A Generalized Model for Analysis and Synthesis of English Intonation

Mahsa Sadat Elyasi Langarani

A DISSERTATION SUBMITTED TO THE FACULTY OF THE CENTER FOR SPOKEN LANGUAGE UNDERSTANDING WITHIN THE OREGON HEALTH & SCIENCE UNIVERSITY SCHOOL OF MEDICINE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE AND ENGINEERING

August 2020

Document compiled on

September 10, 2020

This dissertation is dedicated to silliest goose in the world, Ava.

Acknowledgement

I'm deeply indebted to my advisor, Jan van Santen for his continuous support during my Ph. D. study and being a visionary teacher for me. I'm extremely grateful to Peter Heeman and Esther Klabbers for their unwavering guidance. In addition, I also would like to thank the other members in my committee: Meysam Asgari, Alexander Kain, Xubo Song, and Simon King for their dedicated support, guidance, encouragement and their insightful discussion and comments. It was a great honer to have you all in my committee.

My sincere thanks also goes to all of the current and former faculty at CSLU for their important input and motivation. I want to also thank my fellow students, both past and present. You all will hold a special place in my heart. I would be impossible to leave CSLU without thanking Patricia Dickerson who played a decisive role in making the CSLU a fantastic and joyful place to work.

My dearest Hamid – I am truly grateful for your love, support, and patience. I would not have been able to accomplish my Ph. D. program or balance my research with everything else without you by my side. My family: my parents, Reza and Golsa. You were the first community that encouraged me find my passion in learning. For this, I thank you. Many thanks to my parent-in-laws, Alireza and Saeed who never wavered in their support. A heartfelt thanks goes to my dearest friend Aiste who provided support, inspiration, and motivation along the way.

Contents

List of Tables

List of Figures

Abstract

A Generalized Model for Analysis and Synthesis of English Intonation

Mahsa Sadat Elyasi Langarani

Doctor of Philosophy

Center for Spoken Language Understanding, Oregon Health & Science University

Thesis Advisor: Jan van Santen

August 20, 2020

Intonation provides a means to convey information in speech that is independent of the words and their sounds. Finding a way to automatically describe this non-verbal information is important for developing sophisticated speech technology applications. One leading approach to model intonation is using a superpositional approach that assume intonation has a hierarchical structure, and models the intonation by decomposing it's physical representative (F_0 contours) into component curves with simpler intonation patterns in multi-level manner. However, it is not clear what the set of component curves should be, and how they can be defined with few free parameters, that will allow them to be used in analysis and synthesis of English for a wide range of tasks.

The central objective of this thesis is to propose a generalized model for analysis and synthesis of English intonation. Our model is a quantitative superpositional intonation model that estimates F_0 contour by decomposing it into two levels; a phrase curve for each intermediate phrase and an accent curve for each foot. We keep the shape of the phrase curve as simple as possible to let the accent curves capture the F_0 dynamic patterns. Even though parameters of a specific accent curve are proportional into a specific foot, we have the accent curve span across the entire phrase. The formulation of component curves lets us to model the F_0 contour with a very small set of free parameters. Having a limited number of parameters and having all curves span across the entire phrase facilitates us to optimize the parameters simultaneously to estimate the component curves. We name this model GENIE: GENeralized Intonation model for English.

We investigated GENIE's potential to accurately represent intonational characteristics of the English language in both synthesis and analysis tasks through a variety of speech processing applications. In a direct comparison with the ToBI system, we showed that GENIE's component curves are able to capture the underlying patterns of English intonation. In order to test the ability of GENIE to synthesize high-quality and more natural sounding F_0 contours, we created two

different approaches based on GENIE for generating F_0 contours in a Text-to-Speech system. We investigated the effectiveness of these approaches through objective and subjective evaluations. To examine GENIE's capability to be used as an analysis tool, we created two different approaches to differentiate two speaker groups through their F_0 dynamic differences. Due to the success of these two studies, we proposed a speaker group classifier using the Non-negative Matrix Factorization algorithm and the Gini coefficient. We evaluated our classifier in an English dialect classification task. We also examined the ability of GENIE to adapt to new intonational patterns by performing several perceptual tests with a variety of speech corpora and by creating an intonation adaptation task to generate speaker-specific F_0 contours. Thus, in this dissertation we examined the performance of GENIE in two areas: 1) Predictiveness: does the model produce high-quality prediction of F_0 contours, while being linguistically descriptive? 2) Coverage: is the model flexible to subtle intonational variations?

Chapter 1

Introduction

1.1 What is intonation?

The term intonation refers to the concept of conveying non-verbal information in speech. Intonation is often considered the same as prosody; however, prosody is a more general term which in addition to intonation also consists of rhythm (stress and timing patterns) and intensity patterns. These non-verbal aspects usually can be distinguished as phonological features or acoustic features. Even though there is some disagreement on which properties of spoken language are considered prosodic, there have been many studies that show the presence of prosodic constituents through phonological observations and acoustic measurements [115, 139, 23]. Figure 1.1 illustrates that prosody involves many phonological features (e.g., pitch, stress pattern and loudness) and acoustical features (e.g., fundamental frequency, intensity, and duration).

Pitch is a perceptual feature, which allows for a listener to perceive how high or low someone speaks. Similar to musical melodies, pitch changes over time during an utterance and it is closely

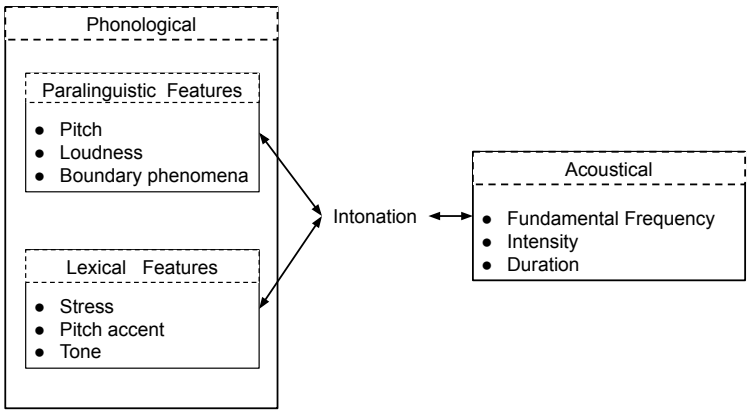


Figure 1.1: Within-group and between-group interaction of prosodic features.

correlated with duration and loudness patterns. The duration patterns or timing differ from one language to another; for example in English, timing is related to the syllable stress and pitch accents, which are two lexical features used to create prominence patterns of an utterance. Loudness is another perceptual feature which allows for a listener to perceive how quiet or loud someone speaks. These perceptual features cannot be measured directly. Instead, we can measure the fundamental frequency (or F_0 values) of the vocal cords during sound production (an acoustic feature measured in Hz). The intensity of a speech signal is the acoustic equivalent of the perceptual loudness feature. There is a strong correlation between all these prosodic features, which makes it practically impossible to define intonation as only involving one of these features. Therefore, we will use the term intonation in this thesis to refer to within-group interactions (e.g., relation between pitch and loudness) between prosodic features in each aspect and between-group interactions (e.g., relation among pitch, stress pattern and F_0) between all aspects.

1.2 How to Represent Intonation?

In spoken language, a speaker transfers a variety of information beyond lexical and syntactic information to convey a specific meaning to a target audience. Prosodic features, such as pitch patterns, prominence, timing, intensity, and phrasing give a speaker the ability to convey different meanings without changing the context (the words that were said). These non-verbal cues in an utterance are called intonation. This section is not intended to be a literature review, which is in Section 2.2 and Section 2.3; rather, this section explains the complexity of intonation in a simple example that starts with the simplest representation of intonation and progressively add more information into it.

Consider an environment consisting of two speakers: A and B. Speaker B always says the sentence "This is an expensive car" as an answer to speaker A's question who inquires about specific information available in the mentioned sentence. These questions are shown below:

Speaker A	Speaker B
What did you say?	This is an expensive car.
Is this an expensive house?	This is an expensive car.
What kind of car is this?	This is an expensive car.
Is this a cheap car?	This is an expensive car.

Even though speaker B says the same sentence in an answer to different questions, he/she puts different levels of emphasis on different parts in the stream of speech to make them prominent, which results in conveying different meanings. To answer "What did you say?", certain stressed syllables will be more prominent than others – using a so-called pitch accent or accent, in speaker B's response "This IS an exPENsive CAR". Small capitals indicate the locations of pitch accents in the sentence:

Speaker A	Speaker B
What did you say?	This IS an exPENsive CAR.
Is this an expensive house?	This IS an exPENsive CAR.
What kind of car is this?	This IS an exPENsive CAR.
Is this a cheap car?	This IS an exPENsive CAR.

Not all intonational information can be transferred through the locations of pitch accents, for instance, speaker B might puts more emphasis on the word "car" to answer "Is this an expensive house?" than when he/she emphasizes the word "car" in response to "What did you say?" By increasing the pitch range on the word "car", speaker B conveys more intonational information to speaker A. Here, pitch range information is an effective intonation characteristic. In the second through to the fourth responses where speaker B does emphasized a syllable, we represent it with an underline:

Speaker A	Speaker B
What did you say?	This IS an exPENsive CAR.
Is this an expensive house?	This IS an exPENsive <u>CAR</u>.
What kind of car is this?	This IS an ex<u>PEN</u>sive CAR.
Is this a cheap car?	This IS an ex<u>PEN</u>sive CAR.

Using English orthography (such as small capitals and underlining) to highlight important intonational information has been part of English intonation analysis for a long time. In an English

pronunciation lexicon, every word has a stressed syllable, and these syllables are more likely to be prominent – having a pitch accent – than other syllables in an utterance. Putting a pitch accent on every word of an utterance would make it sound unnatural to human ears. In general, it is more likely that the stressed syllables in content words get a pitch accent (small capitals in the example above). As mentioned previously, some intonational characteristics are produced by speaker B to add more clarity to the answer according to the communication needs of speaker A (underlined syllables in above example.) However, not all types of intonational differences can be easily captured by English orthography. When speaker B puts more emphasis on the word "car" in answer to "Is this an expensive house?" due to clarifying that this is a car not a house, he/she does not adjust the amount of prominence in the other words. However, when the emphasis shifts to the word "expensive" in response to "Is this a cheap car?", speaker B will lower the amount of emphasis on the word "car" to specify that new intonational information carried by the word "expensive" is more important than intonational information carried by the word "car". Speaker B can convey this information by using two levels of tones, a high tone (H) and a low tone (L) which correspond to either a peak or a dip in intonation. Additionally speaker B can produce a sharp rise in intonation by combining these tones into a bitonal event (LH). These tones only get assigned to the prominent syllables of the emphasized words (stressed syllables of pitch accented words).

Speaker A	Speaker B		
	H	H	H
What did you say?	This IS an exPENsive CAR.		
	H	H	LH
Is this an expensive house?	This IS an exPENsive CAR.		
	H	H	L
What kind of car is this?	This IS an exPENsive CAR.		
	H	LH	L
Is this a cheap car?	This IS an exPENsive CAR.		

This representation brings out more information about intonational characteristics. The accented words *before* the most emphasis word (the word in focus) have the same tones as in the neutral condition (compare the words "expensive" and "is" in the responses to these questions: "What did you say?", "Is this an expensive house?"). The accented words *after* the word in focus

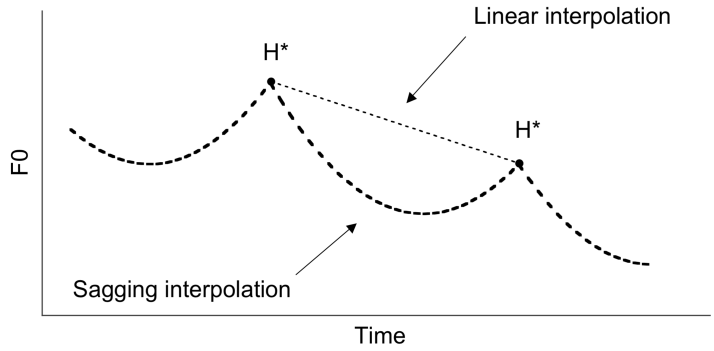
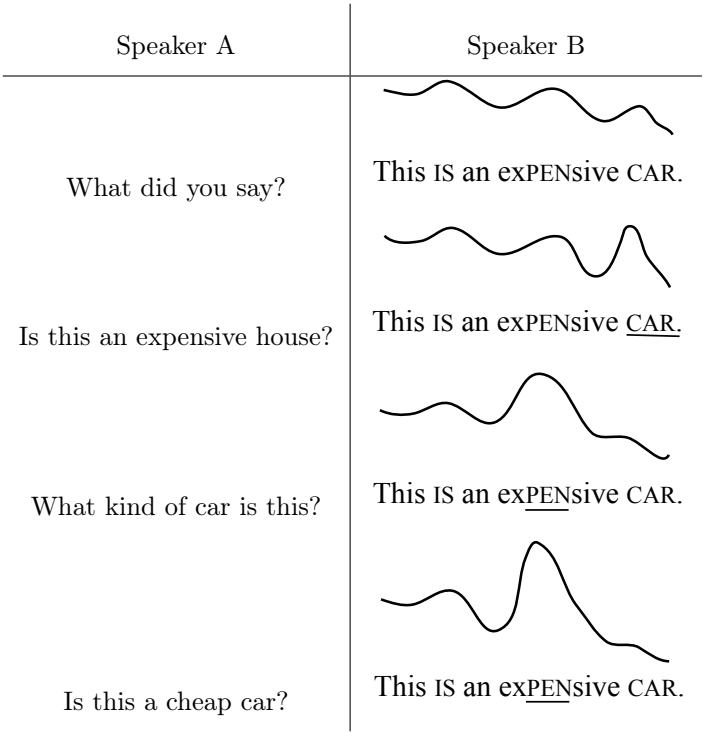


Figure 1.2: The F_0 in weak syllables (between the two peaks) are derived using different functions. A linear function like $F_0 = at + b$ for linear interpolation, and a parabolic function like $F_0 = at^2 + bt + c$ for sagging interpolation. From [187]

have a low tone as opposed to having a high tone in the neutral condition (compare the word "car" in the responses to these questions: "What did you say?", "What kind of car is this?"). This also helps to differentiate the answer of speaker B in response to "What kind of car is this?" from "Is this a cheap car?". Speaker B uses more emphasis on the whole word "exPENsive" in response to "What kind of car is this?". Therefore, the word "exPENsive" becomes the only part of the utterance that contains new information. In contrast, in response to "Is this is a cheap car?", speaker B can place a sharp rise on the second syllable of the word "exPENsive" ; thus clarifying that this is indeed an expensive car not a cheap one. Here, speaker B uses contrastive stress to not only convey more information but also to correct the information ("cheap") that was presented by speaker A.

The main drawback about this representation schema is that only obvious pitch movements are translated and more subtle ones (e.g., syllables without stress) are ignored. The assumption behind these models is that the pitch movement between the two tones are not meaningful (or are not perceptible). This raises the question of how the pitch movement in the weak syllables (e.g., syllable "this" in our target sentence) can be modeled? To answer this question, researchers have proposed different approaches, such as linear interpolation [157] or sagging transition [123]. These theories are illustrated in Figure 1.2. They suggest that pitch movement between the two tones is just a function of distance and can be modeled with any interpolation function. However, it has been shown that listeners are sensitive to changes in F_0 dynamics due to temporal alignment changes [116]. This suggests that the pitch movements in weak syllables are not captured by a simple interpolation between tones, and that they have certain patterns. Therefore, pitch movement carries detailed intonation movements that cannot be captured only by static target tones (L or H) [149]. One way to represent the pitch movement is to directly consider the physical

representation of intonation, which is known as the *fundamental frequency (F_0) contour.*



By looking at the F_0 contours in the example above, it should be clearer how higher frequency values are associated with more prominent stressed syllables, and how speaker B adjusts his/her pitch range and pitch span to convey different intonational information. For instance, the word "expensive" conveys progressively greater emphasis in response to each of the following questions: "What did you say?", "What kind of car is this?", and "Is this a cheap car?."

Through this short example, we pointed out the complexity and richness of English intonation as represented by the F_0 contour of an utterance. It should be noted that choosing F_0 contours to represent the intonation does not mean that we are ignoring other intonational features (such as duration, pitch accent, lexical stress, etc.). As mentioned earlier, intonational features are closely related to each other and one can not be considered in isolation from the others.

1.3 How to Model Intonation?

In the previous section, we discussed that the F_0 contour of an utterance can be used to represent the complexity and richness of English intonation. We have shown that representing intonation by marking up text or by adding tone annotations can not fully convey complexity of English

intonation; how can we model intonation quantitatively? We can use phonetic models, where intonational features are represented numerically in terms of vectors of acoustic features or continuous parameters. More specifically, they represent intonation as a sequence of pairs (time, F_0). There are two main categories of phonetic models: sequential and superpositional models. The sequential approach characterizes the F_0 contour as a sequence of distinct intonational events that are generated left to right. A widely used sequential intonation model is Taylor's TILT model [157], which considers the F_0 contour as a sequence of intonational accents (rising and falling) with linear connections. Superpositional models, starting with the work of Fujisaki [41], posit that the F_0 contour can be described as a superposition of several simpler component curves. Depending on whether the model is sequential or superpositional, the F_0 contour of an utterance results from interpolation between the estimated intonational events or the superposition of components of different temporal scopes.

An assumption behind sequential models is that F_0 contours are directly determined by their surface patterns in small phonological units (mainly at the syllable level). However, intonation is a suprasegmental phenomenon which is influenced by factors at different levels of a hierarchy. At the lowest level in the hierarchy, there are syllables, which are grouped together into prosodic phrases, and eventually utterances. The resulting effect of the intonational hierarchical structure on the F_0 contour cannot be modeled by a sequential approach. For example, a stressed syllable with a pitch accent with a certain amount of emphasis will result in different F_0 values in different parts of a prosodic phrase. Changes in F_0 values are not only related to local factors in smaller phonological units (such as stress at the syllable level or pitch accents at the word level), but also to more global factors in higher phonological units (such as phrasing at the prosodic phrase level).

Characterizing the F_0 contour at different phonological levels is crucial to the definition of the superpositional approach. The superpositional approach characterizes the F_0 contour as an overlay (or superposition) of several component contours of different temporal scopes. Long scope components represent the global patterns of F_0 contour over the length of a prosodic phrase. Shorter scope components represent local F_0 contour changes associated with syllables (commonly stressed accented syllables). Due to superpositional approach capturing the hierarchical structure of intonation by estimating underlying patterns of the F_0 contour in a multi-level manner, the superpositional approaches are more suitable than the sequential approaches for analyzing and synthesizing English intonation. This advantage leads us to the use of a superpositional phonetic approach to model English intonation.

1.4 Problems in Superpositional Approach

In modeling intonation, there are several theoretical and practical concerns due to two factors: (1) the description of intonation, and (2) the approach used to simulate the described intonation. Here, we narrowed down these concerns in the context of superpositional intonation models.

1.4.1 Theoretical Concerns

Using a superpositional approach to decompose an F_0 contour into its component curves – where each component is tied to a distinct phonological unit, leads to various theoretical concerns, which can be summarized as follows:

Hierarchical dependency: Due to the hierarchical structure of intonation, there are multi-level interactions between intonational features. In the superpositional approach each level has its own unique patterns, and they are superimposed on top of each other to estimate underlying patterns of the given F_0 contour. Although various intonation theories agree on the hierarchical structure of intonation, they differ in terms of how many levels should be used to represent the multi-level interaction between intonational features. In general, any superpositional approach should consist of at least two levels: one level for representing global intonational patterns at a prosodic phrase level, and one level for representing local intonational patterns at a shorter temporal scope. However, some theories suggest more than two levels to represent the intonation hierarchy (e.g., three levels [41, 169, 106], more than three levels [105, 151, 178], or even one level for each phonological unit [129].

Adaptive decomposition: Decomposing a F_0 contour into its component curves is challenging since there is no unique solution to the decomposition of a given F_0 contour, because different component curves can combine to produce the same sum curve, unless certain assumptions are made. The way in which component curves are superimposed determines the outcome of the model. For example, component curve estimated at the lower level in the intonation hierarchy, associated with smaller phonological units (commonly syllables), should only be concatenated together and then added to component curves at a higher level. Some overlap can also be applied before adding them to the component curves at the higher level.

Relevancy of component curves: The relevancy of the component curves relies on the purpose for intonation modeling. For instance, if the purpose is to have a generative model to be used in a speech synthesis system (e.g., Text-To-Speech), having an accurate estimation of the F_0 contour is more important than knowing if the component curve shapes are

linguistically meaningful or not. In contrast, if the purpose is to use the intonation model as an analysis tool, then having linguistically meaningful component curves is the central assumption. Ideally, we want to have linguistically meaningful component curves that can be used to generate the same intonation characteristics that would be produced by a specific speaker. This leads us to the following question: which phonological units (syllable, sequence of syllables, words, phrases, and utterances) are more relevant for capturing meaningful intonational movements? This issue is referred to as the lack of reference in intonation research [125, 124, 185]. Xu [185] suggested that the relevant unit for studying underlying meaningful intonational movements is the syllable. However, as also pointed out by Xu, for languages that are not lexically tonal languages, such as English, considering a specific intonation movement per syllable might cause overfitting, but it does not necessarily mean that intonation movements are unspecified in weak syllables (which usually is assumed in phonological-based approaches). This suggests that for English, the relevant prosodic unit should be syllable-based but not specifically limited to the boundary of one syllable. Such a unit, which is known as the foot, was proposed by Abercrombie et al [1]. The reason the foot is a more relevant prosodic unit than the syllable in English will be discussed in detail in the next chapter.

1.4.2 Practical Concerns

In addition to the theoretical concerns, there are two concerns that are critical when it comes to practical usage of the intonation model in different speech processing applications, namely: the level of predictability of the intonation model, and the degrees of freedom of the model [185].

Predictability: Intonation can vary substantially across different languages, making it practically impossible to have one intonation model which can achieve both high predictiveness and high descriptiveness for every existing language. Therefore, depending on the problem, it is important to assess the trade-offs between predictiveness and descriptiveness of the method. Achieving better predictiveness while being linguistically descriptive could lead us to two insights: 1) In the synthesis phase, it is important to not only test the similarity between the natural and estimated F_0 contour but also the intonation characteristics of the input categories. 2) In the analysis phase, the model should be powerful enough to be used in detection and classification of intonational characteristics. Therefore, the model should be able to accurately reconstruct the F_0 contour that makes it a useful tool for detecting prosodic phrase boundary and pitch accent events. Going further, the model should also be able to

capture hidden intonational characteristics of a speaker, which usually can not be easily represented by a phonological-based approaches. This ability would make the model a useful classification tool that covers a variety of cases: classification of individuals with dysarthria vs. neurotypical individuals, clear vs. conversational speaking styles, dialect classification, or differentiating any speaker groups regardless of speaking style, speech data, or any other variation in patterns.

Degrees of freedom: The degrees of freedom of a model refer to the number of independent free parameters that control the model for data estimation. If the number of independent free parameters is too high then the model might become too complex and that causes overfitting problems. If the number of independent free parameters is too low, then the model might be too general and it might not capture the data distribution, which causes underfitting problems. Therefore, the most important choice related to the degrees of freedom of the intonation model is whether each parameter can be meaningfully justified. For example, for a use-case with a small data size, using a simple model with only a few meaningful parameters will be more beneficial than using a complex machine learning method which requires many more parameters. In superpositional-based modeling, mainly two factors affect the degrees of freedom of the model: the number of levels used for representing the hierarchical intonational structure, and the number of parameters needed for each component curve.[1]

1.5 Thesis Problem

Superpositional approaches assume intonation has a hierarchical structure, and models the intonation by decomposing it's physical representative (F_0 contours) into component curves with simpler intonation patterns in multi-level manner. However, it is not clear what the set of component curves should be, and how they can be defined with few free parameters, that will allow them to be used in analysis and synthesis of English for a wide range of tasks.

1.6 Thesis Statement

In this thesis, we create a quantitative superpositional intonation model that provides the high-quality prediction of F_0 contours with few free parameters that the component curves are being linguistically descriptive.

[1] Please note that due to the theoretical concerns there might be additional factors, such as the level of overlap between component curves, association of the levels with phonological units, etc.

Our model decomposes a given F_0 contour into its component phrase and accent curves at two levels: the prosodic phrase level (or intermediate phrase level) and the foot level. We used two connected linear segments to model the phrase curve. We kept the shape of the phrase curve as simple as possible to let the accent curves capture the F_0 dynamic patterns. We used a combination of the skewed normal distribution and a sigmoid function to model three different types of accent curve. First, the skewed normal distribution is used to model rise-fall accents that occur in non-final foot as well as in final foot for declarative utterances. Second, a sigmoid function is used to model the rise at the end of yes-no question utterances. Third, the sum of the skewed normal distribution and the sigmoid function is used to model continuation accents at the end of a non-utterance-final phrase. Even though parameters of a specific accent curve are proportional into a specific foot, we have the accent curve span across the entire phrase. This formulation of component curves lets us to model the F_0 contour with a very small set of free parameters. Having a limited number of parameters and having all curves span across the entire phrase facilitates us to optimize the parameters simultaneously to estimate the component curves. We name this model GENIE (GENeralized Intonation model for English). We show the proposed method can be used as an analysis and synthesis tool of intonational characteristics in a variety of speech processing applications, and it can model real world variations, such as: different speaking styles, different intonational functions, different speech data, etc.

1.7 Contributions of this Thesis

In this dissertation, we propose a generalized model for analysis and synthesis of the English intonation. The proposed model is a superpositional-based model that decomposes a continuous F_0 contour into its linguistically meaningful component curves. We propose several different frameworks to examine the performance of the proposed model in terms of the objective of this thesis.

In Chapter 2 we presents the literature review. First, some definitions are discussed. Then, fundamental and more recent intonational models will be reviewed. The rest of the chapter will focus on the usage of intonation models in speech processing applications. In Chapter 3, we propose the generalized intonation model for English language (GENIE). We present the methodology and mathematical formulation of GENIE. In Chapter 4, we propose a framework which combines GENIE with a regression-based duration model for detection of intonational events. In Chapter 5, we propose two approaches – data-driven-based and neural-network-based – for generating F_0 contours using GENIE in a TTS application. In the second part of this chapter, we propose a new intonation adaptation method using GENIE to transform the perceived identity of a TTS system

to that of a target speaker with a small amount of training data. This chapter tests predictability of the model using both objective and subjective evaluations. In Chapter 6, we propose an approach to perform speaker classification which exclusively uses features derived from the F_0 contour by using GENIE. A special aspect of our approach is the focus on F_0 contour dynamics – often underused in speaker group classification. Finally, the last chapter gives a summary of the main findings of the research carried out in the scope of this dissertation. The dissertation ends with an outlook on future work.

The following are accepted articles that came from the research performed for this thesis:

1. **M. S. Elyasi Langarani**, J. van Santen, E. Klabbers, A novel pitch decomposition method for the generalized linear alignment model, **ICASSP**, 2014 [33]

2. **M. S. Elyasi Langarani**, J. van Santen, Modeling fundamental frequency dynamics in hypokinetic dysarthria, **Spoken Language Technology (SLT)**, 2014 [28]

3. **M. S. Elyasi Langarani**, J. van Santen, S. H. Mohammadi**,** A. Kain, Data-driven Foot-based Intonation Generator for Text-to-Speech Synthesis, **Interspeech,** 2015 [34].

4. **M. S. Elyasi Langarani**, J. van Santen, Speaker intonation adaptation for transforming text-to-speech synthesis speaker identity, **ASRU**, 2015 [29]

5. **M. S. Elyasi Langarani**, J. van Santen, Automatic, model-based detection of pause-less phrase boundaries from fundamental frequency and duration features, **9th ISCA Speech Synthesis Workshop (SSW9)**, 2016 [30]

6. **M. S. Elyasi Langarani**, J. van Santen, Foot-based Intonation for Text-to-Speech Synthesis using Neural Networks, **Speech Prosody**, 2016 [31]

7. **M. S. Elyasi Langarani**, J. van Santen, Recurrent convolutional networks for classification of speaker groups based on prosodic information, **Women in Machine learning Workshop (WiML)**, 2017 [32]

The following are planned submissions covering some of the contributions of the dissertation:

1. **M. S. Elyasi Langarani**, J. van Santen, Investigating prosodic unit effects of fundamental frequency dynamics in clear and conversational speech

2. **M. S. Elyasi Langarani**, J. van Santen, Prosody based dialect classification using NMF and sparsity criteria

Chapter 2

Literature Review

In this chapter, we present the literature review. In Section 2.1, we discuss the relationship between intonation and several prosodic features that are used in this thesis. In Section 2.2, we discuss relevancy of prosodic unit in English intonation. In Section 2.3, we review fundamental and more recent intonational models. In Section 2.4, we focus on the usage of intonation models in TTS and TTS adaptation. In Section 2.5, we discuss how intonational features are extracted for speaker group classification tasks. Finally in Section 2.6, we discuss about evaluation metrics.

2.1 The Phonology of English Intonation

The previous chapter gave an introduction to what intonation consists of, and how we can visualize and model it. It also drew attention to the aspects of prosody that are characteristic of the English language. As can be seen in Figure 1.1 that it is also represented here as Figure 2.1, intonation refers to within-group interactions between prosodic features in each aspect and between-group interactions between all aspects. As discussed in the previous chapter, intonational features are closely related to each other and one can not be considered in isolation from the others. A comprehensive account of the relationship of intonation to other prosodic features lies outside the scope of this thesis, but in this section we discuss the relationship between intonation and several of prosodic features through each aspect: paralinguistic, lexical and acoustical.

From a paralinguistic point of view, intonation is defined as the interaction between pitch, loudness and prosodic boundary phenomena. These paralinguistic features help listeners make inferences about a speaker's state or attitude, such as enthusiasm or friendliness and depression or happiness. It also can help in regulating turn-taking in communication: a speaker can naturally use an F_0 pattern to prompt the listener that it is their turn, or that the speaker does not want to be interrupted. For example, consider this sentence "AVA does not eat ANY burger" in reply to

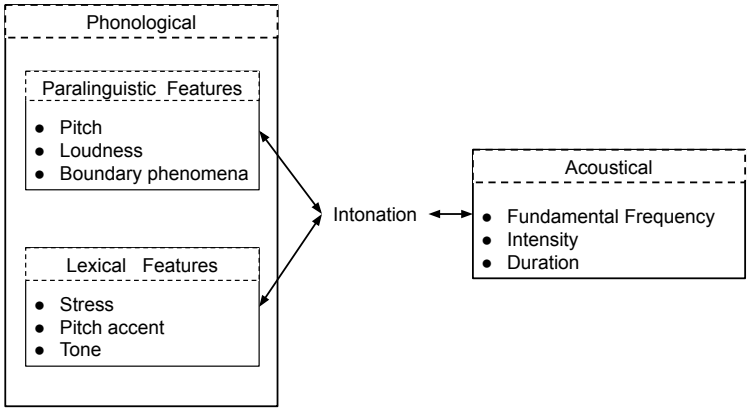


Figure 2.1: Within-group and between-group interaction of prosodic features.

"Why did she not eat her burger?" in the following two cases.

The response on the left, with a falling F_0 at the end, indicates that Ava is a vegetarian and will not eat any meat without exception. The response on the right, with a rising F_0 at the end, means that speaker is not done yet and wants to explain why Ava does not eat any burger and continues to explain: "Ava is selective (or picky); she does not like just any burger".

From a lexical point of view, intonation is defined as the interaction between stress pattern and pitch accent. In English, listeners pay attention to the most prominent syllables to understand the message. For example, the rhythmic pattern in word "iDENtifiCAtion" is identical to the phrase "we TOOK a vaCAtion" since they both share the same stress pattern. Not all syllables are pronounced with the same degree of force. For instance, stressed syllables of emphasized (or accented) words are higher in energy, longer in duration, and have a greater change in F_0 values compared to stressed syllables of unemphasized words. Stress patterns of syllables in American English are predetermined. For example, in the noun "present", the stress falls on the first syllable ('pre sent). As a verb, the second syllable of "present" carries the stress (pre 'sent). However, speakers choose different intonation patterns to emphasize different words for conveying different meanings. For example, in the following sentence "The boy was there when the sun rose," every word (except "the") consists of one stressed syllable. In Figure 2.2 top plot, a speaker emphasizes the words "BOY" and "ROSE" to highlight new information in the conversation. In Figure 2.2 bottom plot, in addition to emphasizing the words "BOY" and "ROSE" also the speaker gives special emphasis

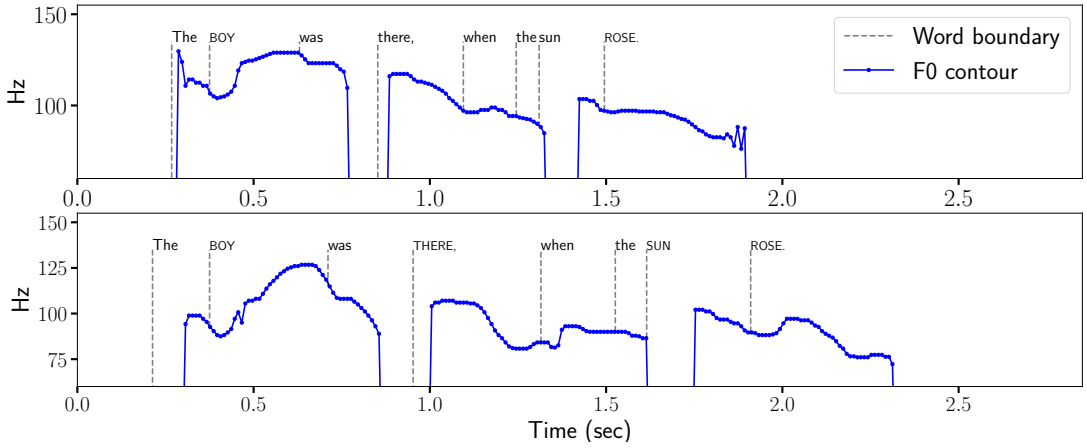
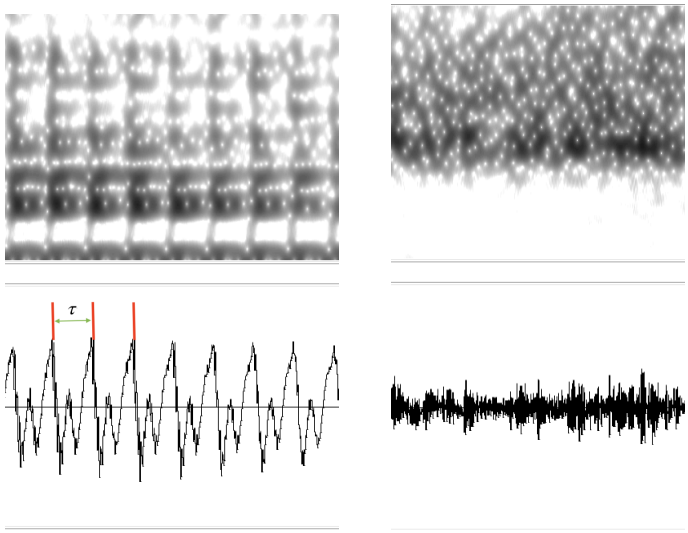


Figure 2.2: An example that shows how a speaker emphasize different words to change the F_0 contour dynamics of an utterance. Small capitals indicate stressed-accented syllables in each utterance.

to more specific details (the words "THERE" and "SUN") to make the sentence clearer. Therefore, utilizing the stressed syllable of an accented word (stressed-accented-syllable) is a key component of a speaker's ability to convey a subtle meaning.

