Suprasegmental representations for the modeling of fundamental frequency in statistical parametric speech synthesis

Manuel Fonseca de Sam Bento Ribeiro

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2018

Abstract

Statistical parametric speech synthesis (SPSS) has seen improvements over recent years, especially in terms of intelligibility. Synthetic speech is often clear and understandable, but it can also be bland and monotonous. Proper generation of natural speech prosody is still a largely unsolved problem. This is relevant especially in the context of expressive audiobook speech synthesis, where speech is expected to be fluid and captivating.

In general, prosody can be seen as a layer that is superimposed on the segmental (phone) sequence. Listeners can perceive the same melody or rhythm in different utterances, and the same segmental sequence can be uttered with a different prosodic layer to convey a different message. For this reason, prosody is commonly accepted to be inherently suprasegmental. It is governed by longer units within the utterance (e.g. syllables, words, phrases) and beyond the utterance (e.g. discourse). However, common techniques for the modeling of speech prosody - and speech in general - operate mainly on very short intervals, either at the state or frame level, in both hidden Markov model (HMM) and deep neural network (DNN) based speech synthesis.

This thesis presents contributions supporting the claim that stronger representations of suprasegmental variation are essential for the natural generation of fundamental frequency for statistical parametric speech synthesis. We conceptualize the problem by dividing it into three sub-problems: (1) representations of acoustic signals, (2) representations of linguistic contexts, and (3) the mapping of one representation to another. The contributions of this thesis provide novel methods and insights relating to these three sub-problems.

In terms of sub-problem 1, we propose a multi-level representation of *f0* using the continuous wavelet transform and the discrete cosine transform, as well as a wavelet-based decomposition strategy that is linguistically and perceptually motivated. In terms of sub-problem 2, we investigate additional linguistic features such as text-derived word embeddings and syllable bag-of-phones and we propose a novel method for learning word vector representations based on acoustic counts. Finally, considering sub-problem 3, insights are given regarding hierarchical models such as parallel and cascaded deep neural networks.

Acknowledgements

I would like to express my gratitude to:

- Rob Clark, who supervised my work during the first half of my PhD. Thank you for believing in my skills during a difficult first year and for guiding me through a period when it is so easy to feel hopeless with disappointing results. Many thanks for providing comments on the completed draft of this thesis!
- Oliver Watts, who officially supervised me for the second half of my PhD, but who was always available since the beginning to comment and discuss random ideas. Many thanks for proof-reading this thesis and for the truly invaluable feedback throughout my work! I am also grateful to Oliver for many resources, such as the Active Learning tool used in Chapter 9 of this thesis.
- Junichi Yamagishi, also my supervisor, for his feedback, for being a relentless source of new ideas, and for always finding the innovative side of an idea. Many thanks for providing the code to the Voice Cloning Toolkit (VCTK) and for guiding me through its inner workings.
- Simon King and James Kirby, my internal examiners for the annual reviews, for commenting on my work and steering it in the right direction.
- Alice Turk, Peter Cahill, and Martti Vainio, my examiners, for peer-reviewing this work and for invaluable comments, which have undoubtedly improved this thesis.
- Antti Suni, Martti Vainio, and colleagues, for their work with the Continuous Wavelet Transform (CWT), which helped shape this thesis. I am very grateful to Antti for kindly sharing their implementation of the CWT.
- Zhizheng Wu, for his early implementations of the Merlin Neural Network Toolkit and for his many insights and discussions regarding neural networks.
- Gustav Eje Henter, for finding a simple implementation for the MUSHRA evaluation used throughout this thesis.

- Rosie Kay, Alexandra Delipalta, and Michael Hobart, for their help recruiting participants and guiding them through some of my listening tests.
- Rosie Kay, for kindly agreeing to lend her voice to the recording of the small dataset presented in Chapter 5.
- Toshiba Research Europe Limited (Cambridge Research Laboratory), for the pre-processed audiobook *A Tramp Abroad*, which was used in some chapters of this thesis.
- The research team for the SIWIS project: Phil Garner, Alexandros Lazaridis, Pierre-Edouard Honnet, and Milo Cerak at the Idiap Research Institute; Eric Wehrli, Jean-Philippe Goldman, and Maria Ivanova at the University of Geneva; and Beat Pfister and Hui Liang at ETH Zurich.
- Everyone at CSTR, for truly being among the most knowledgeable people working in Speech Technology. Your feedback and ideas are invaluable!
- Everyone else in the Informatics Forum, Edinburgh, and beyond, with whom I spent time with. There's too many to name, but you know who you are. A PhD student's life is certainly not limited to work and you all helped this thesis come to life by keeping me sane.

This thesis was supported by the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS) and the EPSRC Standard Research Studentship (EP/K503034/1).

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Manuel Fonseca de Sam Bento Ribeiro*)

The surface of the ocean responds to forces that act upon it in movements resembling the ups and downs of the human voice. If our vision could take it all in at once, we would discern several types of motion, involving a greater and greater expanse of sea and volume of water: ripples, waves, swells, and tides. It would be more accurate to say ripples 'on' waves 'on' swells 'on' tides, because each larger movement carries the smaller ones on its back (...) In speech (...) the ripples are the accidental changes in pitch, the irrelevant quavers. The waves are the peaks and the valleys that we call 'accent'. The swells are the separations of our discourse into its larger segments. The tides are the tides of emotion.

(Bolinger (1964), Bolinger (1972))

Contents

Chapter 1

Introduction

1.1 Motivation

Speech synthesis is the automatic conversion of a language's written form into its spoken form. Statistical parametric speech synthesis consists of various statistical approaches for the modeling of parameterized representations of speech signals (Zen et al., 2007; Tokuda et al., 2013; Zen et al., 2013; Qian et al., 2014; Zen and Senior, 2014).

Statistical parametric speech synthesis techniques are capable of achieving high levels of intelligibility (King, 2014). Although synthetic speech can be clear and understandable, it is often bland and monotonous. Therefore, proper modeling and generation of natural speech prosody is still a largely unsolved problem (Yu, 2012). This is relevant especially in the context of expressive speech synthesis, where speech is expected to be fluid and captivating.

The study of prosody describes the mechanisms that speakers use to structure information into meaningful constituents of varying temporal resolution (e.g. syllables, words, or phrases) as well as the mechanisms used to highlight the relative importance of such units (e.g. through prominence, emphasis). These phenomena are of particular importance for speech synthesis, as good control of prosodic phenomena leads to more natural synthetic speech.

In general, prosody is concerned with aspects of the speech signal that relate to units that are longer than phones (segment) such as vowels or consonants. Such aspects may be, for example, pitch accenting, distribution of pauses, boundary tones, among others (see Section 3.1 for further details). Listeners may perceive the same melody or rhythm in different utterances, and the same segmental sequence can be spoken with different prosodic properties to convey a different message. For this reason, prosody is commonly accepted to be inherently *suprasegmental* (Nooteboom, 1997; Ladd, 2008; Yu, 2012; Xu, 2012). It is manifested above the segment (phone) and governed by longer units within the utterance (for example, at the syllable, word, or phrase levels), and beyond the utterance, at the discourse level (Wennerstrom, 2001). In the context of this thesis, we generalize the term suprasegmental to denote acoustic phenomena as well as textual units that are above the segment.

In speech synthesis, common techniques for the modeling of speech prosody, and speech in general, operate mainly on very short intervals, either at the state or frame levels, which may span several milliseconds of speech. These techniques allow highly intelligible text-to-speech systems to focus on short-term variation that is mainly related to the phone-level. Even though wider context is taken into account, speech still sounds bland and monotonous, with such context having limited impact on the naturalness of synthetic speech (Cernak et al., 2013).

Recent approaches in statistical parametric speech synthesis have attempted to explore the hierarchical nature of prosody through the integration of multiple temporal levels. However, very few approaches focus on signal and/or linguistic representations. That is, they have tried to learn better prosody, but without a clear and proper understanding of both linguistic and signal effects.

The main hypothesis I propose to explore in this thesis can be summarized in the following statement:

> *More natural synthesis of fundamental frequency can be achieved by exploring complex interactions of suprasegmental units in terms of linguistic representations, acoustic representations, and the mapping between them.*

It is the main claim of this thesis that a stronger understanding of suprasegmental contexts is essential for the natural modeling and generation of speech prosody. This understanding must occur at the level of linguistic inputs, acoustic outputs, as well as the function mapping one to the other. We may consider a simplification of this scenario by dividing the main claim into three separate, but overlapping, sub-problems.

Sub-problem 1*: representations of acoustic signals*

In statistical parametric speech synthesis, speech frames can be defined over 5 ms intervals. Within each interval, 40 to 60 parameters might be used to represent the spectral envelope, and 5 parameters to describe aperiodic excitation (King, 2011). Dynamic features are added to these two acoustic streams, corresponding to the first and second derivatives (termed frequently delta and delta-deltas). However, fundamental frequency uses only 1 parameter per speech frame and 2 parameters corresponding to their dynamic features.

In the case of HMM-based speech synthesis, acoustic models are defined over states, with each state corresponding to sub-phone acoustic properties. In the case of DNN-based speech synthesis, acoustic models are normally defined over states or frames. At synthesis time, generation is done separately for each speech frame, essentially being a framework modeling speech only in very short intervals.[1]

Work in this sub-problem aims at identifying accurate representations of acoustic signals that are capable of capturing variation over units of differing temporal resolution. For the purposes of this thesis, we limit ourselves to the fundamental frequency (*f0*) of the waveform, which is one of the acoustic correlates of prosody (see Section 3.1 for details).

[1] There is, naturally, a large amount of work approaching this problem, using, for example, recurrences in deep neural networks (e.g. Fernandez et al. (2014)). Sections 2.2.2 and 2.2.3 give an overview of acoustic modeling frameworks for statistical parametric speech synthesis. Section 3.2.2 overviews hierarchical modeling of fundamental frequency for speech synthesis.

***Sub-problem 2**: representation of linguistic contexts*

Current methodologies model context-dependent phones, where the context is defined by a set of shallow segmental and suprasegmental features. An example from a conventional text-to-speech system can be found in Appendix A. This approach works well with the spectral envelope, as it is mostly marked by short-term variation defined over a clear symbolic representation (phonetic representation of segments that can be inferred from the text).

However, prosody does not have a clear symbolic representation that can be easily inferred from the text. This has been previously termed *the lack of reference problem* (Xu, 2012). In current systems, most suprasegmental features are typically related to the number of segments/syllables/words within prosodic phrases or utterances, and may add little to the naturalness of synthetic speech (Cernak et al., 2013). These representations of suprasegmental contexts can be inferred directly from the text. For example, syllable stress and syllable boundaries may be taken from a lexicon and phrase boundaries may be predicted by models that are trained on small out-of-domain datasets (see Section 2.1.1).

One possible reason for the minor impact of these features on the naturalness of synthetic speech may be the automatic labeling of higher-level phenomena, such as pitch accents or phrase-breaks. Because manual annotation of large databases such as audiobook recordings tends to be costly, researchers make use of automatic labeling methods to annotate their databases. These approaches tend to generate prediction errors, which lead to mismatches between annotated labels and acoustic events in the training data, generally leading to poor modeling of the suprasegmental aspects of speech (see Section 3.2.3).

This sub-problem is concerned with the task of finding representations of prosodic contexts that would allow the model to better describe the variations in the speech signal associated with longer temporal units (such as syllables or words), while remaining close to the acoustic data used for training.

Sub-problem 3*: mapping acoustic and context representations*

This sub-problem is concerned with the mapping of context and acoustic representations. This is the most widely explored problem in previous work and it involves the modeling of prosodic variations at suprasegmental levels.

In this work, we propose to adopt the findings of sub-problems 1 and 2, and explore methodologies to predict one as a function of the other. For example, given representations of the acoustic signal and linguistic contexts, we propose to explore modeling architectures that could leverage their hierarchical structure. Even though a large body of work has been done in this topic, we still consider sub-problem 3 to be unsolved, as current approaches either fail to yield improvements at longer temporal domains or were designed with traditional acoustic and context representations in mind. An overview of these methods and their findings will be given in section 3.2.2.

Each of these three sub-problems is an aspect of the larger question proposed above. And, as would be expected, each of them contains a great deal of complexity and numerous lines of research. This thesis does not aim to solve them, but instead, it aims to provide some contributions to their understanding and begin to provide solutions to more effective ways of modeling prosodic properties in the context of statistical parametric speech synthesis.

We therefore title this thesis

> *Suprasegmental representations for the modeling of fundamental frequency in statistical parametric speech synthesis*

The term *suprasegmental* is chosen to refer to the inherent property of prosody being governed at temporal intervals longer than the phonetic segment. In the context of this thesis, it refers to linguistic units at levels higher than the phone (segment), such as syllables, words, clitic-groups, phrases, and utterances. The term *representations* refers to transformations performed on linguistic and acoustic contexts in order to capture relevant variation at these suprasegmental levels.

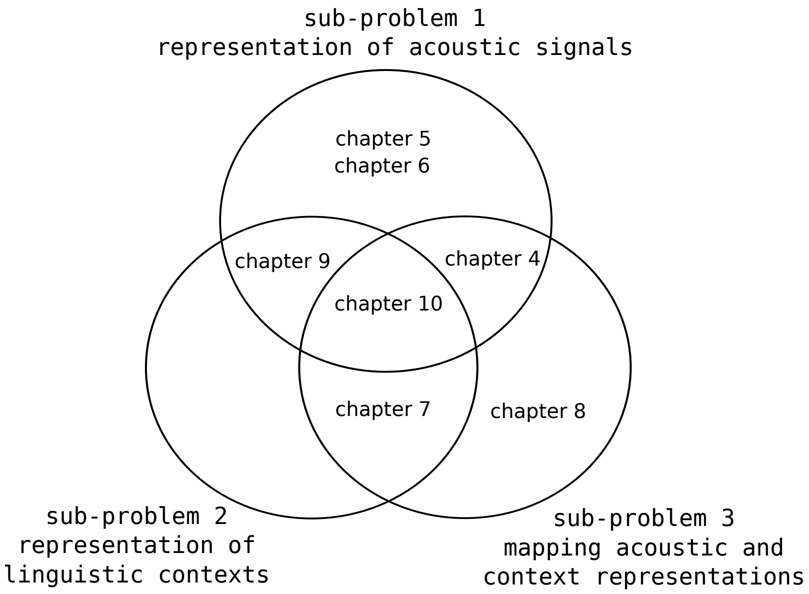

Figure 1.1: Venn diagram illustrating the organization of this thesis with respect to the 3 sub-problems of the main problem.

Two of the acoustic properties most commonly associated with prosody are the fundamental frequency and duration. These are modeled explicitly as acoustic parameters by typical acoustic models for statistical parametric speech synthesis (cf. Chapter 2). For the purposes of the proposed investigations, we opt to focus solely on *fundamental frequency*. We leave investigations focusing on duration and other acoustic correlates of prosody for future work, although we do acknowledge that their interaction should not be disregarded. Finally, most of the work will be developed for *statistical parametric speech synthesis*, considering state-of-the-art techniques in this framework. Chapters 4 and 5 focus on HMM-based speech synthesis, while the remaining chapters focus on DNN-based speech synthesis.

1.2 Thesis overview

This thesis is based primarily on work published in various papers, with chapters 4-9 covering and/or extending the work of a single paper. Figure 1.1 illustrates how the contributing chapters of this thesis interact with the main claim and its three sub-problems. An alternative representation is shown in Figure 1.2, illustrating the dependencies between the work presented in each chapter. Below, we provide a brief summary of the contents and contributions of each chapter, as well as previous presentations of work that they are based on.

Chapters 2 and 3 provide overall background to the techniques used throughout this thesis. Chapter 2 gives an overview of common approaches to text-to-speech synthesis, focusing on Hidden Markov Model and Deep Neural Network based methods. Chapter 3 provides a brief introduction to speech prosody, covering the theoretical foundations that motivate the main claim and contributions of this thesis. The same chapter gives an overview of suprasegmental approaches to the modeling of *f0* for statistical parametric speech synthesis.

Chapters 4, 5, and 6 can be grouped under the topic **multi-level acoustic representations of fundamental frequency**. The main theme throughout these chapters is the hierarchical decomposition of *f0* contours and its evaluation under standard text-to-speech systems. This corresponds to sub-problem 1, *representations of acoustic signals*. Chapter 4 proposes a multi-level representation of *f0* using the Continuous Wavelet Transform (CWT) and the Discrete Cosine Transform (DCT). Because this proposed representation allows models to be defined at suprasegmental levels, we include this chapter at the intersection of sub-problem 1 and sub-problem 2 in Figure 1.1. The proposed multi-level representation is governed by two strong assumptions. The first assumption is investigated with a set of perceptual experiments, presented in Chapter 5. The second is investigated in Chapter 6, where earlier findings are used to derive a linguistically and perceptually motivated wavelet-based decomposition of *f0*. The work that forms the basis of these chapters was initially presented in the following three publications:

1. **A multi-level representation of *f0* using the continuous wavelet transform and the discrete cosine transform.**
   Manuel Sam Ribeiro and Robert A. J. Clark
   In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Brisbane, Australia. April, 2015

2. **A perceptual investigation of wavelet-based decomposition of *f0* for text-to-speech synthesis.**
   Manuel Sam Ribeiro, Junichi Yamagishi, and Robert A. J. Clark
   In Proceedings of Interspeech. Dresden, Germany. September, 2015

3. **Wavelet-based decomposition of *f0* as a secondary task for DNN-based speech synthesis with multi-task learning.**
   Manuel Sam Ribeiro, Oliver Watts, Junichi Yamagishi, and Robert A. J. Clark
   In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Shanghai, China. March, 2016

Chapters 7, 8, and 9 group work and contributions under the topic **hierarchical systems and suprasegmental input representations**. This work explores hierarchical models and their ability to leverage information that is defined at higher linguistic levels, such as syllables or words. This corresponds to work falling under sub-problem 2, *representations of linguistic contexts*, and sub-problem 3, *mapping acoustic and linguistic representations*. Chapter 7 proposes a top-down hierarchical modeling approach in the form of a cascaded deep neural network where suprasegmental effects are modeled first at syllable-level and then channeled to a frame-level model. Additional linguistic representations are evaluated with the proposed hierarchical model, thus falling under sub-problems 2 and 3 in Figure 1.1. The results of this work are further investigated in Chapter 8, which evaluates the behavior of parallel and cascaded deep neural networks with different subsets of linguistic features. Finally, in Chapter 9, a simple approach to learn feature representations at suprasegmental levels (words and syllables) by taking counts over acoustic signals is presented. Because the linguistic representations described in Chapter 9 require a discretized representation of acoustic signals over textual units, this work falls under sub-problems 1 and 2, as shown in Figure 1.1.

The work that forms the basis for these chapters was initially presented in the following three publications:

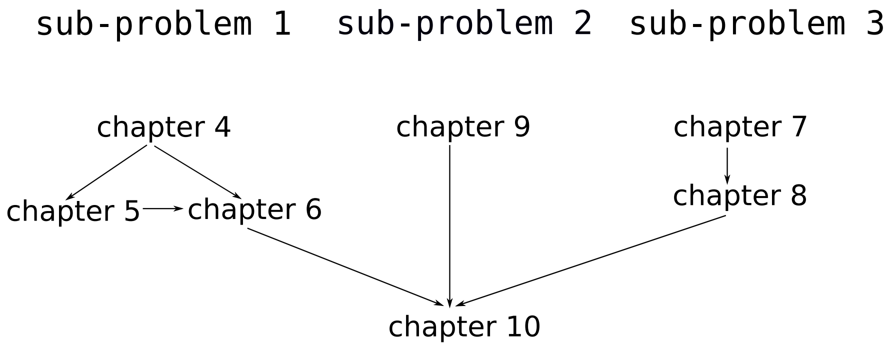

Figure 1.2: Hierarchical structure of each chapter with respect to other chapters and to the 3 sub-problems of the main thesis claim.

1. **Syllable-level representations of suprasegmental features for DNN-based text-to-speech synthesis.**
   Manuel Sam Ribeiro, Oliver Watts, and Junichi Yamagishi
   In Proceedings of Interspeech. San Francisco, United States. September, 2016

2. **Parallel and cascaded deep neural networks for text-to-speech synthesis.**
   Manuel Sam Ribeiro, Oliver Watts, and Junichi Yamagishi
   In Proceedings of Speech Synthesis Workshop (SSW). Sunnyvale, United States. September, 2016

3. **Learning word vector representations based on acoustic counts**
   Manuel Sam Ribeiro, Oliver Watts, and Junichi Yamagishi
   In Proceedings of Interspeech. Stockholm, Sweden. August, 2017

Finally, Chapter 10 provides a global discussion of the main contributions of the thesis, as well as a perceptual evaluation of the main techniques proposed within a single experiment.

Chapter 2

Statistical parametric speech synthesis

This chapter presents background material for statistical parametric speech synthesis. The overall architecture of text-to-speech systems is given, covering common modules found in the front-end and various approaches to the back-end of such systems. Details are given for speech synthesis based on hidden Markov models (HMMs) and deep neural networks (DNNs). The chapter ends with a brief overview of objective and subjective evaluation methodologies used in this thesis and commonly used in the field of speech synthesis.

2.1 Text-to-speech synthesis

A text-to-speech system is often divided into two large components: the front-end and the back-end. The front-end is concerned with text processing, converting raw text input into an intermediate symbolic representation, while the back-end is concerned with the transformation of that intermediate representation into a waveform. The remainder of this section elaborates on the details of these components.

2.1.1 Front-end

The purpose of the front-end, or *text processor*, of a speech synthesizer is to narrow the gap between text and speech. That is, given a text input, it is the goal of the front-end to generate what is often termed a *linguistic specification*. Various modules within the front-end, given the input text, generate a symbolic representation of the speech utterance that identifies details about how it will be spoken. Typically, the information found in this representation may contain a phonetic transcription, stress assignment, syllable boundaries, word part of speech, and intonational phrase breaks. For this reason, the front-end tends to be heavily language-dependent and its creation often requires linguistic expertise.

Conventional systems mostly rely on knowledge-based features dependent on annotated linguistic resources. There have been, however, various attempts at reducing the front-end's need for expert knowledge. For example, Watts (2012) proposes several unsupervised approaches to derive front-end modules for low-resource languages.

Other work has focused on reducing the traditional front-end/back-end modularity of text-to-speech systems. The work of Oura et al. (2008), still relying on linguistic features, proposed a method to simultaneously learn front-end and back-end components. Recently, various approaches have proposed end-to-end methodologies, in which an acoustic model accepts as input sequences of characters and outputs the corresponding waveform, thus entirely bypassing the typical front-end modules (Sotelo et al., 2017; Wang et al., 2017b).

We adopt in this thesis the traditional chained modular front-end. Each module is tasked with extracting some information from the input text before passing it along to the subsequent module in the chain. The output is the linguistic specification for the input text. We provide here a brief overview of common components found in this framework.

Given an input string to process, the first task is often **text segmentation**. This involves segmenting the input into meaningful units such as words. **Text normalization** is concerned with the task of normalizing non-standard words (NSWs) such as abbreviations, numbers, addresses, contractions, among others.

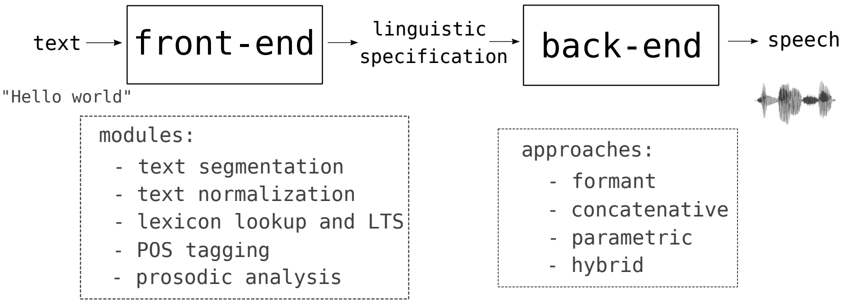

Figure 2.1: Pipeline of a text-to-speech system, including common modules found in the front-end and various approaches to the back-end.

A **lexicon** is a dictionary mapping words to their phonetic transcription, syllabification, stress, and occasionally, part of speech. The lexicon tends to be a resource of high value, as it often requires expert knowledge, and needs to be carefully annotated and reviewed by language experts. However, it is the goal of text-to-speech synthesis to synthesize *any* sentence or word. And the lexicon will not be able to cover all examples in a given language. Therefore, if a word is not covered by the lexicon, a sequence of phones is predicted by a **letter-to-sound** (LTS) module, also called *grapheme-to-phoneme* (G2P) conversion. Automatic syllabification and stress assignment is also applied to the phone sequence. Similarly, a **part of speech tagger** can be applied if the word is not found in the lexicon, if the lexicon does not cover part of speech, or if some type of disambiguation is necessary for phonetic purposes (e.g. 'to *present* a case' or 'to buy a *present*'). In some languages, **post-lexical rules** may be applied to the phone sequence to account for *sandhi* phenomena. These cover cases where the phone sequence of a word changes due to its context. For example, phones at the boundaries of words may change depending on the phones of the neighboring word.

The final group of modules is often considered to deal with prosodic analysis. These involve **phrase-break prediction**, whereby short pauses and breaks are inserted in the sentence, based on punctuation or on part of speech tag sequences. **Pitch accent and boundary tone prediction** may assign labels describing the intonational pattern and prosodic structure of the spoken utterance. For example, a common labeling convention for the English languages is the ToBI (Tone and Break Indices) framework (Silverman et al., 1992). This type labeling is often expensive, as it requires carefully annotated databases. Predictors of ToBI labels that generalize from small annotated datasets may be used. The Boston Radio News Corpus (BURN, Ostendorf et al. (1995)) is a commonly used dataset to train predictors of ToBI labels for the English language.

2.1.2 Back-end

The task of the back-end of a speech synthesizer is to convert the intermediate linguistic specification into a synthetic speech waveform. For this reason, the back-end can also be called the *waveform generator*. Given that most linguistic processing has been done by the front-end, the back-end tends to be non-language-specific. To solve the problem of mapping from a linguistic representation to a waveform, various approaches have been used throughout the years.

Formant synthesis (also **rule-based synthesis**) is one of the earliest fundamental paradigms proposed for the back-end of a text-to-speech system. These approaches typically define a set of rules to artificially generate a waveform from a set of acoustic parameters such as formant frequencies, amplitudes, or bandwidths (Holmes et al., 1964; Klatt, 1980; Balyan et al., 2013). Formant synthesizers have the advantage of being very compact, thus requiring a low memory footprint. But because these systems are based on a set of rules, they are language-specific and typically require the knowledge of experts. Similarly, the technique limits the quality of synthetic speech, which, although intelligible, often sounds unnatural (Taylor, 2009, §13.2.7).

Concatenative synthesis (also **sample-based synthesis**) is perhaps the most common method for waveform generation in commercial systems. This approach focuses on the concatenation of pre-recorded units of speech to generate new waveforms. At one extreme, this class of models includes basic systems in domain-specific scenarios, whose task is to string together words from a small closed vocabulary. We might find such systems, for example, in train stations or customer support phone lines. However, systems aiming for more generic synthesis focus on the concatenation of smaller units. Such units may begin, for example, at the mid-point of a phone and end at the mid-point of the following phone, thus capturing the co-articulation between the two phones. These units are often termed diphones. Systems using a minimal database with a single diphone sample are said to be *diphone synthesizers* (Moulines and Charpentier, 1990).

Extensions of this idea vary the type and number of units present in the database. This generalization is referred to as *unit selection* (Iwahashi et al., 1992; Hunt and Black, 1996; Campbell and Black, 1997) . Unit selection systems use a very large database with multiple samples of the same unit. The task of the waveform generator is then to select the optimal unit sequence given an input linguistic specification.

Parametric synthesis (also **model-based synthesis**) uses parametric representations of speech waveforms, which are modeled via statistical frameworks. For this reason, these systems are often grouped under the term *statistical parametric speech synthesis* (SPSS, Black et al. (2007); Zen et al. (2009b)). This thesis is concerned with this approach to waveform generation. Therefore, Section 2.2 will provide further details on this class of approaches.

Parametric systems offer various advantages over concatenative systems. For example, it is easy to see how unit selection systems can be limited by their database: larger databases allow the system to be more flexible, but also increase the amount of resources needed. Parametric voices are flexible when it comes to the manipulation and control of acoustic parameters. This flexibility makes them attractive for various tasks such as speaker adaptation or transformations of speaking styles and emotion. Additionally, parametric systems tend to benefit from a very small footprint when compared to standard unit selection systems. However, parametric voices suffer from various disadvantages. In fact, when directly compared with unit selection systems, they consistently underperform (King, 2014). Zen et al. (2009b) list three main causes for this underperformance in terms of speech quality and speaker similarity: the buzziness caused by parameterization of the speech signal, the accuracy of the acoustic models, and the over-smoothing of speech parameters.

Hybrid synthesis is a term covering various techniques to combine unit selection and parametric methodologies. The most common hybrid approach uses a statistical framework to generate a sequence of acoustic parameters that are then used to guide the selection of units from the database. This is a different approach from the original unit selection proposal by Hunt and Black (1996), in which the *target* unit was selected based on the linguistic specification. Common approaches use acoustic parameters generated from HMMs (Yan et al., 2010; Qian et al., 2013) although recently DNNs have also been used (Merritt et al., 2016; Wan et al., 2017; Capes et al., 2017). A different method allows parametrically synthesized units to be selected when the speech database does not contain suitable natural candidates (Okubo et al., 2006; Aylett and Yamagishi, 2008; Pollet and Breen, 2008). Although this idea is still hybrid in principle, these methods have often been termed *multiform synthesis*.

2.2 Statistical parametric speech synthesis

2.2.1 Parametric representations of speech

In the context of speech synthesis, a *vocoder*, or voice encoder, transforms the speech waveform into a set of parameters that may be modeled statistically. Common approaches are based on the *source-filter model* of speech production. This model makes the assumption that speech is produced by first generating a source signal, which can be intuitively understood as air exiting the lungs and passing through the vocal folds. The position of the vocal tract articulators (tongue, lips, oral, and nasal cavities) then act as a filter on the source signal. The source-filter model assumes these two components are independent and vocoders aim to find representations that separate the effects of source and filter.

Vocoders extract parameters over speech windows, referred to as a *speech frame*, and may, in common implementations, span 25ms. Each frame is assigned source (or excitation) parameters such as fundamental frequency and voicing information. Some vocoders include extra excitation parameters, such as band aperiodicities: this is the case of STRAIGHT (Kawahara et al., 1999, 2001) or the WORLD (Morise et al., 2016) vocoders. For the filter (or spectral envelope) parameters, mel-cepstrum coefficients (Fukada et al., 1992) are often used. Alternatives use mel-generalized cepstral coefficients (Tokuda et al., 1994) or line-spectral pairs (LSPs, Itakura (1975)). The speech waveform can be analyzed and reconstructed with minimal error via these speech parameters. Other approaches using the source-filter model have been proposed in the GlottHMM (Raitio et al., 2011) and GlottDNN (Airaksinen et al., 2016) vocoders. Alternative approaches to the source-filter model use a sinusoidal model of speech (Stylianou et al., 1995; Shechtman and Sorin, 2010; Degottex and Stylianou, 2012; Erro et al., 2014; Hu, 2016).

2.2.2 Hidden Markov model based speech synthesis

Hidden Markov models (HMMs) are statistical time series models commonly used in a variety of speech and language processing applications. An N-state HMM is defined by the set of parameters:

$$\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi}) \tag{2.1}$$

where $\mathbf{A}$ is a transition probability matrix denoting the probability of moving from state i to state j, such that $\sum_{j=1}^{N} a_{ij} = 1$ for all states i. $\mathbf{B}$ is the set of output probability distributions $b_i(\mathbf{o}_t)$, where $b_i(\mathbf{o}_t) \in \mathbf{B}$ for all states i. $b_i(\mathbf{o}_t)$ denotes the probability of state i generating an observation $\boldsymbol{o}$ at time t. $\mathbf{\Pi}$ is an initial probability distribution over states, where π_i denotes the probability of state i being an initial state.

We assume we are given a collection of T observation vectors $\boldsymbol{O} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, ..., \boldsymbol{o}_T^\top\right]^\top$, corresponding to vocoder parameters extracted from natural speech, and their corresponding linguistic specifications $\mathcal{W}$. It is the goal of the training stage of the HMM to find the set of parameters λ that maximize the likelihood of the training data:

$$\lambda_{max} = \underset{\lambda}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, \mathcal{W}) \tag{2.2}$$

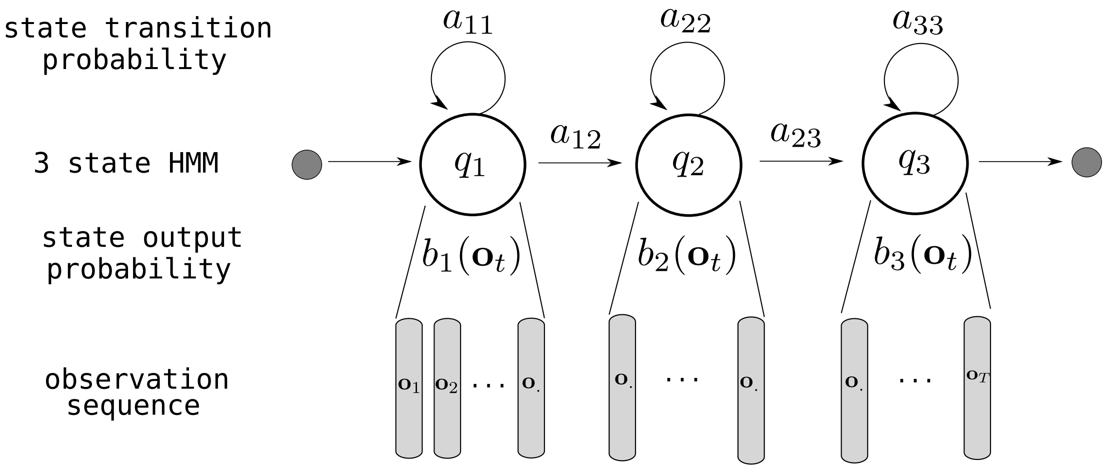

Figure 2.2: Three state left-to-right HMM. State transition probabilities are denoted by a_{ij} and state output probabilities by $b_i(\boldsymbol{o}_t)$.

Figure 2.2 illustrates a typical left-to-right HMM. In the figure, 3 emitting states are illustrated: $Q = q_1, q_2, q_3$, with 2 additional dummy states denoting beginning and end. State transition probability is modeled via the transition probability matrix A and state output probabilities by $b_i(\boldsymbol{o}_t)$.

State output probabilities

For continuous data such as speech parameters, the state output probability distribution is often modeled by a single multivariate Gaussian distribution. This can be represented as:

$$b_i(\mathbf{o}_t) = \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.3}$$

$$= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_t)\right\} \tag{2.4}$$

where $\boldsymbol{o}_t$ is an observation vector of d vocoder parameters at time frame t, $\boldsymbol{\mu}_i$ is a $d \times 1$ Gaussian mean vector associated with state i and $\boldsymbol{\Sigma}_i$ is the $d \times d$ covariance matrix associated with state i.

A Gaussian mixture model (GMM) may be used instead to model the state output probabilities (Rabiner, 1989). In this case, Equation 2.3 is expressed as a weighted sum of M Gaussian components, each one given as in Equation 2.4.

The observation vector $\boldsymbol{o}_t$ may be organized into multiple data streams. These could correspond to the various features extracted by the vocoder. For example, *f0* and mel-cepstral coefficients can be treated as separate data streams. If there are S data streams, then $\boldsymbol{o}_t = \left[\boldsymbol{o}_{t,1}^\top, \boldsymbol{o}_{t,2}^\top, ..., \boldsymbol{o}_{t,S}^\top,\right]^\top$ and the state output probabilities are then defined by:

$$b_i(\mathbf{o}_t) = \prod_{s=1}^{S} b_{is}(\mathbf{o}_{ts}) \tag{2.5}$$

Training HMMs

In the context of text-to-speech synthesis, HMMs are normally defined at the phone-level. Observation sequences are frames of speech parameters extracted by a vocoder. Each state of the model therefore corresponds to a sub-phone sequence of speech frames. We let $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$ be an observation sequence of T frames and $\boldsymbol{q^*} = (q_1^*, q_2^*, ..., q_T^*)$ be the optimal state sequence aligned with $\boldsymbol{O}$.

If this alignment between HMM states and observations were given, deriving the parameters of the HMM would be trivial. For state i, transition probabilities can be estimated simply by counting the number of times we move from state i to state j and then dividing by the total number of transitions out of state i. Initial state probabilities are not relevant, since, as illustrated in Figure 2.2, left-to-right HMMs are typically used for text-to-speech synthesis. The state output probabilities can simply be estimated by taking the mean and covariance of all observations aligned with state i.

However, for HMM training, the states and observations are not initially aligned. Common recipes therefore begin with a **flat start** initialization of the HMMs. This assumes a uniform segmentation of each observation sequence with each state of the model. Parameters can then be initialized given this rough alignment. This generally produces a weak model, given that the observation sequence is poorly segmented and allocated to HMM states.

With flat-started models, it is possible to find the most likely state sequence for observations. The best state sequence involves computing $\boldsymbol{q^*} = \text{argmax}_q\, P(\boldsymbol{O}, \boldsymbol{q}|\lambda)$. This can be achieved efficiently using dynamic programming with the Viterbi algorithm. Knowledge of the optimal state sequence allows us to identify a new alignment between the HMM states and the observations. A new set of model parameters $\hat{\lambda}$ can them be re-estimated using this hard-alignment, such that $P(\boldsymbol{O}|\hat{\lambda}) \geq P(\boldsymbol{O}|\lambda)$. This training phase is often called **Viterbi training**.

Given fully trained HMMs, finding the most likely state sequence can also be used to align model states with observation sequences. In this case, we do not re-estimate the parameters of the model, but we use the hard boundaries identified by the model for further processing. If the observation sequence corresponds to a speech utterance, the method allows likely boundaries for linguistic units such as phones or syllables to be estimated. The process is commonly referred to as *forced alignment*.

A downside of the Viterbi training stage is that we are forced to make hard decisions at state boundaries. A particular observation is allocated to a single state and does not contribute to the parameter estimation of the remaining states. This is because this approach fails to account for all possible state sequences. A stronger optimization algorithm therefore considers that an observation can, in fact, be generated by any state of the HMM.

These constraints can be incorporated in training using an expectation maximization (EM) algorithm: the Baum-Welch algorithm. In this case, all state and observation alignments are considered in the re-estimation of the model parameters. The *hard alignment* of the Viterbi phase causes each observation to be assigned to a single state. In **EM training**, each observation is assigned to all states, and it is weighted by the probability of the assigned state having generated it. This generates a *soft alignment* between states and observations.

Common training recipes begin with a flat-start initialization of the HMMs. Several iterations of Viterbi training are then performed until the likelihood of the training data cannot be further improved. This is then followed by several iterations of EM training.

The description of the training process for HMMs is given here in passing. For a complete understanding, a good grasp of Viterbi and the Baum-Welch algorithms are necessary. These are well-established and well-documented tech-

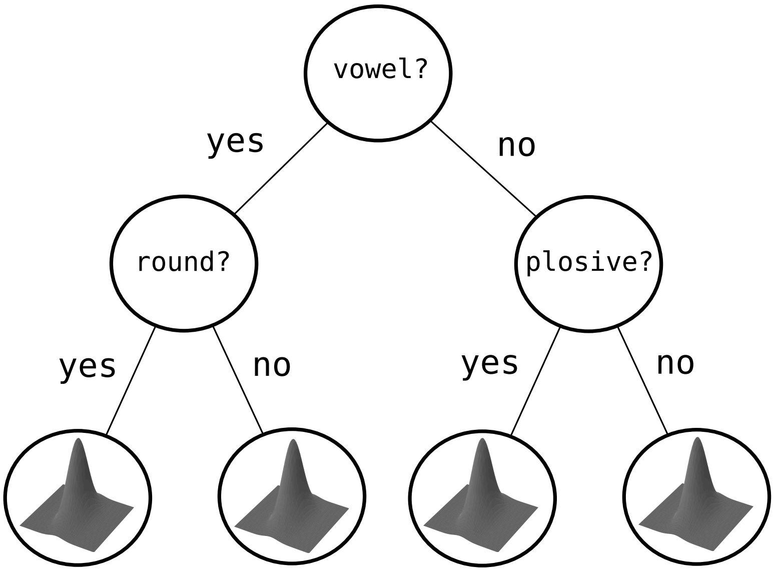

Figure 2.3: Sample decision tree used for state tying. Binary questions are asked at each node to partition the data. Each leaf node models all clustered observation vectors with a Gaussian distribution.

niques in the literature. For this reason, we do not cover them in detail and instead we refer the reader to more detailed descriptions. Taylor (2009) provides a good example for text-to-speech cases and Rabiner (1989) provides good details for the general cases.

Context dependency and parameter tying

In speech synthesis, acoustic parameters are affected by a large variety of linguistic contexts. Although spectral parameters are mostly influenced by the local phone context, prosodic features such as *f0* or duration are mostly affected by prosodic structure. In the examples above, we have mentioned that HMMs can be defined at phone level. However, in an ideal scenario, given enough data, a separate HMM would be trained for each observed unique linguistic context. In reality, however, there are many contextual factors that are either not relevant for a given acoustic parameter or are unseen during training. This would also cause the training data to be too sparse across models to estimate reliable parameters.

To deal with this problem, parameter tying techniques are employed (Odell, 1995). This approach ties model parameters across states to derive robust distributions that can generalize to unseen contexts. Parameter tying is performed automatically by clustering HMM states. This is done in a hierarchical fashion using stream-dependent multivariate classification and regression trees (CARTs). Figure 2.3 illustrates this process. Each HMM state is clustered with other states based on the features occurring in the corresponding linguistic specification. Because different speech parameters might be affected by different contextual factors, they are clustered separately. Given S data streams, S stream-dependent decision trees are learned during the training process (Yoshimura et al., 1999). Tree size can be determined based on the minimum description length (MDS) criterion (Shinoda and Watanabe, 2001). Approaches using multiple CART models, such as Random Forests, have also been proposed (Black and Muthukumar, 2015).

Common training recipes, such as the one used in this thesis, first train context-independent monophone HMMs. After this initial optimization, the models are decoupled and further optimized as context-dependent HMMs. Each HMM state i and data stream s is therefore associated with a leaf of a decision tree. The acoustic parameters that correspond to that state are pooled to derive $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ used in Equations 2.3 and 2.4. If using multiple data streams, then the state output probability of state i is computed according to Equation 2.5.

Duration and fundamental frequency modeling

An HMM, as described so far, may model duration via the state transition probability matrix $\boldsymbol{A}$. However, in such a framework, as the duration increases, the state duration probability decays exponentially. This is not an accurate model of speech duration. An alternative approach models state duration explicitly with a Gaussian distribution (Yoshimura et al., 1998). Because the Markov assumption is violated if duration is modeled explicitly, such models are properly termed **hidden semi-Markov models** (HSMMs, Zen et al. (2004)).

The representation of fundamental frequency used for statistical modeling is composed of voiced frames and unvoiced frames. Voiced frames are considered to be samples from a one-dimensional continuous distribution. Unvoiced frames do not have values associated with them, so they are considered to be samples from a zero-dimensional discrete distribution. To account for this property of the *f0* signal, **multi-space probability distribution HMMs** (MSD-HMMs) can be used (Tokuda et al., 2002).

For the modeling of *f0*, MSD-HMMs consider two sample spaces, corresponding to voice and unvoiced distributions. Each space is associated a space weight w_g such that $\sum_{g=1}^{2} w_g = 1$. The one-dimensional space associated with voice frames is modeled by a normal probability density function, as Equations 2.3 and 2.4. And the zero-dimensional space associated with unvoiced frames contains only one element.

In HMM-based speech synthesis, duration and *f0* are modeled as independent data streams. An illustration of this process is given in Figure 2.4. Furthermore, because of the inability to estimate dynamic features at the boundaries of voice and unvoiced frames, first and second order time derivatives of the *f0* signal are also treated as separate data streams. Alternative approaches that consider the *f0* to be continuous have been explored (Yu and Young, 2011). In the continuous *f0* HMM (CF-HMM), the signal is modeled as a single stream, together with dynamic features. Voicing decision is modeled explicitly as a separate data stream. A very similar approach was later adopted for deep neural network speech synthesis (Zen et al., 2013).

Parameter generation and synthesis

At synthesis time, it is our goal to generate acoustic parameters for an unseen test utterance. For this task, we are given an utterance-level linguistic specification $\mathcal{W}$. The linguistic specification is used to traverse the decision trees and find a sequence of context-dependent HMMs with parameters λ. These models are concatenated to form an utterance-level HMM. Although $\mathcal{W}$ tells us which HMMs to use, it does not tell us which state sequence to follow through the large chain of models. The generation problem involves selecting the acoustic observation sequence $\boldsymbol{O}$ that maximizes $P(\boldsymbol{O}|\lambda)$.[1]

[1] The derivation of the parameter generation process presented here closely follows those of (Zen et al., 2009b; Tokuda et al., 2013). Additional notes and examples were also taken from (Watts, 2012, §2.1.2) and (Valentini-Botinhao, 2013, §3.2.4).

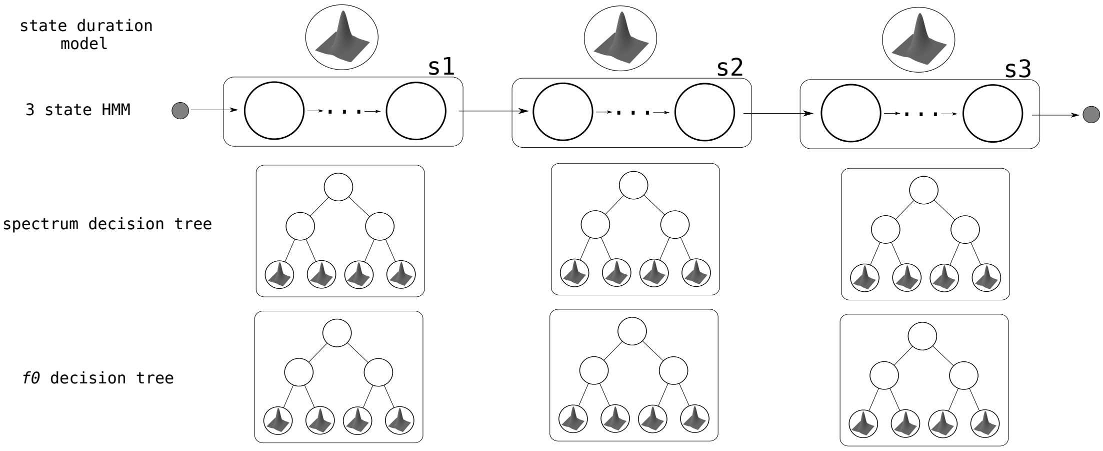

Figure 2.4: Illustration of a three state context-dependent HSMM. State duration is modeled with a Gaussian distribution and tied with decision trees (not illustrated). State output probabilities are modeled with Gaussian distributions and tied with decision trees. Spectrum and source parameters are tied separately.

This can be formulated as:

$$\begin{aligned}
\boldsymbol{O}_{max} &= \operatorname*{argmax}_{\boldsymbol{O}} p(\boldsymbol{O}|\lambda) && (2.6)\\
&= \operatorname*{argmax}_{\boldsymbol{O}} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{O}, \boldsymbol{q}|\lambda) && (2.7)\\
&\approx \operatorname*{argmax}_{\boldsymbol{O},\boldsymbol{q}} p(\boldsymbol{O}, \boldsymbol{q}|\lambda) && (2.8)\\
&= \operatorname*{argmax}_{\boldsymbol{O},\boldsymbol{q}} p(\boldsymbol{O}|\boldsymbol{q}, \lambda) P(\boldsymbol{q}|\lambda) && (2.9)
\end{aligned}$$

where $\boldsymbol{q} = (q_1, ..., q_T)$ is a state sequence corresponding to the observation sequence $\boldsymbol{O} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, ..., \boldsymbol{o}_T^\top\right]^\top$. This derivation results is two maximization problems:

$$\begin{aligned}
\boldsymbol{q}_{max} &= \operatorname*{argmax}_{\boldsymbol{q}} P(\boldsymbol{q}|\lambda) && (2.10)\\
\boldsymbol{O}_{max} &= \operatorname*{argmax}_{\boldsymbol{O}} p(\boldsymbol{O}|\boldsymbol{q}_{max}, \lambda) && (2.11)
\end{aligned}$$

If using HSMMs, the problem denoted by Equation 2.10 can be solved through the state duration probability distributions. That is, the most likely state sequence for the utterance can be found using the most likely state durations:

$$\boldsymbol{q}_{max} = \underset{\boldsymbol{q}}{\operatorname{argmax}} P(\boldsymbol{q}|\lambda) \tag{2.12}$$

$$= \underset{\boldsymbol{q}}{\operatorname{argmax}} \prod_{k=1}^{K} p_k(d_k) \tag{2.13}$$

where K is the total number of unique states in the utterance-level HSMM and $p_k(d_k)$ is the probability that state k generated d_k observations. Note that, if there are T observations in the utterance, then $\sum_{k=1}^{K} d_k = T$. If state durations are modeled with a Gaussian distribution, then:

$$d_k = \mu_k + \rho\sigma_k^2 \tag{2.14}$$

where

$$\rho = (T - \sum_{k=1}^{K} \mu_k) / \sum_{k=1}^{K} \sigma_k^2 \tag{2.15}$$

and μ_k and σ_k^2 are the mean and variance of the distribution associated with state k, and ρ is a parameter that controls the total duration T and the overall speaking rate of the utterance (Yoshimura et al., 1998). The parameter ρ can be set to 0 if there is no intention to manipulate the speaking rate of the utterance.

Given the most likely state sequence $\boldsymbol{q}_{max}$, we can now find the most likely observation sequence:

$$\boldsymbol{O}_{max} = \underset{\boldsymbol{O}}{\operatorname{argmax}} p(\boldsymbol{O}|\boldsymbol{q}_{max}, \lambda) \tag{2.16}$$

$$= \underset{\boldsymbol{O}}{\operatorname{argmax}} \mathcal{N}(\boldsymbol{O}; \boldsymbol{\mu}_{q_{max}}, \boldsymbol{\Sigma}_{q_{max}}) \tag{2.17}$$

Because each state models observations with a Gaussian distribution, the most likely generated sample will always be its mean vector. If a given state generates several samples, then those samples will be constant throughout the state's duration. This is caused by the conditional independence assumption in the HMM's state output probabilities. That is, at each timestep, the state output probability is independent of the state output probability at the previous and the next timestep. This results is a step-wise trajectory along the utterance with abrupt transitions at state boundaries. To account for the temporally dynamic and smooth transitions of natural speech parameters, additional constraints are imposed on the speech parameter generation algorithm.

Parameters extracted directly from the vocoder are termed *static* features. Additional constraints on the speech parameters are placed through the inclusion of *dynamic features*. These typically correspond to the first and second order time derivatives (delta and delta-deltas) of the static speech parameters. The observation vector is set to be: $\boldsymbol{o}_t = [c_t^\top, \Delta c_t^\top, \Delta^2 c_t^\top]^\top$, where c_t corresponds to the static parameters. The corresponding dynamic features can be determined according to:

$$\Delta^{(n)} c_t = \sum_{\tau=-L(n)}^{L(n)} w_\tau^{(n)} c_t \qquad 0 \leq n \leq 2 \tag{2.18}$$

where $2L^{(n)} + 1$ is the size of window coefficients used to compute the n^{th} order dynamic features. For the 0^{th} order window, we set $L^0 = 0$ and $w_0^{(0)} = 1$.

Given the window coefficients, we can express the observation vector $\boldsymbol{o}_t$ as a linear transformation of the static parameters:[2]

$$\boldsymbol{o}_t = \begin{bmatrix} \mathbf{c}_t \\ \Delta \mathbf{c}_t \\ \Delta^2 \mathbf{c}_t \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_{t-2} \\ c_{t-1} \\ c_t \\ c_{t+1} \\ c_{t+2} \end{bmatrix} \tag{2.19}$$

Which can be rewritten as:

$$\boldsymbol{o}_t = \boldsymbol{W} \boldsymbol{c}^* \tag{2.20}$$

where $\boldsymbol{c}^*$ is a vector denoting a window around the static features c_t.

The window coefficient matrix $\boldsymbol{W}$ is determined for the full utterance of T observations. This results in a $3T$xT matrix, where the 3 corresponds to the

[2] The static parameter c_t is here considered to be a scalar, following Tokuda et al. (2013). In practice, $\boldsymbol{c}_t$ is a vector and the window coefficient matrix $\boldsymbol{W}$ is adjusted accordingly.

order of the acoustic features, which typically are static, delta, and delta-deltas. In this case, the window around c_t spans the full utterance, and this relationship can be written as:

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C} \tag{2.21}$$

Because the relationship between $\boldsymbol{O}$ and $\boldsymbol{C}$ is deterministic via $\boldsymbol{W}$, maximization of Equation 2.17 is equivalent to:

$$\boldsymbol{C}_{max} = \underset{\boldsymbol{C}}{\operatorname{argmax}} \mathcal{N}(\boldsymbol{W}\boldsymbol{C}; \boldsymbol{\mu}_{q_{max}}, \boldsymbol{\Sigma}_{q_{max}}) \tag{2.22}$$

$$= (\boldsymbol{W}^\top \boldsymbol{\Sigma}_{q_{max}}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{q_{max}}^{-1} \boldsymbol{\mu}_{q_{max}} \tag{2.23}$$

Equation 2.23 can be determined by setting the partial derivative of the log of Equation 2.22 to $\mathbf{0}$ (Tokuda et al., 2000). Note that all terms on the right hand side of Equation 2.23 are known. They correspond to the utterance-level mean and covariance matrices for the full set of parameters, and W corresponds to the utterance-level matrix expressing the relationship between static and dynamic parameters.

