–48 (cit. on pp. 32, 35).

Fan, Angela et al. (2018). "Hierarchical Neural Story Generation." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 889–898 (cit. on p. 73).

Farrús, Mireia et al. (2016). "Paragraph-based prosodic cues for speech synthesis applications." In: *Proceedings of the International Conference on Speech Prosody*. Boston, MA, USA, pp. 1143–1147 (cit. on p. 96).

Faure, Marc (1980). "Results of a contrastive study of hesitation phenomena in French and German." In: *Studies in Honour of Frieda Goldman-Eisler*. Ed. by Hans W Dechert and Manfred Raupach. De Gruyter Mouton, pp. 287–290 (cit. on p. 50).

Ferreira, Fernanda (1991). "Effects of length and syntactic complexity on initiation times for prepared utterances." In: *Journal of Memory and Language* 30 (2), pp. 210–233 (cit. on p. 23).

Ferreira, Fernanda and Suphasiree Chantavarin (2018). "Integration and Prediction in Language Processing: A Synthesis of Old and New." In: *Current Directions in Psychological Science* 27 (6), pp. 443–448 (cit. on p. 50).

Ferreira, Fernanda and Matthew W. Lowder (2016). "Prediction, Information Structure, and Good-Enough Language Processing." In: *Psychology of Learning and Motivation - Advances in Research and Theory* 65, pp. 217–247 (cit. on pp. 8, 9).

Ferreira, Fernanda et al. (2001). "Misinterpretations of garden-path sentences: Implications for models of sentence processing and re analysis." In: *Journal of Psycholinguistic Research* 30 (1), pp. 3–20 (cit. on p. 8).

Ferreira, Maria Fernanda (1988). "Planning and timing in sentence production: The syntax-to-phonology conversion." PhD thesis. University of Massachusetts Amherst (cit. on p. 93).

Féry, Caroline (2013). "Focus as prosodic alignment." In: *Natural Language & Linguistic Theory* 31, pp. 683–734 (cit. on p. 14).

Fitzpatrick, E. and J. Bachenko (1989). "Parsing for prosody: what a text-to-speech system needs from syntax." In: *Proceedings of the Annual AI Systems in Government Conference*. Washington, D.C., USA, pp. 188–194 (cit. on p. 73).

Fougeron, Cécile and Sun Ah Jun (1998). "Rate effects on French intonation: Prosodic organization and phonetic realization." In: *Journal of Phonetics* 26 (1), pp. 45–69 (cit. on p. 117).

Fougeron, Cécile and Patricia A. Keating (1997). "Articulatory strengthening at edges of prosodic domains." In: *The Journal of the Acoustical Society of America* 101 (6), pp. 3728–3740 (cit. on p. 21).

Fowler, C. A. (1981). "A relationship between coarticulation and compensatory shortening." In: *Phonetica* 38 (1-3), pp. 35–50 (cit. on p. 22).

Fuchs, Susanne and Jelena Krivokapić (2016). "Prosodic Boundaries in Writing: Evidence from a Keystroke Analysis." In: *Frontiers in Psychology* 7 (cit. on p. 134).

Fujihara, Riki et al. (2022). "Topicalization in Language Models: A Case Study on Japanese." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea, pp. 851–862 (cit. on p. 35).

Fukuda, Ryo et al. (2021). "Simultaneous Speech-to-Speech Translation System with Transformer-Based Incremental ASR, MT, and TTS." In: *24th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. Singapore, pp. 186–192 (cit. on p. 48).

Gallegos, Pilar Oplustil et al. (2021). "Comparing acoustic and textual representations of previous linguistic context for improving text-to-speech." In: *The 11th ISCA Speech Synthesis Workshop (SSW11)*. Budapest, Hungary, pp. 205–210 (cit. on p. 96).

Gee, James Paul and François Grosjean (1983). "Performance structures: A psycholinguistic and linguistic appraisal." In: *Cognitive Psychology* 15 (4), pp. 411–458 (cit. on p. 93).

Glennen, S and D C DeCoste (1997). *The Handbook of Augmentative and Alternative Communication*. Singular Publishing Group (cit. on p. 45).

Goldberg, Yoav (2019). "Assessing BERT's Syntactic Abilities." In: *ArXiv preprint* abs/1901.05287 (cit. on p. 34).

Goldstein, E. Bruce (1996). *Sensation and Perception*. 4th. Brooks/Cole Publishing Company, p. 124 (cit. on p. 7).

Goodwin, Charles (Jan. 1981). *Conversational Organization: Interaction Between Speakers and Hearers*. New York: Academic Press (cit. on p. 51).

Govender, Avashna and Simon King (2018a). "Measuring the Cognitive Load of Synthetic Speech Using a Dual Task Paradigm." In: *Proceedings of Interspeech*. Hyderabad, India, pp. 2843–2847 (cit. on pp. 17, 119).

— (2018b). "Using Pupillometry to Measure the Cognitive Load of Synthetic Speech." In: *Proceedings of Interspeech*. Hyderabad, India, pp. 2838–2842 (cit. on p. 17).

Gries, Stefan Th (2016). *Quantitative corpus linguistics with R: A practical introduction*. Routledge (cit. on p. 56).

Griffiths, Roger (1992). "Speech Rate and Listening Comprehension: Further Evidence of the Relationship." In: *TESOL Quarterly* 26 (2), p. 385 (cit. on p. 118).

Grillo, Nino et al. (2018). "Prosody of classic garden path sentences: The horse raced faster when embedded." In: *Proceedings of the International Conference on Speech Prosody*. Poznan, Poland, p. 284 (cit. on pp. 25, 83).

Grosz, B and C Sidner (1986). "Attention, intention, and the structure of discourse." In: *Computational Linguistics* 12, pp. 175–204 (cit. on p. 28).

Grosz, Barbara and Julia Hirschberg (1992). "Some intonational characteristics of discourse structure." In: *2nd International Conference on Spoken Language Processing (ICSLP)*. Banff, AB, Canada, pp. 429–432 (cit. on p. 28).

Grosz, Barbara J. et al. (1995). "Centering: A Framework for Modeling the Local Coherence of Discourse." In: *Computational Linguistics* 21.2, pp. 203–225 (cit. on p. 27).

Guo, Haohan et al. (2021). "Conversational End-to-End TTS for Voice Agents." In: *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*. Shenzhen, China, pp. 403–409 (cit. on p. 96).

Gussenhoven, Carlos (1983). "Testing the reality of focus domains." In: *Language and Speech* 26 (1), pp. 61–80 (cit. on p. 14).

Halliday, M. A. K. (2015). *Intonation and grammar in British English*. De Gruyter Mouton (cit. on p. 13).

Halliday, M. A.K. (1967). "Notes on transitivity and theme in English: Part 2." In: *Journal of Linguistics* 3 (2), pp. 199–244 (cit. on pp. 11, 120).

Hansen, Martin and Birger Kollmeier (1999). "Continuous assessment of time-varying speech quality." In: *The Journal of the Acoustical Society of America* 106 (5), pp. 2888–2899 (cit. on p. 131).

Harris, Zellig S. (1954). "Distributional Structure." In: *WORD* 10 (2-3), pp. 146–162 (cit. on p. 31).

Hart, Johan 't (1981). "Differential sensitivity to pitch distance, particularly in speech." In: *The Journal of the Acoustical Society of America* 69.3, pp. 811–821 (cit. on p. 82).

Hayashi, Tomoki et al. (2019). "Pre-trained text embeddings for enhanced text-to-speech synthesis." In: *Proceedings of Interspeech*. Graz, Austria, pp. 4430–4434 (cit. on p. 94).