The acoustical features of intonation are defined as the interaction between fundamental frequency, duration and intensity. Among acoustic features, the fundamental frequency is mostly considered as a primary physical-prosodical feature that can be measured. There is a periodic pattern at the time-domain representation of a human speech waveform when a voiced sound (e.g., a vowel) is pronounced. Figure 2.3 shows periodic (voiced) and noisy (unvoiced) regions in the word "easy". Each peak in the periodic region is called a glottal pulse. The duration of one glottal cycle is represented by the symbol τ. The fundamental frequency of a periodic signal is the inverse of this duration ($1/\tau$) and is measured in Hz. When a person produces a voiced sound, one's vocal folds produce a set of frequencies (fundamental and its harmonics). The fundamental frequency is the lowest frequency (starting from zero) which is also perceived as the loudest frequency by human ear. The fundamental frequency is usually referred to as F_0. Many factors can affect the F_0 of someone's voice, such as: age (usually kids have a high-pitched voice compared to adults), gender (usually men speak in lower-pitched voice than women), and emotion (people may use high-pitched voice when they are angry or excited, or they may use low-pitched voice when they are sad).



(a) Waveform and spectrogram of /i:/

(b) Waveform and spectrogram of /z/

Figure 2.3: Waveform and spectrogram of a voiced and an unvoiced segment in the word "easy". Each red line represents a glottal pulse. The duration between two back-to-back glottal pulses is represented by the symbol τ. There is no periodic pattern inside the/z/.

2.2 Intonation Segmentation into Prosodic Units

Meaningful prosodic movements can be perceived and expressed differently from one language to another. In this thesis, we only focus on English (mostly on American English pronunciation).

In American pronunciation, every prosodic unit consists of at least one stressed-accented-syllable. The largest prosodic unit that has one complete intonation pattern is called an intonational phrase [125]. Every intonational phrase consists of at least one intermediate phrase and every intermediate phrase consists of at least one stressed-accented-syllable (therefore every intonational phrase does as well); However it is unclear which prosodic unit (syllable, sequence of syllables, words, intermediate phrase, or intonational phrase) is more relevant for representing a single meaningful intonational movements? Many studies used the syllable as the smallest prosodic unit [185, 157]; their motivation was that the syllable is a smallest common prosodic unit across languages (e.g., in Mandarin Chinese every syllable has a meaningful intonation pattern). As discussed in the previous chapter, considering a specific intonation pattern per syllable in English is not necessary since weak syllables in English do not show strong intonation movements like stressed-accented syllables.

After discussing the details of ToBI transcription system In Section 2.2.1, we then discuss its view of the smallest prosodic unit. In Section 2.2.2 we discuses about different candidates for the

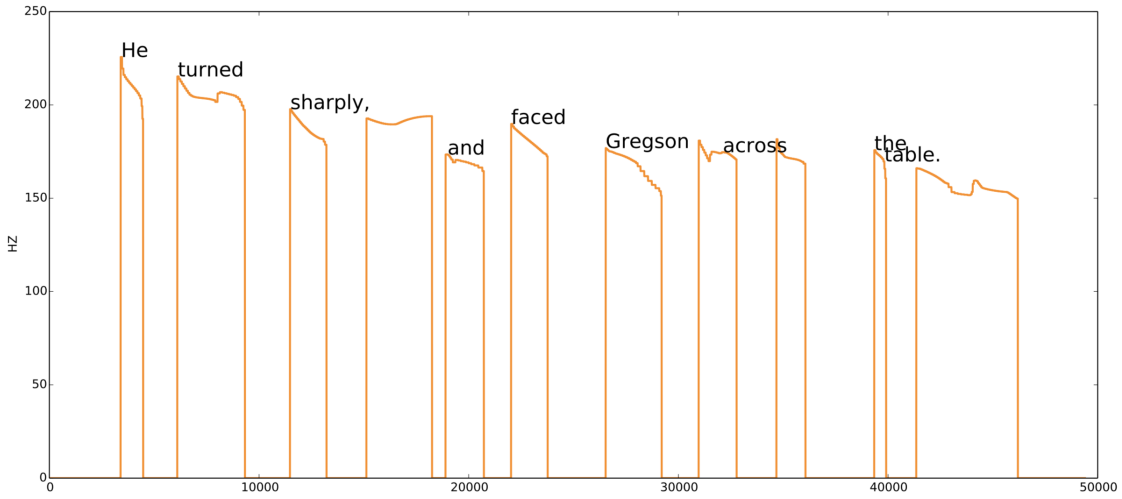


Figure 2.4: Representation of the F_0 contour (orange curve) with Bolinger's notation (black words).

smallest prosodic unit in English. Then in Section 2.2.3, we show how total number of intonational patterns in ToBI can be reduced under the smallest relevant prosodic unit2.2.3.

2.2.1 ToBI Transcription System

Bolinger proposed one of the first and simplest notations for prosody [14]. He aligned the word sequences with their real F_0 values (Figure 2.4, black words). It is much easier for readers to capture intonation from this notation than from plaintext, but Bolinger's notation requires hand labeling; it is almost impossible to automatically analyze or synthesize it. Under the influence of Pierrehumbert research [125, 123], autosegmental-metrical (AM) analysis framework became the dominant in intonational research (for an introduction to AM and a critique see [78]). A modified version of AM was proposed by Silverman and his coworkers [141] as, Tones and Break Indices (ToBI), which is still commonly used.

The ToBI transcription system provides a set of symbolic labels (Table 2.1) for distinguishing between all categorical intonation patterns. To achieve this aim, ToBI considers two aspects of prosody:

1. Accent: contributes to the prominence of a word in an utterance
2. Phrasing: divides sentences into groups of words, which consists of four levels:
   (a) First level: the word boundary within a phrase
   (b) Second level: which is used to mark a mismatch

(c) Third level: the end of an intermediate phrase

(d) Fourth level: the end of an intonational phrase

Symbol	**Description**
H	High tone is associated with pitch that occurs in upper part of a speaker's pitch range
L	Low tone is associated with pitch that occurs in lower part of a speaker's pitch range
*	stressed-accented-syllable
-	End of an intermediate phrase
%	End of an intonational phrase
!	Pitch movement that lowers F_0 from any H tone into a downstep, which is not necessary in the lowest part of the pitch range (not as low as L)

Table 2.1: ToBI symbols

ToBi annotation	intonational phrase **type**
L-L%	Statement sentence and Wh-question
L-H%	Continuation
H-L%	Listing and enumeration (or plateau contour)
H-H%	Yes-No question
!H-L%	Listing and enumeration (or calling contour)
!H-H%	Continuation

Table 2.2: ToBI annotation for phrasal tone

There are two main levels of phrasing: the full intonational phrase level (intonational phrase, fourth level), and the intermediate intonational phrase level (intermediate phrase, third level). ToBI uses this symbol "-" followed by a tone to represent intermediate phrasal tone. The end of one intonational phrase by default is aligned with the end of an intermediate phrase, therefore ToBI categories intonational phrasing patterns through bitonal symbols; a tone plus symbol "%" followed by an intermediate phrasal tone. There are four basic intonational phrasal tone combinations: L-L%, L-H%, H-L%, and H-H%. Also a downstep[1] can only happen in the first H tone of the following

[1]Downstep is a pitch movement that iteratively lowers F_0 peaks of successive accented-syllable with a constant proportion of the previous peak[78]. However this downstep never reaches the lowest part of pitch range (not necessary as low as L tone).

phrasal tones: H-L%, and H-H%. Therefore, there are two more phrasal tones to consider: !H-L%, and !H-H% (Table 2.2). It should be noted that !H-H% is theoretically possible, but it is also hard to distinguish from L-H%.

ToBI categorizes word pitch accents into five accent types. For each accent, the pitch movement of the stressed-accented-syllable is illustrated with a starred tone. These accent types are: H*, L*, L+H*, L*+H, and H+!H*. All high tone accents can be downstepped: !H*, L+!H*, and L*+!H. However, these downstepped accents show the same pattern of their non-downstepped version.

In Figure 2.5 that adapted from [173] illustrates all 28 possible ToBI intonational patterns for a phrase with single stressed-accented-syllable. Each cell illustrates an intonational pattern under certain combinations of an accent tone and a phrasal tone. The theoretical pitch movement of a target tone is illustrated with a horizontal solid line. The starred target tone (pitch movement of stressed-syllable) is differentiated from other tones by a bold solid line. The dotted line shows transition between tone targets.

Most phonological-based approaches (e.g., ToBI transcription system) use words with a pitch accent as the smallest prosodic unit; the main drawback of these approaches is that the intonation movement in unaccented words is unspecified. As we also discussed in the previous chapter, weak syllables (either unstressed or stressed in unaccented word) have enough intonational movements to be specified but also not strong enough to be individualized.

2.2.2 Candidates for Smallest Prosodic Unit in English

There is consensus that each stressed-accented-syllable needs to be specified in English. There are some researchers that advocate that the smallest prosodic unit in English should consist of exactly one stressed-accented-syllable [72, 168, 164, 188, 6], but there is uncertainty about its boundaries. Are they tied to the stressed-accented-syllable boundaries or can they span multiple syllables (not stressed-accented syllable)? For example, in the sentence "I am HAPpy about imPROVEment," with three stressed-accented syllables (represented by uppercase typeface), what is the smallest prosodic unit that can convey meaningful intonation? A few examples are given next:

- Syllable: "I am HAP py a bout im PROVE ment."
- Word: "I am HAP py a bout im PROVE ment."
- Sequence of syllables:
  - Right-headed: "I am HAP py a bout im PROVE ment."
  - Left-headed: "I am HAP py a bout im PROVE ment."

Figure 2.5: Combination of accent types and phrasal tones. The starred target tone is differentiated from other tones by a bold solid line. The dotted line shows transition between tone targets. Adapted from [173]

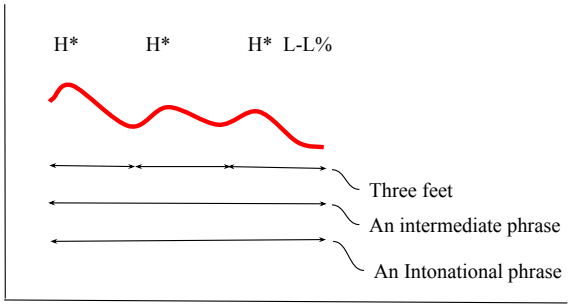


Figure 2.6: Foot structure in a statement utterance

- Both directions: there are many possibilities in this case
  - "I am HAP py a bout im PROVE ment."
  - "I am HAP py a bout im PROVE ment."
  - ...
  - "I am HAP py a bout im PROVE ment."

We use the definition of a left-headed prosodic unit to capture a meaningful prosodic movement since English is a left-dominant language [50]; In English there is a tendency for the first syllable of words to be strong and the remaining to be weak, that is, left-dominant. Therefore, the left-headed prosodic unit preferred over the right-headed prosodic unit for English due to two main reasons. First, multi-syllabic words with primary stress on the final syllable are less common than other words of the same length [24, 20]. Second, most of the intonational function (such as focus) in English have a post-effect rather than a pre-effect. For example it has been shown that if an initial-stressed word in a sentence is focused, any unstressed syllables after the stressed-accented-syllable of the first word will be assigned a higher pitch compared to when there is no focus [88].

The left-headed prosodic unit, which will be referred to as a foot, starts with a stressed-accented-syllable and ends before the next stressed-accented-syllable or with a prosodic phrase boundary [1]. For example in Figure 2.6, each foot start with a stressed-accented-syllable with a H tone and ends before the next one, or in Figure 2.7 the final foot in the first intermediate phrase start with a stressed-accented-syllable with a H tone and ends with a intermediate phrase boundary with H tone.

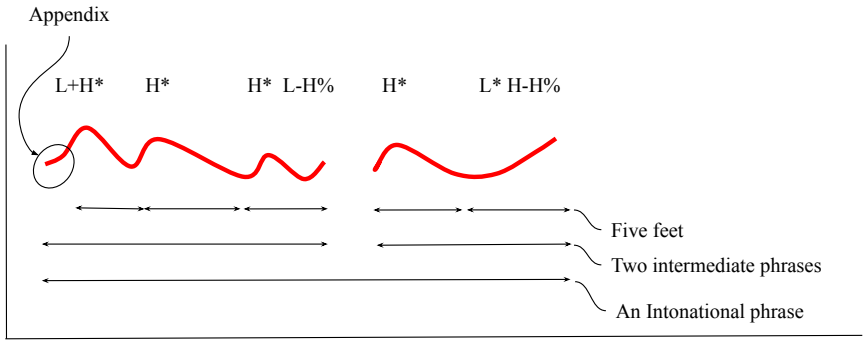


Figure 2.7: Foot structure in a yes-no question utterance that consists of two intermediate phrases.

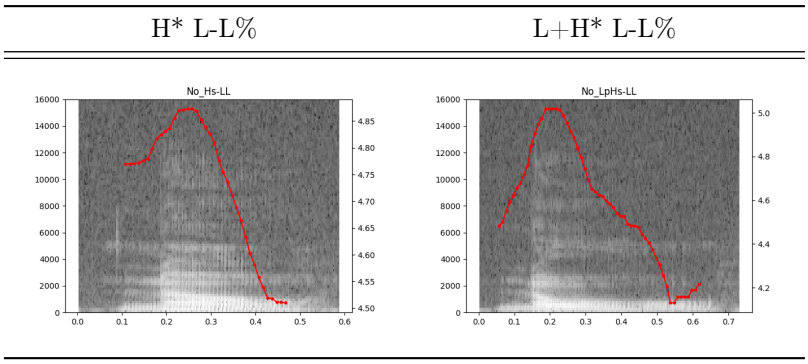


Figure 2.8: Difference between H* and L+H* for the word "NO"

2.2.3 ToBI Intonation Patterns under Foot Segmentation

In the previous section, we argued that the foot is a suitable prosodic unit for studying English intonation patterns. In this section we investigate how ToBI intonation patterns can be categorized using the foot structure. A core difference between an accent and a foot is that an accent is defined as a word containing a prominent syllable and not necessarily as a (left-headed) foot, which requires that the first syllable be the prominent syllable. Feet and accents have overlapping but not necessarily matched boundaries. We will describe this difference through three examples.

First, consider a one-word single-phrase utterance with a stressed-syllable at the beginning, e.g., "NO". In this example, the foot and accent share the same boundaries and only three accent types can occur (H*, L*, and L*+H). In the case of L+H*, because there is not a non-stressed syllable preceding the stressed syllable, the unstarred tone (L) cannot occur. This accent is matched with its monotone accent (H*). In Figure 2.8, one speaker produces the word "NO" under two different intonation patterns H* L-L% and L+H* L-L%. Since there are no unstressed syllables before the prominent syllable, there is only a sharp rise from the mid-pitch range to a high F_0 peak. In the

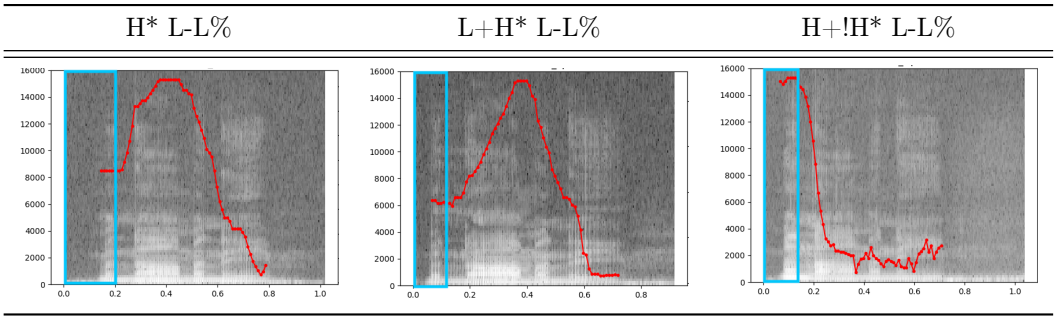


Figure 2.9: Difference between two pairs (H* and L+H*) and (H* and H+!H*) for the word "meNOmonee". The red dots inside the light blue box show the F_0 contour points for the appendix "me-"

case of H+!H*, similar to the case of L+H*, this accent should also be matched with its monotone accent (!H*). However, downstep cannot happen at the beginning of an intonational phrase, since it is required to follow a high tone. Therefore, the accent type H+!H* cannot happen in this situation.

Second, consider a one-word, single-phrase utterance with at least one unstressed syllable at the beginning, e.g., "meNOmonee". The segmentation for this utterance using three different prosodic units (phrase unit, accent unit and foot) are given as follow:

- Phrase: me NO monee.
- Accent: me NO monee.
- Foot: me NO monee.

Accent boundaries are matched with intonational phrase boundaries, while the foot starts at the stressed syllable "-NO-" and ends at the end of the intonational phrase. According to the foot definition, intonational movement in phrase-initial unstressed syllables, which is called the "appendix", is not part of the foot (in this example, F_0 contour points in the unstressed syllable "me-").

Also in this example, only three accent types can occur (H*, L*, and L*+H) in the foot since in case of L+H*, and H+!H*, the unstarred tone is not part of the foot (L in L+H*, and H in H+!H are appendix). However, F_0 contour points under these tones are still part of the intonational phrase. In Figure 2.9, one speaker produces word "meNOmonee" under three intonation patterns to differentiate between two pairs (H* and L+H*) and (H* and H+!H*) in a statement sentence.

Third, consider a multi-word, single-phrase utterance with at least two accented words (e.g., "I am HAPpy about imPROVEment."). In this single-phrase example, the phrase starts with the accented-stressed syllable "I", and as such all unaccented syllables in this example are not considered an appendix.

- Intonational Phrase: I am HAP py a bout im PROVE ment.
- Accent: I am HAP py a bout im PROVE ment.
- Foot: I am HAP py a bout im PROVE ment.

In the third accented word, "imPROVEment," there is a mismatch between the start point of the accent and the foot. The third accent starts with the unstressed syllable "im-" while in the foot segmentation this syllable belongs to the second foot. Therefore, in this situation, three accent types can occur (H*, L*, and L*+H) in the foot. In the case of L+H*, and H+!H*, the unstarred tone will move into the previous foot.

The three examples above show that in foot segmentation only three ToBI accent types can occur (H*, L* and L*+H). The first advantage of ToBI under foot segmentation over the original ToBI is that it decrease ambiguity in differentiating L+H* and H+!H* from H*. As we saw, there is some similarity between H* and L+H* tones and between H* and H+!H* tones. Because of these similarities, utterances often contain regions with more than one valid transcription which decreases the reliability of annotations. ToBI is a qualitative model with low inter-annotator agreement even for trained annotators, and this disagreement becomes even more extreme when ToBI annotations are applied to expressive speech or spontaneous speech. The second advantage is that by using foot segmentation, the total number of intonational patterns can be reduced to 16. Figure 2.10 illustrates these 16 intonational patterns for a single phrase with one stressed-accented-syllable. Each cell illustrates an intonational pattern under certain combinations of an accent tone and a phrasal tone. Since pitch transition between tones are more smooth in real speech, red curves show more realistic F_0 contours for each situation. In chapter 3 we show the proposed intonation model is capable of capturing and predicting all intonation patterns mentioned in the ToBI system (even F_0 contour points in the appendix) by using only foot-based information and reduce the total number of patterns from 24 to three.

2.3 Intonation Models

Intonation models can be distinguished in terms of phonetic vs. phonological models.

Phonological models: In phonological models intonation is considered as a sequence of distinctive discrete tonal categories. Therefore, these models are qualitative and sequential. The ToBI model plays an important role in the popularization of phonological models in intonation description and analysis.

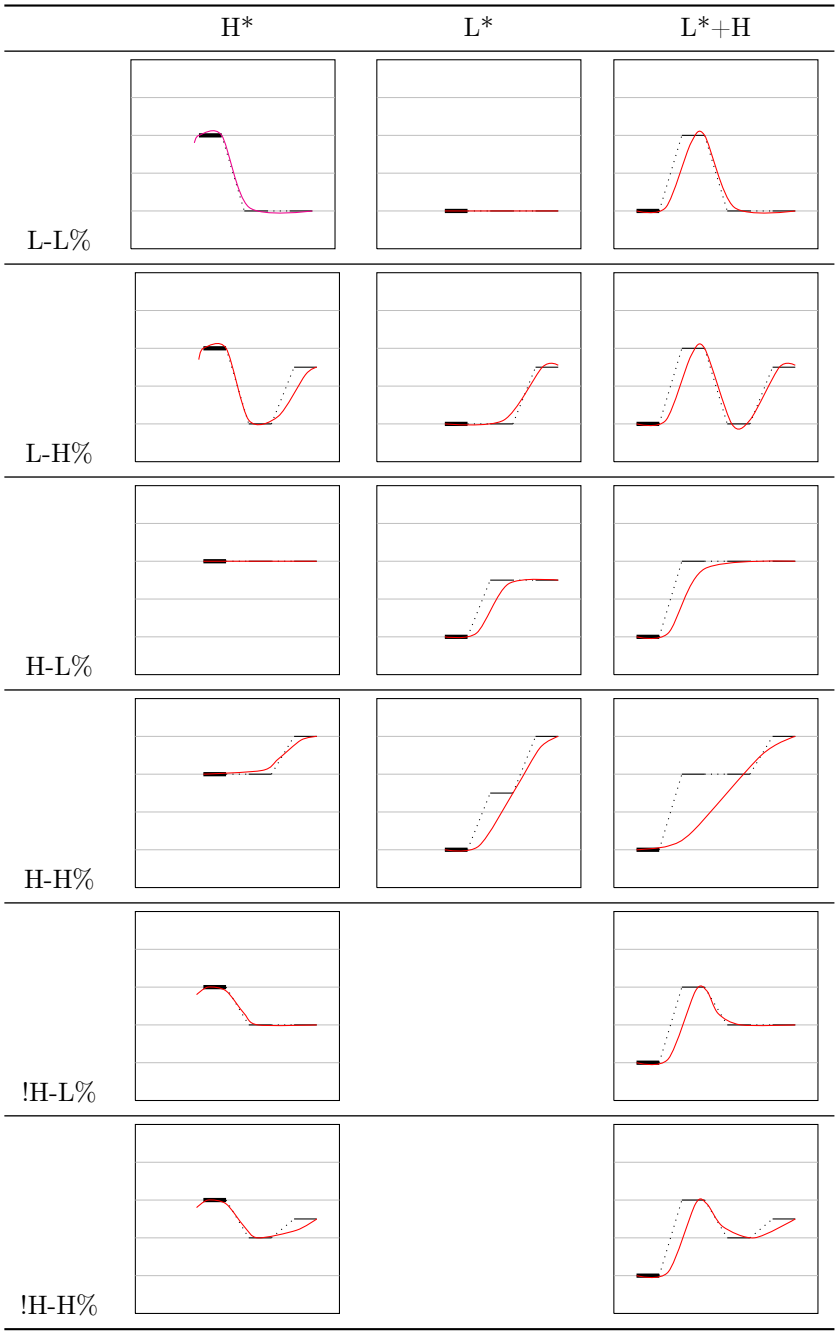


Figure 2.10: Total intonational patterns suggested by the ToBI system under foot segmentation. Each cell illustrates an intonational pattern under certain combinations of accent tone and phrasal tone in an one-foot intonational phrase. The theoretical pitch movement of a target tone is illustrated by a short black horizontal solid line. The starred target tone (pitch movement on the stressed syllable) is differentiated from other tones by a bold solid line. The red lines represent the theoretical smooth pitch contour.

The motivation behind the phonological-based approaches is that the current intonational models can not precisely predict all intonational characteristics. Therefore it is beneficial to determine the observed intonational characteristics of a given utterance as much as possible; however, a drawback of the current phonological approaches is that they are not sufficient. Phonological-based approaches cannot fully model the intonational properties due to their limitation to represent the F_0 contour changes between level tones. In this section we only focus on phonetic-based approaches.

Phonetic models: In phonetic models (which are also regarded as quantitative models), intonational features are represented numerically in term of vectors of acoustic features or continuous parameters. More particularly, they represent intonation as a sequence of (time, F_0) pairs.

In phonetic-based approaches, some reasonable disagreement stems from the fact that intonational aspects are supra-segmental, which lead us into the second dimension:

Sequential vs. superpositional models: In sequential models, intonation is characterized as a sequence of distinct intonational events or targets that are generated left to right. The superpositional approach characterizes the F_0 contour as an overlay (or superposition) of several component contours of different temporal scopes. Longer scope components (such as phrase curves) model the global shape of F_0 contour over length of an IP. The shorter scope components (accent curves) model local F_0 contour changes associated with accented-stressed syllables.

The phonetic models can be sequential that the F_0 contour of an utterance results from interpolation between the estimated intonational events, superpositional that the F_0 contour of an utterance results from superposition of the components of different temporal scopes, or even combination of both.

In the following sections, Section 2.3.1 and 2.3.2 some basic theoretical assumptions underlying the traditional and more recent models are presented.

2.3.1 Traditional Intonation Models

2.3.1.1 Tilt intonation model

The Tilt model is a widely used sequential phonetic intonation model [157, 158]. This model considers the F_0 contour as a sequence of intonational events (pitch accents and boundary tones) with linear connections. Taylor proposed a continuous feature – tilt-value, which jointly uses amplitude (A) and duration (D) of rising and falling pitch movements to model each rise-fall

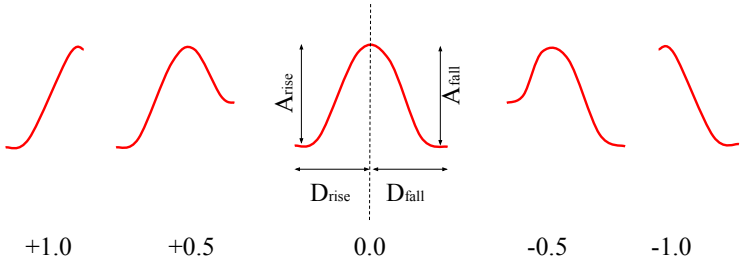


Figure 2.11: Example of five accent types with the continuous tilt-value ranging from +1 to -1

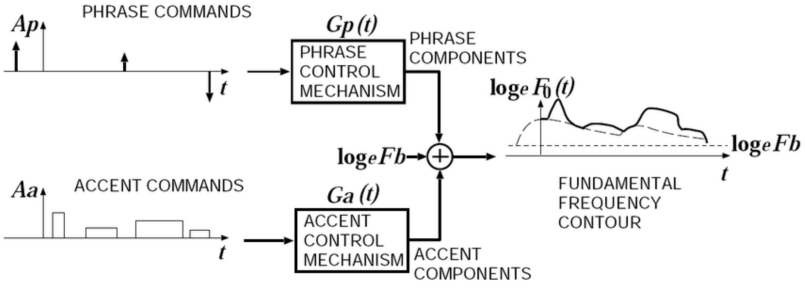


Figure 2.12: Block diagram of the Fujisaki model. From [40]

intonation event. Figure 2.11 shows an example of five pitch accents with continuous tilt-values ranging from +1 to -1. The tilt-value is formulated as follows:

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| - |A_{fall}|)} + \frac{D_{rise} + D_{fall}}{2(D_{rise} + D_{fall})} \tag{2.1}$$

2.3.1.2 The Fujisaki model

The Fujisaki model [41, 39, 40] is a superpositional phonetic model which is applied to the analysis and synthesis of intonation of different languages. This model has three major components: a baseline, a phrase, and an accent component (Figure 2.12). The baseline is equal to the minimum value of the log F_0 for the speaker. The phrase and accent components are modeled using second-order linear filters. The F_0 contour of an utterance results from the superposition (or sum) of the phrase accent components and the baseline.

The Fujisaki model only explains F_0 movements on "declining" utterances—those in which the F_0 contour starts at a higher value and gradually decreases during the phrase—while in some cases a rise in tone happens at the end of a question utterance. The treatment of declination as a fixed component of the model has been often criticized [54, 8], because declination is observed mainly in laboratory recorded speech. However, the biggest disadvantage of the Fujisaki model that it is

not entirely related to the linguistic structure. The phrase curve starts and ends with the start and end of a prosodic phrase, and it is not affected by which syllables are accented. The accent curves are not linguistically tied to a temporal scope since the starting point of an accent curve coincides with the start of an accented syllable, but the end point does not necessarily correspond to any syllable boundary.

2.3.1.3 The Generalized Linear Alignment Model (GLAM)

As previously mentioned the shared assumption in all superpositional approaches is that the F_0 contours can be described as an overlay (or superposition) of component curves that belong to one of several component curve classes. The General Superpositional Model (GSM) proposed by van Santen [166] ties these components to specific phonological entities, namely phrases and left-headed feet.

The definition of GSM can be formulated as follows where C corresponds to a set of curve classes, c represents a particular curve class, k stands for an individual curve and $\oplus$ is an operator.

$$F_0(t) = \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(f) \tag{2.2}$$

The $\oplus$-operator represents an addition-like (or in some cases multiplication-like) function of C. Therefore, this operator can satisfy the usual properties of generalized addition (or multiplication), such as monotonicity and commutativity:

$$monotonicity: \quad if\ a \geq b\ then \quad \begin{cases} a + c \geq b + c \\ a \times c \geq b \times c \end{cases} \Rightarrow a \oplus c \geq b \oplus c$$

$$commutativity: \quad \begin{cases} a + b = b + a \\ a \times b = b \times a \end{cases} \Rightarrow a \oplus b = b \oplus a$$

From the theoretical GSM model came the implementation of the Generalized Linear Alignment Model (GLAM) also developed by van Santen at Bell Labs [169, 164, 136]. This model considers a phonetic superpositional approach to intonation modeling. In this model, the intonation contours consist of three layers: phrase curves, accent curves and perturbation curves. For each layer, a different component curve class was considered. The model was implemented into a multilingual TTS system developed for English, French, German, Italian, Spanish, Romanian, Russian and Japanese.

Phrase curve: Similarly to the Fujisaki model, the phrase curve represents the long-term shape of the F_0 contour. However, unlike the Fujisaki model, it does not have any fixed gradients and

Algorithm 2.1 Accent curve generation algorithm for the linear alignment model

1: Determine the accent template type
2: Determine the F_0 peak position
3: Get the anchor values T_P
4: Calculate the anchor points T_A
5: Calculate the frequency values P_A corresponding to the T_A
6: Apply linear interpolation between P_A values
7: Return the interpolated curve multiplied by an amplitude parameter

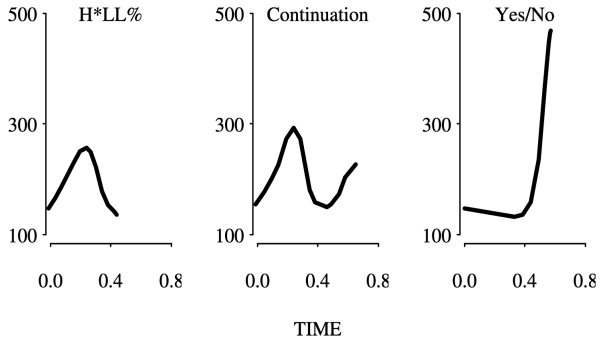


Figure 2.13: Averages of Declarative, Continuation, and Yes/No contours. From [169]

therefore the model has more degrees of freedom compared to the Fujisaki model. Phrase curves are modeled as piece-wise quasi-linear (or log-linear) curves consisting of a start point of the intermediate phrase, an inflection point at the start of the syllable containing the nuclear pitch accent, and an end point of the intermediate phrase.

Accent curve: Accent curves consist of pitch peaks and pitch movements associated with a foot segmentation. It is modeled by parameterized time warps of an accent curve template.

Algorithm 2.1 shows the required steps to generate accent curves. The first step is to determine the accent curve type based on the location of the foot in the intermediate phrase. The accent curve types are declarative template, continuation rise template, and interrogative template (Figure 2.13). In the second step, a template accent curve is defined using a sequence of anchor values. These values describe the archetypical shape of the associated template type. For example, for the declarative template, which employs a rise-fall pattern, the value of the template might be as follows:

$$T_p =< 0, 0.05, 0.2, 0.8, 0.9, 1, 0.9, 0.8, 0.2, 0.05, 0 >$$

The third step consists of determining F_0 peak position using information related to the foot duration and foot structure. This information includes: duration and phonetic composition of the

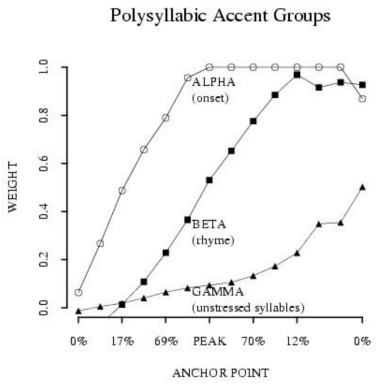


Figure 2.14: Alignment parameters in the linear alignment model From [164]

accented syllable's onset (D_o and C_o respectively), rhyme duration of the accented syllable (D_{rh}), and combined duration of the unaccented syllables (D_{rs}). Peak location is calculated using the equation below:

$$T_{peak} = \alpha_{C_o} \times D_o + \beta_{C_o} \times D_{rh} + \gamma_{C_o} \times D_{rs} \tag{2.3}$$

The fourth step creates a number of anchor points to obtain a good approximation of the accent curve shape (Algorithm 2.1). The Ath anchor point is located at a point on the time axis as computed in Equation 2.4. The alignment parameters (α, β,and γ) are extracted from Figure 2.14.

$$T_A = \alpha_{C_o,A} \times D_o + \beta_{C_o,A} \times D_{rh} + \gamma_{C_o,A} \times D_{rs} \tag{2.4}$$

The fifth step calculates the frequency points P_A using a linear time-warp function considering anchor points T_A, and anchor values T_P. A complete accent curve results from linear interpolation between successive P_A values. Finally, it is multiplied by an amplitude parameter that reflects the degree of emphasis. Figure 2.15 shows the prediction of two different normalized accent curves for two words "spot" and "noon" from one common template. The predicted normalized accent curve can be viewed as a time-warped version of a common template.

Segmental Perturbation Curves: These are short-term curves associated with those parts of the modeled F_0 contour where segmental effects occur e.g., initial parts of a sonorant following a transition from an obstruent.

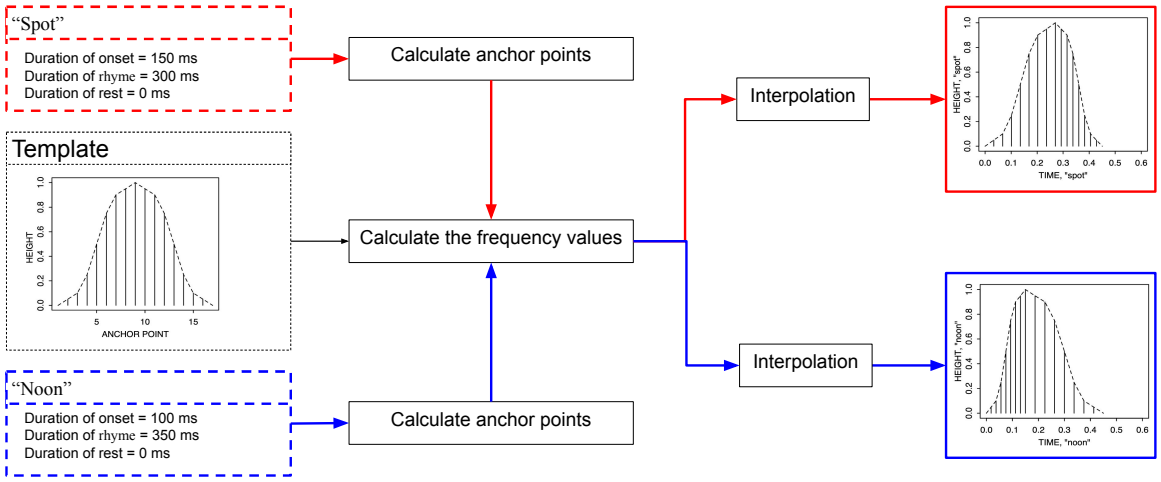


Figure 2.15: The prediction of two normalized accent curves for the words "Spot" and "Noon".