This method takes into account how speech parameters vary across time, while generating the most likely observation sequence given the learned models. This **Maximum Likelihood Parameter Generation** (MLPG) algorithm provides smoother trajectories of speech parameters that result in an increased quality of synthetic speech (Tokuda et al., 1995, 2000; Tomoki and Tokuda, 2007) . Given the generated acoustic parameters, the synthesis process can be completed with a vocoder, as described in Section 2.2.1.

2.2.3 Deep neural network based speech synthesis

Artificial neural networks (ANN) have regained popularity in recent years in a wide variety of applications, such as computer vision (e.g. Krizhevsky et al. (2012)), speech recognition (e.g. Hinton et al. (2012)), natural language processing (e.g. Collobert and Weston (2008)), among others. These are powerful models that capture the complex interactions between input and output features, given enough training data.

An artificial neural network can be seen as a collection of small processing units called nodes. Each node receives a set of inputs and produces an activation, which is then passed along to other nodes. Typically, nodes can be grouped into layers, and information is passed along the network through the connections between the nodes.

A broad distinction of ANN types can be made by considering whether the network connections are cyclic or acyclic. The group of models that allow cycles in the network are often called *recurrent networks*. Some examples include Elman networks (Elman, 1990), Jordan networks (Jordan, 1990), recursive networks (Goller and Kuchler, 1996), or long short term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). A distinct type of ANNs does not allow network connections to form cycles. These acyclic models are referred to as *feedforward networks* and include, among others, perceptrons (Rosenblatt, 1958), autoencoders (Hinton and Zemel, 1994), or convolutional neural networks (LeCun et al., 1995). A very common type of feedforward neural network is the **multilayer perceptron** (MLP, Rumelhart et al. (1985)). Given its wide usage, it is common in the literature to refer to it simply as a feedforward neural network. The techniques investigated throughout this thesis rely on MLPs, therefore the following sections provide further intuition regarding this model.

Forward propagation

An MLP arranges nodes into layers, with connections being made between each layer in the network. Information is then passed through the network in a feedforward manner. Each node receives information from all the nodes in the layer before and passes its activation to the layer after. A multilayer perceptron can be seen as a function $f(\boldsymbol{x})$ mapping input to output vectors. The parameters of the function are determined by the weights associated with the connections between nodes. This process is illustrated in Figure 2.5.

We are given an input vector $\boldsymbol{x} = [x_1, x_2, ..., x_N]^\top$. For the k-th node in the first hidden layer, we have the corresponding connection weights $\boldsymbol{w_k} =$

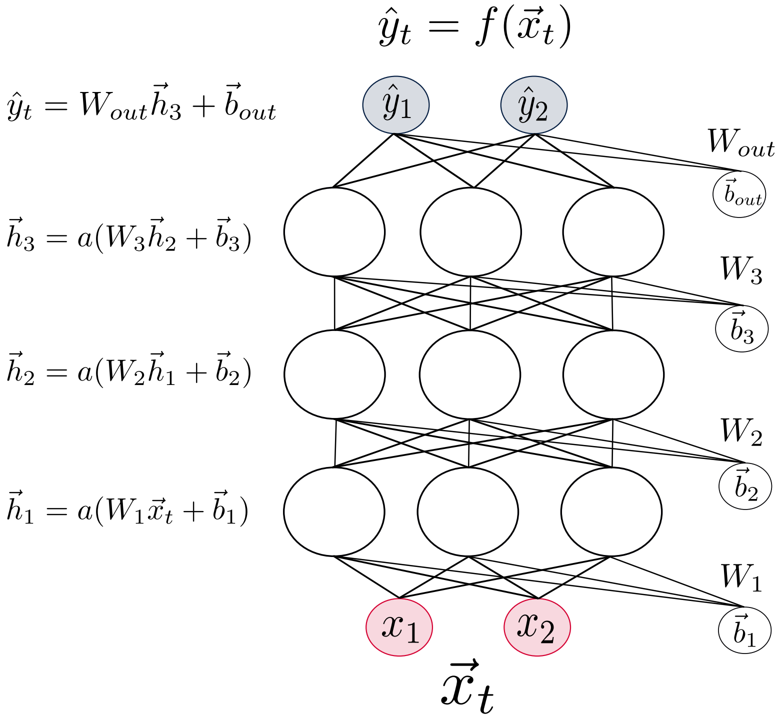
$\hat{y}_t = f(\vec{x}_t)$
$\hat{y}_t = W_{out}\vec{h}_3 + \vec{b}_{out}$
$\hat{y}_1$
$\hat{y}_2$
W_{out}
$\vec{b}_{out}$
$\vec{h}_3 = a(W_3\vec{h}_2 + \vec{b}_3)$
W_3
$\vec{b}_3$
$\vec{h}_2 = a(W_2\vec{h}_1 + \vec{b}_2)$
W_2
$\vec{b}_2$
$\vec{h}_1 = a(W_1\vec{x}_t + \vec{b}_1)$
W_1
$\vec{b}_1$
x_1
x_2
$\vec{x}_t$

Figure 2.5: Multilayer Perceptron (MLP) with three hidden layers. Weight matrices W are represented by the connections between nodes. The output layer uses a linear activation function, which is omitted in the illustration. The model parameters λ correspond to the weight matrices W and bias vectors b. Inputs to the network are denoted by $\vec{x}_t$ and outputs by $\hat{y}_t$

$[w_1, w_2, ..., w_N]^\top$ and a bias scalar b_k. For that node, we compute its pre-activation z_k and its activation h_k as:

$$z_k = \sum_{n=1}^{N} w_{kn}x_n + b_k = \boldsymbol{w}_k^\top \boldsymbol{x} + b_k \tag{2.24}$$

$$h_k = a(z_k) \tag{2.25}$$

where a denotes the activation function associated with the k-th node. In this thesis the activation function is constant throughout the network. In all cases, we use the hyperbolic tangent function $tanh$ defined as:

$$tanh(z) = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{2.26}$$

The hyperbolic tangent function has the property of *squashing* an infinite domain to the finite range $(-1, 1)$. It is a nonlinear function, which gives the model the ability to learn nonlinear mappings between inputs and outputs. Other alternatives for activation functions include the sigmoid, rectifier (with rectified linear units (ReLUs), Nair and Hinton (2010)), or linear activation functions, among others.

For all nodes in a hidden layer, we can organize connection weights and biases as a weight matrix and bias vector, respectively. The pre-activation vector $\boldsymbol{z}^l$ and the activation vector $\boldsymbol{h}^l$ of the l-th hidden layer then become:

$$\boldsymbol{z}^l = \boldsymbol{W}^l \boldsymbol{h}^{l-1} + \boldsymbol{b}^l \tag{2.27}$$

$$\boldsymbol{h}^l = a(\boldsymbol{z}) \tag{2.28}$$

where $\boldsymbol{W}^l$ is the connection weight matrix and $\boldsymbol{b}^l$ is the bias vector associated with the l-th hidden layer, and we let $\boldsymbol{h}^0 = \boldsymbol{x}$.

Note that, from Equation 2.27, the pre-activation for the hidden layer is an affine transformation of the hidden layer's input. The nonlinear activation function is then applied element-wise to the transformed input. Stacking multiple hidden layers allows the model to learn complex nonlinear transformations of the input vectors.

The output vector $\hat{\boldsymbol{y}}$ of a multilayer perceptron is given by an *output layer*, placed after the hidden layers. The implementation of this layer is dependent on the type of problem that is being modeled. For regression problems, the output layer is similar to the hidden layers, but a is defined to be a linear activation function (e.g. identity). This is the case for all systems trained in the context of this thesis. Other choices of output layers are also available. For example, a binary classification problem may use a logistic sigmoid activation function, while a classification problem with more than two classes uses a softmax activation function, giving a probability distribution over all possible classes.

Given the output layer, it is now possible to compute the output vector from the input vector. **Forward propagation** involves computing the activation vector for each layer sequentially. We let $\boldsymbol{out}$ be the network's output layer and we now let $\boldsymbol{h}^l$ be a function denoting the activation vector of the l-th hidden layer. The entire forward pass can then be implemented as a nested function of L hidden layers and 1 output layer.

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}, \lambda) = \boldsymbol{out}(\boldsymbol{h}^L(\cdots \boldsymbol{h}^2(\boldsymbol{h}^1(\boldsymbol{x})))) \tag{2.29}$$

The model parameters λ are the weight matrices and the bias vectors associated with each layer:

$$\lambda = (\boldsymbol{W}^1, \boldsymbol{b}^1, \boldsymbol{W}^2, \boldsymbol{b}^2, \cdots, \boldsymbol{W}^L, \boldsymbol{b}^L, \boldsymbol{W}^{out}, \boldsymbol{b}^{out}) \tag{2.30}$$

where $\boldsymbol{W}^{out}$ and $\boldsymbol{b}^{out}$ are the weight matrix and bias vector associated with the output layer.

Parameter initialization and objective function

In the previous section, we assumed we know the model parameters λ. This section describes how these are learned from a set of data points. We are given T input and output pairs, which we call the training data: $\{(\boldsymbol{x}^t, \boldsymbol{y}^t)\ : 1 \leq t \leq T\}$. We define initial values for all parameters λ. Weight matrices $\boldsymbol{W}$ are initialized by sampling independently from a Gaussian distribution with a mean of 0 and a standard deviation of $\frac{1}{\sqrt{n_{in}}}$, where n_{in} is the number of nodes connecting to the current layer. Bias vectors $\boldsymbol{b}$ are initialized to 0. This approach reflects the implementation of the Merlin Neural Network Toolkit (Wu et al., 2016), although other initialization recipes have been proposed (Glorot and Bengio, 2010).

For the training set, the overall error of the model is computed by taking the average squared error over all samples:

$$E = \frac{1}{2T} \sum_{t=1}^{T} \left\| f(\boldsymbol{x})^t - \boldsymbol{y}^t \right\|^2 \tag{2.31}$$

The problem can then be set as the selection of the optimal λ such that the sum of squares error is minimized. Under certain conditions, it can be shown that a least squares approximation is equivalent to maximization of the likelihood of the training data (Bishop, 2006, §3.1.1). This least squares approximation is suitable for regression tasks, such as the one used in this thesis. Other approaches can also be used depending on the task, such as the cross-entropy objective function commonly used for classification problems.

Backward propagation

Multilayer perceptrons are differentiable operators and they can be optimized with *gradient descent*. The gradient descent process is an iterative optimization algorithm that finds the local minimum of a function. It involves finding the gradient of the function, or the partial derivative of the function with respect to each of its parameters: $\nabla E = \frac{\partial E}{\partial \lambda}$. Updating each parameter in the direction of the negative of the gradient will result in a movement in the direction of a local minimum.

The process begins by finding the partial derivative of the error function with respect to the output layer parameters. Because we can conceptualize the forward pass as a nested function, we can apply the chain rule repeatedly to find the partials of the intermediate functions. This can be repeated for every parameter as we move backward through each layer in the network.

The partial derivative of the least squares error function with respect to the generated output vector can be computed as:

$$\frac{\partial E}{\partial \hat{\boldsymbol{y}}} = \hat{\boldsymbol{y}} - \boldsymbol{y} \tag{2.32}$$

Each layer contains an activation function. For a regression task, the output layer typically contains a linear activation function, whose derivative is 1. The derivative of the nonlinearity used in this thesis is:

$$\frac{\partial \tanh(\boldsymbol{x})}{\partial \boldsymbol{x}} = 1 - \tanh(\boldsymbol{x})^2 \tag{2.33}$$

Finally, the pre-activation for each layer is computed as an affine transformation, as given in Equation 2.27. The partial derivatives for each of those terms is:

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \boldsymbol{W}, \quad \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}} = \boldsymbol{x}, \quad \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}} = 1 \tag{2.34}$$

Let us say we wish to compute the partial derivative of the error function w.r.t the weight matrix $\boldsymbol{W}^{out}$ of the output layer:[3]

$$\frac{\partial E}{\partial \boldsymbol{W}_{out}} = \frac{\partial E}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_{out}} \frac{\partial \boldsymbol{z}_{out}}{\partial \boldsymbol{W}_{out}} \tag{2.35}$$

$$= ((\hat{\boldsymbol{y}} - \boldsymbol{y}) \ \circ a'_{out}(\boldsymbol{z}_{out}))\boldsymbol{h}_L^\top \tag{2.36}$$

$$= (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{h}_L^\top \tag{2.37}$$

where the operator $\circ$ denotes the Hadamard product, or element-wise multiplication. Equation 2.35 applies the chain rule to unroll each of function components of the objective function. And Equation 2.36 replaces each term based on the definitions given above. Because the activation function a in the output layer is the identity function, its derivative is dropped.

The same process can be repeated to the last hidden layer L of the network. We can let $\delta_{out} = \frac{\partial E}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_{out}} = (\hat{\boldsymbol{y}} - \boldsymbol{y})$, as these terms will reappear in the remaining layers of the network.

$$\frac{\partial E}{\partial \boldsymbol{W}_L} = \frac{\partial E}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_{out}} \frac{\partial \boldsymbol{z}_{out}}{\partial \boldsymbol{h}_L} \frac{\partial \boldsymbol{h}_L}{\partial \boldsymbol{a}_L} \frac{\partial \boldsymbol{a}_L}{\partial \boldsymbol{W}_L} \tag{2.38}$$

$$= [(\boldsymbol{W}_L^\top \delta_{out}) \circ a'_L(\boldsymbol{z}_L)]\boldsymbol{h}_{L-1}^\top \tag{2.39}$$

$$= \delta_L \boldsymbol{h}_{L-1}^\top \tag{2.40}$$

Defining δ_l for each layer l, we observe that there is a recursive pattern in the parameter update process. This pattern allows us to compute the partial derivatives of lower layers based on pre-computed terms of top layers. Because the update process begins at the output layer with the error function and moves backwards to the input layer, this method is referred to as **backpropagation** (Rumelhart et al., 1985; Williams and Zipser, 1995).

[3] Common derivations of the backpropagation algorithm are typically given in terms of individual parameters. We choose instead to follow a derivation in matrix form as it proceeds layer by layer and it reflects common software implementations. We follow the rather intuitive and simple explanation given in https://sudeepraja.github.io/Neural

Parameter update and stochastic gradient descent

Given the gradient of the error function, the model parameters λ can be updated as:

$$v = -\alpha \nabla E \tag{2.41}$$

$$\hat{\lambda} = \lambda + v \tag{2.42}$$

where α denotes a step size in the direction of steepest descent, commonly termed the **learning rate**.

In difficult problems, the weight space governing the objective functions tends not to be smooth. With gradient descent, it is easy to get suck in shallow local minima. This can be averted with the addition of a **momentum** term (Plaut et al., 1986), which essentially adds an acceleration term to the weight updates based on earlier iterations. The term v during the i-th weight update is defined as $v_i = mv_{i-1} - \alpha \nabla E$, where m is the momentum term and typically $0 \leq m \leq 1$.

During optimization, it is possible for the learned model to learn spurious patterns from the training data. This typically leads to poor generalization to unseen data and it is said that the model overfits the training data. **Regularization** methods add a penalty for complex solutions, thus forcing the training process to prefer simpler solutions. In this thesis, we use $L2$ regularization, which penalizes solutions employing larger weights. We define the regularization error term as

$$E_{reg} = \frac{1}{2} \sum_i w_i^2 \tag{2.43}$$

where w_i corresponds to each model parameter in λ. With the addition of this complexity penalty, the objective function can be redefined as:

$$E = E_{train} + \beta E_{reg} \tag{2.44}$$

where E_{train} corresponds to Equation 2.31 and β is a small positive constant weighting the regularization error term. The complexity term is differentiable and is included in the estimation of the gradient of the redefined objective.

A very similar approach uses $L1$ regularization, which, although very similar to $L2$, tends to prefer sparse weights rather than smaller weights. A different method of regularization has been proposed with dropout (Srivastava et al., 2014), which forces the model to ignore randomly selected hidden units during training.

So far the training process defined the objective function as computing an error measure over all training samples. Stochastic Gradient Descent (SGD) is similar to gradient descent, but the gradient is approximated on a single training sample. However, using a single example can lead to a high variance in the estimated gradients. To avoid such large variances, most implementations instead use a subset of training samples, termed a *mini-batch*. **Mini-batch Stochastic Gradient Descent** therefore estimates the errors over B samples in a mini-batch and updates the parameters accordingly. A full pass over the training set of T samples, or *epoch*, consists of $\frac{T}{B}$ parameter updates.

The optimization process continues iteratively over several epochs until a stopping condition is met. This condition may be a pre-defined maximum number of epochs or an error measurement on a validation set. If validation error is logged over epochs, loss of generalization may be identified when this error increases. Choosing to stop training based on such a condition is often referred to as *early stopping*.

Deep neural networks for speech synthesis

Recent approaches using neural networks for statistical parametric speech synthesis aim to overcome the limitations given by the conventional modeling of speech parameters with decision trees and Gaussian distributions. There has been a considerable amount of earlier work using neural networks for speech synthesis (e.g. Weijters and Thole (1993); Tuerk and Robinson (1993); Gerson et al. (1996); Sonntag et al. (1997); Chen et al. (1998)). However, recent improvements in software, hardware, and data availability have caused a resurgence of these methods.

In 2013, various studies appeared investigating neural networks for statistical parametric speech synthesis. The feedforward neural network described in this section was proposed as a replacement for decision tree clustering and Gaussian Mixture Models (Zen et al., 2013). Generative models such as restricted

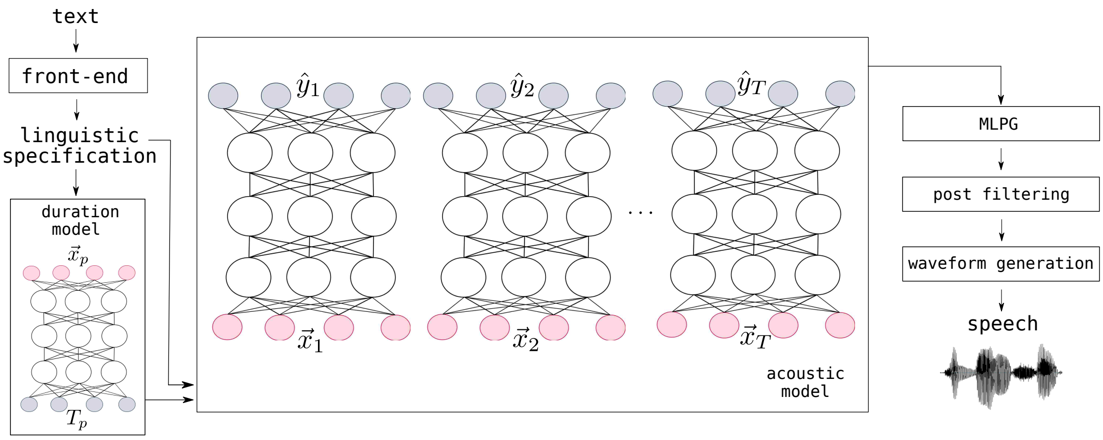

Figure 2.6: Speech synthesis with feedforward neural networks. Illustration inspired by Figure 1 in Zen and Senior (2014). The duration model generates phone-level durations T_p from phone-level labels $\vec{x}_p$. The acoustic model generates frame-level parameters for T frames in the utterance. Note that the neural network used for the acoustic model is applied repeatedly to each frame-level input $\vec{x}_t$.

Boltzmann machines (RBMs) or deep belief networks (DBN) have also been investigated. These studies have proposed such models as replacements for GMMs at the leaves of decision trees (Ling et al., 2013b,a) or as replacements for decision trees and GMMs entirely (Kang et al., 2013). In Fernandez et al. (2013), DBNs are used in conjunction with Gaussian Processes (GPs) to model fundamental frequency.

In Zen and Senior (2014), deep mixture density networks (MDNs, Bishop (1994)) were proposed as alternatives to the conventional multilayer perceptron. MDNs combine mixture models and artificial neural networks through the use of a mixture density output layer. Recurrent architectures have also been proposed, such as long short term memory (LSTM) networks (Fan et al., 2014). More recently, direct modeling of waveform samples was proposed with convolutional neural networks with the *Wavenet* approach (van den Oord et al., 2016).

In this thesis, we use a framework such as the one described in Zen et al. (2013) and Wu et al. (2015), illustrated in Figure 2.6, as this is the method implemented in the Merlin Neural Network Toolkit (Wu et al., 2016). During the data preparation stage, Hidden Markov Models with decision tree clustering and Gaussian distributions are inferred for the training data, as described in Section 2.2.2. The learned models are then used to force align the data at the state-level, from which frame alignment can be inferred. Given this alignment between linguistic features and acoustic parameters, a feedforward neural network called the *acoustic model* can be trained using mini-batch SGD as described in the previous sub-section. An additional neural network called the *duration model* may be trained in a similar fashion to model phone durations. Recently, rather than using a separate model for duration, Henter et al. (2016) included an extra output parameter in the acoustic model representing phone transition probability.

At synthesis time, for unseen data, the duration model estimates context-dependent phone durations in terms of the number frames. For each frame, the acoustic model then generates vocoder parameters, which are then smoothed with MLPG and post-filtered. Finally, a vocoder is used to synthesize the waveform.

This framework for DNN-based speech synthesis is different to the HMM-based framework discussed in the previous section, but one that has consistently led to improvements in the quality of synthetic speech (Zen et al., 2013; Wu et al., 2015; Hashimoto et al., 2015; Qian et al., 2014). In order to understand the differences between paradigms, Watts et al. (2016) proposed a continuum between HMM-based and DNN-based speech synthesis systems. It was observed that the differences leading to clearer improvements in synthetic speech were the change in regression model (decision trees to deep neural networks) and the change in modeling unit (HMM states to speech frames).

Throughout this thesis, we adopt feedforward neural networks for the acoustic model rather than recurrent neural networks. This choice was made as these models are quick to converge and still achieve good performance when compared with recurrent ones (Wu et al., 2016). With respect to model topology, this thesis uses an architecture similar to the benchmarked *DNN* system in Wu et al. (2016). This is a model with 6 hidden layers, each containing 1024 nodes with the hyperbolic tangent function. This topology has been adopted by a variety of other studies (Wu et al., 2015; Valentini-Botinhao et al., 2015; Hu et al., 2015; Watts et al., 2015; Merritt et al., 2016; Henter et al., 2016). The choice of hyper-parameters does not appear to contain any special properties, but it is presumed to be widely used as it is the default configuration of the Merlin Toolkit. Informal experimentation has observed minor fluctuations of acoustic parameters in terms of objective measures by exploring various model settings, but these were not expected to be large enough to affect subjective evaluations. Nonetheless, we acknowledge the importance of formally investigating alternative hyperparameter settings (number of layers, neurons, activation functions, etc), as well as more complex architectures (RNN, LSTM, B-LSTMs). We leave such investigations for future work (see Section 10.3).

We further adopt the implementation of the Merlin neural network toolkit to jointly model source and filter parameters. The assumption that source and filter are independent in vocoding and modeling leads to naturalness limitations (Henter et al., 2014). Furthermore, the parameters themselves are not entirely independent, as it was shown that *f0* can be predicted from MFCCs (Milner and Shao, 2007). Watts et al. (2016) investigated the effect of modeling spectrum and *f0* with separate networks and have found no improvements. For these reasons, in this thesis, we jointly model all acoustic parameters. The techniques that are explored here, although aimed at improving the generation of fundamental frequency, may have the ability to positively affect the remaining acoustic parameters. This is particularly relevant for the methods discussed in Chapter 9, which use *f0* and energy to learn vector representations for statistical parametric speech synthesis.

The framework implemented by the Merlin Neural Network toolkit (Wu et al., 2016) uses a separate model for duration and acoustics parameters, illustrated in Figure 2.6. In this framework, *f0* and duration are modeled and generated independently. Although recent work has investigated joint modeling of *f0* and duration (Ronanki et al., 2015; Henter et al., 2016), the standard approach still remains the separate modeling of duration and acoustic parameters. In this thesis, we adopt this standard approach, and we limit our scope to the acoustic model, focusing particularly on fundamental frequency. Some of the methods proposed could be extended to the duration model, such as those of Chapter 9. However, we leave this extension to duration modeling for future work, and we propose additional lines of research in this direction in Section 10.3.

2.3 Evaluation of speech synthesis

2.3.1 Objective evaluation

In statistical parametric speech synthesis, objective evaluations compare a sequence of acoustic parameters generated from a model with a reference sequence extracted from a waveform. Most objective metrics are distance measures between the two sequences. The underlying assumption is that the distance between the sequences is meaningful in terms of the quality of the model. That is, the smaller the distance between generated and reference parameters, the *better* the model.

However, it is not always the case that objective measures are representative of the quality of the acoustic parameters. Averaging over datasets might dilute otherwise perceptible acoustic differences between systems. In terms of intonation, for example, generated *f0* contours might still considered natural and yet different from the available references (see Section 3.1 for further details regarding this claim).

Objective measures can still be useful as they are fairly easy to compute and they facilitate comparisons over a large number of systems. In the main chapters of this thesis, the proposed hypotheses are initially discussed with respect to objective measures and a large number of systems. Those results are later validated with subjective evaluations on selected systems.

In this section, we give a brief overview of the main measures used in this work. When appropriate, these are computed according to the Merlin Neural Network Toolkit (Wu et al., 2016) and the equations presented here reflect that implementation. Note that objective metrics are sensitive to the vocoder used. In this thesis, we use STRAIGHT (Kawahara et al., 1999, 2001) and these measures are computed accordingly.

Mel cepstral distortion (MCD) measures the distance between two sequences of mel-cepstral coefficients. We are given a reference vector $\boldsymbol{x}$ and a generated vector $\hat{\boldsymbol{x}}$ of mel-cepstral coefficients. MCD is then computed as an extension of the standard Euclidean distance:

$$MCD = \frac{\alpha}{T} \sum_{t=1}^{T} \sqrt{\sum_{d=2}^{D} (x_d(t) - \hat{x}_d(t))^2} \tag{2.45}$$

$$\alpha = \frac{10\sqrt{2}}{ln 10} \tag{2.46}$$

where T is the total number of frames in the data set and D is the dimensionality of the mel-cepstral coefficients extracted at each frame. In this thesis, we use 60 coefficients per speech frame. Following Kominek et al. (2008), the constant α is included for historical reasons. Note that we exclude the first coefficient, commonly associated with the energy of a speech frame. This prevents the distance measure from being influenced by loudness, which may affect some datasets, such as non-professional audiobooks (Kominek et al., 2008).

Band aperiodicity distortion follows the same intuition (and notation) as MCD. For each frame, a D-dimensional vector of parameters is extracted to represent the source excitation signal. In this thesis, we extract 25 band aperiodicities and we compute the distortion between natural and predicted parameters as:

$$BAP = \frac{1}{10T} \sum_{t=1}^{T} \sqrt{\sum_{d=1}^{D} (x_d(t) - \hat{x}_d(t))^2} \tag{2.47}$$

In terms of objective measures related to the *f0* signal, we have used the root-mean-square error and Pearson's product-moment correlation. These are standard measures in the literature, although alternatives have been suggested (Clark and Dusterhoff, 1999). For the purpose of this thesis, these measures are computed at utterance-level on voiced-frames only and the average of all utterances in the test set is reported.[4]

For a given utterance u, **root-mean-square error** of the *f0* signal is deter-

[4] At the time of the writing of this thesis, the Merlin Neural Network Toolkit computes *f0*-related objective measures over all frames in the test set. The method used in this thesis reflects an earlier implementation that takes the average over utterances.

mined as

$$RMSE_u = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n))^2} \tag{2.48}$$

where N is the total number of reference voiced frames for utterance u. Given a total of U utterances in the dataset, the mean value is computed as

$$RMSE = \frac{1}{U}\sum_{u=1}^{U} RMSE_u \tag{2.49}$$

Similarly, the **correlation** of the *f0* signal is determined as

$$r_u = \frac{\sum_{n=1}^{N}(\boldsymbol{x}_u(n) - \bar{x}_u)(\hat{\boldsymbol{x}}_u(n) - \bar{\hat{x}}_u)}{\sqrt{\sum_{n=1}^{N}(\boldsymbol{x}_u(n) - \bar{x}_u)^2}\sqrt{\sum_{n=1}^{N}(\hat{\boldsymbol{x}}_u(n) - \bar{\hat{x}}_u)^2}} \tag{2.50}$$

$$CORR = \frac{1}{U}\sum_{u=1}^{U} r_u \tag{2.51}$$

where $\bar{x}_u$ and $\bar{\hat{x}}_u$ denote the mean value of the reference and the generated *f0* signal for utterance u, respectively. Note that *f0* correlation is here implemented as Pearson's product-moment correlation coefficient. Intuitively, this measure captures the similarity between the overall shape of generated and reference *f0* signals, which is particularly relevant for intonation. While distance-based objective measures aim to be minimized, the signal's correlation aims to be maximized.

Finally, in some chapters of this thesis, **voicing error** is reported as the percentage of frames that were assigned the incorrect voicing label.

2.3.2 Subjective evaluation

Objective evaluation methodologies are often used as an indication of the quality of synthetic speech, especially when a large number of systems is being developed, and reference acoustic parameters are available. However, it is widely agreed that subjective listening tests still remain the standard method for the evaluation of synthetic speech. The subjective evaluation of synthetic speech is not a trivial task and it still remains an active area of research. According to a recent overview of the Blizzard challenge (King, 2014), most system evaluations focus on naturalness and intelligibility. With developments in speaker adaptation and voice conversion techniques, speaker similarity has been adopted as a third dimension in the evaluation of speech synthesis systems.

Subjective evaluation methods are able to provide more accurate quality measurements than objective evaluation methods, but they also tend to be costly. Listening tests typically require a large investment in terms of time and resources, as they require well-designed experiments and the recruitment of human participants. Various factors should also be considered when designing and conducting listening tests for the evaluation of synthetic speech. For example, factors known to affect the results are the type of test, the question being asked, or the type and number of listeners. See Wester et al. (2015) for further details and a complete checklist of issues to consider when designing listening tests.

Although we acknowledge the wide variety of listening tests proposed, in this brief survey we focus on the methods used in this thesis. We provide some detail on well-established protocols for the evaluation of naturalness, with additional notes on methods used for the evaluation of intelligibility and comprehension.

Naturalness evaluation

Protocols for the evaluation of naturalness may be grouped into referenced methods, in which a synthetic sample is judged against an available natural reference, and non-referenced methods, in which synthetic samples do not have an available reference and are instead judged against the listener's expectations. The term naturalness refers to a property that is associated with naturally occurring speech, as produced by human speakers. It has been observed that this property heavily contributes to the overall quality of synthetic speech (Hinterleitner et al., 2011a).

MOS (Mean Opinion Score) is a non-referenced evaluation methodology from the field of speech coding (ITU-T Recommendation P.800, 1996). Because listeners are not given a speech reference to anchor their judgments, this method is also termed absolute category rating (ACR) (as seen in (ITU-T Recommendation P.800, 1996)). In a MOS evaluation, listeners are presented with one speech sample at a time. They are then asked to judge that sample on a 5-point scale in terms of quality, where 1 indicates *bad* and 5 indicates *excellent*. The question asked might instead focus on a different speech attribute, such as naturalness or pleasantness.

Variations of the MOS test are also possible, depending on the hypothesis being investigated. DMOS (Differential MOS) is a referenced version of the MOS test. Listeners provide their judgments for individual samples with respect to a reference sample. CMOS (Comparison MOS) presents the listeners with two randomized samples from different conditions. The task is to judge the second condition with respect to the first in terms of quality on a 6-point scale ranging from -3 (much worse) to 3 (much better). This method allows participants to give two answers with a single judgment – *which sample is better* and *by how much* (ITU-T Recommendation P.800, 1996).

MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor, ITU-R Recommendation BS. 1534-1 (2015)), like the MOS test, is an evaluation methodology inherited from the speech coding literature. In this approach, listeners are presented with all conditions at once and they are tasked with ranking them with respect to each other and an explicit reference. A copy of the explicit reference is hidden within the remaining experimental conditions, which fixes an upper bound for the listeners' judgments. Typically, in the MUSHRA test, an anchor such as low-pass filtered speech sets a lower bound. However, in speech synthesis, defining a lower bound is not a trivial task. The low-pass filtered speech used in speech coding, for example, is not applicable, since it remains a natural version of the speech utterance when compared to a generated version. When MUSHRA is applied in the context of speech synthesis, the anchor is normally dropped (Henter et al., 2014). Listeners are asked to rank each utterance on a 100-point scale, where 1 indicates a *poor* match to the reference and 100 indicates an *identical* match to the reference.

In the MUSHRA paradigm, listeners provide absolute scores measuring the similarity of synthetic samples with respect to a reference. But because all conditions are rated simultaneously, multiple comparisons across conditions are also provided. This implicitly creates a ranking of systems, which might be preferable over an absolute score. Ranking scores can be interpreted as a *preference* judgment, while absolute scores can be interpreted as a measurement of that preference.

There may be experimental scenarios where side-by-side comparisons are useful, but references are not appropriate. A MUSHRA evaluation with neither reference nor anchor may be considered a misuse of the term (ITU-R Recommendation BS. 1534-1, 2015), although the paradigm is still valid. With the lack of proper terminology, these methods have been called hybrids between MUSHRA and preference tests (Henter et al., 2016) or between MUSHRA and MOS tests (Dall et al., 2016a).

Perhaps the simplest evaluation methodology is the **AB test**, also called a **preference** test. In this paradigm, listeners are presented with randomized pairs of samples and are asked to express their preference with respect to some speech attribute. The most common question under this framework asks listeners to select the sample which *sounds more natural.* Depending on the experimental design, listeners may be given a third option indicating that they have *no preference.* If listeners are only given the option to choose between A and B, then it is common to refer to this method as a forced preference test. Variations of the preference test may include a reference, commonly termed an ABX test. In this case, the judgment expresses the listener's preference with respect to the reference.

A **similarity** evaluation aimed at understanding perceptual similarities across multiple systems may also be used (Mayo et al., 2005). Note that the term *similarity* is here used somewhat freely. Any referenced methodology is, in a sense, a measurement of the degree of perceptual similarity between an experimental and a reference condition. An alternative term for this method may be a *"same or different" task* (Merritt, 2016). In this evaluation paradigm, two randomized speech samples are presented to the listener. Unlike other methods, these are not two instances of the same utterance. Listeners are instructed to judge whether the two samples are similar or different in terms of naturalness. All responses are pooled to form a dissimilarity matrix, where each cell in the matrix corresponds to a condition pair. Each condition pair is represented by the number of times the two conditions were judged to be different. The dissimilarity matrix can then be embedded onto a lower-dimensional space with multidimensional scaling (MDS, Borg and Groenen (2005); Mayo et al. (2005)). The distance between conditions in the lower dimensional space is meaningful in terms of the the perceptual distance given by the test participants. This method can be useful for clear visualizations of the perceptual differences of multiple systems. This evaluation paradigm is used in Section 5.4 of this thesis.

Intelligibility and comprehension evaluation

Other well-established evaluation methodologies are also available, focusing instead on the intelligibility and comprehension of synthetic speech. A survey of the main findings until 2004 can be found in Winters and Pisoni (2004). We do not go into detail regarding these alternatives, as they are not used directly by this thesis, but we will mention some of them briefly. Evaluation methods focusing on intelligibility measure how comprehensible speech is under noisy conditions. Noise can be interpreted as originating from an external source (such as background noise from a train, car, etc.) or as artifacts from the synthesis process. Examples of evaluation methods may focus on individual words, such as the modified rhyme test (MRT, House et al. (1965)). Listeners may also be asked to transcribe utterances, from which word error rates can be computed. But because listeners can often infer words from contextual information, semantically unpredictable sentences (SUS) may be used (Benoît et al., 1996).

Alternative methods to evaluate the comprehensibility of synthetic speech use post-perceptual tasks. Participants are first asked to listen to synthetic speech samples and then to carry out simple tasks. These tasks may take the form of word-monitoring, multiple-choice questions, summarization or verification tasks (following the prior work review presented in Wester et al. (2016)).

Although the methods described so far are useful in the measurement of intelligibility and naturalness, they have drawn criticism because they do not evaluate a system under real-world conditions (Taylor, 2009, §17.2.2). It has been mentioned that listening tests commonly used in speech synthesis lack *ecologi-*

cal validity (King, 2014) – the property that the conditions in which a study is conducted approximate real-world conditions. Addressing this problem, a recent approach has allowed users to interact with an avatar in order to evaluate synthetic speech (Mendelson and Aylett, 2017). However, the authors did not find the results to be different to those from an audio-only evaluation paradigm.

On the methods adopted by this thesis

This thesis investigates the modeling of prosody for statistical parametric speech synthesis, focusing on intonation through the modeling of fundamental frequency. Although some methods have been proposed specifically for prosody and audiobook scenarios (Hinterleitner et al., 2011b), we mostly rely on traditional evaluation protocols for the naturalness of speech.

The decision to prefer these methods was made for several reasons. Firstly, the methods used in the evaluation of naturalness described in this section, although with known issues, are well-established throughout the speech community. These are effective in the measurement of clear differences between systems and suitable for the techniques proposed here. The techniques developed in the context of this work mostly focus on a general improvement of generated *f0* signals with respect to a given reference. Furthermore, within each chapter, hypotheses are clearly defined, and the choice of listening test is motivated by those hypotheses. Although this thesis does not directly investigate the evaluation of speech prosody for speech synthesis, we do acknowledge the need for such novel methods.

The work of Hinterleitner et al. (2011b) is appropriate for the evaluation of complete systems, as is the case of the Blizzard challenge (King, 2014). In this thesis, we evaluate very specific speech attributes (such as intonation) in carefully designed experimental conditions. Simpler methodologies, such as an AB or MUSHRA paradigm, are preferable in such scenarios. Throughout the thesis, when appropriate, further comments and discussion regarding evaluation methodologies will be given. For example, Section 5.6.3 of Chapter 5 provides further comments on referenced and non-referenced tests such as MOS and MUSHRA. And Chapter 10 conducts a final evaluation of the main contributions and provides concluding thoughts on the adopted evaluation protocols.

Chapter 3

Speech prosody and fundamental frequency

This chapter presents a brief overview of speech prosody, providing the theoretical foundations motivating the main claim of this thesis and the three sub-problems underlying it. A summary of earlier work in the context of statistical parametric speech synthesis is given with respect to those three sub-problems: representations of fundamental frequency, hierarchical modeling, and representations of linguistic contexts.

3.1 Speech prosody

3.1.1 Introduction

Information that is conveyed through the speech signal spans a variety of domains. These are of a linguistic, para-linguistic, and extra-linguistic nature (Obin, 2011). Information belonging to the *linguistic* domain expresses variation that is related to the underlying linguistic structure of an utterance. These may reflect lexical, syntactic, semantic, or discursive constraints. For example, lexical constraints relate, in some languages, to lexical or word stress. The syntax of the utterance, although not entirely, may affect how the speaker signals the grouping of units in the speech signal. Semantic constraints are related to focus or cohesion, while discursive constraints reflect the overall organization of the discourse.

The para-linguistic and extra-linguistic domains may be grouped under a more general non-linguistic domain. *Para-linguistic* information is related to the context of the communication in which the speech utterance is being produced. For example, listeners may make assumptions regarding the intent or the emotional state of the speaker. *Extra-linguistic* information introduces variation that is normally identified with the speaker as an individual, such as physiological (gender or age), idiolectal, or geographical characteristics.

These domains span multiple dimensions of variation within the speech signal. We may, for example, conceptualize a speech utterance as corresponding to an abstract sequence of phones over time. Each of the dimensions underlying the speech domains places constraints on the signal. Even though the identity of the phones in the sequence is not necessarily changed, the overall information contained in the speech signal may be manipulated in a variety of ways. For example, modal constraints (e.g. declarative, exclamative, interrogative sentences), emotion (e.g. happy, sad, neutral, ...), and age all have an effect on the speech corresponding to the same sequence of phones.

The study of **prosody** is concerned with the linguistic domain and with the description of aspects of the speech signal that are not directly explained by observing individual segments such as vowels or consonants in isolation from one another.[1] Such aspects stem from the mechanisms by which the speaker organizes the speech signal into a coherent structure of linguistic units (syllables, words, phrases, or utterances). Similarly, prosody is concerned with the aspects of the speech signal used to assign and signal the prominence of such linguistic units. Because these aspects are conveyed through acoustic or perceptual units at higher levels than the segment, prosody is widely agreed to be **suprasegmental** (Shattuck-Hufnagel and Turk, 1996; Nooteboom, 1997; Wennerstrom, 2001; Ladd, 2008; Obin, 2011; Xu, 2012; Turk and Shattuck-Hufnagel, 2014).

Traditionally, there are two perspectives on how to approach prosodic phenomena (Shattuck-Hufnagel and Turk, 1996; Nooteboom, 1997). A *phonetic* ap-

[1] The term *segment* may be somewhat ambiguous, as it may be applicable to any clearly identifiable unit in speech. We follow the traditional usage in phonetics and phonology, where segment refers to phones or phonemes.

proach is centered on the acoustic signal, focusing particularly on speech variation that cannot be directly related to individual segments. These observations made of speech data may then be related to information conveyed by the speech domains described above or to overall theories of prosodic structure. A *phonological* approach, on the other hand, focuses instead on the definition of higher-level abstract constituents and general patterns of prominence within them, which have the potential to be realized phonetically in the spoken utterance. Shattuck-Hufnagel and Turk (1996) propose a third approach that finds a compromise between the two views, which is related to the claim of Nooteboom (1997) that both approaches essentially have the same goal, but different starting points. While the phonetic view starts from the acoustic signal and generalizes to abstract representations, the phonological view supports its proposed abstractions with the acoustic signal.

3.1.2 Prosodic structure

When listening to spoken utterances, listeners may perceive that linguistic units are naturally grouped together, with levels of prominence within them. These groups, called prosodic constituents, define the prosodic structure of the speech utterance. Whether approaching prosody from a phonological or phonetic point of view, researchers aim to understand this prosodic structure and how it may be signaled acoustically.

Prosodic structure is composed of two parts. The first part, called the *prosodic constituent structure*, describes how prosodic constituents are signaled acoustically and how they relate to each other hierarchically. The second part, the *prosodic prominence structure*, is concerned with a description of the several degrees of prominence within prosodic constituents.

In terms of acoustic properties, there are three main correlates of prosodic structure: *fundamental frequency*, *duration*, and *intensity*. Fundamental frequency (or *f0*) refers to the vibrations of the vocal folds over time. This is a physical property that can be estimated directly from the acoustic signal. It is useful to separate it from its perceptual (or psychoacoustic) counterpart, often termed *pitch*. For example, speakers, limited by physical constraints, are bounded to an *f0* range, but listeners can perceive the same melody or pitch contour across speakers. These two distinct speech dimensions can be related to the production and perception of speech. These are a physical (or acoustic) dimension – that of *f0*, intensity, and duration – and a perceptual (or psychoacoustic) dimension – that of pitch, loudness, and quantity, respectively.[2]

The remainder of this section will provide a brief discussion of prosodic constituent and prosodic prominence structures as well their key acoustic correlates. We will focus primarily on fundamental frequency, as it is the acoustic property investigated throughout this thesis.

Prosodic constituent structure

It is believed that **prosodic constituent structure** is a linguistic universal (Turk and Shattuck-Hufnagel, 2014), with different languages selecting different constituents and levels from a universal hierarchy to signal different types of linguistic information. Units may be formed at multiple levels and incorporate constituents of differing length. For example, syllables or words may form constituents at lower levels and phrases may form constituents at higher levels. Regardless of the type of constituent, it is generally agreed that prosodic structure is hierarchical in nature (Turk and Shattuck-Hufnagel, 2014). That is, constituents are typically formed by constituents at lower levels.

There is some debate regarding the definition of constituents and how they interact hierarchically. Shattuck-Hufnagel and Turk (1996) provide an overview of four proposed hierarchies of prosodic constituency, with a selection illustrated in Figure 3.1. At lower levels, constituents such syllables, feet, or prosodic words may be identified. Although their exact definition may vary, these relate to acoustic phenomena such as lexical stress and pitch accents. At higher levels, phonological or intonational phrases may be identified. These are studied with respect to phrasal stress, boundary tones, or with respect to the overall syntactic

[2] We use here the terminology of (Ladd, 2008, §1.1), although we replace the terms *physical* and *psychophysical* with *acoustic* and *psychoacoustic*, respectively. Furthermore, we note that Hirst and Di Cristo (1998) use the term *length* as the psychoacoustic counterpart of *duration* and add *spectral tilt* and *timbre* as additional acoustic and psychoacoustic correlates of prosody.

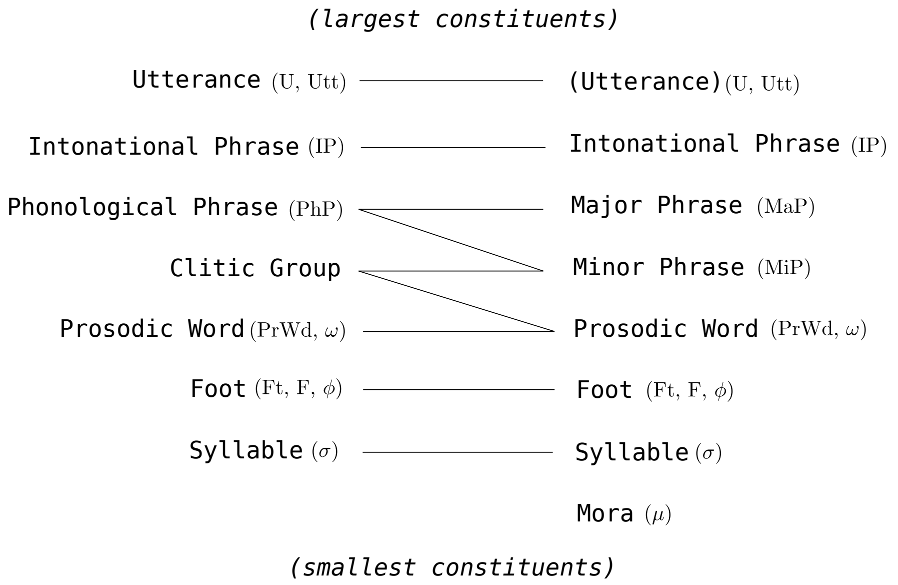

Figure 3.1: Two views of the prosodic hierarchy, inspired by Figure 2 of Shattuck-Hufnagel and Turk (1996). The left column illustrates the views of Nespor and Vogel (1986) and Hayes (1989) and the right column the views of Selkirk (1980) and Selkirk (1986). Horizontal lines indicate correspondence between the two views and parenthesis indicate common representations for a prosodic constituent.

structure of the sentence. Although syntax does impose some constraints on prosodic structure, it is generally believed that the two are not isomorphic (see Shattuck-Hufnagel and Turk (1996) for an overview of this claim).

Figure 3.1 defines the utterance as the highest level constituent. However, researchers have also investigated prosodic phenomena at a supra-sentential level. Such studies focus on finding evidence for supra-sentential prosodic constituents that can be related to theories of discourse structure (Grosz and Hirschberg, 1992; Hirschberg, 1993; Sluijter and Terken, 1993; Swerts and Geluykens, 1994; Nakatani et al., 1995; Wichmann, 2000; Wennerstrom, 2001; Smith, 2004; Tyler, 2013).

Prosodic prominence structure

Although no less important, **prosodic prominence structure** has received less attention than prosodic constituency (Shattuck-Hufnagel and Turk, 1996). As in the case of prosodic constituency, prominence is thought to be organized hierarchically (Turk and Shattuck-Hufnagel, 2014), with different levels of prominence identified within words or phrases.

One level of prominence occurs with syllables. For example, considering the sentences 'to *present* a case' or 'to buy a *present*', we may perceive that, in the words *present*, the location of the prominent syllable varies (e.g. *preSENT* and *PREsent*). In this case, the prominent syllable is said to carry **word stress** (also lexical stress). Syllables which are stressed may be called strong, while unstressed syllables may be called weak.

A different level of prominence occurs with phrases. For example, we may perceive two realizations of a phrase, such as 'to PRESENT a case' or 'to present a CASE'. The capitalized word is perceived to be more prominence than the remaining words. The word that is perceived to be more prominent is said to carry **phrasal stress** (also termed phrasal accent or sentence stress).

Note that, in the example above, even though the phrasal stress is shifted from 'present' to 'case', the word stress is still perceived. That is, one of the syllables in 'present' is still perceived to be more prominent than the other. This hierarchical prominence structure may be conceptualized using grid-like (Hayes, 1983) or tree-like (Liberman and Prince, 1977) representations.

Prominence encodes information regarding the structure and role of the utterance in the discourse (Wagner and Watson, 2010). The presence of prominence is influenced by factors related to discourse and information structure. For example, it is argued that prominence may be placed on units that are *new* (also called non-given or unpredictable in some studies) in the context in which the utterance is spoken (see Wagner and Watson (2010) for a brief review).

Acoustic correlates of prosodic structure

Studies investigating prosodic constituent structure are normally concerned with acoustic effects at constituent boundaries. These effects are expressed by a variety of acoustic correlates, such as duration, fundamental frequency, voice quality, or articulation degree (Turk and Shattuck-Hufnagel, 2014). It has been argued that durational cues are a reliable indicator of the presence and strength of constituent boundaries. This can be observed primarily through initial lengthening, final lengthening, and pauses. Initial lengthening occurs in the onset of the first syllable after a boundary, while final lengthening occurs in the rhyme of last syllable before a boundary (Wightman et al., 1992; Turk and Shattuck-Hufnagel, 2007). Related to lengthening, the presence of pauses is a strong indicator of constituent boundaries. Furthermore, it was also shown that the degree of lengthening (segment duration or pause) correlates well with the strength of the boundary (Wagner and Watson, 2010). Other acoustic correlates of prosody boundaries are voice quality (Dilley et al., 1996) – e.g. breathy, pressed, tense, creaky, etc.; segment articulation degree (Fougeron and Keating, 1997) – e.g. hypo or hyper articulation; and, to a lesser extent, intensity (Wagner and Watson, 2010).

Prominence structure is signaled through similar acoustic properties, with the addition of intensity and spectral tilt. It is agreed that the overall effects are different than those signaling prosodic constituency (Wagner and Watson, 2010; Turk and Shattuck-Hufnagel, 2014). For example, in terms of duration, final lengthening in constituent boundaries increases the duration of the nucleus and then the coda of the syllable, while prominence primarily lengthens the nucleus and then onset of the syllable (Turk and Shattuck-Hufnagel, 2014). The study of Kochanski et al. (2005) showed evidence that intensity and duration constitute good predictors of prominence in syllables, with intensity being the stronger of the two. In terms of spectral tilt, stressed syllables tend to show an even intensity distribution across the frequency spectrum, while unstressed syllables tend to have lower intensities for higher frequencies (Gussenhoven, 2004).

But perhaps more importantly in the context of this thesis, fundamental frequency is a key acoustic correlate of both constituent boundaries and prominence. A **pitch accent** is a local pitch event associated with a unit, such as a syllable, signaling some level of prominence relative to the surrounding units in the utterance (Ladd, 2008). These events may be represented in terms of two primitive pitch targets, denoted High (L) or Low (L). The acoustic realization of such events tends to be fairly local, although their presence and form may be determined by the context in which the utterance is produced (relating it to discourse and information structure).

Pitch events at the edges of constituents are termed **boundary tones**. In the English language, these are commonly associated with intonational phrases. These pitch excursions may be used to signal semantic, pragmatic, or discursive information (Wagner and Watson, 2010), such as indicating surprise, cuing turn-taking, or, for example, to differentiate declarative/interrogative sentences.

Additionally, it is commonly accepted that pitch events are scaled relative to each other (Wagner and Watson, 2010). Accents falling on individual words may be downstepped in order to signal overall constituent structure. This may be seen, for example, in structures such as *A but (B and C)* and *(A and B) but C* (Féry and Truckenbrodt, 2005). In this case, the second constituent may carry a lower pitch event relative to the first constituent to signal their dependency. Fundamental frequency may also be used to signal boundaries by resetting the overall pitch reference line, in what is often termed ***f0* reset**, which is a phenomenon related to boundary tones. Also, related to *f0* resets, it can be observed that *f0* tends to drift downward over long constituents, an effect often called ***f0* declination**.

Finally, it should be noted that the acoustic correlates of prosody are also affected by short-term variation associated with the segment, often called **micro-prosody**. Voiceless segments, for example, lack explicit *f0* values, and high vowels generally have higher *f0* than low vowels. Similarly, some segments - e.g. vowels, fricatives - are intrinsically longer than others - e.g. plosives, liquids (Nooteboom, 1997).

The acoustic correlates of prosody can therefore be observed at various temporal levels. For example, micro-prosodic variation at the segment level, lexical stress at the syllable level, boundary tones or phrase stress at the phrase level, or overall *f0* declination at the sentence level. Quoting other work, Wu et al. (2008) suggest that this phenomenon can be regarded as "small ripples on top of big waves".[3] This conceptualization of the acoustic signal as ripples on waves on swells on tides motivates some of the choices made throughout this thesis.

[3] The origin of this idea appears to have come from Bolinger (1964) (reprinted in Bolinger (1972)). The epigraph of this thesis was taken from Bolinger (1972), but formatted in a similar fashion to that of Ladd (2008).

3.1.3 Intonation

Intonation is concerned with the tonal patterns used to convey discourse meaning and to signal phrasal structure in speech communication (Gussenhoven, 2004). Although on occasion used interchangeably with the term prosody, the term intonation, in its narrower definition, refers to suprasegmental effects focusing on non-lexical characteristics (Ladd, 2008). This notion excludes factors such as lexical stress or word tone and focuses instead on phenomena such as pitch accents, boundary tones, or overall declinations. These phenomena are often called supra-lexical or post-lexical (Hirst and Di Cristo, 1998). According to Ladd (2008), this definition of intonation also excludes paralinguistic phenomena (such as the speaker's involvement in the speech communication).

Intonation is thought to have a phonological organization with tonal units describing pitch events such as High (H), Low (L), and their combinations (Jun, 2006). This view of intonation is the foundation of the Autosegmental-Metrical (AM) model of intonational phonology (Ladd, 2008). This model assumes a discrete sequence of phonological tones that can be aligned with syllables or placed at the edges of phrases. The tones aligned with syllables indicate relative prominence levels between the various syllables and describe **pitch accents** and **phrasal accents**. The tones placed at phrase boundaries are associated with prosodic constituents and describe **boundary tones**.