Hayashi, Tomoki et al. (2020). "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain, pp. 7654–7658 (cit. on p. 75).

Hayes, Bruce (1984). "The Phonology of Rhythm in English." In: *Linguistic Inquiry* 15 (1), pp. 33–74 (cit. on p. 21).

Hewitt, John and Christopher D. Manning (2019). "A Structural Probe for Finding Syntax in Word Representations." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, pp. 4129–4138 (cit. on p. 34).

Higginbotham, D. Jeffery et al. (2009). "The effect of context priming and task type on augmentative communication performance." In: *AAC: Augmentative and Alternative Communication* 25 (1), pp. 19–31 (cit. on p. 70).

Hirschberg, Julia and Cinzia Avesani (1997). "The role of prosody in disambiguating potentially ambiguous utterances in English and Italian." In: *Intonation: Theory, models, and applications*. Athens, Greece, pp. 189–192 (cit. on p. 27).

Hirschberg, Julia and Diane Litman (1993). "Empirical Studies on the Disambiguation of Cue Phrases." In: *Computational Linguistics* 19.3, pp. 501–530 (cit. on pp. 73, 93, 97).

Hirschberg, Julia and Owen Rambow (2001). "Learning prosodic features using a tree representation." In: *7th European Conference on Speech Communication and Technology (EUROSPEECH)*. Aalborg, Denmark, pp. 1175–1178 (cit. on p. 73).

Hodari, Zack et al. (2021). "CAMP: A two-stage approach to modelling prosody in context." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Toronto, ON, Canada, pp. 6578–6582 (cit. on pp. 16, 41, 95).

Holtzman, Ari et al. (2020). "The Curious Case of Neural Text Degeneration." In: *8th International Conference on Learning Representations, (ICLR)*. Addis Ababa, Ethiopia (cit. on p. 74).

Honnibal, Matthew et al. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. `https://spacy.io/` (cit. on p. 104).

Hono, Yukiya et al. (2020). "Hierarchical multi-grained generative model for expressive speech synthesis." In: *Proceedings of Interspeech*. Shanghai, China, pp. 3441–3445 (cit. on p. 92).

Howell, Peter and Karima Kadi-Hanifi (1991). "Comparison of prosodic properties between read and spontaneous speech material." In: *Speech Communication* 10 (2), pp. 163–169 (cit. on p. 28).

Howes, Christine et al. (2012). "Responding to incomplete contributions in dialogue." In: *Cogsci* 34 (34) (cit. on p. 69).

Hsu, Wei-Ning et al. (2017). "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, pp. 1878–1889 (cit. on p. 92).

Hu, Na et al. (2016). "Discourse prosody and its application to speech synthesis." In: *Proceedings of 10th International Symposium on Chinese Spoken Language Processing, ISCSLP 2016*. Tianjin, China, pp. 1–5 (cit. on p. 96).

Hunt, Andrew J. and Alan W. Black (1996). "Unit selection in a concatenative speech synthesis system using a large speech database." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Atlanta, GA, USA, pp. 373–376 (cit. on p. 40).

Hwang, Hyekyung and Karsten Steinhauer (2011). "Phrase length matters: The interplay between implicit prosody and syntax in Korean "garden path" sentences." In: *Journal of Cognitive Neuroscience* 23 (11), pp. 3555–3575 (cit. on p. 120).

Irwin, Patricia (2011). "Intransitive sentences, argument structure, and the syntax-prosody interface." In: *Proceedings of the 28th West Coast Conference on Formal Linguistics (WCCFL)*. Los Angeles, CA, USA, pp. 275–284 (cit. on p. 26).

Ito, Keith (2017). *The LJ Speech Dataset*. `https://keithito.com/LJ-Speech-Dataset/` (cit. on pp. 54, 75).

ITU-R (2015). *Recommendation BS.1534 and BS.1116: Methods for the subjective assessment of small impairments in audio systems* (cit. on pp. 60, 80).

Jaccard, Paul (1908). "Nouvelles recherches sur la distribution florale." In: *Bull. Soc. Vaud. Sci. Nat.* 44, pp. 223–270 (cit. on p. 110).

Jadoul, Yannick et al. (2018). "Introducing Parselmouth: A Python interface to Praat." In: *Journal of Phonetics* 71, pp. 1–15 (cit. on p. 77).

Jasinskaja, Ekaterina et al. (2007). "Nuclear accent placement and other prosodic parameters as cues to pronoun resolution." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 1–14 (cit. on p. 27).

Jawahar, Ganesh et al. (2019). "What Does BERT Learn about the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 3651–3657 (cit. on pp. 32, 33).

Jelinek, Frederick (1976). "Continuous Speech Recognition by Statistical Methods." In: *Proceedings of the IEEE* 64 (4), pp. 532–556 (cit. on p. 29).

Jescheniak, Jörg D. and Willem J.M. Levelt (1994). "Word Frequency Effects in Speech Production: Retrieval of Syntactic Information and of Phonological Form." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20 (4), pp. 824–843 (cit. on p. 8).

Jia, Ye et al. (2021). "PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 151–155 (cit. on p. 40).

Jillings, Nicholas et al. (2016). "Web Audio Evaluation Tool: A framework for subjective assessment of audio." In: *Proceedings of the International Web Audio Conference*. Atlanta, GA, USA (cit. on pp. 61, 114, 126).

Johnson, Neal F (1965). "The psychological reality of phrase-structure rules." In: *Journal of Verbal Learning and Verbal Behavior* 4 (6), pp. 469–475 (cit. on p. 118).

Judge, Simon and Mark Landeryou (2007). "Disambiguation (Predictive Texting) for AAC." In: *Communication Matters Journal* 22 (2) (cit. on p. 48).

Jurafsky, Dan et al. (2008). "Probabilistic Relations between Words: Evidence from Reduction in Lexical Production." In: *Typological studies in language* 45, pp. 229–254 (cit. on pp. 8, 69).

Kakouros, Sofoklis and Okko Räsänen (2016). "Perception of Sentence Stress in Speech Correlates With the Temporal Unpredictability of Prosodic Features." In: *Cognitive Science* 40 (7), pp. 1739–1774 (cit. on p. 16).

Kakouros, Sofoklis et al. (2018). "Making predictable unpredictable with style – Behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech." In: *Neuropsychologia* 109, pp. 181–199 (cit. on pp. 11, 16).

Kalchbrenner, Nal et al. (2018). "Efficient Neural Audio Synthesis." In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Proceedings of Machine Learning Research, pp. 2415–2424 (cit. on p. 43).

Kameyama, Megumi (1999). "Stressed and unstressed pronouns: Complementary preferences." In: *Focus: Linguistic, cognitive and computational perspectives* (cit. on p. 27).

Kane, Shaun K. and Meredith Ringel Morris (2017). "Let's Talk about X: Combining image recognition and eye gaze to support conversation for people with ALS." In: *Proceedings of Designing Interactive Systems (DIS)*. Edinburgh, United Kingdom, pp. 129–134 (cit. on p. 70).

Kane, Shaun K. et al. (2012). "What we talk about: Designing a context-aware communication tool for people with aphasia." In: *ASSETS'12 - Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY, USA, pp. 49–56 (cit. on p. 70).

Kastner, Kyle et al. (2019). "Representation Mixing for TTS Synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Brighton, UK, pp. 5906–5910 (cit. on p. 40).

Katz, Jonah and Elisabeth Selkirk (2011). "Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English." In: *Language* 87 (4), pp. 771–816 (cit. on p. 10).