2.3.2 Recent Intonation Models

2.3.2.1 Quantitate target approximation

Quantitate target is a multi-language phonetic approach since it models the continuous F_0 contour with regards to intonational features at the syllable level [188]. However, it cannot be purely categorized as a sequential or superpositional approach. It models the intonation (or tone in tonal languages) as a sequence of target approximations (TA) which are syllable-synchronized. The F_0 contour in each syllable is modeled using two curves. A base line which represents the pitch target as a straight line with slope m and height b, and a combination of polynomial and exponential curves for representing the dynamic pitch target (Equation 2.5). Therefore, this model could be considered as a superpositional approach using a syllable segmentation (note that t in equation 2.5 is limited to a syllable). This model could also be considered as a sequential approach according to the sequentiality and syllable synchronization assumption, as it processes a syllable at a time.

$$F_0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \tag{2.5}$$

One of the advantages of this model is that the polynomial coefficients (c_1, c_2, and c_3) are not optimized as independent variables. They are a function of m, b, and λ. Therefore, this model consists of three independent parameters per syllable. Therefore, this model has potential to capture the F_0 dynamics in syllables level with few parameters that makes it a suitable method for analysis and synthesis of syllable-time languages (e.g., Mandarin Chinese). As we discussed in the first chapter, English is a stress-time language and considering a specific intonation movement

for weak syllables might cause overfitting.

2.3.2.2 Statistical phrase and accent models

Anumanchipalli et al. introduced the statistical phrase and accent model (SPAM) [7], which is based on a superpositional approach that decomposes the F_0 contour into phrase and accent components (the residual of the F_0 contour minus the phrase curve). They used an iterative Expectation Maximization algorithm to train the phrase and accent components. The phrase component was initialized by the minimum value of F0 over a syllable. They used the TILT representation for the accent shape at the syllable level; the F_0 is not modeled for unaccented syllables. Each accented syllable is represented as a tuple of four values: peak location, amplitude, duration and tilt-value. At each iteration, first, a decision tree (CART) is applied and K-means clustering performed for modeling phrase and accent components, respectively. Second the F_0 contour is estimated by adding the phrase and accent curves together. Third, the residual of the real F_0 contour and the estimated F_0 contour is added to the phrase curve. Finally, the residual of the real F_0 contour minus the updated phrase curve is used to update the accent estimations (TILT parameters). The main purpose of this model is to synthesis a high-quality and natural sounding F_0 contour.

2.3.2.3 F_0 contour decomposition using discrete cosine transform

Teutenberg and colleagues [159] use the discrete cosine transform (DCT) to model the F_0 contour. They propose a two-level model, one level for estimating the general movement of the F_0 contour (phrase curve), and a second level for estimating the details of the voiced regions, which is equal to the F_0 contour minus the phrase curve. They use the mean (the first DCT coefficient) of each voiced region's signal (as the phrase curve value), and the sum of the weighted cosine functions with zero phase for approximating the DCT. During analysis, they extract the DCT coefficients from the F_0 contour for voiced regions and apply a linear interpolation for filling in the unvoiced regions. For synthesis, they apply the inverse-DCT at each level separately. The estimate of the F_0 contour is equal to the sum of the result of the inverse-DCT at the two levels. The disadvantage of this model is that it does not consider textual information that affects the F_0 dynamic range (variance), such as lexical stress patterns in the current, previous, and next voiced regions, and the length of voiced regions based on the number of syllables and phonemes. Furthermore, the number of DCT coefficients is not fixed, and can be different for different speakers, which makes the modeling more challenging.

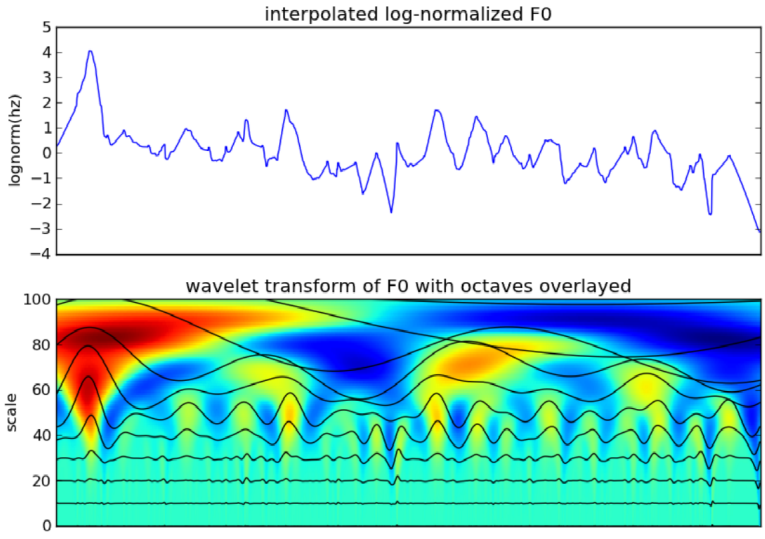


Figure 2.16: (adopted from [151]): Example of F_0 decomposition using continuous wavelet transform with 10 scales.

2.3.2.4 F_0 contour decomposition using continuous wavelet transform

Continuous wavelet transform (CWT) decomposes the F_0 contour into several frequency components where each component is distinguished through a scale. Based on the application, different number of scales are used for modeling the F_0 contour. Ming et al. used a five-scale CWT to model F_0 contours for emotional conversion[105]. In [151] ten distinct scales are used to model F_0 contours in different linguistic levels for synthesis purposes. The scales 0 and 1 correspond to the phone level, scales 2 and 3 correspond to the syllable level, scales 4 and 5 correspond to the word level, scales 6 and 7 correspond to the intonational phrase level, and scales 8 and 9 correspond to the utterance level. Figure 2.16 shows an example of an utterance decomposition using this method.

Ribeiro et al. [129] combined both DCT and CWT to explore a multi-level representation of F_0. The decomposition process can be summarized by the following steps: 1) A ten-scale CWT-based decomposition approach (identical to [151]) is applied to decompose F_0. 2) The number of scales is reduced to five corresponding to different linguistic levels: phone, syllable, word, intonational phrase and utterance level. 3) The contour in each scale is segmented by considering the corresponding linguistic level, e.g., the contour in the third scale is segmented at word boundaries. 4) For parametrizing each segment, an individual DCT is applied. Different coefficients are used at different levels: 6, 6, 4, 4, 3 coefficients are used for the phone, syllable, word, intonational phrase, and utterance level, respectively. Combining both DCT and CWT results in

more contribution of the higher linguistic levels in the naturalness of the synthesized speech. The successful use of this method in speech synthesis was the inspiration for other studies [105, 178, 92]. It is unclear if this method is suitable for analysis since these component curves were not meant to be linguistically meaningful.

2.3.2.5 Gamma distribution based decomposition

This superpositional model [56, 138] decomposes the F_0 contour into two component classes: the phrase curve class and the accent curve class (called "atom" by the author). The phrase component is the same as in Fujisaki's model while the accent components are modeled using the gamma distribution (Equation 2.6).

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma} t^{k-1} e^{-t/\theta}, \quad k = 2, \theta = 1/\alpha \tag{2.6}$$

The phrase curve and accent curves are estimated through two separate processes. First, an F_0 contour is decomposed into phrase and residual curve components using a greedy algorithm. Then, the same greedy algorithm is applied on the residual to estimate the accent curve parameters.

2.3.2.6 Procedure for Representing Intonation in the Superpositional Model

Procedure for Representing Intonation in the Superpositional Model (PRISM) is a superpositional phonetic model inspired by GLAM. PRISM decomposes a F_0 contour into three components curves: phrase curve, accent curve and perturbation curve [106].

Phrase curve: The Phrase curve is piecewise-linear, consisting of foot-length line segments. Compared to GLAM's phrase curve, this curve requires additional parameters ($n + 1$ compared to GLAM's three parameters per phrase curve containing n feet).

Accent curve: Accent curves in this model are a simplified version of accent curves used in the GLAM model. This simplification is applied in two steps, calculating anchor values (T_P, templates) and anchor points T_A. Unlike GLAM, the anchor points T_A are not extracted using information related to the foot structure (they are not calculated through Equation 2.4). Anchor points are n values sampled at equal time points (nine points was recommended by the author). The template corresponding to a rise-fall pattern, continuation rise pattern, and interrogative pattern are implemented by a Gaussian curve, summation of a Gaussian curve and a rising exponential curve, and rising exponential curve, respectively.

Perturbation Curves: These curves are modeled by a negative exponential curve.

Algorithm 2.2 PRISM two-phase decomposition algorithm

Phase 1: wavelet decomposition .....

1: Smoothing the F_0 contour
2: $Phr \leftarrow$ *Applying wavelet decomposition to the smoothed* F_0 *contour*
3: $Res \leftarrow F_0 - Phr$

Phase 2: template based decomposition .....

4: Determine the Res peak position
5: **for** each foot **do**
6: $RawAcc \leftarrow$ *get raw Res contour for current foot*
7: Get the template curve with n point values
8: $TmpAcc \leftarrow maximum(RawAcc) * template\ curve$
9: $EstAcc \leftarrow$ *apply n-point linear time-warp between TmpAcc and RawAcc*
10: Segmental Influence curves are parameterized in a manner similar to accent curves
11: Apply optimizer on accent curves and Segmental Influence curves parameters

PRISM's algorithm has two phases. In the first phase, a given F_0 contour is decomposed into a phrase curve and a residual curve using the discrete wavelet transform. The second phase consists of template based decomposition of the residual into accent curves and segmental perturbation curves. Algorithm 2.2 shows the steps required for F_0 decomposition using PRISM.

Similar to GLAM, PRISM has three component curves where each of the component curves is tied to a distinct phonological segmentation. There are certain aspects to PRISM that must be examined more closely. First, PRISM allows negative accent curves to model F_0 values that fall under the phrase curve. American English generally does not have negative accents. Second, PRISM uses nine parameters for estimating each accent, which, given the generally regular shapes of local pitch excursions should not be necessary – fewer parameters, such as location, width, and asymmetry, should suffice. Third, it uses n + 1 parameters to model phrase curve that undermine the perceptual relevance of the phrase curve because there is no global declination. Fourth, PRISM optimizes the phrase and accent curves separately, which is prone to local minimum problems.

2.4 Intonation in Text-To-Speech (TTS) Systems

The aim of TTS systems is to synthesize intelligible and natural sounding speech waveforms from the input text. Most traditional TTS systems consist of two phases: a front-end, which converts the input text into an abstract linguistic representation, and a back-end, which generates the speech waveform along with the prosody of the sentence to be spoken using the linguistic information. Figure 2.17 shows the general schema of a TTS system.

The objective of this section is not to provide a comprehensive account; rather it samples the

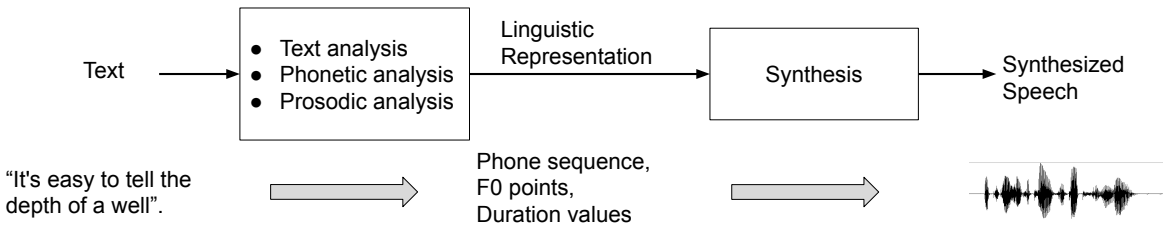


Figure 2.17: Schematic diagram of a speech synthesis system

common prosodic features and models that are used in the back-end phase of TTS systems under two categories: intonation synthesis and intonation adaptation.[2]

2.4.1 Synthesis

The methods for synthesizing F_0 in speech synthesis are very diverse, ranging from rule-based methods in older systems whereby F_0 contours are generated by rule and then imposed onto a concatenated sequence of stored acoustic units [146], to statistical parametric based synthesis in which F_0 is generated frame-wise in parallel with spectral frame generation and is, similarly, imposed onto spectral frames [192], to unit selection systems where the database is sufficiently rich that stored F_0 can be used as-is [126].

Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) were the two most common acoustic models used in parametric TTS systems. However, their limitations, such as data disjointness caused by decision-trees (which are used to represent complex, nonlinear relationships between the input text and the acoustic features) have motivated researchers to use deep neural networks (DNNs).

There are two main challenges when it comes to modeling F_0 for synthesis purposes. The first challenge is that there are only F_0 observations within voiced speech regions. The question is how F_0 values in unvoiced regions should be represented. The second challenge is capturing the suprasegmental properties in F_0 movements. As we discussed in Section 2.2, considering a phonological unit that is larger than the syllable but does not coincide with word boundaries is more suitable for capturing the suprasegmental properties of F_0 movements. However, most HMM-based synthesizers predict F_0 at the frame level using limited linguistic contextual information. This frame-by-frame prediction of F_0 results in an overly-smooth F_0 contour that cannot properly

[2] A more comprehensive account is given in [71] which provided an overview of the evolution of the TTS system from its early ages till today.

Reference	Unvoiced F_0 representation	Intonation model	Model domain
[38]	Random values generated from a probability density function (pdf) with large variance	continuous HMM	Frame-level
[62]	Zero	continuous HMM	Frame-level
[194, 195]	Interpolated	GMM	Frame-level
[154]	Interpolated + Low Pass Filter	GMM	Frame-level
[80, 128]	Interpolated	DCT	Syllable-level and phrase-level
[129]	Interpolated	CWT	Syllable-level word-level and phrase-level

Table 2.3: A comparison of several approaches for F_0 contour modeling of HMM-based TTS systems. Approaches are classified into three categories: unvoiced F_0 representation, intonation model, and model domain

represent the suprasegmental properties of F_0 movements. In our review of approaches, we will focus on these two issues.

2.4.1.1 HMM-based approaches

Hidden Markov Models (HMMs) in synthesis are stochastic generative acoustic models that generate an observation sequence given a discrete hidden state sequence. Typically, the spectrum and F_0 are modeled in separate streams due to their different characteristics and time scales. To model missing F_0 values in unvoiced regions, multi-space probability distributions (MSD) are usually used in HMM-based synthesis systems [160, 190, 101]. The MSD-HMM uses a discrete HMM to model the F_0 values for unvoiced frames and a continuous mixture HMM to model the F_0 values for voiced frames. The first limitation of this approach is that it is sensitive to voicing classification errors. One solution to this is to assume that F_0 is continuous in unvoiced regions as well [194]. The second limitation is that frame-by-frame prediction of F_0 values results in overly-smooth F_0 contours. In order to capture the prosodic patterns on a larger scale and to generate more natural F_0 contours, superpositional approaches are used[80, 129, 128]. Table 2.3 gives a summary of these approaches.

2.4.1.2 DNN based approaches

Many articles on speech synthesis report that the usage of deep learning techniques shows improvement over HMM-based approaches in terms of naturalness, similarity, and quality of the generated

Reference	Unvoiced F_0 representation	Intonation model	Model domain
[87, 86, 198]	Undefined	MSD	Frame-level
[198]	Interpolated	[194]	Frame-level
[91]	Interpolated	DNN applied on a vector-space representation of input texts	Frame-level or state-level
[68]	Zero	hybrid approach between DNN and Gaussian process based regression	syllable-level
[131]	Interpolated	SPAM + LSTM	syllable-level

Table 2.4: A comparison of several approaches for F_0 contour modeling of DNN-based TTS systems. Approaches are classified into three categories: unvoiced F_0 representation, intonation model, and model domain.

speech [87, 91]. Some only apply DNN models on spectral modeling and keep the prediction of F_0 values identical to HMM-based approaches [86, 198], while others use DNN models directly for predicting F_0 contours [36, 91, 68]. Kang et al. used a deep belief network as a generative model for the joint distribution of linguistic and acoustic features [68]. They suggested that the low-dimensional F_0 features are not modeled well when combined with high-dimensional spectrum features. They used a combination of discriminative DNN and Gaussian process-based regression to predict log F_0 values. First, a DNN is trained to map linguistic feature to log F_0 values. The activations at the last hidden layer are then used as the input for the Gaussian process based non-parametric regression. In [91], a DNN is trained on vector-space representations of linguistic context. This vector-space representation was derived without using any linguistic resources. In [131] a template-based approach was explored. A simplified LSTM classifier was used to predict a template at the syllable-level using textual information. These templates are extracted using the SPAM model (see Section 2.3.2.2) from training data. Table 2.4 gives a summary of these approaches.

The main problem of statistical parametric TTS systems is that they are typically composed of many domain-specific modules (e.g., a text analyzer, an F_0 generator, a spectrum generator, etc.). These modules usually are trained independently, so errors from each module may compound and result in a complex TTS system [179, 153]. More resent methods use the sequence-to-sequence deep learning technique to merge these internal modules into a single model that directly connects the input text to the output audio (this technique is called end-to-end TTS). The end-to-end TTS systems based on sequence-to-sequence techniques are commonly RNN-based [145]. However,

due to RNN-based disadvantages (long-term dependencies, and CPU time consuming), attention-based mechanisms [177, 83] and CNN-based learning models have been proposed [117, 42]. The end-to-end TTS models obtained better performance over a statistical parametric speech synthesis system in terms of naturalness; However, it still remains a challenge to control the synthesis model to generate speech with desired intonational characteristics (e.g., emotion) [176].

2.4.2 Intonation Adaptation

In TTS adaptation, the aim is to transform the perceived identity of a TTS voice to that of another speaker. To clarify, in the case of TTS, the source speaker is the speaker whose recordings were used to generate the acoustic units (for unit selection approaches), acoustic inventory (for diphone based synthesis), or acoustic features for HMM or DNN approaches. This speaker's recordings may also be used as training data for prosody mimic. Thus, the speech generated by a TTS system generally sounds like the source speaker. For prosody mimic (intonation adaptation), the challenge is to compute a transformation that, when applied to the speech data or to any representations thereof, generates output speech mimicking a target speaker.

Most TTS adaptation papers are focused on spectral features, and they use trivial methods to modify prosody [108, 182, 107]. Typically, F_0 is represented by just its mean and the standard deviation (SD); thus, during synthesis, the output utterance will match only these target speaker features without attempting to capture the dynamic details of the speaker's prosodic style [18]. In a more sophisticated approach, Chappell proposed a linear transformation that globally maps mean and standard deviation of F_0 values in utterance level [18]. Patterson went a step beyond Chappell's approach and used four types of data points in an utterance to represent F_0 [119]. For given an utterance, they selected the sentence-initial F_0, the sentence-final F_0 values, all the non-initial pitch accent peaks and all the post pitch accent valleys. The main drawback of these mapping methods is that they cannot fully capture dynamic patterns of F_0 contour. HMM-based [155, 60] and superpositional [169, 37, 33, 172] approaches are potentially more accurate and practical methods for capturing intonation.

Intonation can be transformed at different levels (listed is column in Table 2.5): frame [155, 18, 44, 35], tone [175] syllable [52, 90, 156, 60, 175, 59], word [3], sequence of syllables [60, 59, 61] and sentence [18] with different methods (listed is column in Table 2.5). As mentioned, the most common method to transform F_0 is by globally matching the mean and SD of the target speaker's F_0 contour. The mean and SD values of the source and target speaker's F_0 contours are used to define a linear transformation that is applied to the source speaker's F_0 contour, typically in

Approach	Adaptation method	Adaptation domain	Intonation model	model domain
[18]	Linear	Frame-level	Average(mean and SD) of raw F_0 contour	sentence-level
[18]	Polynomial conversion	Frame-level	Scatterplot model of mean F_0	Phone-level
[44]	Piecewise linear mapping	Frame-level	Pitch range model	Accent-level and sentence-level
[156]	Linear modification	Syllable-level	Raw F_0	Syllable-level and phrase-level
[156]	GMM	Syllable-level	Pitch target model	Syllable-level
[172]	GMM	Syllable-level	DCT + multi-level dynamic features	Syllable-level and phrase-level
[60]	Data-driven F_0 segment selection	Sequence of syllable	MSD-HMM	Syllable-level
[156]	CART	Syllable-level	Pitch target model	Syllable-level
[52]	Codebook+CART	Syllable-level	DCT	Syllable-level
[18]	Contour codebook + DTW	Sentence-level	Raw F_0	Sentence-level
[155]	MSD-MLLR	Frame-level	MSD-HMM	Frame-level
[90]	MLLR	Syllable-level	GMM	Syllable-level

Table 2.5: A comparison of several prominent intonation transformation approaches. These techniques are classified into four categories: adaptation method, adaptation domain, intonation model, and model domain.

the log domain [18]. Extensions of this approach include higher-order polynomial [18], piecewise linear transformation [44] and linear modification based on hand-labeled intonational (syllable-phrase) features. Another class of methods predict intonation by modeling F_0 and spectral features jointly [93, 49, 184]. In cases where limited amounts of data are available, statistical techniques are usually utilized to extract the mapping function. The most popular technique is based on a Gaussian mixture model (GMM) [156, 172, 35, 59, 9]. Two other methods use F_0 contour codebooks [18] and parametrized codebooks [52, 59]. Weighting multiple contours has shown a minor performance improvement [162]. Various other methods, such as hierarchical models [180],

CART [52, 156] and MLLR [155, 90] are proposed.[3] However, it still remains a challenge to generate speech with intonational characteristics of the target speaker when data are limited.

2.5 Intonation in Speaker State Classification

The aim in speaker state classification (SSC) is to recognize the speaker's state using paralinguistic features (and/or linguistic features). Typical problems include the recognition of a speaker's emotion, age, gender, identity, and health. The process of this classification usually consists of three steps: feature extraction, feature selection, and classification. We focus only on the feature extraction step.

Most approaches in feature selection extract a large number of acoustic features from the speech signal and use standard machine learning techniques as a black box often achieving good classification accuracy. However, there are two drawbacks to this common approach. First, they are often not informative for scientists working in the domain field (e.g., autism researchers), because they are interested in finding which features are the most important ones for classification and why. For example, just knowing that a classifier performs at 90% accuracy fails to answer these questions. Of course, in certain industrial or governmental applications, classification accuracy is the primary or even sole interest. Second, these approaches require that the recording conditions – microphone, room acoustics, distance to microphone – are not in the least bit confounded with the classes under consideration. The large number of acoustic features may capture differences in recording conditions, so that the final classification result may have little to do with the classes of interest. This is particularly dangerous in multi-site data collection efforts in which each site is responsible for recording a specific class. To combat these issues, some researchers have turned to the use of prosodic features, which is discussed next.

In the last two decades, the usage of prosodic features have shown an improvement in the performance of classification systems. Prosodic features can be grouped with respect to two factors:

1. The temporal structure used for feature extraction: a distinction is drawn between short-term and long-term temporal structures. The short-term features, which are also referred to as segmental features, are extracted for every frame (typically 25ms in length). Long-term features, which are also called suprasegmental features, are extracted at the utterance level (or continuously voiced regions separated by a pause). However, other linguistic units (e.g., syllables) have gotten more attention in recent years (first two columns in Table 2.6).

[3] A more comprehensive account is given in [181].

Features	Level	Low-Level-Descriptors (LLDs)	Functional
segmental	Frame	Frame-energy, Frame-intensity	
suprasegmental	Utterance	F_0, energy, intensity, harmonics-to-noise ratio, shimmer, jitter, normalized amplitude quotient, duration	Extreme values (maximum, minimum), mean, moments (standard deviation, variance, kurtosis, skewness), percentiles of non-zero frames, duration in seconds
	Syllable	F_0, intensity, duration, F_0 residual, F_0 regression, intensity regression	Extreme values (maximum, minimum), mean, local range (span), gradient, voiced-unvoiced ratio

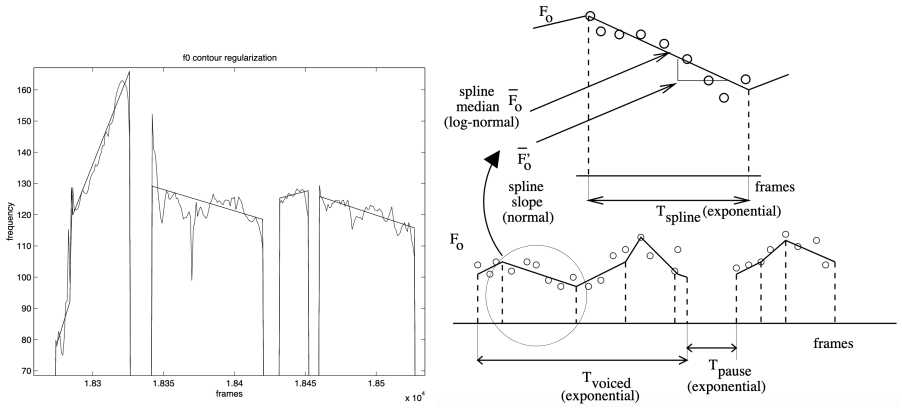
Table 2.6: Categorization of prosodic features in terms of the linguistic unit and parametrization.

2. The level of feature descriptors: the features can be in two levels: Low-Level-Descriptors (LLDs), and functionals. LLD features consist of prosodic features at both the segmental and suprasegmental level. The functional features are statistical features that are derived from suprasegmental LLD features (last two columns in Table 2.6).

Table 2.6 summarizes categorization of common prosodic features according to the above factors. As we discussed in Section 2.2 and Section 2.4, prosodic features (especially F_0 contours) have suprasegmental properties and frame-level segmentation is too short for capturing these properties. As in found in [186, 150], in order to produce the smallest meaningful F_0 movement, a longer span of time is required (in average $100ms$). Even though short-term features cannot properly represent the prosodic characteristics, these features could be effective when: large amounts of data are available [2], they are combined with spectral features (especially in noisy conditions) [70], log F_0 or normalized F_0 are used instead of the raw F_0 [69, 113, 144, 140].

It has been shown that the prosodic features (such as F_0, intensity, and energy) are more effective when they are extracted at the utterance level [140, 17, 67, 45]. However, utterance level features listed in Table 2.6 cannot properly convey suprasegmental properties of prosody since these statistics fail to capture local F_0 dynamic changes (specifically in long duration utterances with multiple pitch accents) [143]. Generally, researchers took two ways to face this issue. The first one is to consider using features that represents local F_0 dynamics, and the second one is to consider analyzing prosody in linguistic units shorter than an utterance and longer than a frame or phoneme (e.g., syllable).

When using short-term based techniques, adding the delta F_0 can help capture some of the local



(a) F_0 regularization using a piecewise linear model (b) Piecewise linear model's parameters

Figure 2.18: F_0 regularization and feature extraction using a piecewise linear model in a long-term segment (continuously voiced regions separated by a pause). From [143]

F_0 dynamic information. When using long-term based techniques, several methods are used for capturing local F_0 dynamics. For example, in [2] F_0 and energy trajectories are used to determine the corresponding slope of the contour in the voiced region. (+ for rising and - for falling slope). For each segment, F_0 and energy symbols are joined together (e.g., ++, +-, -+, --, and uv for unvoiced regions). Then, a sequence of these symbols are used to represent the long-term features. In [143], first a regularization is applied on the F_0 contour using a Piecewise Linear Model (see Figure 2.18a), then Piecewise Linear Model features (segment median, segment slope, and segment duration) and durational features (duration of the voiced segment and pause duration) are extracted from each continuously voiced region (see Figure 2.18b).

Regarding the second solution, usually a segmentation method is applied on the F_0 contour to split the F_0 contour into a sequence of smaller segmented F_0 contours. Each segment is represented by a set of features. Typically syllables are used as the segmentation unit. The most common way to produce segmentation automatically is by using automatic speech recognition (ASR); however due to ASR limitations in some areas, such as emotion and language classification, some ASR-free approaches have been proposed. Segmentation into syllable-like regions is usually accomplished with the knowledge of vowel onsets [100, 4, 98] or F_0/energy contour valley points [26].

In recent years, there have been a number of studies on syllable-based analysis of prosodic features. In [99], the author used Tilt parameters to represent the dynamics of F_0 contours in syllable-like regions. A total of seven parameters were used for each segment: mean value of F_0, peak F_0, change of F_0 (delta), distance of F_0 peak to vowel onset, amplitude tilt, duration

tilt, change of log energy (delta). A polynomial function is used to approximate F_0 and energy contours in syllable-like segments [26]. Each segment is represented by a feature vector consisting of the polynomial coefficients (for both F_0 and energy contours) and segment duration. The same method is also used in [27] except that in this case the coefficients are time-normalized. Raymond et al. have investigated a large set of prosodic features in syllable-like units [113, 112, 114]. These features consisted of F_0, energy, and duration. In addition to the raw F_0 and raw energy contour, normalized, regression, and residual contours(where the phrase curve is subtracted from the corresponding contour) were also included. Five features are extracted for each of the following features of F_0/energy: value in the syllable nucleus, maximum, minimum, range (or span) and gradient. Three duration features are considered: the time interval between two neighboring syllable nuclei, syllable duration, and the ratio between the length of the F_0 contour and the syllable duration. Figure 2.19 demonstrates these prosodic features.

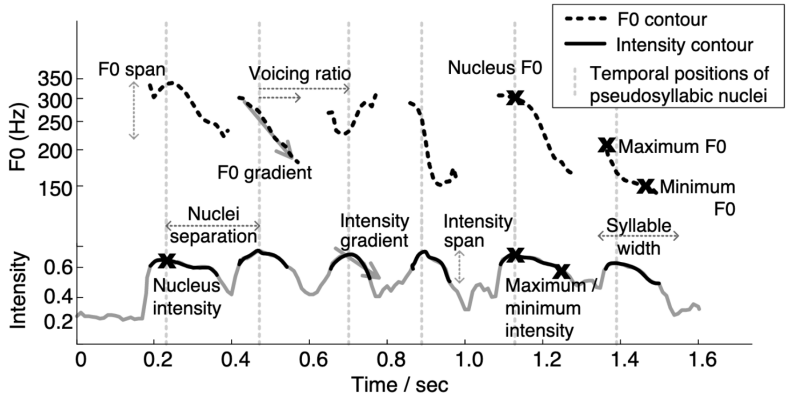
2.6 Evaluation

Evaluation methods can be split into two categories: objective and subjective. Objective evaluation in intonation measures the goodness of fit between estimated F_0 contours and the original contours. The root mean square error ($RMSE$), which is known as the standard error, is a widely used objective measure in intonation. Based on a rule of thumb, the lower the value the better the model can relatively estimate the F_0 contour. Objective measures are popular because they are objectively unchangeable and easy to calculate; however, in intonation research it is important to realize how these results are being interpreted.

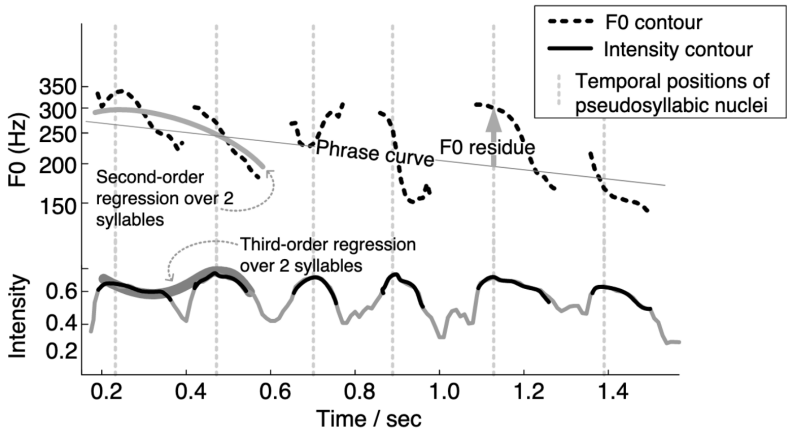
RMSE can not explain type of measurement error. For example, if the $RMSE$ is $15Hz$, it means the square root of the average squared difference between the estimated F_0 contour and the original one is $15Hz$. RMSE does not clarify that it results from high error in few outlier points (e.g., caused by halving/doubling error or gross error) or very small error across all points (e.g., random error), which both could have the same RMSE but might be perceptually quite different.

RMSE does not show relative values. For instance, a $15Hz$ difference is considered a bad fit when studying a synthetic F_0 contour of a male adult in read speech data, while a $15Hz$ difference is considered a suitable fit when studying a synthetic F_0 contour of a child in emotional-based data. In other words, human ears are not sensitive to a $15Hz$ difference when the base frequency is really high (e.g., a happy child).

In order to determine how the human ear can distinguish between frequencies, experts divide the frequency range of human hearing (20-20kHz) into eleven octaves. Octaves are not equally



(a) Raw prosodic features at the syllable level



(b) Prosodic regression features at the syllable level

Figure 2.19: Representation of prosodic features at the syllable level for an example utterance. From [114].

spaced in frequency. Lower octaves cover a narrower frequency range than higher octaves, since the human ear can more easily distinguish between frequencies in lower frequencies (the human ear is more sensitive to low frequency changes). This division is done in such a way that each octave covers double the frequency range of the previous octave. To more closely match how humans distinguish frequency, each octave can be split into 12 semitones. The human ear can distinguish only one semitone differences. The semitone is formulated, as in Equation 2.7, using a *baseline* frequency.

$$S : semitone = 12 * \log_2(f/baseline) \tag{2.7}$$

For example, if we are studying a male adult, the *baseline* is $50Hz$, while the *baseline* is $300Hz$ in

an happy child. The first semitone difference corresponds to a frequency differential of $3Hz$, and $18Hz$ for a male adult and a happy child, respectively, as we will now show.

$$f_0 = baseline \tag{1.2}$$

$$\begin{aligned}
\Delta S = S_{n+1} - S_n &= 1 \\
12 * \log 2(f_{n+1}/baseline) - 12 * \log 2(f_n/baseline) &= 1 \\
\log 2(f_{n+1}/baseline) - \log 2(f_n/baseline) &= 1/12 \\
\log 2(f_{n+1}/f_n) &= 1/12 \\
f_{n+1}/f_n &= 2^{1/12} \\
f_{n+1} &= 2^{1/12} * f_n
\end{aligned} \tag{1.3}$$

From Equation 1.2 and Equation 1.3 we can conclude that $f_1 - f_0 = baseline(2^{1/12} - 1)$. Therefore, the smallest frequency difference perceivable by the human ear at the speaker's baseline for:

- male adult is $50 * (2^{1/12} - 1) = 2.9731$ Hz ≈ 3 Hz
- happy child is $300 * (2^{1/12} - 1) = 17.8389$ Hz ≈ 18 Hz

Due to the above calculation, in many studies the use of log F_0 is preferred over raw F_0. The RMSE provides a sense of how close (or far) estimated F_0 values are from the raw F_0; However fails to clarify how well the model explains the shape of F_0 contour. In this case, correlation between estimated F_0 contour and raw F_0 contour can be used as an evaluating measure. In above example, if the correlation is high (e.g., 0.8) regardless of RMSE value, then the model considered as a good estimation of the F_0 contour that explains 80% of the shape of F_0 contour. The higher the correlation value is, the more precise is the model. Furthermore, when dealing with the objective evaluation of real speech data, it is important to either normalize the data before analysis or find a reliable measure for comparison afterwards.

Another way to evaluate the effectiveness of a method is by using subjective evaluation approaches. These methods have been quite diverse, since subjective evaluation is subject to human interpretation. Evaluating the naturalness of the estimated F_0 contour has been the most frequently used subjective evaluation approach. A common method is to ask subjects to listen to a generated utterance and judge the naturalness of the utterance on a five-point scale (e.g., 1:

bad, 2: poor, 3: fair, 4: good, and 5: excellent). Another method is to ask subjects to do an A/B testing, when two systems are being tested against each other, to pick which of two is most natural on a five-point scale (e.g., -2: definitely prefer the first utterance, -1: probably prefer the first, 0: neither, 1: probably prefer the second utterance, 2: definitely prefer the second). Also we could ask subjects questions to get at their understanding/interpretation of the utterance: "Is the speaker sad?", or "What word is most emphasized?".

During subjective evaluation, it is important to control how subjects are being instructed. For example, in evaluation of speech in a news-reading speaking style, asking the subjects to rate the naturalness of an utterance is adequate, and further clarification may not be needed. However, in evaluation of other types of speech (e.g., clear vs conversational speech, emotional speech), a slight change in instruction (e.g., providing content of utterance with or without punctuations) may result in different ratings. Although it is important to clarify the instructions given to the subjects, it is unethical to lead them in a specific direction. For example, assume that we are interested in evaluating the ability of our model to handle marked-up input to design a contrastive emphasis test. The content of the utterance under test is "This is a CHEAP car", where capitals indicate an emphasized word according to a contrastive choice. It would be leading if the subjects are asked to answer this question "Is the word "cheap" emphasized?" or "Is this utterance the answer to: What kind of car is this?". In the first question, we asked the subjects to pay particular attention to the word "cheap", therefore even a small emphasis on the word "cheap" would lead the subjects to answer positively to the question. In the second question, we asked the subjects to pay particular attention to the adjective, therefore any noticeable emphasis on the word "cheap" would lead the subject to answer positively to the question. In both cases, a high rating suggests that our model is emphasizing the word "cheap", but does not indicate that our model correctly puts contrastive emphasis on the word "cheap".