A common framework for the annotation of tonal events and constituent boundaries is the Tones and Break Indices (ToBI, Silverman et al. (1992)). This framework defines two tiers of annotation. The first tier, called the *tonal tier* is associated with tonal events describing accents and boundary tones, while the second tier, called the *break indices*, indicates the strength of each word boundary (Ladd, 2008). Proposed initially for the English language, various extension to ToBI have since been made in an effort to describe the intonational phonology of multiple languages (Jun, 2006).

Alternative approaches to ToBI have also been proposed. For example, INTSINT (INternational Transcription System for INTonation, Hirst and Di Cristo (1998)) is a coding system that describes relative pitch movements within a speech utterance. Intuitively, this can be compared to the narrow phonetic transcription of the utterance, rather than the broad phonological transcription of ToBI (Hirst and Di Cristo, 1998). However, Hirst et al. (2000) linked this transcription system to a distinct intermediate representation level, placed between a phonetic and a deep phonological level, which the authors call a *surface phonological representation* level. Note that the ToBI framework requests the human annotators to transcribe the intonational contour with respect to multiple acoustic cues by listening to the speech waveform. INTSINT, on the other hand, is a data-driven method relying only on fundamental frequency.

3.1.4 The lack of reference problem

One of the fundamental problems of prosody is based on the knowledge that it is an inherently phonological and phonetic phenomenon without a clear orthographic representation. This has been called the **lack of reference problem** (Xu, 2012), which is of particular importance for text-to-speech synthesis. Although some knowledge may be inferred from punctuation or other diacritics, this knowledge remains incomplete and ambiguous. For example, the segmental sequence may be easily inferred from the orthographic representation. This can be achieved, among other approaches, with lexica or G2P modules (see Section 2.1.1 for details). But devising a representation for the prosodic layer still remains a challenging problem. Annotation with protocols such as ToBI is often costly, and high inter-annotator agreement is not always achieved. The automatic detection of prosodic events is not a trivial task and often requires annotated databases for training, which are task- or domain-dependent (Rosenberg, 2009, 2010).

3.1.5 On the scope of this thesis

Given the complex interaction of acoustic properties, levels, and theoretical frameworks in speech prosody, a work such as this thesis is necessarily limited in scope. For this reason, throughout this thesis, we adopt a purely phonetic approach to the the modeling of speech prosody. This limits our attention to one of the its acoustic correlates: *fundamental frequency*, although we do acknowledge the importance of the remaining acoustic properties. In statistical speech synthesis, fundamental frequency is represented explicitly by the acoustic parameters extracted from the vocoder (see Section 2.2.1). This is in contrast, for example, to intensity, which is represented implicitly via the spectrum parameters. Additionally, duration and fundamental frequency are normally modeled separately with independent acoustic and duration models (see Section 2.2.3). For these reasons, it is convenient to limit our investigations to the acoustic model, focusing particularly on fundamental frequency. We leave extensions of the proposed techniques to the remaining acoustic correlates of prosody for future work.

Furthermore, we acknowledge the necessity of a good understanding of discourse for a more natural generation of prosody in speech synthesis. However, the techniques explored in this work are bounded by the sentence and make no attempt to move to discourse level.

The choices made throughout this thesis are motivated by the general belief that prosody is suprasegmental and that the acoustic signal is affected by information conveyed at multiple levels (e.g. phones, syllables, words, phrases). We do not use an annotation protocol for the representation of prosodic events, although automatically predicted ToBI labels are used as input features in the HMM-based systems presented in Chapters 4 and 5. Additionally, the notion of phrase is used throughout this work. Phrase boundaries are inferred by the front-end of a text-to-speech system and may correspond to the intonational or phonological phrases of Figure 3.1.

The scope of this thesis is derived from its main claim: *More natural synthesis of fundamental frequency can be achieved by exploring complex interactions of suprasegmental units in terms of linguistic representations, acoustic representations, and the mapping between them*, and its three sub-problems.

The conceptualization of acoustic effects being realized as *ripples on top of waves* motivates the work presented in Chapters 4, 5, and 6. In this work, parametric representations of the *f0* signal at multiple levels are investigated, for example, in an effort to separate micro-prosodic variation (say, ripples) from the overall sentence declination (say, the waves). The understanding of the acoustic signal as a collection of segmental and suprasegmental phenomena motivates the investigations presented in Chapters 7 and 8. These chapters investigate hierarchical architectures that decouple the two components (segmental and suprasegmental) and explore different integration methodologies. Finally the *lack of reference* problem motivates the work presented in Chapter 9, which proposes a method for the representation of suprasegmental units such as syllables and words.

3.2 Suprasegmental modeling of fundamental frequency

In this section, we provide a brief overview of recent approaches to the modeling of fundamental frequency in the context of text-to-speech synthesis. There is, of course, a long and well-established tradition of such methodologies. However, a complete review of these approaches is beyond the scope of this thesis, and work reviewed here is selected and limited to the recent approaches that have influenced our contributions.[4]

We prioritize recent work that focuses on representations of suprasegmental models of *f0* in the context of statistical parametric speech synthesis. This is divided according to the three sub-problems that guide this thesis, proposed in Section 1.1. In general, an attempt is made to divide previous work according to those sub-problems, although at times these are not clearly separable and may intersect. The overview presented here is brief and mostly focused on identifying general trends and observations that directly or indirectly influenced this thesis. Detailed reviews of relevant contributions are left to the introductory sections of subsequent chapters.

[4] Further reading may begin with Yu (2012), which provides a general overview of *f0* modeling and generation in the context of HMM-based speech synthesis. More general sources include (Taylor, 2009, Ch. 6) and Rao (2012), which describe general approaches to predict prosodic information from text. Additionally, Chapter 9 of Taylor (2009) gives an overall description on the synthesis of prosody.

3.2.1 Representations of fundamental frequency

The stylization of speech prosody is the process of decomposing a prosodic signal into several components: those that contain variation meaningful to the listener and those that account for residual variation (Obin, 2011). This is typically applied to fundamental frequency in an effort to find a compact representation of the signal. The prosodic signal is transformed in such a way that it is described at various temporal domains. These domains may reflect, for example, *f0* effects associated with multiple linguistic levels.

Various methods to stylize the *f0* signal have been proposed. The TILT intonation model (Taylor, 1998) was designed to identify intonational events (such as pitch accents and boundary tones) and represent them with interpretable parameters. MoMel (*Modeling Melody*, Hirst et al. (2000)) does not make assumptions regarding linguistic units, therefore it does not require a priori segmentation and identifies prosodic targets (inflections in *f0* signal) which may span temporal domains of variable length. The ProsoGram (Mertens, 2004) method is perceptually motivated and captures tonal events over syllable-sized segments. The recently proposed SLAM (*Stylization and LAbeling of speech Melody*, Obin et al. (2014)) is a data-driven method that identifies a set of discrete units representing shapes of an acoustic signal over various temporal domains. This method has been proposed for the analysis of speech prosody and it is not necessary invertible, although it has been applied to speech synthesis (Dall and Gonzalvo, 2016).

The Fujisaki model of intonation (Öhman, 1967; Fujisaki and Hirose, 1982; Fujisaki, 1983; Fujisaki et al., 1998) is a production-motivated approach to the decomposition of fundamental frequency. The signal identifies two components: the *phrase* or global component, assumed to be an impulse response, and the *accent* or local component, assumed to be a stepwise function. This approach inspired additional production-based methods, such as the atom-decomposition method (Honnet et al., 2015).

Parametric decomposition methods are a subset of a larger body of methodologies that have been proposed for the stylization of prosodic signals. With these methods, the signal is decomposed into a set of elementary contours, which may be associated with macro- and micro-prosodic variations (Obin, 2011). This decomposition is typically performed with a pre-defined set of time-varying functions: Legendre polynomials (Hsia et al., 2010), cubic spline functions (Wu et al., 2008; Qian et al., 2011), or cosine functions (Discrete Cosine Transform, Teutenberg et al. (2008)).

Previous work has compared several parametric stylization methodologies, either informally (Obin, 2011) or formally (Wu et al., 2008; Qian et al., 2011). In most of these analyses, the Discrete Cosine Transform (DCT) has been shown to be the most promising method. Common applications of the DCT in the context of speech synthesis are focused on suprasegmental modeling of speech prosody (Teutenberg et al., 2008; Latorre and Akamine, 2008; Wu et al., 2008; Qian et al., 2011; Obin et al., 2011; Stan and Giurgiu, 2011; Ronanki et al., 2016; Ijima et al., 2017).

More recently, the Continuous Wavelet Transform (CWT) has been used to represent the *f0* signal for speech synthesis (Vainio et al., 2013; Suni et al., 2013). The DCT and the CWT are of particular interest to this thesis, as they are used in the work presented in Chapters 4, 5, 6, and 9. The following sections provide further intuition and applications regarding these two methods.

3.2.1.1 The discrete cosine transform (DCT)

The Discrete Cosine Transform (DCT) was introduced in the context of *f0* modeling and synthesis by Teutenberg et al. (2008). The DCT stylizes a contour of N discrete samples with a weighted sum of zero phase cosine functions. The signal is represented by N DCT coefficients $C = [c_1, c_2, c_3, ...c_N]$. If x is a signal of length N, then:

$$c(k) = w(k) \sum_{n=1}^{N} x(n) cos(\frac{\pi(2n-1)(k-1)}{2N}) \tag{3.1}$$

where

$$w(k) = \begin{cases} \sqrt{\frac{1}{N}} \text{ if } k = 1 \\ \sqrt{\frac{2}{N}} \text{ if } 1 < k \leq N \end{cases} \tag{3.2}$$

The DCT is an invertible transform. The signal can be reconstructed with the Inverse Discrete Cosine Transform (IDCT).

$$x(n) = \sum_{k=1}^{N} w(k)c(k)cos(\frac{\pi(2n-1)(k-1)}{2N}) \tag{3.3}$$

With all coefficients, the IDCT is able to perfectly reconstruct the signal. However, one of the advantages of the transform is the decomposition of the initial signal into macro and micro variations (Obin, 2011). Most of the energy (macro variation) is stored in the initial coefficients, which often leads to an approximation of the signal with minimal loss by truncating the representation to the first M coefficients.

Most applications of the DCT to the *f0* signal for speech synthesis are concerned with finding a parametric representation over long temporal spans. This allows a fixed-sized compact representation of the signal, which then facilitates its integration with hierarchical models. Typically, these approaches have used the first 5–7 DCT coefficients (Teutenberg et al., 2008; Latorre and Akamine, 2008; Wu et al., 2008; Qian et al., 2011; Obin et al., 2011; Stan and Giurgiu, 2011; Ronanki et al., 2016; Ijima et al., 2017).

3.2.1.2 The continuous wavelet transform (CWT)

Wavelets have been used in a variety of applications in speech processing (Farouk, 2014). One of the main ideas behind wavelets is the analysis of a signal according to scale or resolution. A wavelet is a short waveform with finite duration averaging to zero. Larger windows are used to observe the long-term or coarser features of the signal, while smaller windows are used to observe the short-term variations.

To provide an intuition for the wavelet transform, it is common to compare it with the Fourier Transform. The latter uses the sine and cosine functions at various frequencies and measures their similarity with the input signal using a constant-sized window. It outputs a set of coefficients that represent the contribution of each frequency to the signal, essentially transforming the signal from the time domain to the frequency domain. Wavelet transforms are similar in principle. Instead of the sine and cosine functions, the Continuous Wavelet Transform

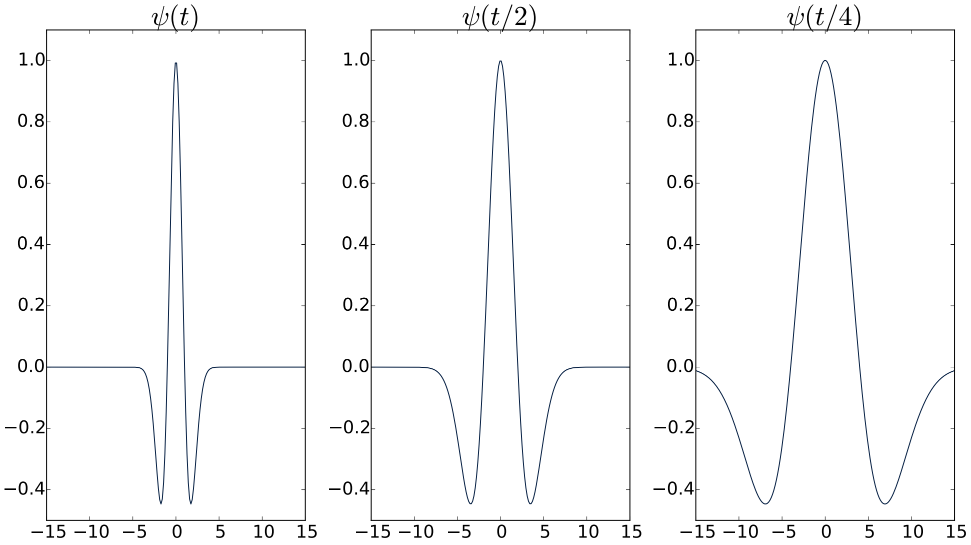

Figure 3.2: Mexican hat wavelet at various levels of resolution. Left figure illustrates $\psi(t)$, corresponding to scale 1. Middle figure shows $\psi(t/2)$, corresponding to scale 2. Right figure illustrates $\psi(t/4)$, corresponding to scale 4.

(CWT) uses transformations of one analyzing function, commonly referred to as the *mother wavelet* (Daubechies et al., 1992; Fugal, 2009; MATLAB, 2014).

The mother wavelet can be chosen or developed according to the analysis type or the input signal, and various families of wavelets have been used, such as the Haar, Daubechies, or Coiflet families (Graps, 1995). In speech synthesis, recent work has used the Mexican hat wavelet as the mother wavelet for *f0* processing (Vainio et al., 2013; Suni et al., 2013; Sanchez et al., 2014; Suni et al., 2017). The Mexican hat wavelet is given by:

$$\psi(t) = \frac{2}{\sqrt{3\pi^{1/4}}}(1-t^2)e^{\frac{-t^2}{2}} \tag{3.4}$$

Two transformations of the mother wavelet are allowed: *scaling* and *translation*. The first transformation, *scaling* (or dilation) is the stretching or compression of the wavelet in time. An example is shown in Figure 3.2, where the mother Wavelet (left) is scaled by a factor of 2 (middle), and by a factor of 4 (right). *Translation* is the shifting of the wavelet in time.

For a given scale, the CWT measures the similarity between the signal and the scaled wavelet. It then shifts the wavelet in time by one sample and repeats the process. The output is an MxN matrix, where M is the number of scales and N is the length of the signal. The CWT coefficient at scale a and position b is given by:

$$C(a,b;f(t);\psi) = a^{-1/2}\int_{-\infty}^{\infty} f(t)\psi(\frac{t-b}{a})dt \tag{3.5}$$

where $f(t)$ is the input signal at time t and ψ is the analyzing wavelet. It is easy to see that the three parameters that affect the coefficients are the scale, the position, and the analyzing wavelet. By varying a and b, we obtain all CWT coefficients.

Note that even though this transform is referred to as continuous, it still operates on a discrete signal. The name merely distinguishes it from other wavelet transforms, such as the Discrete Wavelet Transform (DWT). One main difference is how the CWT is allowed to transform the mother wavelet. When scaling it, the CWT allows any input as long as $a \geq 1$, and, when translating it, the CWT typically shifts the wavelet smoothly one data point at a time.

As the scale gets smaller, the more compressed the wavelet is, thus being more susceptible to higher frequencies and short-term variation of the input signal. Similarly, as the scale increases, the more stretched the wavelet is, thus making it more susceptible to lower frequencies and long-term variations of the signal. Therefore, there is an inverse relationship between scale and frequency, which is reflected in the output of the transform. Typically, higher frequencies are captured by the smaller scales, and lower frequencies are captured by the larger scales. This provides some intuition as to how the wavelet transform can be seen as a filtering technique.

Scaling the wavelet is essentially varying the size of the analysis window. As mentioned above, larger windows (scales) capture long-term coarse features, while smaller windows (scales) capture short-term finer features. This shows how the transform can be seen as a windowed transform. Figure 3.2 exemplifies the usage of the mother wavelet at different scales. On the left, the mother Wavelet is not scaled, and corresponds to scale 1, while the other two sub-figures illustrate the wavelet scaled by some factor. The larger the scaling factor, the more stretched is the wavelet, thus enlarging the window size and capturing coarser features of the input signal.

If the Continuous Wavelet Transform is to be used as a stylization method for analysis, modeling, and synthesis, it must be invertible. A reconstruction formula is suggested by Suni et al. (2013). Assuming the signal has been decomposed into ten scales $i = [1, ..., 10]$, with each pair of neighboring scales one octave apart, then each scale can be approximately recovered by:

$$C_i'(x) = C_i(x)(i + 2.5)^{-5/2} \tag{3.6}$$

The original signal can then be approximately reconstructed by summing over all scales:

$$\hat{f}_0(x) = \sum_{i=1}^{10} C_i'(x) \tag{3.7}$$

Figure 3.3 illustrates the 10 scale decomposition strategy proposed by Suni et al. (2013). This leads to another interpretation of the CWT when applied to the *f0* signal. The stylization method separates the signal into its lower frequencies (or macro-prosodic variation) and its higher frequencies (or micro-prosodic variation).

In speech synthesis, the use of the CWT is not new, but there has been renewed interest. Most work with the CWT used the Mexican hat wavelet to analyze the *f0* signal (Kruschke and Lenz, 2003; Vainio et al., 2013; Suni et al., 2013; Sanchez et al., 2014). Earlier work used the CWT to estimate phrase and accent commands under a generalized superpositional model (Kruschke and Lenz, 2003) or as a stylization method (Wang and Narayanan, 2005). More recently, the CWT was proposed as a method for the analysis of prosody (Vainio et al., 2013; Suni et al., 2017), as a representation of *f0* in HMM-based speech synthesis (Suni et al., 2013), for annotation and control of prominence in speech synthesis (Vainio, 2014; Vainio et al., 2015), and as a parameterization method for *f0* in voice conversion (Sanchez et al., 2014).

This thesis contributes to signal representation approaches with work presented in Chapters 4, 5, and 6. These chapters inherit the findings of methods using the discrete cosine transform and the continuous wavelet transform for the

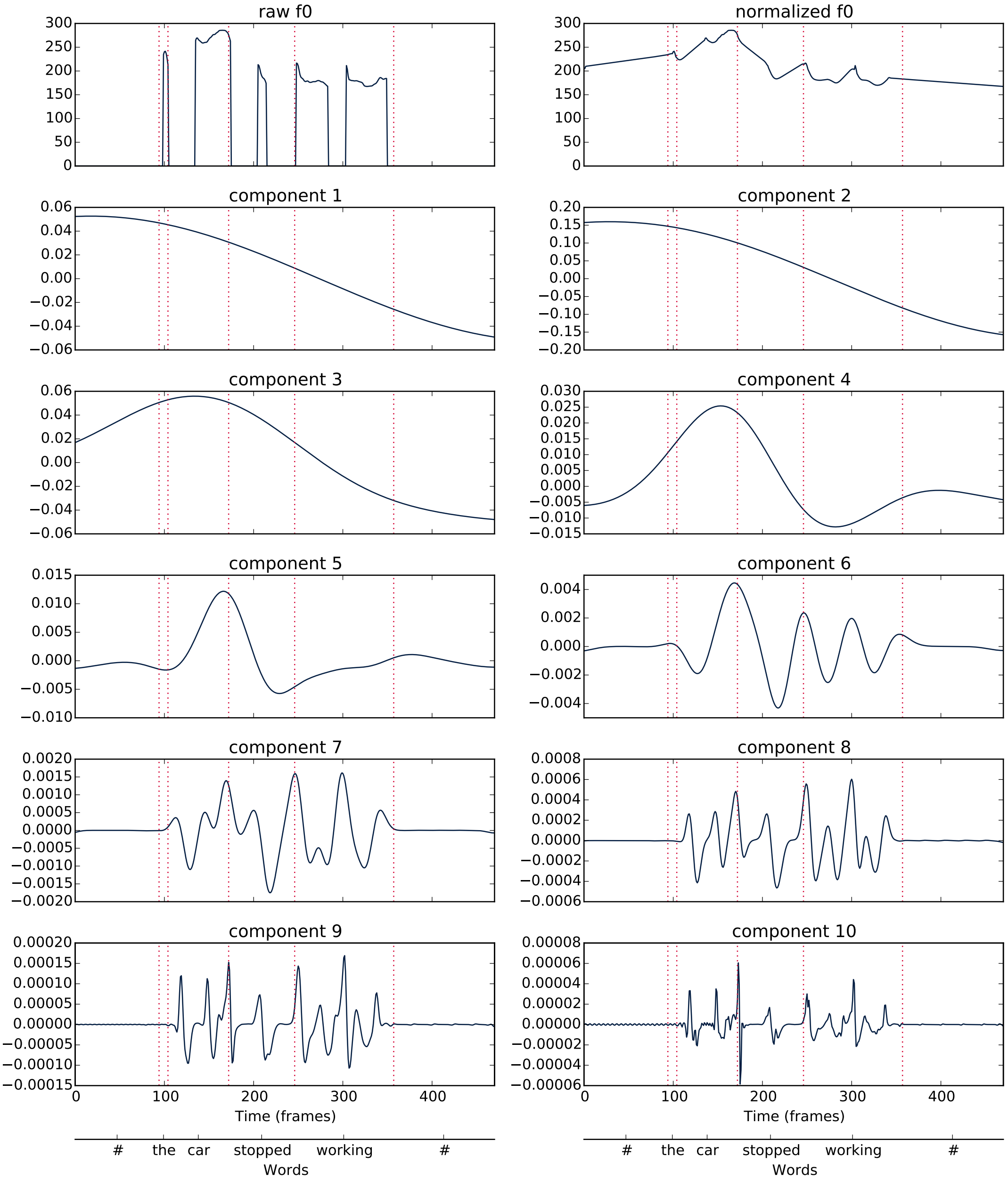

Figure 3.3: Continuous Wavelet Transform (CWT) applied to the *f0* signal using the 10-scale decomposition strategy proposed by Suni et al. (2013). Top sub-figures illustrate the raw and linearly interpolated *f0* signal in Hz. Remaining sub-figures illustrate the 10 CWT components sorted by increasing frequency of the mother Wavelet. Red vertical lines indicate word boundaries.

representation of *f0* in statistical parametric speech synthesis.

3.2.2 Hierarchical models

Because prosody is agreed to be hierarchical and suprasegmental, methods focusing on modeling prosodic variations at various temporal domains have been proposed. These approaches attempt to move away from the short-term models that are often used in statistical parametric speech synthesis. Although terminology varies, there is often a distinction between two types of methods: *joint* (or *asynchronous*) models and *superpositional* (or *synchronous*) models.

Joint models jointly describe prosodic variations over several temporal domains. The signal is modeled separately at different levels with no interaction between them. Typically, the syllable or phrase levels are added to the traditional phone level. For parameter generation, the lower level parameters are maximized under the constraint of the higher level models. However, these approaches are weakened by using the same signal across linguistic levels. For example, models at the lower levels (such as phones) are still influenced by acoustic effects associated with long temporal domains (such as overall sentence declination). Similarly, models at higher levels (such as syllables or phrases) still have to account for micro-prosodic variation.

For statistical parametric speech synthesis, earlier joint models were primarily based on a stylization of the *f0* signal over long temporal domains using the Discrete Cosine Transform (Teutenberg et al., 2008; Latorre and Akamine, 2008; Wu et al., 2008; Qian et al., 2011; Obin et al., 2011; Latorre et al., 2013). These have been proposed in the context of HMM-based speech synthesis. Long temporal units are typically defined over syllables, although some work has focused on longer units, such as phrases (Teutenberg et al., 2008; Obin et al., 2011).

For DNN-based speech synthesis, early work used a recurrent neural network with recurrences defined at syllable and word levels (Chen et al., 1998). Recently, joint hierarchical models in various forms have been proposed. In the approach of Ronanki et al. (2016) the *f0* signal at syllable-level is represented with the DCT and clustered to find a small set of *f0* templates. These templates are predicted with LSTMs and used as input to a frame-level acoustic model. A different method focuses instead on the input features and defines a top-down hierarchical encoder-decoder at word, syllable, and phone levels (Ronanki et al., 2017). The work of Wang et al. (2017a) proposes a model for *f0* using feedback links defined over multiple linguistic units. These links transfer aggregated quantized *f0* features over phones and syllables.

Superpositional models consider *f0* to have an additive structure with its components influenced by separate factors. In this class of models, signal effects are estimated separately at various temporal domains, and then superimposed, in an additive or in a multiplicative fashion. Note that we use the term "superpositional" quite loosely here. Traditionally, superpositional models refer to variants of the Generalized Superpositional Model used in the context of concatenative speech synthesis such as Fujisaki (2008), Van Santen and Möbius (2000), or Bailly and Holm (2005).[5] This type of models are also called *overlay* in Ladd (2008) or *superimpositional* in Taylor (2009). In statistical parametric speech synthesis, other types of additive models of *f0* have been proposed, and therefore we adopt the terminology of Obin (2011), which uses "superpositional" to describe this type of modeling approach.

Various additive models have been proposed in the context of HMM-based speech synthesis (Teutenberg et al., 2008; Zen and Braunschweiler, 2009; Hsia et al., 2010; Lei et al., 2010; Yin et al., 2014). For DNN-based speech synthesis, the work of Yin et al. (2016) investigated deep neural networks structured hierarchically using a cascaded (or joint) and parallel (or superpositional) approach.

Although duration is beyond the scope of this thesis, it is worth noting some work investigating hierarchical models of duration (Gao et al., 2008; Obin et al., 2009; Qian et al., 2011). These models have shown a decrease in prediction error, although subjective evaluations were disappointing. Only one approach has shown perceptual differences, and only when changing speaking rate (Zen et al., 2012). Remaining work has either failed to show significant improvements (Qian

[5] We omit from this overview the large body of work exploring superpositional models for concatenative speech synthesis, although we do acknowledge their influence on more recent approaches.

et al., 2011) or did not conduct perceptual experiments (Obin et al., 2009). However, the results of Qian et al. (2011) suggest that simultaneous modeling of *f0* and duration over various temporal domains might be able to have a stronger impact on the naturalness of synthetic speech than modeling either of them separately in those temporal domains.

In the modeling of *f0*, studies showing a comparison of short temporal domains (e.g. phones, syllables) and long domains (e.g. words, phrases, utterances) seem to agree that there is little to gain by considering long domains (Wu et al., 2008; Zen and Braunschweiler, 2009; Obin et al., 2011; Stan and Giurgiu, 2011; Qian et al., 2011). These studies claim that after including large temporal domains, no relevant improvements were observed either over the frame-level baseline or over the shorter domains. However, some work using long domains has claimed significant improvements (Hsia et al., 2010), but they do not appear to report any explicit analysis of temporal domains.

It is possible that the minimal influence of long domains in these studies is a consequence of not having the proper text-derived features at those levels. The vast majority of the work summarized here is limited to the features conventionally used in statistical parametric speech synthesis, which are mostly related to the length and position of syllables, words, or phrases. Hierarchical modeling is covered in this thesis by Chapters 7 and 8, where we investigate a cascaded and parallel deep neural network architectures with various types of linguistic features.

3.2.3 Representations of linguistic contexts

Prosody is an inherently phonological and phonetic process that operates at a suprasegmental level. However, it suffers from what has been termed the *lack of reference problem* (Xu (2012), see also a brief discussion on p. 55). This is of particular importance for text-to-speech, as acoustic parameters are generated from textual information. In most systems, prosodic phenomena are inferred from modules learned from small annotated datasets or from a set of shallow suprasegmental features. Examples include lexical stress, pitch accents, intonational phrase breaks, ToBI labels, or part-of-speech tags, which are predicted from individual modules in the front-end. The remaining features mostly describe the position of syllables, words, or phrases in the utterance. See Appendix A for a full description of linguistic features.

In data-driven systems, this process may lead to noise in the representation of linguistic contexts. When dealing with large databases, such as audiobooks, manual labeling of higher-level features tends to be costly. This is because the automatic annotation of prosodic information is not a trivial problem and is prone to error (Rosenberg, 2009, 2010). For this reason, researchers often make use of automatic labeling for the annotation of the training data: for example, by using pitch accent or intonational phrase boundary predictors trained on small, potentially out-of-domain, datasets. Such methods inevitably generate prediction errors and cause mismatches between acoustic events in the training data and their annotated labels, which is harmful to the acoustic model used for synthesis.

Earlier work has shown that using manually-annotated labels is largely preferred to using automatically-annotated labels at training and test time (Watts et al., 2010). This is also observed when dealing with the more realistic scenario of using manual labels at training time and automatic labels at test time.

The impact of high-level features was also investigated by Watts et al. (2010), but limited to the system using manual labels. It was observed that good annotation of pitch accents, boundary tones, and location of intonational phrase boundaries all contribute to the naturalness of synthetic speech. However, Watts et al. (2010) did not explicitly evaluate the impact of the features on the automatically-labeled system, although it was hypothesized that their importance would not be as strong.

A separate study has investigated the relative impact of features at various linguistic levels on the naturalness of HMM-based synthetic speech (Cernak et al., 2013). The authors repeatedly added contextual features (phone, syllable, word, phrase, and utterance) to their system. It was observed in a listening test that most of the naturalness associated with synthetic speech is given by features at the lower levels, mostly related to phones and syllables. It was mentioned in the study that manually-annotated training data was used, which contradicts some of the observations reported by Watts et al. (2010), although details regarding the origin and type of annotation are omitted.

These studies suggest that contextual factors describing prosodic phenomena can be beneficial for speech synthesis, provided they match the training data and achieve good accuracy at synthesis time. This, however, may still pose a problem when using very large databases. especially *found data* such as audiobooks. The studies of Cernak et al. (2013) and Watts et al. (2010) use experimental datasets with approximately 1 hour of speech data for training.

One approach is to use *signal-driven labelling* to annotate the training data, while using predicted labels for synthesis. Two recent studies have adopted this approach, either using automatically annotated ToBI labels (Tesser et al., 2013) or SLAM, a data-driven stylization method (Dall and Gonzalvo, 2016).

Alternatively, researchers have sought to learn data-driven representations of higher-level linguistic units. In terms of discrete representations, some studies focused on specific prosodic phenomena such as prominence or emphasis (Badino et al., 2009, 2012), or on additional text representations through parsing (Obin et al., 2010; Dall et al., 2016b).

With an increased interest in deep neural networks, recent approaches began to explore high-dimensional continuous-valued representations of contexts. These methods have been used as input to intermediate front-end modules, such as phrase-break predictors (Watts et al., 2011; Vadapalli and Prahallad, 2014; Watts et al., 2014), or as direct input to text-to-speech systems (Lu et al., 2013; Wang et al., 2015a, 2016b,a; Ijima et al., 2017). Because continuous-valued representations are typically data-driven, some work has investigated them as replacement of expensive knowledge-based features, typically POS tags (Watts et al., 2011, 2014; Wang et al., 2015a).

It is worth noting the work of Wang et al. (2016b), which investigated the impact of text-based continuous representations at various linguistic levels on the naturalness of synthetic speech. The authors found that phone and syllable-level representations added little to the acoustic model, while phrase-level representations gave the most promising results. This is an interesting observation, as most hierarchical models described earlier found that longer-temporal levels have less impact on the quality of synthetic speech. At higher levels, sentence-level control vectors learned jointly with the acoustic model have also been proposed (Watts et al., 2015).

This thesis contributes to this body of work with the methods proposed in Chapter 9. A data-driven method to learn vector representations of words and syllables based on acoustic counts is proposed, and these are used directly as input to a DNN-based acoustic model for text-to-speech synthesis. This method uses acoustic evidence from the training data for the acoustic model, thus also being signal-driven, and reducing the mismatch between context labels and training data.

Chapter 4

A multi-level representation of fundamental frequency

This chapter is an extended version of the work described in "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform" (Ribeiro and Clark, 2015) presented at ICASSP 2015.

We propose a representation of f0 *using the Continuous Wavelet Transform (CWT) and the Discrete Cosine Transform (DCT). The CWT decomposes the signal into various scales of selected frequencies, while the DCT compactly represents complex contours as a weighted sum of cosine functions. The proposed approach has the advantage of combining signal decomposition and higher-level representations, thus modeling low-frequencies at higher linguistic levels and high-frequencies at lower linguistic levels.*

4.1 Introduction

Standard techniques in statistical parametric speech synthesis are still focused on short-term approaches, such as Multi-Space Distribution HMMs (MSD-HMM, Tokuda et al. (2002)) or Continuous F0 HMMs (CF-HMM, Yu and Young (2011)). These approaches typically focus on short-term variations and capture suprasegmental effects somewhat implicitly through context dependent models.

In order to leverage the suprasegmental characteristics of prosody, some work has started to explore multiple temporal domains in the modeling of *f0* (Teutenberg et al., 2008; Latorre and Akamine, 2008; Qian et al., 2011; Obin et al., 2011). These approaches typically model *f0* over larger units, such as syllables or phrases, adding them to the traditional phone-level models. To represent *f0* at these higher levels, the Discrete Cosine Transform (DCT) is used, which is able to compactly represent complex contours.

Common findings within these approaches show that, although multi-level models improve synthesized speech, higher levels contribute little to the naturalness of synthetic speech. However, in most of these approaches there is no attempt to separate the long-term from short-term effects of *f0*.

Recently, the Continuous Wavelet Transform (CWT) has been proposed for the analysis and modeling of *f0* within an HMM-framework (Suni et al., 2013). In the work of Suni et al. (2013), some improvements were seen in the accuracy of *f0* modeling, but these effects were still being modeled only locally at frame-level. Conversely to this CWT model, in the class of models using the DCT, the same signal is modeled both on lower and higher levels. That is, long-term intervals still have to deal with short-term effects and short-term models have to deal with long-term effects.

In this work, we propose to explore a multi-level representation of *f0* by combining both transforms. This allows us to represent *f0* by first decomposing it into several scales and then model each at their hypothesized respective levels. That is, short-term effects are modeled with short-term units and long-term effects are modeled with long-term units.

4.2 Multi-level representation

4.2.1 Signal decomposition into multiple linguistic levels

This section details the steps applied to the *f0* signal in order to achieve the proposed representation. The following illustrates the chain of processes from the *f0* signal to the multi-level representation, with each of the following subsections providing further details:

$$f_0 \rightarrow normalization \rightarrow continuous\ wavelet\ transform \rightarrow discrete\ cosine\ transform$$

Signal normalization

Normalization of *f0* signal follows the steps of Vainio et al. (2013) and Suni et al. (2013). The signal is first transformed to the logarithmic scale. To reduce artifacts from the *f0* tracker, values below two standard deviations of the mean of *log-f0* are removed. Unvoiced regions are then linearly interpolated and re-introduced into the original *log-f0* signal. Finally, the interpolated *log-f0* contour is reduced to zero mean and unit variance, as this is required by the wavelet transform.

Continuous wavelet transform

To decompose *f0* into multiple linguistic levels, we use a wavelet based decomposition approach identical to that described in Suni et al. (2013), with 10 wavelet scales, each one octave apart. To reduce the number of scales, adjacent scales were combined, which resulted in a 5 scale representation of the signal, each approximately 2 octaves apart. This is similar to the steps described in Section 3.2.1 of this thesis.

The use of these particular scales is motivated by an attempt to relate scales to levels of linguistic structure, and each one of these scales is labeled with an approximate representation in a linguistically-motivated hierarchical structure. We assume that high frequencies (lower scales) capture short-term variations associated with the phone and that low frequencies (higher scales) capture long-term variations associated with the utterance.

Therefore, we associate components 1-2 with the utterance level, components 3-4 with the phrase level, components 5-6 with the word level, components 7-8 with the syllable level, and components 9-10 with the phone level. Figure 3.3 (p. 64) illustrates these components. The association of each component to a particular linguistic level is approximate and inspired by the work of Suni et al. (2013). The use of these linguistic levels was a practical decision in Suni et al. (2013), but they do span the linguistic structure that is conventionally used in TTS systems (phones, syllables, words, phrases, utterances).

Level	Coefficients
Utterance	3
Phrase	4
Word	4
Syllable	6
Phone	6

Table 4.1: Number of selected DCT coefficients at each linguistic level.

Note, however, that these associations remained assumptions at the time this work was conducted. There is no evidence to support the idea that, in this decomposition, for example, the wavelet components associated with the syllable correspond in fact to *f0* effects at syllable-level. Further work focusing on assessing the linguistic meaningfulness of these components will be discussed in Chapters 5 and 6. The steps performed so far on *f0* are similar to those described in Suni et al. (2013).

Discrete cosine transform

To model each linguistic level, the signal extracted with the CWT is first segmented appropriately at each level, e.g. the syllable component is segmented at syllable boundaries – bootstrapped by forced alignment at the state level. The DCT is then applied to parameterize each segment individually.

The usefulness of the DCT derives from the fact that we can truncate its representation to a few coefficients without any relevant signal loss. This allows the signal associated with each linguistic level to be modeled at its respective level. Previous work typically uses the first 3–7 coefficients, depending on the modeling approach, as seen in Section 3.2.1. A quick evaluation was performed on a held-out set in order to determine an appropriate number of DCT coefficients at each of the linguistic levels to minimize signal loss at the same time as representing the signal compactly. Correlation and RMSE were used to measure the signal before and after reconstruction.

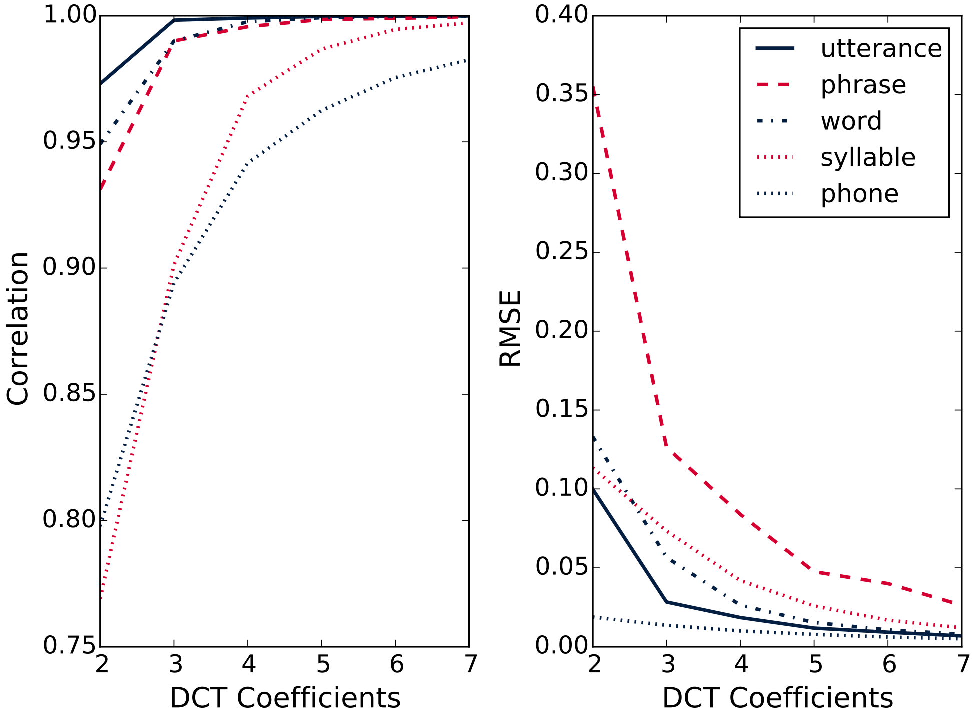

Figure 4.1: Reconstruction error in terms of Correlation and RMSE with varying DCT coefficients for each wavelet component.

Each component from the CWT was represented with the DCT at its corresponding level with varying number of coefficients (2 to 7 coefficients). Figure 4.1 shows the results for each of the linguistic levels. Not surprisingly, the lower linguistic levels (high frequencies) require more coefficients for an accurate representation than the higher levels (low frequencies). The number of chosen coefficients for each level is detailed in Table 4.1, based on evidence from Figure 4.1.

4.2.2 Analysis of the proposed multi-level representation

To test this representation, we measured RMSE (Hz) and correlation of reconstructed linear *f0* before and after parameterization. Table 4.2 shows that a CWT/DCT representation is comparable to either the CWT or the DCT separately. We lose less than 1% of the signal and RMSE is 2.66Hz.

At this point, the signal is represented by segments at 5 linguistically moti-

Representation	RMSE (Hz)	Correlation
CWT-only	1.96	.997
DCT-only (5 coeffs at syllable level)	2.43	.995
CWT-DCT	2.66	.995

Table 4.2: Representation reconstruction error and correlation.

vated levels, each with a fixed number of coefficients. Every phone in the utterance has an observation vector of 6 components representing high-frequencies, each syllable an observation vector of 6 components representing mid-frequencies, and so on.

Figure 4.2 illustrates the decomposition for one utterance. Each of the five scales is associated with a linguistic level (listed in the lower left-hand side corner of each subplot). Vertical dashed lines indicate forced-aligned boundaries at each of the levels, bootstrapped by forced alignment at the state level. For each window, demarcated by the segment boundaries, the discrete cosine transform is applied and the top N coefficients are preserved (where N is shown in Table 4.1 and on the lower left-hand side corner of each subplot of Figure 4.2).

The main motivation for the proposed representation is that each linguistic unit can model effects of varying frequency at different levels. Higher frequencies are modeled at phone-level, while slow-varying phenomena are modeled at phrase and utterance levels. The mid-frequencies should capture *f0* variation that would correspond to word or syllable-level effects. For example, considering the word-level signal in Figure 4.2, the word *ordered* appears to be the most prominent in the sentence, followed by *show me*, and the less prominent content word *place*. In terms of syllable variation, it can be seen from the same figure that the stressed syllables appear to have more prominent signal peaks. It is expected that these phenomena are captured over the entire data through the proposed representation.

Furthermore, in Suni et al. (2013), the sentence mean was removed while normalizing the signal and was ignored at training time. For synthesis, the authors have used the overall sentence mean inherited from the baseline model. In this

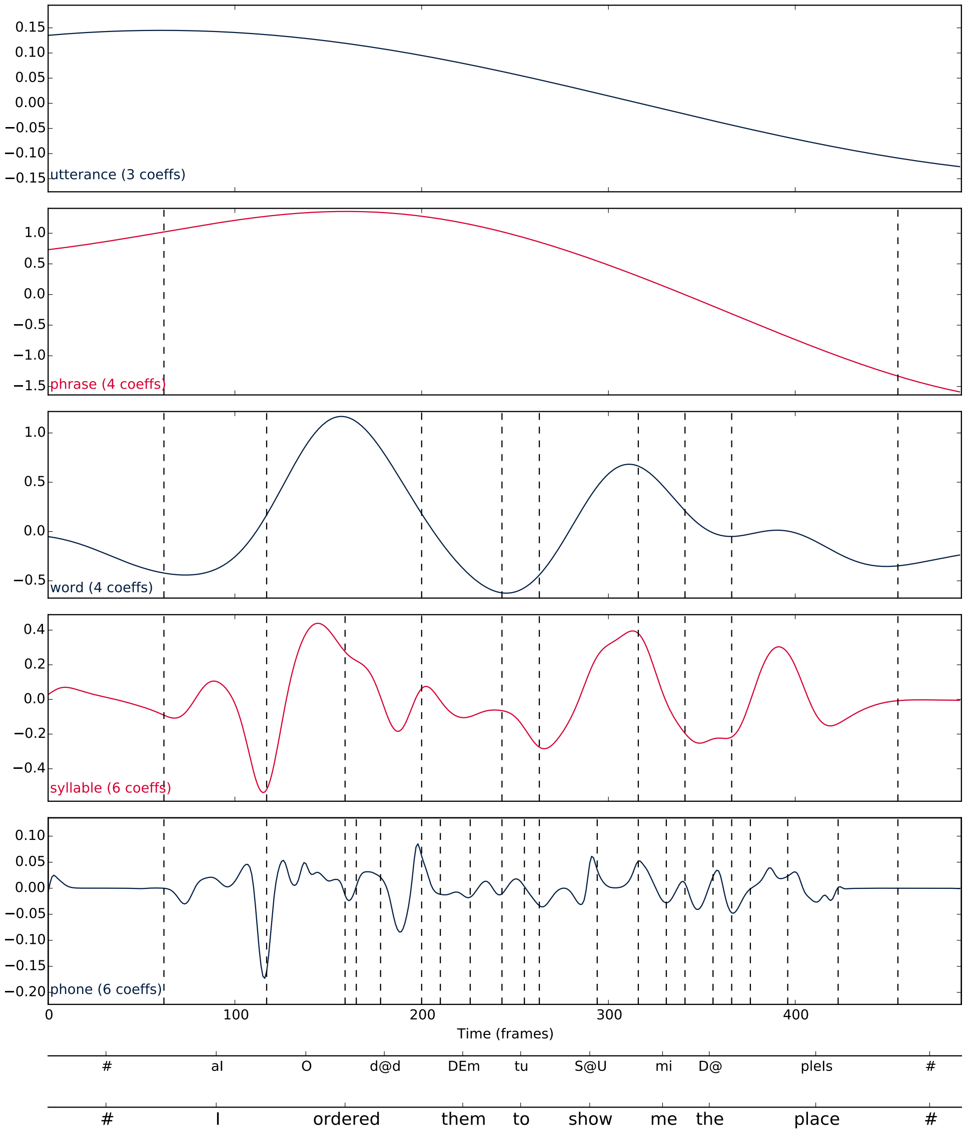

Figure 4.2: Five scale CWT decomposition with force-aligned boundaries (vertical dashed lines) at each linguistic level. The top DCT coefficients at each level (listed at the lower left-hand side of each subplot) are extracted for the signal between each boundary. The bottom axes list the syllables and the words for this example, with the axis marks placed at the mid-point of each syllable or word, considering the force-aligned boundaries .

work, we include sentence mean as the fourth component in the utterance-level observation vector, which is modeled jointly with the proposed representation.

This representation allows us to move beyond frame-level modeling, used by Suni et al. (2013), and model utterance-level effects at utterance level, and phone-level effects at phone-level.

4.3 Data

4.3.1 Audiobook data

For this task, we have used the freely available audiobook *A Tramp Abroad*, written by Mark Twain and first published in 1880, available from *Librivox*.[1] Audiobooks are a rich source of speech data, as the speaker often reads full chapters sequentially, thus making it ideal to explore prosodic phenomena such as phrase breaks or word prominences motivated by the discourse. It is also very expressive data, as the reader mimics the voices of characters and attempts to convey some type of emotion depending on the circumstances. The data has been pre-processed according to the methods described in Braunschweiler et al. (2010) and Braunschweiler and Buchholz (2011). We have used the manually selected subset consisting only of narrated speech described in Braunschweiler and Buchholz (2011), thus setting aside direct speech data. The reason for this is that we intend to focus only on expressive read speech that is influenced by higher-level phenomena, and avoid possible changes of speaking style and voice characteristics contained within the direct speech portions of the book.

4.3.2 ARCTIC data

The CMU ARCTIC speech synthesis database consists of phonetically balanced independent sentences designed for unit selection speech synthesis (Kominek and Black, 2004), available from *Festvox*.[2] For this task, we have used the female SLT voice, which was divided into 1000 training utterances and 113 testing utterances. The main motivation for the selection of this data is the contrast with the more expressive audiobook data. We hypothesize that the representation we intend to

[1] http://librivox.org

[2] http://festvox.org/cmu_arctic

explore is not useful with less expressive datasets. In this dataset, sentences were recorded separately and are not semantically related. The speaker often places a fairly neutral intonation on each sentence, which is ideal, for example, for unit selection systems, which rely on the concatenation of similar units to minimize join artifacts. Even though this data is expressive in its own way, it will contain less prosodic variation. Therefore, multi-level modeling is perhaps less relevant for these types of data.

4.3.3 Data preparation

Acoustic features were extracted every 5 ms using the STRAIGHT vocoder (Kawahara et al., 1999, 2001) via the Voice Cloning Toolkit (VCTK, Yamagishi et al. (2014)). These are 60 mel-cepstral coefficients, 1 *f0* value, and 5-band aperiodicities. To these features, the corresponding delta and delta-deltas were appended. HTK/HTS (Zen et al., 2007, 2009a) was then used via the Voice Cloning Toolkit to force-align the data. Context-dependent 5-state left-to-right hidden Markov models were used with monophone pretraining. The models learned on the training data were then used to force-align the train, validation, and test sets at the state-level. Higher-level alignment, such as phone, syllable, word, or phrase boundaries were inferred from the state-level alignment. All systems were trained using the linguistic feature set described in Appendix A.0.1.

4.4 Systems trained

To test the proposed *f0* representation, the following systems were trained.

MSD-HMM Standard *f0* MSD model using 5-state left-to-right HMMs at phone-level.

CF-HMM Continuous-F0 HMM using the interpolated *f0* signal. *f0* is modeled in a single data stream with joint dynamic features. A variant of this system controls for the number of parameters (denoted by the suffix **-pctrl*), which limits the size of the tree to be similar to that of the MSD-HMM.

CWT-HMM The 5-scale wavelet representation is modeled with HMMs, similarly to Suni et al. (2013). Each scale is modeled by a separate data stream, with joint dynamic features. A variant of this system also controls for the number of parameters (denoted with **-pctrl*), limiting the size of each wavelet scale decision tree to be similar to that of the MSD-HMM decision tree.

DCT-phn and DCT-syl Interpolated log-*f0* is represented at phone or syllable levels using 6 DCT coefficients. Since the CWT is not used for this representation, the signal was not normalized for zero mean and unit variance. Observation vectors were clustered with multivariate regression trees. For generation, the *f0* contour is found by traversing the decision tree and finding the predicted observation vector at its leaf node. The signal is then reconstructed using the IDCT at phone or syllable levels with force-aligned duration.

CWT/DCT-MRT and CWT/DCT-URT Normalized interpolated log-*f0* is first decomposed with the CWT, then each scale is represented by the DCT at each level. Multivariate Regression Trees (denoted by **-MRT*) are used to cluster observation vectors. Therefore, the model consists of 5 trees, one at each level. An alternative clusters each vector component using Univariate Regression Trees (denoted by **-URT*) to a total of 24 regression trees (one per vector component). For generation, the signal is found by first applying the IDCT, concatenating wavelet contours, and applying the wavelet reconstruction formula.

CWT/DCT-HMM Initial experiments have shown that higher frequencies are harder to predict than lower frequencies. This approach models the high frequencies (phone-level component) with 5 state left-to-right HMMs, using an individual data stream with joint dynamic features. The remaining components are modeled similarly to the *CWT/DCT-MRT* system, using multivariate regression trees.

4.5 Results

4.5.1 Objective results

The synthesized *f0* contours were compared to the reference contours for all 50 utterances in the test set for each dataset. As objective measures, we have used the traditional root-mean-square-error (RMSE) and Pearson's correlation coefficient. These measures are sensitive to duration, so to make all models comparable, segment durations for all systems were taken from the force-aligned natural speech from the held out test set. Each measure is computed at sentence-level over voiced-frames only, and the arithmetic average is taken over the entire test corpus. Results for the two datasets are shown in Table 4.3.

Given that the CF0-HMM and the CWT-HMM use considerably more parameters (larger trees) than the MSD-HMM, we have trained two systems (denoted with the suffix **-pctrl*) controlling for tree growth, such that tree size is comparable to the tree generated by the MSD-HMM in terms of the total number of parameters.

Audiobook data

Considering the results shown in Table 4.3, we observe that, when using expressive audiobook data, objective measures indicate that our proposed representation combining both the CWT and the DCT performs better than all other systems. The CWT-HMM using a smaller tree is the only system that surpasses it in terms of correlation. The CWT-HMM shows relevant improvements over the MSD and CF0 HMMs, which reinforces the relevance of performing signal decomposition for *f0* modeling. The DCT models perform the worst out of all systems, which suggests that some improvements might be achieved using more complex models with this representation, such as using dynamic features, as shown by earlier work (Teutenberg et al., 2008; Latorre and Akamine, 2008; Qian et al., 2011; Obin et al., 2011).

A series of t-tests show that, in terms of RMSE, there are significant differences between the MSD-HMM and the models using the CWT. The system

System	audiobook data		ARCTIC data	
	RMSE	Correlation	RMSE	Correlation
MSD-HMM	40.895	0.327	15.174	0.705
CF0-HMM	54.997	0.336	58.678	0.722
CF0-HMM-pctrl	54.984	0.469	58.751	**0.784**
CWT-HMM	**34.066**	0.455	**13.135**	**0.786**
CWT-HMM-pctrl	65.811	**0.534**	**12.905**	**0.799**
DCT-phn	43.737	0.238	35.043	0.268
DCT-syl	41.338	0.223	28.144	0.297
CWT/DCT-MRT	**32.462**	**0.493**	**13.949**	**0.760**
CWT-DCT-URT	**32.536**	**0.504**	**14.018**	**0.762**
CWT/DCT-HMM	**32.617**	**0.493**	**14.040**	**0.756**

Table 4.3: Objective measures for audiobook and ARCTIC data. Bold highlights indicate best performing results.

CWT/DCT-MRT shows a significant decrease in error when compared to the MSD-HMM system (t(49)=-9.124, p<.001). Similar results are observed for the CWT/DCT-URT, CWT/DCT-HMM, and CWT-HMM.

In terms of correlation, a different pattern is observed. All CWT systems improve significantly over the systems using the MSD-HMM. Looking at the CWT/DCT-MRT system, we observe a significant increase in correlation when compared to the MSD-HMM (t(49)=4.494, p<.01). The only baseline system with a higher correlation appears to be the CF0-HMM controlling for tree growth, showing no significant differences when compared to the CWT systems. This method, however, appears to increase RMSE. Comparing the CWT-HMM with the proposed CWT/DCT representation showed no significant improvements in terms of RMSE and correlation.