Kendra's Language School (2021). *Easy Slow English Conversation Practice for Super Beginners*. URL: `https://www.youtube.com/watch?v=VF-FYgfPdW4` (cit. on p. 125).

— (2022). *Easy Slow English Speaking Practice – Essential Phrases You Can Use for a Lifetime*. URL: `https://www.youtube.com/watch?v=IXlDRfZ8hAA&t=10s` (cit. on p. 125).

Kenter, Tom et al. (2019). "CHiVE: Varying Prosody in Speech Synthesis with a Linguistically Driven Dynamic Hierarchical Conditional Variational Network." In: *Proceedings of the 36th International Conference on Machine Learning, (ICML)*. Long Beach, CA, USA, pp. 3331–3340 (cit. on p. 16).

Kenter, Tom et al. (2020). "Improving the prosody of RNN-based English text-to-speech synthesis by incorporating a BERT model." In: *Proceedings of Interspeech*. Shanghai, China, pp. 4412–4416 (cit. on pp. 41, 94).

Khosla, Sopan et al. (2021). "Evaluating the Impact of a Hierarchical Discourse Representation on Entity Coreference Resolution Performance." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online, pp. 1645–1651 (cit. on p. 28).

Kilgarriff, Adam (2001). "Comparing corpora." In: *Int. Journal of Corpus Linguistics* 6.1, pp. 97–133 (cit. on p. 55).

Kilgarriff, Adam et al. (2004). "The Sketch Engine." In: *Proceedings of the Eleventh EURALEX International Congress*, pp. 105–116 (cit. on p. 55).

Kilgarriff, Adam et al. (2014). "The sketch engine: Ten years on." In: *Lexicography* 1 (1), pp. 7–36 (cit. on p. 55).

Kim, Jaehyeon et al. (2020). "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Online, pp. 8067–8077 (cit. on p. 41).

Kim, Jaehyeon et al. (2021). "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." In: *International Conference on Machine Learning*. Vienna, Austria, pp. 5530–5540 (cit. on p. 40).

Kisler, Thomas et al. (2017). "Multilingual processing of speech via web services." In: *Computer Speech & Language* 45, pp. 326–347 (cit. on p. 58).

Kiss, Katalin É (1998). "Identificational focus versus information focus." In: *Language* 74 (2), pp. 245–273 (cit. on p. 12).

Kitaev, Nikita and Dan Klein (2018). "Constituency Parsing with a Self-Attentive Encoder." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 2676–2686 (cit. on p. 60).

Kjelgaard, Margaret M. and Shari R. Speer (1999). "Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity." In: *Journal of Memory and Language* 40 (2), pp. 153–194 (cit. on p. 25).

Klafka, Josef and Allyson Ettinger (2020). "Spying on Your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, pp. 4801–4811 (cit. on pp. 32, 33).

Klatt, Dennis H. (1973). "Interaction Between Two Factors that Influence Vowel Duration." In: *The Journal of the Acoustical Society of America* 54 (1), pp. 313–313 (cit. on p. 22).

Kleinhans, Janine et al. (2017). "Using Prosody to Classify Discourse Relations." In: *Proceedings of Interspeech*. Stockholm, Sweden, pp. 3201–3205 (cit. on p. 28).

Koc, Wai Wan et al. (2021). "Text-to-Speech with Model Compression on Edge Devices." In: *The 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 114–119 (cit. on p. 48).

Koester, Heidi Horstmann and Sajay Arthanat (2018). "Text entry rate of access interfaces used by people with physical disabilities: A systematic review." In: *Assistive Technology* 30 (3), pp. 151–163 (cit. on pp. 47, 125).

Kong, Jungil et al. (2020). "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Online (cit. on pp. 43, 125).

Kong, Zhifeng et al. (2021). "DiffWave: A Versatile Diffusion Model for Audio Synthesis." In: *International Conference on Learning Representations*. Vienna, Austria (cit. on p. 43).

Kousidis, Spyros et al. (2014). "A multimodal in-car dialogue system that tracks the driver's attention." In: *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*. Istanbul, Turkey, pp. 26–33 (cit. on p. 51).

Krauss, Robert M. et al. (1977). "The role of audible and visible back-channel responses in interpersonal communication." In: *Journal of Personality and Social Psychology* 35 (7), pp. 523–529 (cit. on p. 6).

Krifka, Manfred (2008). "Basic notions of information structure." In: *Acta Linguistica Hungarica* 55.3-4, pp. 243–276 (cit. on pp. 11, 12).

Krivokapic, Jelena (2010). "Speech planning and prosodic phrase length." In: *Proceedings of the International Conference on Speech Prosody*. Chicago, Il, USA, paper 311 (cit. on p. 61).

Krivokapić, Jelena (2014). "Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1658) (cit. on p. 10).

Kudo, Taku (2018). "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 66–75 (cit. on p. 31).

Kudo, Taku and John Richardson (2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium, pp. 66–71 (cit. on p. 31).

Ladd, D. Robert (1984). "Declination: A review and some hypotheses." In: *Phonology* 1, pp. 53–74 (cit. on p. 52).

Lam, Perry et al. (2022). "EPIC TTS Models: Empirical Pruning Investigations Characterizing Text-To-Speech Models." In: *Proceedings of Interspeech*. Incheon, South Korea, pp. 823–827 (cit. on p. 48).

Lambrecht, Knud (1994). *Information Structure and Sentence Form*. Cambridge University Press (cit. on pp. 11, 12).

Latif, Siddique et al. (2021). "Controlling Prosody in End-to-End TTS: A Case Study on Contrastive Focus Generation." In: *CoNLL 2021 - 25th Conference on Computational Natural Language Learning, Proceedings*. Punta Cana, Dominican Republic, pp. 544–551 (cit. on pp. 92, 98).

Laures, Jacqueline S. and Kate Bunton (2003). "Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions." In: *Journal of Communication Disorders* 36 (6), pp. 449–464 (cit. on p. 16).

Le Maguer, Sébastien et al. (2013). "Evaluation of contextual descriptors for HMM-based speech synthesis in French." In: *Proceedings of 8th ISCA Workshop on Speech Synthesis (SSW 8)*, pp. 153–158 (cit. on p. 44).

Lee, Hung Yi et al. (2017). "Personalizing Recurrent-Neural-Network-Based Language Model by Social Network." In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 25 (3), pp. 519–530 (cit. on p. 69).

Lee, Yoonhyung et al. (2021). "Bidirectional Variational Inference for Non-Autoregressive Text-to-Speech." In: *Iclr 2021* (cit. on p. 41).

Lehiste, I. et al. (1975). "Role of duration in disambiguating syntactically ambiguous sentences." In: *The Journal of the Acoustical Society of America* 57 (S1), S47–S47 (cit. on p. 25).

Lei, Wenqiang et al. (2021). "Have We Solved The Hard Problem? It's Not Easy! Contextual Lexical Contrast as a Means to Probe Neural Coherence." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Online, pp. 13208–13216 (cit. on p. 101).

Leonarduzzi, Laetitia and Sophie Herment (2013). "Non canonical syntactic structures in discourse: tonality, tonicity and tones in English (semi-) spontaneous speech." In: *Proceedings of Interspeech*. Lyon, France, pp. 1453–1457 (cit. on p. 13).

Levelt, Willem J M (1989). *Speaking: From intention to articulation*. The MIT Press, pp. 566, xiv, 566–xiv (cit. on p. 52).

Levelt, Willem J.M. (1993). "Timing in Speech Production with Special Reference to Word Form Encoding." In: *Annals of the New York Academy of Sciences* 682 (1), pp. 283–295 (cit. on p. 41).