Chapter 3

GENeralized Intonation model for English (GENIE)

In this chapter, we propose a new generalized intonation model for English (GENIE).[1] GENIE is inspired by the General Superpositional Model (GSM) [169, 164, 136]. In Section 3.1, we explain what the shared assumptions are between GENIE and GSM, and how it differs from GSM's other implementations (namely GLAM and PRISM). In Section 3.2, we present the details of GENIE. GENIE like GLAM and PRISM is a superpositional intonational model that provides the high-quality prediction of F_0 contours, but unlike them it uses a very small set of parameters which are optimized simultaneously, and it focuses not only on the synthesis of English intonation but also on the analysis of English intonation. Finally, in Section 3.3 we use GENIE to show it can produce accurate and linguistically meaningful results.

3.1 GENIE model properties

In Section 3.1.1, we summarize the underlying assumptions of GSM, which directly inspired the creation of three models: GLAM, PRISM, and GENIE. Then, we discuss the shared assumptions and differences between GENIE and its cousins. In Section 3.1.2, we introduce the additional assumptions of GENIE.

3.1.1 Fundamental assumptions

GSM is a theoretical framework. It was developed by van Santen [164], which we discussed it in details in Chapter 2. The idea behind GSM was that although there are several superpositional

[1] This chapter is based on work published in ICASSP [33].

approaches with different assumptions, they are all part of one bigger family and they can be represented by one single formula. GSM has two core assumptions:

Assumption 1: F_0 contours can be described as an overlay (or superposition) of underlying component curves that belong to one of several component curve classes

Assumption 2: Each class of curves corresponds to a distinct temporal scope.

Based on these GSM assumptions, the F_0 contour of an utterance can be modeled using an overlay of component curves that belong to one of several component curve classes where each corresponds to a distinct temporal scope. This definition is sufficiently general to formulate any superpositional model as follows where: C corresponds to a set of curve classes (e.g., phrase curve class and accent curve class), c represents a particular curve class of C (e.g., accent curve class), k stands for an individual curve of c's class (e.g., rise-fall accent curve), and $\oplus$ is an operator.

$$F_0(t) \approx \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(t)$$

This formula is very general and may not have obvious testable predictions. A successive narrowing down would lead to such predictions.

The Generalized Linear Alignment model (GLAM) is the first direct implementation of GSM. It was also developed by van Santen at Bell Labs [169, 164, 136]. The objective of GLAM was to be used as a generative intonational model in a multilingual TTS system. We characterize GLAM through an additional set of assumptions and compare them with GSM's. GLAM has stronger assumptions than GSM since it focuses on TTS applications, but is still general enough to cover multiple languages.[2]

Assumption 1.1: This assumption refines Assumption 1 from GSM. An F_0 contour — interpolated in unvoiced regions — can be decomposed into component curves: a phrase curve ($P(t)$ in Equation 3.1) and a sum of one or more accent curves ($A(t)$ in Equation 3.1).

$$F_0(t) \approx P(t) + A(t) \tag{3.1}$$

Assumption 2.1: This assumption refines Assumption 2 from GSM. The Phrase class is tied to an intermediate phrase (discussed in Section 2.2.1), and the Accent class corresponds to a foot(discussed in Section 2.2.3), which is a shorter scope than an intermediate phrase and consists of an accented-stressed syllable followed by with zero or more unaccented syllables.

[2]The validity of these assumptions has been tested in many research projects [164, 169, 166, 72, 136, 137].

(a) Rise_fall accent curve (b) Yes-no question accent curve (c) Continuation accents

Figure 3.1: Three different accent categories

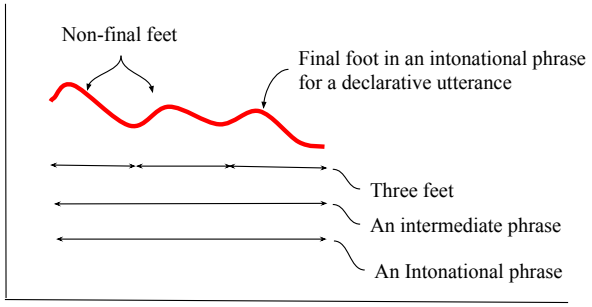


Figure 3.2: Foot structure in a statement utterance

Assumption 3: The Phrase class should be smooth over long time stretches, which enables us to determine the effect of intonational characteristics and functions on the component curves.

GLAM uses a phrase curve consists of two quasi-linear segments, the first from the phrase start (p_s) to the start of the final foot in the the phrase (generally associated with the nuclear pitch accent, p_f), and the second from the latter to the end point of the last voiced segment of the phrase (p_e).

Assumption 4: Three different accent categories are used to estimate the three intonational patterns: Rise-fall accent groups (e.g., H*L-L%, Figure 3.1a), yes-no question contours (e.g., L*H-H%, Figure 3.1b), and continuation contours (e.g., H*L-H%, Figure 3.1c). Rise-fall accents occur in any non-final feet as well as in the final foot in an intonational phrase for a statement utterance. Figure 3.2 shows occurrence of three rise-fall accents in an intonational phrase. Continuation contours consist of a dual motion in which an early peak is followed by a valley and a final rise. A continuation accent occurs at the final foot in an intermediate phrase that is not aligned with the end of an intonational phrase. In Figure 3.3, the accent curve in the final foot of the first intermediate phrase is a continuation accent. Yes-no question accents occur on the final foot in an intonational phrase for a yes-no question and consist of an accelerated decrease starting at the onset of the accented syllable, followed by a steep increase in the nucleus (in Figure 3.3, the accent curve in the last foot).

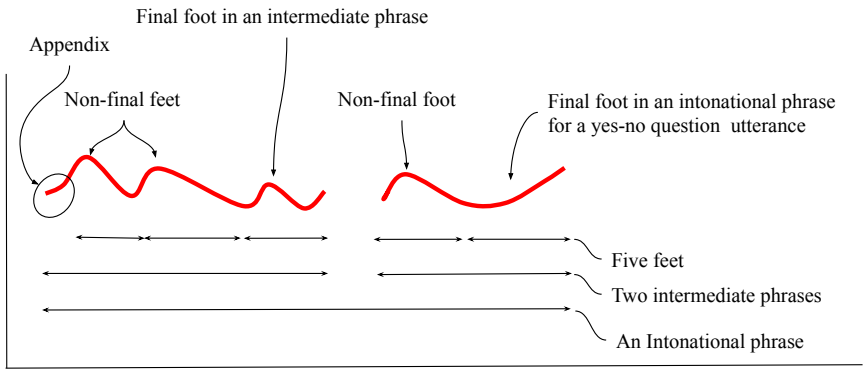


Figure 3.3: Foot structure in a yes-no question utterance that consists of two intermediate phrases.

Assumption 5: A specific accent curve cannot be assigned to phrase-initial unstressed syllables – also called appendix – since these syllables are not part of any foot. In Figure 3.3, we can see the appendix does not belong to any particular foot, but it is still part of the intermediate phrase.

The Above assumptions were developed intentionally to make GLAM a suitable and flexible synthetic intonational model. These assumptions are shared among GLAM and GENIE. Implementation-wise GLAM has three more assumptions to make it suitable across languages. The rest of assumptions in this section are not part of GENIE's shared assumption.

- Accent curves consist of pitch peaks and pitch movements associated with syllables (e.g., a foot in English or one syllable in Mandarin). It is modeled by parameterized time warps of an accent curve template.
- Overlap is allowed only between successive accent curves
- Segmental Influence Curves are also considered as a component class, which are added to the Phrase and Accent class to estimate F_0 contours

The first item was developed to give GLAM flexibility to consider all intonational patterns in different languages. The second and third items were added to improve the voice quality of synthesized speech. There are several drawbacks to these assumptions when it comes to real-world cases. 1) Optimization of component curve parameters can not be done simultaneously since some curves require independent preprocessing (pitch peak detection for accent curves and vowel onset detection for segmental influence curves). 2) Beside the actual parameters, there are some hyper-parameters that need to be tuned, such as the number of anchor points, and the degree of overlap

between two successive accent curves.

The Procedure for Representing Intonation in the Superpositional Model (PRISM) was a second direct implementation based on GSM. Like GLAM, PRISM is a synthetic intonational model for a TTS system, but it only considered American English. PRISM adopted most of GLAM's assumptions, but differed from GLAM in terms of the following items:

- The phrase curve is piecewise-linear, consisting of foot-length line segments instead of the two line segments allowed by GLAM. However, this introduces additional parameters in the process (n + 1 instead of 3 parameters per prosodical phrase containing n feet), and may also undermine the perceptual relevance of the phrase curve because there is no global declination.
- PRISM allows negative accent curves to model F_0 values that fall under the phrase curve. Generally, American English does not have negative accents.

3.1.2 GENIE's additional assumptions

The word "General" in GSM denotes that its assumptions are general enough to define any superpositional model for any language regardless of whether it is used for synthesis or analysis. In practice, there might not be one model for solving every possible problem (based on the no free lunch theorem). Intonation can vary substantially across languages. Probably, there is not a single intonation model that can achieve both high predictiveness (used as a synthesis tool) and high descriptiveness (used as an analysis tool) for every language. Therefore, depending on the problem, it is important to assess the trade-offs. GENIE and GLAM define these trade-offs differently. The word "Generalized" in GLAM denotes that its assumptions are general enough to define any superpositional model across any language, but its assumptions were specified to make GLAM a practical synthesis tool. The word "Generalized" in GENIE denotes that its assumptions are intended to make it general enough to make GENIE a practical synthesis and analysis tool, but its assumptions are specific only for the English language. GENIE (like GLAM) uses stronger assumptions than GSM. Some of these assumptions are shared with GLAM, which we discussed in the previous section. Some of these assumptions are specific to GENIE, which we discuss in this section.

Assumption 3.1: This assumption refines Assumption 3 from GLAM. The shape of the phrase curve should be kept as simple as possible. GENIE uses a phrase curve consisting of two linear segments (not quasi-linear or log-linear segments like GLAM).

With regards to Assumption 3.1, a fundamental issue in superpositional-based approaches is that

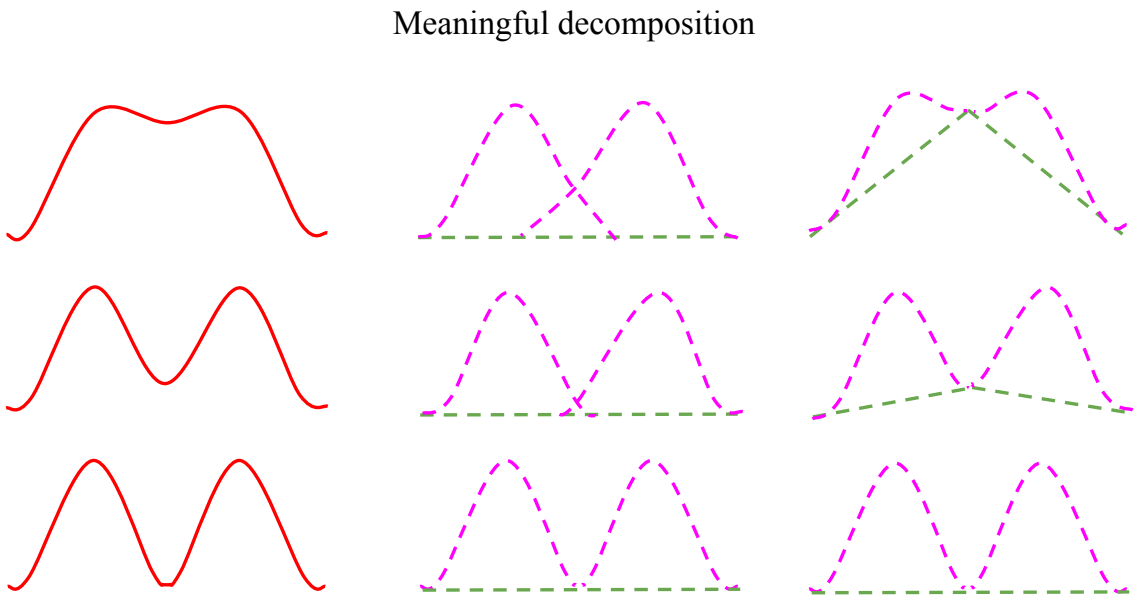


Figure 3.4: F_0 contour decomposition example, comparing when the phrase curve is a horizontal line versus when it has to capture the local minima. Each red curve represents a F_0 contour of a one-phrase utterance consisting of two feet, with different amounts of overlap. Green curves and magenta curves represent phrase curves and accent curves, respectively.

there is no unique way to decompose a curve into its components curves. One reason for this is that in real speech intonational movements are more complex than in theory. For example, we are interested in decomposing a given F_0 contour with two feet with rise-fall patterns into its component curves. One common sense solution is that the F_0 contour can be estimated by the concatenation of two identical rise-fall patterns with a coinciding start and end point. Figure 3.4 bottom row shows this approach; however, in real speech there can be an overlap between successive intonation movements (top and middle red solid curves).[3] Figure 3.4 shows two different ways of decomposition for three F_0 contours (red solid curves). Both ways are mathematically valid since the sum of the component curves in both ways results in a perfect fit to the F_0 contour; we prefer the first decomposition that by keeping the shape of the phrase curve as simple as possible lets the accent curves capture the meaningful intonation patterns. As the second decomposition, the phrase curve has to capture the local minima that prevents the accent curves from capturing meaningful intonation patterns.

[3]The degree of overlap depends on many factors, such as the number of syllables, duration of syllables, level of emphasis (or even focus), etc.

Assumption 6: All accent curves of an intermediate phrase span the full length of the intermediate phrase.

GLAM allows an overlap only between successive accent curves but it does not clarify what the extent of overlap allowed is. In some implementations based on GLAM, it was suggested that non-phrase-final accents should have no more than 20% of the foot duration overlapped with the next accent curve. Therefore, the amount of overlap is dependent on the foot duration and needs to be determined individually for each accent. Defining the amount of overlap as a free parameter results in an increase in the degrees of freedom of the model. In order to have overlap without any changes to the degrees of freedom, the overlap has to be tied to a segmental unit. In GLAM, the overlap can be tied to one of the following units, the syllable, foot, or intermediate phrase. We suggest to tie the overlap to the intermediate phrase boundaries. Allowing accent curves to span the full length of an intermediate phrase results in a bidirectional overlap between all accent curves of the intermediate phrase. The advantage of this amount of overlap is as follows. First, it allows a simpler mathematical formulation for analysis and synthesis. Second, the model is able to account for the pitch movement in appendices, which we discuss next.

Assumption 7: F_0 values in an appendix are predictable by the model.

The motivation behind Assumption 7 is that in ToBI it is uncertain whether L+H* and H* are distinct phonological categories in English intonation patterns. By showing that this uncertainty is not an issue in the proposed model, predictability of the F_0 values in appendices is necessary. With regards to the fifth assumption, an accent curve cannot be assigned to an appendix since it does not belong to any particular foot, but it is part of the intonational phrase. Therefore, recovering the pitch movement in an appendix would be possible by using information of both the phrase curve and the first accent curve which falls into the appendix segment.[4] Figure 3.5 shows how bidirectional overlap enables the model to predict F_0 values in an appendix.

Earlier in this chapter, we mentioned that our core goal for GENIE is that it be used as an intonational analysis and synthesis tool for the English language. Therefore, for making GENIE a practical tool, we consider two terms that are not assumption but are more like guiding principal.

- The number of parameters should be minimal, and each parameter should be meaningful.

The degrees of freedom of a model refers to the number of independent free parameters required to control the model for data estimation. When considering a model that fits to data, it is common

[4]Even though all accent curves are contributed due to sixth assumption, but just the first accent curve has any real influence.

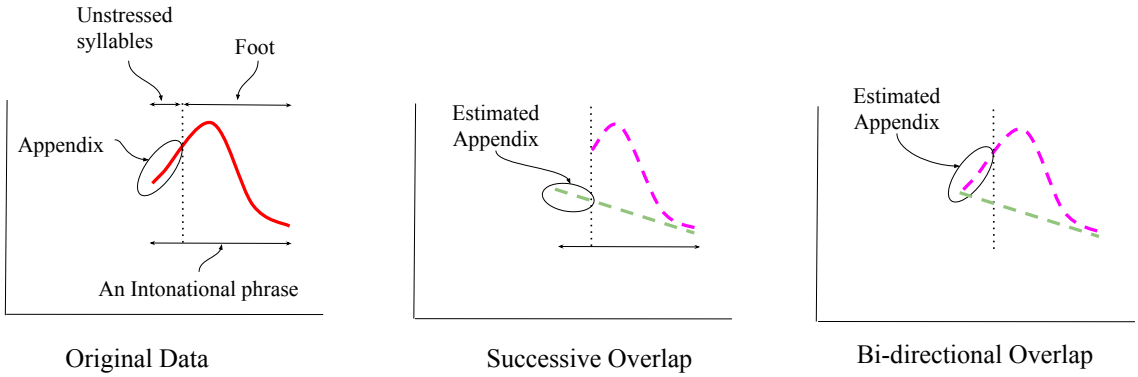


Figure 3.5: Letting an accent curve span the entire intonational phrase in both directions (bidirectional overlap) results in more accurate estimation of F_0 values in the appendix by GENIE.

practice to pay attention to three things. 1) Higher degrees of freedom implies better fit but decreases generalizability of the model. 2) If the size of the database is small, meaningful parameters will be more beneficial. 3) Parameters should be independent (highly correlated parameters contain redundant information). Building an efficient yet accurate and predictable parametric model is not easy. If GENIE satisfies these conditions, it would make it a practical tool for synthesizing English intonation.

- The model should be able to quantitatively capture all intonational patterns in English.

In order to make GENIE a practical intonational analysis tool two goals must be satisfied at the same time. First, GENIE's component curves should be linguistically descriptive. The general shape of GENIE's component curves are adapted from GLAM, and their validity has been tested in many research projects [164, 169, 166, 72, 136, 137]. Second, GENIE should decompose a given F_0 contour into linguistically meaningful component curves. As discussed above, not all decomposed component curves are meaningful, even if they add up to a very accurate estimation of the F_0 contour. If GENIE satisfies these goals, it should be able to quantitatively capture all intonational patterns in English, which makes it a useful tool for analyzing English intonation.

3.2 GENIE model methodology

In the previous section we summarized the underlying assumptions of GENIE that lead us to its creation. In Section 3.2.1 we discuss the mathematics behind GENIE's component accent and phrase curves. Then in Section 3.2.2, we introduce a way to implement GENIE.

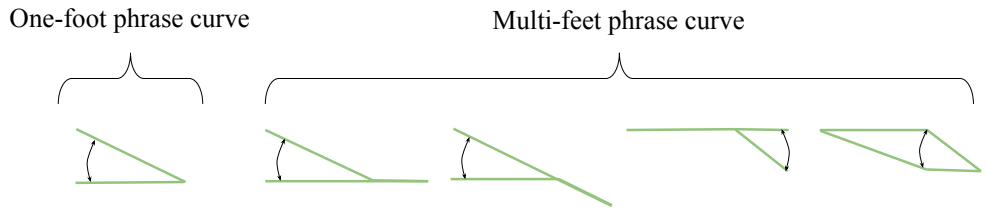


Figure 3.6: Each green line represents a phrase curve which indicates the general underlying F_0 contour for any type of utterance. Each black two-headed arrow shows how a specific parameter can change while other parameters kept unchanged.

3.2.1 Component curve classes

The proposed GENIE model is a superpositional-based intonation model that decomposes a continuous F_0 contour — interpolated in unvoiced regions — into two component curve classes: phrase curve and accent curve classes.

Phrase curve class: a phrase curve consists of two connected linear segments, the first from the phrase start (t_s) to the start of the final foot in the intermediate phrase (t_f), and the second from the latter to the end point of the last voiced segment of the intermediate phrase (t_e). The phrase curve does not account for how the F_0 changes to mark the end of the intermediate phrase. For these three time points, we associate three parameters p_s, p_f, and p_e to represent the phrase curve value. The phrase curve is constructed by linear interpolation between the three parameters (Equation 3.2). Please note that in an intermediate phrase with only one foot the phrase curve can be calculated by linear interpolation between the two points (p_s, p_e).

$$P(t) = interpolate(p_s, p_f, p_e) \tag{3.2}$$

This definition satisfies Assumptions 1-3, 1.1, and 3.1; however, in order to satisfy Assumption 7, some limitations on phrase curve class parameters are required. The phrase curve class represents the general underlying movement of the F_0 contour: 1) in statements, the phrase curve is a descending curve function. Therefore, $p_s \geq p_f \geq p_e$. 2) In yes-no questions, the final phrase curve is an ascending curve function. Therefore, $p_s \leq p_f \leq p_e$. Figure 3.6 shows all possible movements for the phrase curve parameters. Each plot represents how a specific parameter can change while other parameters are kept unchanged. For example, the top-left plot shows that p_s can have a value equal or higher than p_e. Therefore, the phrase curve is allowed to be either a horizontal curve or a pure descending curve.

Accent curve class: Accent curves are described by certain parametric curves. In order to satisfy

Assumptions 1.1, 2.1, 4, 5, and 7 and also motivated by the two guiding principles, we use a combination of the skewed normal distribution and a sigmoid function to model three different types of accent curves. First, the skewed normal distribution is employed to model rise-fall accents that occur in non-final positions as well as in final positions in statements ($f(t)$ in Equation 3.3). Second, a sigmoid function is used to model the rise at the end of a yes-no question ($g(t)$ in Equation 3.4). And, third, the sum of the skewed normal distribution and the sigmoid function is used to model continuation accents at the end of a non-utterance-final intermediate phrase ($h(t) = f(t) + g(t)$).

$$f(t) = C\frac{2}{\omega}\phi(\frac{t-\xi}{\omega})\Phi(\alpha(\frac{t-\xi}{\omega})) \tag{3.3}$$

$$g(t) = D\frac{1}{1+e^{-\beta(t-\gamma)}} \tag{3.4}$$

In Equations 3.3 C, ω, ξ, and α represent the amplitude, scale, location, and skewness of the rise-fall accent curve, respectively. In Equations 3.4 D, β, and γ indicate amplitude, slope, and location of the yes-no question accent curve, respectively. Figure 3.7 shows the effect of a change in one parameter on the shape of the rise-fall accent model ($f(t)$), while other parameters are kept unchanged. In comparison, in each plot the darkest curve represents the normal distribution (by setting $C = 1,\ \xi = 0,\ \omega = 1,$ and $\alpha = 0$ in $f(t)$).

Since this model is a superpositional-based model, the F_0 contour of a one-phrase utterance results from an overlay of component curve classes (Equation 3.1). The accent curve class ($A(t)$) is formulated below in Equation 3.5 where n is the total number of feet in the intermediate phrase:

$$A(t) = \sum_{i=1}^{n} A_i(t) = \sum_{i=1}^{n-1} f_i(t) + A_n(t) \tag{3.5}$$

$$A_n(t) = \begin{cases} f_n(t) & \text{statement} \\ g_n(t) & \text{yes-no question} \\ f_n(t) + g_n(t) & \text{non-utterance-final intermediate phrase} \end{cases} \tag{3.6}$$

Even though accent curve types are separated by their position i (in Equation 3.5) in an intermediate phrase and intermediate phrase type (e.g., statement vs. yes-no question in Equation 3.6), they are a function of t not a subsegment of t. This allows for bidirectional overlap between accent curves. Therefore, the parameters of a specific accent curve are proportioned to a specific foot but it spans across the entire intermediate phrase.

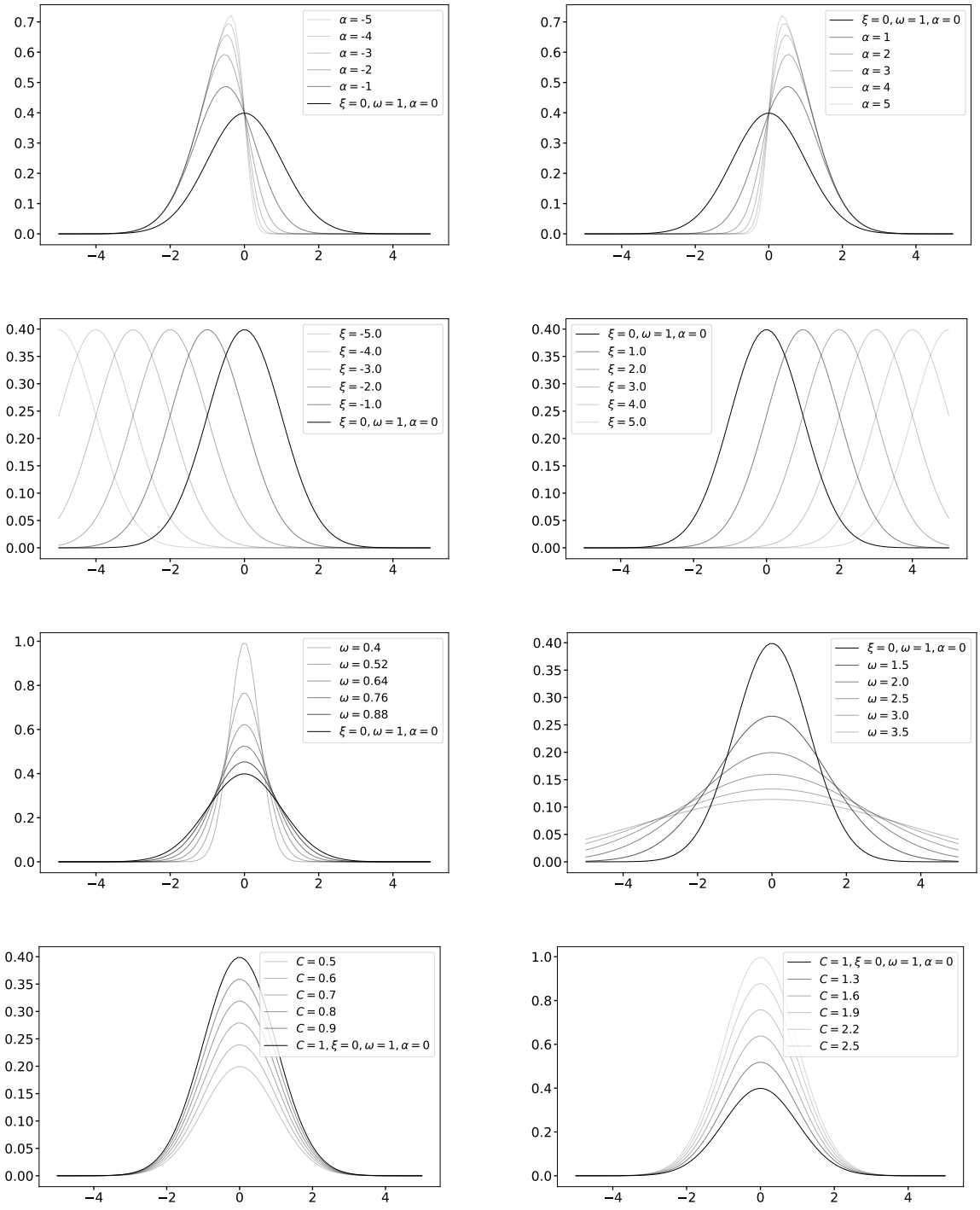


Figure 3.7: Each plot represents the effect of changing a specific parameter of a rise-fall accent curve while other parameters are kept unchanged. The darkest curve in each plot represents the normal distribution.

3.2.2 A Decomposition Implementation for GENIE

Decomposing an F_0 contour into its component curves is core to any superpositional approach including GENIE; however, there are many ways that this decomposition can be implemented. Below, we outline two ways that an implementation can be done.

In general, every decomposition method is actually a curve fitting problem, in which a mathematical function (Y) is constructed in such a way as to obtain the best fit for the data points (F_0).

$$F_0(t) = Y(t) \tag{3.7}$$

The simplest technique for solving this fitting problem is brute-force search (or exhaustive search). This technique considers all combination of candidates (parameters of Y) to check which combination results in an exact match. The brute-force search is very easy to implement, but it can be a time-consuming process; given the length of the utterance, the number of solutions is probably exponential.

In most speech processing applications, speed is more important than the simplicity of the implementation. This leads us to look for an approximate solution instead (Equation 3.8).

$$F_0(t) \approx Y(t) \tag{3.8}$$

One way to come up with an approximate solution is using an iterative approach, which consists of the following steps: initializing the parameters of Y, and updating them in each iteration until there is no significant improvement in a cost function. The root weighted mean square error (RWMSE) is one way to compare the deviations between the observed F_0 contour and the estimated one Y. In Equation 3.9, $F_0(i)$ represents the continuous F_0 contour with i frames, and w represents a weight vector. The weight w is computed as the multiplication of the voicing flag and the signal energy.

$$RWMSE(X,Y) = \sqrt{\frac{\sum w_i (F_0(i) - Y_i)^2}{\sum w_i}} \tag{3.9}$$

Below we discuss how we use above iterative approach to decompose an F_0 contour for GENIE. This decomposition requires the foot structure of the observed F_0 contour, which includes intermediate and intonational phrase boundaries, as well as the utterance type. By knowing the foot structure we can constrain the parameters' search boundaries.

A good initial guess for the parameters would speed up convergence to the optimal solution.

In order to initialize the phrase curve's three parameters, we use the actual F_0 values as an initial guess if the speech is voiced at two time points (t_s, and t_f). If the start of the phrase is unvoiced, the initial phrase start value is set to match the p_f. The p_e is set to match the minimum between p_f and the F_0 value at the time point t_e. These points are adjusted downwards if there are any F_0 values falling under the phrase curve. This prevents optimization from being stuck in the local minimum. Next, the initialized phrase curve (P_0) is subtracted from the F_0 contour to obtain the initial values for the accent curves (Equation 3.10).

$$Raw\,accent : R(t) = F_0(t) - P_0(t) \tag{3.10}$$

For example, to initialize a rise-fall accent, we compute the skewness (Equation 3.11), the mean (Equation 3.12), and the variance (Equation 3.13) of the raw accent values (R) in a foot as the initial values of the rise-fall accent parameters: α, ω, and ξ.

$$\text{skewness of } R(t) = \frac{4-\pi}{2}\frac{(\delta\sqrt{\frac{2}{\pi}})^2}{(1-\frac{2\delta^2}{\pi})^{3/2}} \qquad \text{where } \delta = \frac{\alpha}{\sqrt{1+\alpha^2}} \tag{3.11}$$

$$\text{mean of } R(t) = \xi + \omega\delta\sqrt{\frac{2}{\pi}} \tag{3.12}$$

$$\text{variance of } R(t) = \omega^2(1-\frac{2\delta^2}{\pi}) \tag{3.13}$$

We use the LMFIT python library (Non-linear least-square minimization and curve-fitting for Python) [111] to optimize GENIE's component curves parameters while minimizing the cost function (Equation 3.9). This library allows for the combination (adding or multiplying) of pre-built model classes with basic algebraic operations. A python implementation of GENIE is available.[5]

In addition to this implementation, we have designed a GUI (Graphical User Interface) toolkit,[6] that provides a framework for manipulating GENIE's parameters and visualizing the effects.

3.3 Experiments to show the efficacy of GENIE

In this section, we examine GENIE's potential to be used as both a synthesis and analysis tool for English intonation through several experiments. In the first part, we discuss how GENIE can reduce total number of intonational patterns defined by the ToBI system from 24 to three, and

[5] Add linke here ???
[6] Add linke here ???

in the second part, GENIE is used for objective testing to show it can produce accurate and linguistically meaningful results.

3.3.1 Linguistically meaningful

In this section, we want to show that GENIE is capable of capturing and predicting all intonation patterns present in ToBI by only three different accent **curves**.

For an intermediate phrase consisting of one accented-stressed-syllable (a one-foot intonational phrase), ToBI can describe 28 different intonational patterns (as described in Section 2.2.1). If our model can fit accent curves to each of the 28 intonational patterns, while keeping the phrase curve as a horizontal line equal to the minimum of the F_0 contour, then we can claim that the component curve classes are linguistically meaningful; in other words, GENIE can phonologically represent English intonation patterns. In the second chapter, we showed that using foot segmentation, the total number of ToBI intonational patterns can be reduced to 16. In Figure 3.8, we show how theoretically these 16 intonational patterns can be decomposed into their component curves using GENIE. Each plot under the "intonational pattern" column represents an individual intonational pattern used by ToBI under certain combinations of accent tone and phrasal tone in a one-foot intonational phrase. Each plot under the "component curves" column represents a decomposition of the individual intonational pattern using GENIE. As we can see, by setting the phrase curve (green line) to the minimum value of the intonational pattern, an accent curve can capture the meaningful F_0 dynamic pattern of the residual.

We previously discussed that the only concern about foot segmentation is that the appendix, a sequence of phrase-initial unstressed syllables, is ignored under this segmentation, while in the ToBI system the appendix is differentiated through a less prominent tone in a bitonal accent type (e.g., an L tone in a L+H* accent tone). In Section 3.1.2, we argued that an appendix does not show as much pitch movement in an accented-stressed-syllable, but that does not mean that the pitch movement in an appendix is unspecified. According to Assumptions 5, 6 and 7, GENIE predicts the pitch movement in the appendix without assigning a specific accent curve to it, and it does it by allowing accent curves to span the full length of an intermediate phrase (Figure 3.5). We show GENIE's ability through the same two examples as in Section 2.2.3

First, consider a one-word single-phrase utterance with a stressed-syllable at the beginning, e.g., "NO". In Figure 3.9, one speaker produces the word "NO" under five ToBI accent types in a continuation phrase (L-H%). As we discussed in Chapter 2, under foot segmentation only three accent types can occur (H*, L*, and L*+H) since there are no unstressed syllables before the

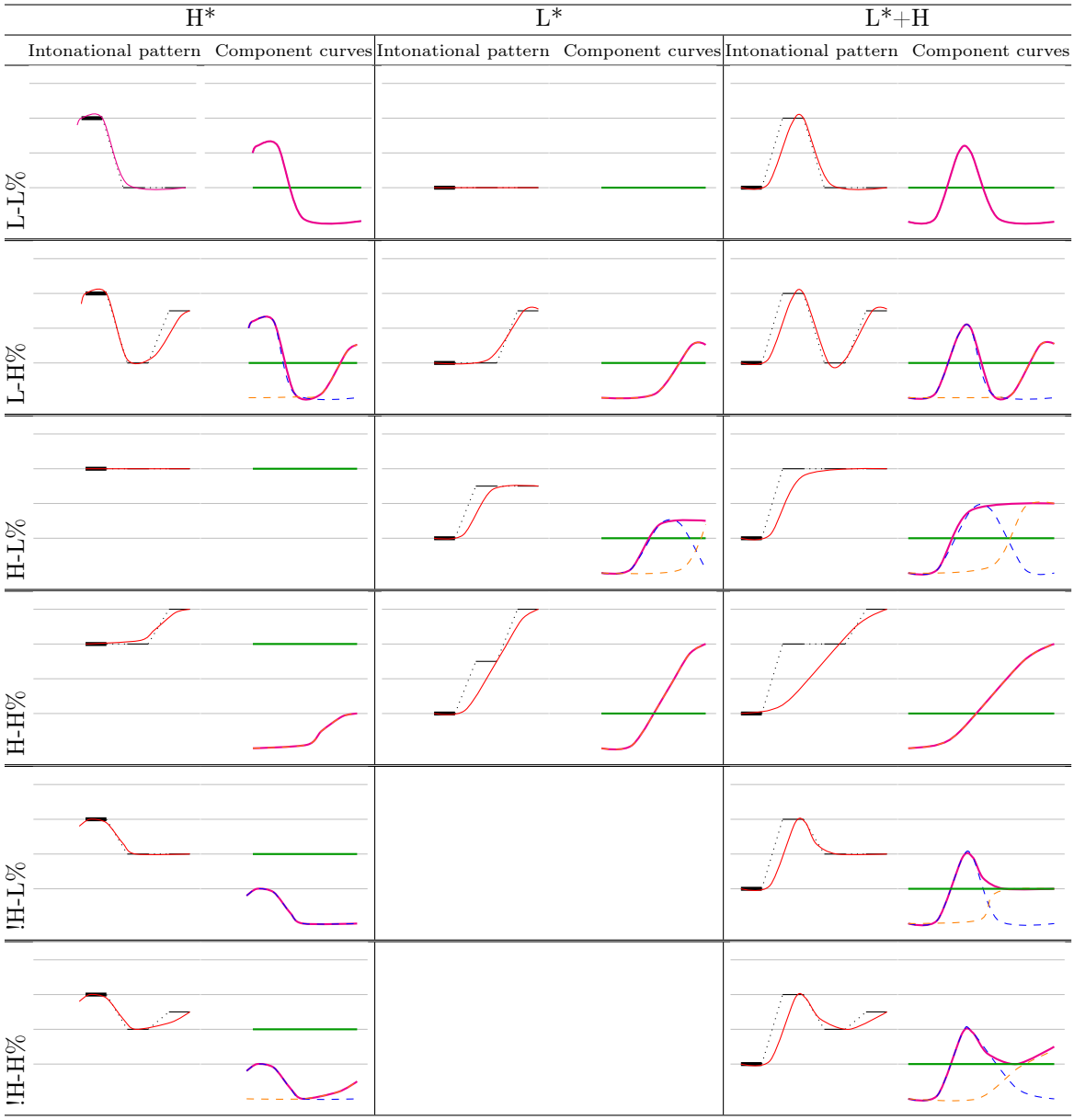


Figure 3.8: Decomposition of all intonation patterns used by the ToBI system under foot segmentation. In each intonational pattern, the theoretical pitch movement of a target tone is illustrated by a short black horizontal solid line. The starred target tone (pitch movement of stressed-syllable) is differentiated from other tones by a bold solid line. The red lines represent the theoretical smooth pitch contour. Next to each intonational pattern, there are the theoretical component curve classes of the proposed model: the green line represents the phrase curve and the magenta line represents the accent curve.