It should be noted that the more stable systems in terms of both measures appear to be the CWT-HMM without controlling for tree growth and the proposed CWT/DCT systems. These models behave consistently in terms of RMSE and correlation. The proposed representation has the advantage of considering long-term representations, as well as achieving similar quality with considerably less parameters.

ARCTIC data

With the more traditional less expressive database, signal decomposition and modeling over longer units appears to be less relevant. Most systems achieve similar error rates, although we do notice some improvement in terms of correlation for systems using signal decomposition (i.e. those that the CWT). As before, models using only the DCT perform the worst. It is unclear why the CF0-HMM models achieve such high error when compared to the alternatives.

In terms of RMSE, the proposed CWT/DCT does not show significant improvements over the baseline system MSD-HMM. However, the CWT-HMM with and without parameter control does show improvements. When compared to the CWT/DCT-MRT, for example, RMSE of the CWT-HMM is significantly lower (t(49)=4.403, p<.01).

A similar pattern is observed in terms of correlation. The proposed representation does not show significant improvements over the baseline MSD-HMM systems. However, the CWT-HMM proposed by Suni et al. (2013) does show a significant increase in correlation when compared to the MSD-HMM (t(49)=-4.836, p<.001).

4.5.2 Subjective results

A perceptual experiment was conducted on 3 selected systems from the objective results. 50 test utterances were synthesized with the *f0* contours predicted from the MSD-HMM, CWT-HMM, and CWT/DCT-MRT systems. Spectral and aperiodicity parameters were used from the MSD-HMM, thus only *f0* is different.

16 native speakers have judged randomized utterance pairs in a preference test with a “no preference” option. Utterance pairs were organized such that each participant only judged the same utterance pair once. Each utterance pair was judged 8 times for a total of 400 judgments per condition. The process was

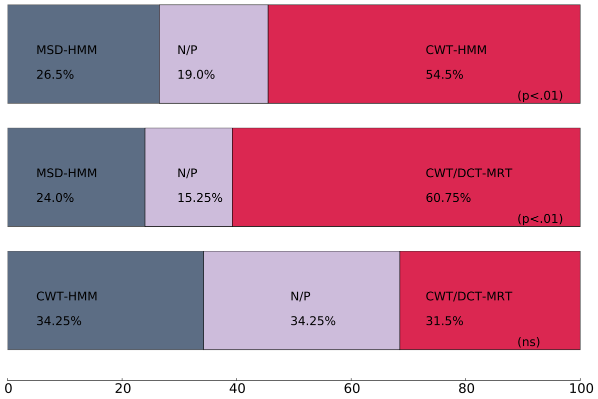

Figure 4.3: Audiobook data preference test results with N/P indicating "No Preference". In parenthesis, p-values indicate the results of 1-tailed binomial tests with an expected 50% split, with the N/P results evenly distributed over the remaining conditions.

identical for both datasets.

Figure 4.3 shows results for the audiobook data. We see percentage preferences and the results of a 1 tailed-binomial test assuming an expected 50% split, with the no-preference judgments distributed equally over the other two conditions. We see a significant preference for the CWT-HMM and CWT/DCT-MRT over the baseline MSD-HMM, but no preference between the CWT-HMM and CWT/DCT-MRT systems.

Figure 4.4 shows results for the ARCTIC data. The values in parenthesis indicate the results of a 1-tailed binomial test assuming an expected 50% split, with the no-preference judgments distributed equally over the other two conditions. Objective results have shown no relevant improvements over the baseline when it comes to the proposed CWT/DCT representation. However, a perceptual experiment does show a significant preference for the systems using signal decomposition over the system that does not. Although with smaller effects we observe the same pattern in the less expressive data as we did with the more expressive dataset. Participants prefer the proposed CWT/DCT system and the CWT-HMM over the traditional MSD-HMM, but they have no preference between the CWT/DCT and the CWT-HMM systems.

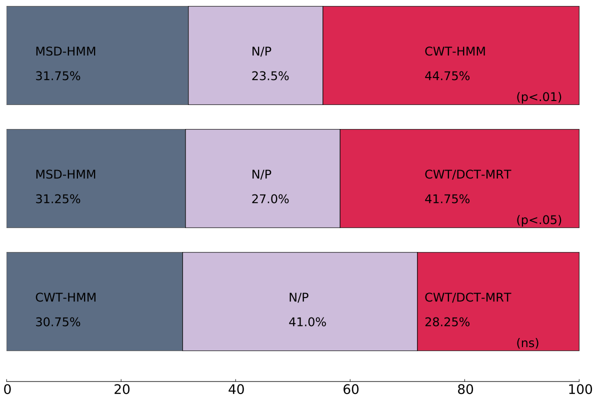

Figure 4.4: ARCTIC data preference test results with N/P indicating "No Preference". In parenthesis, p-values indicate the results of 1-tailed binomial tests with an expected 50% split, with the N/P results evenly distributed over the remaining conditions.

4.6 Discussion

4.6.1 Number of parameters

Table 4.4 shows the number of parameters for the three systems evaluated in Section 4.5.2. It is seen that the CWT-HMM trees tend to grow quite large. In the work of Suni et al. (2013), similar observations were mentioned. The decision trees learned over the wavelet signals tended to become larger and various attempts were made to control tree growth. No results were reported regarding the effect of tree size on the objective measurements. In the datasets we have experimented with, it was observed that controlling for tree size improves results slightly in the case of the less expressive data, but not in the case of the more expressive dataset. However, it should be noted that, even if tree growth is controlled in the CWT-HMM, it will still have 5 times more parameters than the baseline MSD-HMM. Observing the CWT/DCT systems, we notice a considerable decrease in the number of required parameters, having much simpler trees than both MSD-HMM and CWT-HMM, and achieving better or similar performance in terms of naturalness.

		Audiobook Data		ARCTIC Data	
		# Leaf Nodes	# Used Questions	# Leaf Nodes	# Used Questions
MSD-HMM	log-f0	46583	1846	5050	1095
CWT-HMM	phn	66632	1878	5822	1085
	syl	117987	1955	8895	1192
	wrd	159043	2015	14613	1292
	phr	271343	2014	13994	1271
	utt	213919	1961	11808	1267
CWT/DCT	phn	1938	865	318	214
	syl	699	433	108	84
	wrd	426	280	50	41
	phr	173	136	28	23
	utt	13	12	8	7

Table 4.4: Number of parameters for Audiobook and ARCTIC data.

4.6.2 Objective measures for individual wavelet components

Figures 4.5 and 4.6 show objective measurements in terms of RMSE and Correlation for individual scales for the two systems using the CWT. We include the CWT-HMM, modeled at state level, and the CWT/DCT representation, modeled at multiple linguistic levels. Note that typically wavelet components at various scales tend to have different ranges. Therefore, the RMSE results are not comparable across linguistic levels.

Although there appears to be some differences in both datasets in terms of the two systems, these differences were not enough to result in a clear perceptual difference. In the case of the audiobook dataset (Figure 4.5), we observe that, in general, for the lower linguistic levels (phone, syllable, word), the CWT/DCT system appears to outperform the state-level CWT-HMM in terms of RMSE and correlation. However, this trend is not observed for the ARCTIC dataset (Figure 4.6), where the two systems are closer, but the state-level CWT-HMM tends to outperform the multi-level CWT/DCT system.

These observations relate to the hypothesis that a multi-level framework would be suitable for more expressive data, as the speaker would include effects at vari-

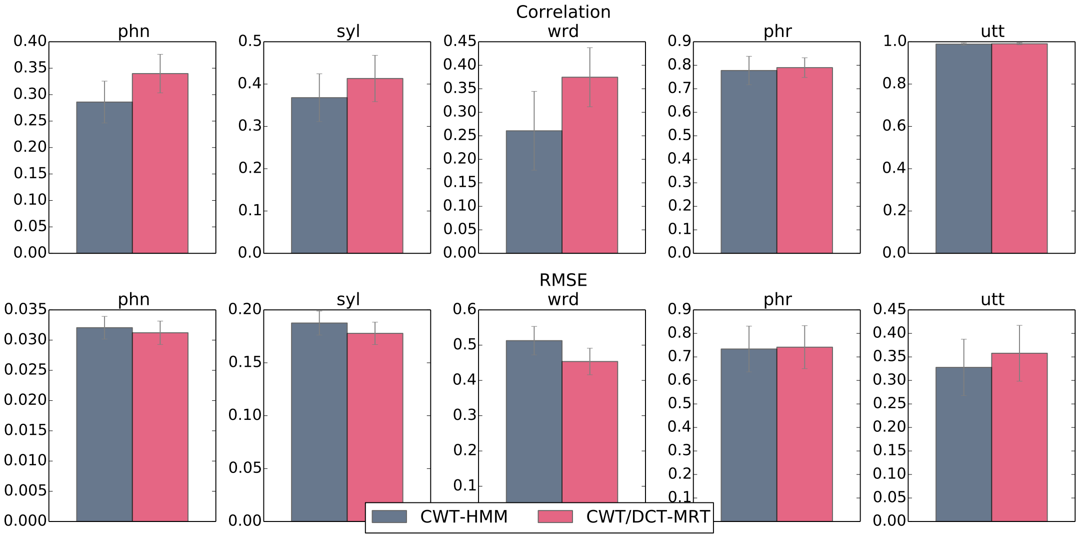

Figure 4.5: RMSE and Correlation of individual wavelet levels for Audiobook data with error bars denoting 95% confidence intervals for the sample mean.

ous linguistic levels, such as various degrees of word prominence or clearer phrase boundaries. The ARCTIC dataset was designed and recorded for unit selection systems (Kominek and Black, 2004). Sentences are spoken in isolation without any semantic relationship between them. The speaker tries to be consistent for each sentence, which leads to a fairly neutral (and similar) intonation over the entire dataset. The audiobook dataset, on the other hand, was not recorded for text-to-speech. Chapters are read sequentially, which motivates the speaker to place contextually-motivated prominences or phrase breaks, either in order to captivate the listener or to convey meaningful semantic information. A representation of *f0* at multiple linguistic levels could be more suitable for this type of data.

However, it is not clear what effect each of these linguistic levels has on the overall *f0* signal. We have observed that two systems using signal decomposition outperformed the system that does not. However, Figures 4.5 and 4.6 show that the lower-frequency components (phrases and utterances) achieve very good correlation when compared with the remaining components. It is not obvious where the improvements originate. For example, we could hypothesize that the advantages of using signal decomposition are mostly due to a direct modeling

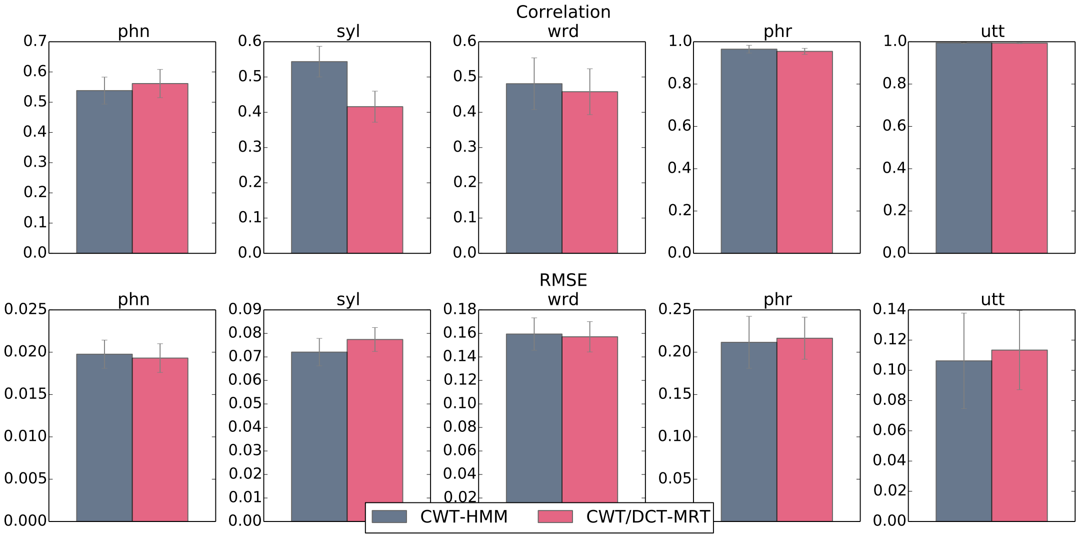

Figure 4.6: RMSE and Correlation of individual wavelet levels for ARCTIC data with error bars denoting 95% confidence intervals for the sample mean.

of middle or lower frequencies. It is unknown if the improvements observed on the systems doing signal decomposition with the CWT originate from a stronger modeling of these frequencies.

This set of experiments has shown that signal decomposition is beneficial and systems using it outperform systems that do not. But beyond this claim, it is unclear from these results what causes these systems to surpass traditional approaches. The following section discusses these hypotheses and how they are handled in subsequent chapters of this thesis.

4.6.3 Overall Discussion

These experiments indicate that signal decomposition is relevant to the modeling of *f0*, as all systems using the CWT performed better for expressive data than systems that do not. In terms of the proposed representation, we did not observe any relevant improvements over the CWT-HMM. However, the CWT-HMM by Suni et al. (2013) is still a state-level model, capturing long-term effects implicitly. It is therefore not surprising that it performs better in terms of objective measures on a less expressive dataset. State-level modeling is most likely a better choice in these cases, as the speech is not influenced by the dependencies related to the discourse. For example, each sentence in the ARCTIC dataset was recorded separately, whereas the sentences in the audiobook data are semantically related and were recorded sequentially. Prominences and phrase-breaks in the audiobook data may be motivated by the discourse (e.g. saliency of a word or a group of words).

We could argue, then, without any evidence to support this hypothesis, that the CWT/DCT representation lacks proper long-term linguistic features (e.g. words, phrases) to improve over the CWT-HMM. In these experiments, we were mostly limited to the standard set of shallow context features commonly used in speech synthesis (Tokuda et al., 2013). It has been shown that this feature set is not very effective when it comes to modeling prosodic naturalness (Cernak et al., 2013). Work exploring more complex representations of linguistic contexts would fall under sub-problem 2 of the main claim of this thesis (see Figure 1.1, p. 6). Chapters 7 and 9 explore additional feature representations that could be useful for the models proposed in this chapter. It is possible that a state-level system such as the CWT-HMM would be unable to leverage them, while a higher-level representation such as the CWT/DCT would.

It is also observed from these experiments that the importance of each scale is not yet clearly understood. In fact, in the work presented in this chapter, there are two main assumptions:

- All wavelet components and their respective linguistic levels, contribute equally to the *f0* signal.
- Each wavelet component, under the current decomposition strategy, can be meaningfully associated with a linguistic level.

The first assumption affects this work, as well as the work described in Suni et al. (2013), as we reconstruct *f0* by weighting all components equally. It was observed informally by the authors of Suni et al. (2013) that weighting components differently affects the overall *f0* contour. For example, giving more weight to the middle-frequencies has the effect of making speech more *resolute*.

The second assumption was acceptable in the work of Suni et al. (2013), as the CWT-HMM models all components at the same state level. However, this might have implications for the representation used in the CWT/DCT system, as this approach relies on the association of a given decomposition component with a specific linguistic level.

It is clear, then, that a deeper understanding of how this wavelet decomposition strategy affects and interacts with the *f0* signal is relevant for this work. In Chapters 5 and 6 we focus on the two assumptions. Chapter 5 investigates the first assumption with a set of perceptual experiments aimed at understanding the role of each decomposition component to the reconstructed *f0* signal. Chapter 6 focuses on the second assumption, attempting to relate decomposition components to linguistic levels.

4.7 Conclusion

This chapter investigated a multi-level representation of *f0* for a text-to-speech system, inspired by initial work with the continuous wavelet transform (CWT) described by Suni et al. (2013). The representation proposed in this chapter has the advantage of decomposing the signal into various frequencies which can be associated with multiple linguistic levels. This allows the system to model prosodic effects at their respective linguistic levels.

It was observed that systems performing signal decomposition perform better than systems that do not. This approach was evaluated on an more expressive and a less expressive dataset. Although the effects are not as strong, signal decomposition was shown to be useful for both types of data and to give considerably more compact parameters.

The lack of improvements of the proposed multi-level CWT/DCT representation over the state-level CWT-HMM might be due to a lack of proper suprasegmental features. Additional contextual features will be investigated in Chapters 7 and 9. CARTs might also not be the appropriate modeling technique, clustering each segment individually. More complex models, capable of leveraging the interactions between scales and over time might be more useful. These types of models will be explored in Chapters 7 and 8. However, two key assumptions affect the work described here: the meaningful association of wavelet components with linguistic levels, and their equal contribution to the overall *f0* signal. These will be investigated in Chapters 5 and 6.

Chapter 5

A perceptual investigation of wavelet-based decomposition of f0

This chapter covers the work described in "A perceptual investigation of wavelet-based decomposition of f0 *for text-to-speech synthesis" (Ribeiro et al., 2015), which was presented at Interspeech 2015.*

In this chapter, we consider a wavelet-based decomposition strategy for f0 *and we investigate the perceptual relevance of the found decomposition components. To achieve this, the* f0 *signal is reconstructed with selected decomposition components and native listeners are asked to judge the naturalness of synthesized utterances relative to that of natural speech. Results indicate that HMM-generated f0 is comparable to the wavelet components with lower frequencies, suggesting it mostly generates utterances with neutral intonation. Middle frequencies achieve very high levels of naturalness, while very high frequencies are perceived to be mostly noise.*

5.1 Introduction

Wavelets have been used in a variety of applications in speech processing for a number of years (Farouk, 2014). Recently, there has been a growing interest in the application of wavelets for the analysis and modeling of prosody in the context of statistical parametric speech synthesis (Vainio et al., 2013; Suni et al., 2013; Ribeiro and Clark, 2015). Chapter 4 introduced a multi-level representation of the *f0* signal using the continuous wavelet transform (CWT) and the discrete cosine transform (DCT). The general conclusion appears to be that approaches using signal decomposition outperform approaches that do not.

However, it was also shown in Chapter 4 that there are two underlying assumptions in the proposed wavelet-based decomposition strategy. The first is that *all wavelet components contribute equally to the reconstructed signal.* The second is that *decomposition components can be meaningfully associated with linguistic levels.* This chapter focuses on the first assumption. A set of listening evaluations is conducted, aimed at understanding the contribution of each wavelet decomposition component to the overall signal.

Based on evidence from the work of Suni et al. (2013), it is known that some components can be manipulated in order to affect the reconstructed *f0* signal. In the previous chapter, we observed that some components are harder to predict than others. We hypothesize that, on expressive datasets, short-term modeling approaches (such as MSD-HMMs) tend to average most *f0* variation and generate a neutral contour that is comparable to the low frequencies of the Continuous Wavelet Transform. Approaches that use signal decomposition outperform these short-term approaches because, on top of the easily predictable low frequencies, there is always some improvement from the explicit modeling of middle or low frequencies.

With this in mind, we propose to explore the following hypotheses in a series of experiments:

- Listeners respond more to middle frequencies (scales 5 to 8) and associate them with higher levels of naturalness when compared to other scales.

- Listeners do not respond much to low frequencies (scales 1 to 4) and they achieve comparable naturalness to *f0* synthesized from an HMM-based system.

- High frequencies (scales 9 and 10) do not contribute significantly to perceived naturalness nor do they contain relevant information for the reconstructed signal.

To evaluate these hypotheses, we will conduct four perceptual experiments. In the *first* experiment we will measure listeners' perception of selected wavelet ranges under a specific task. By asking participants to judge which word appears more prominent in an utterance, we will test whether or not a given wavelet scale contains some type of variation that is perceptible to the listeners.[1]

In the *second* experiment, we give listeners different utterances and ask them to judge whether or not they are similar in terms of naturalness. Considering the ratio of dissimilarity between wavelet scales, we use Multidimensional Scaling (MDS, Borg and Groenen (2005); Mayo et al. (2005)) to establish a perceptual distance between all scales and natural speech.

In the *third* and *fourth* experiments, we will measure naturalness by asking listeners how much each utterance resembles natural speech. In the first of these, we run traditional Mean Opinion Score (MOS) tests, where participants are asked to rate an utterance on a scale of 1–5. In the final experiment, we run a MUltiple Stimuli with Hidden Reference and Anchor test (MUSHRA, ITU-R Recommendation BS. 1534-1 (2015)), in which listeners rate an utterance against a reference and against all other conditions. The key difference between the MOS and MUSHRA evaluations is that, in the first, participants rate an utterance without any reference. In the second test, participants are given a reference and asked to judge each sample against that and against all conditions simultaneously. Since identical stimuli are used for the two evaluation approaches, these results should also give us insights regarding the two testing methodologies.

[1] This is a different approach to the task described in Vainio et al. (2013) or Suni et al. (2017). The work described in Vainio et al. (2013) annotates prominence with features extracted with the continuous wavelet transform. The automatic annotation is then compared against the annotation given by expert native listeners. It is the goal of that evaluation to show that the continuous wavelet transform can be used for the automatic annotation of word prominence. The work of Suni et al. (2017) extends this notion and uses a signal combining *f0*, intensity, and duration to annotate prominence and prosodic constituent boundaries, which is then compared with manually annotated ToBI labels.

5.2 Reconstruction of the f0 signal

The *f0* signal is normalized according to the steps of Vainio et al. (2013) and Suni et al. (2013), briefly described on p. 73. The *log-f0* signal is linearly interpolated over unvoiced regions and reduced to zero mean and unit variance. To decompose *log-f0*, we use a continuous wavelet based decomposition approach identical to that described in Chapter 4, as well as in Suni et al. (2013), using 10 wavelet scales, each one octave apart. A visualization of the decomposition is given in Figure 3.3 (p. 64).

For reconstruction, we use a variation of the *ad hoc* reconstruction formula proposed by Suni et al. (2013):

$$f_0(x) = \sum_{i=1}^{10} w_i C_i(x)(i + 2.5)^{-5/2} \tag{5.1}$$

where scale 1 corresponds to the highest frequency scale and w_i is the weight given to scale i where $w_i \in \{0, 1\}$. There is an inverse relationship between scales and frequency. For clarity, we index components by their frequency. Components with a lower index therefore correspond here to low frequency components. Table 5.1 shows all experimental conditions with these components indexed by increasing frequency. Note that these frequencies are related to the wavelet component and not the pitch range of the speaker. For each condition, *f0* is reconstructed from the wavelet domain zeroing out selected frequencies. For example, for condition *1-2*, the weight vector would be $\vec{w} = [0, 0, 0, 0, 0, 0, 0, 0, 1, 1]$. This guarantees that the signal is reconstructed as before, but preserving only selected components.

5.3 Experiment 1: prominence annotation

5.3.1 Data

For this experiment, we recorded a female native speaker speaking simple utterances in response to different stimuli. All sentences consisted of 3 content words and had similar syntactic structure. The stimuli were chosen in order to suggest

Condition	Description	Freq. (Hz)
natural	Vocoded speech using natural parameters	-
all	All *f0* frequencies.	0.1-50
1-2	Low frequencies. Components indexed at 1 and 2.	0.1-0.2
3-4	Low frequencies. Components indexed at 3 and 4.	0.4-0.8
1-4	All low frequencies. Components indexed at 1, 2, 3, and 4.	0.1-0.8
5-6	Middle frequencies. Components indexed at 5 and 6.	1.6-3.2
7-8	Middle frequencies. Components indexed at 7 and 8.	6.3-13
5-8	All middle frequencies. Components indexed at 5, 6, 7, and 8.	1.6-13
9-10	High frequencies. Components indexed at 9 and 10.	25-50
MSD-HMM	*f0* signal predicted from an MSD-HMM.	-

Table 5.1: Experimental conditions with approximate CWT frequency ranges.

different pitch accent locations in the responses.[2] Table 5.2 exemplifies given stimuli and requested responses for a single utterance. The full set of stimuli and responses for this short database can be found in Appendix B. The absence of stimulus (illustrated by the use of ellipsis) suggests a neutral (or random) response from the speaker, without a clear indication of prominence. The remaining stimuli suggested the placement of prominence in one of the three content words of the response: subject, verb, or object. We recorded 10 different utterances in 4 different contexts resulting in a total of 40 utterances. For each utterance, the stimulus was given to the speaker via a headset and the speaker would respond with the suggested utterance.

5.3.2 Design

For each of the 40 utterances, speech parameters were extracted using the STRAIGHT vocoder (Kawahara et al., 1999, 2001). This evaluation relies on copy synthesis. It does not use artificially generated parameters, other than the *f0* signal used with the MSD-HMM condition. When synthesizing the evaluation data, all conditions except *natural* use spectral envelope and aperiodicity parameters from the neutral response. The condition which we call *natural* is, in fact,

[2] We thank Robert Clark for suggesting the recording of this dataset.

stimulus	response
...	John won at Mary's.
Paul won at Mary's?	John won at Mary's.
John lost at Mary's?	John won at Mary's.
John won at Kate's?	John won at Mary's.

Table 5.2: Stimuli and responses for one utterance in the data set. Ellipsis indicates that no stimulus was given to the speaker.

vocoded speech using all original parameters.

For the remaining experimental conditions, the *f0* signal was parameterized at the syllable level with the DCT and reconstructed to match the corresponding duration of the vocoder parameters extracted from the neutral response. This ensures that speakers will not respond to durational or intensity cues when judging the utterances, as the only difference between them is the fundamental frequency.

We have a total of 400 unique utterances (10 sentences x 4 contexts x 10 conditions). Each of the 25 participants listened to a randomized subset of 80 utterances. They were asked to select which of the 3 words appears most prominent or salient in each utterance, with the option to indicate that all words appear equally prominent.

5.3.3 Results

From the expected 2000 judgments (25 participants x 80 utterances), 23 were missing. This left us with an average of 198 judgments per condition, with each unique utterance having been judged either 4 or 5 times.

To analyze the results, we used an approach similar to that described in Cole et al. (2010). That is, we considered the results from the *natural* condition as the gold set to which we measured the accuracy of all other conditions. Therefore, for each utterance, the response with the most votes across all natural speech judgments was taken to be the correct choice. Across the *natural* condition, we

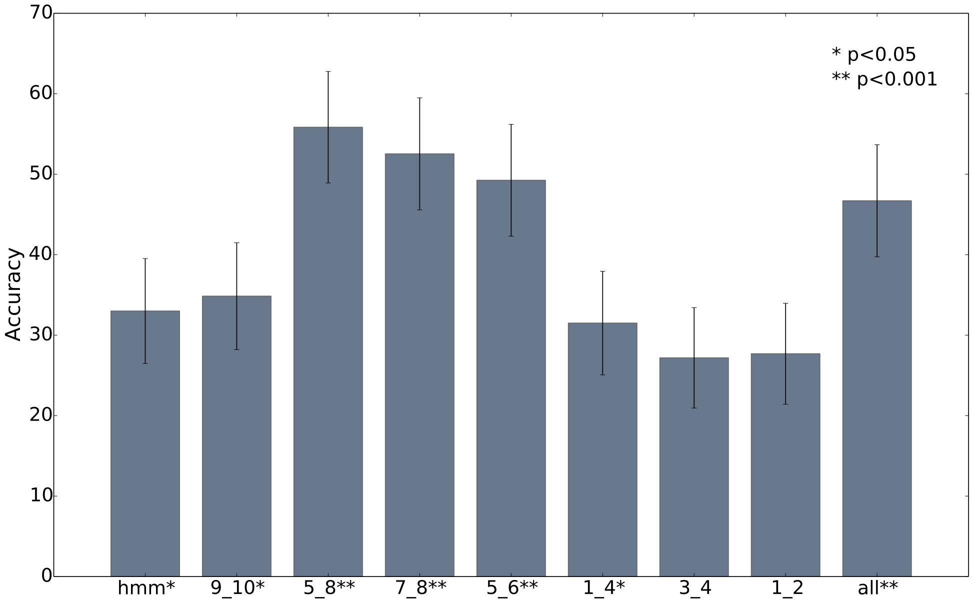

Figure 5.1: Accuracy results for prominence detection evaluation. Vertical error bars indicate 95% confidence intervals estimated with the Normal Approximation Method of the binomial confidence interval. Notes with asterisks indicate the result of a 1-tailed binomial test assuming a chance accuracy of 25%.

observed an *annotator* agreement of 86%.[3]

Figure 5.1 shows the accuracy results for the non-natural conditions, assuming the most likely answer given in the natural speech as the ground truth. The notes on the graph indicate the result of a 1-tailed binomial test given a chance accuracy of 25%. All conditions using the middle frequencies achieve accuracies that are significantly above chance. Some conditions achieve a smaller, although significant effect, where we would expect not to see any (such as the HMM condition, for example). The reason for this might be how we are computing the results. When faced with uncertainty, listeners might default to the same answer. This could be, for example, that all words are equally prominent or that the subject of the sentence carries the prominence. This would not be unlikely,

[3] In this task, the annotators essentially provide a likelihood of a given response being prominent. An alternative approach to analyze these results would be to consider this distribution over the possible choices, rather than picking the most likely one as the ground truth. This method would have been more in line with the approach proposed in the work of Cole et al. (2010).

given that the remaining acoustic parameters are from the neutral response, so we would expect listeners to perform well under this condition. If we compute a 1-tailed binomial test assuming instead a chance accuracy of 1/3, then the three conditions observing a small effect will not deviate significantly from chance.

Results in Figure 5.1 show that the middle frequencies tend to carry a good portion of the *f0* signal and listeners are able to perceive these differences. It was surprising to observe that the condition using only these components scored a higher accuracy than the condition using the reconstructed signal with all components. One interpretation of these results might be that the pitch excursions present in the *f0* signal become more noticeable once we remove short and long term frequency information. In this case, given that the remaining parameters correspond to the neutral response, listeners identify those pitch excursions as a prominence marker. This can be related to the approach for automatic annotation of prominence proposed by Vainio et al. (2013), which uses the middle frequency components associated with the word level.

5.4 Experiment 2: naturalness similarity

5.4.1 Data

To conduct this experiment, we used the freely available audiobook *A Tramp Abroad*, as described in Section 4.3.1 (p. 78). A standard 5-state left-to-right HMM system was trained on roughly 5000 utterances (approximately 9.5 hours of speech). 20 utterances not in the training set were randomly selected for these experiments. The *natural* condition is vocoded speech, which uses all natural acoustic parameters. All remaining conditions use the same mel-cepstral, band aperiodicity, and voicing parameters predicted from the HMM system, and duration is derived from the force-aligned data. These conditions vary only in terms of *f0*, as shown in Table 5.1.

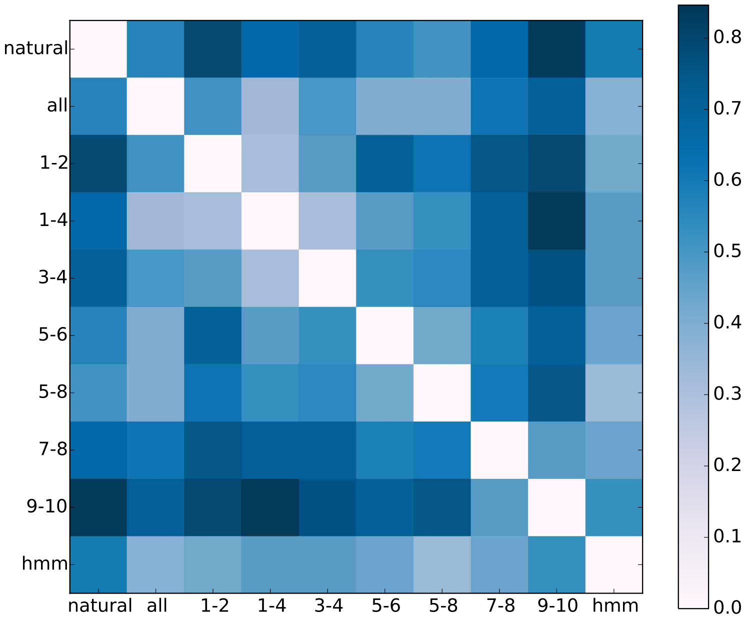

Figure 5.2: Dissimilarity ratio matrix for experiment 2. Lighter colors indicate lower dissimilarity ratio between conditions, while darker colors indicates higher dissimilarity ratio.

5.4.2 Design

The 20 utterances selected from the test set were synthesized according to the 10 conditions described in Table 5.1. 10 native listeners participated in the experiment, each rating a total of 144 utterance pairs. Participants were instructed to listen to each pair carefully and judge if the pair is similar or different in terms of naturalness. Each pair given to the participants consisted of different utterances and different conditions. Within any three consecutive pairs, the same condition and utterance is not repeated. This prevents the task from being too easy and discourages participants from judging all comparisons as different. This follows the methodology described in Merritt and King (2013) and Henter et al. (2014).

5.4.3 Results

Considering the 45 distinct condition pairs, each pair was judged at least 32 times. A 10x10 dissimilarity matrix was constructed, indicating the fraction of

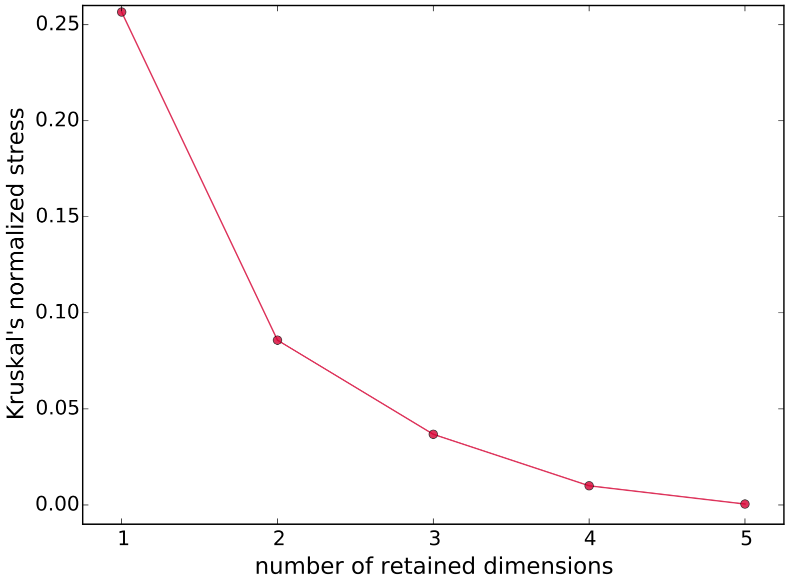

Figure 5.3: Kruskal's normalized stress as a function of the number of dimensions.

times each pair was judged as different, which can be seen in Figure 5.2.

Multidimensional Scaling (MDS, Borg and Groenen (2005); Mayo et al. (2005)) was used to embed the systems into a 2-dimensional space based on the dissimilarity matrix. The Euclidean distances between points in this space is representative of their similarity. We have used the function *mdscale* from the Matlab statistic toolbox with Kruskal's normalized stress1, which returned a stress value of 0.086.[4]

Figure 5.3 shows the stress value as we increase the number of dimensions. As more dimensions are used, the more accurate is the representation of the points in the low-dimensional space. We have chosen two dimensions, as it shows a considerable decrease in error and the visualization of the results appears to be consistent with the dissimilarity matrix. Figure 5.4 shows the two-dimensional representation of the 10 conditions as judged by the participants. Distances between points are representative of their dissimilarity in terms of naturalness.

[4] The stress value represents the loss from the low-dimensional approximation of the data. With a value of 0 indicating no error and a perfect representation of the data. More dimensions naturally lead to better approximations, but they become hard to visualize. Further details can be found in Borg and Groenen (2005).

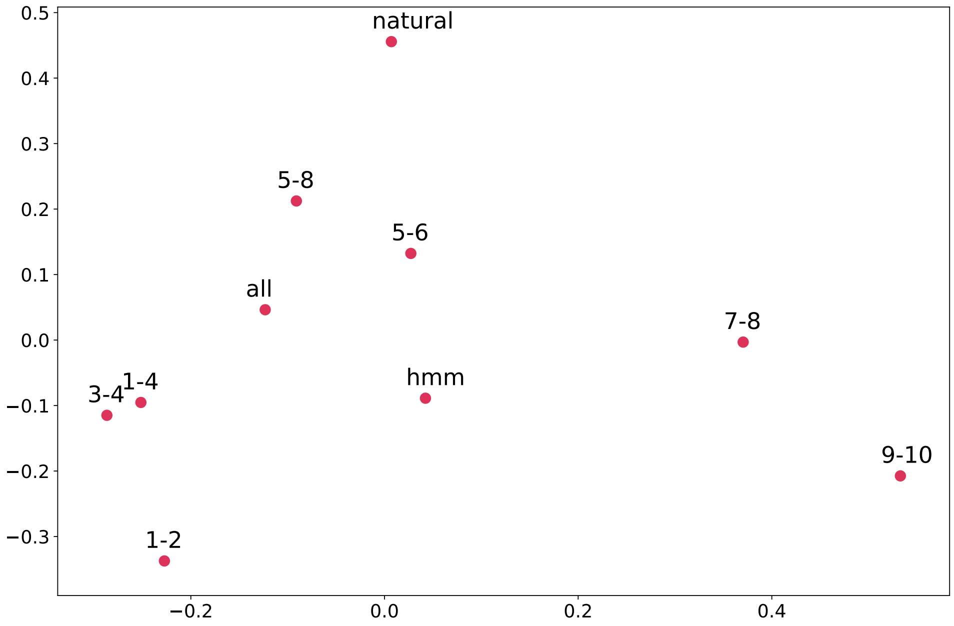

Figure 5.4: Two-dimensional representation of the dissimilarity matrix as estimated by MDS. Each point represents one condition and distances are representative of their dissimilarity in terms of naturalness.

Systems that are close in the embedding space are judged to be similar by the native listeners, while points that are farther apart are judged to be dissimilar.

Listeners appear to have naturally clustered the conditions using most of the low frequency components from the CWT (1-2, 3-4, 1-4), as well as conditions using high frequency components (7-9 and 9-10), and mid-frequency components (5-6 and 5-8). The condition with *all* frequencies (reconstructed *f0* signal) is closer the conditions using only the mid-frequencies, which suggests that these components carry most of the information for the signal. Quite surprisingly, the reconstructed *f0* signal appears farther from *natural* speech than the utterances using only selected components (5-6 and 5-8).

5.5 Experiment 3: referenced evaluation

5.5.1 Design

In the MUSHRA test, participants are asked to rate all conditions of the same utterance in parallel from 0 (very poor) to 100 (very natural). Each condition has one slider and listeners are given the *natural* (vocoded) condition as reference. This utterance is also included in the unlabeled conditions and participants are instructed to judge at least one utterance as completely natural. This fixes the high end of the scale and all conditions are judged in relation to this.

The same data described in Section 5.4.1 was used. 10 native listeners rated all 20 sets of 10 stimuli, each stimulus originating from the conditions detailed in Table 5.1. The order of the stimuli was randomized for each participant.

5.5.2 Results

From the expected 200 sets, 48 were discarded due to the hidden reference being judged as less than completely natural. These were excluded from the analysis. Figure 5.5 illustrates the distribution of the remaining 152 sets for all utterances and participants. Listeners ranked the CWT middle frequencies higher than the CWT low or high frequencies, with the HMM generated *f0* being comparable to the CWT low frequencies. Table B.1 in Appendix B shows the full results from Bonferroni-corrected pairwise Wilcoxon sign rank tests on all conditions.

Comparison of the ranked systems shows that there were no significant differences among all three conditions using low frequency components (conditions 1-2, 3-4, and 1-4). The condition using HMM-generated *f0* also shows no significant differences from any of the systems using low frequencies, which supports one of the initial hypothesis of this work. The system using the components previously associated with the syllable level (condition 7-8) is comparable in terms of naturalness to the system using low frequencies and the systems using HMM-generated *f0*. All other systems are significantly different, including the systems using the fifth and the sixth components, which are ranked higher than all other conditions, except the one using all signal components.

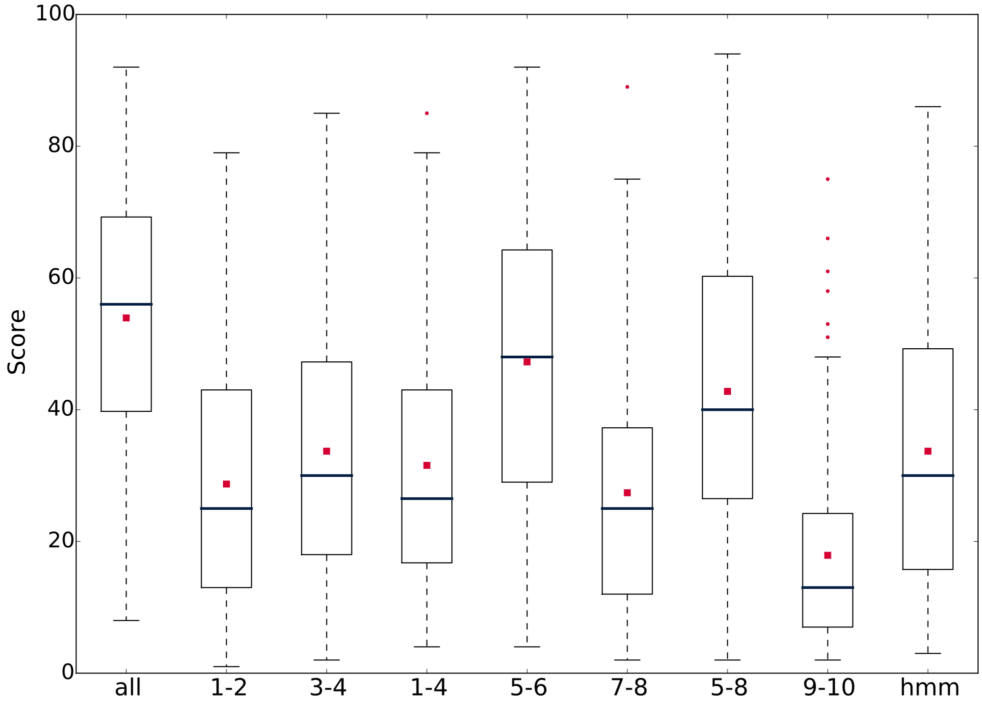

Figure 5.5: Boxplot for the MUSHRA evaluation, where the y-axis denotes the score given by participants on a 0-100 scale and the x-axis denotes the systems that were evaluated. Dark blue horizontal line shows the median and the red square shows the mean.

5.6 Experiment 4: non-referenced evaluation

5.6.1 Design

In the MOS (Mean Opinion Score) test, 25 participants were asked to rate each utterance on a scale of 1 (completely unnatural) to 5 (completely natural). No other instructions were given to the participants, therefore, unlike the MUSHRA evaluation, participants have no reference against which to judge each utterance. All utterances were randomized for each participant. This experiment uses the same data described in Section 5.4.1.

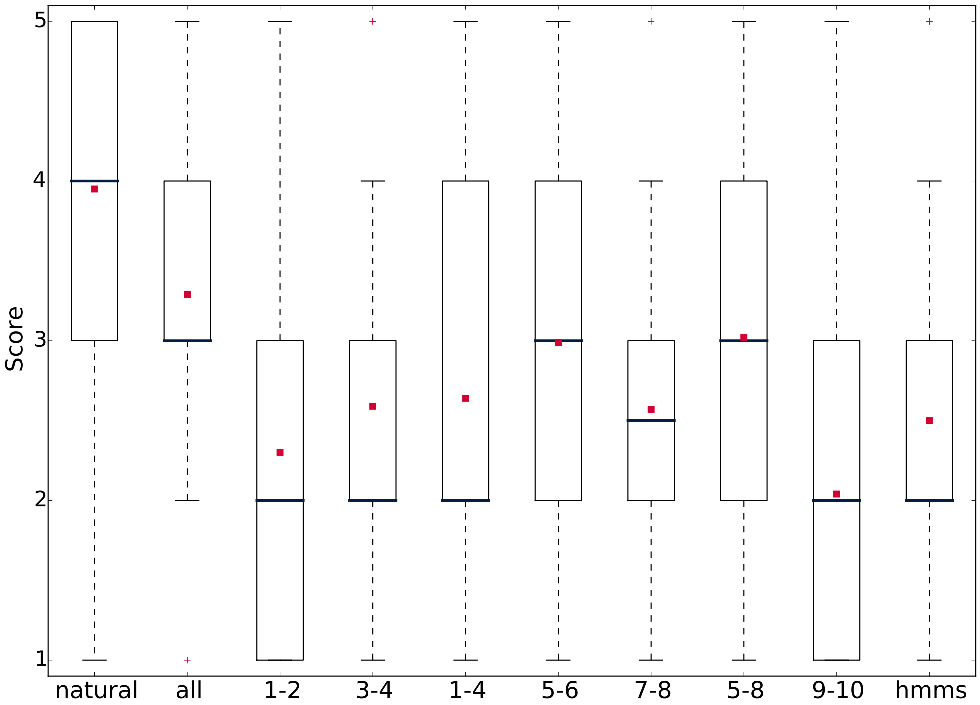

Figure 5.6: Boxplot for the MOS evaluation, where the y-axis denotes the score given by participants on a 1-5 scale and the x-axis denotes the systems that were evaluated. Dark blue horizontal line shows the median and the red square shows the mean.

5.6.2 Results

From the expected 1000 judgments, 1 was missing. Each unique utterance was judged 5 times and all conditions had 100 judgments, except 1 condition which had an utterance with 4 judgments and a total of 99 scores.

Figure 5.6 shows a boxplot with the results for the MOS test. Listeners were quite conservative in the judgment of the natural speech, which has a median of 4, possibly because it was given as vocoded speech. It still performs higher than the remaining conditions, followed by, as expected, the condition that reconstructs *f0* with all frequencies. Table B.2, given in Appendix B shows the full results for Bonferroni-corrected pairwise Wilcoxon sign rank tests on the mean opinion score for all conditions evaluated.

Considering the condition using the reconstructed signal with all components, the only conditions that show no significant differences are those using the middle frequencies 5-6 and 5-8. These are also the only conditions with selected components that are rated significantly higher than the HMM-generated *f0* signal. However, when comparing these two conditions with the remaining frequency ranges, we observe no significant differences with the condition using low frequencies (conditions 1-4, 3-4) and selected middle frequencies (condition 7-8). The two conditions considering only the very low frequencies (condition 1-2) and the high frequency components (condition 9-10) are only significantly different than the systems using the 5th and 6th components (conditions 5-6 and 5-8).

5.6.3 Referenced and non-referenced evaluations

The comparison of the results from the referenced evaluation (MUSHRA) and the non-referenced evaluation (MOS) is worth mentioning. In general, the ranking of the conditions is similar in the two evaluations, which suggests that the two tests are similar in terms of the results. In both cases, the middle frequency conditions using the fifth and the sixth components are ranked higher than all other conditions, suggesting that this particular range carries a large amount of information with respect to naturalness.

However, it is clear that the results from the MOS evaluation are more conservative in terms of the differences between the systems. One obvious difference is that, lacking a reference stimulus, listeners judged whether the stimulus that they listen is likely to occur against their internal references. Referenced tests, such as the MUSHRA, in a sense, measure the deviation from a speech sample. In this case, for the evaluation conducted in this chapter, we evaluate *how much of the signal is carried in a given component.* With a non-referenced test such as MOS, we evaluate instead *how acceptable or natural a given signal is.* In this case, the amount of non-significant differences makes sense. If, for example, as initially hypothesized, the low frequency components correspond to an *f0* signal associated with neutral intonation, then it is still a valid contour, depending on the context in which it is given.

These observations follow the findings reported in Latorre et al. (2014), which investigated the differences between referenced and non-referenced evaluation methodologies. The evaluation of natural and context-appropriate intonational patterns in synthesis applications still remains an open problem (see Section 10.3 of this thesis for thoughts on future work in this direction).

In any case, regardless of referenced or non-referenced methodology, the results indicate the two conditions that stand out by being closer to the condition using all frequency components are those that use the 5th and the 6th components. These were also rated higher than other components in the evaluations reported in Sections 5.3 and 5.4.

5.7 Discussion

The main hypotheses motivating this set of evaluations derived from a key assumption made in Section 4.6.3 of Chapter 4: that *all wavelet components and their respective linguistic levels, contribute equally to the f0 signal.*

Given the observed results, we find evidence to reject this hypothesis. Native listeners tend to prefer the middle scale components of the continuous wavelet transform when used to decompose the *f0* signal. The 5th and 6th components consistently show higher degrees of naturalness when compared to the remaining components. These are the components that have been previously associated with the word level in Chapter 4. A visualization of the sum of these two components is given in Figure 5.7, alongside the original *f0* contour. The utterance shown was extracted from the database used for the prominence evaluation, described in Section 5.3.1.

The 7th and 8th components were also hypothesized to rank higher than others, given that these were expected to capture syllable-level effects, as proposed in Chapter 4. However, the results from this investigation show that the 7th and the 8th components behave similarly to higher frequency components. That is, alone they don't contribute much in terms of naturalness, but when used with the middle components 5 and 6, results tend to improve. This is observed in experiments 1 and 2, in which the condition reconstructing *f0* with all four components achieves better results. However, in experiments 3 and 4, no significant differences are seen in the judgments between the combination of components 5

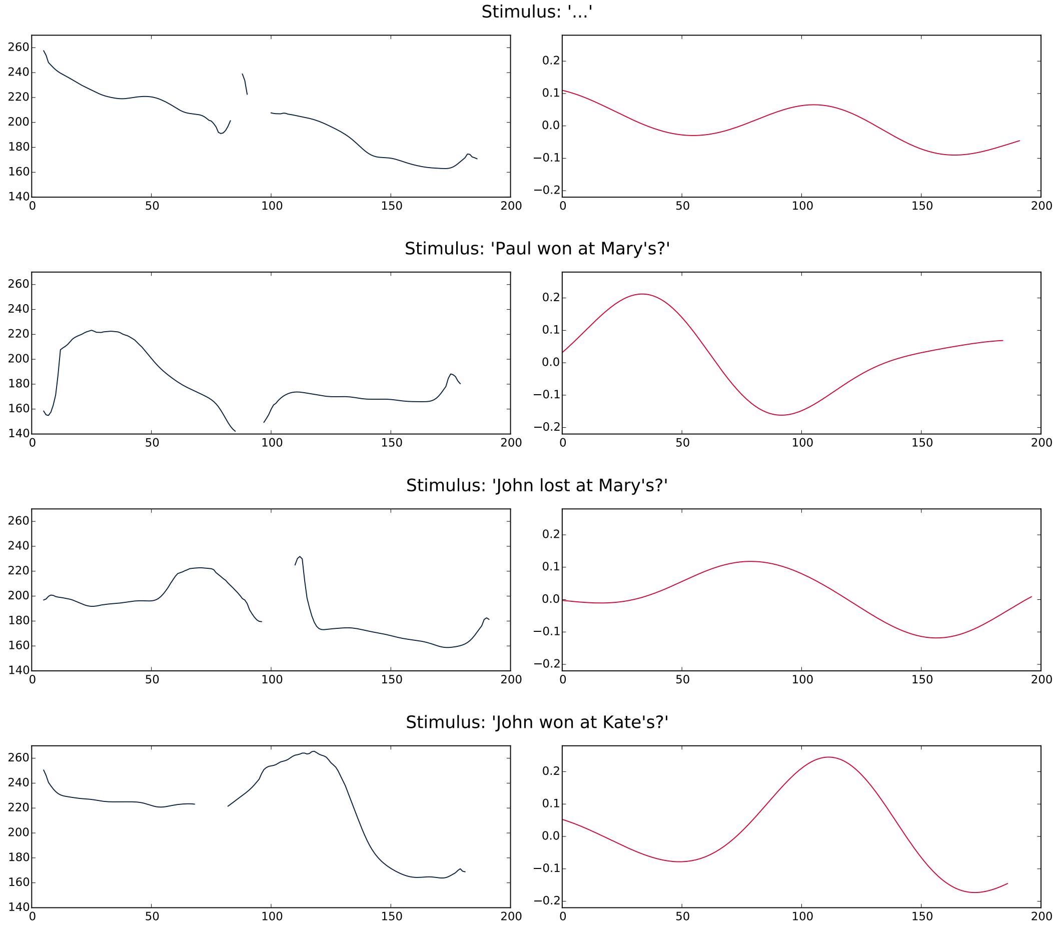

Figure 5.7: Response signals for the sentence "John won at Mary's." given four different stimuli. The left column illustrates the *f0* signal and the right column the sum of the 5th and 6th wavelet components under a 10-scale decomposition strategy.

to 8 and components 5 and 6.

This suggests that listeners prefer components 5 and 6. In the first and second experiments, these perform better than the condition using all frequencies, which was unexpected. In experiments 3 and 4, they appear to behave as expected, ranking below all frequencies, but this difference is not significant in the MOS experiment. In the MUSHRA evaluation, however, this is a significant difference (Table B.2, p, 209). It is strongly suggested by these results that the mid-frequency components of a wavelet-based decomposition contain most of the information that listeners associate with naturalness in expressive speech.

Therefore, such components might provide a suitable candidate when exploring suprasegmental models of *f0*.

The second hypothesis expected the low frequency components to be similar to the *f0* signal generated by an MSD-HMM. These were the components associated with the phrase and utterance levels. Results observed in the MOS and MUSHRA evaluations show that no significant differences were found between the ratings of all the lower scales. Similarly, we have failed to observe significant differences between these components when comparing them to the HMM condition. This suggests that HMMs are not very effective at modeling expressive *f0*. A possible cause for this might be the focus on frame-level modeling combined with a lack of understanding of proper suprasegmental contexts. These models tend to average different effects, causing the generated contour to be neutral and similar to the natural low frequencies. However, it should be noted that HMM-generated *f0* does not appear to be completely similar to the low frequency components, as the results from experiment 2 indicate.

The final hypothesis claimed that the higher frequency components would carry little relevant information to the *f0* signal. The results provide evidence to support this claim, showing that the reconstructed *f0* using only these components is consistently rated lower than all other components. This might explain the lack of improvements when modeling them at frame-level with HMMs, while using suprasegmental approaches with the remaining scales, as proposed in Chapter 4. When rated in isolation in the MOS test, they appear to be comparable to the low frequencies (scales 1 and 2). However, when ranked directly, listeners tend to prefer the low frequencies, possibly due to a more stable intonational pattern over the utterance.