Levis, John M. and Greta Muller Levis (2018). "Teaching high-value pronunciation features: Contrastive stress for intermediate learners." In: *The CATESOL Journal* 30 (1), pp. 139–160 (cit. on p. 99).

Lewis, Mike et al. (2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, pp. 7871–7880 (cit. on p. 30).

Li, Chunrong et al. (2012). "Detection and emphatic realization of contrastive word pairs for expressive text-to-speech synthesis." In: *Proceedings of the 8th International Symposium on Chinese Spoken Language Processing, (ISCSLP)*. Kowloon Tong, China, pp. 93–97 (cit. on pp. 98, 101).

Li, Ke et al. (2020). "An Empirical Study of Transformer-Based Neural Language Model Adaptation." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain, pp. 7934–7938 (cit. on p. 69).

Li, Naihan et al. (2019). "Neural Speech Synthesis with Transformer Network." In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. Honolulu, Hawaii, pp. 6706–6713 (cit. on p. 41).

Liberman, Mark Y and Kenneth W Church (1992). "Text analysis and word pronunciation in text-to-speech synthesis." In: *Advances in speech signal processing*, pp. 791–831 (cit. on pp. 116, 122).

Lieberman, Philip (1963). "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech." In: *Language and Speech* 6 (3), pp. 172–187 (cit. on p. 8).

Lim, Yohan et al. (2021). "A Preliminary Study on Wav2Vec 2.0 Embeddings for Text-to-Speech." In: *International Conference on ICT Convergence*. Jeju Island, Korea, pp. 343–347 (cit. on p. 41).

Liu, Danni et al. (2022). "From Start to Finish: Latency Reduction Strategies for Incremental Speech Synthesis in Simultaneous Speech-to-Speech Translation." In: *Proceedings of Interspeech*. Incheon, South Korea, pp. 1771–1775 (cit. on pp. 48, 49, 52, 71).

Liu, Rui et al. (2021). "GraphSpeech: Syntax-aware graph attention network for neural speech synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Toronto, Canada, pp. 6059–6063 (cit. on p. 73).

Love, Tracy et al. (2008). "How left inferior frontal cortex participates in syntactic processing: Evidence from aphasia." In: *Brain and Language* 107 (3), pp. 203–219 (cit. on p. 119).

Love, Tracy et al. (2009). "Slowed speech input has a differential impact on on-line and off-line processing in children's comprehension of pronouns." In: *Journal of Psycholinguistic Research* 38 (3), pp. 285–304 (cit. on p. 119).

Loáiciga, Sharid et al. (2022). "New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea, pp. 875–886 (cit. on p. 35).

Luke, Steven G. and Kiel Christianson (2016). "Limits on lexical prediction during reading." In: *Cognitive Psychology* 88, pp. 22–60 (cit. on p. 69).

Luo, Renqian et al. (2021). "Lightspeech: Lightweight and fast text to speech with neural architecture search." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Toronto, ON, Canada, pp. 5699–5703 (cit. on p. 48).

Ma, Mingbo et al. (2019a). "STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework." In: *Proceedings of ACL*. Florence, Italy, 3025–3036 (cit. on p. 46).

Ma, Mingbo et al. (2020). "Incremental Text-to-Speech Synthesis with Prefix-to-Prefix Framework." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online, pp. 3886–3896 (cit. on pp. 46, 48, 55, 56).

Ma, Shuang et al. (2019b). "Neural TTS Stylization with Adversarial and Collaborative Games." In: *The 7th International Conference on Learning Representations, ICLR 2019*. New Orleans, LA, USA (cit. on p. 92).

Madureira, Brielen and David Schlangen (2020). "Incremental Processing in the Age of Non-Incremental Encoders: An Empirical Assessment of Bidirectional Models for Incremental NLU." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, pp. 357–374 (cit. on pp. 71, 72).

Makarov, Peter et al. (2022). "Simple and Effective Multi-sentence TTS with Expressive and Coherent Prosody." In: *Proceedings of Interspeech*. Incheon, South Korea, pp. 3368–3372 (cit. on p. 96).

Martos, Alejandro Pérez-González de et al. (2021). "Towards simultaneous machine interpretation." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 3951–3955 (cit. on p. 47).

McAuliffe, Michael et al. (2017). "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi." In: *Proceedings of Interspeech*. Stockholm, Sweden, pp. 498–502 (cit. on p. 104).

McCoy, Tom et al. (2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 3428–3448 (cit. on p. 35).

McFee, Brian et al. (2015). "Librosa: Audio and music signal analysis in Python." In: *Proceedings of the Python in Science Conference*. Austin, Texas, USA, pp. 18–24 (cit. on p. 77).

McLaughlin, Drew J. and Kristin J. Van Engen (2020). "Task-evoked pupil response for accurately recognized accented speech." In: *The Journal of the Acoustical Society of America* 147 2, EL151 (cit. on p. 119).

Meister, Clara et al. (2023). "Locally Typical Sampling." In: *Transactions of the Association for Computational Linguistics* 11, pp. 102–121 (cit. on p. 74).

Michalsky, Jan et al. (2018). "Conversational quality is affected by and reflected in prosodic entrainment." In: *Proceedings of the International Conference on Speech Prosody*. Poznan, Poland, pp. 389–392 (cit. on p. 28).

Mikolov, Tomas and Geoffrey Zweig (2012). "Context dependent recurrent neural network language model." In: *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*. Miami, USA, pp. 234–239 (cit. on p. 50).

Mikolov, Tomas et al. (2013). "Efficient estimation of word representations in vector space." In: *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. Scottsdale, AZ, USA (cit. on p. 29).

Miller, G. A. and J. A. Selfridge (1950). "Verbal context and the recall of meaningful material." In: *The American journal of psychology* 63 (2), p. 176 (cit. on p. 118).

Miller, George A (1951). *Language and communication.* McGraw-Hill (cit. on p. 5).

Mohan, Devang S Ram et al. (2020). "Incremental Text to Speech for Neural Sequence-to-Sequence Models using Reinforcement Learning." In: *Proceedings of Interspeech*. Shanghai, China, pp. 3186–3190 (cit. on p. 46).

Morrill, Tuuli (2012). "Acoustic Correlates of Stress in English Adjective-Noun Compounds." In: *Language and Speech* 55 (2), pp. 167–201 (cit. on p. 22).

Munhall, Kevin et al. (1992). ""Compensatory shortening" in monosyllables of spoken English." In: *Journal of Phonetics* 20 (2), pp. 225–239 (cit. on p. 22).

Murray, Gabriel et al. (2006). "Prosodic Correlates of Rhetorical Relations." In: *Proceedings of the Analyzing Conversations in Text and Speech*. New York City, New York, pp. 1–7 (cit. on p. 28).

Nagel, H. Nicholas et al. (1996). "Prosodic Influences on the Resolution of Temporary Ambiguity during On-Line Sentence Processing." In: *Journal of Psycholinguistic Research* 25 (2), pp. 319–344 (cit. on p. 25).

Nenkova, Ani and Dan Jurafsky (2007). "Automatic detection of contrastive elements in spontaneous speech." In: *Proceedings of 2007 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2007*. Kyoto, Japan, pp. 201–206 (cit. on pp. 93, 101).

Nenkova, Ani et al. (2007). "To Memorize or to Predict: Prominence labeling in Conversational Speech." In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York, pp. 9–16 (cit. on p. 97).

Nicol, Janet and David Swinney (1989). "The role of structure in coreference assignment during sentence comprehension." In: *Journal of Psycholinguistic Research* 18 (1), pp. 5–19 (cit. on p. 119).