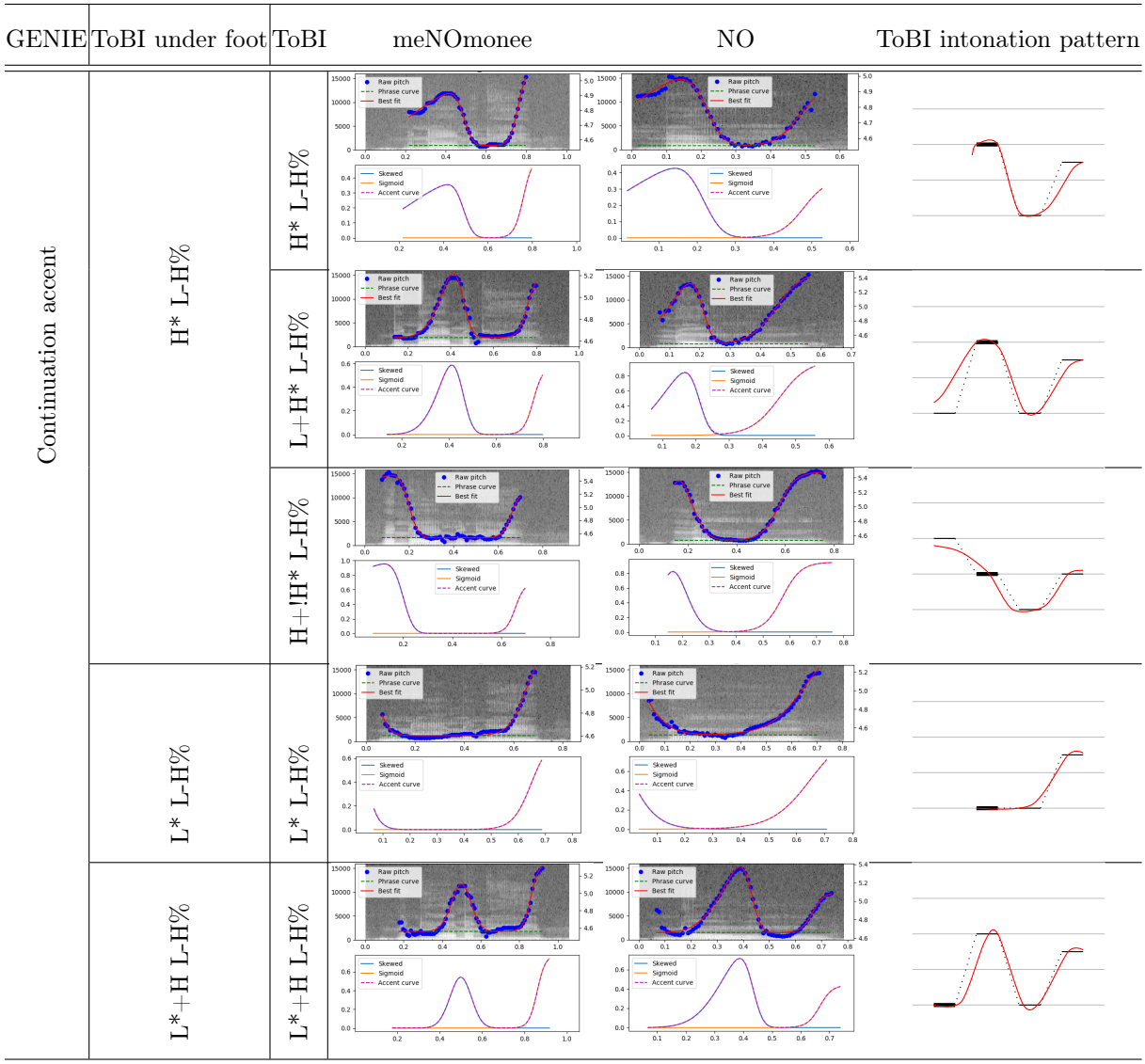


Figure 3.9: Decomposition of two words "meNOmonee" and "NO" for the five accent types in a continuation phrase(L-H%). The red lines represent the estimated pitch contour, green lines represent the estimated phrase curves, magenta lines represent the estimated accent curves. The raw pitch is represented by blue dots.

prominent syllable to carry the unstarred tone in two cases L+H* and H+!H*. Interestingly, all these five different ToBI accent types can be represented by only a continuation accent type due to the inherent flexibility of GENIE's accent curve formulation that is based on skewed normal distribution and sigmoid function.

Second, consider a one-word, single-phrase utterance with at least one unstressed syllable at the beginning, e.g., "meNOmonee." In this example, the foot starts at the stressed syllable "-NO-" and ends at the end of the intonational phrase, therefore again only three accent types can occur (H*, L*, and L*+H) under foot segmentation. In Figure 3.9, the same speaker produces the word "meNOmonee" under five different ToBI intonation patterns. As we can see, even though the pitch values under the appendix are not part of the foot, GENIE can accurately predict them and capture the F_0 dynamics of the intonation pattern due to its strong assumptions. The same logic can apply for the five other intonational phrase types (Figure 3.8).

In this section, first, we showed that GENIE is capable of capturing and predicting all intonation patterns present in the ToBI system. Second we showed that GENIE can represent all 28 different ToBI intonational patterns by only three different accent curves (in Equation 3.5) due two reasons: 1) flexibility of component curve to capture any F_0 dynamics 2) extension of accent curves to the full length of an intermediate phrase. This implies that considering 28 different intonational patterns for an intermediate phrase consisting of one accented-stressed-syllable is not necesseray, and they are all variation of three different accent curves. Further in Chapter 4 and Chapter 6, we use GENIE's ability of decomposing a F_0 contour into linguistically meaningful component curve as an analysis tool in variety of tasks.

3.3.2 Objective evaluation

We evaluated GENIE's potential in producing accurate and linguistically meaningful results. First, we start with the simplest scenario when the corpus contains only synthetically generated F_0 curves. Second, we consider a corpus of all-sonorant utterances. Finally, we consider a more challenging scenario when the corpus contains recordings of one child spoken in four different emotions. The Root Weighted Mean Squared Error (RWMSE) was extracted between the observed F_0 values and the estimated values. In experiments 2 and 3 we compared the performance of GENIE and PRISM (on the corpora for which comparable data are available).

3.3.2.1 Decomposing synthetic intonation contours

The first experiment with the implementation of GENIE was a proof-of-concept using synthetically generated F_0 contours. The contours were generated using a text-to-speech system that used GLAM to generate F_0 contours. We generated synthetic curves for 229 sentences present in the CSLU Emphasis Protocol [110]. This protocol was designed to elicit F_0 contours produced with various linguistic and prosodic features. It prescribes which syllables are accented, where each foot starts and ends, and where phrase boundaries occur. Finally, each utterance in the protocol has a target word that is spoken with a prescribed degree of emphasis. The protocol systematically varied the accent type (standard vs contrastive), the sentence type (declarative, wh-question, or yes-no question), the number of syllables in the foot (1, 2, 3 or more), and the phrasal position of the target word (initial, medial, or final). Here is an example with foot boundaries marked with brackets and the target word marked in all-caps: [Will we] [really know] [MARIO], [when we're in] [Maine?]. For each F_0 contour in the date, we apply the implementation of GENIE and then calculate RWMSE between the F_0 contour and the GENIE's estimated F_0 contour; it results in a very small overall RWMSE of 1.4307 Hz for whole data.

While humans can hear very fine distinctions between two pure tones when listening to them sequentially at a short time interval, in a longer sentence this type of error is not noticeable. Klatt notes that subjects could hear a 0.3 Hz difference in a constant F_0 contour, but when the synthetic F_0 contour is a linear descending ramp (32 Hz/sec) the just-noticeable difference slips to 2.0 Hz [73]. Comparing perceived intonation in two sentences, 't Hart [152] found that there is significant variability in the subjects' sensitivity to intonation differences. Some subjects are able to perceive differences of 1.5 - 2 semitones where others were only able to hear differences when the intonation was more than 4 semitones apart. They conclude that only differences of more than 3 semitones play a part in communicative situations. As we discussed in Section 2.6, semitones are measured on a perceptual scale and the actual frequency difference depends on the frequency range. Suppose the base frequency is 200 Hz, then a 2 semitone difference corresponds to a frequency differential of 24 Hz. But if the base frequency is really high, say 800 Hz, then the same 2 semitone differential corresponds to a frequency differential of 97 Hz.

The slight discrepancy between the generated accent curves and the decomposed curves is due to the fact that the accent curves generated by GLAM are asymmetric curves coupled together via cosine interpolation, whereas GENIE uses a smooth skewed normal distribution. Not only do we suspect that this discrepancy is inaudible, we also suggest that the skewed normal distribution can provide accurate approximations to a broader range of curves.

3.3.2.2 Decomposing all-sonorant speech

This experiment involves actual recordings using all-sonorant speech from the same CSLU Emphasis Protocol. One male speaker spoke a subset of 61 sentences in this protocol. The recordings are forced-aligned to the phonemes using the CSLU Toolkit [57]. We used the YAAPT algorithm [197] to extract F_0 values. We applied linear interpolation between voiced areas to replace the unvoiced areas. In this experiment, we compared the performance of GENIE with that of PRISM. The RWMSE for decomposition using PRISM was 5.40 Hz [106]. We use a similar methodology as in the previous section. The overall RWMSE for GENIE was 2.37 Hz. We applied a one-sample two-tailed t-test to determine whether this difference was significant. The results showed that GENIE performed significantly better than PRISM ($t(60) = 4.21$, $p < 0.05$).

3.3.2.3 Decomposing recordings with voiced and unvoiced speech sounds

In the previous experiments, F_0 values were available for all frames in the speech recordings, so that we could apply GENIE on continuous F_0 contours. A challenge for intonation decomposition of natural speech recordings is the presence of unvoiced regions and pauses where there are no F_0 values, and segmental perturbations. A common way to solve this issue is to use linear interpolation between voiced areas to fill in unvoiced areas. One side-effect of having unvoiced segments in speech is that an unvoiced phoneme preceding a voiced phoneme can cause a segmental perturbation at the start of the voiced phoneme, where the observed F_0 values are slightly higher than they should be [136]. Thus, linear interpolation will give suboptimal results. In order to test GENIE on a speech corpus with voiced and unvoiced segments and compare it directly with PRISM, we use the CSLU affect corpus [72]. This corpus was not specifically designed for synthesis purposes, but was created to study different prosodic and spectral variations using the same affect-neutral text for each sentence spoken in four different affects (Angry, Fearful, Happy, and Sad). One female child actor reading a total of 24 sentences in each affect (96 utterances total). The sentences are fairly short, consisting of a single phrase and 2-5 words in a phrase. The correct affect was prompted by vignettes that preceded each sentence. For this particular speaker, the F_0 ranges from 200-800 Hz.

Figure 3.10 represents the intonation decomposition of the sentence "She was taking a bath" into the component curves for the four affect types based on the proposed model versus PRISM. PRISM detects negative accent curves for two types of affects: Fearful, and Sad. The negative accent in the first foot of the Fearful sentence makes it a slightly better fit between the actual F_0 values and the decomposed values. However, there are doubts regarding the use of negative

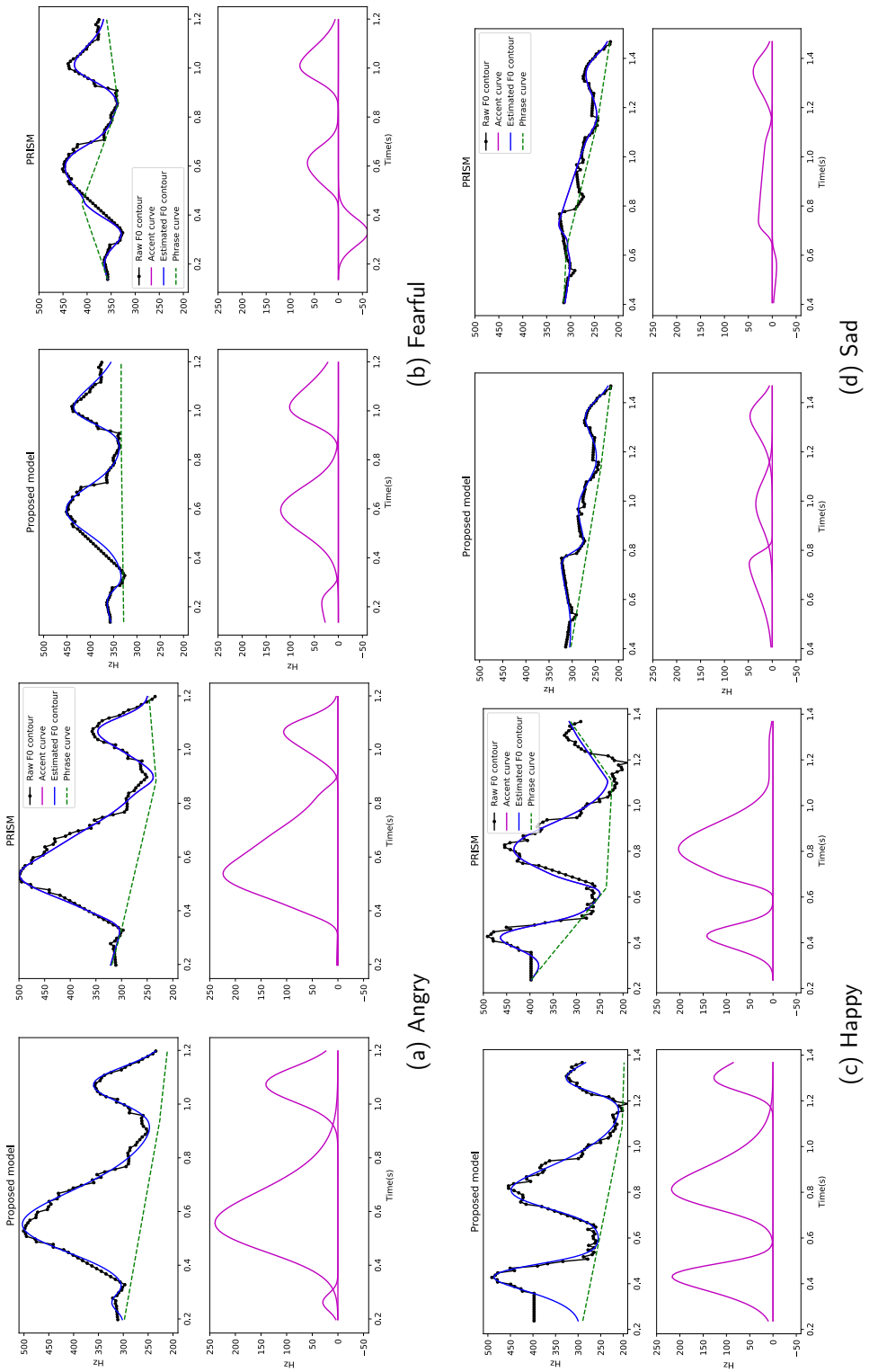


Figure 3.10: Decomposition of the sentence "She was taking a bath" for the four affect types (Angry, Fearful, Happy, and Sad) using both GENIE and PRISM. The blue lines represent the estimated F_0 contour, green lines represent the estimated phrase curves, magenta lines represent the estimated accent curves. The raw F_0 is represented by black dots.

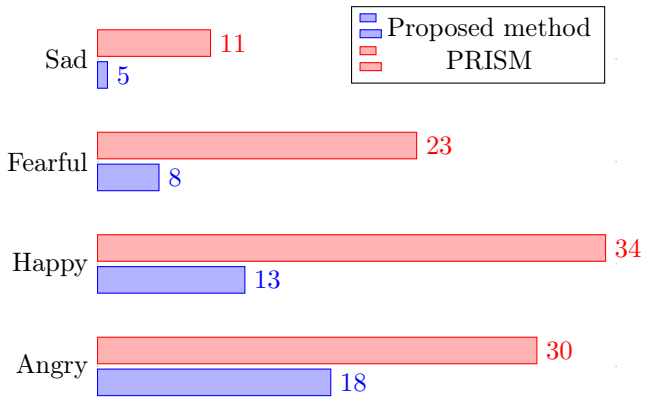


Figure 3.11: The RWMSE of GENIE vs. PRISM in Hz.

accents in American English.

The RWMSE for GENIE and PRISM are shown in Figure 3.11. GENIE performs better than PRISM for all of the affects. The average difference between the RWMSE of the two methods is 9.16 Hz. We applied a one-sample two-tailed t-test to determine whether this difference was significant. The results showed that GENIE performed significantly better ($t(95) = 2.22$, $p = 0.027$). The frequency range of the angry and fearful utterances cover the entire frequency range of the speaker (200 Hz-800 Hz) and there are a few points specifically around each accent peak that have more effect on the RWMSE. But since F_0 perception does not follow a linear scale, F_0 discrepancies at higher frequencies are likely to be less audible. Further in Chapter 5, we investigate GENIE's ability to decompose and generate high-quality F_0 contour through several perceptual studies.

Chapter 4

Intonation Annotation Using GENIE

In the previous chapter, we proposed GENIE, a foot-based superpositional analysis and synthesis intonation model for English. We showed that the implementation GENIE was able to decompose F_0 contours accurately for a few different data sets using a limited set of parameters. In this chapter we demonstrate the use of GENIE as an analysis tool to automatically detect the occurrence of phrase boundaries and show that it can do so reliably.[1]

4.1 Motivation

Humans use phrasing to chunk speech into semantic or syntactic units, not only as a natural by-product of how speech is "computed" by the brain or as a result of limitations of the speech production apparatus (e.g., running out of breath), but also as a device to make it easier for the listener to understand the message.

The acoustic-prosodic correlates of phrase boundaries involve both F_0 and temporal features. Phrase boundaries can be produced by, for example, final lowering of the F_0 at the end of statements, final rises at the end of yes-no questions, and continuation rises for non-utterance-final breaks. In the temporal domain, phrase boundaries can be produced by, for example, the presence of pauses or phrase-final lengthening.

As we discussed in Chapter 2, there are two levels of phrasing: the full intonational phrase level (intonational phrase), and the intermediate intonational phrase level (intermediate phrase). An intonational phrase is frequently followed by a pause and it is indicated by strong phrase-final F_0 changes and strong phrase-final lengthening. An intermediate phrase is not indicated by a pause. The phrasing cues after an intermediate phrase are weaker than phrasing cues after an intonational phrase.

[1] This chapter is based on work published in the 9th ISCA Speech Synthesis Workshop [30].

		Intonational phrase		Intermediate phrase
	ToBI tonal marking	L-L%, L-H%, H-L%, H-H%, !H-L%, and !H-H%		L-, H-, and !H-
	ToBI breaking index	4		3
	Total Number of Intonational phrase	1		-
	Total Number of Intermediate phrase	1 or more		1
	Total Number of feet	1 or more		1 or more
	Followed by pause	Yes, PB^+	No, PB^-	No, PB^-
	Phrasing cues	Very strong	Strong	Less strong
Location	End of statement utterance	Yes	Yes	No
	End of Yes/No question utterance	Yes	Yes	No
	End of non-utterance-final phrase	Yes	Yes	Yes

Table 4.1: Comparison between two levels of phrasing: intonational phrase and intermediate phrase. The term "phrasing cues" associates with phrase-final F_0 changes and phrase-final lengthening.

While phrase boundaries involving pauses (PB^+, intonational phrases) are relatively easy to automatically detect, pauseless phrase boundaries (PB^-, intonational phrases or intermediate phrases) are much harder to detect [132]. The two main reasons why PB^-s detection is a difficult task are: 1) F_0 contours may pass entirely smoothly through the phrase boundary; and 2) lengthening is difficult to assess because phoneme durations depend on many other factors besides the presence of a phrase boundary. For example, a 120 ms schwa (/@/, as in the word "the") is relatively long while a 120 ms /aI/ (as in "by") is relatively short [165].

In the ToBI annotation scheme, there are four break indices, where the two highest indices indicate phrasing. An intermediate phrase (break index 3) is associated with a monotone boundary tone (L-, H-, and !H-) while an intonational phrase (break index 4) is associated with a bitonal boundary tone (listed in Table 2.2). An intonational phrase consists of one or more intermediate phrases where each intermediate phrase consists of one or more feet. Table 4.1 summarizes the differences between intonational and intermediate phrases.

As mentioned previously, it is not easy to detect phrase boundaries when they are not followed by a pause (especially when dealing with intermediate phrases). Agreement among human labelers or between the human and automated labelers is not very high for this task. This is less the case for a phrase boundary involving a pause [132]. A common solution is to hire an expert to label the data, but inter-rater reliability among experts might be low as different experts may use different

acoustic cues to decide on the labeling. One solution is to hire more experts and use their mutual-agreement as ground truth; however, expert annotations are costly and time-consuming to collect.

It would be ideal if we had an automatic annotator that results in accurate correct prediction regardless of speaking style. As we showed in previous chapter, the implementation of GENIE results in highly accurate estimation of F_0 contour; assuming the given inputs (foot structure, phrase boundary, and raw F_0 contour) are accurate. This led us to hypothesize that GENIE has a potential to be used for this task in evaluating flexibility of GENIE in capturing meaningful and underlying intonation patterns.

In order to use GENIE to determine the best phrase boundaries for a sentence, we could generate all possible phrase boundaries for the sentence by considering occurrence/non-occurrence of PB^- after each word. We then could use GENIE to find which variation resulted in the lowest error with respect to the model; however, considering all possible phrase boundaries results in an exponential number of variations for a sentence.

Rather than solving this problem, in this chapter we are going to take a simpler approach. We limit the search space of variations by using a labeling method that over generates PB^- candidates for a sentence. Then, we generate number of phrase boundaries for the sentence by considering occurrence/non-occurrence each of those PB^-s. For a better comparison, we use three labeling methods to constrain the PB^- search space.

The aim of this chapter is to use GENIE as an analysis tool to improve the detection of pauseless phrase breaks by filtering out incorrectly placed pauseless phrase breaks by a labeler or an automatic labeling system. As such, we are proposing a hybrid method that constrains the PB search space, and filters out the false positives by using GENIE. We also investigate using a duration model to further improve the results.

We propose a framework that combines GENIE with a duration model to improve the phrase boundary assignment driven from a labeling methods. In Section 4.2, we use three labeling methods to constrain the PB search space. Each method results in a phrase boundary assignment for a given sentence, by determining which word in the sentence is followed by a phrase boundary. In Section 4.3, we use GENIE to determine which phrase boundary assignment provides the best fit of the model. In Section 4.4, we use a duration model that measures pre-boundary lengthening by predicting the duration of a vowel based on all factors known to affect vowel duration, but excluding boundary related factors. Then, we combine GENIE and the duration model to improve the detection of PB^-s. Finally, in Section 4.5, we introduce a method to derive a ground truth.

4.2 Constraining the Phrase Boundary Search Space

In order to avoid the very large space of possible boundary assignments for a given sentence, we limited the search space using three labeling methods: Expert, Festival, and a combination of both. For each method, in addition to phrase boundaries, pitch accents have also been determined, as they are used for creating foot structure which is needed in GENIE to model the surface F_0 movements.

Expert: Two linguistically informed experts manually indicated phrase boundaries.[2] They each separately used Praat [13] for annotating pitch accent labels and phrase boundary labels. They also had access to phonetic transcriptions and segmentation (e.g., phoneme, syllable, and word boundaries). Then we used they mutual agreement on pitch accent labels and phrase boundary labels as a final outcome of this method.

Festival: Festival was used to predict pitch accents and phrase boundaries. It predicts phrase boundaries at the word level based on an algorithm presented in [11]. It also predicts pitch accents at the syllable level. The pitch accent labels were moved to the word level, such that if one syllable of a word is accented then the whole word is accented. Only textual information was used for this prediction without any acoustic or prosodic information. We placed a period after any word which was followed by a pause, before feeding text to Festival, in order to have the same PB^+ as the original speech.

Combination of Festival and Expert (Comb): We combined phrase boundary and pitch accent labels from Expert and Festival by considering the union between the two. Our objective behind this method is that by giving a bit more possible PB, we might reduce the number of true negatives (present of PB that was missed by Festival or Expert method). Pitch accent labels in this method are obtained via the union of Expert pitch accent labels and Festival pitch accent labels. PB^-s are also obtained via the union of the Expert PB^- labels and Festival PB^- labels. These methods are different in terms of pitch accent labels and the location of PB^- labels but they all have the same PB^+ labels.

4.3 Using GENIE to Filter out False Positives

In this section, we describe how we use GENIE as an analysis tool to select a specific boundary assignment for each sentence. Before providing the method's details, we recall a fact about GENIE

[2] Author of this dissertation and her adviser were the two experts.

Algorithm 4.1 Usage of GENIE

Input

St ← *get Stress label of S from Dictionary*
PB ← *get Phrase Boundary labels from X*
Acc ← *get Accent labels from X*
phrase_boundary_assignments ← *All combinations of occurrence/non-occurrence PB*

Output

Phrase break prediction by X_{F_0}

1: *Index* ← *0*
2: **for** PBA in phrase_boundary_assignments **do**
3: *Feet* ← *Get foot structure(PBA, Acc, St)*
4: *Fitted_*F_0 ← *Fit the* F_0 *model(Feet, Raw* F_0*)*
5: *Error*[*Index*] ← *RWMSE(Fitted_*F_0*, Raw* F_0*)*
6: *Index* ← *Index + 1*
7: *Inx* ← *Index of lowest Error*
8: Report PBA[Inx]

from the third chapter; foot structure and phrase boundaries are GENIE's requirement. A foot starts with a stressed-accented-syllable and ends before the next stressed-accented-syllable or with a prosodic phrase boundary. Therefore, GENIE depends on syllable stress, pitch accent, and phrase boundary labels. Syllable stress labels are predetermined in English; however pitch accent and phrase boundary are variable and based on the speaker's style. Here, for a given sentence, S, the syllable stress labels are dictionary-based while phrase boundary labels and pitch accent labels come from each labeling method X ($X = Expert, Festival, or Comb$) as described in Section 4.2). Now by given the S and the labels from any labeling method in X, we use GENIE to filter out incorrect occurrence of PB^- and report the best phrase boundary assignment.

Algorithm 4.1 shows the required steps to detect PB^-s using GENIE, given a sentence S along with its prosodic labels from X. We illustrate the steps by an example. Consider the S, "I like cooking rice and kids.", which received two set of labels (phrase boundary labels and pitch accent labels) from the *Comb*.

- Input Phrase Boundary labels (PB) from *Comb*: I like cooking$^{PB^-}$ rice$^{PB^-}$ and kids$^{PB^+}$.
- Input Accent labels (Acc) from *Comb*: I like COOKing rice and KIDS.

Using the labels, we consider all combinations of occurrence/non-occurrence of PB^- labels. We call these combinations for a given sentence *phrase boundary assignments*.

1. I like cooking$^{PB^-}$ rice $^{PB^-}$ and kids$^{PB^+}$.
2. I like cooking rice$^{PB^-}$ and kids$^{PB^+}$.

3. I like cooking$^{PB^-}$ rice and kids$^{PB^+}$.

4. I like cooking rice and kids$^{PB^+}$.

Then for each assignment, we generate the foot structure with respect to the *Acc* labels and stress labels.

1. [I like] [cooking]$^{PB^-}$ rice $^{PB^-}$ and [kids]$^{PB^+}$.

2. [I like] [cooking rice]$^{PB^-}$ and [kids]$^{PB^+}$.

3. [I like] [cooking]$^{PB^-}$ rice and [kids]$^{PB^+}$.

4. [I like] [cooking rice and] [kids]$^{PB^+}$.

Foot structure in the first assignment is not valid since it consists of a prosodic phrase with no foot (rice $^{PB^-}$): we discard the first assignment. For the three remaining assignments (2, 3, and 4), we apply GENIE and then calculate a Root Weighted Mean Square Error (RWMSE) between the raw F_0 and GENIE's estimated F_0 contour. At the end we determine the best phrase boundary assignment using a goodness of fit measure (the lowest RWMSE). Let say the second assignment results in the lowest RWMSE, it means using GENIE we determine that the *Comb* incorrectly placed a PB^- after the word "cooking" for the sentence S.

- Input Phrase Boundary labels (PB): I like cooking$^{PB^-}$ rice$^{PB^-}$ and kids$^{PB^+}$.

- Output Phrase Boundary labels: I like cooking rice$^{PB^-}$ and kids$^{PB^+}$.

By applying GENIE to the Expert, Festival, and Comb assignments, we can in principle filter out the incorrect PB^-s. We call these methods: $Expert_{GENIE}$, $Festival_{GENIE}$, and $Comb_{GENIE}$.

4.4 Using a Duration Model to Filter out False Positives

As mentioned in Section 4.1, phrase-final lengthening is a well-established prosodic cue for phrase boundaries, with some of the earlier work reporting lengthening at many types of boundary (e.g., [74]), not just at the boundaries considered by ToBI. We use a simple model from literature that expressed vowel duration as a sum of product terms, with each component of a product depending on a specific factor (e.g., stress, post-vocalic consonant) [74]. Special cases of the sum-of-products model include the additive model (each product term has just one factor) and the multiplicative model (a single product term containing all factors). Using this model, it was shown that phrase-final lengthening is largely confined to phrase-final syllables, with much weaker lengthening for

earlier syllables [165]. We therefore confine our duration modeling to vowels in phrase-final syllables of each potential phrase boundary.

The duration of a vowel depends on many features in addition to the position in the phrase. The sums-of-products model was used to take into account these factors in order to evaluate the presence of lengthening. We fit the additive version of the model using the following features: the current phoneme whose duration is of interest, next phoneme, previous phoneme's stress label (binary), current syllable's stress label (binary), and current word's accent label (binary). The key is that we did not include position in the phrase as a feature in this prediction. Also note that we exclude both sentence-initial and sentence-final vowels, since this would confound the parameter estimates for the features included in the analysis.

By letting D^i_{Obs} be the observed duration of the i^{th} vowel in a sentence and D^i_{Pre} the predicted duration using the duration model, we define the ratio of the observed duration to the predicted duration of the vowel as $R_i = D^i_{Obs}/D^i_{Pre}$. Then, we extract a sequence of ratios, normalized per sentence (Equation 4.1).

$$Sig=\left\{\frac{R_i}{Median\{R_j|j \notin PB\}} \mid i \in Sentence's\ vowels\right\} \tag{4.1}$$

Thus, the sequence Sig is a vector that, by construction, provides hints about which vowels may be lengthened, and thus about possible phrase boundaries. After extracting the Sig vectors for all sentences for each of the six approaches (three labeling methods, and whether or not F_0 information was used), a logistic regression model [122] is trained to predict the phrase boundary assignments. In each case, we split the data into 10 partitions and applied 10-fold cross validation. When we present the results in Section 4.6, we will distinguish methods that use this duration modeling with the suffix "Dur". We note, however, that the estimation of the duration parameters and hence of D^i_{Pre} was not part of the cross-validation procedure. However, given the extremely small number of parameters compared to vowel tokens (30 compared to over 2,500), the risk of over-training was minimal.

4.5 Ground Truth

As we discussed at section 4.1 it is difficult to come up with a correct phrase boundary assignment for a corpus, which also makes it difficult to come up with the ground truth. For this chapter we define ground truth in the following manner. We use a group of native speakers of English and their majority vote as the ground truth. Our assumption – for preferring a group of speakers over

an individual – is that if a pauseless phrase break cannot be perceived by majority of these native speakers then it is not strong enough to be considered as a phrase break even though an expert might argue that there is some evidence in occurrence of a pauseless phrase break.

We used Amazon Mechanical Turk [15] with native speakers of English (master participants who have approval ratings of at least 95%). Their task was to determine the location of phrase boundaries in a sentence, regardless of phrase boundary type. At any given trial, the turkers were presented with the text displayed in normal, horizontal format, accompanied by a vertical list of the words, displayed in the same order, and each word followed by a button. They also listened to the audio of the text and had an option to replay. The task was to click on any words that the turker thought that should be followed by a comma or period.

The reference phrase boundary assignment for a given sentence is calculated by majority vote. A PB after a word in the sentence is included if more than half of turkers click on this word.

4.6 Experiments

In this section, we evaluate the potential of GENIE as an analysis tool to filter out false positives in order to improve PB^- detection. In Section 4.6.1, we introduce the corpora that we used. This corpora consists of two types of speech data, read speech and prosodically rich speech. In Section 4.6.2 we evaluate our assumption about generating the ground truth. We discuss in what degree turkers were reliable by measuring how much agreement exists in the labels given by various turkers. In Section 4.6.3, for each speaker, we extract phrase boundary assignments of each sentence via each labeling method X ($X = Expert$, $Festival$, $or\ Comb$). Then, we filter out false positives of these assignments using GENIE (X_{GENIE}), the duration model (X^{Dur}), and GENIE and the duration model (X^{Dur}_{GENIE}). In total, we compare 12 assignments with the ground truth for each sentence.

One way to evaluate these comparisons is by reporting the percentage of correct predictions. However, in this case the percentage of correct predictions is a biased measure since we are dealing with an unbalanced database. The unbalanced data is when the positive cases (i.e., appearance of a PB^- after a word) are much lower than the negative cases since roughly 90% of words are not followed by a phrase boundary. Analysis of the unbalanced data often results in a large number of false positives, that is, words wrongly identified as a word followed by a PB^-. Therefore, we use the F1 score (Equation 4.2) as a performance measure, since it gives equal importance to precision and recall. In addition, PB^+s are not considered in the results of this study since the location of all PB^+s are the same for all methods (i.e., all the phrase boundaries involving a pause are

correctly detected by all methods and turkers).