Therefore, we could hypothesize that the middle components might provide a good starting point for suprasegmental models of *f0*. We can think of these components as the signal stripped of most of its noise, preserving meaningful information to the listener, such as word prominence. The low frequencies can be captured and modeled by the traditional frame-level approaches such as MSD-HMMs, or even modeled directly with a very simple approach such as the one described in Chapter 4. It is possible that the high frequency components could simply be discarded, although this hypothesis was not investigated in this work.

The second key assumption described in Section 4.6.3 of Chapter 4 claimed that *each wavelet component, under the current decomposition strategy, can be meaningfully associated with a linguistic level.* The work presented in this chapter already showed some evidence to reject this assumption. For example, components associated with the syllable level (scales 7 and 8) tend to rate lower than expected, as observed in experiment 2. Chapter 6 investigates this notion more thoroughly, and, using these findings, proposes an alternative decomposition strategy using the CWT for the *f0* signal.

5.8 Conclusion

This chapter conducted a perceptual investigation of a wavelet-based decomposition strategy of the *f0* signal. It was observed that the mid-frequency components produced by this decomposition strategy are commonly associated with higher levels of naturalness. The lower frequency components are comparable to HMM-generated *f0* and, although they contribute to the reconstructed signal, they mostly reflect neutral intonation. High frequency components do not appear to contribute much to the signal and may be regarded as noise. This suggests that not all components from the decomposition contribute equally to the reconstructed signal, as was initially assumed by earlier work (Suni et al., 2013; Ribeiro and Clark, 2015).

Chapter 6

A dynamic wavelet-based decomposition of f0

This chapter covers the work described in "Wavelet-based decomposition of f0 *as a secondary task for DNN-based speech synthesis with multi-task learning" (Ribeiro et al., 2016c), presented at ICASSP 2016.*

The multi-level representation proposed in Chapter 4 was based on two key assumptions regarding a 10-scale wavelet based decomposition of the f0 *signal. While Chapter 5 investigated the initial assumption that all wavelet components contributed equally to the signal's reconstruction, this chapter will focus on the assumption that wavelet components can be meaningfully related to linguistic units. The distributions of peak rates for each wavelet component are compared with the distributions of linguistic levels. The observations from this evaluation lead to an alternative decomposition strategy, in which the mother wavelet is varied dynamically for each utterance. The two decomposition strategies are then evaluated as the secondary task of a feedforward deep neural network.*

6.1 Introduction

Chapter 4 proposed a multi-level representation of the *f0* signal using the continuous wavelet transform (CWT) and the discrete cosine transform (DCT). The first part of the representation was inspired by the work of Suni et al. (2013), using the CWT to decompose the *f0* signal into components that were associated with linguistic levels. It was argued in Chapter 4 that the work presented was mainly driven by two assumptions: (1) that all wavelet components are equally relevant to the reconstructed *f0* signal; and (2) that these components can be meaningfully associated with various linguistic levels.

Chapter 5 investigated the first assumption in a set of experiments that ranked each decomposition component against a reference or against each other. It was observed that the mid-frequency components were consistently rated higher in terms of naturalness than all other components. This suggests that these components carry most of the meaningful variation in the signal. Evidence found through these experiments does not support the first assumption made in earlier chapters.

The current chapter is motivated by the need to gain further insight into the second assumption. Section 6.2 provides a description and analysis of two wavelet-based decomposition strategies. Section 6.2.1 focuses on the decomposition strategy proposed by Suni et al. (2013) and used in Chapters 4 and 5. For the purposes of the current chapter, this strategy is termed the *static decomposition strategy*. An alternative strategy is proposed in Section 6.2.2, leveraging findings from earlier chapters, and it is termed the *dynamic decomposition strategy*. Section 6.3 evaluates the two decomposition strategies as the secondary tasks of an acoustic model in the context of text-to-speech synthesis.

6.2 Wavelet-based decomposition strategies

6.2.1 Static strategy

The decomposition strategy using the continuous wavelet transform investigated in earlier chapters used a fixed set of 10 scales. Each scale is separated by one octave, with the initial scale being set to be approximately the frame size (5ms). Under this decomposition strategy, the scaling of the mother wavelet (and its *frequency*) for each of the ten components is fixed across the entire dataset. Due to this property, for the purposes of this chapter, we term this approach a *static*

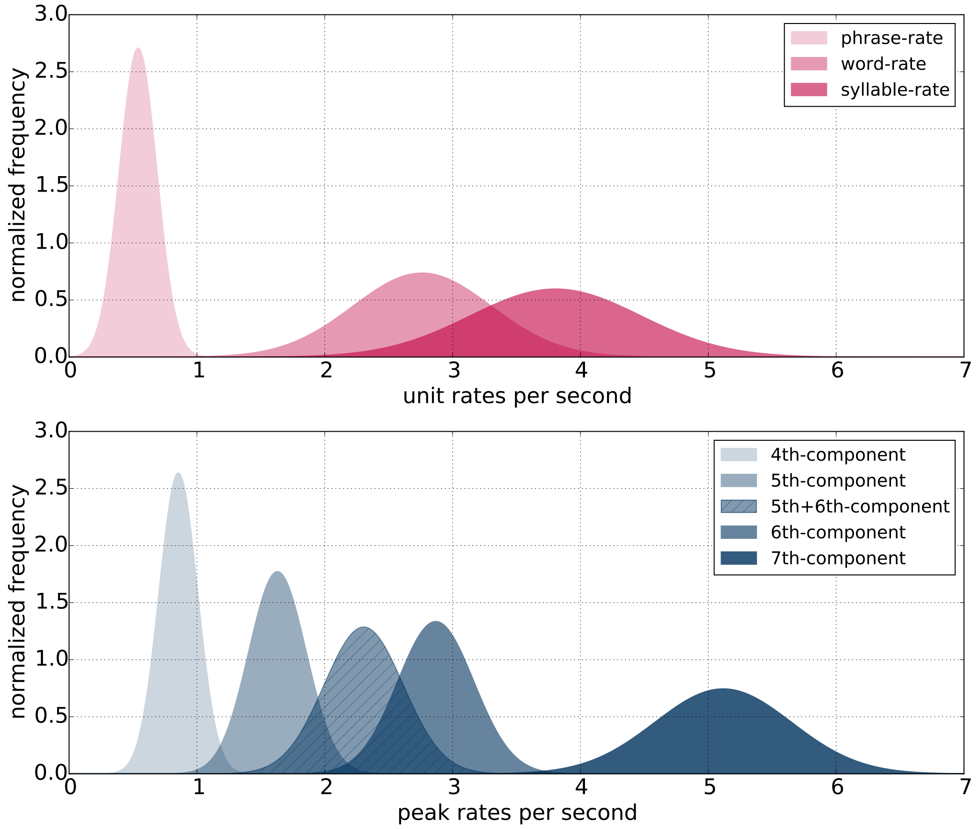

Figure 6.1: Top figure shows unit rate distributions for selected linguistic units. Bottom figure shows peak (local maxima) rates per second for selected components under a static 10-scale wavelet-based decomposition of the *f0* signal, including the sum of the 5th and 6th components.

decomposition strategy.

It is our goal to investigate the relationship between this decomposition strategy and the linguistic levels (e.g. words, syllables, phrases) across a large database. Linguistic levels are found from the front-end and unit boundaries derived from state-level forced alignment, as in Chapter 4. To visualize the relationship between decomposition parameters and linguistic levels, we consider the distributions of word and syllable rates, as well as peak (local maxima) rates per second on selected wavelet components. For any linguistic unit and decomposition component, if their respective unit and peak rates match within an utterance, we infer that the given component captures the variation associated with that linguistic unit. For example, at the word level, a component giving approximately 1 local maximum per word would correspond to an effect of the *f0* signal that could be related to the word frequency.

Rates are computed at the utterance level on 5000 utterances of the training set of the audiobook data described in Chapter 4. Figure 6.1 shows a Gaussian best fit for the normalized distributions given selected components and linguistic units. The selection of components and linguistic units illustrated in the figure was made based on findings from Chapter 5. We focus on the mid-frequency components and their relationship with syllables and words.

Although these might change depending on speaker or speaking rate, the 6th component approximately matches the distribution of words, suggesting it could be modeled at word level. The fact that the distributions match implies that the 6th component contains approximately 1 local maximum per word in the utterance, giving a approximation of, for example, word prominence.[1]

From Figure 6.1, we observe that the 5th component is best modeled at a level that is higher than the word. The distribution of phrases could be associated with the 4th component. Therefore, the 5th component lies at a level higher than the word, but lower than the phrase. The 7th and 8th component were associated with the syllable level, but evidence from Figure 6.1 suggests that their distributions do not match. This might explain the perceptual similarity observed in Chapter 5 between the condition using only the 7th and the 8th components and the condition using higher frequency components.

The distances between these distributions can be formally quantified with the root mean square error.[2] Table 6.1 summarizes key comparisons between linguistic levels and static decomposition components. The informal observations from the distribution visualization are supported by the error measures shown on the table. The phrase rate is well modeled by the sum of the 3rd and 4th components. The word rate is closer to the peak rate of the 6th component

[1] Variations of this idea have been used before using the Continuous Wavelet Transform to identify word-level prominence (Vainio et al., 2013; Vainio, 2014; Vainio et al., 2015).

[2] Although we acknowledge that other distance or divergence measures could be used, we choose here to use the root mean squared error since it allows for direct comparison of samples. While Figure 6.1 visualizes the overall distribution for the dataset, mean squared error is computed taking into account each sample utterance in the dataset.

decomposition component	phrase-rate	word-rate	syllable-rate
4th	0.371		
5th		1.255	
6th		0.566	
7th			1.542
3rd-4th	0.263		
5th-6th		0.714	
7th-8th			2.53

Table 6.1: Root mean squared error for selected decomposition components and respective linguistic levels. The top rows measure on single components given a 10-scale decomposition and the middle rows measure on the sum of adjacent scales, as used in Chapters 4 and 5 of this thesis.

rather than the sum of 5th and 6th. The syllable rate appears to include the largest error with the sum of the 7th and 8th component.

Based on the informal observation from the distributions in Figure 6.1 and the approximation errors of Table 6.1, we find evidence to support the hypothesis that components of the static wavelet-based decomposition strategy is not meaningfully associated with linguistic levels. This would have no impact for frame-levels models, such as the CWT-HMM proposed by Suni et al. (2013), but it could affect suprasegmental representations such as the one proposed in Chapter 4.

6.2.2 Dynamic strategy

An alternative approach to the decomposition strategy investigated in the previous section varies the scale of the mother wavelet such that it matches the rate of a given linguistic unit within utterances. This would constrain the CWT to generate a representation that contains approximately one local maximum per linguistic unit. Because this decomposition strategy is not limited to a fixed set of scale values for the mother wavelet, we term this approach a *dynamic strategy*.

We propose a decomposition using four distinct linguistic levels: syllable, word, clitic-group, and phrase. These linguistic levels are chosen based on segmentation available from a typical text-to-speech front-end. In this work, we make use of **syllable** and **word** boundaries.

Additionally, we consider the **phrase**, which is inferred similarly to those described in Chapter 4. Phrase boundaries (or breaks) are predicted from text using models trained on smaller annotated datasets. This corresponds to the traditional phrase break prediction found in most text-to-speech systems.[3] Conceptually, these phrases may correspond to the intonational or phonological phrases of Nespor and Vogel (1986), following the description of Shattuck-Hufnagel and Turk (1996) (see Section 3.1 and Figure 3.1 for details). Note, however, that there is no attempt to relate these units with theories of prosodic constituency. The linguistic levels we use in this work are simpler and inferred from the text.

The rate for words, syllables, and phrases is easily derivable given an utterance-level alignment of speech with text. However, we lack textual annotation for a linguistic level between the word and the phrase. Such an intermediate level might be useful, since it could capture the variation associated with the fifth component of Figure 6.1. For this reason, we define the rate of this intermediate level to be the average of the word and phrase rates, and we call this level the **clitic-group**.

Note that, in this work, the clitic-group is a conceptual unit capturing *f0* variation between word and phrase units. It does not necessarily correspond to clearly defined textual units, and it is not the purpose of this work to find such a relation. We freely adopt the terminology of Nespor and Vogel (1986), but make no attempt to relate this component with theories of prosodic constituency. The clitic-group of Nespor and Vogel (1986) is described as a content word with an optional adjacent monosyllabic function word (Shattuck-Hufnagel and Turk, 1996). Observing the *f0* component associated with this level in the proposed dynamic decomposition strategy, illustrated in Figure 6.3, it mostly appears to capture variation related to content words, which could potentially be grouped with adjacent function words. This observation alone motivates the usage of this

[3] For analysis and acoustic model training purposes, ideally we would consider acoustically motivated phrase breaks. This was not implemented throughout the work described in this chapter, but it is corrected in the database used in Chapters 9 and 10 of this thesis.

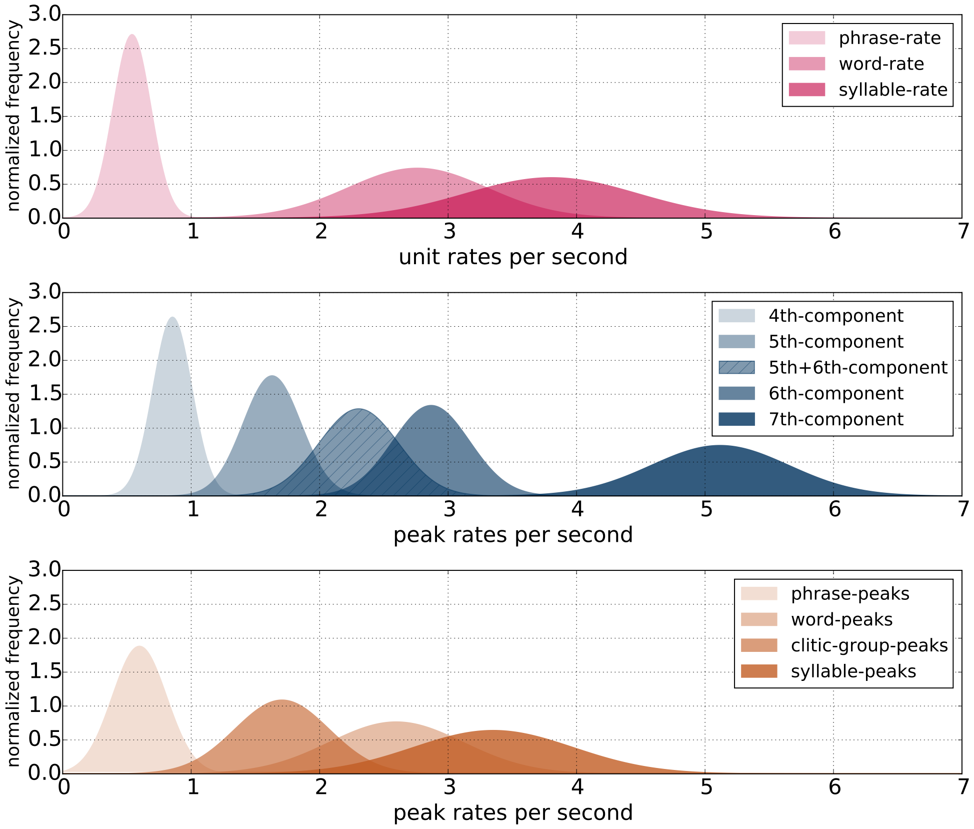

Figure 6.2: Top figure shows unit rate distributions for selected linguistic units. Middle figure shows peak (local maxima) rates per second for selected components under a static 10-scale wavelet-based decomposition of the *f0* signal. Bottom figure shows the peak rates for the proposed dynamic wavelet-based decomposition of the *f0* signal.

terminology.

In the proposed dynamic decomposition strategy, we first compute the unit rate at each linguistic level for each utterance in the database. We then set the wavelet scale a according to:

$$a = \frac{1}{\lambda f}, \text{ where } \lambda = \frac{2\pi}{\sqrt{m + 0.5}} \tag{6.1}$$

a is the wavelet scale, according to equation 3.5 (p. 62), f is the frequency, which is set to the unit rate of each level, and λ is the Fourier wavelength (Torrence and Compo, 1998), where m is set to 2 for the Mexican hat wavelet.

decomposition component	phrase-rate	clitic-group-rate	word-rate	syllable-rate
phrase	0.145			
clitic-group		0.233		
word			0.327	
syllable				0.565

Table 6.2: Root mean squared error for selected decomposition components and respective linguistic levels components from a dynamic decomposition of the *f0* signal.

6.2.3 Analysis

Figure 6.2 compares the proposed dynamic strategy with the static strategy described in the previous section. The figure shows that the proposed method is a better approximation to the observed distribution of linguistic units than the 10-scale decomposition. The clitic-group was included in order to capture the range given by the 6th component, which was judged to capture relevant long-term variation (cf. Chapter 5). Note also that the proposed dynamic decomposition captures the variation associated with components 4 to 6, covering the range of 0.6–3.35 Hz. This falls well within the range that speakers have associated with naturalness (1.6–3.2Hz), according to the results presented in Chapter 5.

Table 6.2 quantifies the distance between the proposed decomposition strategy and the unit rates of the database. Overall, all levels of a dynamic decomposition are strong predictors of the rates within the utterance. Even though we are using observed unit rates, there will always be a non-zero error since in general some units might not carry an accent or prominence.

Figure 6.3 illustrates the dynamic decomposition strategy on an utterance from the database. This figure can be compared with a static decomposition strategy. Figure 4.2 (p. 77) shows a 10-scale static decomposition with adjacent scales added together. This was the approach used for the multi-level representation described on Chapter 4 and the investigation on Chapter 5. For comparison purposes, the same utterance is illustrated.

The curve that had been previously associated with the word level is now associated with the clitic group, and it corresponds, for this particular example, to the content words of the utterance. In the new decomposition strategy, the

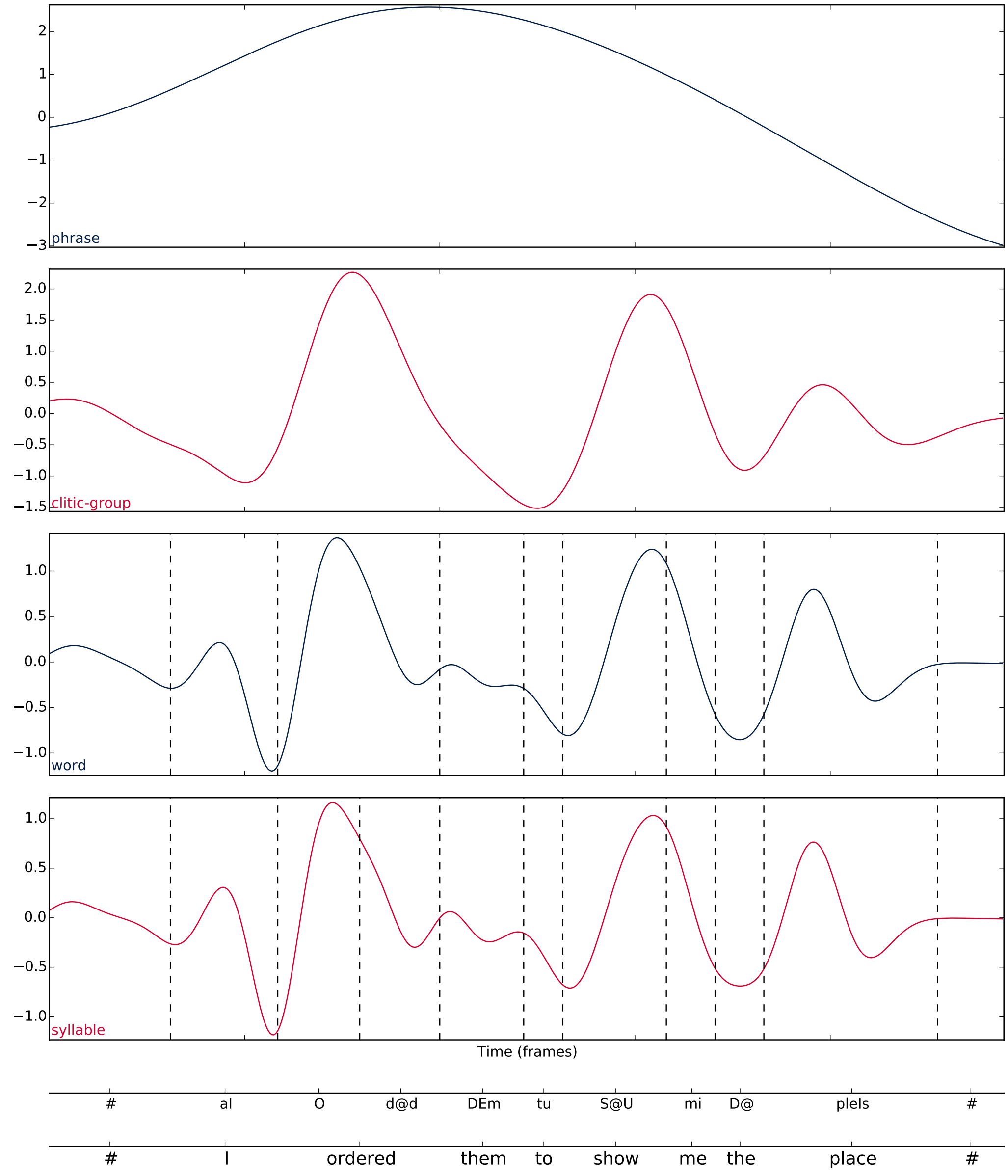

Figure 6.3: Four scale dynamic CWT decomposition strategy with force-aligned boundaries (vertical dashed lines) at each linguistic level. The bottom axes list the syllables and the words for this example, with the axis marks placed at the mid-point of each syllable or word, considering the force-aligned boundaries. This decomposition provides an alternative to the static decomposition strategy illustrated in Figure 4.2 (p. 77).

		all	**5-6**	**5-8**	**4-7**	**dynamic**
rmse	mean	2.588	21.134	11.553	16.647	11.303
	std	1.01	10.078	5.548	8.799	4.847
corr	mean	0.995	0.704	0.904	0.815	0.901
	std	0.001	0.115	0.047	0.091	0.057

Table 6.3: Reconstruction error for selected 10-scale decomposition components and the proposed dynamic decomposition.

curves associated with words and syllables are similar due to the example consisting mostly of monosyllabic words. Looking at the two figures, the dynamic decomposition appears to extract a more stable contour in terms of syllables, while discarding the high frequency information previously associated with the phone-level.

Table 6.3 shows reconstruction error for the proposed dynamic decomposition, as well as for selected wavelet scales given the 10-scale decomposition. A dynamic decomposition method such as the one proposed will not achieve a good reconstruction error, as some components are discarded (e.g. the very low or high frequencies). However, we do observe that it is comparable to the ranges 5-8, which was one of the systems associated with higher naturalness levels, according to the evaluations conducted in Chapter 5. Although there is some reconstruction loss, *f0* could be modeled directly with this representation if supplemented by an additional fifth component. This additional component could be set as the residual of the reconstructed signal not accounted for by the observed *f0* signal.

The main advantage of this representation over the static strategy is that it is *linguistically derived*, as well as *perceptually motivated*. We claim that this dynamic decomposition strategy is linguistically derived because it is based on observed unit rates from the data and the peak rate distributions can be associated with suprasegmental linguistic levels. We also claim that it is perceptually-motivated because it is based on the frequency ranges judged to contain higher levels of naturalness in the set of perceptual experiments described in Chapter 5.

In order to compare these two decomposition strategies, we consider them to be the secondary task in a feedforward neural network. The following sections compare the two approaches as additional output features in the optimization of an acoustic model for text-to-speech synthesis.

6.3 Experiments

6.3.1 Multi-task learning

The main idea behind multi-task learning (MTL, Caruana (1997)) is to train a model on similar tasks using the same shared representation of input features. We provide the model with a secondary task, which could help produce better latent representations, thus improving performance on the primary task, illustrated in Figure 6.4. Multi-task learning has been applied to automatic speech recognition (Seltzer and Droppo, 2013) and to natural language processing (Collobert and Weston, 2008) with various degrees of success.

In speech synthesis, recent work has explored secondary tasks mostly in the spectral domain. In Wu et al. (2015), gammatone spectrum, formant frequencies, line spectral frequencies (LSF), or spectro-temporal excitation patterns (STEP) were used. Although improvements were seen in objective measures, the authors failed to see significant differences between these systems and the baseline in a perceptual evaluation.

In the following sections, we investigate the usefulness of the wavelet-based decomposition strategies described in Section 6.2 as an *f0-based* secondary task.

We hypothesize that a different representation of the *f0* signal will be useful for the optimization of the model on the standard acoustic feature set, especially when the representation is perceptually motivated, as in the case of the mid-range frequencies of the static and dynamic strategies. We also hypothesize that a balance between spectral features and *f0-based* features will be relevant. The standard feature set uses 180 spectral features, 75 band aperiodicities, but only 3 *f0* features, considering deltas and delta-deltas.

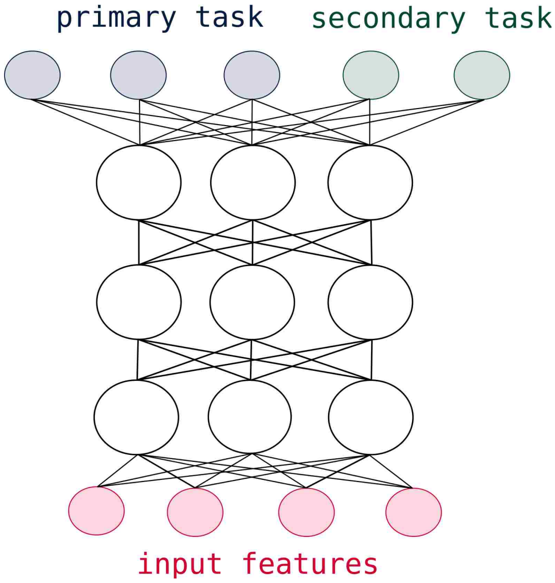

Figure 6.4: Multi-task deep neural network (MTL-DNN). A secondary task is added alongside the primary task during training. At synthesis time, the secondary task is discarded.

6.3.2 Experimental setup

We have used the freely available audiobook *A Tramp Abroad*, processed as described in Section 4.3.1. Acoustic parameters were extracted using the STRAIGHT vocoder (Kawahara et al., 1999, 2001). These were log-*f0*, 60-dimensional mel cepstral coefficients (MCCs), and 25 band aperiodicities (BAPs), extracted at 5ms intervals. Log-*f0* was linearly interpolated and voiced/unvoiced decision (VUV) was stored separately. We further appended dynamic features (deltas and delta-deltas), thus creating a 180 dimensional vector for MCCs, a 3 dimensional vector for log-*f0*, and a 75 dimensional vector for BAPs. With the voiced/unvoiced decision, the full output acoustic feature vector consists of 259 values. We call this the *primary task*.

We further processed the interpolated log-*f0* signal with the CWT, using the two decomposition strategies described in Section 6.2. The various components of these decomposition strategies and their dynamic features are called the *secondary task*. The unit rates used with the proposed dynamic decomposition of *f0* were inferred from syllable, word, and phrase boundaries. These were extracted from the data after forced-alignment. For this alignment, context-dependent 5-state left-to-right hidden Markov models were used with monophone pretraining. The models learned on the training data were then used to force-align the training, validation, and test sets at the state-level.

As input features, we used the set of 592 binary questions at phone and higher-levels plus 9 numerical features related to the state and frame position. The full input feature vector consists of 601 values. Details on the feature set can be found in Appendix A.0.2. Input features were normalized to the range [0.01, 0.99] and output features were normalized to zero mean and unit variance. For these experiments, durations was inferred from the forced-alignment.

The acoustic model for this task is a feedforward deep neural network. We use *tanh* as the activation function in the hidden layers and a linear activation function in the output layer. A total of six layers were used, each with 1024 nodes. For training, we set the mini-batch size to 256. Momentum was set to 0.3 for the first 10 epochs with a learning rate of 0.002. After the first 10 epochs, momentum was set to 0.9 and the learning rate reduced by 50% after each epoch. Each system trained for a maximum number of 25 epochs. Early stopping was used and training and halted once validation error started increasing. L2 regularization weight was set to 10^{-5}. Training parameters and implementation are very similar to those described in Wu et al. (2015). We have used an pre-release version of the Merlin Neural Network Toolkit (Wu et al., 2016). Training, development, and test sets consist of 4500, 300, and 100 utterances, respectively. The training set consists of 9 hours of speech data, while validation contains 33 minutes, and test set 9 minutes. In these experiments, we keep input features, data, and architecture constant. The primary task is the same for all systems and only the secondary task is varied.

6.3.3 Systems trained

We trained a total of 16 systems, which are shown in Table 6.4. They are differentiated only by the secondary task that they use. The system using no secondary task is taken as a baseline.

The first block of systems uses the static 10-scale decomposition. For example, *cwt-5* indicates that the signal corresponding to the fifth component was used as a secondary task. The system *cwt-5, cwt-6* indicates that the fifth and the sixth components were included as two secondary tasks. This specific range was selected as the experiments described in Chapter 5 associated them with higher levels of naturalness. The second block of systems uses the proposed four-level dynamic decomposition. Selected levels of this decomposition strategy are used as the secondary task. When more than one component is included, more than one secondary task was used simultaneously.

The main hypothesis we test is that including *f0* components capturing suprasegmental prosodic variation as secondary tasks will improve the overall quality of speech parameters generated by a system trained on an expressive dataset. We expect that the distribution of the quality improvements seen with each component to be similar to the distribution of their naturalness ratings. That is, components (or ranges) that were judged more natural in Chapter 5 will give better results if used as secondary tasks when compared with a system not using additional tasks.

6.3.4 Objective results

Objective results for all trained systems are shown in Table 6.4. The results presented for all systems are computed only on the primary task. At this point, the output for the secondary task is discarded and no attempt was made to integrate it in the *f0* signal predicted from the primary task, although future work in this direction could be useful.

We observe that including all decomposition components does not improve the results over the baseline. In fact, noticeable decreases are seen, especially in terms of *f0* prediction. Similarly, lower frequency components, such as the phrase component, do not show improvements. This is not surprising, as these components reflect the longer-term variation of the *f0* signal and may not be useful for the short-term variation these frame-level models attempt to describe. The *cwt-5-6* condition, which uses the sum of components 5 and 6 of a 10-

Secondary acoustic features	**MCD (dB)**	**BAP (dB)**	**F0 RMSE (Hz)**	**F0 Corr**	**V/UV Error Rate (% of frames)**
none	4.64	2.18	27.68	0.44	4.42
cwt-1 to cwt-10	4.65	2.20	28.82	0.40	4.53
cwt-5	4.48	2.15	27.31	0.46	4.05
cwt-6	4.48	2.15	27.38	0.48	4.05
cwt-5, cwt-6	4.48	2.16	27.28	0.47	4.07
cwt-5-6	**4.46**	**2.15**	**26.96**	**0.49**	**3.40**
cwt-syl, cwt-wrd, cwt-clg, cwt-phr	4.64	2.20	28.69	0.43	4.48
cwt-syl	**4.47**	**2.15**	27.14	**0.48**	**4.01**
cwt-wrd	4.48	2.15	27.41	0.46	4.07
cwt-clg	4.48	2.15	**26.90**	0.47	4.12
cwt-phr	4.64	2.18	28.07	0.44	4.50
cwt-syl, cwt-wrd	4.66	2.19	28.14	0.44	4.59
cwt-wrd, cwt-clg	4.50	2.16	27.50	0.46	4.09
cwt-clg, cwt-phr	4.67	2.19	28.67	0.42	4.66

Table 6.4: Objective results for trained systems. All systems include MCCs, log *f0*, VUV, and BAPs as primary acoustic features. Secondary acoustic features are added as per the proposed decomposition, using either a dynamic or a 10-scale decomposition. MCD is mel cepstral distortion, BAP is band aperiodicity error, V/UV is voiced/unvoiced error, and RMSE and Corr are the root-mean-squared error and correlation between predicted and original *f0* signal on voiced frames only. Best performing results for each parameter are indicated in bold.

scale static decomposition, outperforms all other systems. This is also not a surprise, as this is the condition judged as most natural by participants in the experiments reported in Chapter 5. The disadvantage of this component is that it is not directly associated with a linguistic unit, unlike the proposed dynamic decomposition.

Quite interestingly, the condition including the syllable and word-level components together as secondary task performs worse than the remaining systems, being equivalent to the lower frequency components. The reason for this might be the large overlap in the syllable and word distributions seen in Figure 6.2,

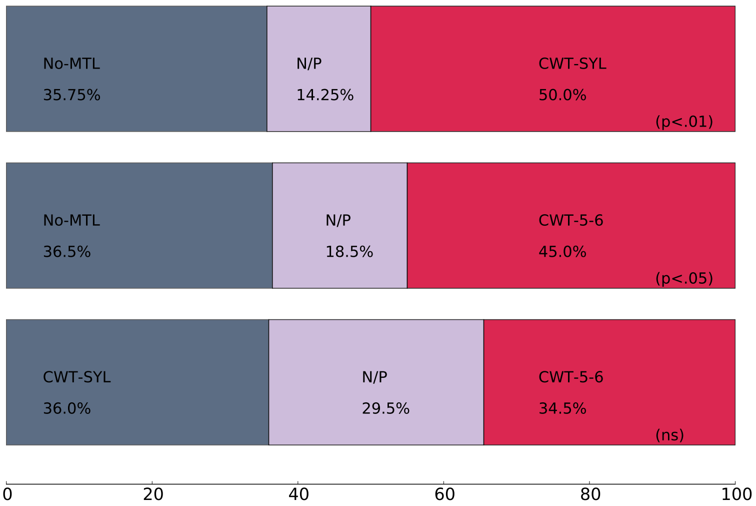

Figure 6.5: Preference test results with N/P indicating "No Preference". In parenthesis, p-values indicate the results of 1-tailed binomial tests with an expected 50% split, with the N/P results evenly distributed over the other two conditions.

which makes these two components highly correlated. It was expected that the clitic-group or the word components would outperform all other systems, as these are approximately in the frequency range judged to contribute most towards naturalness. Instead, we observe that the syllable component yields the best objective measures. Further experiments could investigate how these lower-frequency components (word and clitic-group) behave under models capable of leveraging long-term information, such as recurrent networks with Long Short-Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997; Fernandez et al., 2014).

6.3.5 Subjective results

We conducted a perceptual evaluation of 3 selected systems. We chose the system from each decomposition strategy with the highest *f0* correlation and the baseline system for inclusion in the evaluation. 50 test utterances were synthesized from the primary parameters, and the secondary parameters were discarded. 16 native speakers judged randomized utterance pairs in a preference test with a *no preference* option. Each pair was judged 8 times by different participants and each condition received a total of 400 judgments.

Results are presented in Figure 6.5, where we see preference percentages and the results of a 1-tailed binomial test assuming an expected 50% split, with the no-preference judgments distributed equally over the remaining conditions. The two proposed systems are preferred over the baseline, but no significant differences are seen when they are compared against each other. It was surprising to see a much smaller effect when comparing the 10-scale system with the baseline, as it was expected to achieve higher naturalness.

6.4 Discussion

In Chapter 4 of this thesis, the *f0* signal was decomposed with the continuous wavelet transform (CWT). The components extracted from the signal were then modeled at various linguistic levels within an HMM-based text-to-speech system. However, for modeling, the proposed representation made two fundamental assumptions.

The first assumes that *all decomposition components were equally relevant to the reconstructed signal.* This was investigated in Chapter 5. The results of a set of perceptual evaluations found evidence to support that this assumption does not hold. Some decomposition components can be removed without much loss of the naturalness of the signal, while specific components (those in the mid-frequency ranges) carry most of the signal's information.

The second assumption, that *each decomposition component can be meaningfully associated with a linguistic level*, was investigated in the current chapter. By relating unit rates (e.g. syllables, words) with peak rates (local maxima) of signal components, we have found that the second assumption also does not hold. It was observed that, in the case of phrase distributions, there is an approximate match between unit rate and the peak rate distributions. However, in the case of syllables and words, the unit and peak rate distributions were not very similar. These two linguistic levels were of particular importance as the signal components that were being associated with them were those that listeners associated with higher naturalness.

An alternative decomposition strategy was proposed, which adapts the scale of the mother wavelet to match the unit rates within the utterance. Objective results indicate that including multiple decomposition components as secondary tasks is not beneficial for the primary task. Similar results were observed with low frequency components, such as those associated with the phrase level. A decrease in the objective error measurements was observed when using those components that are in the same range that listeners have associated with naturalness. The distribution of the objective scores is consistent with earlier findings related to the components that were used as secondary tasks. The systems using components that were given higher naturalness ratings (those in the mid-frequency ranges) consistently scored higher when used as secondary tasks. In the dynamic decomposition strategy, these are those that have been extracted with syllable, word, or clitic-group unit rates.

Listening tests have provided further evidence that these components can be beneficial for the modeling of acoustic parameters for text-to-speech systems. Both systems using multi-task learning outperformed the system using only the primary task. And although no preference was observed towards the system using a dynamic decomposition component when compared to a system using a static decomposition component, the proposed strategy has the advantage of being related to linguistic levels.

The proposed dynamic wavelet-based decomposition has two key advantages over a static strategy. It is both perceptually-motivated and linguistically-motivated. We claim that the dynamic decomposition strategy is perceptually-motivated because its key components (syllable, word, clitic-group) are within the frequency range (1.6–3.2 Hz) that listeners have associated with naturalness. Therefore, we know that the fundamental intonational patterns of the *f0* signal are being preserved in the decomposition components. We also claim that this strategy is linguistically-motivated because the mother wavelet is dynamically scaled to match the speech rate of the utterance. This conditions the transform to find approximately one local maximum per unit in the utterance, which could correspond to the prominence level of that unit within the utterance.

Further work could investigate direct modeling of the decomposed *f0* signal by incorporating a fifth component made of the residual of the reconstructed and the original signal. This would be a good step towards a hierarchical and suprasegmental model of the *f0* signal. Future work could also attempt to combine the prediction of the secondary task with the predicted *f0* signal, instead of discarding it. The components at each level may also be used to learn better feature representations at each linguistic level, as they are assumed to capture each level's variation. Some work in this direction is presented in Chapter 9.

6.5 Representations of acoustic signals

The current chapter concludes the evaluation of decomposition approaches to the modeling of the *f0* signal in statistical parametric speech synthesis. Considering the segmentation of the main claim into three sub-problems, this work falls under *sub-problem 1: representations of acoustic signals.* Although we are aware of a wide variety of representation approaches for the *f0* signal (see Section 3.2.1), we have chosen to focus on a wavelet-based decomposition approach. The main outcome of this work is the dynamic decomposition strategy of the *f0* signal, whose components can be meaningfully associated with linguistic levels.

In the following chapters of this thesis, we focus on the analysis of the remaining two sub-problems of the main claim: *representation of linguistic contexts* (Chapters 7 and 9) and *mapping acoustic and context representations* (Chapters 7 and 8). The main contribution of Chapters 4, 5, and 6, focusing on suprasegmental acoustic representations, will be revisited in Chapter 10, together with the main contributions presented in later chapters of this thesis.

6.6 Conclusion

This chapter investigated the association of *f0* components with linguistic levels under two wavelet-based decomposition strategies. While the first strategy (Suni et al., 2013; Ribeiro and Clark, 2015) uses a static set of scales, the second strategy dynamically adapts to the speech rate of an utterance. It was found that the dynamic strategy has the ability to meaningfully relate to linguistic units such as syllables and words. These approaches were evaluated as secondary tasks in multi-task DNNs for expressive speech. We have observed a strong preference for the systems using multi-task learning with selected decomposition components.

Chapter 7

Syllable-level representations of suprasegmental features

This chapter presents an extended version of the work described in "Syllable-level representations of suprasegmental features for DNN-based text-to-speech synthesis" (Ribeiro et al., 2016b), presented at Interspeech 2016.

A top-down hierarchical system based on deep neural networks is investigated. Suprasegmental features are processed separately from segmental features and a compact distributed representation of high-level units is learned at the syllable level. The suprasegmental representation is then integrated into a frame-level network. Objective measures show that a hierarchical framework outperforms a standard frame-level network. Additional features incorporated into the hierarchical system are then tested. At the syllable level, a bag-of-phones representation is proposed and, at the word-level, vector representations are learned from text sources are used. It is shown that the hierarchical system is able to leverage new features at higher-levels more efficiently than a system which exploits them directly at the frame level. A perceptual evaluation of the proposed systems is conducted and followed by a discussion of the results.

7.1 Introduction

Considering the main claim of this thesis and its three sub-problems defined in Chapter 1, the work presented in previous chapters provided contributions and insights into **multi-level representations of fundamental frequency**. These chapters were associated with sub-problem 1: *representations of acoustic signals*. The current chapter focuses on different sub-problems of the main thesis claim. Together with Chapters 8 and 9, the common topic of these contributions is **hierarchical systems and suprasegmental input representations**. These fall under sub-problem 2: *representation of linguistic contexts*, and sub-problem 3: *mapping acoustic and context representations*.

In terms of modeling, multi-level approaches have been proposed for HMM-based systems (Hsia et al., 2010; Qian et al., 2011). In the case of DNN-based systems, recurrent (Fernandez et al., 2014), hierarchical (Yin et al., 2016), or mixed (Chen et al., 1998) approaches have claimed to capture the long-term dependencies of speech.

On the input side, it has been shown that prosodic contexts are very poorly understood. Cernak et al. (2013) revealed that, in HMM-based synthesis, features above the syllable-level do not improve the naturalness of synthetic speech (see Section 3.2.3 for a discussion). In an effort to acquire a better understanding of linguistic contexts, continuous representations of input features have been explored, either at the segment (Lu et al., 2013; Wu et al., 2015) or word level (Watts et al., 2014; Wang et al., 2015a), with various degrees of success.

This investigation adds to the work exploring input features for prosody modeling in text-to-speech synthesis, specifically focusing on continuous representations of suprasegmental contexts. Two main contributions are made:

1. a top-down hierarchical model that pre-processes suprasegmental information and represents it compactly at the syllable level.
2. an investigation of this architecture with a bag-of-phones representation at the syllable level and word embeddings learned from a large text database with the skip-gram model (Mikolov et al., 2013a,b).

There are two possible lines of research in terms of suprasegmental information. A first approach may focus solely on learning transformations of conventional linguistic features (e.g. via a neural network) in an attempt to improve the prediction of the acoustic parameters without adding any extra information. A second line of research attempts to learn additional features and provides the model with new information regarding suprasegmental context. Both approaches are handled by the work presented in the current chapter.

In Section 7.3, an investigation of a cascaded hierarchical framework is conducted. In this architecture, multiple linguistic levels are connected via a *bottleneck layer*, which learns a compact latent representation of suprasegmental features. Since no additional information is given to the model, this work falls under the first line of research. We investigate three variables of interest: position of the bottleneck layer, size of the bottleneck layer, and acoustic features used to optimize the suprasegmental network. In Section 7.4, we evaluate the ability of a hierarchical framework to leverage additional features. This work falls under the second line of research, as new information is injected into the model. This new information consists of a bag-of-phones representation of syllable structure, and, at the word level, text-derived word embeddings learned with the skip-gram model.

7.2 Experimental setup

For these experiments, we have used freely available audiobook data. This is the same dataset described in Section 6.3.2. Training, development, and test sets consist of 4500, 300, and 100 utterances, respectively. The training set consists of 9 hours of speech data, the validation 33 minutes, and the test set 9 minutes.

As in earlier work, we have used the STRAIGHT vocoder (Kawahara et al., 1999, 2001) to extract log-*f0*, 60-dimensional mel cepstral coefficients (MCCs), and 25 band aperiodicities (BAPs) at 5ms intervals. Log-*f0* was linearly interpolated and voiced/unvoiced decision (VUV) was stored separately. Dynamic features (delta and delta-deltas) are added to the static features to form an acoustic feature vector with 259 dimensions. Acoustic features were reduced to zero mean and unit variance. For these experiments, we use natural duration, which is inferred from the force-alignment given by a pre-trained 5-state left-to-right HMM.

As input features, we use a set of 592 binary questions at phone and higher-levels plus 2 numerical features related to the HMM state. These are the linguistic features described in Appendix A.0.2 (p. 207). Input features were normalized to the range [0.01, 0.99]. A subset of the input features are defined at syllable and word-level (and are here termed *suprasegmental features*). These are separated from the remaining features defined at frame and phone-level (and are termed *segmental features*).

Basic network

The basic deep neural network is a simple feedforward multilayer perceptron. We use a configuration identical to the baseline system described in Chapter 6. A network with 6 hidden layers is used, each layer consisting of 1024 nodes. We set *tanh* as the activation function in the hidden layers and we use a linear output layer. During training, we use a mini-batch size of 256 and we set the maximum number of iterations to 25. The network inputs the normalized binary vectors from the concatenated segmental and suprasegmental features.

Top-down hierarchical network

Figure 7.1 illustrates the top-down hierarchical deep neural network. An initial network takes the feature vector corresponding to the normalized binary suprasegmental features and maps it to acoustic parameters defined at the syllable level. The network is set to be a 6 hidden layer triangular network. The lower layers begin with 1024 nodes and this is halved in the next layer such that the top hidden layer reaches the desired dimensionality. We term this top hidden layer the *bottleneck layer*. Further details are given in Section 7.3. The syllable network uses the *tanh* activation function in the hidden layers and a linear output layer. Mini-batch size is set to 16 and the maximum number of iterations is set to 25. The remaining hyperparameters are identical to those used for the basic deep

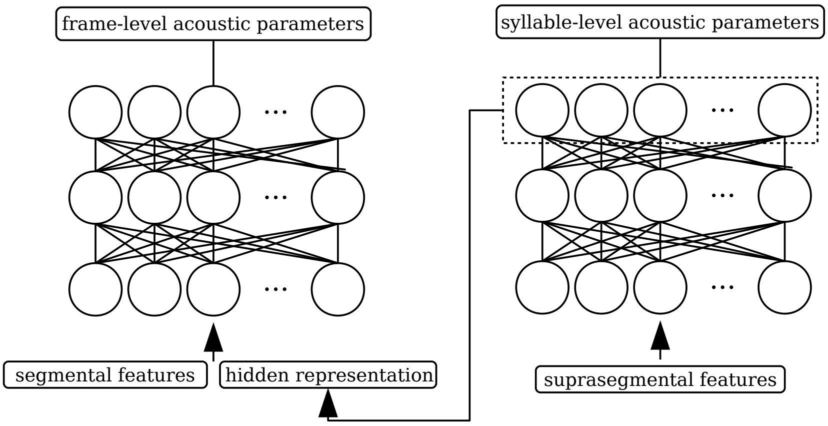
frame-level acoustic parameters
syllable-level acoustic parameters
segmental features
hidden representation
suprasegmental features

Figure 7.1: Integration of suprasegmental bottleneck features into a frame-level deep neural network

neural network.

After the suprasegmental network is trained, the hidden representation of the bottleneck layer is concatenated with the segmental feature vector. The frame-level network is then trained as described in the previous section. Segmental and suprasegmental networks are optimized separately.

7.3 Hierarchical framework

In this section, we explore a variety of configurations with the proposed cascaded hierarchical structure. We begin in Section 7.3.1 by investigating the position of the bottleneck layer within the syllable-level network. Section 7.3.2 investigates how to include the learned representations in the frame-level network, while Section 7.3.3 explores combinations of acoustic features in the output parameters of the suprasegmental network.

7.3.1 Position of the bottleneck layer

We evaluate two strategies for integrating suprasegmental hidden representations into a frame-level network and the position of the bottleneck layer in the suprasegmental network. Regarding integration methodologies, we consider replacing or appending to the original binary set of suprasegmental features. Deviating from

System	Bottleneck Size	MCD	BAP	F0-RMSE	F0-CORR	VUV
baseline	-	4.596	2.197	28.054	0.449	5.19
append - higher layer	128	4.603	2.199	27.625	0.441	5.243
replace - higher layer	128	4.568	**2.182**	**27.312**	**0.462**	5.202
replace - middle layer	128	**4.565**	2.184	27.482	0.441	5.058
replace - lower layer	128	4.580	2.188	27.875	0.455	5.147

Table 7.1: Objective measures for systems varying the position of the bottleneck layer in the suprasegmental network. MCD is mel cepstral distortion, BAP is band aperiodicity error, F0-RMSE and F0-Corr are the root-mean-squared error and correlation between the predicted and original f0 signal on voiced frames, and VUV is voicing error.

Figure 7.1, the *append* system concatenates the learned hidden representation with the full feature set containing segmental and suprasegmental features. The *replace* systems instead removes the suprasegmental features from the frame-level network and use only the hidden representation alongside the segmental features. This follows the diagram illustrated in Figure 7.1.

We further vary the position of the bottleneck layer in the suprasegmental network. *Lower layer* indicates that the first hidden layer (closer to the linguistic features) has been set as the bottleneck. *Higher layer* indicates that the last hidden layer (closer to acoustic features) has been used as the bottleneck. The *middle layer* network places the bottleneck in the third hidden layer, approximately halfway between linguistic and acoustic features, thus diverging from the described triangular architecture.

Table 7.1 shows the results from the trained systems, as well as the baseline, the system with the full set of input features at the frame level. We observe that appending the hidden representations to the full set of features does not change performance, although there are very small improvements in terms of *f0* RMSE. This could be due to the fact that binary and distributed representations contain similar information. However, if we remove the binary suprasegmental features and use only the bottleneck features from the syllable-level DNNs, we

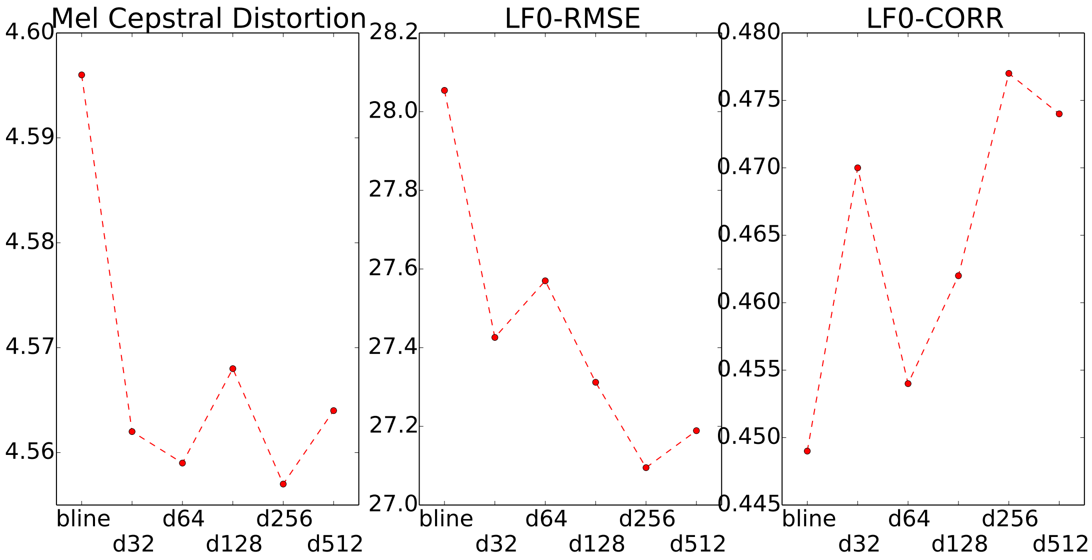

Figure 7.2: Objective measures for mel-cepstral coefficients and *log-f0* as a function of bottleneck size. *bline* represents the baseline system, while *dn* indicates a bottleneck layer of dimensionality *n*.

notice some improvements across all acoustic parameters. This instead suggests that the binary suprasegmental features may actually be noisy and damage the performance of the system. In this case, one could think of the pre-processing network as a denoising process, where we learn a compact distributed representation of suprasegmental context. This hypothesis will be evaluated more carefully in Chapter 8.

In terms of the position of the bottleneck layer, we observe that all positions appear to achieve similar results. But performance increases slightly as we move closer to the acoustic features. This follows the results of the frame-level experiments conducted in Wu and King (2016), which observed that best performance with bottleneck features occurs when the bottleneck layer is placed in the top half of the network.

7.3.2 Size of the bottleneck layer

In this set of experiments, we observe the effect of the bottleneck layer size in the objective measures. We use the framework illustrated in Figure 7.1, where we replace the suprasegmental features with the learned hidden representations. All experiments use the syllable mean for the acoustic features and the top layer as bottleneck layer.

Bottleneck size is varied in powers of 2, from 32 to 512. All syllable-level systems are triangular networks with 6 hidden layers. The lower layer has 1024 nodes and we either maintain that size or we reduce it in half until we reach the desired dimensionality in the top hidden layer. As an example, the triangular network for a bottleneck layer of size 256 would have the following structure, in terms of layer size: [1024, 1024, 1024, 1024, 512, 256]. Similarly, for a bottleneck layer of size 128, we would have [1024, 1024, 1024, 512, 256, 128].

Figure 7.2 shows layer size effect on mel-cepstral distortion (MCD), *log-f0* RMSE and correlation. All systems using hidden representations outperform the baseline, with the best results occurring with a dimensionality of 256. These results make sense, as the segmental features use a vector of 350 dimensions. A vector of 256 dimensions for suprasegmental features balances the two types of features and allows the best prediction for all acoustic parameters.

7.3.3 Suprasegmental acoustic features

In the following experiments, we observe the effect of the acoustic features that are used as output for the syllable-level deep neural network. Mel-cepstral coefficients have a considerably larger number of features than *f0*, which will make them have a much larger impact on the model's error during training. By excluding selected groups of features, we expect to determine which features have the biggest impact on the suprasegmental bottleneck representations. Intuitively, we would expect *f0*-based acoustic features to have a stronger impact on the learned representations, as *f0* is associated with intonation and suprasegmental variation.

All systems in this set of experiments use the bottleneck features as the only suprasegmental information, as shown in Figure 7.1. The syllable-level system consists of a triangular network with 6 hidden layers and a bottleneck size of 256 nodes. The top hidden layer is used as the embedded representation of suprasegmental context.