Nielsen, Elizabeth et al. (2020). "The role of context in neural pitch accent detection in English." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, pp. 7994–8000 (cit. on p. 93).

Nix, Andrew J. et al. (1993). "Phoneme detection as a tool for comparing perception of natural and synthetic speech." In: *Computer Speech and Language* 7 (3), pp. 211–228 (cit. on pp. 17, 121, 129).

Nooteboom, S. G. (1987). "Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information." In: *Language and Cognitive Processes* 2 (3-4), pp. 145–163 (cit. on p. 9).

Nooteboom, S. G. and J. G. Kruyt (1987). "Accents, focus distribution, and the perceived distribution of given and new information: An experiment." In: *The Journal of the Acoustical Society of America* 81 (S1), S67–S67 (cit. on p. 119).

Oord, Aaron van den et al. (2016). "WaveNet: A Generative Model for Raw Audio." In: *Proceedings of the 9th ISCA Workshop on Speech Synthesis (SSW 9)*. Sunnyvale, CA, USA, p. 125 (cit. on p. 43).

OpenAI (n.d.). *ChatGPT* (cit. on p. 84).

— (2023). *GPT-4 Technical Report*. arXiv: `2303.08774 [cs.CL]` (cit. on p. 68).

Ostendorf, Mari et al. (1995). "The Boston University radio news corpus." In: *Linguistic Data Consortium*, pp. 1–19 (cit. on pp. 92, 117).

Pan, Shimei and Kathleen R. McKeown (1999). "Word Informativeness and Automatic Pitch Accent Modeling." In: *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (cit. on p. 93).

Pannekamp, Ann et al. (2005). "Prosody-driven sentence processing: An event-related brain potential study." In: *Journal of Cognitive Neuroscience* 17 (3), pp. 407–421 (cit. on p. 118).

Papineni, Kishore et al. (2002). "Bleu: a method for automatic evaluation of machine translation." In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Philadelphia, PA, USA, pp. 311–318 (cit. on p. 72).

Paris, C. R. et al. (2000). "Linguistic cues and memory for synthetic and natural speech." In: *Human Factors* 42 (3), pp. 421–431 (cit. on p. 17).

Park Kyubyong Kim, Jongseok (2019). *g2pE*. `https://github.com/Kyubyong/g2p` (cit. on p. 54).

Pauker, Efrat et al. (2011). "Effects of cooperating and conflicting prosody in spoken English garden path sentences: ERP evidence for the boundary deletion hypothesis." In: *Journal of Cognitive Neuroscience* 23 (10), pp. 2731–2751 (cit. on p. 119).

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python." In: *The Journal of Machine Learning research* 12, pp. 2825–2830 (cit. on p. 59).

Peirce, Jonathan et al. (2019). "PsychoPy2: Experiments in behavior made easy." In: *Behavior Research Methods* 51 (1), pp. 195–203 (cit. on p. 129).

Peiró-Lilja, Àlex and Mireia Farrús (2018). "Prosodic patterns to enhance text-to-speech naturalness." In: *Proceedings of the International Conference on Speech Prosody*. Poznań, Poland, pp. 612–616 (cit. on p. 96).

Pennington, Jeffrey et al. (2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 1532–1543 (cit. on p. 93).

Perrotin, Olivier et al. (2021). "Evaluating the extrapolation capabilities of neural vocoders to extreme pitch values." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 11–15 (cit. on p. 44).

Pickering, Martin J. and Simon Garrod (2007). "Do people use language production to make predictions during comprehension?" In: *Trends in Cognitive Sciences* 11 (3), pp. 105–110 (cit. on p. 6).

— (2013). "An integrated theory of language production and comprehension." In: *Behavioral and Brain Sciences* 36 (4), pp. 329–347 (cit. on p. 6).

Pierrehumbert, Janet (1980). "The phonology and phonetics of English intonation." PhD thesis. Massachusetts Institute of Technology (cit. on pp. 9, 52).

Pierrehumbert, Janet and Julia Hirschberg (1990). "The meaning of intonational contours in the interpretation of discourse." In: *Intentions in communication* (14), pp. 271–311 (cit. on p. 28).

Ping, Wei et al. (2018). "Deep Voice 3: 2000-Speaker Neural Text-to-Speech." In: *Proceedings of ICLR*. Vancouver, BC, Canada, pp. 214–217 (cit. on p. 41).

Pisoni, David B. and Sharon Hunnicutt (1980). "Perceptual evaluation of MITALK: the MIT unrestricted text-to-speech system." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Denver, Colorado, USA, pp. 572–575 (cit. on p. 17).

Pisoni, David B. et al. (1987). "Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility." In: *Computer Speech and Language* 2 (3-4), pp. 303–320 (cit. on pp. 17, 121).

Pitrelli, John F. and Ellen M. Eide (2003). "Expressive speech synthesis using American English ToBI: Questions and contrastive emphasis." In: *2003 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2003*. Honolulu, Hawaii, USA, pp. 694–699 (cit. on p. 98).

Plag, Ingo (2010). "Compound stress assignment by analogy: The constituent family bias." In: *Zeitschrift fur Sprachwissenschaft* 29 (2), pp. 243–282 (cit. on p. 22).

Pouget, Maël et al. (2015). "HMM training strategy for incremental speech synthesis." In: *Proceedings of Interspeech*. Dresden, Germany, pp. 1201–1205 (cit. on pp. 52, 68, 70).

Pouget, Maël et al. (2016). "Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis." In: *Proceedings of Interspeech*. San Francisco, USA, pp. 2846–2850 (cit. on pp. 46, 52, 53, 68, 85, 117).

Prenger, Ryan et al. (2019). "Waveglow: A Flow-based Generative Network for Speech Synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Brighton, UK, pp. 3617–3621 (cit. on pp. 43, 54).

Price, Patti et al. (1991). "The Use of Prosody in Syntactic Disambiguation." In: *Speech and Natural Language: Proceedings of a Workshop*. Pacific Grove, CA, USA, pp. 2956–70 (cit. on p. 23).

Purver, Matthew et al. (2009). "Split Utterances in Dialogue: a Corpus Study." In: *Proceedings of the SIGDIAL 2009 Conference*. London, UK, pp. 262–271 (cit. on p. 6).

Pynte, Joel (1998). "The role of prosody in semantic interpretation." In: *Music Perception* 16 (1), pp. 79–97 (cit. on p. 23).

Quené, Hugo (2007). "On the just noticeable difference for tempo in speech." In: *Journal of Phonetics* 35.3, pp. 353–362 (cit. on p. 82).

Quené, H. and R. F. Port (2002). "Rhythmical factors in stress shift." In: *IULC Working Papers* 38 (2) (cit. on p. 21).

Radford, Alec et al. (2019). "Language models are unsupervised multitask learners." In: *OpenAI blog* 1.8, p. 9 (cit. on pp. 30, 68, 75, 97).

Raffel, Colin et al. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *Journal of Machine Learning Research* 21, 140:1–140:67 (cit. on p. 30).

Ravanelli, Mirco et al. (2021). *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv:2106.04624. arXiv: `2106.04624 [eess.AS]` (cit. on p. 113).

Reeve, Jonathan (2016). *Chapterize*. `https://github.com/JonathanReeve/chapterize` (cit. on p. 104).

Ren, Yi et al. (2019). "FastSpeech: Fast, Robust and Controllable Text to Speech." In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, BC, Canada, pp. 3165–3174 (cit. on p. 41).

Ren, Yi et al. (2021). "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech." In: *International Conference on Learning Representations*. Online (cit. on pp. 41, 42, 74, 113).