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2\,TP}{2\,TP + FP + FN} \tag{4.2}$$

4.6.1 Corpora

Prosodically Rich Database (PRD): we used a prosodically rich database. In this corpus 100 sentences were selected from the AP Newswire (years 1988-1990), automatically were annotated in terms of factors relevant for duration prediction [165] and greedy methods were used to select text with maximal coverage of the resulting feature space [135]. These sentences contained on average of 19 words. One female American English speaker, who is an experienced voice talent, was given carte blanche as to how to read these sentences as long as her utterances were affectively and prosodically meaningful, natural, and sounded exciting. All sentence internal punctuation were removed, and the speaker was instructed to insert phrase boundaries as judged appropriate; the speaker was not provided any instructions in terms of whether phrase boundaries should contain pauses or involve specific intonational cues. The recordings from the speaker were manually phonetically transcribed and time-aligned. We then followed the exact same procedure for a second speaker except the recordings from Speaker 2 were manually graphemically transcribed (i.e., slight deviations from the read text were corrected) but were segmented automatically using the HTK toolkit [193]; no manual corrections were made in the latter case.

CMU Arctic Speech Database: we also used the CMU Arctic speech database [75]. The database was automatically labeled via CMU Sphinx using the FestVox labeling scripts. We used speaker SLT, a US English female. To perform a fair comparison between CMU Arctic and Prosodically Rich Database (PRD), we wanted to create a collection of sentences most similar to PRD in terms of phoneme sequences. The CMU Arctic contains 1132 utterances from speaker SLT. For each sentence in the PRD, we find the 10 best sentences from the CMU corpus that are most similar in terms of their phoneme sequences using a standard string alignment algorithm (using the Bio.pairwise2.align function from the BioPython library [21]). Finally, from this collection of 100×10 sentences, which can include duplicates, we extract one from each 10 best sentences that was most frequent in all 100×10 sentences. The end-product is a set of 100 unique sentences.

4.6.2 Reliability of the Ground Truth

In Section 4.5 we introduced our process of generating the ground truth; in this section we want to measure the reliability of the ground truth. For any task in which multiple labelers are used, labelers might disagree about the observed target (i.e., appearance of a PB after a word). In order to reduce this issue, we took three precisive steps. First, as described in Section 4.5, we hired master turkers, who have approval ratings of at least 95%, from Amazon Mechanical Turk. Second, we randomly select three sentences to be annotated twice. A turker that did not have the exact same annotation for these three sentences was excluded. Third, we hired 15 unique turkers for each speaker (the two speakers in the PRD and SLT from CMU).

As described in Section 4.5, in each mode we use majority vote to determine the reference phrase boundary assignment. For measuring how reliable these references are, we need to assess the inter-labeler agreement. The inter-labeler agreement indicates the difficulty of the task.

There are a number of measures to estimate the inter-labeler agreement. We apply two of them to ensure reliability of the ground truth. The simplest measure of agreement is Total agreement, also known as Accuracy. Accuracy is the number of equally labeled words by different turkers, divided by the total number of words. In our case, accuracy has a bias towards TN (True Negative, a word that correctly not being labeled as a word followed by a phrase break). The value of TN tends to be high since most words are not followed by a phrase break. The second measure is Occurrence agreement which is not affected by TN.

$$Occurrence\,agreement = \frac{TP}{TP + FP + FN} \times 100 \tag{4.3}$$

$$Total\,agreement = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \tag{4.4}$$

In order to assess the inter-labeler agreement, for each sentence, we split the group of 15 turkers into all possible combinations of two groups with seven and eight members in each (Algorithm 4.2 steps 2-3). We computed the respective unions of the boundary assignments for each group (Algorithm 4.2, steps 5-7), and then computed the group-wise agreement (Algorithm 4.2, steps 8-9) for these unions measured using the Occurrence agreement (Equation 4.3) and Total agreement (Equation 4.4). When the turkers are in perfect agreement, the percentage of group-wise agreement is equal to 100%.

Results of group-wise agreement are presented in Table 4.2. The results show a high agreement level on average in all five modes. The inter-labeler agreement is higher for the CMU database

Algorithm 4.2 Intergroup raw agreement

for S in Sentences **do**
  $L \leftarrow \{l_1, l_2, ..., l_{15}\}$
  $A \leftarrow$ *all 7-combinations of the set L,* $\binom{L}{7}$
  for *subset* in A **do**
    $subset^c \leftarrow L$ - *subset*
    $C(subset) \leftarrow \bigcup_{i=1}^{subset} l_i(S)$
    $C(subset^c) \leftarrow \bigcup_{i=1}^{subset^c} l_i(S)$
    $O \leftarrow Occurrence\ agreement(C(subset), C(subset^c))$
    $T \leftarrow Total\ agreement(C(subset), C(subset^c))$
report average of O and T

		Total	Occurrence
PRD	Speaker 1	96.29	80.92
	Speaker 2	89.40	79.22
CMU	SLT	95.13	87.55

Table 4.2: Percentages of group-wise agreement

than the PRD database with respect to the occurrence agreement measure. One reason is that the CMU database was used to create a TTS database and the speaker was instructed to pronounce the sentences in a news reading style.

4.6.3 Results

In this section we give the results of using GENIE and the duration model to improve the phrase boundary assignments driven from the three labeling methods. For each speaker, we extract phrase boundary assignments of each sentence via each labeling method X ($X = Expert, Festival, orComb$). Then, we filter out false positives of these assignments using GENIE (X_{GENIE}), the duration model (X^{Dur}), and GENIE and the duration model (X^{Dur}_{GENIE}). In total, we compare 12 assignments with the ground truth for each sentence. We compare 12 phrase boundary assignments with the ground truth.

The median F1 scores of the three labeling methods, without use of GENIE or the duration model, are summarized in top three rows in Table 4.3. In the CMU Arctic database we only report $Festival$ results, since the labeling results from $Expert$ were identical to $Festival$. Based on the results for the CMU Arctic database, we conclude that both $Festival$ and $Expert$ are highly accurate in PB^- detection due to high F1 measure (which also implies on reliability of the ground truth). As for the PRD database, $Expert$ performs better than $Festival$. This is undoubtedly due to the experts having access to all the acoustic/prosodic/textual information. There is no surprise

	PRD		CMU
	Speaker 1	Speaker 2	SLT
$Expert$	0.68	0.64	–
$Comb$	0.40	0.50	–
$Festival$	0	0.34	0.98
$Expert_{GENIE}$	0.95	0.80	
$Comb_{GENIE}$	0.50	0.42	–
$Festival_{GENIE}$	0.21	0.39	0.88
$Expert^{Dur}$	0.68	0.50	–
$Comb^{Dur}$	0.86	0.50	–
$Festival^{Dur}$	0.67	0.45	0.5
$Expert^{Dur}_{GENIE}$	0.90	0.64	–
$Comb^{Dur}_{GENIE}$	0.90	1	–
$Festival^{Dur}_{GENIE}$	0.90	1	1

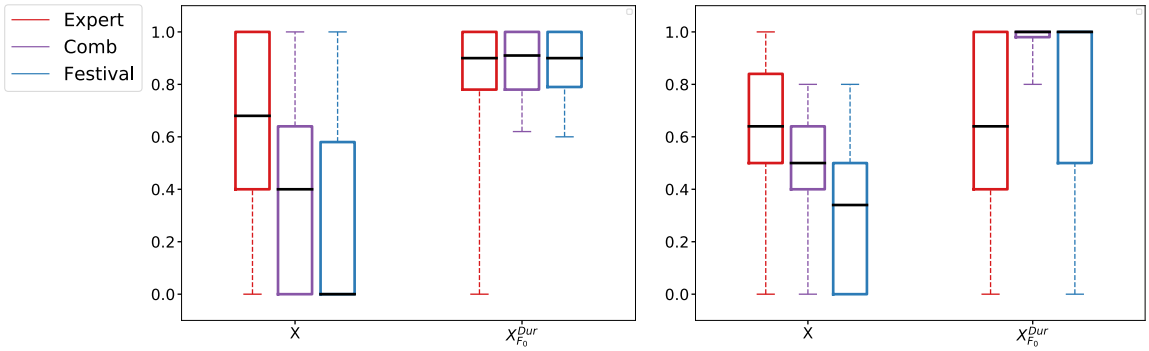
Table 4.3: This table summarizes median F1 scores for all 12 methods in comparison with text+speech ground truth for the three speakers.

that $Comb$ performed worse than $Expert$ and better than $Festival$.

As we discussed in Section 4.3, we used GENIE to select a subset of the PB^- to get the best fit of the F_0 contour. In Table 4.3, we can see an improvement on the F1 scores ($Expert_{GENIE} > Expert$, $Festival_{GENIE} > Festival$), when the speakers were instructed to pronounce the sentences in an exciting-sounding voice (Speaker 1 and Speaker 2), but it did not improve the performance of the $Comb$ method ($Comb_{GENIE} \backsimeq Comb$). A reason for that is the implementation of GENIE that we used is an optimization-based method. In the $Comb_{GENIE}$ method, the number of optimization parameters increased by combining PB^- labeling of two methods ($Festival$ and $Expert$) which caused the model to be overfitted to the F_0 contour.

As we discussed in Section 4.4, we used the duration model to select a subset of the PB^- driven from the three labeling methods. The $Expert^{Dur}$ performed worse than the $Expert$ condition. In the $Festival^{Dur}$ case, we see a significant improvement for the PRD (Speaker 1 and Speaker 2); however, this improvement could not be found in the CMU Arctic database. A reason for that might be the complexity of the PRD sentences compared to the CMU Arctic database.

While using GENIE and the duration model individually produced minor improvements, their



(a) Speaker 1 (b) Speaker 2

Figure 4.1: This figure summarizes the F1 score of each labeling method X ($X = Expert$, $Festival$, *or* $Comb$), and their combination with F_0 and duration information (X^{Dur}_{GENIE}) for the two speakers. Three different colors red, purple, and blue are used to represent results of the Expert, Comb and Festival methods, respectively. Medians are represented by a solid horizontal black line.

combination resulted in major improvements especially in the $Comb$ and $Festival$ cases, In Table 4.3, compare the numbers in the following pairs: $(Expert, Expert^{Dur}_{GENIE})$, $(Comb, Comb^{Dur}_{GENIE})$, and $(Festival, Festival^{Dur}_{GENIE})$. The experts (as well as $Festival$) are performing at 0.98 for the CMU Arctic database, and this performance reaches perfection when GENIE and the duration model were applied. This almost equality of performance with and without GENIE and the duration model implies that the phrase boundaries of the CMU speaker matched the grammatical phrase boundaries.

Side-by-side box-plots in Figure 4.1 show the distribution of the F1 score of each labeling method X, and their combination with GENIE and the duration model (X^{Dur}_{GENIE}) for the two speakers in the PRD database.[3] We mentioned earlier that the higher number of PB^- candidates in the $Comb$ method was the reason that $Comb$ performed worse than $Expert$ (in Figure 4.1a, and 4.1b, compare most left red bot-plot ($Expert$) with most left purple bot-plot ($Comb$)). However, in $Comb^{Dur}_{GENIE}$, GENIE and the duration model appeared to filter out incorrect PB^- assignments, resulting in better performance by decreasing the False Positives (in Figure 4.1a, and 4.1b, compare left purple bot-plot ($Comb$) with right purple bot-plot ($Comb^{Dur}_{GENIE}$)). There is no surprise in Festival's performance in PRD since only textual information was used for the $Festival$ based methods (in Figure 4.1a, and 4.1b, compare most left red bot-plot ($Expert$) with most left blue bot-plot ($Festival$)). However, results for $Festival^{Dur}_{GENIE}$ were as good as the results for $Comb^{Dur}_{GENIE}$

[3]We computed the F1 score for each sentence, and these boxplots are the distribution of these F1 scores.

	PRD		CMU
	Speaker 1	Speaker 2	SLT
$(Expert, Expert_{F_0}^{Dur})$	**0.007**	0.733	–
$(Comb, Comb_{F_0}^{Dur})$	**0**	**0**	–
$(Festival, Festival_{F_0}^{Dur})$	**0**	**0**	0.154

Table 4.4: P-value of Exact Wilcoxon Test between $(X, X_{F_0}^{Dur})$

which makes the impact of using GENIE and the duration model for filtering out incorrect PB^- assignments generated from $Festival$ labels more interesting. This suggests that the proposed model has the potential to detect PB^- for a prosodically rich dataset (such as emotional speech) using only textual information.

We employed the Exact Wilcoxon Test [55] to assess whether the following pairs -– ($Expert$, $Expert_{GENIE}^{Dur}$), $(Comb, Comb_{GENIE}^{Dur})$, and $(Festival, Festival_{GENIE}^{Dur})$ — were from significantly different distributions (the P-values are shown in Table 4.4). The reason we chose this statistic over the standard t-test is that we did not meet the normality assumption for some cases. All pairs are significantly different in term of F1 score distribution except the following two pairs: $(Expert, Expert_{GENIE}^{Dur})$ in Speaker 2 and $(Festival, Festival_{GENIE}^{Dur})$ in CMU.

4.7 Conclusion

In this chapter, we discussed how prosodic information can be used for improving the detection of pauseless phrase breaks. Pauseless phrase breaks associate with two surface phenomena: phrase-final F_0 changes and phrase-final lengthening. We used GENIE as an analysis tool to automatically capture the phrase-final F_0 changes and a duration model to capture the phrase-final lengthening. We showed that using these models individually produced minor improvements, while combining them results in a higher agreement between the labeling method and the ground truth. In prosodically rich speech, we improved the F1 measure by 0.68, 0.47, 0.10 in a paired comparison for $(Festival, Festival_{GENIE}^{Dur})$, $(Comb, Comb_{GENIE}^{Dur})$ and $(Expert, Expert_{GENIE}^{Dur})$, respectively. An interesting finding was that $Festival_{GENIE}^{Dur}$ was as good as $Comb_{GENIE}^{Dur}$. This suggests that with only textual information and using GENIE and the duration model we are able to filter out incorrect pauseless phrase breaks for prosodically rich datasets (such as emotional speech, and spontaneous speech).

The above approach, using GENIE and the duration model, has three advantages. First, it uses very few parameters, making the method usable in cases where few samples are available. This is

in particular the case when collecting speech data from special populations, such as dialect groups or individuals with speech or language challenges. Second, it makes use of global as well as local information available in an utterance. Third, it may allow us to "connect" this line of research with linguistics research, because the models are grounded in such research. We will further use the goodness of fit of the implementation of GENIE in the six chapter to differentiate one speaker group from another.

We limited the search space by not considering the very large space of possible boundary assignments for a given sentence. We showed using GENIE and the duration model resulted in major improvements especially in the *Festival*, which implies that considering such large spaces may not be needed after all.

Chapter 5

Intonation Generation and Adaptation in TTS

In the third chapter, we introduced GENIE as an analysis and synthesis tool for English intonation. In this chapter, we mainly focus on synthesis and discuss how GENIE can be used as an intonation generator model for an English Text-To-Speech (TTS) synthesis system.[1] In Section 5.2, we propose two methods for generating intonation for English based on GENIE. The first method is a data-driven foot-based intonation generator ("DRIFT"). The second method is a foot-based neural network intonation generator ("FONN") that maps foot-based features to GENIE's accent parameters using a simple Artificial Neural Network (ANN). We then turn to intonation adaption in Section 5.3. We use GENIE as an analysis tool to extract underlying prosodic characteristic of source and target speaker, then during test we use GENIE as an synthesis tool to generate target-specific F_0 contours. Finally, in Section 5.4, we give a summary of the main fundings of this chapter.

5.1 Motivation

Research into the analysis and modeling of speech prosody has increased dramatically in recent decades, and speech prosody has emerged as a crucial concern for Text-To-Speech (TTS) synthesis. Every TTS synthesis system needs to model prosodic phenomena to provide both natural and expressive speech. Hence, we want to investigate are the GENIE-based methods capable of generating more natural-sounding speech compare to baseline; if yes, can we go further and show the DRIFT method has potential to be used to generate expressing convincing speech.

[1] This chapter is based on work published in 3 papers [34, 29, 31].

The main challenge in generating natural sounding speech is capturing the suprasegmental properties in F_0 movements. For example, in English, standard L+H*L-L% rising peak accents involve a smooth rise during the course of the accented syllable followed by a descent until the next accented syllable or phrase boundary [169, 85, 78, 163]. A study by Anumanchipalli explicitly addressed this issue [6] by considering various phonological units in a statistical parametric speech synthesis framework, including the frame, syllable, word, accent group, phrase, and sentence. "Accent group" was defined as a sequence of syllables containing an accented syllable and not necessarily as a foot, which requires that the first syllable is accented. Anumanchipalli showed that the best-performing phonological unit in his study was the accent group. However, most HMM-based synthesizers predict F_0 at the frame level using limited linguistic contextual information. This frame-by-frame prediction of F_0 results in an overly-smooth F_0 contour that cannot properly represent the suprasegmental properties of F_0 movements. This motivated us to hypothesize that GENIE has a potential to generate more natural sounding F_0 contour than frame-based methods. We examine this hypothesis in Section 5.2.

One challenge in generating expressive speech is how well an F_0 generation method performs when input text is marked up to create intonation patterns that are not present in the training data. For example, suppose that one instructs the system, via markup, to convey strong contrastive stress, can the system create compelling-sounding contrastive stress when the training data do not contain any instances of contrastive stress? We address this issue in Section 5.2.4.4.

Going further, we also interested to see how GENIE can be used to transfer the perceived intonational identity of a TTS voice to that of a target speaker? To clarify, in the case of TTS, the source speaker is the speaker whose recordings were used to generate the acoustic units (for unit selection approaches), acoustic inventory (for diphone based synthesis), or acoustic features for HMM or DNN approaches. This speaker's recordings may also be used as training data for prosody mimic. Thus, the speech generated by a TTS system generally sounds like the source speaker. For prosody mimic, the challenge is to compute a transformation that, when applied to the speech data or to any representations thereof, generates output speech mimicking a target speaker.

5.2 Proposing a F_0 Generation Method for TTS Systems

There are different ways that we can use GENIE in F_0 generation. So to be fair, we are exploring two different ways. After discussing details of the baseline in Section 5.2.1, we propose two foot-based intonational approaches for F_0 generation based on GENIE: DRIFT and FONN

in Section 5.2.2 and Section 5.2.3, respectively. Then in Section 5.2.4, we compare F_0 contours generated by FONN with the baseline and with DRIFT in a subjective listening experiment with stimuli created by imposing contours generated by the three methods onto natural speech. In this test, we also explore the role of sparsity, by comparing test items whose constituent phoneme sequences, stress patterns, and phrasal structures are well vs. poorly covered by the training data. This exploration is based on the assumption that FONN and DRIFT are less sensitive to sparsity than HTS. Since DRIFT uses templates associated with individual curves in the training data, while FONN computes curves based on multiple observed curves in the training data, we expect DRIFT to have a relative advantage over FONN in well-covered test data sets because such data would provide ample stored templates that closely match the test context in terms of the selection features, but we expect FONN to have a relative advantage over DRIFT in poorly-covered test data. In a second experiment, we determined the ability of DRIFT to convey contrastive stress. This served to demonstrate the ability of DRIFT to generate F_0 contours from marked-up input text.

5.2.1 Baseline: Model-driven frame-based intonation generator

Hidden Markov Model (HMM) is a statistical parametric speech synthesis that takes the linguistic representation of a given text as input and outputs the acoustic features. We use a HMM-based baseline that is a model-driven frame-based intonation generator for comparison purposes.

5.2.1.1 Intonation model

The multi-space probability distribution (MSD) HMM [102] is a special case of using HMMs to model observed F_0 values. MSD-HMM includes discrete and continuous mixture HMMs to model F_0. The state output probability is defined by an MSD, which is a joint distribution of discrete F_0 values and voicing labels [196].

5.2.1.2 Training

We used the HTS toolkit (version 2.2) [199] to perform HMM-based TTS synthesis.[2] HTS uses the Festival speech synthesis architecture to extract a sequence of contextual and phonological features at several levels, such as, for a given utterance, the phrase, word, syllable, phoneme, and frame levels. As a result, there are many combinations of contextual features to consider when obtaining

[2]At the time this research was performed, HTS was the dominant method for statistical parametric speech synthesis.

models. HTS employs decision-tree (DT) based context clustering for handling a large number of feature combinations. The left panel in Figure 5.1 shows independent DT-based context clustering solutions for F_0 and duration, respectively.

5.2.1.3 Synthesis

Synthesis consisted of these steps: A to-be-synthesized sentence was converted into a contextual label sequence; the utterance HMM was constructed by concatenating the context-dependent state HMMs given the label sequence; state durations of the utterance HMM were determined [191]; a sequence of F_0 values (one value per frame), including a voiced/unvoiced label, was generated given the utterance HMM and the state durations.

5.2.2 Data-driven foot-based intonation generator (DRIFT)

In this section, we discuss how DRIFT generates a F_0 contour given a text and its duration information as an input. First in Section 5.2.2.1, we briefly review GENIE, which DRIFT is based on. Then in Section 5.2.2.2, we describe how we train DRIFT. We build a inventory of parameter vectors characterizing the individual shapes of GENIE's component curves; these parameter vectors are labeled in terms of basic linguistic features. Finally in Section 5.2.2.3, we describe DRIFT synthesis a F_0 contour. For a input text, we generate component curves by retrieving parameter vectors whose linguistic labels match those of the test and use these vectors to generate F_0 curves with the same duration as those in the test.

5.2.2.1 Intonation model

GENIE was used to decompose a F_0 contour. In GENIE, the phrase curve consists of two connected linear segments, between the phrase start and the start of the final foot, and between the latter and the end point of phrase, respectively. As we discussed in Section 3.2, GENIE uses a combination of the skewed normal distribution and the sigmoid function to model three different types of accent curves. GENIE allows for simple joint optimization of phrase and accent curve parameters using fewer parameters.

5.2.2.2 Training

For each utterance in the training data (train and test set selection is explained in Sections 5.2.4.1 and 5.2.4.2), we do the following. First, we run Festival to generate accent labels, syllable labels, and phrase boundaries. Second, we derive the foot structure. Third, we apply GENIE to compute

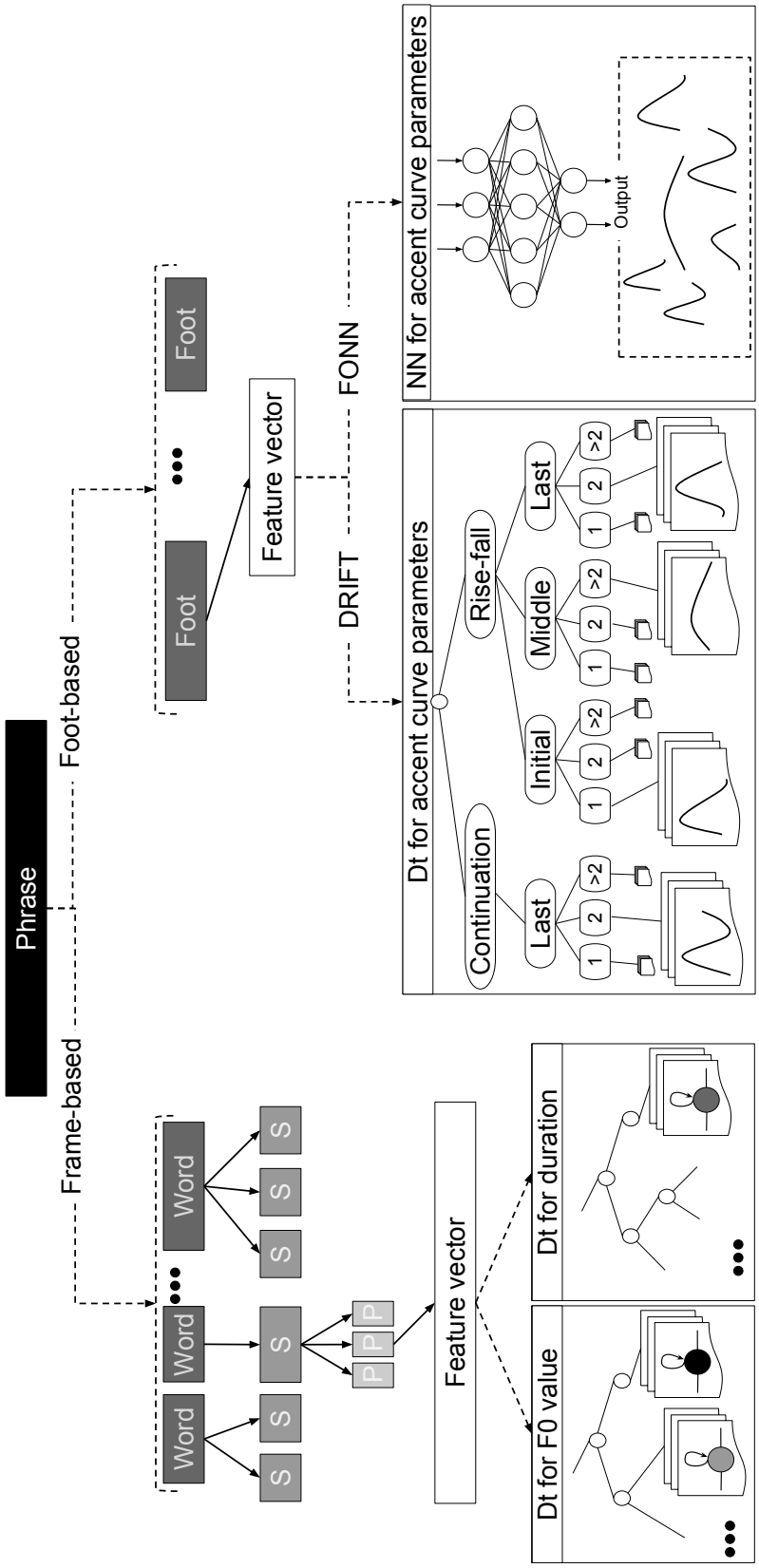


Figure 5.1: Overview of foot-based and frame-based schemes

the component accent and phrase curves. Fourth, the RWMSE between the accent curve and the raw accent contour – defined as the raw F_0 contour minus the phrase curve – is extracted for each foot. We exclude any curve that does not meet a certain threshold on fitness error. Fifth, We create two inventories, one for the accent curves and one the phrase curves, and each uses a different set of features: F_{Acc} and F_{Phr}, respectively.

The accent curve inventory is created as follows. In contrast with HTS, which uses a large number of features per frame, we only extract five features per foot: phrase type, foot position in phrase, number of syllables in foot, onset duration of stressed-accented syllable, and rhyme duration of stressed-accented syllable. We use the first three features for categorizing the accent curves, and we will use the last two features later on in synthesis part for retrieving closest accent curves. We store the vector comprising the estimated accent curve parameters and the values of OD and RD in the inventory. The inventory contains twelve sub-inventories defined in terms of the F_{Acc} features AT, Pos, and $SNum$ (middle panel of Figure 5.1). Because the data were not tagged for yes-no (or any) questions, a yes-no question sub-inventory is not included.

$$F_{Acc} = \begin{cases} AT : accent\,type\,(rise\text{-}fall, continuation) \\ Pos : foot\,position\,in\,phrase\,(initial, middle, final) \\ SNum : number\,of\,syllables\,in\,foot\,(1, 2, > 2) \\ OD : onset\,duration\,of\,stressed\,accented\,syllable \\ RD : rhyme\,duration\,of\,stressed\,accented\,syllable \end{cases}$$

The phrase curve inventory is created as follows. Two contextual features are extracted per phrase: phrase type and number of foot in phrase. We store the vector consisting of the phrase curve parameters (phrase start, the start of the final foot in phrase, and phrase end) is stored in the inventory. Note that if a phrase contains just one foot, then the phrase is modeled by two parameters (phrase start and phrase end). The inventory contains four sub-inventories, differentiated in terms of the F_{Phr} features, PT and $FNum$.

$$F_{Phr} = \begin{cases} PT : phrase\,type\,(statement, continunation) \\ FNum : number\,of\,feet\,in\,phrase\,(1, > 1) \end{cases}$$

In order to determine wether the 12 sub-inventories differ from each other, we performed a classification experiment. An RBF kernel based SVM [121] was used to classify each pair of sub-inventories by using these features: all accent curve parameters plus OD and RD. The $F1$

average over all inter sub-inventories for continuation- final is 0.4917. This low $F1$ score in case of the continuation class indicates that the accent curve parameters in this category could not be differentiated through *SNum*. Therefore, we could ignore *SNum* and merged the three continuation sub-inventories into one. For rise-fall the average $F1$ score for initial, middle, and final were 0.8228, 0.8595, and 0.6325, respectively. These high $F1$ score show that accent curves varied systematically as a function of the F_{Acc} features. Therefore, we kept the nine sun-inventories under rise-fall as is.

5.2.2.3 Synthesis

In this method, we run Festival on an input sentence to generate accent labels, syllable labels, and phrase boundaries. Then, we derive the foot structure, and determine *AT*, *Pos*, and *SNum* for each foot. The values *OD* and *RD* are predicted using forced alignment [159] applied to original test utterances[3]. A suitable accent sub-inventory is chosen for that foot by traversing the proposed DT using the first three features: *AT*, *Pos*, and *SNum* (middle panel of Figure 5.1). We calculate the Euclidean distance between the *OD*, and *RD* of the current foot, and the stored accent curves in the chosen sub-inventory. The five candidate accent curves with the lowest distance in that sub-inventory are retrieved. To minimize the differences between successive accent curve heights in a phrase, we apply a Viterbi search to the sequence of candidate accent curves; the observation matrix consists of the normalized duration distances and the transition matrix consists of the normalized accent curve height differences.

For the current phrase, the suitable phrase sub-inventory is chosen by using these two features: *PT* and *FNum*. We use the average of the stored phrase curves parameters in the chosen sub-inventory as synthetic phrase curve parameters.

5.2.3 Foot-based F_0 Generator using Neural Networks (FONN)

In this section, we discuss how FONN generates a F_0 contour given a text and its duration information as an input. Similar to DRIFT, we use GENIE to compute the component curves. Also we use similar feature sets as in DRIFT for accent curves. Dissimilar to DRIFT which uses a structured inventory of accent curve parameters, we use an ANN to compute accent curve parameters. We describe training and synthesis steps in Section 5.2.3.1 and Section 5.2.3.2, respectively.

[3]To ensure that the comparison strictly focused on the quality of the F_0 contours and was not affected by other aspects of the synthesis process

5.2.3.1 Training

Similar to the DRIFT model, for each utterance in the training data we do the following. First, we run Festival to generate accent labels, syllable labels, and phrase boundaries. Second, we derive the foot structure. Third, we apply GENIE to compute the component curves. Fourth, we calculate the RWMSE between the accent curve and the raw accent contour. We exclude any curve that does not meet a certain threshold on fitness error.

For the fifth step which differs from DRIFT, we store two vectors for each foot, an input and a target vector. The input vector consists of the features from feature F_{Acc}. We normalize the OD and RD by foot duration. The target vector consists of the parameters of the accent curve. Before storing the target vector, we normalize the parameters.

We use the input and target vector to train an ANN. The ANN consists of two layers as shown in the right panel of Figure 5.1. The input dimension is 10 which represents the first three binary features in F_{Acc} and the last two features in F_{Acc}. The output dimension is 7 which represents the accent curve parameters. The hidden layer size is 200. The hidden layer uses a sigmoid activation function and the output layer uses a linear activation function.

5.2.3.2 Synthesis

Like the DRIFT method (Section 5.2.2.3), an input sentence is segmented into phrases, each phrase is segmented into a foot sequence, and for each foot the F_{Acc} features are extracted. These feature vectors are given to the trained ANN sequentially to predict accent curves parameters. We use the predicted parameters to create accent curves for each foot. In order to create phrase curve, we use the DRIFT's phrase inventory by taking average over the stored phrase curves parameters in the chosen sub-inventory.

5.2.4 Experiments

We ran two experiments to evaluate the performance of the three intonation generation approaches subjectively: the first test measured the naturalness and the second test measured the ability to convey contrastive stress. We used Amazon Mechanical Turk [15], with turkers who have approval ratings of at least 95% and were located in the United States.

5.2.4.1 Database

We use the CMU Arctic speech database [75]. The database was automatically labeled via CMU Sphinx using the FestVox labeling scripts. We use speaker SLT, a US English female. This corpus

Algorithm 5.1 Automatic selection of test data

1: **for** 2000 iterations **do**
2: $A \leftarrow$ *Choose 50% of database randomly for training set*
3: **for** each token in A **do**
4: $Frq[token] \leftarrow$ *Extract the frequency of token*
5: $B \leftarrow$ *database - A*
6: **for** S in B **do**
7: $C \leftarrow$ *Replace tokens of the S with number from Frq)*
8: $x1 \leftarrow$ *median of the C divided by maximum of the C*
9: $x2 \leftarrow$ *number of zeroes in the C divided by totall number of tokens in the S*
10: $DisWell \leftarrow$ *Euclidean((1,0)(x1,x2))*
11: $DisPoor \leftarrow$ *Euclidean((0,1)(x1,x2))*
12: $Well \leftarrow$ *Choose 50 sentences with lowest DisWell*
13: $Poor \leftarrow$ *Choose 50 sentences with lowest DisPoor*
14: $wellSET \leftarrow$ *Choose 50 more frequent sentences from Wells*
15: $poorSET \leftarrow$ *Choose 50 more frequent sentences from Poors*
16: $randomSET \leftarrow$ *Choose 50 sentences randomly from remaining data*
17: $trainSET \leftarrow$ *remaining data*

contains 1132 utterances, which are recorded at 16bit 32KHz, in one channel.

5.2.4.2 Set coverage

In data driven approaches, data sparsity is a pervasive challenge [135]. We want to evaluate the impact of sparsity on the three methods by using a test data selection algorithm. We create three test sets that differ in terms of how they are covered by train set. Units used to compute coverage included the diphone, which is commonly used as a feature for set coverage [79] because it does not have sparsity of triphone and context independency of phonemes. They also included prosodic context via syllable (lexical) stress and word accent labels. Thus, each sentence was represented as a sequence of diphone/stress/accent tokens. We are interest to investigate the effect of whether train and test set are matched in terms of coverage of those tokens. We created four subsets of data: *trainSET*, containing training data; *wellSET*, containing test data that are well covered by *trainSET*; *poorSET*, containing test data that are poorly covered by *trainSET*; and *randomSET*, a random selection from the test data.

We create an algorithm (Algorithm 5.1) to select the four subsets. We randomly select half of database as a train set (A), and we calculate and store the occurrence frequency of each token. (Algorithm 5.1 step from 2 to 5). Then, for each sentence in the remaining data (B) we do followings. First, we replace each token in the sentence with its occurrence frequency value in A or with zero. Second, we calculate two distance metrics, *DisWell* and *DisPoor*, to measure the sentence coverage by A (Algorithm 5.1 step from 7 to 10). For example, if the sentence is

well covered by A, we expect to have low *DisWell* value and high *DisPoor* value, and vice-versa. Lower a distance the stronger evidence that how the sentence is covered by A. Third, we choose 50 sentences with lowest *DisWell* value and 50 sentences with lowest *DisPoor* value for well test set and poor test set (Algorithm 5.1 steps 12 and 13).

Since a randomization is involved, we need to repeat the process several times in order to lend credibility of the data selection. At the end of iterations, we select more frequent sentences of each sets from all iterations as the final sets (Algorithm 5.1 step from 14 to end).

5.2.4.3 Naturalness test

We ran three separate tests to compare each pair of three synthesis methods (HTS vs. DRIFT, HTS vs. FONN, and DRIFT vs. FONN). For each pair, we used a comparison test to evaluate the naturalness of the F_0 contours synthesized by the two methods. In this test, turkers heard two stimuli with the same content back-to-back and then were asked which they prefer using a five-point scale consisting of -2 (definitely First one), -1 (probably First one), 0 (unsure), +1 (probably Second one), +2 (definitely Second one). We randomly switched the order of the two stimuli. The experiment included 50 utterance pairs for each of the three test sets (total 150 pairs). Three control utterance pairs, which were trivial to judge, were added to the experiment to filter out unreliable turkers. Each turker only judged pairs from one test set (i.e., *poorSET*, *randomSET*, and *wellSET*). We employed a total of 150 turkers.