System	MCD	BAP	F0-RMSE	F0-CORR	VUV
average on full syllable [all]	**4.557**	2.176	**27.095**	**0.477**	5.012
average on syllable nucleus [all]	4.573	2.181	27.347	0.465	5.185
average on full syllable [mgc]	4.558	**2.173**	27.286	0.460	5.026
average on full syllable [lf0]	4.575	2.188	27.280	0.459	5.095
average on full syllable [bap]	4.566	2.183	27.431	0.464	5.189
average on full syllable [mgc, bap]	4.561	2.176	27.353	0.452	**5.004**
average on full syllable [mgc, lf0]	4.563	2.177	27.406	0.461	5.095
average on full syllable [bap, lf0]	4.568	2.181	27.414	0.461	5.058

Table 7.2: Objective results for various combinations of acoustic parameters in the syllable-level deep neural network. The acoustic parameters used as output features are denoted in square brackets. Abbreviations are identical to those in table 7.1. For each objective measure, best performing results are highlighted in bold.

We vary only the syllable-level acoustic features. In the first set, we compare averaging over the full syllable and averaging only over the syllable nucleus. Averaging only over the syllable nucleus might provide more stable acoustic values over the syllable, which also motivates this experiment. In the remaining sets, we use the full syllable average and we explore various combinations of acoustic parameters. The results are detailed in Table 7.2.

Results indicate that averaging over the full syllable gives better results than averaging only over the syllable nucleus. And including all acoustic parameters gives better results than including selected combinations. However, we do observe that the variation within each of the measures is very small. This suggests that it is the learning of hidden representations that contributes to the small performance gains we observe, rather than the acoustic features we use.

We were, however, limited to averaging over the syllable for existing acoustic features. It is possible that an alternative configuration could be capable of learning more useful representations. For example, we could consider hidden representations learned using only prosodic parameters in the output features. These could range from various statistics of *f0* over the syllable or other type of representations. Various possibilities of other representations include transforms such as the Discrete Cosine Transform (DCT) or the Continuous Wavelet Transform (CWT). A simple representation of *f0* over the syllable using the DCT could be attempted (Teutenberg et al., 2008; Stan and Giurgiu, 2011), or more complex representations of *f0*, such as the dynamic multi-level representation proposed in Chapter 6.

7.3.4 Discussion

The experiments conducted in this section should be considered exploratory as they were not initially motivated by a clearly defined set of hypotheses. However, some interesting observations are inferred from the results, which may lead to further work.

It was observed in Section 7.3.1 that, when pre-processing suprasegmental features, *replacing the original feature set with the transformed set tends to give better results than combining original and transformed feature sets.* This observations suggests that the hierarchical framework may be operating as a denoiser or feature extractor for the suprasegmental binary representation. Although this hypothesis was not directly tested, some of the results indicate that this may be a possibility. Chapter 8 will evaluate that hypothesis in controlled conditions using various subsets of linguistic inputs and hierarchical frameworks. It as also observed that *placing the bottleneck layer closer to the acoustic features generally gives better results than placing it closer to the linguistic features.* This observation agrees with the findings described in Wu and King (2016).

In Section 7.3.2, we observe that a *balanced segmental and suprasegmental features gives best results.* This is an interesting finding as it suggests that blindly adding new features to a frame-level network may not be a good idea. We may argue that a good balance between the two types of features is needed for more accurate results. Note, however, that varying bottleneck layer size was performed as an optimization of the suprasegmental network. We did not directly test the hypothesis that a good balance between the two types of features generally gives better results. This observation may not be so clear, as results described in Chapter 9 indicate that dimensionality is not necessarily a problem if the representations are rich enough.

In Section 7.3.3, we observe that, when defining the suprasegmental acoustic feature vector, *averaging over the full syllable is better than averaging over the nucleus.* This was a surprising observation, as we would expect the acoustic parameters of the nucleus to be more stable and consistent across syllables. This was expected to be relevant especially when including mel-cepstral coefficients, as these features attempt to represent the position of the articulators, which is correlated with phone identity. Averaging over an entire syllable (rather than just the nucleus) was expected to give unstable acoustic representations that would lead to underperforming systems.

In this set of experiments, we have also observed that *averaging on all acoustic features is better than averaging on selected features.* This result was, to a certain extent, also surprising. We would expect *f0* to be the strongest feature, as it is often associated with prosodic and suprasegmental variation.

However, these observations should be taken carefully. These are based on objective measurements from the frame-level network, after integration of the suprasegmental representation. It is not measured directly on the suprasegmental network. These observations are also solely based on objective measures. It is not clear if the improvements observed are perceptually meaningful. In the following section, we provide a further investigation into the hierarchical framework by including additional features defined at syllable and word level.

7.4 Additional features

In the previous section, we investigated various combinations of acoustic parameters in the syllable-level DNN. Results seem to indicate that choice of acoustic features is less important than the learning process, although we were limited to combinations of syllable averages to optimize the suprasegmental network. In this section, we investigate the addition of new features to the frame-level and hierarchical networks. The hypothesis is that *the hierarchical network will be able to leverage the information given by the new features, while the frame-level network* *will depend mostly on segmental features and ignore the new high-dimensional representations.*

One of the main claims of the work presented in this chapter is that suprasegmental features may need to be processed differently in order to fully leverage their potential. If we choose to investigate new features, adding them to a frame-level DNN may not produce the desired results, as the system will mostly depend on phone-level information.

The proposed framework, illustrated in Figure 7.1, may be able to leverage this type of new information and may be ideal to test new high-dimensional representations of suprasegmental units that would otherwise be diluted alongside the frame-level features. We evaluate this claim with two sets of additional features, one defined at the syllable level and another defined at the word level. Section 7.4.1 uses a bag-of-phones representation of syllables, while Section 7.4.2 uses text-derived word embeddings.

7.4.1 Syllable bag-of-phones

We propose a bag-of-phones representation for syllables in the form of an *n-hot encoding*, a binary vector with n active bits. We use 3 bags of phones, each defined for the onset, nucleus, and coda, and containing phone identity and articulatory features. Taking the onset as an example, we define a binary vector where each component indicates either an articulatory feature or a phone identity. For all phones belonging to the onset of the current syllable, we activate the respective articulatory and identity component in the onset vector. The same applies to the nucleus and coda subvectors. This approach allows us to define a fixed-size representation of syllables that accounts for the variable number of phones in the onset and coda, while still including all of their defining features. An illustration of this representation is given in Figure 7.3.

We compare four systems:

frame frame-level deep neural network with all available segmental and suprasegmental features.

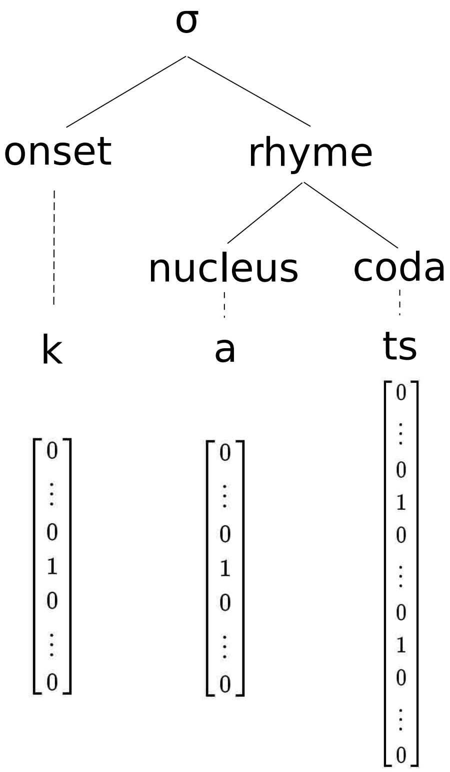

Figure 7.3: Syllable bag-of-phones illustration for the monosyllabic word *cats*. The syllable is segmented into onset and rhyme. The rhyme is then further divided into nucleus and coda. An *n-hot* encoding for phone identity and articulatory features is then created for the *n* phones. The concatenation of these three vectors is the *syllable bag-of-phones*.

frame-BoP baseline DNN with all features plus bag-of-phones representation for onset, nucleus, and coda of current syllable.

syl suprasegmental features are trained separately at the syllable level using a hidden layer with 256 nodes. Input features to syllable DNN consist of the suprasegmental features used with the baseline system. That is, this system does not use the bag-of-phones for onset, nucleus, and coda.

syl-BoP similar to *syl*, except we include the bag-of-phones features in the input to the syllable-level DNN. That is, we extend the suprasegmental features used with the baseline with the proposed bag-of-phones for onset, nucleus, and coda.

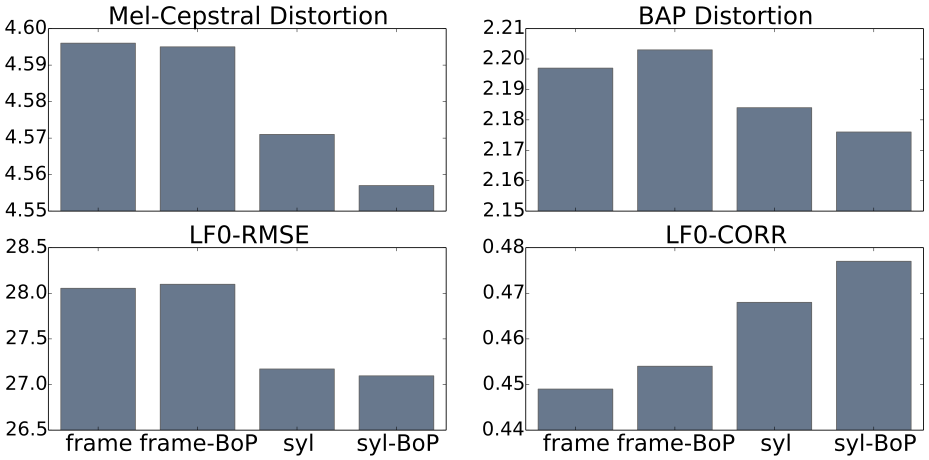

Figure 7.4: Objective measure for mel-cepstral distortion, band aperiodicity distortion, and *log-f0* RMSE and correlation testing the proposed *bag-of-phones* representation.

Figure 7.4 plots the results for mel-cepstral distortion, band aperiodicity distortion, *f0* RMSE, and *f0* correlation. Results indicate that adding the bag-of-phones representation to the frame-level DNN does not affect the results. This follows our initial hypothesis that simply appending more suprasegmental features to a frame-level DNN might limit the impact of those features.

The difference between *frame* and *syl* shows the changes we get by training suprasegmental representations separately. This causes the biggest improvements among the four systems. The difference between *syl* and *syl-BoP* measures the changes caused by the bag-of-phones features when training suprasegmental features separately. In the case of RMSE, we do not see many improvements, but we do notice better results in terms of correlation and MCD. It's interesting to observe that, in the BAP case, adding bag-of-phones features to the baseline slightly decreases performance, while adding the same features to a syllable-level DNN slightly increases it. This set of experiments shows that not only is the separate training of suprasegmental feature predictors useful, but it creates a framework that is able to leverage additional features in a more robust manner.

7.4.2 Text-derived word embeddings

We extend the previous investigation to incorporate word-level features in the hierarchical framework defined at syllable-level.[1] For this task, we use word embeddings, which are high-dimensional continuous vector representation of words. In this vector space, words that have similar distributions are closer together than words with different distributions, given some distance measure. This type of setup has been shown to capture relevant syntactic and semantic properties of words, and they have been successfully applied to various tasks (Collobert and Weston, 2008; Socher et al., 2011; Mikolov et al., 2013c).

There are a number of proposed methods to learn word embeddings for large text databases (Turian et al., 2010; Mikolov et al., 2013a; Huang et al., 2012; Pennington et al., 2014). For this work, we choose to use the skip-gram model (Mikolov et al., 2013a,b) via the publicly available implementation of *word2vec*.[2]. The reason for this choice is motivated by the work of Levy et al. (2015), which compares various methods to learn word embeddings. The authors indicate that, in general, most methods learn similar word representations and achieve similar performance in a variety of similarity and analogy tasks. However, the skip-gram model, with some hyperparameter configurations, tends to outperform techniques such as continuous bag-of-words (CBOW, Mikolov et al. (2013a)) or Global Vectors (GloVe Pennington et al. (2014)).

In speech synthesis, continuous representations of words learned through models such as the skip-gram or one of its variants have been previously explored (Wang et al., 2015a, 2016b). While Wang et al. (2015a) claimed that these embeddings are a useful unsupervised replacement for hand-annotated word-level features, Wang et al. (2016b) failed to see any improvements at this level.

To learn these embeddings, we have used the freely available English

[1] Less formal experiments were conducted on word-level deep neural networks. These architectures were similar to the ones investigated in this chapter, except they were defined at the word level rather than the syllable level. Acoustic parameters of the suprasegmental network were defined to be the average over the entire word and input features were set to be those extracted on words. Results showed similar improvements to the syllable-level network, further reinforcing the thought that it is the hierarchical framework rather than the addition of new features giving the strongest results.

[2] https://code.google.com/archive/p/word2vec

Wikipedia data dump from September 2015.[3] This data has been pre-processed and cleaned and we have kept the first 500 million words. Two models were trained on this dataset, one using an embedding size of 100 and another an embedding size of 300. The systems use the publicly available *word2vec* implementation of the skip-gram model with negative sampling and they were trained for 15 epochs with a window of 5 words.

We consider five systems, whose identifiers are given in Table 7.3. The ***frame*** system is the basic frame-level DNN using no additional features. ***frame-w100*** uses 100-dimensional word embeddings and appends them to the input of the basic DNN. The remaining systems use the framework illustrated in Figure 7.1, including the *bag-of-phones* representation described in the previous section. The first model (***syl-BoP***) is trained without word embeddings, and the final two with 100 and 300-dimensional embeddings (***syl-BoP-w100*** and ***syl-BoP-w300***)

Table 7.3 summarizes the results for each system. Adding word embeddings to the frame-level DNN does not show any improvements over the baseline. This is surprising, as we would expect some improvements, given the work described in Wang et al. (2015a). However, the authors in Wang et al. (2015a) used a carefully annotated non-expressive database for their experiments. In these experiments, we used a shallow feature set learned automatically from various text sources. They have also used a bi-directional LSTM, while we have used a feedforward DNN. We do not observe any relevant differences in terms of objective measures for the hierarchical systems. However, adding a larger word feature vector allows the system to slightly improve *f0* prediction. This is interesting, as it suggests that higher-dimensional features may be useful to learn more complex relationships between suprasegmental units. In that case, the proposed hierarchical system might be useful in processing them.

7.4.3 Discussion

Regarding the main hypothesis evaluated in this section, we find some evidence to support the notion that a hierarchical framework is able to use high-dimensional

[3] http://dumps.wikimedia.org/enwiki/20150901

System	**MCD**	**BAP**	**F0-RMSE**	**F0-CORR**
frame	4.596	2.197	28.054	.449
frame-w100	4.598	2.204	28.048	.448
syl-BoP	4.557	**2.176**	27.095	.477
syl-BoP-w100	**4.55**	2.177	27.086	.463
syl-BoP-w300	4.565	2.178	**26.850**	**.479**

Table 7.3: Objective results for word embedding systems. Abbreviations are identical to those in table 7.1.

suprasegmental features more efficiently than a frame-level network. We observe that *a hierarchical framework generally leads to more accurate generation of speech parameters.* This observation was clear from Figure 7.4 and follows the findings reported in Section 7.3. In terms of additional features, we observe that *syllable bag-of-phones provide slight improvements in the objective measures when used with a hierarchical framework.* However, it should be noted that the improvements were small when compared to a hierarchical system not using syllable bag-of-phones and may not be perceptually significant. According to Figure 7.4, mel-cepstral distortion and *f0* correlation are the acoustic features that are more influenced by this representations. When using word embeddings with a hierarchical framework, we observe *no clear changes in the objective scores, although higher-dimensional representations appear to be more promising.*

We find some evidence to support the initial hypothesis, indicating that there is some effect related to the addition of high-dimensional representations, although the results were not as strong as we expected them to be. Chapter 8 will continue this line of research and provide more solid results regarding these observations. In the following section, we evaluate two of the best performing systems against a baseline system with a listening test.

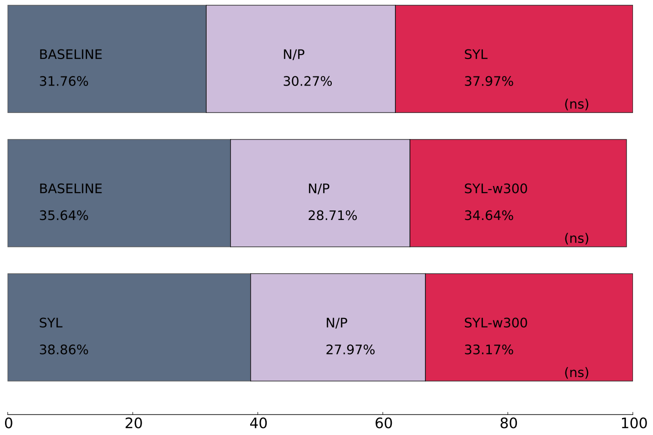

Figure 7.5: Preference test results with N/P indicating "No Preference". In parenthesis, p-values indicate the results of 1-tailed binomial tests with an expected 50% split, with the N/P results evenly distributed over the other two conditions.

7.5 Subjective evaluation

A listening test was conducted on selected systems described in previous sections. For brevity, henceforth we will refer to the system *syl-BoP* as system *syl*. This system pre-processes suprasegmental features separately with bags-of-phones and uses a bottleneck layer with 256 nodes. System *syl-w300* is similar, but it adds 300-dimensional word embedding representations to the input of the syllable-level network. The *baseline* system processes suprasegmental features directly. From a held-out set, 50 test utterances were synthesized from the parameters predicted by the frame-level network. 16 native speakers judged randomized utterance pairs for the two conditions in a preference test. Each pair was judged 8 times and each condition received a total of 400 judgments. Participants were allowed to play utterance pairs as many times as they wished and were asked to select the utterance that they perceived to be more natural.

Figure 7.5 shows the results of the listening test. We conduct a 1-tailed binomial test on the results, assuming an expected 50% split with the *no-preference* judgments distributed equally over the remaining conditions. Although there appears to be a slight preference for the *syl* system, the results are not significant in any of the conditions. This is surprising, as listening to the synthesized wave-

ID	syl-preference	ID	syl-preference
1	48.15%	9	44.44%
2	60.87%	10	65.38%
3	59.26%	11	42.59%
4	56.52%	12	52.17%
5	53.70%	13	55.56%
6	63.04%	14	65.22%
7	38.89%	15	46.30%
8	56.52%	16	45.65%

Table 7.4: Preference results by participant. Percentages indicate participant's preferences for the system *syl* over the baseline.

forms informally showed clear differences between the systems. In the following section, we provide further insights into these results.

7.5.1 Analysis of subjective results

The results observed in the listening test were surprising. Clear differences were perceived when listening to the synthesized speech samples informally. The subjective evaluation, however, showed no overall significant preference, although some participants seem to prefer the proposed hierarchical methods. Table 7.4 shows aggregated results per participant with the no-preference option divided equally among competing conditions. While some listeners prefer the hierarchical systems (2, 6, 10, 14), others prefer the baseline system (7, 11). Some participants do not have a clear preference (1, 12).

The 50 utterances used for evaluation were randomly selected from a held-out set and are not part of the same set used to compute objective measurements. Therefore, we revisit the objective measurements for these utterances and we conduct paired t-tests on the *syl* and *baseline* systems. We observe that, in terms of mel-cepstral distortion, there is a significant difference between the baseline (M=4.631, SD=0.429) and hierarchical (M=4.577, SD=0.42) systems, t(49)=4.1744, p<.001. However, in terms of band aperiodicity distortion, we do not observe a significant difference between baseline (M=2.196, SD=0.174) and hierarchical (M=2.184, SD=0.189) systems, t(49)=0.8873, *ns.* Considering the *f0* signal, there is a significant improvement in terms of RMSE between the two systems: baseline (M=25.06, SD=10.44) and hierarchical (M=24.13, SD=10.88), t(49) = 2.1276, p<.05. But we failed to observe significant differences in terms of *f0* correlation for the baseline (M=0.504, SD=0.153) and the hierarchical approach (M=0.529, SD=0.180), t(49)=1.2706, *ns.*

Objective measurements for the two proposed systems showed statistically significant improvements over the baseline in terms of mel-cepstral distortion and *f0* RMSE. In order to understand what listeners are responding to when submitting their judgments, we compare their preferences and the differences between the two systems in terms of objective measures. As before, we compare the *baseline* and *syl* systems.

Figure 7.6 shows a visualization of these trends. Each point on the figure represents an utterance in the test set given in the listening test. For each utterance, we take the relative number of times it was preferred in a judgment, with the no-preference option distributed among the two conditions. Therefore, the x-axis indicates the preference towards the proposed *syl* system. A lower value shows a preference towards the baseline and a higher value a preference towards the proposed system. The y-axis indicates the different between the hierarchical and baseline system for selected acoustic parameters. Positive values indicate higher error for the *syl* system, while negative values indicate higher error for the *baseline* system. Because we are trying to maximize correlation, this pattern is shifted in the visualization of *f0* correlation.

Correlating objective measures with perceptual scores, we observe that there is no significant correlation in terms of mel-cepstral distortion (r=-0.0075, n=50, *ns*) and band-aperiodicity distortion (r=-0.174, n=50, *ns*). However, we do observe a significant correlation in terms of *f0* RMSE (r=-0.355, n=50, p<.01) and *f0* correlation (r=0.323, n=50, p<.05). These results show that listeners are judging the utterances in terms of the *f0* signal. A significant correlation is observed in terms of *f0* correlation even though no statistically significant differences were

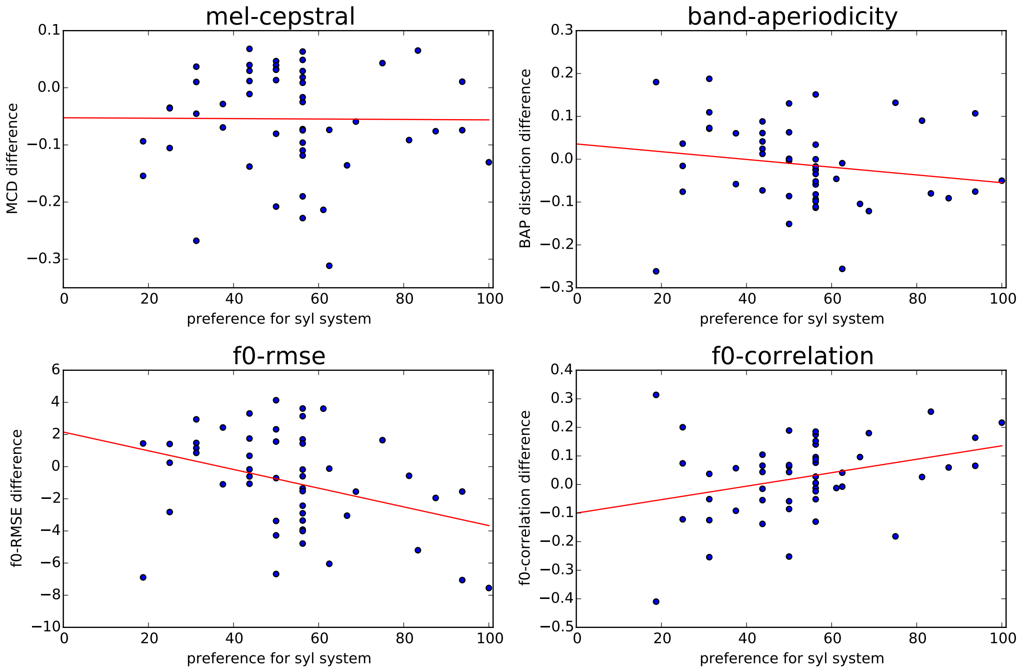

Figure 7.6: Subjective judgments and objective measurement distortion for acoustic features over the 50 test utterances. Each point indicates an utterance in the listening test. Preference for the *syl* system is indicated in the x-axis. Difference in objective measurement between *syl* and *baseline* systems is indicated in the y-axis. Red line shows a best fit line.

observed in the objective scores. On the other hand, listeners do not respond to mel-cepstral coefficient variation, even though there is a statistical significant improvement in terms of mel-cepstral distortion.

It appears, in any case, that the *f0* signal is the decisive factor that listeners respond to when submitting their judgments. This is reassuring, as when learning representations of suprasegmental context, we are essentially focusing on a better understanding of prosody. The proposed systems, therefore, do modify the *f0* signal in a way that affects listeners' preferences. This was observed in terms of the objective scores differences and in terms of listener's judgments.

From this analysis, we can also infer that listeners are rating utterances according to their similarity to the natural parameters. When comparing two samples, lower RMSE and higher correlation are preferred over higher RMSE and lower correlation. This is also an interesting result, as it is unclear what listeners will prefer when judging prosody. Similarity to natural parameters does not necessarily imply listener preference. A prosodic layer that is different to the one that was naturally produced may still be acceptable for a particular listener. The results obtained might be a side effect of the methodology used. A preference test without context might not be the most adequate to evaluate natural intonational patterns. Section 10.3 of this thesis provides a short discussion on the need for evaluation protocols focusing on hypotheses centered on the generation of speech prosody.

7.6 Overall discussion

For a better insight into how the syllable-level DNN manipulates the suprasegmental features, we visualize the hidden representations learned by the network. For this task, we use t-SNE (t-Distributed Stochastic Neighbor Embedding, van der Maaten and Hinton (2008)), an efficient technique for dimensionality reduction. We randomly sample 1500 syllable embeddings from the *syl* system. These are then reduced with t-SNE to two dimensions. The results are plotted in Figure 7.7. In the figure, two sections are enlarged for clarity. Colors are assigned according to syllable nucleus and syllables are represented textually as the concatenation of its phones separated by underscore (e.g. *I_n*). We observe that some syllables, which could potentially be different, are closer in the embedding space. This is the case, for example, for *T_r_i* and *T_r_u*. The onset of the syllable seems to be the main similarity between the two samples. On the left-hand section, it appears to be the nucleus and coda the main point of similarity for *I_N* and *n_I_N*.

The syllable-level network can be thought of as a feature extractor for suprasegmental features. Appending new representations of context to a frame-level feature vector could introduce more noise than useful information. This could be the reason why we failed to observe improvements when adding word embeddings to the frame-level network. Chapter 8 investigates this hypothesis. In any case, pre-processing such features separately may allow us to learn useful

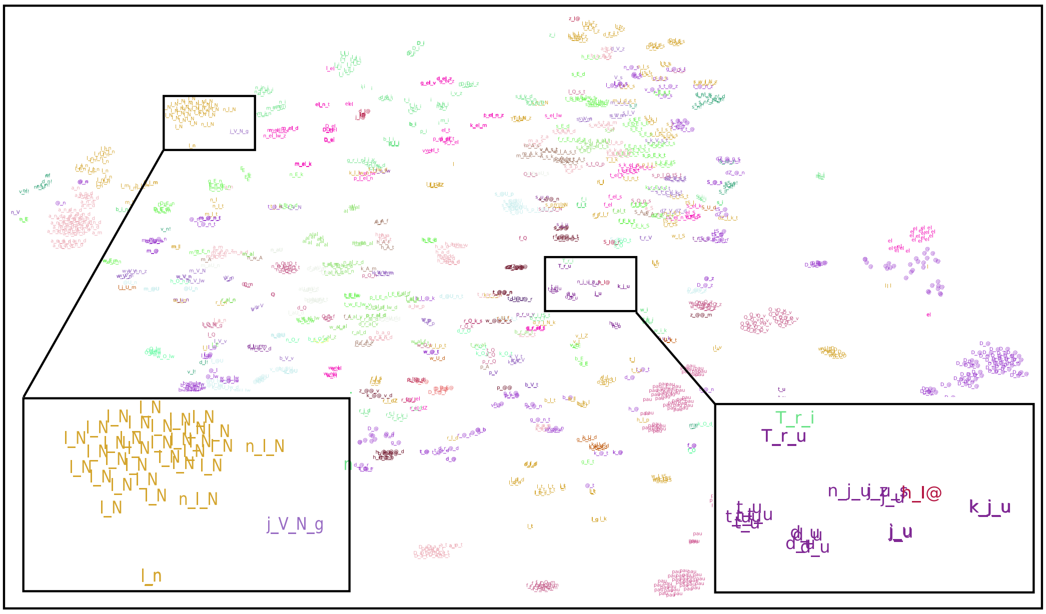

Figure 7.7: 2-dimensional visualization of suprasegmental embeddings at syllable-level using t-SNE (van der Maaten and Hinton, 2008). Colors are assigned to syllables based on nucleus identity. Syllables are represented textually as the concatenation of its phones joined by an underscore (e.g. *l_n*).

and compact representations of suprasegmental context for frame-level prediction. The *syl* system achieves good accuracy when computing objective measures on the syllable-level parameters (MCD: 4.699, BAP: 1.252, F0-RMSE: 26.717). Although we failed to observe an overall significant preference for the proposed systems, the hierarchical systems are still capable of learning meaningful embeddings of suprasegmental features.

The results observed in this work raise questions regarding the evaluation methodology of synthetic speech. In Section 7.5.1, we showed that, in a perceptual evaluation, there is a slight preference towards the proposed system, although that preference is not statistically significant. Participants judged various utterances in a preference test. The lack of significant preference could be due to the large *no-preference* between the two compared systems. But it could also be due to a participant effect, where we see a trend to one system or another depending on participant. Increasing the number of judgments per participant could help us identify the significance of these trends. Alternatively, a different evaluation methodology, such as a MUSHRA test, might be more adequate, or designing evaluation protocols aimed specifically at assessing how synthetic speech handles prosodic properties, for example, by including some type of context (see Section 10.3 of this thesis for further lines of research on this topic).

Finally, in Section 7.1, we mentioned two scenarios in which suprasegmental features could be investigated. The first explores methods to transform and manipulate features that are already available. The second scenario focuses on deriving additional suprasegmental features for speech synthesis.

It is commonly accepted that expressive speech, and prosody in general, are mostly affected by suprasegmental effects. But current systems operate mainly at the frame or phone level. Adding more features to frame-level systems may not lead to more accurate acoustic parameter generation, as the results of Wang et al. (2016b) indicate. Therefore, it is essential to have a framework that can leverage suprasegmental information efficiently without affecting segmental prediction.

In this chapter, we have provided contributions to each of these scenarios. In Section 7.3, we have investigated the first scenario with a top-down hierarchical framework. The results showed some evidence that the proposed hierarchical structure may be operating as a feature extractor. This hypothesis will be evaluated under more controlled conditions in Chapter 8.

Section 7.4 provided some results considering the second scenario. Two sets of features were evaluated with the current framework: syllable bag-of-phones and text-derived word embeddings. Future work could focus simply on learning more appropriate representations of context for DNN-based speech synthesis. With more relevant features, it would be interesting to observe how this method would perform. Chapter 9 provides a novel way to infer suprasegmental features based on acoustic counts. These features are evaluated with a frame-level network. In Chapter 10, we integrate these additional features in a hierarchical framework to test the hypothesis that these structures are able to leverage new information.

Finally, in terms of future work, it is unclear at this point how this method performs with recurrent systems. It has been argued that systems using recurrent neural networks (RNNs) or long short term memory networks (LSTMs, Fernandez et al. (2014)) are capable of leveraging suprasegmental information much more efficiently. Understanding how this method could be used within such systems could be useful, although a mixed hierarchical and recurrent model such as the one presented in Chen et al. (1998) might be more interesting to explore. Future work could also focus on the acoustic features in the suprasegmental network, such as *f0* representations or wavelet-based decompositions of acoustic signals.

7.7 Conclusion

In this work, we have proposed a system that pre-processes suprasegmental features separately from segmental features. The system can be thought of as a top-down hierarchical model, where a deep neural network transforms suprasegmental features first and then integrates them with a frame-level network.

We have investigated the hierarchical framework and the effect of various configurations such as bottleneck layer size or position. The addition of new features in the form of syllable bag-of-phones and text-derived word embeddings was also evaluated. Although objective results seem to show a clear trend towards the proposed hierarchical systems, a subjective test failed to see significant results. Further work presented in this thesis will focus on understanding the hierarchical network and learning new suprasegmental features from a speech database.

Chapter 8

Parallel and cascaded deep neural networks

This chapter covers the work described in "Parallel and cascaded deep neural networks for text-to-speech synthesis" (Ribeiro et al., 2016a), which was presented at the 9th Speech Synthesis Workshop (SSW9).

Motivated by the findings of Chapter 7, we conduct an investigation of cascaded and parallel deep neural networks for speech synthesis with a focus on the input linguistic features. In these systems, suprasegmental linguistic features (syllable-level and above) are processed separately from segmental features (phone-level and below). Cascaded networks input suprasegmental features alongside frame features. Parallel networks process segmental and suprasegmental features separately and concatenate them at a later stage. These experiments are conducted with a standard set of high-dimensional linguistic features as well as a hand-pruned one.

8.1 Introduction

Chapter 7 investigated a hierarchical structure in the form of a cascaded deep neural network. This work was motivated by the idea that frame-level networks might underperform when given high-dimensional input linguistic features of suprasegmental contexts. Hierarchical architectures are attractive alternatives, as they could have the ability to extract meaningful information at suprasegmental levels by pre-processing them separately. This property would be useful when evaluating additional high-dimensional representations of linguistic context. For example, injecting a high-dimensional representation of words into a frame-level network might cause the model to underperform when compared to an identical model that does not use such representations.

In Chapter 7, an exploratory analysis of architecture configurations observed that, when using additional features, the hierarchical systems outperformed the non-hierarchical systems in terms of objective measures. However, we did not find relevant improvements in terms of a subjective evaluation. Results indicated that it was the hierarchical framework rather than the addition of new features that contributed to improvements in terms of objective scores. An hypothesis put forward was that these hierarchical architectures were operating as denoisers or feature extractors on the linguistic features. That is, a suprasegmental network pre-processes high-dimensional representations of context and extracts only information *useful* for predicting acoustic features. Channeling these denoised representations to a frame-level model instead of the noisy representations allows the model to generate more natural acoustic parameters.

In the current chapter, we provide further insights into this hypothesis. This investigation focuses mostly on the modeling architecture and how it processes linguistic features. For this reason, according to the three sub-problems detailed in Chapter 1, this work provides contributions to sub-problem 3: *mapping acoustic and context representations*. While Chapter 7 investigated only a cascaded hierarchical framework, this chapter will investigate cascaded and parallel neural networks. This was inspired by the work described in Yin et al. (2016).

Section 8.2 details the cascaded and parallel hierarchical architectures, as well as the two sets of linguistic inputs to be used in the evaluation of the proposed hypotheses. Section 8.3 describes detailed hypotheses and systems trained, as well as the objective and subjective evaluations. Finally, Sections 8.4 and 8.5 discuss the results and future lines of research.

8.2 Deep neural network architectures

8.2.1 Basic network

The basic deep neural network is identical to the frame-level model described in Section 7.2. This is a simple multilayer perceptron with 6 hidden layers, each layer with 1024 nodes. The activation function is set to be *tanh* in the hidden layers and it is linear in the output layer. For training, a mini-batch size of 256 is set and the maximum number of iterations is set to 25.

We use the same output features described in Section 7.2. These are log-*f0*, 60-dimensional mel cepstral coefficients (MCCs), and 25 band aperiodicities (BAPs), extracted with the STRAIGHT vocoder (Kawahara et al., 1999, 2001) at 5 ms intervals. Dynamic features (deltas and delta-deltas) are appended to these static features. The log-*f0* signal is linearly interpolated and a binary voiced/unvoiced decision is appended to the acoustic feature vector. The complete output vector has a total of 259 dimensions, which are then normalized to zero mean and unit variance.

8.2.2 Cascaded and parallel networks

We define *segmental features* to be those that describe the input at the level of the segment and below, at the phone and frame level. We term features that represent the input at linguistic levels above the segment *suprasegmental features*: features at the syllable, word, phrase, and utterance levels.

In the cascaded and parallel approaches, segmental and suprasegmental features are decoupled and processed separately. In both systems, distributed representations of suprasegmental contexts are learned and later integrated into a frame-level system. An initial suprasegmental network is defined at the syllable level. This network takes as input representations of context at the syllable and above levels and maps them to acoustic parameters defined over syllables. For the current experiments, the output of this network consists of a 258 dimensional vector obtained by averaging the frame-level acoustic features over the entire syllable. This follows best architecture investigated in the experiments detailed in

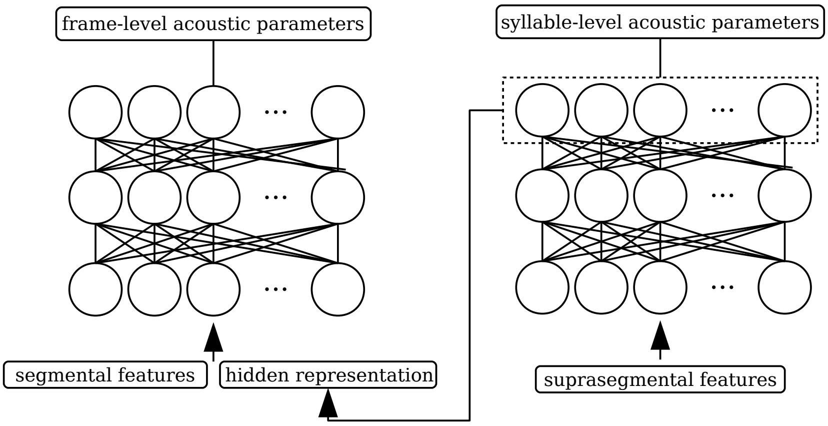
frame-level acoustic parameters
syllable-level acoustic parameters
segmental features
hidden representation
suprasegmental features

Figure 8.1: Hierarchical cascaded deep neural network.

Section 7.3.3.

The suprasegmental network is set to be a 6 hidden layer triangular network. In terms of layer size, it is defined as (1024, 1024, 1024, 1024, 512, 256). That is, the top hidden layer is a bottleneck layer with 256 dimensions. The hidden activation function is set to be *tanh* and the output layer uses a linear activation function. Mini-batch size is set to 16 and the maximum number of iterations is set to 25.

Figure 8.1 illustrates the *cascaded deep neural network* (Yin et al., 2016), which can be thought of as a top-down hierarchical network. This was the architecture used in the work detailed in Chapter 7. The distributed representation of suprasegmental features is concatenated with the segmental feature vector. A second network is then trained to generate source and spectral parameters at the frame level.

Figure 8.2 illustrates the *parallel deep neural network* (Yin et al., 2016). In this integration strategy, segmental and suprasegmental features are joined at a later stage. The second network inputs only segmental features and its architecture is similar to that of the suprasegmental network. The distributed representation learned from both networks, each with 256 dimensions, is used to drive a single layer network that generates acoustic parameters at the frame level. The hyperparameters of the single-layer network are similar to those of the frame-level network.

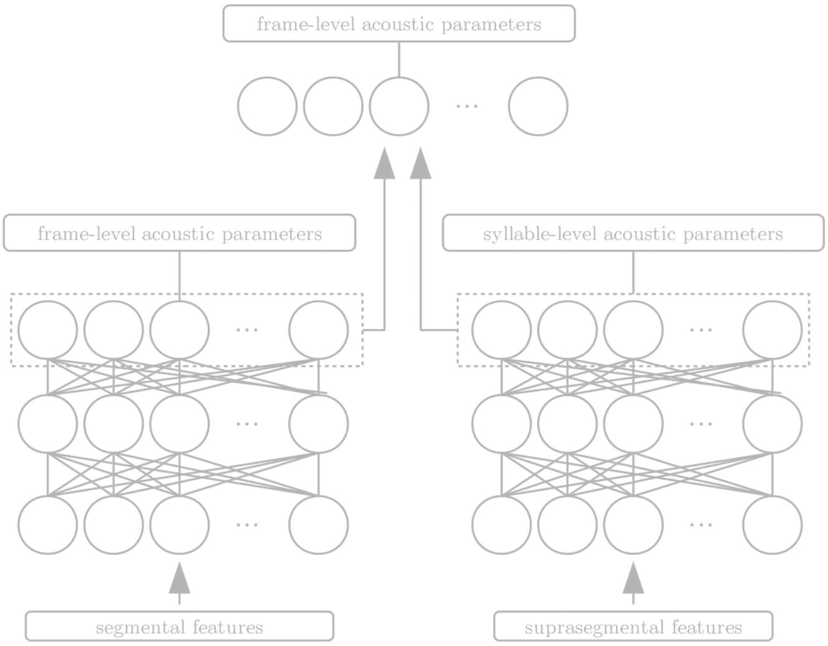
frame-level acoustic parameters
frame-level acoustic parameters
syllable-level acoustic parameters
segmental features
suprasegmental features

Figure 8.2: Hierarchical parallel deep neural network.

8.2.3 Linguistic features

As input to the deep neural networks, we will consider two sets of linguistic features. These two feature sets are fully detailed in Appendix A.

The first feature set is inherited from a conventional question set used for tree clustering in HMM-based synthesis (Appendix A.0.1). These are the questions used in the work described in Chapters 4 and 5. Linguistic contexts obtained through a common front-end such as the one distributed with Festival[1] are defined at phone, syllable, word, phrase, and utterance levels. Questions are defined in terms of quinphone identity, syllable stress or accent, part-of-speech, predicted phrase boundaries, ToBI labels, or positional information in words, phrases, and utterances. To these, we add two additional features defined at the state level. These refer to the state number (absolute and relative position) within the current phone after forced alignment of the data. We term this set of linguistic features the *standard* feature set.

A major concern with the standard set of linguistic features is its high di-

[1] http://www.cstr.ed.ac.uk/projects/festival

linguistic level	hand-selected	standard
state	2	
phone	350	
syllable	152	426
word	92	184
phrase	-	211
utterance	-	300

Table 8.1: Dimensionality of input features per linguistic level.

mensionality. There is an imbalance between the segmental and suprasegmental feature sets with respect to the number of features. With respect to suprasegmental text representations, some features may not be useful for frame-level prediction. Therefore, the question set was pruned through optimization on a small text-to-speech database. Various input feature combinations were selected and used as input during acoustic model training. Questions were discarded if their absence led to an improved acoustic model or if the model did not underperform in terms of objective results.[2] Based on this analysis, features at phrase and utterance levels were removed. Various features within the syllable and word level sets were ignored, such as forward or backward context and several positional features. This is the linguistic feature set used for DNN-based speech synthesis by the Merlin Toolkit (appendix A.0.2). This smaller pruned set of linguistic features is here termed the *hand-selected* feature set.

Binary representations of these question sets were used and all features were normalized to the range of [0.01, 0.99]. Segmental features were kept constant for the standard and hand-selected sets. Thus, only suprasegmental features vary between the two sets. Table 8.1 summarizes the dimensionality at each linguistic level of each of the feature sets. Note that there is a large number of features at the phrase or utterance levels because we choose to use a binary representation of these features. The number of dimensions increases because

[2] This work was performed during early development stages of the Merlin Toolkit (Wu et al., 2016) by Zhizheng Wu, to whom we are grateful.

each feature is represented as a binary question (e.g. “Is the total number of words in the current sentence 3?”). For most features, this could be avoided by treating them as numerical features rather than binary. For example, the number of words in a sentence could be represented as an integer (e.g. *3*). This is an artifact from tree clustering in HMM-based systems, which requires all questions to be binary. However, this representation is well suited to the current problem, whose main hypothesis requires the introduction of some noise in high-dimensional representations of linguistic features. Representing these features in a binary way allows us to achieve that result because the acoustic model is forced dedicate more parameters to the processing of large high-dimensional sparse vectors. For this reason, we claim that the *standard* feature set is a high-dimensional noisy input when compared to the *hand-selected* feature set.

8.3 Experiments

As in earlier chapters, these experiments were conducted on expressive audiobook data. A detailed description of the dataset is given in Section 6.3.2 (p. 122). Training, development, and test sets are similar to those in Chapter 7. They consist of 4500, 300, and 100 utterances, respectively. The training set consists of 9 hours of speech data, while validation contains 33 minutes, and test set 9 minutes. The data used for the listening test was randomly drawn from a held-out set.

8.3.1 Systems and hypotheses

Given three network architectures and two sets of linguistic features, a total of six systems were trained. Two systems employed the basic feedforward deep neural network architecture, two systems the cascaded deep neural network architecture, and two systems the parallel network architecture. Within each of these system pairs, we vary the input feature vector, either using the standard set or the hand-selected subset.

These systems were constructed to test the following hypotheses:

Addition of noisy suprasegmental features: Adding more (suprasegmental) features to a frame-level DNN will degrade the performance of the model. It is expected that the baseline system with the standard feature set will perform worse than the baseline system with the hand-selected features, as saturating a subsegmental model with noisy suprasegmental inputs is likely to be harmful. This hypothesis should provide further evidence for the observations described in Chapter 7.

Hierarchical systems: Hierarchical architectures will outperform non hierarchical systems. Previous investigations have suggested that handling various linguistic levels separately tends to be beneficial for speech synthesis systems. We expect cascaded and parallel deep neural networks to outperform their corresponding basic feedforward network.

Parallel and cascaded deep neural networks: Parallel architectures will outperform cascaded architectures. Although using a different setup, previous work using these methodologies has found that parallel systems tend to outperform cascaded systems (Yin et al., 2016). One of the disadvantages of processing suprasegmental information directly with a subsegmental network is that the system might learn to depend heavily on segmental features and ignore long-term unit information. In a cascaded approach, even though segmental and suprasegmental feature sets are decoupled, a frame-level network still has to account for them. In a parallel architecture, this may not be the case, as the system processes the two feature sets separately and only concatenates them in the top hidden layer.

8.3.2 Objective results

Table 8.2 shows objective measures on the test set for all six systems. The first block in the table denotes the three networks operating with the *standard* set of linguistic features. The second identifies those that use the *hand-selected* feature set. Observing only the baseline feedforward networks (systems FS and FH), we note a small improvement when moving to the hand-selected feature set, especially in terms of mel-cepstral distortion. All hierarchical systems outperform

system	architecture	features	MCD	BAP	F0-RMSE	F0-CORR
FS	feedforward	standard	4.68	2.22	28.23	.43
CS	cascaded	standard	4.60	2.19	27.43	.45
PS	parallel	standard	**4.59**	**2.17**	**26.97**	**.45**
FH	feedforward	hand-selected	4.61	2.20	27.66	.44
CH	cascaded	hand-selected	**4.57**	2.19	27.48	.45
PH	parallel	hand-selected	4.59	**2.17**	**27.16**	**.45**

Table 8.2: Objective results for trained systems. MCD is mel cepstral distortion, BAP is band aperiodicity error, and F0-RMSE and F0-Corr are the root-mean-squared error and correlation between the predicted and original *f0* signal on voiced frames.

their respective baselines, although the impact appears to be less for the systems using hand-selected features.

The parallel architecture gives the best results. It is interesting to observe that, when using this architecture with a standard feature set, we achieve performance that is comparable to the same architecture using a hand-selected feature set. In terms of *f0* RMSE, the parallel system with the standard feature set (system PS) gives the lowest error. This is reassuring, as we provide the syllable-level network with a larger number of input features. This suggests that hierarchical architectures are capable of leveraging high-dimensional representations of suprasegmental contexts. Such is not the case for frame-level networks. In the following section, we report a listening evaluation aimed at validating these observations.

8.3.3 Subjective results

To assess the naturalness of speech samples produced by the trained systems, we conducted a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test (ITU-R Recommendation BS. 1534-1, 2015). This methodology allows the simultaneous comparison of multiple samples. Each sentence to be tested is assigned a set of stimuli. In our case, a single set of stimuli includes 7 samples: one

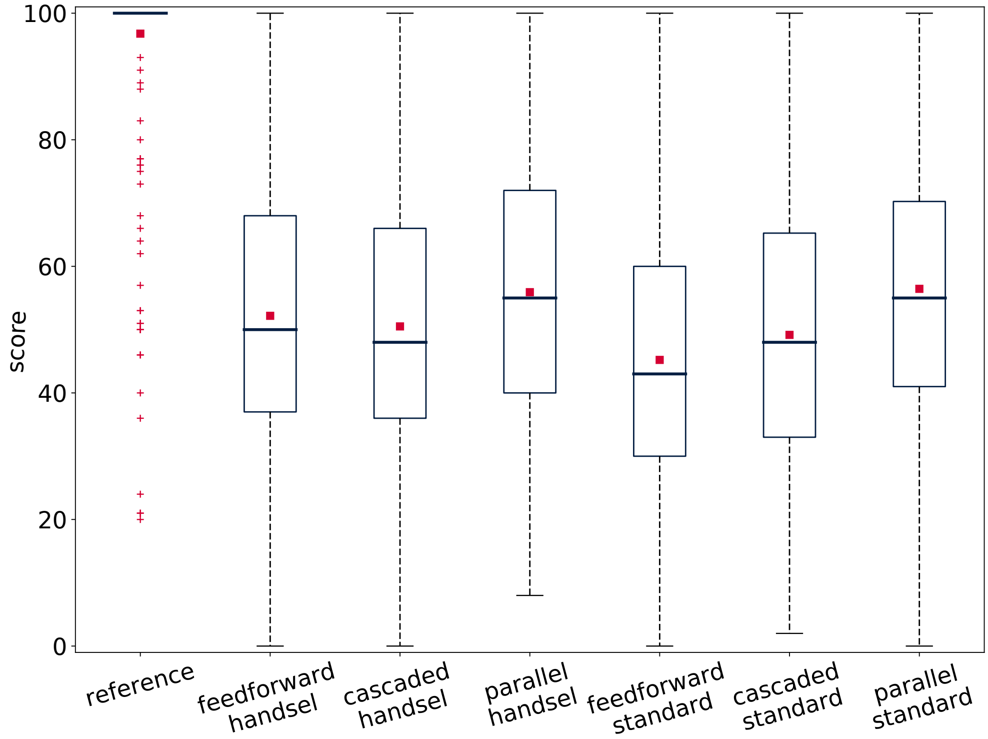

Figure 8.3: Boxplot for the absolute results from the MUSHRA evaluation. The y-axis denotes the absolute score given by participants on a scale 0-100 and the x-axis indicates evaluated systems. Dark blue horizontal line shows the median and the red square shows the mean.

from each system described in Section 8.3 plus a final sample of matching vocoded speech. This final sample is termed the *reference*. Within each set, samples are unlabeled and, for each participant, the order of the samples is randomized. Participants are then asked to judge the set of parallel samples on a scale from 0 (completely unnatural) to 100 (completely natural) with respect to the reference sample. The reference sample itself is included in the unlabeled samples. This ensures that participants provide accurate judgments and fixes the high end of the scale.

A total of 20 native English listeners participated in the listening test. Each participant rated 20 sets of stimuli produced from sentences taken from a set held out from the training data. Sentence order was randomized for each participant. This allowed us to gather a total of 400 parallel comparisons. All tests were

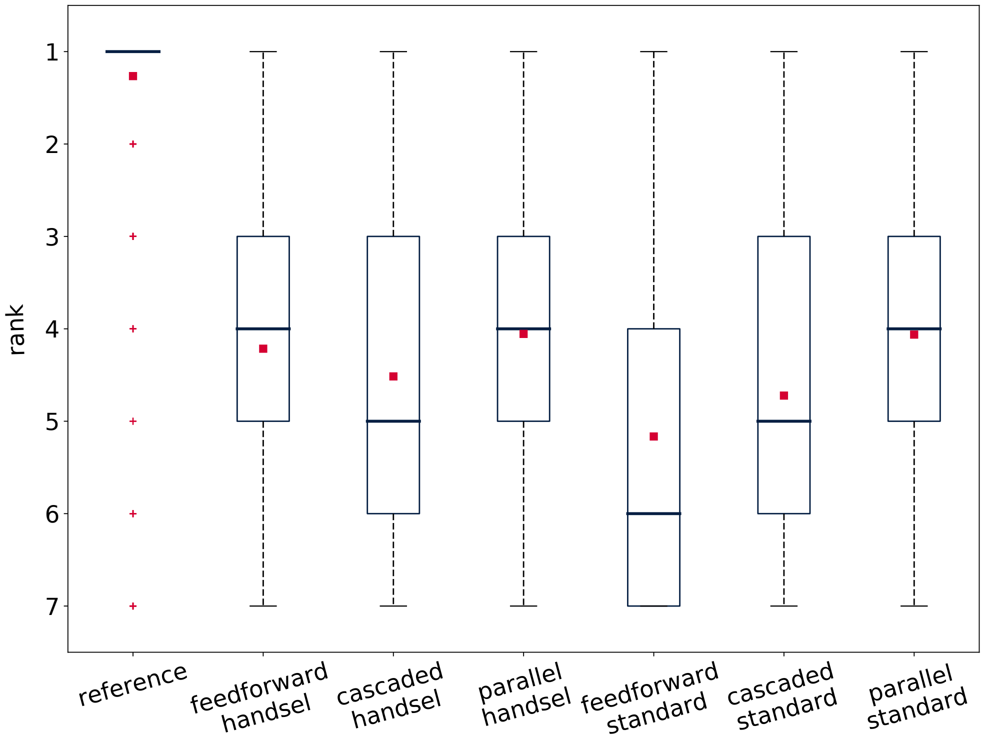

Figure 8.4: Box plot for the rank order results from the MUSHRA evaluation. The y-axis denotes the rank of each system, as given by participants in terms of absolute scores, and the x-axis indicates evaluated systems. Dark blue horizontal line shows the median and the red square shows the mean.

conducted in sound-insulated booths and all listeners were remunerated for their time.

Figure 8.3 shows the distribution of the stimuli for each condition in terms of the absolute values given by the test participants. Figure 8.4 shows the distribution in terms of their rank order, as derived from the absolute values.

8.4 Discussion

To better understand the results, we conduct a two-tailed paired t-test on the absolute values given by the listeners. To reduce Type I error whose likelihood is greater for multiple comparisons, we perform a Holm-Bonferroni correction on all results. All system pairs are significantly different at the level of p<.05, except systems FH and CH, systems CH and CS, and systems PH and PS. Furthermore, we conducted a double-sided Wilcoxon signed-rank test on the rank order results with a Holm-Bonferroni correction. The same pattern was observed, with the addition of two system pairs not showing statistically significant differences: systems FH and PH, and systems FH and PS.