Repp, Sophie (2016). "Contrast : Dissecting an Elusive Information-structural Notion and its Role in Grammar." In: *The Oxford Handbook of Information Structure* (October) (cit. on p. 100).

Roland, Douglas et al. (2012). "Semantic similarity, predictability, and models of sentence processing." In: *Cognition* 122 (3), pp. 267–279 (cit. on p. 69).

Rooth, Mats (1992). "A theory of focus interpretation." In: *Natural Language Semantics* 1 (1), pp. 75–116 (cit. on pp. 11, 26, 97, 99).

Rosenberg, Andrew (2010). "AuToBI - A tool for automatic ToBI annotation." In: *Proceedings of Interspeech*. Makuhari, Japan, pp. 146–149 (cit. on p. 93).

Ross, K and M Ostendorf (1996). "Prediction of abstract prosodic labels for speech synthesis." In: *Computer Speech Language* 10 (3), pp. 155–185 (cit. on p. 93).

Rudnicky, Alexander I. (2015). *The CMU pronouncing dictionary*. URL: `URL:http://www.speech.cs.cmu.edu/cgibin/cmudict` (cit. on p. 54).

Saeki, Takaaki et al. (2021a). "Incremental Text-to-Speech Synthesis Using Pseudo Lookahead with Large Pretrained Language Model." In: *IEEE Signal Processing Letters* 28, pp. 857–861 (cit. on pp. 45, 47, 49, 70, 74, 82, 87).

— (2021b). "Low-Latency Incremental Text-to-Speech Synthesis with Distilled Context Prediction Network." In: *Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021*. Cartagena, Colombia, pp. 749–756 (cit. on pp. 49, 70).

Sanderman, Angelien A. and René Collier (1997). "Prosodic Phrasing and Comprehension." In: *Language and Speech* 40 (4), pp. 391–409 (cit. on pp. 119, 129).

Sanders, Eric and Paul Taylor (1995). "Using statistical models to predict phrase boundaries for speech synthesis." In: *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, pp. 1811–1814 (cit. on p. 73).

Sanford, Alison J S et al. (2006). "Shallow Processing and Attention Capture in Written and Spoken Discourse." In: *Discourse Processes* 42 (2), pp. 109–130 (cit. on p. 9).

Schadle, Igor (2004). "Sibyl: AAC system using NLP techniques." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3118, pp. 1009–1015 (cit. on p. 70).

Schafer, Amy et al. (1996). "Focus in relative clause construal." In: *Language and Cognitive Processes* 11 (1-2), pp. 135–164 (cit. on p. 25).

Schafer, Amy Jean (1997). "Prosodic parsing: The role of prosody in sentence comprehension." PhD thesis. University of Massachusetts Amherst (cit. on p. 24).

Schober, Michael F. and Herbert H. Clark (1989). "Understanding by addressees and overhearers." In: *Cognitive Psychology* 21 (2), pp. 211–232 (cit. on p. 6).

Selkirk, Elisabeth (1984). *Phonology and Syntax- The Relation between Sound and Structure*. The MIT Press (cit. on pp. 14, 21).

— (1995). "Sentence prosody: Intonation, stress and phrasing." In: *Handbook of phonological theory*. Ed. by John Goldsmith. Blackwell, pp. 550–569 (cit. on pp. 13, 14).

— (2000). "The Interaction of Constraints on Prosodic Phrasing." In: *Prosody: Theory and Experiment: Studies Presented to Gösta Bruce*, pp. 231–261 (cit. on p. 120).

Selkirk, Elisabeth O. (2008). "Contrastive focus, givenness and the unmarked status of "discourse-new"." In: *Acta Linguistica Hungarica* 55 (3-4), pp. 331–346 (cit. on p. 62).

Sennrich, Rico et al. (2016). "Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pp. 1715–1725 (cit. on p. 31).

Setter, Jane (2015). *Aspects of connected speech*. University of Reading. URL: `https://www.youtube.com/watch?v=VM0cNDxBySc` (cit. on p. 20).

Shannon, Claude E and Warren Weaver (1949). "The mathematical theory of information." In: *Urbana: University of Illinois Press* 97 (6), pp. 128–164 (cit. on pp. 5, 29).

Shechtman, Slava and Alex Sorin (2019). "Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities." In: *Proceedings 10th ISCA Speech Synthesis Workshop (SSW 10)*. Vienna, Austria, pp. 275–280 (cit. on p. 43).

Shen, Jonathan et al. (2018). "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Calgary, AB, Canada, pp. 4779–4783 (cit. on pp. 41, 42, 53).

Shi, Wei and Vera Demberg (2019). "Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 5790–5796 (cit. on p. 36).

Shields, Joyce L. et al. (1974). "Reaction time to phoneme targets as a function of rhythmic cues in continuous speech." In: *Journal of Experimental Psychology* 102 (2), pp. 250–255 (cit. on p. 9).

Shwartz, Vered and Ido Dagan (2019). "Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition." In: *Transactions of the Association for Computational Linguistics* 7, pp. 403–419 (cit. on p. 33).

Silverman, Kim et al. (1992). "TOBI: a standard for labeling English prosody." In: *2nd International Conference on Spoken Language Processing, ICSLP 1992*. Banff, AB, Canada, pp. 867–870 (cit. on p. 9).

Simantiraki, Olympia et al. (2018). "Impact of different speech types on listening effort." In: *Proceedings of Interspeech*. Hyderabad, India, pp. 2267–2271 (cit. on p. 17).

Sityaev, Dmitry et al. (2007). "Some aspects of prosody of friendly formal and friendly informal speaking styles." In: *XVI International Conference on Phonetics Sciences*. Saarbrücken, Germany., pp. 6–10 (cit. on p. 28).

Siuzdak, Hubert et al. (2022). "WavThruVec: Latent speech representation as intermediate features for neural speech synthesis." In: *Proceedings of Interspeech*. Incheon, Korea, pp. 833–837 (cit. on p. 41).

Skantze, Gabriel and Anna Hjalmarsson (2013). "Towards incremental speech generation in conversational systems." In: *Computer Speech and Language* 27 (1), pp. 243–262 (cit. on pp. 50, 51).

Skerry-Ryan, R. J. et al. (2018). "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron." In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*. Stockholm, Sweden, pp. 4700–4709 (cit. on p. 92).

Smith, Caroline L (2004). "Topic transitions and durational prosody in reading aloud: production and modeling." In: *Speech Communication* 42 (3), pp. 247–270 (cit. on p. 28).

Sorodoc, Ionut-Teodor et al. (2020). "Probing for Referential Information in Language Models." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, pp. 4177–4189 (cit. on p. 35).

Speer, Robyn et al. (Oct. 2018). *rspeer/wordfreq: v2.2 [Computer software]* (cit. on p. 76).

Stalnaker, Robert (2002). "Common ground." In: *Linguistics and Philosophy* 25 (5-6), pp. 701–721 (cit. on pp. 11, 12).

Steinhauer, Karsten and Angela D. Friederici (2001). "Prosodic boundaries, comma rules, and brain responses: The closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers." In: *Journal of Psycholinguistic Research* 30 (3), pp. 267–295 (cit. on p. 118).

Steinhauer, Karsten et al. (1999). "Brain potentials indicate immediate use of prosodic cues in natural speech processing." In: *Nature Neuroscience* 2 (2), pp. 191–196 (cit. on p. 118).

Stephenson, Brooke et al. (2020). "What the future brings: Investigating the impact of lookahead for incremental neural TTS." In: *Proceedings of Interspeech*. Shanghai, China, pp. 215–219 (cit. on pp. 4, 76).