We evaluated the two approaches by imposing the F_0 contours generated by the two approaches onto recorded natural speech, thereby ensuring that the comparison strictly focused on the quality of the F_0 contours and was not affected by other aspects of the synthesis process. To ensure that the F_0 contours were properly aligned with the phonetic segment boundaries of the natural utterances, the contours were time warped such that the predicted phonetic segment boundaries corresponded to the segment boundaries of the natural utterances. Note that the predicted phonetic segment boundaries were the same for the two approaches. To compute the segment boundaries of the natural utterances, we used the HTS state durations and phoneme durations. Finally, we used PSOLA to impose the synthetic contours onto the natural recordings.

Figure 5.2 shows the results of the pairwise comparisons between the naturalness of the F_0 contours synthesized by the two configuration pairs (HTS-DRIFT, HTS-FONN, and DRIFT- FONN). In general, perceptual results indicated superior performance of DRIFT and FONN over HTS. DRIFT performed better than FONN in random and well coverage cases.

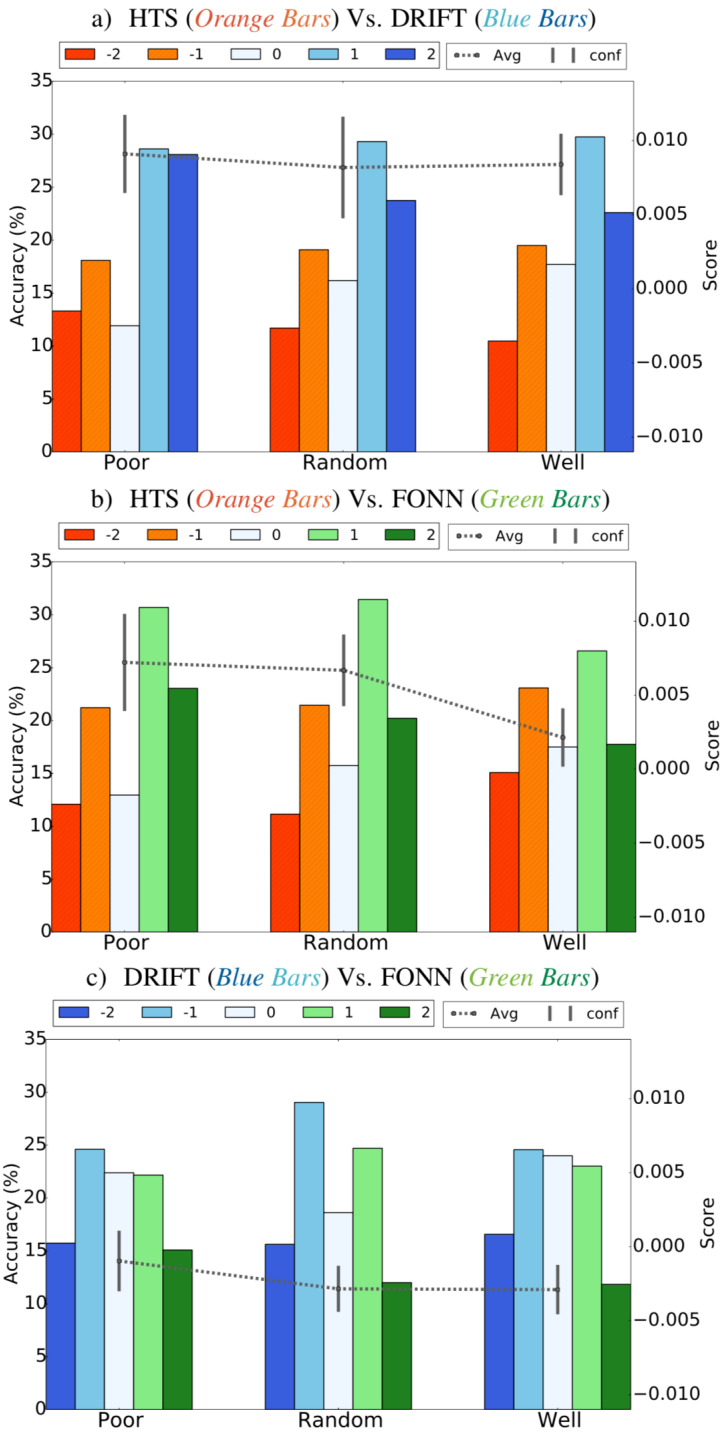


Figure 5.2: Each of the group bars (poor, random, and well) represent the histogram (in percentage (left y-axis)) of the related preference points: The five-point scale consists of -2 (definitely first version), -1 (probability first), 0 (unsure), +1 (probability second), +2 (definitely second). The dotted line and the confidence intervals correspond to the values (right y-axis) computed via Equation 5.1.

			HTS vs. DRIFT	HTS vs. FONN	DRIFT vs. FONN
Poor	t-test	t(49)	7.9034	6.7803	-0.6974
		p-value	***	***	-
	Randomization	mean	1.3277	0.4120	-0.8512
		SD	1.2454	1.2189	0.8353
Random	t-test	t(49)	5.9978	5.7140	-2.2792
		p-value	***	***	*
	Randomization	mean	1.1718	0.2137	-0.1916
		SD	1.0709	1.1669	0.9297
Well	t-test	t(49)	4.9139	2.0512	-2.3892
		p-value	***	*	*
	Randomization	mean	0.6584	0.5868	-0.1571
		SD	1.4475	0.9291	1.0863

– p > 0.05 * p < 0.05 ** p < 0.01 *** p < 1.0e-10

Table 5.1: Results of one-sample t-tests [t-value(df), p-value], and mean and standard deviation (SD) of the randomization-based t-statistic distribution for three pairwise comparisons in three test sets that vary in how well they are covered by the training data.

For significance testing, we first computed a score for each utterance using Equation 5.1, and then, separately for each test set, applied a one-sample t-test (Results are summarized in Table 5.1). In Equation 5.1, j, n, m, and C_{ji} stand for thej^{th} utterance in the current test set, the number of listeners, the number of utterance in the current test set, and the rating of the i^{th} listener for the j^{th} utterance, respectively, and || indicates the absolute values.

$$score_j = \frac{\sum_{i=1}^{n} (C_{ji}|C_{ji}|)}{\sum_{j=1}^{m} (\sum_{i=1}^{n} (|C_{ji}|))} \qquad , Cji \in \{-2, -1, 0, 1, 2\} \tag{5.1}$$

Conventional t-test results for the first and second comparisons (Table 5.1, first and second rows) show that the scores for DRIFT and FONN are significantly better than those for HTS for all test sets. The third comparison (Table 5.1, last row) indicates that the scores for DRIFT and FONN differed significantly from each other for two test sets (random and well), but were the same for *the poorSET*. The superiority of FONN over HTS, but not that of DRIFT over HTS, was reduced in the *wellSET*.

In order to show the robustness of the t-test results, we also performed a randomization test for each comparison in each test set. We randomly changed the signs of all ratings, computed the scores for each utterance, and then calculated the t-statistic. We repeated these steps 2000 times. The means and standard deviations of the resulting distributions are reported in Table 5.1, confirming the conventionally obtained significance levels. For example, the t-value of the first

comparison (HTS-DRIFT) for *the poorSET* is far from chance (e.g., 7.9034 deviates by 6.5757 standard deviations from the randomization mean of 1.3277, for a normal t(49) distribution with mean 1.3277 and SD of 1.2454, this yields a chance level less than $1.0e-10$).

In another experiment, we performed a test in which we compared the systems based on the impact of coverage. We first computed a difference score for each utterance, defined by the difference between the scores for the two approaches, and subsequently performed a two-sample t-test comparing these difference scores between the *poorSET* and *wellSET* data. We only found statistically significant results for the HTS-FONN comparison ($t(49) = -3.5675$, $p = 2.8036e-4$, one-tailed; these results were again confirmed using a randomization test). This result showed a powerful significant trend for the impact of coverage to be stronger for the HTS approach than for FONN. Figure 5.2 (gray curve, right y-axis) also showed the results of comparing the two systems in terms of the impact of coverage of each test set by the *trainSET*.

5.2.4.4 Testing the ability to synthesize text marked up for contrastive stress

To evaluate the ability of DRIFT to handle marked-up input, we created a contrastive emphasis test. First, we selected 22 sentences from the test data that contained a pair of noun-adjective words for which contrastive stress is meaningful. Then, for each of these sentences, we generated two utterances such that in each utterance one of the two words was emphasized. For example, for the sentence "This is a red house", with capitals indicating stress, we considered "This is a RED house" and "This is a red HOUSE". We used DRIFT for generating the F_0 curves, and then implemented a simple rule whereby we increased and decreased amplitudes of the accent curves associated with the emphasized and non-emphasized words by multiplication with factors of 3 and 0.5, respectively.

In the perceptual test, each turker was asked to imagine the following situation: "Two people, John and Mary, are having a dialogue; unfortunately, John is not a good listener so that Mary has to repeat what she just said, emphasizing the word that John— apparently —got wrong. Your task is to figure out which word John got wrong." The experiment was administered to 50 turkers with each turker judging 44 (22×2) sentences. The percentage of emphasized words conveyed correctly was 84.85%. We also applied the same test for a recorded natural voice (female native American English speaker) for the 44 sentences, and obtained a nearly identical accuracy of 85.15%. We concluded that DRIFT's ability to convey contrastive stress is comparable to that of natural speech.

5.3 Proposing an F_0 Adaptation Method for TTS Systems

In this section, we propose a new intonation adaptation method to transform the perceived intonational identity of a TTS voice to that of a target speaker with a small amount of training data. For modeling intonation, we use GENIE that captures F_0 contours with a small number of parameters at two levels: the foot level and the phrase level. For generating F_0 contours, we used the DRIFT method which is based on GENIE. Because the number of parameters to be estimated is relatively small, it is feasible to adapt the speaking style using any mapper function, such as the Joint distribution Gaussian mixture model (JDGMM). We compare our proposed method with a baseline adaptation method in which the source F_0 contour is transformed linearly such that the per-utterance mean and variance of the target F_0 contour is unaltered; yet, this generated F_0 contour still has the dynamics of the source F_0 contour. Thus, in this part we address two questions. First, is adapting just the mean and SD enough? And, if not, does DRIFT succeed in capturing extra, dynamic information that is lost in the linear transformation approach?

In Section 5.3.1, after introducing the baseline we briefly review JDGMM. Then in Section 5.3.2, we discuss how we train the JDGMM mapper, and how we use this mapper on the estimated source and target component curves derived from DRIFT to generate target F_0 contour. Finally, in Section 5.3.3, we ran two subjective listening experiments (speech similarity and speech quality) to study the performance of the two methods for two male target speakers.

5.3.1 Intonation Mapping

5.3.1.1 Baseline: Mean-Variance Linear Mapper

In Voice Conversion (VC) and TTS literature, it is often assumed that the F_0 mean and standard deviation (SD) are adequate to capture prosodic style [148]. The most common method for transforming F_0 values is to globally match the average mean and SD of the target speaker's F_0 contour, while maintaining the dynamic intonation pattern of the source. With this assumption, intonation can be transformed by mapping $log - F_0$ using a linear transformation, where μ and σ represent the average mean and SD of the $log - F_0$ of the training set [18].

$$F_{mimicked} = \frac{\sigma_{target}}{\sigma_{source}}(F_{source} - \mu_{source}) + \mu_{target} \tag{5.2}$$

For the baseline method, we used a slightly different linear transformation in which the baseline does not have a training stage. Therefore, in the baseline method μ and σ represent the mean and SD of the original utterances of the test set. This assumption gives the linear model a strong

opportunity to overfit the target speaking style in a given sentence, making it in principle more effective than the average-mean-and-SD linear mapper.

5.3.1.2 Joint Distribution GMM Mapper

In this section, we present a brief overview of the GMM mapping function [63]. Let $X = x_1, ..., x_n$ and $Y = y_1, ..., y_n$ be sets of parameters vectors for n segments (foot or phrase in the case of mapping accent parameters or phrase parameters, respectively) from the source and target model. Note that each vector is normalized using the maximum and minimum values of X and Y. Let $Z = [X, Y]$ be the joint source-target parameters vector. A GMM represents the distribution using M multivariate Gaussians;

$$P(z) = \sum_{m=1}^{M} \alpha N(z; \mu_m, \sum_m)$$

where $N(z; \mu_m, \sum_m)$ is a normal distribution with mean μ_m and covariance $\sum_m$ of component m. The prior probability of the component m is represented by α_m. The parameters of the GMM are calculated using the Expectation Maximization (EM) algorithm on the joint vector Z.

During transformation, for each component, we estimate the weighted mixture of the maximum likelihood estimator of the target vector given the source vector for each component;

$$\hat{y}_i(x_i) = E[Y|X = x_i] = \sum_{m=1}^{M} \omega_m^x(x_i)[\mu_m^y - \sum_m^{xy} \sum_m^{xy-1} (xi - \mu_m^y)]$$

where $\omega_m^x(x_i)$ is a posterior probability that the segment x_i belongs to the class described by the component m.

$$\omega_m^x(x_i) = \frac{\alpha_m N(x_i; \mu_m^x, \sum_m^{xx})}{\sum_{k=1}^{M} \alpha_k N(x_i; \mu_k^x, \sum_k^{xx})}$$

5.3.2 Intonation Adaptation

5.3.2.1 Mapper Training Procedure

The aim of F_0 adaptation is to predict the intonation style of the target speaker with a small amount of parallel training data, since otherwise one might just as well obtain a complete set of speech recordings of the target speaker and avoid the transformation process all together. We randomly select a small set of recordings (section 5.3.3.1, 28 parallel utterances) from the source

and target speakers. For each utterance, we apply GENIE to decompose the F_0 contour of the utterances into component accent and phrase curves. We use the estimated source and target accent curve parameters to train a JDGMM mapper with two components (M = 2). This process is performed similarly for phrase curve parameters. Thus, the mapper operates in the parameter space defined by the DRIFT model and indirectly mapped source F_0 contours onto target F_0 contours (Top block-diagram in Figure 5.3a).

5.3.2.2 Adaptation Procedure

Similar to the DRIFT model, for an input sentence we do the following. First, we run Festival to generate accent labels, syllable labels, and phrase boundaries. Second, we derive the foot structure, and determine *AT*, *Pos*, and *SNum* for each foot. Third, we predict the values of *OD* and *RD* using forced alignment applied to the original utterance [10]. Forth, we retrieve the five candidate source accent curves with the lowest distance in the selected sub–inventory.

For the fifth step which is not part of DRIFT, we apply the accent mapper to each of those five candidates to predict five transformed accent curves per foot. At the end similar to the DRIFT model, we apply a Viterbi search to minimize the differences between successive transformed accent curve heights in a phrase.

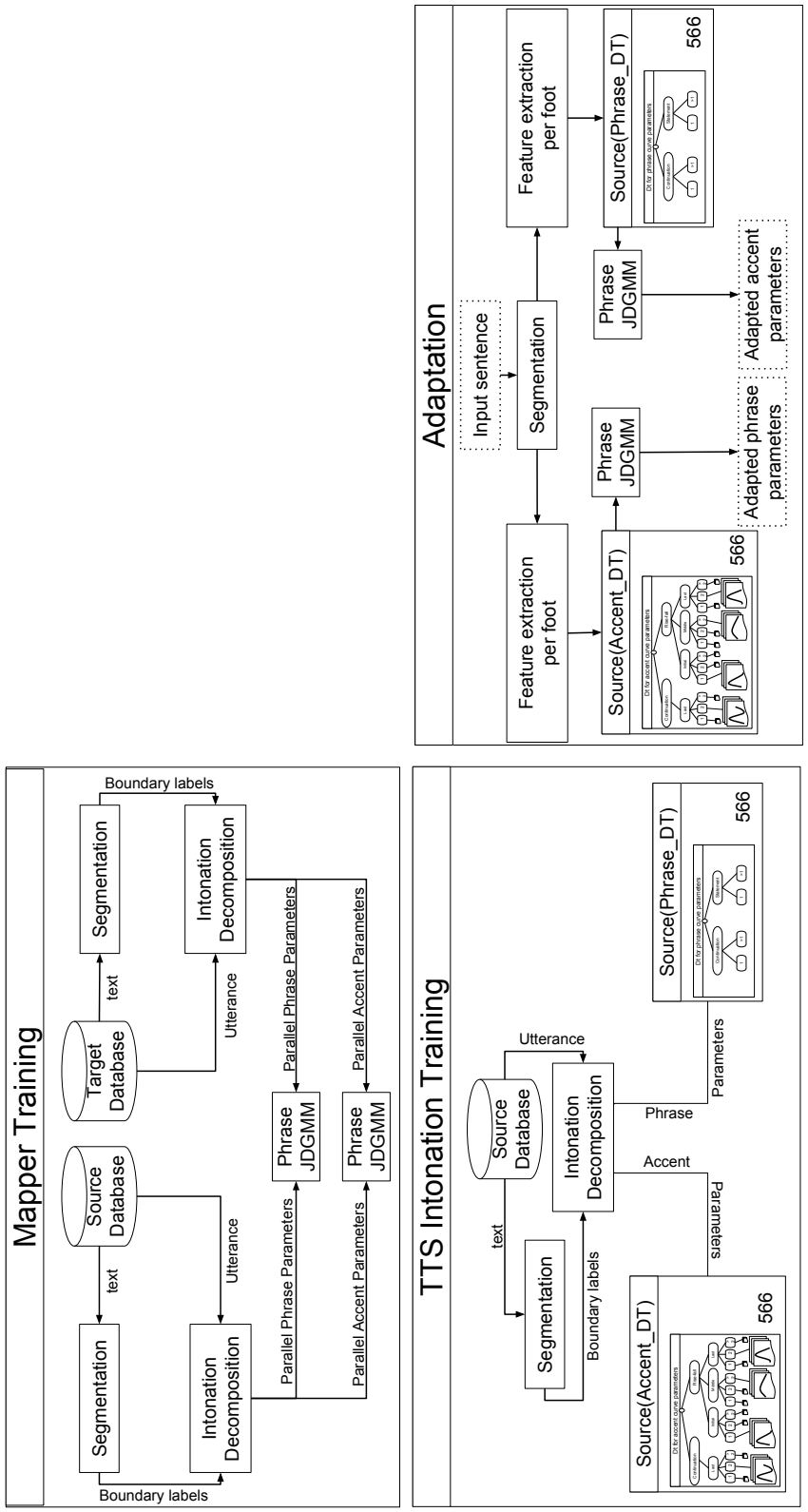
For the current phrase, the F_{Phr} features are extracted. Parameters of the source phrase are predicted by calculating the average of the stored phrase curves parameters in the selected sub-inventory. Transformed phrase parameters are estimated by applying the phrase mapper to predicted source phrase parameters. (Figure 5.3b)

5.3.2.3 Synthesis Procedure

During synthesis, for an input sentence we do the following. First, we apply the mapper to the source speaker's DRIFT model parameters (i.e., the parameters that would be used to generate TTS output during normal operation, see bottom block diagram in Figure 5.3a) to generate predicted target speaker DRIFT parameters (described in Section 5.3.2.2). Second, we use these predicted parameters to generate the accent and phrase curves, which are added together to generate a target F_0 contour. Finally, we use this target contour in the process of generating output speech.

5.3.3 Experiments

We ran two tests to perform a subjective evaluation of the intonation generation performance of the two approaches: the first test measures speech quality and the second test measures speech



(a) Training (b) Adaptation

Figure 5.3: Block-diagrams of training and adaptation of proposed method

similarity between a stimuli and the target speaker. We used Amazon Mechanical Turk with turkers who have approval ratings of at least 90% and were located in the United States.

In each test, we evaluated the two approaches by imposing the F_0 contours generated by the two approaches onto recorded natural speech, thereby ensuring that the comparison strictly focused on the quality of the F_0 contours and was not affected by other aspects of the synthesis process. To ensure that the F_0 contours were properly aligned with the phonetic segment boundaries of the natural utterance, the contours were time warped so that the predicted phonetic segment boundaries corresponded to the segment boundaries of the natural utterance. To compute the segment boundaries of the natural utterance, we used the phoneme durations predicted by forced alignment [10]. Finally, we used PSOLA to impose the synthetic contours onto the natural recordings.

5.3.3.1 Databases

For the TTS adaptation experiment, we use the CMU Arctic database [75] as in Section 5.2.4.1. We consider the speaker SLT as the source speaker and two male speakers (English speaker: BDL, and Scottish speaker: AWB) as the target speakers. Utterances of SLT and BDL were recorded in a sound proof room while AWB's utterances were recorded in a quiet office.

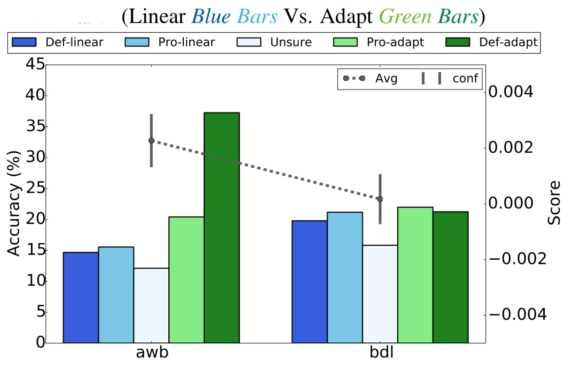
We use two training sets for the subjective evaluation: a large set, which included 566 training utterances, and a small set, which included 28 (5% of the large set) training utterances. We use the large set for training the source model and the small set for training the mapper. A set of 150 utterances is selected randomly for test purposes.

5.3.3.2 Speech Quality Test

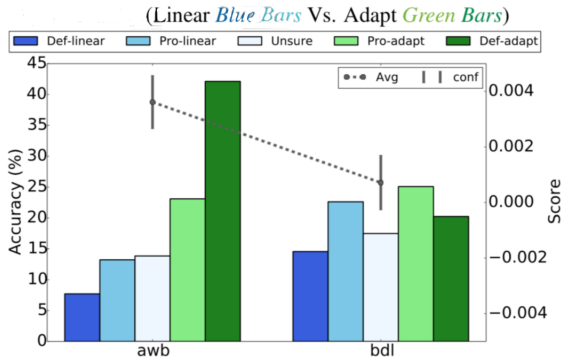
We used a comparison test to evaluate the quality of the F_0 contours synthesized by the two approaches. In this test, turkers heard two stimuli with the same content back-to-back and then were asked which they preferred using the same five-point scale as in Section 5.2.4.3. We randomly switched the order of the two stimuli. Three trivial-to-judge utterance pairs were added to filter out unreliable turkers. Each turker judgs 50 utterance pairs. We ended up with judgements from 150 turkers in total.

Figure 5.4a shows the results for the test sets for two target speakers. For significance testing, we first computed a score for each utterance using Equation 5.1, and then, separately for each test set, we applied a one-sample t-test.

Conventional t-test results show that the scores of the two methods differed significantly from each other for AWB (first two rows of Table 5.2). We also performed a randomization test for the



(a) Quality



(b) Similarity

Figure 5.4: Speech quality and similarity test. Dashed curves correspond to the values computed via Equation 5.1

same difference by randomly changing the signs of all ratings 2000 times, computing the scores for each utterance, and calculating the t-statistic. (This randomization test is more conservative than the conventional t-test.) The means and standard deviations of the resulting distributions are summarized in Table 5.2, and yield conclusions similar to those based on the conventional t-tests. For speaker BDL, the baseline worked as well as our proposed method. This suggests that both the source (SLT) and target (BDL) have similar intonation patterns: matching the mean and SD appeared sufficient.

5.3.3.3 Speech Similarity Test

To evaluate speaker mimic accuracy, we ran a speaker similarity test. In this test, turkers heard three stimuli. First, a natural recording of the target speaker to convey the target speaking style. Second, two stimuli with the same content (but contents differing from that of the natural recording) back-to-back. They were then asked which of these two stimuli provided the best mimic

			Linear vs. Adapt (Quality)	Linear vs. Adapt (Similarity)
AWB	t-test	t(149)	5.7749	8.8257
		P-value	***	***
	Randomization	Mean	1.5082	2.0077
		SD	2.0859	3.2153
BDL	t-test	t(149)	0.4874	1.9756
		P-value	-	*
	Randomization	Mean	0.6518	0.7022
		SD	0.6406	1.0415

– $p > 0.05$ * $p < 0.05$ ** $p < 0.01$ *** $p <$ 1.0e-10

Table 5.2: Quality and similarity experiment results: one-sample t-tests [t-value(df), p-value], and mean and standard deviation (SD) of the randomization-based t-statistic distribution comparing the linear and Adapt methods, for two speakers (AWB and BDL)

		Linear vs. Natural		Adapt vs. Natural	
		Mean of F_0	SD of F_0	Mean of F_0	SD of F_0
AWB	t(149)	-0.5502	-2.5262	-11.4206	-13.3240
	P-value	-	**	***	***
BDL	t(149)	-0.1684	-1.2474	-10.7368	-7.9373
	P-value	-	-	***	***

– $p > 0.05$ * $p < 0.05$ ** $p < 0.01$ *** $p <$ 1.0e-10

Table 5.3: Differences in mean and SD between transformation methods and natural target speech: one-sample t-tests [t-value(df), p-value]of two speakers (AWB and BDL) for two pairwise comparisons of linear and Adapt methods with Natural method.

of the target using the same five-point scale as in Section 5.2.4.3. We randomly switched the order of the two stimuli. The experiment was administered to 150 turkers, with each turker judging 50 utterance pairs. Three trivial-to-judge utterance pairs were added to the experiment to filter out unreliable turkers.

Figure 5.4b shows the results for the test sets for two target speakers. Our proposed method was clearly superior for speaker AWB, and marginally superior for BDL (last column of Table 5.2).

Interestingly, for both speakers, our proposed method produced means and SDs that differed far more from those of the target speaker than the linear method (Table 5.3). For example in first row, the small t-value of the mean of F_0 comparison in linear vs. natural indicates that the mean difference between linear and natural is insignificant. In case of adapt vs. natural, the high t-value indicates that the mean difference between adapt and natural is highly significant. By

		Natural	Linear	Adapt
AWB	Mean of F_0	150.2694	149.0506	131.3975
	SD of F_0	57.5239	47.2186	17.7554
BDL	Mean of F_0	126.5171	126.3289	117.3713
	SD of F_0	23.0289	21.0202	11.7965

Table 5.4: Mean of the mean and standard deviation (SD) of F_0 of two speakers AWB and BDL from linear, Adapt, and Natural speech.

looking at the actual mean of the mean and standard deviation of F_0 in Table 5.4, we can see that means and SDs of linear method is very close to those form natural speech. Even though there is a significant difference between means and SDs of our proposed method and natural speech, yet, for both speakers, our proposed method was perceived as producing a significantly better mimic than the linear method. Apparently, copying the mean and SD of a target speaker is neither sufficient nor necessary for prosody mimic.

5.4 Conclusion

In the first half of this chapter, we proposed two foot-based intonational approaches for F_0 generation: DRIFT and FONN. The key characteristics of DRIFT are as follows. During training, it creates two structured inventories of component accent and phrase curves using GENIE. During test, it retrieves component curves from those two structured inventories. The accent curves are selected based on (1) the distance in a low-dimensional feature space between a foot in the to-be-synthesized sentence and the feet associated with the accent curves in the inventory and (2) height differences between successive accent curves. Third, usage of a superpositional model in which selected accent curves are added to a phrase curve. The phrase curve is created by taking average over the stored phrase curves parameters in the chosen sub-inventory. The key characteristics of FONN are as follows. First, like DRIFT, FONN uses GENIE to compute the component curves. Second, it uses similar feature sets as in DRIFT. Third, unlike DRIFT which uses accent curve parameter templates, it uses a trainable parametric method to compute accent curve parameters.

Both FONN and DRIFT methods result in F_0 curves that are guaranteed to have the desired smooth suprasegmental shapes and are well-suited to handle sparse training data. Perceptual results indicated superior performance of FONN and DRIFT compared to a frame-based approach (HTS). Using our test data selection algorithm, we show that FONN and DRIFT outperform HTS across all the three test sets. This shows the usefulness of GENIE for speech synthesis. As we

predicted, FONN handles impact of sparse training data in *poorSET* better than DRIFT. We also showed that DRIFT can generate compelling contrastive stress via markup.

We surmise that, for speech synthesis, template based approaches, such as DRIFT that create accent curves that inherently preserve natural detail are to be preferred over approaches that compute accent curves. It remains to be seen, however, whether FONN may nevertheless outperform template based approaches in exceptionally sparse data conditions where several slots in the template tree are missing.

In the second half of this chapter, our proposed intonation adaptation method showed promise as a way to capture the dynamics of the F_0 contours of a target speaker. Whether it performs better than a much simpler linear transformation of the source speaker's F_0 contours depends on the degree and type of differences between the source and target contours. Given the pronounced intonation differences between the North American and (Glasgow) Scottish dialects [85, 22], it is perhaps no surprise that the linear model fared less well for speaker AWB. We need to take into account that the linear model as applied in this study did not accurately reflect its actual use in synthesis, in which the per-token mean and SD are — obviously — not given and where thus estimates need to be used. Thus, we do now know whether the linear model as used in practice might have produced significantly worse results than our proposed method for speaker BDL, and not only, as was the case in the linear method employed in this study, for speaker AWB. Finally, our results may have implications for the role in speaker mimic of copying the mean and SD, or, in fact, of any approach based on copying statistical moments of the F_0 distribution and that does not take dynamic patterns into account.

Chapter 6

Towards Intonation Based Classification

6.1 Motivation

Speaker (or speaker state) classification covers a variety of cases: emotion classification [147, 81], speaker verification [183], classification of individuals with autism spectrum disorder vs. neurotypical individuals [170], classification of individuals with dysarthria vs. neurotypical individuals [43, 174], clear vs. conversational speaking styles, dialect classification[48], etc. In general, speaker classification involves using spectral and prosodic features extracted from the speech. Typically, these studies extract a large number of acoustic features from the speech signal and use machine learning in a standard train and test classification paradigm, often achieving good accuracy of classification.

There are two drawbacks to these general approaches. First, they are often not informative for scientists working in the field in question (e.g., autism researchers), because they are interested in which features are the most important ones for classification and why. Just knowing that a classifier performs at 90% accuracy hardly serves their scientific enterprise. Of course, in certain industrial or governmental applications, accuracy of classification is the primary or even sole interest. Second, these approaches require that the recording conditions – microphone, room acoustics, distance to microphone – are not in the least confounded with the classes under consideration. The large number of acoustic features may capture differences in recording conditions, so that the final classification result may have little to do with the classes of interest. This is particularly dangerous in multi-site data collection efforts in which each site is responsible for recording a specific class. In fact for the VoxForge database of languages and dialects, researchers have achieved high accuracy classification using only one second of silence [47].

Our proposed approach to speaker classification exclusively uses features derived from the F_0 contours. A potential advantage of this is that these contours are generally less sensitive to

recording conditions than spectral features, unless these conditions are so poor that for some classes the F_0 contours are systematically more inaccurate than for other classes. However, this problem can be addressed by verifying that no systematic differences between classes exist in the quality of the F_0 contours, either by manually inspecting sufficiently large subsamples or by automatically measuring typical manifestations of pitch tracking errors, such as doubling, halving, or values outside of some reasonable range. A special aspect of our approach is the focus on F_0 contour dynamics – often underused in speaker group classification, which typically focuses just on basic features, such as utterance-level mean and standard deviation.

In chapter 3, we proposed a generalized intonation model (GENIE), and in Section 3.3.1 we showed that GENIE's component curves are linguistically meaningful. Having this ability gives us an advantage to investigate the prosodic features in a more relevant prosodic structure. This could be very suitable for classification purposes especially when there are no speech intelligibility differences between the speaker groups under classification, the speaker groups are not different in terms of basic common prosodic features, or the spectral features are not reliable.

In this chapter, we focus on answering the following question: Can F_0 dynamics differences between two speaker groups be used to differentiate one from another? In section 6.2 we show that statistically there are differences between F_0 dynamics of participants with Parkinson's Disease (PD) and healthy control participants. In addition, we show that GENIE was better in bringing out these F_0 dynamic differences than raw F_0 contours and other less sophisticated baseline methods.[1] In section 6.3, we investigate F_0 dynamic differences between clear speech (CLR) and conversational speech (CNV). We show this differentiation is attributed to F_0 dynamics and is independent of utterance duration and F_0 range. Finally in section 6.4, we propose a new prosody based approach to classify at least two speaker groups. This classification uses the assumption that two speaker groups can be differentiated through their F_0 dynamic differences.

6.2 F_0 Dynamics in Hypokinetic Dysarthria

Hypokinetic dysarthria (Hd), which often accompanies Parkinson's Disease (PD), is characterized by hypernasality and by compromised phonation, prosody, and articulation. In this section, we propose automated methods for the detection of Hd. Whereas most such studies focus on measures of phonation, the focus of this section is on prosody, specifically on F_0 dynamics. Intonation in Hd is clinically described as involving mono pitch, which has been confirmed in numerous

[1] This section is based on work published in 2014 IEEE Spoken Language Technology Workshop[28].

studies reporting reduced within-utterance pitch variability; However, past work has failed to quantitatively measure it.

In terms of GENIE, reduced variability could result from atypical values of multiple components. First, there could be a reduced slope of the phrase curve: whereas in typical speech, there generally is a declination in F_0, perhaps the underlying factor for reduced within-utterance variability in PD is the lack of such declination. Second, there could be reduction in the number of feet (or in other words, fewer words receiving emphasis). Third, reduction in the height of either all accent curves or specific (e.g., phrase-initial, final) accent curves. To assess these components, we defined four explicit F_0 methods in Section 6.2.1. In Section 6.2.1.2, we compute statistical features from each methods and use them to distinguish participants with Hd from healthy controls. In Section 6.2.2, We show a new measure of F_0 dynamics, based on GENIE, which performs Hd vs. Control classification more accurately than simpler versions of the model or conventional variability statistics.

6.2.1 Method

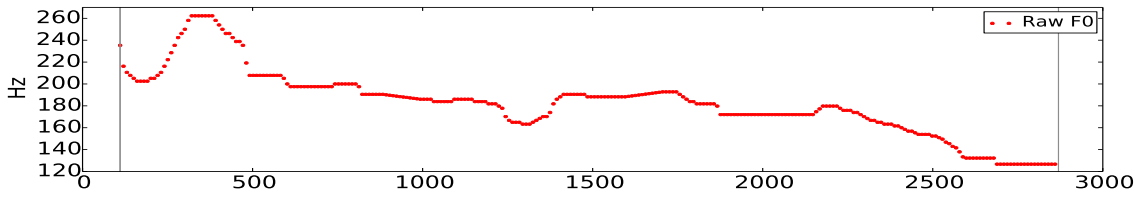
Baseline, Global Pitch Method: We use the per-utterance mean and standard deviation (SD) of the raw F_0 values as features.

Local Pitch Method: We define *Local Pitch Method* as a superpositional approach, where the phrase curve is discontinuous, consisting of linear segments that each have a zero slope. In other words, the frequency value of the phrase curve in each foot is equal to the minimum F_0 value in a foot (Figure 6.1b). The accent curves are obtained by, for each foot, subtracting this phrase curve from the F_0 values (Figure 6.1d). This method is used to assess the importance of a sloping, continuous phrase curve.

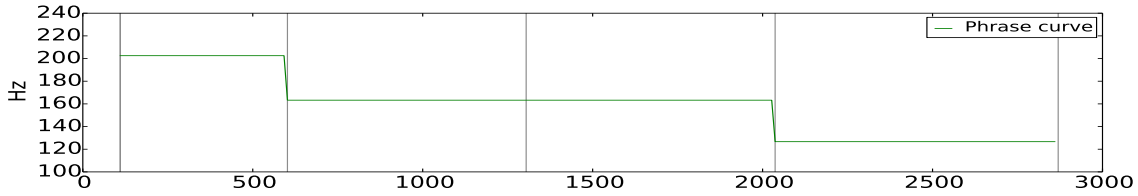
Raw Accent Method: The *Raw Accent Method* is similar to the *Local Pitch Method* in that accent curves are obtained by subtraction of a phrase curve from the raw F_0 curve; what differs is the phrase curve shape (Figure 6.1e), which we use GENIE to estimate the phrase curve.