We begin by considering the first hypothesis: **addition of noisy suprasegmental features**. Looking at the results of the two baseline systems, we observe that there is a statistically significant difference between system FH, using a hand-selected feature set, and system FS, using a noisy standard feature set (z=-6.607, p<.001, r=-0.33). The comparison relevant to this hypothesis is illustrated on the upper left-hand subplot of Figure 8.5. Results suggests that adding a larger number of suprasegmental features to a frame-level network significantly damages performance. This might be problematic when exploring a better understanding of longer context for prosody modeling. This evidence supports the motivation for the work presented in Chapter 7. If we are investigating high-dimensional representations of context, using a frame-level network might mask potential benefits from rich representations.

The second hypothesis focuses on the behavior of **hierarchical systems**. It was expected that hierarchical approaches would consistently improve over their non-hierarchical counterparts. A related claim, motivated by observation made in Chapter 7, hypothesized that these architectures would function as feature extractors or denoisers.

For this analysis of the results, we consider the systems using different feature sets separately. We observe that, with systems using a *hand-selected* feature set, hierarchical systems do not significantly differ from a non-hierarchical system. A comparison of a feedforward system with a cascaded system does not show statistically significant differences in the ranking of the systems (z=-2.341, *ns*, r=-0.12). Similarly, a parallel integration strategy does not differ from a standard feedforward network using the same feature set (z=1.039, *ns*, r=0.052). However, this pattern is not observed if we consider the three systems using a high-dimensional *standard* feature set. A cascaded hierarchical framework significantly outperforms

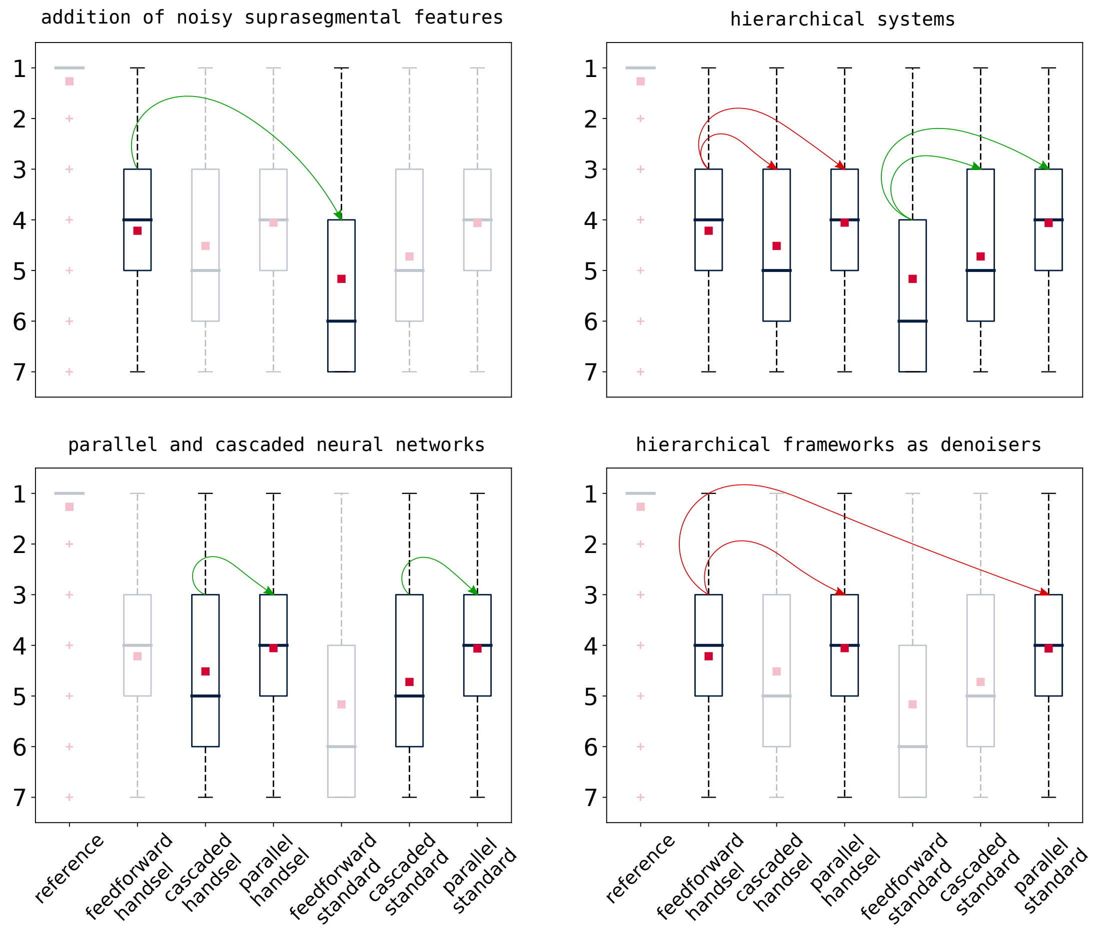

Figure 8.5: Rank order results from the MUSHRA evaluation annotated with relevant system comparisons. The y-axis denotes the rank of each evaluated system, as given by participants in terms of absolute scores. Green arrows indicate statistically significant comparisons based on the holm-bonferroni-corrected double-sided Wilcoxon signed-rank tests. Red arrows indicate non significant differences. For clarity, comparisons related to different hypotheses are illustrated in different subplots.

a frame-level system (z=3.226, p<.05, r=0.161). Similarly, a parallel hierarchical framework outperforms the same frame-level system (z=7.354, p<.001, r=0.368). Therefore, hierarchical systems improve over a non-hierarchical system when using high-dimensional noisy representations of context. These comparisons are visualized in the upper right-hand subplot of Figure 8.5. These results suggest that these hierarchical frameworks might be operating as feature selectors for high-dimensional suprasegmental features. In our results, the hierarchical systems using the standard set are comparable to most systems using hand-selected features.

The third hypothesis concerned the comparison of **parallel and cascaded deep neural networks**. Although using a hand-selected feature set, Chapter 7 used a cascaded approach to integrate suprasegmental representations with a frame-level network. Objective measures showed some improvements in terms of the generation of acoustic parameters. However, a listening test failed to detect significant differences between the systems.

In the work of Yin et al. (2016), which focuses on a multi-level approach to modeling the *f0* signal, it was seen that parallel architectures tend to outperform cascaded architectures. This observation is supported by our results. There is a statistically significant preference for the parallel systems over cascaded systems. The preference for a parallel architecture over a cascaded one is seen with systems using a *hand-selected* feature set (z=3.74, p<.001, r=0.187) and with systems using a *standard* feature set (z=4.663, p<.001, r=0.233). The relevant system comparisons is illustrated on the lower left-hand subplot of Figure 8.5.

We could hypothesize that the frame-level network of the cascaded systems ends up depending too much on segmental features instead of balancing both sets. This is not the case for the parallel integration, as only one layer is used after concatenation. Further work could investigate this interpretation of the results by observing how the network weighs the various groups of features using techniques such as the ones described in Sim (2015).

Finally, we observe that a parallel architecture using a noisy feature set is comparable to a feedforward system using a hand-selected feature set. We do not observe a statistically significant difference when comparing the ranks of a feedforward systems with a hand-selected feature set and a parallel system using a standard feature set (z=1.302, *ns*, r=0.184). This is illustrated on the lower right-hand subplot of Figure 8.5. This supports the claim made in Chapter 7 hypothesizing that *hierarchical frameworks are working as denoisers of suprasegmental features.*

8.5 Future work

As future work, the parallel neural network could be the focus of further research. It is unknown at this point whether decoupling the various linguistic levels could be useful. As suggested above, an attempt to visualize the impact of each linguistic level in the networks could be attempted (Sim, 2015). Other lines of research could investigate how these hierarchical networks operate with recurrent systems, in a framework similar to that described in Chen et al. (1998). Alternative acoustic features for the suprasegmental networks were not investigated. An obvious choice would be the use of selected components of a wavelet-based decomposition of the *f0* signal, such as the approach described in Chapter 6.

Finally, based on the work presented in Chapter 7, it would be interesting to investigate how these architectures behave with less noisy additional features. We have attempted to used syllable bag-of-phones and text-derived word embeddings in Chapter 7. Although some improvements in objective measures were seen, we did not observe a preference for the systems using these features. We propose therefore to first investigate additional representations of linguistic contexts that we know are useful for the acoustic model. In Chapter 9, we propose a novel method to learn continuous representations of words and syllables based on acoustic events. Although these features are initially evaluated with frame-level networks, Chapter 10 integrates them into a parallel network architecture.

8.6 Conclusion

Hierarchical systems structured as cascaded or parallel deep neural networks were investigated for decoupling segmental and suprasegmental features in statistical parametric speech synthesis. We observed that, on expressive data, hierarchical systems are preferred over a standard feedforward network if using high-dimensional noisy features. This preference was not observed when using a hand-selected feature set. Hierarchical systems with a standard feature set are comparable to all systems using hand-selected features, which suggests they operate mostly as denoisers or feature extractors. We also observed that a parallel integration of segmental and suprasegmental features is preferred over a cascaded integration. This preference was observed for both feature sets.

Chapter 9

Word vector representations based on acoustic counts

This chapter is an extended version of the work described in "Learning word vector representations based on acoustic counts" (Ribeiro et al., 2017), which was presented at Interspeech 2017.

This chapter presents a simple count-based approach to learning word vector representations by leveraging statistics of co-occurrences between text and speech. This type of representation requires two discrete sequences of units defined across modalities. Two possible methods for the discretization of an acoustic signal are presented, which are then applied to fundamental frequency and energy contours of a transcribed corpus of speech, yielding a sequence of textual objects (e.g. words, syllables) aligned with a sequence of discrete acoustic events. Constructing a matrix recording the co-occurrence of textual objects with acoustic events and reducing its dimensionality with matrix decomposition results in a set of context-independent representations of word types. We observe that the more discretization approaches, acoustic signals, and levels of linguistic analysis are incorporated into a TTS system via these count-based representations, the better that TTS system performs.

9.1 Introduction

In statistical parametric speech synthesis, acoustic parameters are generated by an acoustic model and then used to drive a vocoder in order to obtain an artificially-generated speech waveform. The acoustic model has, in recent years, typically taken the form of a deep neural network (Zen et al., 2013; Qian et al., 2014). The input to this model is often referred to as the *linguistic specification*, which is a representation designed to bridge the gap between text and speech. Common feature sets for English data, such as those described in Appendix A, mostly involve context-dependent phones, syllable stress, word part-of-speech, as well as various positional features describing phonetic and prosodic contexts within a text sentence. The group of modules that processes a text sentence and generate the corresponding linguistic specification is often termed the *front-end* (cf. Section 2.1.1).

However, earlier work in the context of HMM-based speech synthesis found that features defined at linguistic levels above the syllable have little impact on the prediction of acoustic parameters (Cernak et al., 2013). Good representations of higher-level phenomena (often related to syllables or words) are essential for an accurate generation of natural speech prosody, especially in the context of expressive audiobook speech synthesis, where speech is expected to be more fluid and pleasing.

Following these notions, work presented in Chapter 7 investigated a cascaded hierarchical framework with additional features defined at suprasegmental levels. Although improvements were observed in terms of objective scores, a listening test failed to see a preference for systems using the proposed syllable bag-of-phones or text-based word embeddings. Chapter 8 elaborated on those results and found that hierarchical frameworks work mostly as feature extractors for noisy input representations. But it is still unclear whether hierarchical frameworks have the ability to leverage high-dimensional suprasegmental information more efficiently than subsegmental frameworks. To investigate that hypothesis, however, we require a strong set of suprasegmental features that is known to improve subsegmental systems. It is the goal of the current chapter to propose a novel method to learn these high-dimensional representations of linguistic contexts. Considering the segmentation of the main claim of this thesis into the three sub-problems described in Section 1.1, this chapter provides contributions focused on sub-problem 2: *representation of linguistic contexts*.

In this work, we investigate a method to learn acoustically-motivated representations of words and syllables for text-to-speech synthesis. For this task, we explore vector space models (VSM), which are a well-established approach for obtaining semantic representations in the field of Natural Language Processing (NLP). VSMs are rooted in the distributional hypothesis, which claims that words that have similar contexts tend to have similar meanings (Lenci, 2008; Turney et al., 2010). The approaches chosen to learn such representations can be grouped into two main classes. Following the terminology of Baroni et al. (2014), these can be *count models* and *predictive models*.

The first class of models is defined by extracting co-occurrence statistics over large text corpora. Various transformations can be applied to the raw counts, such as context weighting or dimensionality reduction techniques (Bullinaria and Levy, 2007, 2012; Lebret and Collobert, 2015). Conversely, the second class of models frames the problem as a context prediction task. That is, given a word, it is the model's objective to determine the context with which it occurs (i.e. its neighboring words). Therefore, it is expected that words that have similar contexts will be mapped to similar representations in the low-dimensional dense space learned by the model (Turian et al., 2010; Mikolov et al., 2013a; Huang et al., 2012; Pennington et al., 2014). This method has been previously used and discussed in Section 7.4.2. Investigations have been made into these two approaches, comparing them with various configurations on a set of semantic tasks. Although earlier work showed a clear preference for predictive models (Baroni et al., 2014), more recent work has shown that their superiority might not be as obvious (Levy et al., 2015).

In terms of their application to speech synthesis, various approaches have been proposed. Representations learned with count-based methods have been explored as input features to modules within a TTS front-end (e.g. phrase-break prediction (Watts et al., 2011, 2014)), replacement of those modules (e.g. part-of-speech tagging (Watts et al., 2011)), or as direct input for acoustic modeling (Watts, 2012; Lu et al., 2013). With recent developments in neural network architectures, predictive approaches have gained popularity. Recent work investigated representations of words derived from large text databases (Wang et al., 2015a, 2016b) and in combination with acoustic parameters (Wang et al., 2016a; Ijima et al., 2017).

In this chapter, we investigate a simple approach inspired by the traditional class of models based on co-occurrence statistics. Such statistics are extracted over a parallel corpus of text and speech and common transformations are applied to the raw count matrices. In a real-world scenario, these representations could be easily included in the *front-end* of a text-to-speech system as simple look-up tables.

Section 9.2 describes the methodology for learning the proposed count-based word and syllable vector representations. Section 9.3 defines the dataset used, while Section 9.4 details a set of experiments investigating the effect of the learned representations on a text-to-speech acoustic model. A perceptual evaluation is described in Section 9.5, followed by a discussion of the results in Section 9.6.

9.2 Count-based representations

Given a fixed vocabulary V and a fixed set of acoustic classes A, we define a count matrix $M \in \mathbb{R}^{|V| \times |A|}$. M_{ij} denotes the number of times the j^{th} class is observed occurring with the i^{th} vocabulary unit. The vocabulary can be defined over textual objects (e.g. words, syllables). The classes can be defined by discretizing an acoustic signal, such as *f0* or energy. Sections 9.2.1 and 9.2.2 provide details on how these classes are determined.

Because occurrences can be context-dependent, we can extend the set of classes over a unit type[1] to account for neighboring occurrences of the acoustic class. If we set a window of size w, then $M \in \mathbb{R}^{|V| \times w|A|}$.

[1] The term *type* is used to indicate an element in the set of observed textual units, where no repetitions are allowed. The term *token* indicates an instance of a type. For example, in terms of words, the sentence "I saw what I saw" contains 5 tokens and 3 types.

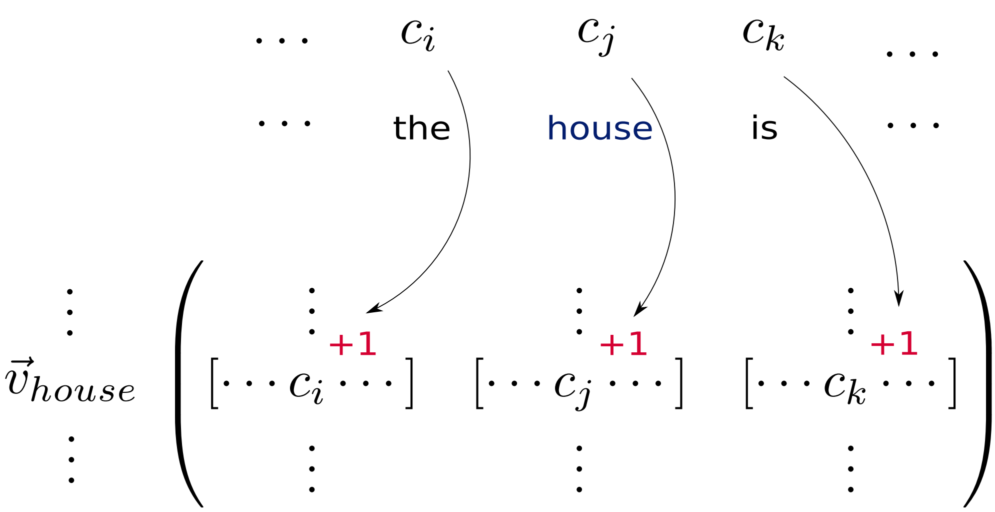

Figure 9.1: The co-occurrence count matrix is populated by taking counts of acoustic classes C and textual objects such as words. This example uses a window of size $w = 3$. For all instances of the word token $house$, we take counts of the co-occurring acoustic event c_j and its neighboring events c_i and c_k. Note that the word tokens co-occurring with $house$ do not participate in the count vectors. Only the co-occurring acoustic tokens. Each of the 3 sub-vectors of $house$ is then normalized by the total number of counts.

For example, consider an utterance for which U is a sequence of linguistic units and C is the corresponding sequence of acoustic classes. If $w = 3$, then at timestep t we count the occurrence of C_{t-1}, C_t, and C_{t+1} in the i^{th} row of M, for which i is the vocabulary index of the unit U_t. Note that, in this case, the tokens U_{t-1} and U_{t+1} are not used for the counts of U_t. This process is illustrated in Figure 9.1.

Each row of the raw count matrix M is then normalized by the total number of counts within each sub-vector of occurrences. Therefore, each sub-vector of the i^{th} row is a probability distribution over the acoustic classes A. Each row consists of the concatenation of w probability distributions.

Finally, we reduce the dimensionality of the normalized count matrix M by finding the Singular Value Decomposition (SVD) of the matrix, such that $M = U\Sigma V^T$. We take k left singular vectors of M, such that the sum of squares of the retained singular values is at least 90% of the sum of squares of all singular values.

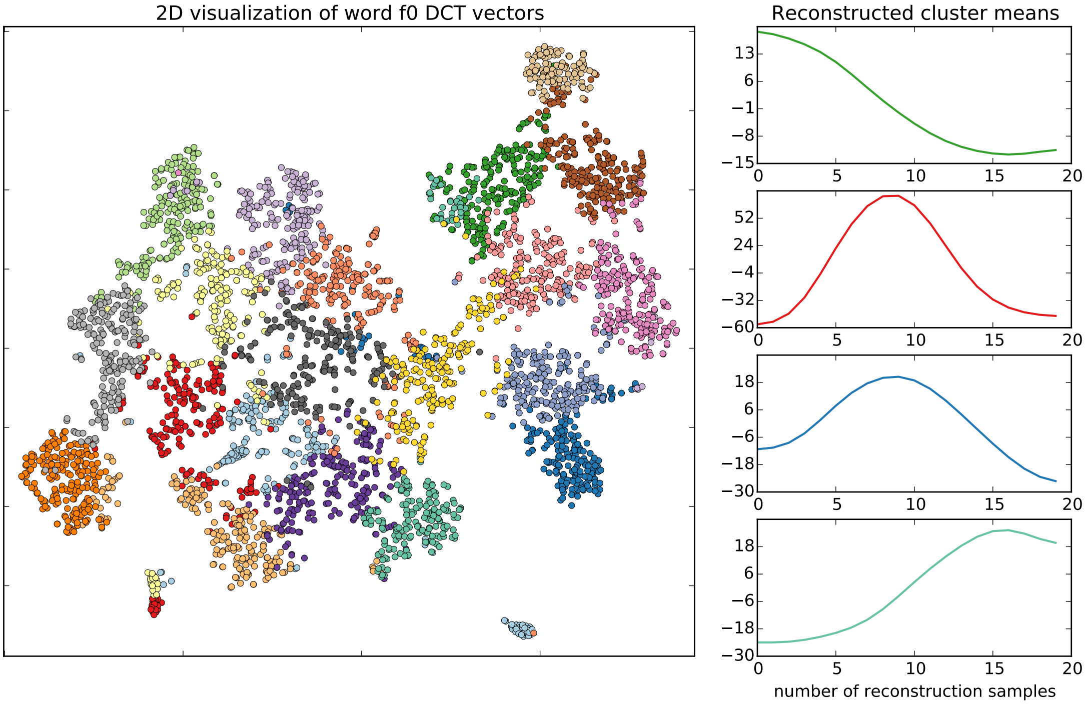

Figure 9.2: On the left, the figure illustrates DCT vectors color-coded by their corresponding cluster, reduced to 2-dimensions with t-SNE (van der Maaten and Hinton, 2008). Each cluster in this illustration contains 200 randomly selected DCT vectors. Axis labels and marks are purposefully not included. On the right, sub-figures illustrate the average of all DCT vectors in four sample clusters, reconstructed with 20 samples with the zeroth coefficient set to 0.

The result of this operation is a matrix $\hat{U} \in \mathbb{R}^{|V| \times k}$. Each row of this matrix corresponds to a entry in the vocabulary V, and we let that be the representation for that linguistic unit.

9.2.1 Cluster-based class definition

This section describes a possible approach to the quantization of an acoustic signal into a set of classes A. We assume we are given a set of linguistic units, corresponding to entries in a vocabulary V, and its acoustic signal, such as *f0*, with known unit boundaries. The boundaries can be found by force-aligning the data at state or phone-level. Higher-level linguistic units can then be inferred via the lower-level units.

Within each utterance, the signal is normalized to zero mean and unit variance. For each unit (e.g. syllable or word), the Discrete Cosine Transform (DCT) (cf. Section 3.2.1) is applied to the samples associated with the corresponding linguistic unit. The first d coefficients are preserved and we let that be a vector representation of the signal for a given unit.

We then use k-means clustering to group the acoustic vectors into classes. For clustering, we exclude the zeroth DCT coefficient, as that is approximately the mean energy of the signal and can heavily bias the clustering step. We can regard the clustered vectors as a representation of the *shape* of the signal for a given linguistic unit. The acoustic classes A are defined to be the clusters identified by k-means. An additional class is added to represent silences such as pauses or hesitations. Figure 9.2 shows a visualization of clustered DCT vectors, as well as the average *shape* for four sample clusters using this method.

9.2.2 Mean-based class definition

The cluster-based representation ignores the mean value of the unit when defining the acoustic classes. Therefore, a simpler approach quantizes the mean value of the signal over the entire linguistic unit. If we consider the *f0* signal, we might observe that a speaker's range is mostly within 100–300Hz, as shown in Figure 9.3. We then define 100 classes over this interval, each spanning a range of 2Hz. Two additional classes are added to include occurrences below and above the interval. An additional class is added to represent silences. Note that there are several hyperparameters required by the two proposed class definitions, such as number of clusters, number of retained DCT coefficients, and bin size. Details of hyperparameter choices are given in Section 9.4.

9.3 Data

We use the data made available for the Blizzard Challenge 2013 (King and Karaiskos, 2013), provided by Lessac Technologies Inc. and originally available

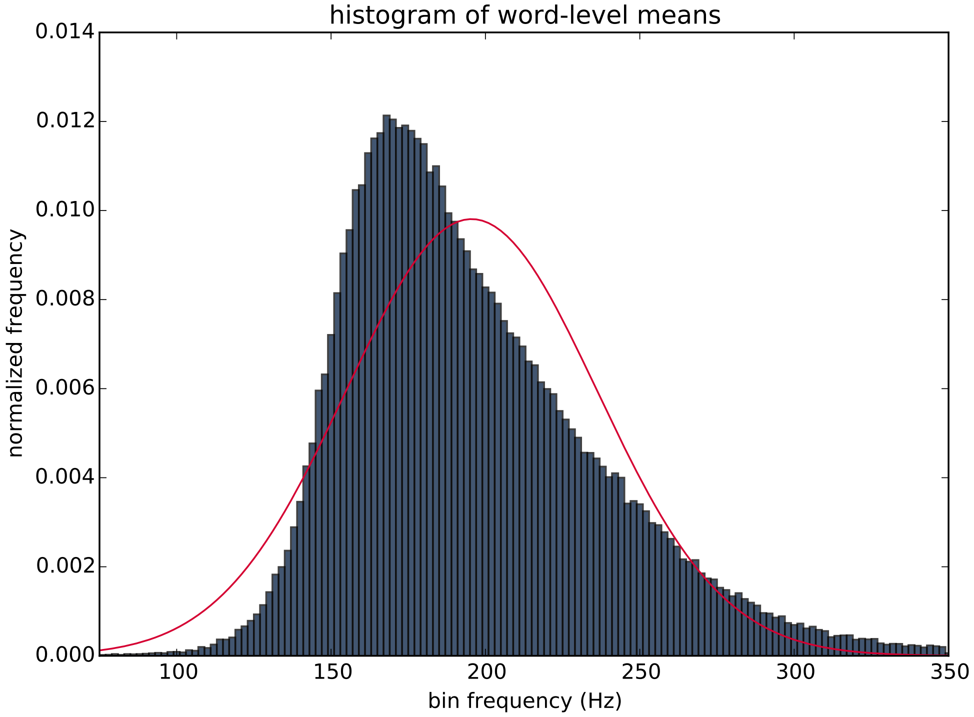

Figure 9.3: Normalized histogram with 2Hz bins of *f0* means at word-level with a best fit line.

from Voice Factory International Inc. The data consists of a single female speaker reading the text of classic novels. This database is of particular interest as the speaker is a professional narrator and actress, which suggests the prosodic variation correlates meaningfully with the text being read. It is also a large dataset, which is interesting for this type of study. However, as the speaker mimics character voices over several books, there is a large variance in terms of speaking styles. Therefore, utterance selection using an active learning approach (Yong et al., 2015; Watts et al., 2013) was performed and a subset of utterances corresponding mostly to narrated speech were selected.

Given the utterance-level segmentation already available from the Blizzard challenge, state-level forced alignment was obtained using context independent HMMs with Festival[2] and HTK[3] via the Merlin toolkit (Wu et al., 2016). Pauses

[2] `http://www.cstr.ed.ac.uk/projects/festival`

[3] `http://htk.eng.cam.ac.uk`

were inserted motivated by acoustic evidence, using Festvox's *ehmm* (Prahallad et al., 2006). The training set consists of approximately 18 hours of speech over 13000 utterances, with approximately 220k word and 300k syllable tokens. We set aside an additional 300 and 100 utterances for validation and test purposes.

9.4 Experiments

To evaluate and gain further insight into the proposed method, we will consider three levels of variation when learning vector representations:

Discretization method This level of variation is focused on the discretization methods described in Section 9.2. We consider a *cluster-based* approach and a *mean-based* approach to define the set of acoustic classes A. While the cluster-based class definition focuses on the *shape* of the signal for a given linguistic unit, the mean-based class definition focuses on the mean value of the linguistic unit.

Linguistic level Most experiments with the acoustic signal learn representations at the word level. This is mainly influenced by earlier work that learns text-based representations by considering words as textual units. However, the proposed method is data-driven and can be applied to any linguistic level. Therefore, we further evaluate our method with textual units defined over syllables. We choose to use syllables because syllabification is given by conventional TTS front-ends and it is readily available at test time.

Acoustic signals Given a linguistic level and a discretization strategy, the proposed approach can be applied to any acoustic signal. We consider two distinct signals: the interpolated *f0* signal and the zeroth mel-cepstral coefficient ($c0$) which we here term *energy*. Following the findings of Chapter 6, we also consider a dynamic wavelet-based decomposition of the *f0* signal.

We begin in Section 9.4.1 by describing the main architecture and hyperparameters used throughout the proposed set of experiments. The same configuration is used for baseline and systems using additional features. Sections 9.4.2

and 9.4.3 are concerned with varying the acoustic signal, *f0* and *energy*, respectively. In Section 9.4.4, we extend the analysis to syllable-level representations. All sections consider the two proposed discretization methods.

9.4.1 Baseline

The baseline system is a simple multilayer perceptron. The network contains 6 hidden layers, each with 1024 nodes. The hidden layers use *tanh* as the activation function and the output layer uses a linear activation function. For training, minibatch size is set to 256 and we set a maximum number of epochs to 25 with 5 warmup epochs. Learning rate is initially set to 0.002 for warmup epochs and after that reduced by 50% with each epoch. Momentum is set to 0.3 for warmup epochs and 0.9 for all others. L2 regularization weight is set to 10^{-5}. Training is done with the Merlin Toolkit (Wu et al., 2016).

For output features, we use *log-f0*, 60-dimensional mel cepstral coefficients (MCCs), and 25 band aperiodicities (BAPs) at 5 ms intervals, extracted using STRAIGHT (Kawahara et al., 1999, 2001). To these features, we append their respective dynamic features (deltas and delta-deltas). The *log-f0* signal is linearly interpolated through unvoiced regions and a binary voiced/unvoiced decision is appended to the acoustic feature vector. The complete output vector has a total of 259 dimensions, which are then normalized to zero mean and unit variance.

Input features to the network are derived from the labels extracted with Festival and they correspond to a set of 592 binary questions defined at phone, syllable, and word levels. These are quinphone identity, syllable stress, and guessed part-of-speech, as well as all positional features. To these questions, we append 2 features indicating frame number relative to phone size and state number. We let this feature set be the input to the baseline system. This is the same feature set used in Chapters 6, 7, and 8. Full details on these features are available in Appendix A.0.2.

Additional models are trained under identical conditions, but append the learned representations to the baseline feature set, using a window of 3 units. That is, if using word-level features, we append the representations for previous, current, and next word. All input features are normalized to the range [0.01, 0.99].

9.4.2 Fundamental Frequency

In this set of experiments, we learn word vector representations using the *f0* signal. We consider a cluster-based (Section 9.2.1) and a mean-based discretization method (Section 9.2.2).

We further consider a transformation of the *f0* signal that is claimed to capture word level variation. We revisit the wavelet-based dynamic decomposition strategy proposed in Chapter 6. The word-level component of this decomposition strategy is extracted and we let this component be the acoustic signal from which we learn our representations.

The training data contains approximately 220k word tokens. The vocabulary is defined by taking all word types that occur at least 5 times. All other words are mapped to a token symbolizing out-of-vocabulary entries, *UNK*. This generates a set of units V with 4468 word types. With this vocabulary, we map 8.7% of the total tokens to *UNK*.

For *cluster-based representations*, we set the number of DCT coefficients to 8, excluding the zeroth coefficient. We then use k-means clustering to map the DCT coefficients at word-level to 20 clusters. An additional class accounts for silence or pause tokens. This gives us the set of acoustic classes A. For *mean-based representations*, we set the *f0* range to be between 100Hz and 300Hz. With a bin size of 2Hz, this gives us 100 acoustic classes.[4] To these we append 2 additional classes for any word-means occurring above or below the range and 1 additional class for silence or pause tokens. Figure 9.2 shows an example of 4 DCT clusters and Figure 9.3 shows the histogram of word means for the *f0* signal.

We set the count window w to be 3, which results in a count vector of size $w|A|$ for each word. Note that silence or pause tokens are not in the vocabulary V, but they are taken into account because $w > 1$. Their counts participate in the neighboring acoustic classes of in-vocabulary units.

[4] We choose to use here linear *f0* rather than log *f0* as we don't expect this transformation to have an impact on the learned representations.

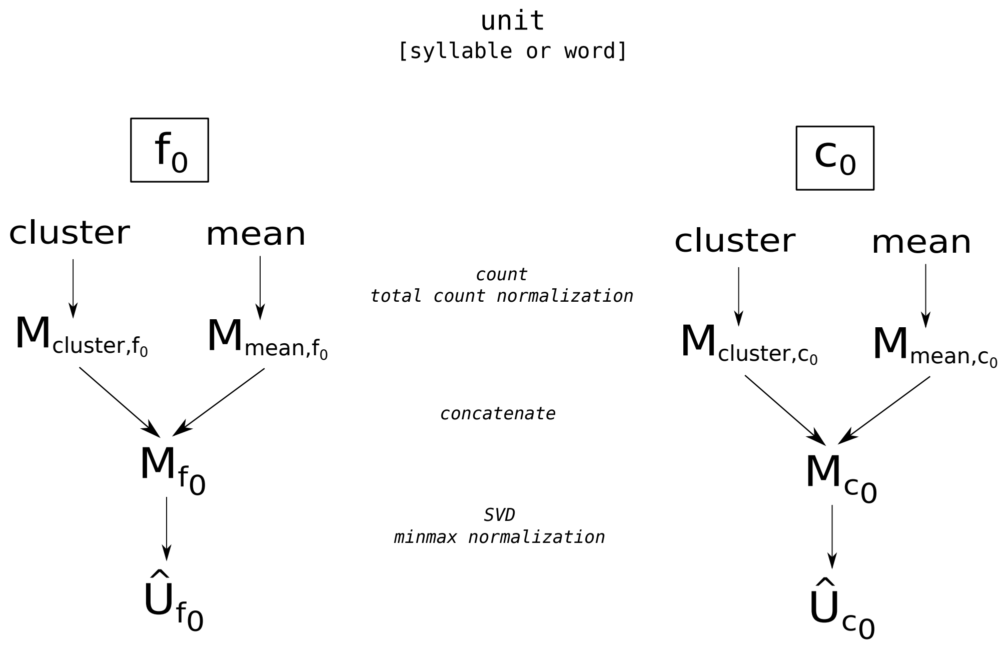

Figure 9.4: Integration of multiple discretization methods, acoustic signals, and linguistic levels.

We also consider representations using both cluster and mean-based class definitions. An illustration of the integration strategy of multiple discretization approaches and multiple acoustic signals is given in Figure 9.4. When considering multiple discretization methods, counts are extracted and normalized separately to form two distinct matrices. These matrices are then concatenated row-wise to form the matrix M_{f_0}. We then find the SVD of this matrix to get the reduced matrix $\hat{U}_{f_0}$. When considering multiple acoustic signals, we treat those as separate decomposed matrices, such as $\hat{U}_{f_0}$ or $\hat{U}_{CWT}$.

Table 9.1 shows the results in terms of objective measures for the trained systems. We observe that all systems using additional features from the proposed vector representations outperform the baseline. In terms of the two proposed discretization methods, the mean-based approach appears to outperform the cluster-based representations. When using the *f0* signal, surprisingly, combining cluster and mean-based vector representations appears to improve MCD, but not *f0* RMSE. This is not the case for the CWT-based representations, which improve over all objective measures when combining discretization methods.

Unit	**Signal**	**Discretization**	**Vector Dim.**	**DNN Input Dim.**	**MCD**	**BAP**	**F0-RMSE**	**F0-CORR**
baseline	-	-	-	594	5.717	2.538	38.137	0.455
word	*f0*	cluster	50	744	5.688	2.531	37.629	**0.472**
word	*f0*	mean	150	1044	5.673	2.520	**37.095**	0.483
word	*f0*	cluster+mean	200	1194	**5.656**	**2.516**	**37.263**	**0.473**
word	wavelet	cluster	50	744	5.684	2.524	38.133	0.458
word	wavelet	mean	100	894	5.700	2.532	38.176	0.460
word	wavelet	cluster+mean	100	894	5.664	2.525	37.716	0.473
word	*f0*+wavelet	cluster+mean	300	1494	**5.643**	**2.515**	**36.878**	**0.481**

Table 9.1: Objective results for count-based representations at the word level using the *f0* signal and the word-level component of a dynamic wavelet-based decomposition strategy. Vector Dim. indicates the dimensionality of the representation on the decomposed count matrices. DNN Input Dim. denotes the dimensionality of the input features to the network, which includes a window of 3 units. MCD is mel-cepstral distortion, BAP is band aperiodicity error, and RMSE and CORR are the root-mean-squared error and correlation between predicted and original *f0* signals on voiced frames only.

We do not observe large differences in terms of objective measures when comparing systems learning representations directly over the *f0* signal or over the word-level component of a wavelet-based decomposition strategy. This is perhaps not surprising, given that the cluster-based method focuses on the *shape* of the signal for a given word. The wavelet transform correlates the signal with the mother wavelet (cf. p. 60), emphasizing the signal's variation associated with the word. It is possible that the application of the DCT followed by clustering (Figure 9.2) ends up representing similar signal contours. We do note, however, that combining both signals leads to the best performing configuration. These improvements might come from the mean-based representations, which, if using the CWT-decomposed *f0* signal, might be related to word prominence (Vainio et al., 2013). The observed results indicate that including additional signals might be worth exploring.

Unit	**Signal**	**Discretization**	**Vector Dim.**	**DNN Input Dim.**	**MCD**	**BAP**	**F0-RMSE**	**F0-CORR**
baseline	-	-	-	594	5.717	2.538	38.137	0.455
word	energy	cluster	50	744	5.692	2.521	38.217	0.452
word	energy	mean	100	894	5.690	2.529	38.211	0.473
word	energy	cluster+mean	150	1044	5.680	2.527	38.245	0.462
word	*f0*+energy	cluster+mean	350	1644	**5.637**	**2.517**	**37.194**	**0.479**

Table 9.2: Objective results for count-based representations at the word level using the zeroth mel-cepstral signal, counting over classes defined over means or clustered vectors. Notation is identical to that of Table 9.1.

9.4.3 Energy

We experiment with the zeroth mel-cepstral coefficient, which may be regarded as a measure of the energy of a speech frame. Table 9.2 details objective results for systems appending representations learned with the zeroth mel-cepstral signal. The same details described in Section 9.4.2 are used for these representations. Cluster-based representations use the same hyperparameters. Mean-based representations use 80 classes over the range $[3, 7]$ with a bin size of 0.05, and we include 3 additional classes. Figure 9.4 illustrates the steps taken when combining discretization methods or acoustic signal. As before, the system combining both discretization approaches concatenates the normalized counts before applying SVD. For the system combining both signals (*f0+energy*), we use SVD to produce two separate matrices $\hat{U}_{f_0}$ and $\hat{U}_{c_0}$. These are treated as separate features and we simply concatenate the learned representations to the features of baseline system using a window of 3 words.

As expected, using the zeroth mel-cepstral signal provides little improvement in terms of *f0*. However, quite surprisingly, it does not outperform representations based on *f0* in terms of mel-cepstral distortion, as seen on Table 9.1. No clear difference is observed in terms of the mean and cluster-based methods, but we again observe slight improvements through their interaction. Combining both signals results in the best improvements of all representations defined at the word level.

Unit	Signal	Discretization	Vector Dim.	DNN Input Dim.	MCD	BAP	F0-RMSE	F0-CORR
baseline	-	-	-	594	5.717	2.538	38.137	0.455
syllable	*f0*	cluster+mean	180	1134	5.686	2.527	38.018	0.476
syllable	energy	cluster+mean	150	1044	5.673	2.52	38.029	0.474
syllable	*f0*+energy	cluster+mean	330	1584	5.645	2.505	37.264	0.483
word	*f0*+energy	cluster+mean	350	1644	5.637	2.517	37.1194	0.479
syllable+word	*f0*+energy	cluster+mean	680	2634	**5.612**	**2.501**	**36.927**	**0.498**

Table 9.3: Objective results for count-based representations at the syllable and word levels using the *f0* and energy signals, counting over classes defined over means or clustered vectors. Notation is identical to that of Table 9.1.

9.4.4 Syllable-level representations

We can easily extend this approach to other types of linguistic units, such as the syllable. We represent syllable types textually as the concatenation of the phones present in a given syllable and we build V by mapping all units with fewer than 5 occurrences over the training data to the unknown token *UNK*. From the approximately 300k tokens, a vocabulary of 3447 unit types is defined. This maps 1.9% of the total tokens to *UNK*. The remaining parameters are similar to those of the word-level representations, except we vary the number of singular vectors kept after SVD such that at least 90% of the singular values are preserved.

For brevity, we do not include all system combinations and we evaluate only representations using both cluster-based and mean-based approaches. Table 9.3 shows objective results for the trained systems. In terms of combination of different counts, discretization methods were concatenated before matrix decomposition. All other levels of variation assume separate matrices, which were then added to the linguistic specification using a window of 3 textual units. The combination methodology can be visualized in Figure 9.4.

We observe some improvements with syllable-level representations, but they do not outperform their equivalent systems at the word level. As before, the interaction of both *f0* and energy representations shows the best results. The system including representations at both syllable and word levels gives the best results of all configurations. Although there might be some correlation between representations, their interaction is still useful to the acoustic model.

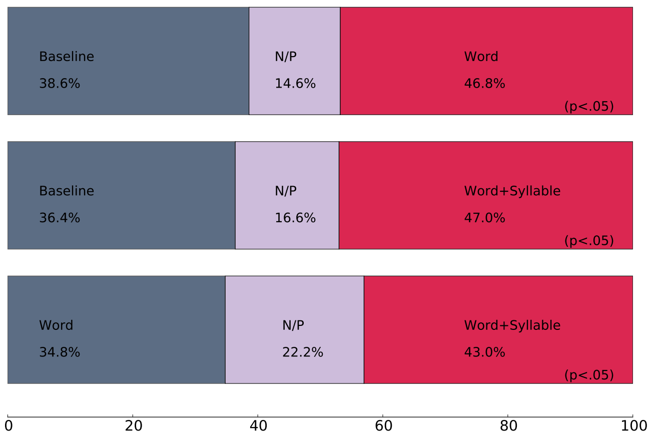

Figure 9.5: Preference test results with N/P indicating "No Preference". In parenthesis, p-values indicate the results of 1-tailed binomial tests with an expected 50% split, with the N/P results evenly distributed over the other two conditions.

9.5 Subjective evaluation

Given the large number of system configurations, we opt to conduct a perceptual evaluation on selected systems, based on the objective results presented in the previous section. Besides the baseline, we consider a system using the best combination of word-level features and the system using the best combination of word and syllable level features. The proposed systems use both acoustic signals (*f0* and energy) and both discretization methods (cluster and mean) for their vector representations. Therefore, we vary only the linguistic level for the listening evaluation.

A preference test with a *no preference* option was conducted on the three selected systems. From the test set, 50 utterances were randomly selected and synthesized with the acoustic parameters generated from each system. 20 native speakers judged randomized utterance pairs for each pair of systems. Each utterance pair was judged 10 times and each condition received a total of 500 judgments. Percentage preferences are shown in Figure 9.5, which includes the results of a 1 tailed-binomial test assuming an expected 50% split, with the *no preference* judgments distributed equally over the other two conditions.

The results are consistent with the objective results presented in Section 9.4.

Systems using the proposed additional features are preferred over the system using no features. The system using multiple linguistic levels (e.g. words, syllables) is preferred over using only representations learned at word level.

9.6 Discussion

In the previous section, we experimented with three main factors pertaining to how the representations were learned: *discretization method* (e.g. cluster, mean), *acoustic signal* (e.g. *f0*, energy), and *linguistic level* (e.g. words, syllables). Figure 9.6 visualizes the trained systems in terms of objective measurements. We selected mel-cepstral distortion (MCD) and *f0* root-mean-squared-error (RMSE) as these are related to the two acoustic signals used while learning the representations. The two axes are determined by the difference over the baseline. Farther from 0 in the positive direction indicates an improvement.

In terms of the first factor, we observe that the mean-based approach outperforms the cluster-based approach, but combining both methods provides better results than either method separately. This pattern is also observed when varying the acoustic signal: *f0* is shown to have a stronger impact on the objective results than energy, but the interaction of both signals gives the best results. In terms of linguistic units, we observe that word representations outperform syllable representations when using *f0*, but not when using energy. Combining both levels appears more useful than either level in isolation. We further observe that the system using representations learned over word *f0* means performs quite well when compared to all other systems in terms of *f0*-RMSE. Improvements in terms of MCD mostly originate from the inclusion of additional information. In general, we observe that the more information is incorporated into the *linguistic specification* of a TTS system using the proposed representations, the better that TTS system performs.

However, as we provide the acoustic model with more information, we also increase the complexity of the acoustic model. While the baseline system only needs to process a vector of 594 dimensions, the best performing systems uses a vector of 2634 dimensions. This increases training and testing time, which might

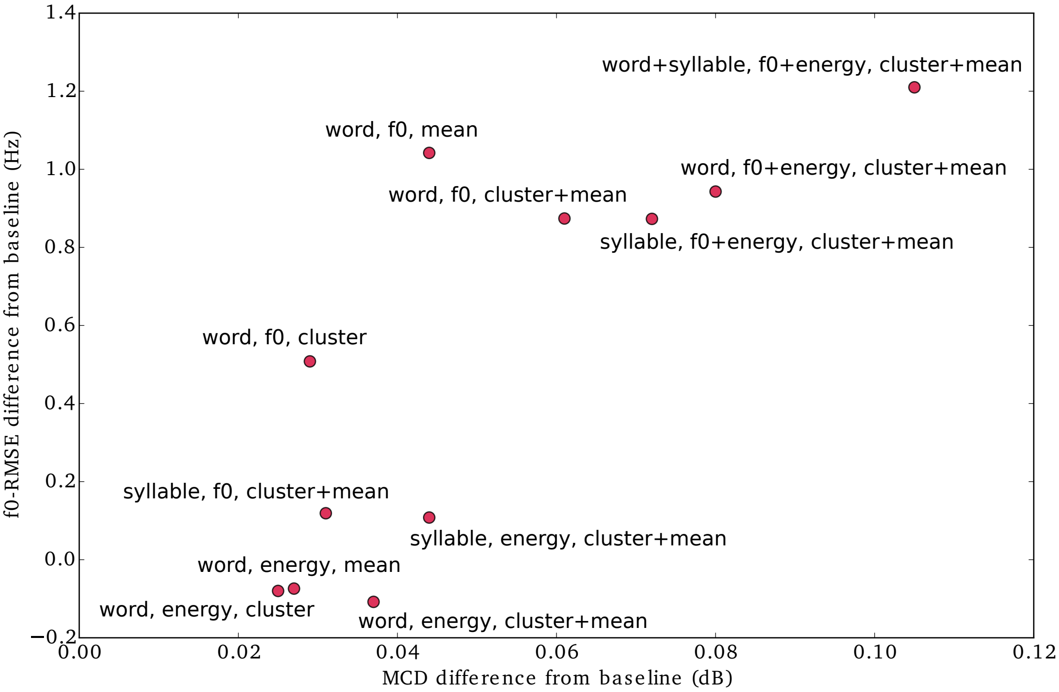

Figure 9.6: Visualization of trained systems in terms of objective measurements. Each axis denotes the difference between a system using learned vector representations and the baseline in terms of an objective measure. Horizontal axes shows difference for MCD (mel-cepstral distortion) and vertical axis the difference for *f0* RMSE (root-mean-squared-error) on voiced frames.

be an issue in a commercial setting.

In any case, the proposed approach is still attractive as it is a fairly cheap way to learn suprasegmental representations that are dependent on the data used for acoustic modeling. This method could be useful, for example, to systems that are limited by weaker front-ends. Further investigation could consider replacing features such as part-of-speech tags or syllable stress. Similar scenarios have been previously investigated with text-derived word embeddings (Wang et al., 2015a).

Corpus size is also a relevant factor when learning word vector representations. In this work, we have used a fairly large text-to-speech database. It is unknown how useful these representations would be on a smaller corpus. In a similar line of research, work focusing on the speaker dependency of these representations would be very interesting. For example, such analysis could learn count-based vector representations on a larger dataset and apply them on an acoustic model trained on a different and smaller dataset. These types of investigations could be performed together with a comparison of count-based and predicted-based approaches for multi-modal embeddings. For example, evaluating our proposed methods with the proposal of Ijima et al. (2017).[5]

With respect to the current methodology, it should be noted that no optimization of hyperparameters was attempted. No tuning was performed, for example, on the number of clusters or the number of bins for each discretization method. It was surprising to observe such improvements with fairly arbitrary initial choices of setting for these hyperparameters. Further improvements might be observed with careful optimization.

In terms of discretization methods, we might consider earlier approaches for the *f0* signal, such as Tilt (Taylor, 1998), MoMel (Hirst et al., 2000), ProsoGram (Mertens, 2004), or SLAM (Obin et al., 2014). Acoustic signals such as jitter and shimmer might also be useful in the context of a TTS system. Learning representations at multiple levels might be useful, as recent work showed that using continuous representations of phrases as input to acoustic models for text-to-speech synthesis can lead to improved quality in synthetic speech (Wang et al., 2016b).

9.7 Conclusion

This chapter presented a novel method for learning vector representations by leveraging statistics of co-occurrences between text and speech. The proposed approach requires two discrete sequences of units across modalities and two discretization methodologies are described. This count-based method for learning representations is data-driven and can be applied to any linguistic unit or acous-

[5] The work of Ijima et al. (2017) uses a very large database (700 hours over 5372 speakers) with a word-level bi-directional LSTM-based architecture. The latent representation learned by a bottleneck layer in the model is treated as the word vector representation, which is then used with the main acoustic model. Although a different method than the one we propose in this chapter, which learns context-independent vectors, the approach Ijima et al. (2017) could be a starting point for the comparison of different techniques.

tic signal. In the set of experiments presented, we have evaluated the proposed method with two discretization approaches (e.g. cluster-based and mean-based), two acoustic signals (fundamental frequency and the zeroth mel-cepstral coefficient), and two linguistic levels (syllables and words). Improvements were observed in terms of objective measures across all systems using learned vector representations. It was also observed that the more information is injected into the acoustic model via the learned representations, the better that acoustic model performs. A subjective evaluation showed a preference for systems using additional features learned using the proposed count-based method over a system not using them.

Considering the three sub-problems of the main thesis claim described in Chapter 1, we have proposed contributions to *sub-problem 2: representation of linguistic contexts*. Chapter 7 provided an initial investigation under this topic. Syllable bag-of-phones and text-derived word embeddings were proposed and evaluated with subsegmental and hierarchical frameworks. It was shown that these representations could be useful with a hierarchical architecture, but not with a subsegmental model. The current chapter has extended the investigation into linguistic context representations and proposed a novel way to learn vector representations of suprasegmental units over a large database.

Chapter 10

Conclusions and future work

The final chapter of this thesis provides a summary of its main contributions. This is followed by a global evaluation of those contributions and a discussion with respect to the main claim of the thesis. The chapter concludes with a proposal of lines for future research focusing on the natural generation of speech prosody.

10.1 Overview of main contributions

This thesis aims to provide evidence that more appropriate suprasegmental representations lead to more natural generation of fundamental frequency in statistical parametric speech synthesis. The claim is motivated primarily by a set of observations about traditional approaches to text-to-speech. In terms of acoustic signals, fundamental frequency is represented at the frame level, but affected by a variety of prosodic phenomena associated with multiple linguistic levels. For example, *f0* is affected by segment identity and co-articulation at lower levels and lexical stress, pitch accents, and boundary tones at higher levels. Linguistic representations are often positional or predicted by separate modules trained on potentially out-of-domain data. Finally, traditional acoustic models are defined over sub-segmental intervals such as states or frames.

A segmentation of the main claim into clearer sub-problems was proposed in Section 1.1. These are: representations of acoustic signals, representation of linguistic contexts, and the mapping between acoustic and context representations.

The relationship of each chapter of this thesis to each sub-problem was illustrated in Figures 1.1 and 1.2 (p. 6 and 9). With this in mind, the overall contributions of this thesis and its corresponding chapters can be summarized in the following points:

- A representation of *f0* that is defined at multiple linguistic levels using the continuous wavelet transform and the discrete cosine transform. [**Chapter 4**].
- A stronger understanding of the role of the various frequencies in wavelet-based decomposition and their relation to the perception of naturalness in synthetic speech. [**Chapter 5**]
- A perceptually and linguistically motivated decomposition of *f0* using the continuous wavelet transform. [**Chapter 6**].
- An investigation of cascaded deep neural networks with additional features defined at a suprasegmental level. [**Chapter 7**].
- Insights into the role of hierarchical deep neural networks using cascaded and parallel approaches with a variety of features defined at various linguistic levels. [**Chapter 8**].
- A data-driven method to learn vector representations of linguistic units such as syllables or words by taking counts over acoustic events. [**Chapter 9**].

Each of these contributions deals with one or two of the proposed sub-problems. Although there is some level of interaction between these sub-problems, illustrated by Figure 1.1, there is no direct comparison of each presented technique. The following section therefore describes a final evaluation of the main contributions of this thesis.

10.2 Global evaluation

In order to conduct a global evaluation, we have selected one main contribution to each of the three sub-problems. These main contributions are also illustrated as the leaves of the hierarchical diagram on page 9. These are the main findings described in Chapters 6, 8, and 9.

Chapter 6 proposed a representation of *f0* using the continuous wavelet transform that was then used as a secondary task with a feedforward neural network. This approach is related to the first sub-problem of the main claim – *representations of acoustic signals*.

Chapter 9 proposed a method to learn syllable and word vector representations by counting events from acoustic signals. This approach learns a matrix encoding a representation of linguistic units and can be used as input to an acoustic model. This work falls under the second sub-problem of the main claim – *representations of linguistic contexts*.

Chapter 8 investigated a hierarchical approach using parallel deep neural networks, with each side of the network focusing on separate linguistic levels. This work is focused on the third sub-problem – *the mapping between input and output representations*. It was shown that this architecture has the ability to denoise input features, but it is not known if it can leverage new features more efficiently than a single feedforward neural network.

10.2.1 Systems trained

Table 10.1 describes the main systems and how they vary with respect to the main sub-problems. It proposes a final analysis of systems with additions at various levels: input, output, and acoustic model architecture. Two combinations were omitted, as they were either already investigated or they were not expected to be meaningful. The first would be a simple hierarchical parallel network. This was investigated in Chapter 8 and it was observed that this architecture performs similarly to a feedforward network if using a pruned set of linguistic features. The second omitted combination would be a system using the wavelet-based decomposition of Chapter 6 with a parallel network. We would expect this configuration to perform similarly to the system using the decomposition strategy as a secondary task, given the denoising properties of the parallel architecture.