Stephenson, Brooke et al. (2021). "Alternate endings: Improving prosody for incremental neural TTS with predicted future text input." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 3865–3869 (cit. on pp. 4, 67).

Stephenson, Brooke et al. (2022). "BERT, can HE predict contrastive focus? Predicting and controlling prominence in neural TTS using a language model." In: *Proceedings of Interspeech*. Incheon, South Korea, pp. 3383–3387 (cit. on pp. 4, 91).

Strobl, Carolin et al. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution." In: *BMC Bioinformatics* 8, pp. 25 –25 (cit. on p. 60).

Strom, Volker et al. (2007). "Modelling prominence and emphasis improves unit-selection synthesis." In: *Proceedings of Interspeech*. Antwerp, Belgium, pp. 1282–1285 (cit. on pp. 92, 98).

Sun, Guangzhi et al. (2020). "Fully-Hierarchical Fine-Grained Prosody Modeling For Interpretable Speech Synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain, pp. 6264–6268 (cit. on p. 92).

Suni, Antti et al. (2017). "Hierarchical representation and estimation of prosody using continuous wavelet transform." In: *Computer Speech and Language* 45, pp. 123–136 (cit. on pp. 94, 95, 104, 116, 117, 123).

Suni, Antti et al. (2020). "Prosodic prominence and boundaries in sequence-to-sequence speech synthesis." In: *Proceedings of the International Conference on Speech Prosody*. Tokyo , Japan, pp. 940–944 (cit. on pp. 98, 113).

Swerts, Marc and Emiel Krahmer (2008). "Facial expression and prosodic prominence: Effects of modality and facial area." In: *Journal of Phonetics* 36 (2), pp. 219–238 (cit. on p. 134).

Syrdal, Ann K. and Yeon Jun Kim (2008). "Dialog speech acts and prosody: Considerations for TTS." In: *Proceedings of the 4th International Conference on Speech Prosody*. Campinas, Brazil, pp. 661–665 (cit. on p. 96).

Syrdal, Ann K et al. (2010). "Speech acts and dialog TTS." In: *Proceedings of 7th ISCA Workshop on Speech Synthesis (SSW 7)*. Kyoto, Japan, pp. 179–183 (cit. on p. 96).

Tachibana, Hideyuki et al. (2018). "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Calgary, AB, Canada, pp. 4784–4788 (cit. on p. 43).

Taglicht, Josef (1998). "Constraints on intonational phrasing in English." In: *Journal of Linguistics* 34 (1), pp. 181–211 (cit. on p. 120).

Talman, Aarne et al. (2019). "Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations." In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland, pp. 281–290 (cit. on pp. 41, 95).

Tam, Derek et al. (2022). "Isochrony-Aware Neural Machine Translation for Automatic Dubbing." In: *Proceedings of Interspeech*. Incheon, South Korea, pp. 1776–1780 (cit. on p. 134).

Tan, Xu et al. (2021). "A Survey on Neural Speech Synthesis." In: *ArXiv preprint arXiv:2106.15561* (cit. on p. 40).

Taylor, Wilson L. (1953). ""Cloze Procedure": A New Tool for Measuring Readability." In: *Journalism Quarterly* 30 (4), pp. 415–433 (cit. on p. 69).

Tenney, Ian et al. (2019). "BERT Rediscovers the Classical NLP Pipeline." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 4593–4601 (cit. on p. 34).

Tiedemann, Jörg and Yves Scherrer (2017). "Neural Machine Translation with Extended Context." In: *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark, pp. 82–92 (cit. on p. 50).

Tokuda, Keiichi et al. (2000). "Speech parameter generation algorithms for HMM-based speech synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Istanbul, Turkey, pp. 1315–1318 (cit. on p. 40).

Tolins, Jackson and Jean E. Fox Tree (2014). "Addressee backchannels steer narrative development." In: *Journal of Pragmatics* 70, pp. 152–164 (cit. on p. 6).

Trnka, Keith et al. (2006). "Topic Modeling in Fringe Word Prediction for AAC." In: *Proceedings of the 11th International Conference on Intelligent User Interfaces*. Sydney, Australia, pp. 276–278 (cit. on p. 69).

Trouvain, Jürgen and Martine Grice (1999). "The Effect of Tempo on Prosodic Structure." In: *Proceedings of the 14th International Conference in Phonetic Sciences*. San Francisco, CA, USA, pp. 1067–1070 (cit. on p. 117).

Trueswell, John C. et al. (1993). "Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference From Garden-Paths." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19 (3), pp. 528–553 (cit. on p. 24).

Tukey, John W. (1949). "Comparing Individual Means in the Analysis of Variance." In: *Biometrics* 5.2, pp. 99–114 (cit. on p. 130).

Tun, Patricia A (1998). "Fast noisy speech: age differences in processing rapid speech with background noise." In: *Psychology and aging* 13.3, p. 424 (cit. on p. 119).

Tyagi, Shubhi et al. (2020). "Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection." In: *Proceedings of Interspeech*. Shanghai, China, pp. 4407–4411 (cit. on p. 16).

Tyler, Joseph (2013). "Prosodic correlates of discourse boundaries and hierarchy in discourse production." In: *Lingua* 133, pp. 101–126 (cit. on p. 28).

— (2014). "Prosody and the Interpretation of Hierarchically Ambiguous Discourse." In: *Discourse Processes* 51 (8), pp. 656–687 (cit. on p. 28).

Umbach, Carla (2004). "On the notion of contrast in information structure and discourse structure." In: *Journal of Semantics* 21 (2), pp. 155–175 (cit. on p. 100).

Valin, Jean-Marc and Jan Skoglund (2019). "LPCNET: Improving Neural Speech Synthesis through Linear Prediction." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Brighton, UK, pp. 5891–5895 (cit. on p. 43).

Vartabedian, A. G. (1966). "The Effects of Transmission Delay in Four-Wire Teleconferencing." In: *Bell System Technical Journal* 45 (10), pp. 1673–1688 (cit. on p. 7).

Venditti, Jennifer J et al. (2002). "Discourse constraints on the interpretation of nuclear-accented pronouns." In: *Proceedings of the 1st International Conference on Speech Prosody*. Aix-en-Provence, France, pp. 675–678 (cit. on p. 27).

Voita, Elena et al. (2019). "The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 4396–4406 (cit. on p. 33).

Vulić, Ivan et al. (2020). "Probing Pretrained Language Models for Lexical Semantics." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, pp. 7222–7240 (cit. on p. 31).

Wagner, Michael (2006). "Givenness and Locality." In: *Semantics and Linguistic Theory* 16, p. 295 (cit. on p. 99).

Wagner, Petra et al. (2019). "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program." In: *Proceedings of 10th ISCA Workshop on Speech Synthesis (SSW 10)*. Vienna, Austria, pp. 105–110 (cit. on p. 15).

Wang, Changhan et al. (2021a). "FAIRSEQ S^2: A Scalable and Integrable Speech Synthesis Toolkit." In: *arXiv preprint arXiv:2109.06912* (cit. on pp. 113, 124).

Wang, Dagen and Shrikanth S. Narayanan (2004). "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Montreal, Quebec, Canada, pp. 525–532 (cit. on p. 94).

Wang, Disong et al. (2021b). "Fcl-TaCO2: Towards fast, controllable and lightweight text-to-speech synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Toronto, ON, Canada, pp. 5714–5718 (cit. on p. 48).

Wang, Jie et al. (2021c). "Adversarially learning disentangled speech representations for robust multi-factor voice conversion." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 846–850 (cit. on p. 92).