For this evaluation, we have used the dataset described in Section 9.3 (p.

System	Input	Architecture	Output
baseline	DNN-standard	feedforward	MGC, f0, BAP, VUV
vectors	DNN-standard, word and syllable vectors	feedforward	MGC, f0, BAP, VUV
CWT	DNN-standard	feedforward	MGC, f0, BAP, VUV, CWT-syl
vectors, CWT	DNN-standard, word and syllable vectors	feedforward	MGC, f0, BAP, VUV, CWT-syl
vectors, parallel	DNN-standard, word and syllable vectors	parallel feedforward	MGC, f0, BAP, VUV
vectors, parallel, CWT	DNN-standard, word and syllable vectors	parallel feedforward	MGC, f0, BAP, VUV, CWT-syl

Table 10.1: Systems evaluating combinations of main contributions.

178). Input features denoted *DNN-standard* in Table 10.1 correspond to a conventional feature set used for DNN-based speech synthesis, which is detailed in Appendix A.0.2. A *feedforward* architecture corresponds to a 6 hidden layer network with 1024 nodes per layer. Output acoustic features denoted *MGC, f0, BAP, VUV* correspond to the standard mel-cepstral coefficients, fundamental frequency, band aperiodicities, and voice/unvoiced decision extracted with the STRAIGHT vocoder (Kawahara et al., 1999, 2001).

Each additional contribution generally corresponds to the best performing configuration in its respective chapter. Additional input representations (denoted *vectors* in Table 10.1) are learned over words and syllables and consider the *f0* and energy signals with the two proposed discretization methods of Chapter 9. Additional output acoustic representations (*CWT-syl* in Table 10.1) considers the syllable-level component of the dynamic wavelet-based decomposition strategy as a secondary task, according to Chapter 6. The parallel architecture is identical to that of Chapter 8, using triangular pre-processing networks at segmental and suprasegmental levels, which are integrated via a 1-hidden-layer network. When integrating multiple contributions with a parallel architecture (system *vectors, parallel, CWT*), word vector input representations are used with the suprasegmental network and wavelet-based output representations are used as the secondary task of the 1-hidden-layer network combining latent representations.

10.2.2 Subjective results

We assess the differences between systems using a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor, ITU-R Recommendation BS. 1534-1 (2015)) evaluation. Participants were asked to rate samples generated from the six systems of Table 10.1 side-by-side, along with a seventh reference sample of matching vocoded speech. Each sample was rated with respect to the reference, but also against one another, along a 100-point scale where 0 denotes completely unnatural and 100 denotes completely natural.

A total of 16 native English speakers participated in the evaluation. The test was conducted in a sound-insulated booth with headsets and all participants were remunerated for their effort. Each listener rated a total of 20 sets of samples, thus providing us with a total of 320 parallel comparisons. One test participant, inadvertently or not, scored all samples at zero in 9 out of the 20 observed sets. For this reason, the remaining judgments given by this participant were excluded from the analysis. Additionally, 6 sets from two other participants failed to correctly identify the hidden reference. These sets were also excluded from the analysis. This allowed us to consider a total of 294 sets of stimuli over 15 participants.

Figure 10.1 illustrates the distributions of the rated stimuli in terms of the absolute score given by the test participants. Figure 10.2 illustrates the same data in terms of system rank order, derived from the 100-point scale provided by test participants.

10.2.3 Discussion

To understand the differences between the systems, we conduct a two-tailed paired t-test on the absolute distributions over the 100-point scale illustrated in Figure 10.1. All results are adjusted for multiple comparisons with a Holm-Bonferroni correction. We observe that all systems using proposed contributions significantly outperform the baseline at the level of $p<.05$. No significant differences are observed when comparing systems using additional techniques.

We further analyze the results in terms of their rank order by conducting a double-sided Wilcoxon signed-rank test with a Holm-Bonferroni correction. All systems using proposed contributions outperform the baseline at the level of $p<.05$, except *vectors-parallel*. As before, non-baseline systems do not show significant differences in terms of their rank order when compared against one

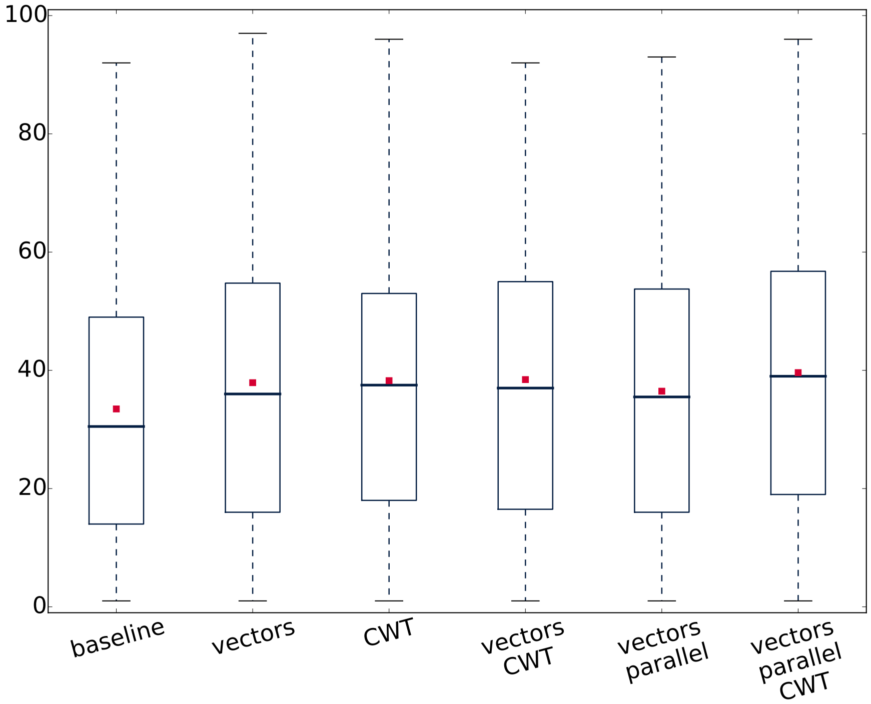

Figure 10.1: Absolute results from the MUSHRA evaluation. A dark blue horizontal line indicates the median and a red square indicates mean. The hidden reference stimulus is omitted from the visualization, as it was always ranked as completely natural.

another.

It was surprising to observe a non-significant difference between the system using count based representations with a hierarchical framework (*vectors-parallel*) and the baseline. This was observed in the rank-order analysis, but not in the absolute score analysis, although the Wilcoxon signed-rank test is more conservative than the pairwise t-test. A similar observation was made in the analysis conducted in Section 8.4. Given the findings detailed in Chapter 8, the parallel architecture with additional features (*vectors-parallel*) was expected to perform similarly to the corresponding non-hierarchical system (*vectors*). Although Figure 10.2 shows a slightly lower rank for system *vectors-parallel*, this system's score is not significantly different from those of the remaining systems using contributions proposed by this thesis.

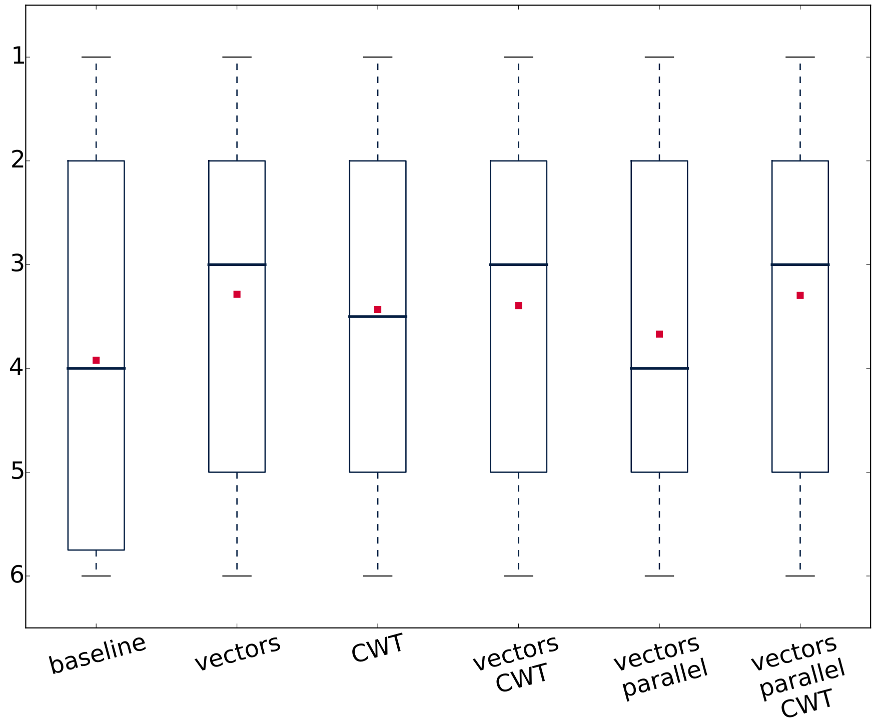

Figure 10.2: Rank order results from the MUSHRA evaluation. A dark blue horizontal line indicates the median and a red square indicates mean. The hidden reference stimulus is omitted from the visualization, as it was always ranked as completely natural.

The main observations indicate that additional input and output representations (system *vectors* and *CWT*) appear to provide the stronger results. Unexpectedly, however, their interaction (system *vector-CWT*) does not perform better than either of them separately. It should be noted that the integration of these methods followed a simple approach, incorporating them together via feedforward neural networks. It is possible that different architectures may be required to leverage the full potential of stronger representations of both inputs and outputs. This particular integration method was not directly investigated by the work presented in this thesis. For example, the integration of the proposed input and output representations with stronger hierarchical models might lead to interesting results. It was mentioned in Chapter 8 that the main advantage of a hierarchical parallel architecture was its ability to denoise high-dimensional representations. Since this analysis, various promising hierarchical frameworks have been proposed (Wang et al., 2017a; Ronanki et al., 2017), which could provide a starting point for stronger hierarchical modeling with additional input and output representations.

With respect to the main thesis claim, we observe that *techniques exploring suprasegmental representations lead to more natural generation of speech parameters.* This was shown through various listening tests in this thesis and validated with a global evaluation.[1] However, although this work provides interesting methods and insights into the main problem, it also generates further questions and ideas. In the following section, we detail possible next steps for the contributions and research topics covered by this thesis.

10.3 Future work

In this section, we suggest directions for future research focusing on each of the three sub-problems governing the main claim of this thesis. Specific details for additional work are given in the discussion of each contributing chapter. Those are directly related to the hypotheses and results described with each contribution. The ideas proposed here instead aim to go beyond the scope of this thesis and generalize to the overall problem of natural generation of speech prosody in text-to-speech synthesis.

Sub-problem 1*: representations of acoustic signals*

The notion that prosodic information is embedded in the speech signal hierarchically was discussed in Section 3. This describes the signal as manifesting effects that can be associated with multiple temporal domains, conceptualized as *small ripples on top of big waves* (Wu et al., 2008).

[1] We purposefully avoid generalizing these observations to a comparison of the three sub-problems: input representations, output representations, or acoustic model. The global evaluation conducted in this chapter is limited to the techniques proposed in the scope of this thesis. It should be clear that the observed results are a consequence of the proposed methods and not evidence to the importance of each sub-problem to the main claim.

This thesis made no claims or assumptions regarding a prosodic constituent hierarchy and it is limited to considering the *f0* signal. Future work could combine the findings of Chapters 4, 5 and 6 with recent work using the Continuous Wavelet Transform (Vainio et al., 2013; Vainio, 2014; Vainio et al., 2015; Suni et al., 2017).

The work of Suni et al. (2017) is of particular relevance to this sub-problem. The authors have used a combined signal of *f0*, intensity, and duration with the CWT to infer prosodic prominence and constituent structure. Results indicate that their approach is comparable to supervised methods for the annotation of such structures. This annotation could be used directly with an acoustic model (Tesser et al., 2013) or used to learn continuous representations of context, such as those proposed in Chapter 9.

This thesis was limited to constituents that are easily inferred from textual sources (e.g, words or syllables) or that are readily available in common text-to-speech front-ends (e.g. predicted phrase breaks). These lines of future work therefore aim to go beyond those limitations and focus on a hierarchy that is grounded on established theories of prosodic structure as well as on multiple acoustic correlates of speech prosody.

Sub-problem 2*: representation of linguistic contexts*

Stronger data-driven techniques that infer complex representations of linguistic contexts would be beneficial for text-to-speech systems. This is motivated by what has been called the *lack of reference problem* (Xu, 2012).

In Chapter 9, we have investigated a count-based method to learn vector representations. Recent work focused instead on predictive-based methods such as the skip-gram model (Mikolov et al., 2013a). Various studies have proposed multimodal variations of these models that have been successfully applied in speech and vision (Lazaridou et al., 2015). Alternatives such as deep canonical correlation analysis (DCCA) for representation learning also might provide interesting results (Andrew et al., 2013; Wang et al., 2015b). Similarly, the sequence-to-sequence LSTM based autoencoder described in

Wan et al. (2017) to learn phone-level embeddings might be interesting to explore at higher linguistic levels.

Other approaches could combine the method proposed in Chapter 9 with a dependency parser, for example. This would be a multi-modal approach related to the text-centered dependency-based word embeddings of Levy and Goldberg (2014). Rather than considering sequential context, this method considers textual units (e.g. words) that are adjacent on a dependency graph.

Additionally, context representations could be explored with longer units within the utterance, such as phrases. Chapter 9 focused on syllables and words, but continuous representations of phrases have shown to be beneficial for speech synthesis (Wang et al., 2016b).

For future work in the context of sub-problem 1, we have described the potential benefits of inferring a prosodic constituent hierarchy from the acoustic signal. These findings could be used to learn vector representations of those constituents. Recent work with recursive models have successfully learned representations of hierarchical structures (Socher et al., 2011; Socher, 2014).

Ultimately, context representations for text-to-speech should be directed at event longer temporal units, spanning multiple sentences. Although this thesis was limited to sentential phenomena, speech prosody is known to be influenced by discourse-level effects (Grosz and Hirschberg, 1992; Hirschberg, 1993; Sluijter and Terken, 1993; Swerts and Geluykens, 1994; Nakatani et al., 1995; Wichmann, 2000; Wennerstrom, 2001; Smith, 2004; Tyler, 2013). When using expressive databases such as audiobooks, representations inferred with larger contexts could be able to capture complex interactions of lower-level units. These might capture effects related, for example, to information structure (Steedman, 2000; Calhoun, 2007).

***Sub-problem 3**: mapping acoustic and context representations*

It is widely agreed that prosody is suprasegmental and hierarchical in nature. However, most text-to-speech systems still operate on very short-term intervals. In Chapters 7 and 8, we investigated cascaded and parallel deep neural networks, in which two networks process segmental and suprasegmental layers separately. This is a simple two-tier approach that collapses the rich higher-level prosodic hierarchy into a single suprasegmental layer. Ideally, the global prosodic constituent structure would be processed at multiple linguistic levels.

Similarly, in our method, segmental and suprasegmental networks were optimized separately. Because the multiple layers of meaning manifested through the acoustic correlates of prosody are embedded into the same speech signal, jointly optimizing all model networks might be beneficial.

Recent hierarchical models have already begun to steer in this direction, jointly optimizing multiple recurrent linguistic levels (Ronanki et al., 2017; Wang et al., 2017a). Such models could be linked with additional representations of context and stronger representations of prosodic structure. However, these methods are still limited to individual sentences. Extending recurrences over multiple sentences might be useful. For example, in a sample-level approach such as Wavenet (van den Oord et al., 2016), context is introduced into the model via dilated convolutions. An extension of this idea to parametric speech synthesis could be attempted, where multiple linguistic levels leverage discourse-level context to jointly generate prosodic acoustic parameters.

Similarly, instead of operating on clearly defined linguistic levels (such as syllables or words), a hierarchical architecture defined over prosodic constituents might be worth exploring. Such work could combine the signal-driven prosodic annotation proposed in Suni et al. (2017) with a hierarchical model similar to that of Ronanki et al. (2017).

Besides the three sub-problems that stem from the main claim of this thesis, there are parallel areas of research that deserve additional notes. One such case relates to the evaluation protocols used for text-to-speech synthesis. For systems and hypotheses focusing on the natural generation of speech prosody, evaluation methods need to go beyond the currently established methodologies discussed in

Section 2.3.

An initial step has already been taken in this direction with the work of Latorre et al. (2014). The authors test the hypothesis that, given a speech sample without proper context, the mental reference of the listeners for the prosodic layer might vary. This potentially invalidates non-referenced evaluation protocols such as MOS or AB tests for prosody evaluation.

Future work could therefore investigate evaluation protocols that account for supra-sentential context. These might be referenced or non-referenced. It is unknown, for example, how context affects listeners' responses. The type of context (e.g. text, audio) or size of context (e.g. number of sentences or paragraphs) might generate different responses. Further work investigating these questions might be useful for speech synthesis. Alternatively, Winters and Pisoni (2004) argue that the naturalness of synthetic speech in terms of prosody might need to be evaluated indirectly. This would involve measuring the listeners' response in terms of memory, attention, or cognitive load.

Although this thesis aimed to isolate hypotheses to clearly identify differences between systems, some of the effects were not as strong as expected. This might be due to evaluation methodologies focusing on out-of-context sentences. In the listening tests conducted during this thesis, various listeners have informally mentioned the difficulty of evaluating sentences that differ primarily in terms of intonation. A protocol that can accurately evaluate synthetic speech in terms of prosody would be beneficial for future work focusing on the generation of prosodic phenomena.

10.4 Final remarks

This thesis provided evidence that stronger suprasegmental modeling of fundamental frequency is essential for more natural generation of speech prosody. This evidence was given through contributions across three sub-problems of this main claim: suprasegmental representation of acoustic signals, suprasegmental representations of linguistic contexts, and suprasegmental and hierarchical acoustic modeling.

Natural generation and control of speech prosody remains one of the fundamental problems in text-to-speech synthesis. A stronger understanding of speech phenomena is indispensable if researchers hope to bridge the gap between natural and synthetic speech. This thesis has sought to provide insights into this problem and hopefully foster the discussion of topics that could eventually lead to more natural speech synthesis.

Appendix A

Linguistic Features

This appendix provides a full description of the linguistic features used in this work. Section A.0.1 includes a description of the full set of features used for the HMM-based systems in Chapters 4 and 5. Section A.0.2 provides a description of the subset of features used in the DNN-based systems presented in the remaining chapters.

A.0.1 Feature set used with HMM-based systems

This is a typical set of features for the English language used for HMM-based speech synthesis (Tokuda et al., 2013). The set is defined by 2926 binary questions extracted from the features defined below. The binary questions are used for decision tree clustering within the HTS framework. Further details can be found in the question set distributed with HTS English recipes (Zen et al., 2007, 2009a).

Features defined at the level of the *phone*

- current phone.
- previous and next phones.
- previous and next phones at a distance of 2 from current phone.

The concatenation of the 5 phones is termed the *quinphone*. For each phone in the quinphone, various features are defined. These are related to the phone identity and the phone's articulatory characteristics (place and manner of articulation).

Features defined at the level of the ***syllable***

- number of phones since/until the end of the syllable.
- lexical stress of the previous, current, and next syllable.
- number of phones in the previous, current, and next syllable.
- number of syllables since/until a lexically stressed syllable.
- number of syllables since/until a pitch accented syllable.
- previous, current, and next syllable contains a pitch accent.
- number of syllables since/until a word boundary.
- number of syllables since/until a phrase boundary.
- number of stressed syllables since/until a phrase boundary.
- number of pitch accented syllables since/until a phrase boundary.
- identity and articulatory features of the nucleus of the current syllable.

Features defined at the level of the ***word***

- guessed part of speech of the previous, current, and next word.
- number of syllables in the previous, current, and next word.
- number of words since/until a phrase boundary.
- number of content words since/until a phrase boundary.
- number of words since/until a content word.

Features defined at the level of the ***phrase***

- number of syllables in the previous, current, and next phrase.
- number of words in the previous, current, and next phrase.
- number of phrases since/until an utterance boundary.
- predicted ToBI endtone for current phrase.

Features defined at the level of the ***utterance***

- number of syllables in the utterance.
- number of words in the utterance.
- number of phrases in the utterance.

A.0.2 Feature set used with DNN-based systems

This is a binary *hand-selected* feature set distributed with earlier versions of the Merlin Neural Network Toolkit (Wu et al., 2016). The questions are a subset of the features available with HTS and described in Section A.0.1. This subset reduces the 2926 HTS questions to 592 binary questions based on the features defined below.

The current subset was hand-selected after experimentation aimed at the optimization of objective measures on a British English dataset.[1] Various systems have since been presented using this convention (Wu et al., 2015; Wu and King, 2016; Henter et al., 2016; Ronanki et al., 2016, 2017). For the release of the Merlin Neural Network Toolkit (Wu et al., 2016), the feature set was further optimized. For example, by allowing numerical features (e.g. positional or count-based) to be continuous rather than binary. Binarization was inherited from the HTS question set, which required binary features for decision tree clustering. Except where noted, all DNN-based systems reported in this thesis use the early hand-selected question set of 592 binary features.

Features defined at the level of the *phone*

- identity and articulatory characteristics of current phone.
- identity of previous and next phones.
- identity of previous and next phones at a distance of 2 from current phone.

Features defined at the level of the *syllable*

- number of phones since/until the end of the syllable.
- lexical stress of the previous, current, and next syllable.
- number of phones in the current syllable.
- number of syllables since/until a lexically stressed syllable.
- number of syllables since/until a pitch accented syllable.
- previous, current, and next syllable contains a pitch accent.
- number of syllables since/until a word boundary.

[1] This feature selection work was carried out during early development of the Merlin Toolkit by Zhizheng Wu, to whom we are grateful.

- identity and articulatory features of the nucleus of the current syllable.

Features defined at the level of the *word*

- guessed part of speech of the previous, current, and next word.
- number of syllables in the current word.
- number of words since/until a phrase boundary.
- number of content words since/until a phrase boundary.
- number of words since/until a content word.

Appendix B

Additional materials for chapter 5

This appendix provides additional material for the perceptual investigation of a wavelet-based decomposition of *f0* described in Chapter 5

	1-2	**3-4**	**1-4**	**5-6**	**7-8**	**5-8**	**9-10**	**hmm**
all	p<.001	p<.001	p<.001	p<.05	p<.001	p<.001	p<.001	p<.001
1-2	-	ns	ns	p<.001	ns	p<.001	p<.001	ns
3-4		-	ns	p<.001	p<.05	p<.01	p<.001	ns
1-4			-	p<.001	ns	p<.001	p<.001	ns
5-6				-	p<.001	ns	p<.001	p<.001
7-8					-	p<.001	p<.001	ns
5-8						-	p<.001	p<.01
9-10							-	p<.001

Table B.1: Bonferroni-corrected pairwise Wilcoxon sign rank test for the MUSHRA evaluation (Section 5.5).

	all	**1-2**	**3-4**	**1-4**	**5-6**	**7-8**	**5-8**	**9-10**	**hmm**
natural	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001	p<.001
all	-	p<.001	p<.001	p<.01	ns	p<.001	ns	p<.001	p<.001
1-2		-	ns	ns	p<.01	ns	p<.01	ns	ns
3-4			-	ns	ns	ns	ns	p<.01	ns
1-4				-	ns	ns	ns	p<.01	ns
5-6					-	ns	ns	p<.001	p<.05
7-8						-	ns	p<.05	ns
5-8							-	p<.001	p=.05
9-10								-	ns

Table B.2: Bonferroni-corrected pairwise Wilcoxon sign rank test for the Mean-Opinion-Score evaluation (Section 5.6).

	Context 1	Context 2	Context 3	Context 4
stimulus	...	Paul won at Mary’s?	John lost at Mary’s?	John won at Kate’s?
response	John won at Mary’s.	John won at Mary’s.	John won at Mary’s.	John won at Mary’s.
stimulus	...	Michael played the drums?	David brought the drums?	David played the guitar?
response	David played the drums.	David played the drums.	David played the drums.	David played the drums.
stimulus	...	Richard took the train?	Bill missed the train?	Bill took the bus?
response	Bill took the train.	Bill took the train.	Bill took the train.	Bill took the train.
stimulus	...	Peter lives in Paris?	Mike visited Paris?	Mike lives in London?
response	Mike lives in Paris.	Mike lives in Paris.	Mike lives in Paris.	Mike lives in Paris.
stimulus	...	Rose baked a cake?	Paul bought a cake?	Paul baked a pie?
response	Paul baked a cake.	Paul baked a cake.	Paul baked a cake.	Paul baked a cake.
stimulus	...	Rachel traveled to Italy?	Maria moved to Italy?	Maria traveled to Spain?
response	Maria traveled to Italy.	Maria traveled to Italy.	Maria traveled to Italy.	Maria traveled to Italy.
stimulus	...	Rodney wrote a book?	Kate bought a book?	Kate wrote a letter?
response	Kate wrote a book.	Kate wrote a book.	Kate wrote a book.	Kate wrote a book.
stimulus	...	The man sells bananas?	The woman buys bananas?	The woman sells apples?
response	The woman sells bananas.	The woman sells bananas.	The woman sells bananas.	The woman sells bananas.
stimulus	...	The goose chased the cat?	The dog played with the cat?	The chased the goose?
response	The dog chased the cat.	The dog chased the cat.	The dog chased the cat.	The dog chased the cat.
stimulus	...	The microwave stopped working?	The car started working?	The stopped talking?
response	The car stopped working.	The car stopped working.	The car stopped working.	The car stopped working.

Table B.3: Stimuli and responses used for the recording of the prominence database.

Bibliography

Airaksinen, M., Bollepalli, B., Juvela, L., Wu, Z., King, S., and Alku, P. (2016). GlottDNN a full-band glottal vocoder for statistical parametric speech synthesis. In *Proceedings of Interspeech*.

Andrew, G., Arora, R., Bilmes, J. A., and Livescu, K. (2013). Deep canonical correlation analysis. In *ICML (3)*, pages 1247–1255.

Aylett, M. P. and Yamagishi, J. (2008). Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning. *Proceedings of LangTech*.

Badino, L., Andersson, J. S., Yamagishi, J., and Clark, R. A. (2009). Identification of contrast and its emphatic realization in HMM-based speech synthesis. In *Proceedings of Interspeech*, Brighton, United Kingdom.

Badino, L., Clark, R. A., and Wester, M. (2012). Towards hierarchical prosodic prominence generation in TTS synthesis. In *Proceedings of Interspeech*.

Bailly, G. and Holm, B. (2005). SFC: a trainable prosodic model. *Speech Communication*, 46(3):348–364.

Balyan, A., Agrawal, S., and Dev, A. (2013). Speech synthesis: A review. *International Journal of Engineering Research & Technology (IJERT)*, 2(6):57–75.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Benoît, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18(4):381–392.

Bishop, C. M. (1994). Mixture density networks. Technical report, Aston University.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Black, A. W. and Muthukumar, P. K. (2015). Random forests for statistical speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Black, A. W., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–1229.

Bolinger, D. L. (1964). Around the edge of language: Intonation. *Harvard Educational Review*, 34(2):282–293.

Bolinger, D. L. (1972). *Intonation: selected readings*. Penguin.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Braunschweiler, N. and Buchholz, S. (2011). Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality. In *Proceedings of Interspeech*, pages 1821–1824.

Braunschweiler, N., Gales, M. J., and Buchholz, S. (2010). Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Proceedings of Interspeech*, pages 2222–2225.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, 44(3):890–907.

Calhoun, S. (2007). *Information structure and the prosodic structure of English: A probabilistic relationship*. PhD thesis, University of Edinburgh.

Campbell, N. and Black, A. W. (1997). Prosody and the selection of source units for concatenative synthesis. In *Progress in speech synthesis*, pages 279–292. Springer.

Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M., et al. (2017). Siri on-device deep learning-guided unit selection text-to-speech system. *Proc. Interspeech 2017*, pages 4011–4015.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Cernak, M., Motlicek, P., and Garner, P. N. (2013). On the (un) importance of the contextual factors in HMM-based speech synthesis and coding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8140–8143.

Chen, S.-H., Hwang, S.-H., and Wang, Y.-R. (1998). An RNN-based prosodic information synthesizer for mandarin text-to-speech. *IEEE Transactions on Speech and Audio Processing*, 6(3):226–239.

Clark, R. A. and Dusterhoff, K. E. (1999). Objective methods for evaluating synthetic intonation. In *Proceedings of Eurospeech*, volume 4, pages 1623–1626.

Cole, J., Mo, Y., and Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2):425–452.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Dall, R., Brognaux, S., Richmond, K., Valentini-Botinhao, C., Henter, G. E., Hirschberg, J., Yamagishi, J., and King, S. (2016a). Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5155–5159.

Dall, R. and Gonzalvo, X. (2016). JNDSLAM: A SLAM extension for speech synthesis. In *Proceedings of Speech Prosody*, pages 1024–1028.

Dall, R., Hashimoto, K., Oura, K., Nankaku, Y., and Tokuda, K. (2016b). Redefining the linguistic context feature set for HMM and DNN TTS through position and parsing. In *Proceedings of Interspeech*, pages 2851–2855.

Daubechies, I. et al. (1992). *Ten lectures on wavelets*, volume 61. SIAM.

Degottex, G. and Stylianou, Y. (2012). A full-band adaptive harmonic representation of speech. In *Proceedings of Interspeech*, pages 382–385.

Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24(4):423–444.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Erro, D., Sainz, I., Navas, E., and Hernaez, I. (2014). Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):184–194.

Fan, Y., Qian, Y., Xie, F.-L., and Soong, F. K. (2014). TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Farouk, M. H. (2014). *Application of Wavelets in Speech Processing*. Springer.

Fernandez, R., Rendel, A., Ramabhadran, B., and Hoory, R. (2013). F0 contour prediction with a deep belief network-gaussian process hybrid model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6885–6889.

Fernandez, R., Rendel, A., Ramabhadran, B., and Hoory, R. (2014). Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In *Proceedings of Interspeech*.

Féry, C. and Truckenbrodt, H. (2005). Sisterhood and tonal scaling. *Studia Linguistica*, 59(2-3):223–243.

Fougeron, C. and Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The journal of the acoustical society of America*, 101(6):3728–3740.

Fugal, D. L. (2009). *Conceptual wavelets in digital signal processing: an in-depth, practical approach for the non-mathematician*. Space & Signals Technical Pub.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech*, pages 39–55. Springer.

Fujisaki, H. (2008). In search of models in speech communication research. In *Proceedings of Interspeech*.

Fujisaki, H. and Hirose, K. (1982). Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. In *Proceedings of 13th International Congress of Linguists*, pages 57–70.

Fujisaki, H., Ohno, S., and Wang, C. (1998). A command-response model for f0 contour generation in multilingual speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 137–140.

Gao, B., Qian, Y., Wu, Z., and Soong, F. K. (2008). Duration refinement by jointly optimizing state and longer unit likelihood. In *Proceedings of Interspeech*, pages 2266–2269.

Gerson, I., Karaali, O., Corrigan, G., and Massey, N. (1996). Neural network speech synthesis. In *Sixth Australian International Conference on Speech Science and Technology.*

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Goller, C. and Kuchler, A. (1996). Learning task-dependent distributed representations by backpropagation through structure. In *IEEE International Conference on Neural Networks*, volume 1, pages 347–352.

Graps, A. (1995). An introduction to wavelets. *IEEE Computational Science & Engineering*, 2(2):50–61.

Grosz, B. and Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In *Second International Conference on Spoken Language Processing.*

Gussenhoven, C. (2004). *The phonology of tone and intonation.* Cambridge University Press.

Hashimoto, K., Oura, K., Nankaku, Y., and Tokuda, K. (2015). The effect of neural networks in statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4455–4459.

Hayes, B. (1983). A grid-based theory of english meter. *Linguistic Inquiry*, 14(3):357–393.

Hayes, B. (1989). The prosodic hierarchy in meter. *Rhythm and meter. Phonetics and phonology*, 1:201–260.

Henter, G. E., Merritt, T., Shannon, M., Mayo, C., and King, S. (2014). Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. In *Proc. Interspeech*, pages 1504–1508.

Henter, G. E., Ronanki, S., Watts, O., Wester, M., Wu, Z., and King, S. (2016). Robust TTS duration modelling using DNNs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5130–5134.

Hinterleitner, F., Möller, S., Norrenbrock, C., and Heute, U. (2011a). Perceptual quality dimensions of text-to-speech systems. In *Twelfth Annual Conference of the International Speech Communication Association.*

Hinterleitner, F., Neitzel, G., Möller, S., and Norrenbrock, C. (2011b). An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Proceedings of Blizzard Challenge Workshop. International Speech Communication Association (ISCA).*

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10.

Hirschberg, J. (1993). Studies of intonation and discourse. In *ESCA Workshop on Prosody*.

Hirst, D. and Di Cristo, A. (1998). A survey of intonation systems. *Intonation systems: A survey of twenty languages*, pages 1–44.

Hirst, D., Di Cristo, A., and Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and experiment*, pages 51–87. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holmes, J. N., Mattingly, I. G., and Shearme, J. N. (1964). Speech synthesis by rule. *Language and speech*, 7(3):127–143.

Honnet, P.-E., Gerazov, B., and Garner, P. N. (2015). Atom decomposition-based intonation modelling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4744–4748.

House, A. S., Williams, C. E., Hecker, M. H., and Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America*, 37(1):158–166.

Hsia, C.-C., Wu, C.-H., and Wu, J.-Y. (2010). Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):1994–2003.

Hu, Q. (2016). *Statistical parametric speech synthesis based on sinusoidal models.* PhD thesis, The University of Edinburgh.

Hu, Q., Wu, Z., Richmond, K., Yamagishi, J., Stylianou, Y., and Maia, R. (2015). Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning. In *Proceedings of Interspeech*, pages 854–858.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 373–376.

Ijima, Y., Hojo, N., Masumura, R., and Asami, T. (2017). Prosody aware word-level encoder based on BLSTM-RNNs for DNN-based speech synthesis. *Proceedings of Interspeech*.

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57(S1):S35–S35.

ITU-R Recommendation BS. 1534-1 (2015). Method for the subjective assessment of intermediate quality level of coding systems. *International Telecommunication Union.*

ITU-T Recommendation P.800 (1996). Methods for subjective determination of transmission quality. *International Telecommunication Union.*

Iwahashi, N., Kaiki, N., and Sagisaka, Y. (1992). Concatenative speech synthesis by minimum distortion criteria. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 65–68.

Jordan, M. I. (1990). Attractor dynamics and parallelism in a connectionist sequential machine. In Diederich, J., editor, *Artificial Neural Networks*, pages 112–127. IEEE Press, Piscataway, NJ, USA.

Jun, S.-A. (2006). *Prosodic typology: The phonology of intonation and phrasing*, volume 1. Oxford University Press on Demand.

Kang, S., Qian, X., and Meng, H. (2013). Multi-distribution deep belief network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8012–8016.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *MAVEBA*, pages 59–64.

Kawahara, H., Masuda-Katsuse, I., and De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3):187–207.

King, S. (2011). An introduction to statistical parametric speech synthesis. *Sadhana*, 36(5):837–852.

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, 1(1).

King, S. and Karaiskos, V. (2013). The blizzard challenge 2013. In *Proceedings of Blizzard Challenge Workshop. International Speech Communication Association (ISCA).*

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, 67(3):971–995.

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2):1038–1054.

Kominek, J. and Black, A. W. (2004). The CMU Arctic speech databases. In *Fifth ISCA Workshop on Speech Synthesis.*

Kominek, J., Schultz, T., and Black, A. W. (2008). Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kruschke, H. and Lenz, M. (2003). Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis. In *Proceedings of Interspeech.*

Ladd, D. R. (2008). *Intonational phonology.* Cambridge University Press.

Latorre, J. and Akamine, M. (2008). Multilevel parametric-base f0 model for speech synthesis. In *Proceedings of Interspeech*, pages 2274–2277.

Latorre, J., Gales, M. J., Knill, K., and Akamine, M. (2013). Training a suprasegmental parametric f0 model without interpolating f0. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6880–6884.

Latorre, J., Yanagisawa, K., Wan, V., Kolluru, B., and Gales, M. J. (2014). Speech intonation for TTS: Study on evaluation methodology. In *Proceedings of Interspeech*, pages 2957–2961.

Lazaridou, A., Pham, N. T., and Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Lebret, R. and Collobert, R. (2015). Rehabilitation of count-based models for word vector representations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 417–429. Springer.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Lei, M., Wu, Y.-J., Ling, Z.-H., and Dai, L.-R. (2010). Investigation of prosodie fo layers in hierarchical fo modeling for HMM-based speech synthesis. In *IEEE 10th International Conference on Signal Processing (ICSP)*, pages 613–616.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *ACL (2)*, pages 302–308.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry*, 8(2):249–336.

Ling, Z.-H., Deng, L., and Yu, D. (2013a). Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2129–2139.

Ling, Z.-H., Deng, L., and Yu, D. (2013b). Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7825–7829.

Lu, H., King, S., and Watts, O. (2013). Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. *8th ISCA Workshop on Speech Synthesis (SSW8)*, pages 281–285.

MATLAB (2014). Wavelet toolbox documentation (r2014a). `http://www.mathworks.co.uk/help/wavelet/index.html`. Accessed: 2014.

Mayo, C., Clark, R. A., and King, S. (2005). Multidimensional scaling of listener responses to synthetic speech. In *Proceedings of Interspeech*, Lisbon, Portugal.

Mendelson, J. and Aylett, M. (2017). Beyond the listening test: an interactive approach to TTS evaluation. In *Proceedings of Interspeech*.

Merritt, T. (2016). *Overcoming the limitations of statistical parametric speech synthesis*. PhD thesis, The University of Edinburgh.

Merritt, T., Clark, R. A., Wu, Z., Yamagishi, J., and King, S. (2016). Deep neural network-guided unit selection synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5145–5149.

Merritt, T. and King, S. (2013). Investigating the shortcomings of HMM synthesis. In *8th ISCA Workshop on Speech Synthesis (SSW8)*, pages 185–190.

Mertens, P. (2004). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Speech Prosody 2004, International Conference*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Milner, B. and Shao, X. (2007). Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction. *IEEE transactions on audio, speech, and language processing*, 15(1):24–33.

Morise, M., Yokomori, F., and Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nakatani, C. H., Hirschberg, J., and Grosz, B. J. (1995). Discourse structure in spoken language: Studies on speech corpora. In *Working notes of the AAAI spring symposium on empirical methods in discourse interpretation and generation*, pages 106–112.

Nespor, M. and Vogel, I. (1986). Prosodic phonology. *Dordrecht: Foris.*

Nooteboom, S. (1997). The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, 5:640–673.

Obin, N. (2011). *Melos: Analysis and modelling of speech prosody and speaking style.* PhD thesis, Université Pierre et Marie Curie-Paris VI.

Obin, N., Beliao, J., Veaux, C., and Lacheret, A. (2014). SLAM: Automatic stylization and labelling of speech melody. In *Speech Prosody*, pages 246–250.

Obin, N., Lacheret, A., Rodet, X., et al. (2011). Stylization and trajectory modelling of short and long term speech prosody variations. In *Proceedings of Interspeech.*

Obin, N., Lanchantin, P., Avanzi, M., Lacheret, A., Rodet, X., et al. (2010). Towards improved HMM-based speech synthesis using high-level syntactical features. In *Speech Prosody*, pages 2000–2004.

Obin, N., Rodet, X., Lacheret-Dujour, A., et al. (2009). A multi-level context-dependent prosodic model applied to duration modeling. In *Proceedings of Interspeech*, pages 512–515.

Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition.* PhD thesis, The University of Cambridge.

Öhman, S. (1967). *Word and sentence intonation: A quantitative model.* Speech Transmission Laboratory, Department of Speech Communication, Royal Institute of Technology.

Okubo, T., Mochizuki, R., and Kobayashi, T. (2006). Hybrid voice conversion of unit selection and generation using prosody dependent HMM. *IEICE transactions on information and systems*, 89(11):2775–2782.

Ostendorf, M., Price, P. J., and Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. *Linguistic Data Consortium*, pages 1–19.

Oura, K., Nankaku, Y., Toda, T., Tokuda, K., Maia, R., Sakai, S., and Nakamura, S. (2008). Simultaneous acoustic, prosodic, and phrasing model training for TTS conversion systems. In *IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–4.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.

Plaut, D. C., Nolan, S. J., and Hinton, G. E. (1986). Experiments on learning by back propagation. Technical report, Carnegie-Mellon University.

Pollet, V. and Breen, A. P. (2008). Synthesis by generation and concatenation of multiform segments. In *Proceedings of Interspeech*, volume 8, pages 1825–1828.

Prahallad, K., Black, A. W., and Mosur, R. (2006). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Qian, Y., Fan, Y., Hu, W., and Soong, F. K. (2014). On the training aspects of deep neural networks (DNN) for parametric TTS synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Qian, Y., Soong, F. K., and Yan, Z.-J. (2013). A unified trajectory tiling approach to high quality speech rendering. *IEEE transactions on audio, speech, and language processing*, 21(2):280–290.

Qian, Y., Wu, Z., Gao, B., and Soong, F. K. (2011). Improved prosody generation by maximizing joint probability of state and longer units. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1702–1710.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P. (2011). HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):153–165.

Rao, K. S. (2012). *Predicting Prosody from Text for Text-to-Speech Synthesis*. Springer Science & Business Media.

Ribeiro, M. S. and Clark, R. A. J. (2015). A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia.

Ribeiro, M. S., Watts, O., and Yamagishi, J. (2016a). Parallel and cascaded deep neural networks for text-to-speech synthesis. In *9th ISCA Workshop on Speech Synthesis (SSW9)*, Sunnyvale, United States.

Ribeiro, M. S., Watts, O., and Yamagishi, J. (2016b). Syllable-level representations of suprasegmental features for DNN-based text-to-speech synthesis. In *Proceedings of Interspeech*, San Francisco, United States.

Ribeiro, M. S., Watts, O., and Yamagishi, J. (2017). Learning word vector representations based on acoustic counts. In *Proceedings of Interspeech*, Stockholm, Sweden.

Ribeiro, M. S., Watts, O., Yamagishi, J., and Clark, R. A. J. (2016c). Wavelet-based decomposition of f0 as a secondary task for DNN-based speech synthesis with multi-task learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China.

Ribeiro, M. S., Yamagishi, J., and Clark, R. A. J. (2015). A perceptual investigation of wavelet-based decomposition of f0 for text-to-speech synthesis. In *Proceedings of Interspeech*, Dresden, Germany.

Ronanki, S., Henter, G. E., Wu, Z., and King, S. (2016). A template-based approach for speech synthesis intonation generation using LSTMs. *Proceedings of Interspeech*, pages 2463–2467.

Ronanki, S., Watts, O., and King, S. (2017). A hierarchical encoder-decoder model for statistical parametric speech synthesis. *Proceedings of Interspeech*.

Ronanki, S., Wu, Z., and Clark, R. A. (2015). Joint modeling of f0 and duration in deep neural network based speech synthesis.

Rosenberg, A. (2009). *Automatic detection and classification of prosodic events*. PhD thesis, Columbia University.

Rosenberg, A. (2010). AutoBI-a tool for automatic ToBI annotation. In *Proceedings of Interspeech*, pages 146–149.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sanchez, G., Silen, H., Nurminen, J., and Gabbouj, M. (2014). Hierarchical modeling of f0 contours for voice conversion. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology*, 3(1):371–405.

Selkirk, E. O. (1980). *On prosodic structure and its relation to syntactic structure*. Indiana University Linguistics Club.

Seltzer, M. L. and Droppo, J. (2013). Multi-task learning in deep neural networks for improved phoneme recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6965–6969.

Shattuck-Hufnagel, S. and Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of psycholinguistic research*, 25(2):193–247.

Shechtman, S. and Sorin, A. (2010). Sinusoidal model parameterization for HMM-based TTS system. In *Proceedings of Interspeech*, pages 805–808.

Shinoda, K. and Watanabe, T. (2001). MDL-based context-dependent subword modeling for speech recognition. *Acoustical Science and Technology*, 21(2):79–86.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling english prosody. In *Second International Conference on Spoken Language Processing*.

Sim, K. C. (2015). On constructing and analysing an interpretable brain model for the DNN based on hidden activity patterns. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 22–29.

Sluijter, A. M. and Terken, J. M. (1993). Beyond sentence prosody: Paragraph intonation in dutch. *Phonetica*, 50(3):180–188.

Smith, C. L. (2004). Topic transitions and durational prosody in reading aloud: production and modeling. *Speech Communication*, 42(3):247–270.

Socher, R. (2014). *Recursive Deep Learning for Natural Language Processing and Computer Vision*. PhD thesis, Computer Science Department. Stanford University.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics.

Sonntag, G. P., Portele, T., and Heuft, B. (1997). Prosody generation with a neural network: Weighing the importance of input parameters. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 931–934.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., and Bengio, Y. (2017). Char2wav: End-to-end speech synthesis. In *ICLR 2017*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Stan, A. and Giurgiu, M. (2011). A superpositional model applied to f0 parameterization using DCT for text-to-speech synthesis. In *6th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic inquiry*, 31(4):649–689.

Stylianou, Y., Laroche, J., and Moulines, E. (1995). High-quality speech modification based on a harmonic+noise model. In *Fourth European Conference on Speech Communication and Technology*.

Suni, A., Šimko, J., Aalto, D., and Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Suni, A. S., Aalto, D., Raitio, T., Alku, P., Vainio, M., et al. (2013). Wavelets for intonation modeling in HMM speech synthesis. In *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*.

Swerts, M. and Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and speech*, 37(1):21–43.

Taylor, P. (1998). The tilt intonation model. In *International Conference on Spoken Language Processing*, pages 1383–1386.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press.

Tesser, F., Sommavilla, G., Paci, G., and Cosi, P. (2013). Experiments with signal-driven symbolic prosody for statistical parametric speech synthesis. In *Eighth ISCA Workshop on Speech Synthesis*.

Teutenberg, J., Watson, C., and Riddle, P. (2008). Modelling and synthesising f0 contours with the discrete cosine transform. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3973–3976.

Tokuda, K., Kobayashi, T., and Imai, S. (1995). Speech parameter generation from HMM using dynamic features. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 660–663.

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994). Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. In *ICSLP*, volume 94, pages 18–22.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (2002). Multi-space probability distribution HMM. *IEICE TRANSACTIONS on Information and Systems*, 85(3):455–464.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. (2013). Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1315–1318.

Tomoki, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, 90(5):816–824.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78.

Tuerk, C. and Robinson, T. (1993). Speech synthesis using artificial neural networks trained on cepstral coefficients. In *Proceedings of Eurospeech*.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Turk, A. and Shattuck-Hufnagel, S. (2014). Timing in talking: what is it used for, and how is it controlled? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1658):20130395.

Turk, A. E. and Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in american english words. *Journal of Phonetics*, 35(4):445–472.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Tyler, J. (2013). Prosodic correlates of discourse boundaries and hierarchy in discourse production. *Lingua*, 133:101–126.

Vadapalli, A. and Prahallad, K. (2014). Learning continuous-valued word representations for phrase break prediction. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Vainio, M. (2014). Phonetics and machine learning: Hierarchical modelling of prosody in statistical speech synthesis. In *International Conference on Statistical Language and Speech Processing*, pages 37–54. Springer.

Vainio, M., Suni, A., and Aalto, D. (2015). Emphasis, word prominence, and continuous wavelet transform in the control of HMM-based synthesis. In *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, pages 173–188. Springer.

Vainio, M., Suni, A., Aalto, D., et al. (2013). Continuous wavelet transform for analysis of speech prosody. *TRASP 2013-Tools and Resources for the Analysys of Speech Prosody, An Interspeech 2013 satellite event, August 30, 2013, Laboratoire Parole et Language, Aix-en-Provence, France, Proceedings.*

Valentini-Botinhao, C. (2013). *Intelligibility enhancement of synthetic speech in noise*. PhD thesis, The University of Edinburgh.

Valentini-Botinhao, C., Wu, Z., and King, S. (2015). Towards minimum perceptual error training for DNN-based speech synthesis. In *Proceedings of Interspeech*.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85.

Van Santen, J. P. and Möbius, B. (2000). A quantitative model of f0 generation and alignment. *IntonationAnalysis, Modelling and Technology*, pages 269–288.

Wagner, M. and Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945.

Wan, V., Agiomyrgiannakis, Y., Silen, H., and Vit, J. (2017). Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence LSTM-based autoencoders. In *Proceedings of Interspeech*, Stockholm, Sweden.

Wang, D. and Narayanan, S. (2005). Piecewise linear stylization of pitch via wavelet analysis. In *Ninth European Conference on Speech Communication and Technology*.

Wang, P., Qian, Y., Soong, F. K., He, L., and Zhao, H. (2015a). Word embedding for recurrent neural network based TTS synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. A. (2015b). Unsupervised learning of acoustic features via deep canonical correlation analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4590–4594.

Wang, X., Takaki, S., and Yamagishi, J. (2016a). Enhance the word vector with prosodic information for the recurrent neural network based TTS system. In *Proceedings of Interspeech*, San Francisco, United States.

Wang, X., Takaki, S., and Yamagishi, J. (2016b). Investigation of using continuous representation of various linguistic units in neural network based text-to-speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, 99(10):2471–2480.

Wang, X., Takaki, S., and Yamagishi, J. (2017a). An RNN-based quantized f0 model with multi-tier feedback links for text-to-speech synthesis. *Proceedings of Interspeech*.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017b). Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*.

Watts, O. (2012). *Unsupervised learning for text-to-speech synthesis*. PhD thesis, The University of Edinburgh.

Watts, O., Gangireddy, S., Yamagishi, J., King, S., Renals, S., Stan, A., and Giurgiu, M. (2014). Neural net word representations for phrase-break prediction without a part of speech tagger. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2599–2603.

Watts, O., Henter, G. E., Merritt, T., Wu, Z., and King, S. (2016). From HMMs to DNNs: where do the improvements come from? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China.

Watts, O., Stan, A., Clark, R. A., Mamiya, Y., Giurgiu, M., Yamagishi, J., and King, S. (2013). Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. In *8th ISCA Workshop on Speech Synthesis (SSW8)*, pages 101–106.

Watts, O., Wu, Z., and King, S. (2015). Sentence-level control vectors for deep neural network speech synthesis. In *Proceedings of Interspeech*.

Watts, O., Yamagishi, J., and King, S. (2010). The role of higher-level linguistic features in HMM-based speech synthesis.

Watts, O., Yamagishi, J., and King, S. (2011). Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger. In *Proceedings of Interspeech*, pages 2157–2160.

Weijters, T. and Thole, J. (1993). Speech synthesis with artificial neural networks. In *IEEE International Conference on Neural Networks*, pages 1764–1769.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.

Wester, M., Valentini-Botinhao, C., and Henter, G. E. (2015). Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations. In *Proceedings of Interspeech*.

Wester, M., Watts, O., and Henter, G. E. (2016). Evaluating comprehension of natural and synthetic conversational speech. In *Proceedings of Speech Prosody 2016*.

Wichmann, A. (2000). *Intonation in text and discourse: Beginnings, middles, and ends*. Longman Harlow, UK.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3):1707–1717.

Williams, R. J. and Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation: Theory, architectures, and applications*, 1:433–486.

Winters, S. J. and Pisoni, D. B. (2004). Perception and comprehension of synthetic speech. *Progress Report Research on Spoken Language Processing*, 26:1–44.

Wu, Z. and King, S. (2016). Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1255–1265.

Wu, Z., Qian, Y., Soong, F. K., and Zhang, B. (2008). Modeling and generating tone contour with phrase intonation for mandarin chinese speech. In *IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–4.

Wu, Z., Valentini-Botinhao, C., Watts, O., and King, S. (2015). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wu, Z., Watts, O., and King, S. (2016). Merlin: An open source neural network speech synthesis system. In *9th ISCA Workshop on Speech Synthesis (SSW9)*, Sunnyvale, United States.

Xu, Y. (2012). Speech prosody: a methodological review. *Journal of Speech Sciences*, 1(1):85–115.

Yamagishi, J., Bakos, G., Clark, R., and Veaux, C. (2014). User guide for voice cloning toolkit (VCTK) version 1.0. *The University of Edinburgh*.

Yan, Z.-J., Qian, Y., and Soong, F. K. (2010). Rich-context unit selection (RUS) approach to high quality TTS. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4798–4801.

Yin, X., Lei, M., Qian, Y., Soong, F. K., He, L., Ling, Z.-H., and Dai, L.-R. (2014). Modeling DCT parameterized f0 trajectory at intonation phrase level with DNN or decision tree. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Yin, X., Lei, M., Qian, Y., Soong, F. K., He, L., Ling, Z.-H., and Dai, L.-R. (2016). Modeling f0 trajectories in hierarchically structured deep neural networks. *Speech Communication*, 76:82–92.

Yong, L. C., Watts, O., and King, S. (2015). Combining lightly-supervised learning and user feedback to construct andimprove a statistical parametric speech synthesizer for malay. *Research Journal of Applied Sciences, Engineering and Technology*, 11(11):1227–1232.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1998). Duration modeling for HMM-based speech synthesis. In *ICSLP*, volume 98, pages 29–31.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*.

Yu, K. (2012). Review of f0 modelling and generation in HMM based speech synthesis. In *IEEE 11th International Conference on Signal Processing (ICSP)*, volume 1, pages 599–604.

Yu, K. and Young, S. (2011). Continuous f0 modeling for HMM based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1071–1079.

Zen, H. and Braunschweiler, N. (2009). Context-dependent additive log f0 model for HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 2091–2094.

Zen, H., Gales, M. J., Nankaku, Y., and Tokuda, K. (2012). Product of experts for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):794–805.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of of Sixth ISCA Workshop on Speech Synthesis*, pages 294–299.

Zen, H., Oura, K., Nose, T., Yamagishi, J., Sako, S., Toda, T., Masuko, T., Black, A. W., and Tokuda, K. (2009a). Recent development of the HMM-based speech synthesis system (HTS). In *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA).*

Zen, H. and Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7962–7966.

Zen, H., Tokuda, K., and Black, A. W. (2009b). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2004). Hidden semi-markov model based speech synthesis. In *Proceedings of Interspeech.*