Wang, Lucy Lu et al. (2020). "CORD-19: The COVID-19 Open Research Dataset." In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online (cit. on p. 56).

Wang, Yuxuan et al. (2017). "Tacotron: Towards End-to-End Speech Synthesis." In: *Proceedings of Interspeech*. Stockholm, Sweden, pp. 4006–4010 (cit. on p. 41).

Wang, Yuxuan et al. (2018). "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis." In: *International Conference on Machine Learning*. Stockholm, Sweden, pp. 5180–5189 (cit. on pp. 92, 96).

Watson, Duane and Edward Gibson (2004). "The relationship between intonational phrasing and syntactic structure in language production." In: *Language and Cognitive Processes* 19 (6), pp. 713–755 (cit. on p. 93).

Watson, Duane G. et al. (2008). "Interpreting pitch accents in online comprehension: H* vs. L+H*." In: *Cognitive Science* 32 (7), pp. 1232–1244 (cit. on p. 29).

Wei, Jason et al. (2021). "Frequency Effects on Syntactic Rule Learning in Transformers." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Punta Cana, Dominican Republic, pp. 932–948 (cit. on pp. 31, 34).

Weiss, Ron J. et al. (2021). "Wave-Tacotron: Spectrogram-Free End-To-End Text-To-Speech Synthesis." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Toronto, ON, Canada, pp. 5679–5683 (cit. on p. 40).

Wester, Mirjam et al. (2016). "Evaluating comprehension of natural and synthetic conversational speech." In: *Proceedings of the International Conference on Speech Prosody*. Boston, MA, USA, pp. 736–740 (cit. on p. 17).

Wester, Mirjam et al. (2017). "Real-time reactive speech synthesis: Incorporating interruptions." In: *Proceedings of Interspeech*. Stockholm, Sweden, pp. 3996–4000 (cit. on p. 51).

Whalen, D. H. et al. (1995). "The effects of breath sounds on the perception of synthetic speech." In: *Journal of the Acoustical Society of America* 97 (5), pp. 3147–3153 (cit. on p. 17).

Wiedemann, Gregor et al. (2020). "Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings." In: *Proceedings of the 15th Conference*

on Natural Language Processing, KONVENS 2019. Erlangen, Germany, pp. 161–170 (cit. on p. 33).

Wightman, C. W. and M. Ostendorf (1991). "Automatic recognition of prosodic phrases." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Toronto, Ontario, Canada, pp. 321–324 (cit. on p. 94).

Wightman, Colin W. et al. (1992). "Segmental Durations In The Vicinity Of Prosodic Phrase Boundaries." In: *Journal of the Acoustical Society of America* 91 (3), pp. 1707–1717 (cit. on p. 21).

Winkler, Susanne (2005). *Ellipsis and Focus in Generative Grammar*. De Gruyter Mouton, pp. 22 –25 (cit. on p. 7).

Wisenburn, Bruce and D. Jeffery Higginbotham (2008). "An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances: Objective results." In: *AAC: Augmentative and Alternative Communication* 24 (2), pp. 100–109 (cit. on p. 70).

Wolf, Thomas et al. (2020). "Transformers: State-of-the-art natural language processing." In: *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*. Online, pp. 38–45 (cit. on p. 75).

Wolters, Maria K. et al. (2014). "Can older people remember medication reminders presented using synthetic speech?" In: *Journal of the American Medical Informatics Association* 22 (1), pp. 35–42 (cit. on p. 17).

Wu, Zhiyong et al. (2020). "Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, pp. 4166–4176 (cit. on p. 32).

Xiao, Yujia et al. (2020). "Improving Prosody with Linguistic and Bert Derived Features in Multi-Speaker Based Mandarin Chinese Neural TTS." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain, pp. 6704–6708 (cit. on pp. 41, 95).

Yamamoto, Ryuichi et al. (2020). "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain, pp. 6199–6203 (cit. on pp. 43, 75, 113).

Yamasaki, Tomohiro (2022). "Grapheme-to-Phoneme Conversion for Thai using Neural Regression Models." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, WA, USA, pp. 4251–4255 (cit. on p. 54).

Yanagita, Tomoya et al. (2019). "Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework." In: *Proceedings of Interspeech*. Graz, Austria, pp. 183–188 (cit. on pp. 39, 45, 52, 64, 116).

Yao, Kaisheng and Geoffrey Zweig (2015). "Sequence-to-sequence neural net models for grapheme-to-phoneme conversion." In: *Proceedings of Interspeech*. Dresden, Germany, pp. 3330–3334 (cit. on p. 41).

Yngve, V. H. (1970). "On getting a word in edgewise." In: *Chicago Linguistics Society, 6th Meeting*, pp. 567–578 (cit. on p. 6).

Yoon, Hyun-Wook et al. (2022). "Language Model-Based Emotion Prediction Methods for Emotional Speech Synthesis Systems." In: *Proceedings of Interspeech*. Incheon, South Korea, pp. 4596–4600 (cit. on p. 95).

Yu, Lang and Allyson Ettinger (2020). "Assessing Phrasal Representation and Composition in Transformers." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, pp. 4896–4907 (cit. on p. 34).

Yu, Zhou et al. (2015). "Incremental Coordination: Attention-Centric Speech Production in a Physically Situated Conversational Agent." In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic, pp. 402–406 (cit. on p. 51).

Yule, George (1980). "Speakers' topics and major paratones." In: *Lingua* 52 (1-2), pp. 33–47 (cit. on p. 28).

Zeman, Daniel et al. (2017). "CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies." In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada, pp. 1–19 (cit. on p. 122).

Zen, Heiga et al. (2019). "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech." In: *Proceedings of Interspeech 2019*. Graz, Austria, pp. 1526–1530 (cit. on pp. 55, 76).

Zhang, Hao et al. (2019). "Neural Models of Text Normalization for Speech Applications." In: *Computational Linguistics* 45.2, pp. 293–337 (cit. on p. 41).

Zhang, Shaolei et al. (2020). "Future-Guided Incremental Transformer for Simultaneous Translation." In: *AAAI Conference on Artificial Intelligence*. New York, NY, USA, pp. 14428–14436 (cit. on p. 47).

Zhang, Tong et al. (2006). "Extraction of pragmatic and semantic salience from spontaneous spoken English." In: *Speech Communication* 48 (3-4), pp. 437–462 (cit. on p. 101).

Zheng, Renjie et al. (2019). "Speculative Beam Search for Simultaneous Translation." In: *Proceedings of EMNLP-IJCNLP*. Hong Kong, China, pp. 1395–1402 (cit. on p. 71).

Zheng, Renjie et al. (2020). "Fluent and Low-latency Simultaneous Speech-to-Speech Translation with Self-adaptive Training." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online, pp. 3928–3937 (cit. on p. 48).

Zhou, Xuehao et al. (2020). "End-to-End Code-Switching TTS with Cross-Lingual Language Model." In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Barcelona, Spain, pp. 7614–7618 (cit. on p. 95).

Zimmermann, Malte (2008). "Contrastive focus and emphasis." In: *Acta Linguistica Hungarica* 55 (3-4), pp. 347–360 (cit. on p. 15).

Zou, Yuxiang et al. (2021). "Fine-grained prosody modeling in neural speech synthesis using ToBI representation." In: *Proceedings of Interspeech*. Brno, Czech Republic, pp. 3146–3150 (cit. on pp. 41, 95).

Zvonik, Elena and Fred Cummins (2003). "The effect of surrounding phrase lengths on pause duration." In: *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*. Geneva, Switzerland, pp. 777–780 (cit. on p. 61).