Weighted Raw Accent Method: The *Weighted Raw Accent Method* is similar to the *Raw Accent Method* in that accent curves are obtained by subtraction of a phrase curve from the raw F_0 curve; what differs is that we apply a weight to each frame obtained by the multiplication of the voiced/unvoiced flag and the energy. We use GENIE to estimate the phrase curve.

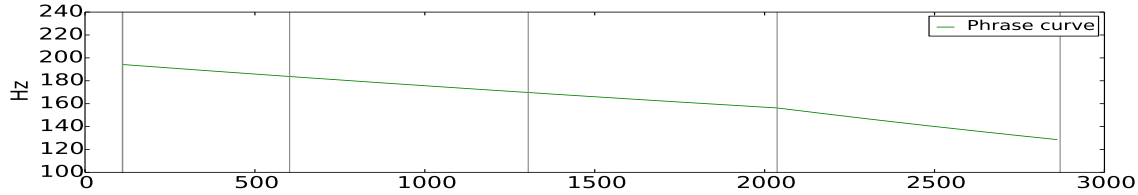
GENIE Accent Curve: We use GENIE for modeling F_0 contours. In GENIE, the phrase curve



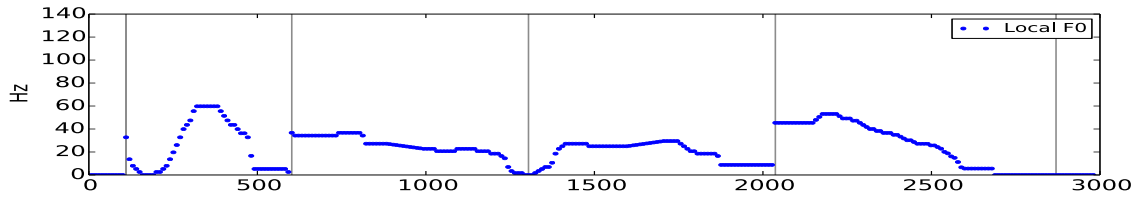
(a) Raw F_0 contour



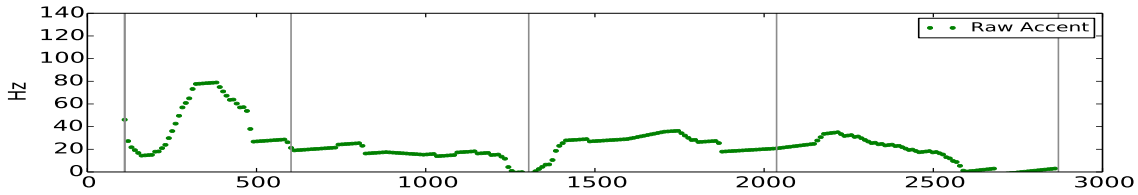
(b) Phrase curve used in the *Local Pitch Method*



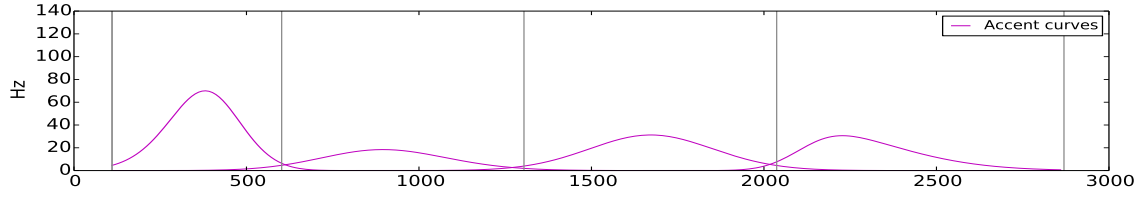
(c) Phrase curve used in the *Raw Accent Method, the Weighted Raw Accent Method*, and *GENIE*



(d) Accent curves of the *Local Pitch Method*



(e) Accent curves of the *Raw Accent Method*



(f) Accent curves of *GENIE*

Figure 6.1: Example F_0 decomposition contours of a 49 year old female in the Control group using three F_0 models. Decomposition applied to a sentence with foot boundaries marked with brackets "[All are be][lieved to be][embassy emplo][yees]."

consists of two connected linear segments, between the phrase start and the start of the final foot, and between the latter and the end point of the phrase, respectively (Figure 6.1c). We use a combination of a skewed normal distribution and a sigmoid function to model three different types of accent curves (Figure 6.1f).

6.2.1.1 Participants and data preparation

Participants were ten individuals with PD (age 42-80) and ten healthy controls (age 49-71). The average ages did not differ significantly ($t(18) = 1.08, p > 0.25$). Participants were selected to have good speech intelligibility. And indeed, the average Speech Intelligibility values, as measured via Yorkston and Beukelman (1996)'s *Sentence Intelligibility Test* [189], were 96.3 and 97.4 in the PD and control groups, respectively (t(18)=1.21, p>0.2, two-tailed). Thus, these were groups whose speech problems, if present at all, were subtle and hence pose a challenging test for any classification algorithm. Using greedy text selection methods [167], we selected 37 sentences from the Gigaword Corpus [46] to maximally cover a (symbolic) feature space defined by features known to affect F_0, such as predicted sentence and word stress, sentence length, and word length [168].

We used the YAAPT algorithm [197] to extract F_0 contours. We applied linear interpolation between voiced areas to replace the unvoiced areas. Roughness, hoarseness, and breathiness, typical not only in Hd but also more generally in older individuals [134] increases F_0 halving and doubling [25]. Therefore, we manually corrected the extracted F_0 curves, blind as to diagnostic status (PD vs. Control). Finally, we converted the F_0 values into a logarithmic scale to reduce the impact of the unequal gender distributions in the two groups.

We used Festival toolkit [11] to extract syllable stress, pitch accent, and phrase boundary labels for each sentence. We used these labels to generate foot structure which is GENIE's requirement.

6.2.1.2 Feature extraction

In this section, we computed features for each extracted accent curve (via the four methods), distinguishing between feet in phrase-initial (first foot in an intermediate phrase), phrase-final (last foot in an intermediate phrase), and phrase-medial (i.e., neither initial nor final) position. We refer to this variable that says whether a foot is initial, final or medial as *Pos*. We computed four statistical features per extracted accent curve in each foot: 1) Location (*loc*): location of the peak normalized by foot duration. 2) Magnitude (*mag*): the amplitude of the accent curve. 3) Weighted temporal standard deviation ($WTSD$): the $WTSD$ of the accent curve's distribution (Equation 6.1). In equation 6.1, t_i and x_i are the ith sample of time and accent curve value. $\bar{x}_t$

is the weighted average of time computed by $\sum t_i x_i / \sum x_i$. 4) Weighted temporal skewness ($WTSk$): the $WTSk$ of the accent curve (Equation 6.2).

$$WTSD = \sqrt{\sum x_i (t_i - \bar{x_t})^2 / \sum x_i} \tag{6.1}$$

$$WTSk = \frac{\sum x_i (t_i - \bar{x_t})^3 / \sum x_i}{WTSD^3} \tag{6.2}$$

To explore the discriminatory power of each feature, we applied t-tests to the per-speaker means of these accent features.

For SVM based classification we used larger sets of features, which also included per-speaker standard deviations (SD). $Set1$ was used for the *Global Pitch Method* and $Set2$ for the other methods, where:

$$\bullet Set_1 = \begin{cases} \text{(Per-speaker) } median\, of\, pitch\, mean,\ SD \\ \text{(Per-speaker) } SD\, of\, pitch\, mean,\ SD \end{cases}$$

$$\bullet Set_2 = \begin{cases} Pos \\ \text{(Per-speaker) } median\, of\ loc, mag, WTSD, WTSk \\ \text{(Per-speaker) } SD\, of\ loc,\ mag, WTSD, WTSk \end{cases}$$

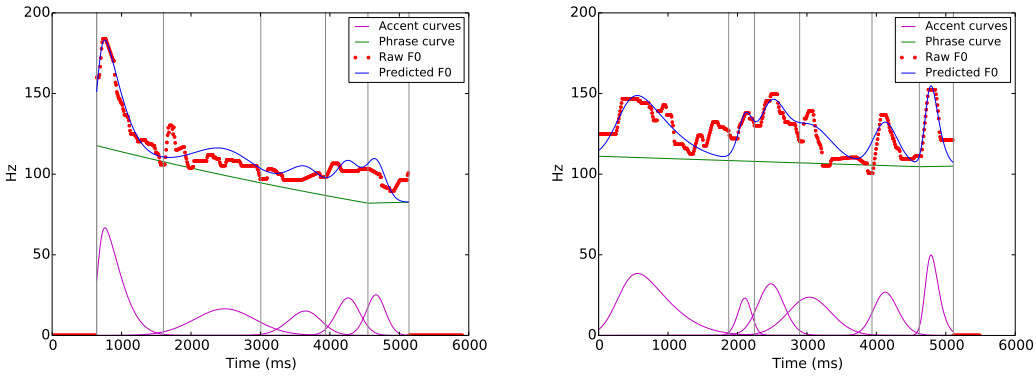
6.2.2 Experiments

6.2.2.1 Performance of the Global Pitch, Local Pitch, and Raw Accent methods.

For the *Global Pitch Method*, we extracted two commonly used prosodic features, the mean and SD of the F_0 curve for each utterance (37 utterances for each speaker), and four features (loc, mag, $WTSD$, and $WTSK$) for the *Local Pitch* and *Raw Accent* methods, for each foot of each utterance (ranging from 96 to 118 feet per speaker).

Before applying the classification, we first want to determine whether there is some significant differences between two groups (PD vs. Control) in terms of the extracted features or not. We applied two-group, two-tailed t-tests (PD vs. Control) to these features. For the *Global Pitch Method* features, no significant differences were found. The third and fourth rows in the Table 6.1 evaluate the features derived from the *Local Pitch* and *Raw Accent* methods, and present some marginally significant results. Interestingly, only the phrase-initial feet seem to matter.

We next employed an RBF kernel based SVM using the scikit-learn toolkit [120] to classify PD vs. Control for each method. We set the *gamma* and C SVM parameters to 10^{-1} and 10^5,



(a) Participant in Control group (b) Participant in PD group

Figure 6.2: Fitted curves for two 66-year old male participants.

respectively. We used Set_1 for the *Global Pitch Method* and Set_2 for the other methods. We used accuracy and $F1$ measures to evaluate the SVM results. The accuracy is the average of the true positive (TP, the percentage accurate classification of participants with PD) and true negative (TN, the percentage accurate classification of control participants) rates; $F1$ is computed from Equation 6.3.

$$F1 = \frac{2\,TP}{2\,TP + FP + FN} \tag{6.3}$$

where FP is the false positive rate $(100 - TP)$, and FN the false negative rate $(100 - TN)$. Table 6.2 shows for each method the averages over all selections of two held-out participants.Based on the t-test results, features extracted from the *Global Pitch Method* essentially yield chance performance (with 52% accuracy in classification). In contrast, features extracted from the two other methods perform better than chance. This suggests that foot-based features are more informative than global, whole-phrase features.

6.2.2.2 Performance of the *GENIE Accent Curve*

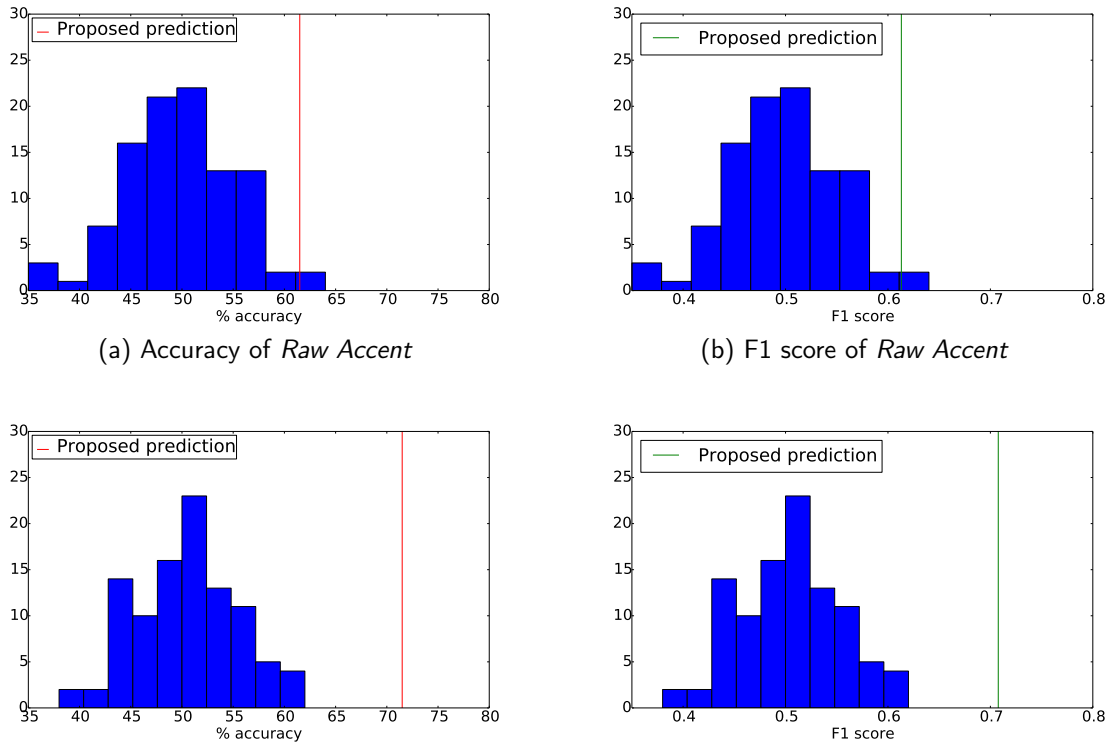
We now turn to the *GENIE Accent Curve*. To ensure that any results are not due to a better model fit in one group, we applied a t-test to the per-participant means of the root mean square (RMS) deviation of the predicted and observed F_0 values. No significant difference between the groups was found, with the RMS values for the PD and control groups at 0.82 and 0.88, respectively. Figure 6.2 shows an example of F_0 decomposition of the sentence "Afghan government officials were not immediately available to confirm the decision" into accent curves and phrase curve for

			GENIE Accent		*Local Pitch*	*Raw Accent*		*Weighted Raw Accent*	
Foot position (Pos)	Initial	Feature	*loc*		*WTSD*	*loc*	*WTSD*	*loc*	*WTSD*
		Mean_{PD}	0.677		29.627	0.688	24.247	0.669	24.074
		$\text{Mean}_{Control}$	0.629		25.395	0.643	21.218	0.630	20.800
		P-value	0.008		0.080	0.060	0.100	0.100	0.090
	Medial	Feature	*loc*	*WTSK*	–	–		*WTSK*	
		Mean_{PD}	0.435	0.108	–	–		0.239	
		$\text{Mean}_{Control}$	0.462	0.053	–	–		0.143	
		P-value	0.080	0.090	–	–		0.100	
	Final			–	–	–		–	

Table 6.1: P-values and means for two-group, two-tailed t-tests (PD vs. Control) as a function of Pos, method, and feature; p-values larger than 0.1 are omitted.

Method	$TN(\%)$	$TP(\%)$	*Accuracy* (%)	$F1$(score)
Global Pitch	30	75	52.5	0.612
Local Pitch	70	62	66.0	0.646
Raw Accent	62	61	61.5	0.613
Weighted Raw Accent	58	68	63.0	0.645
GENIE Accent	74	69	**71.5**	**0.708**

Table 6.2: Classification performance for each method



(a) Accuracy of *Raw Accent*

(b) F1 score of *Raw Accent*

(c) Accuracy of *GENIE Accent*

(d) F1 score of *GENIE Accent*

Figure 6.3: Reliability of the classification's result

two 66 year old male subjects in each group. We note the difference in the shape of the accent curves (e.g., location of the peak, skewness and SD), especially for the phrase-initial foot.

After extracting the four standard features from the estimated accent curves (i.e., loc, mag, $WTSD$, and $WTSK$), we applied t-tests in the same way as was done for the other models (Table 6.1, row labeled "Modeled Accent"). Results indicate that the groups differed significantly in the peak location of phrase-initial feet. We next employed an RBF kernel based SVM to classify PD vs. Control. Table 6.2 illustrates that the features (Set_2) extracted via the *GENIE Accent Curve* yielded the highest $F1$ score, accuracy, TN, and TP values of all methods.

In order to determine the significance of the classification result, we performed a randomization test in which the diagnostic status of the 20 participants was randomized 100 times and the SVM training and test procedures were applied to each randomization. Figure 6.3 shows the histogram of the randomized SVM results and the observed results; we display these histograms to show that the distributions resulting from randomization are well-behaved, lending credibility to this significance testing method. The histograms show that the observed results are far better than can

be expected by chance for the *GENIE Accent Curve*, with marginally significant results for the *Raw Accent Method*. (We used a randomization test because the assumptions underlying conventional statistical methods, such as Hotelling's T^2 test are unlikely to be met.)

6.2.2.3 Improving the Raw Accent Method using Frame Weighting

As described in Section 6.2.1.1, we applied linear interpolation between voiced areas to replace the unvoiced areas. This might causes some disadvantages for the *Raw Accent Method* since there may be regions that, while not fully voiceless, are nevertheless low in sonorous and thus may contribute minimally to perceived pitch and/or may include substantial segmental perturbations that are not modeled by accent curves. To address this, we compared results of the *Weighted Raw Accent Method* with the *Raw Accent Method* using the four features (loc, mag, $WTSD$, and $WTSK$). We applied the t-test on these four features (Table 6.1, last row). We found marginally significant results not only on phrase-initial feet but also on phrase-medial feet. We next employed an RBF kernel based SVM to classify PD vs. Control. The results are slightly more accurate (Table 6.2): The *Weighted Raw Accent Method* with the features (Set_2) improved the $F1$ score and accuracy 0.03 points and 1.5 percent compared to the *Raw Accent Method*. Yet, the results are still not as good as for the *GENIE Accent Curve*.

6.3 F_0 Dynamics in Clear and Conversational Speech

In perceptually difficult environments, speakers naturally and spontaneously tend to speak in a speaking style that will be perceived more easily and clearly by their targeted audience. Such a speaking style is referred to as clear speech (CLR). In contrast, the speaking style that speakers use to casually communicate with a normal listener in a quiet environment about an understood topic is referred to as conversational speech (CNV). Factors affecting the perception of speech and thus the invocation of a clear speech speaking style involve the presence of background noise, whether the listener has a hearing impairment, the age of the listener, or whether the listener is a native speaker of the language.

Previously, it has been shown that compared to CNV speech, CLR speech can be characterized by the following acoustic-prosodic features: a decrease in speaking rate, and an increase in the number of accented words, number of pauses and prosodic phrases, and range and mean of F_0 contour, [142]. Motivated by these acoustic-prosodic characteristics of CLR speech, researchers have attempted to find out how they can modify CNV speech to make it more intelligible like CLR speech. Liu and Zeng in [89] examined whether speech intelligibility improved by modifying the

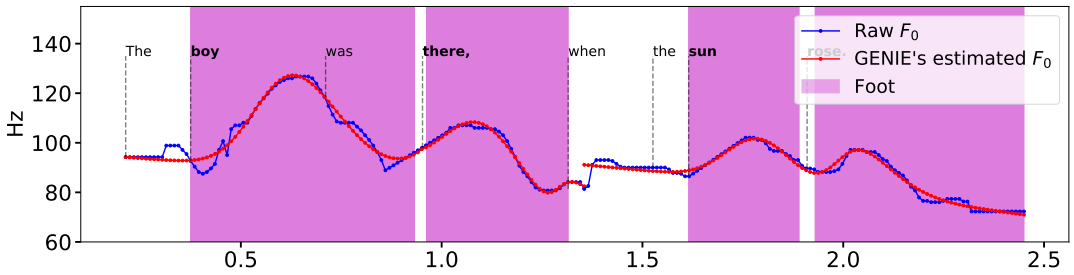
rate of speech in CNV speech or not. For the same sentences spoken in both CNV and CLR styles, authors used the pitch-synchronous overlap and add method to decrease the speaking rate in CNV speech by increasing silent gaps (pauses) between phonetic segments to match the duration of the CLR speech. They did not find statistically significant improvement in speech intelligibility. In series of articles, Kain and his colleagues [5, 65, 66, 64, 77, 109] examined the contribution of a variety of acoustic features to improve intelligibility of speech. In parallel recordings of CNV and CLR speech, they replace a combination of acoustic features of CNV speech with those extracted from CLR speech, to generate a synthesized CNV speech. Then using perceptual experiments, they examine the intelligibility of the unmodified and transformed speech. Results indicated that transformed CNV speech with acoustic-prosodic features (the pausing patterns, F_0, and energy were extracted from CLR speech) does not improve the speech intelligibility over unmodified CNV speech. These are both important and interesting findings, given that unmodified CNV and CLR speech differed in speech intelligibility that were characterized by acoustic-prosodic features. These results motivated us to further investigate which prosodic features between CLR and CNV speech are more relevant to these characteristics. If we find a prosodic feature that can differentiate CLR speech from CNV speech, it might be the feature that is responsible for improved intelligibility CLR speech over CNV speech. In our first experiment (section 6.3.3), we show that two common prosodic features (F_0 range and F_0 mean) at the utterance and phoneme level do not differentiate CLR from CNV speech. In the previous section, even though participants with Parkinson's Disease and healthy control participants did not differ in terms of F_0 range and F_0 mean, we found there are differences between F_0 dynamics of the groups. This led us to hypothesize that the prosodic characteristics of CLR speech are attributed to F_0 dynamics and are independent of speaking rate, F_0 range and F_0 mean.

The objective of this section is to show an increase in the number of prosodic units (number of feet) in CLR speech is attributed to F_0 dynamics and is independent of speaking rate, F_0 range and F_0 mean. To address this, we use the similar methodology as in the Chapter 4 (section 6.3.1). In Section 6.3.2, we describe details of two data sets used in both experiments. In the second experiment (section 6.3.4), we investigate F_0 dynamics differences that are due to different prosodic structure between CLR and CNV speech, to find out whether the increase in number of feet in CLR speech is independent of the F_0 range and duration of the utterance.

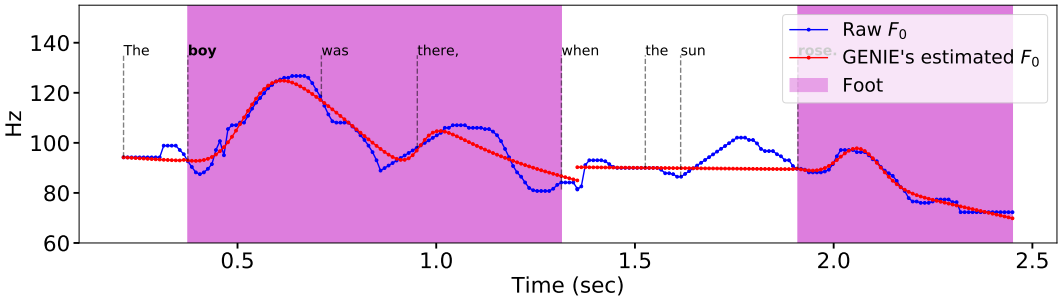
6.3.1 Using GENIE to Find the Best Foot Structure

In order to investigate whether CNV and CLR speaking styles can be differentiated by their F_0 differences due to different prosodic structures (different foot structure and number of feet) or not, we use a similar methodology as in Chapter 4. In Chapter 4, to determine the best phrase boundaries for a sentence, we generated a number of phrase boundaries for the sentence. We then used GENIE to find which variation resulted in the lowest error with respect to the model. In this section, we generate all possible foot structures of each sentence and use GENIE to find the best one.

Before explaining our method in details, we illustrate how different foot structures and the goodness-of-fit measure are related in an example. Figure 6.4 represents the F_0 decomposition of



(a) Example F_0 contour and GENIE's estimated F_0 contour for a sentence with foot boundaries marked with brackets "The [boy was][there,] when the[sun][rose]."



(b) Example F_0 contour and GENIE's estimated F_0 contour for a sentence with foot boundaries marked with brackets "The [boy was there,] when the sun [rose]."

Figure 6.4: Example F_0 contour and GENIE's estimated F_0 contour using two different foot structures. The raw F_0 values are represented by the blue dotted line and GENIE's estimated F_0 contour is represented by the red dotted line. Each magenta region represents one foot. In 6.4a, GENIE's estimated F_0 contour is much closer to the raw F_0 contour of the utterance than in 6.4b.

an utterance: "The boy was there, when the sun rose." into component curves. In this example, two different foot structures are examined by GENIE. Foot boundaries are being represented by the magenta regions in the figure and the stressed-accented-syllables are marked in bold. In 6.4a,

GENIE's estimated F_0 contour is much closer to the raw F_0 contour of the utterance than in 6.4b. This indicates that the given foot structure, in Figure 6.4a, is close to the actual foot structure of the utterance. Therefore, GENIE generates a more accurate estimation of the F_0 contour than when the given foot structure is far removed from the actual foot structure of the utterance (Figure 6.4b).

We use GENIE for this study as follows. For each sentence, the syllable stress labels are extracted from a pronunciation lexicon. In order to determine the accent labels, we consider all combinations of occurrence/non-occurrence of the accent label, only for the stressed syllables (since only stressed-accented syllable can signal the start of a new foot). Using the stress label, the generated accent labels, and the phrase boundaries label, we generate all possible foot structures. We call these foot structures for a given sentence *foot assignments*. For each foot assignment, we apply GENIE and then calculate in a Root Mean Square Error (RMSE) between the raw F_0 and GENIE's estimated F_0 contour. A decrease in RMSE should indicate a closer correspondence between the foot assignment and the actual foot structure of the utterance. Ideally, in comparison between the foot assignment relevant to the lowest RMSE in CLR and CNV styles, we want to show the foot count in the CLR foot assignment is higher than in CNV.

A problem with above methodology is that there is an overfitting issue, which is RMSE continues to decrease as number of foot increases in foot assignments. Therefore using the lowest RMSE as a comparison measure may not result in differentiating CLR style from CNV style based on their foot structure. To address this issue, after calculating all the RMSEs that each has a specific foot assignment, we sort these RMSEs based on foot count in their foot assignment. Then we compare the results of the RMSE in both CLR and CNV styles with respect to the foot counts.

6.3.2 Speech Corpus

Following Kain and et al [66], we use two types of sentence sets in the two experiments in section 6.3.3 and section 6.3.4. In the first set, 70 phonetically balanced sentences were extracted from the IEEE Harvard Psychoacoustic Sentence Set (database H) [133]. These sentences are syntactically and semantically normal (e.g., Cars and busses stalled in snow drifts). In the second set, 70 sentences were generated by randomly exchanging words in the first set (e.g., Slide the cars through a stray blue lake). Therefore, in the second set, sentences are syntactically correct but semantically anomalous (database A) [171].

One male native speaker of American English was asked to record 140 sentences (from both data sets) in two speaking styles (CNV and CLR) to create a total of 280 recordings. When

recording CNV speech, he was instructed to speak in the way that he used to communicate in his daily life. When recording CLR speech, he was instructed to speak clearly as he would when communicating with hearing-impaired listeners.

Earlier in section 6.3, we discussed several of Kain's research projects; one of them investigated the relationship between different acoustic features and the intelligibility properties of CLR speech for the speaker in both data sets [66]. The results indicated that for database H, using acoustic-phonetic features can improve the speech intelligibility of the synthesized CNV speech over original CNV, while using acoustic-prosodic features did not improve the speech intelligibility. For database A, considering different combinations of acoustic-phonetic and acoustic-prosodic features did not result in any significant intelligibility improvements.

6.3.3 Experiment 1: Differences in F_0 mean and range at the utterance and phoneme levels

In this experiment, we focus on answering this question: Can we differentiate CLR speaking style from CNV speaking style by using two common prosodic features (F_0 range and F_0 mean) at the utterance and phoneme level? We compared the F_0 mean and range differences between CLR and CNV speaking styles using three scales: Hertz (Hz), Logarithmic (Log), and normalized, and at two levels: the utterance level and the phoneme level. The results from the Hertz scale did not differ from the Logarithmic scale, therefore we ignored the results of the Hertz scale in this study. The normalized scale was created by subtracting the F_0 minimum of the whole utterance from the raw F_0 values and dividing by the F_0 range of the whole utterance to minimize the effect of pitch range. At the utterance level, the mean F_0 of the whole utterance was calculated by taking the average over all voiced regions, while at the phoneme level, the average was calculated over the F_0 values of vowels and diphthongs.

We extracted 560 F_0 mean values (140 utterances and 4 conditions: two scales in two levels) for both CNV and CLR. Then, we performed a paired t-test to determine whether the F_0 mean values of the two populations differed significantly. We also performed a paired t-test for the F_0 range values. The results of these p-values are shown in Table 6.3. As can be seen, at the utterance level, we found no significant difference between CLR and CNV in terms of the F_0-mean (except for database A on the normalized scale). The effect of the F_0-range was minimized on the normalized scale. At the phoneme level, CLR and CNV styles are significantly different in terms of F_0 mean and F_0 range for database A. For database A, the F_0-mean was not significantly different on the normalized scale.

	F_0 mean				F_0 range			
	Utterance level		Phoneme level		Utterance level		Phoneme level	
	Log	Normalized	Log	Normalized	Log	Normalized	Log	Normalized
Data A	–	**	***	***	**	–	***	***
Data H	–	–	***	–	***	–	***	***

– $p > 0.05$ * $p < 0.05$ ** $p < 0.01$ *** $p <$ 1.0e-10

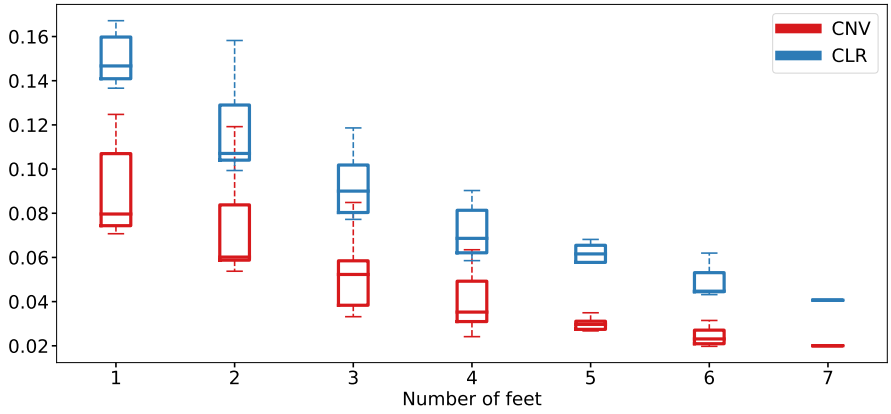
Table 6.3: Results of paired t-test between CLR and CNV speech in terms of F_0 mean and F_0 range. Comparisons were made in six conditions by considering two F_0 scales (Logarithmic (Log), and normalized) and two levels (utterance and phoneme).

6.3.4 Experiment 2: F_0 dynamic differences due to different prosody structures

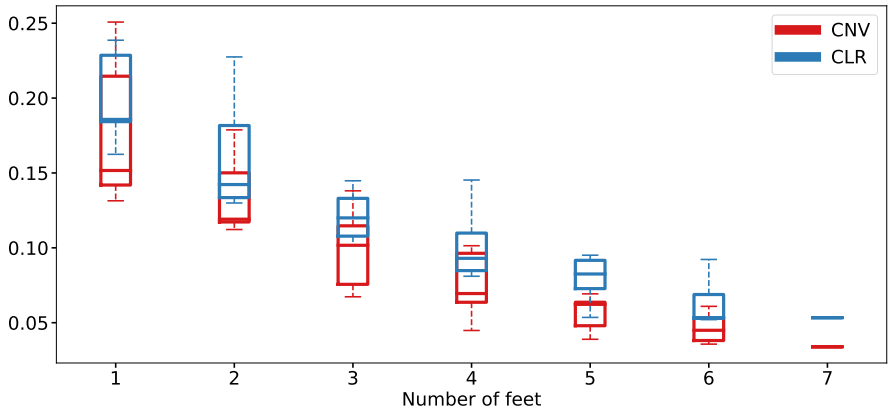
In Section 6.3.1, we described how we can use the goodness of fit of the implementation of GENIE to show that more feet are needed in CLR style than in CNV style. In other words, speakers who speak in a CLR speaking style emphasize more words than in CNV style. In this section, we describe how we use the proposed method on both data sets A and H.

We first illustrate the process for this sentence "The fish twisted and turned on the bent hook." First, we extract the syllable stress labels and the phrase boundaries from Festival. Then, we consider all combinations of occurrence/non-occurrence of an accent label, only for the stressed syllables. Using these labels, we generate all possible foot assignments. For each foot assignment, we apply GENIE and then calculate RMSE between the raw F_0 and GENIE's estimated F_0 contour. Then we sort the RMSEs based on foot count in their foot assignment in both CLR and CNV styles. Figures 6.5a shows the distribution of the RMSEs in CLR (blue box-plots) and CNV(red box-plots) styles for different foot counts. As we can see, the blue box-plot is above the red box-plot, across all numbers of feet. This implies CLR speech requires more feet than CNV in foot assignment to achieve a lower RMSE. For instance, in Figure 6.5a, to archive a fit with 0.06 RMSE (on Log scale) CLR style needs five or four feet while CNV style can be modeled with two feet with the same RMSE.

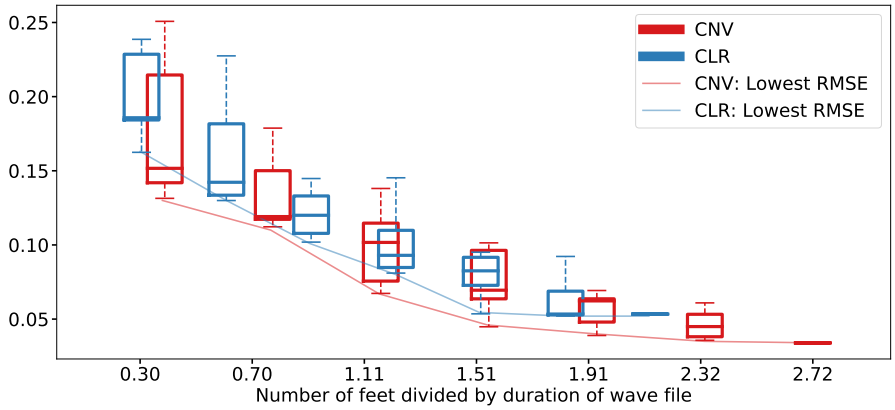
From the above, can we conclude that the RMSE difference between the two styles is because of the foot structure differences? The answer is not yet. In the previous experiment, we showed that the F_0 range of utterances in CLR style is significantly higher than CNV style. Therefore, there is a chance that the RMSE difference between the two styles is caused by F_0 range differences and not by foot structure difference. Thus we want to minimize the F_0-range difference: we apply our process to the normalized F_0 scale as well. As can be seen in Figure 6.5b, even though the normalized scale decreased the gap between the distributions of the RMSEs in CLR and CNV, still the mean of the distribution of the RMSEs in CLR is higher than the mean of the distribution



(a) Log scale



(b) Normalized scale



(c) Normalized scale, adjusted duration

Figure 6.5: Distribution of the RMSEs in CLR (blue box-plots) and CNV(red box-plots) styles for different foot counts for this sentence "The fish twisted and turned on the bent hook."

of the RMSEs in CNV, across all numbers of feet. This implies more feet are needed in CLR style than in CNV style, and this increase in foot count is independent of the F_0-range of the utterance. For instance, in Figure 6.5b, to archive a fit with 0.06 RMSE (on normalized scale) CLR style needs five feet while CNV style can be modeled with four feet with the same RMSE.

From the above, can we conclude that the RMSE difference between the two styles is because of the foot structure differences? The answer still is not yet. It has been noted in many studies [142, 104, 76] that speakers usually significantly reduce their speaking rate in the CLR speaking style. It also has been shown that durations increase when more words are accented – durations are longer for phonemes in accented words regardless of whether they are stressed or not [161, 165, 16]. Therefore, there also is a chance that the above results are correlated with speaking rate. The foot count increase in the CLR style might be attributed to utterance duration and not a characteristic of the CLR style. In order to minimize the effect of duration, we repeat the above experiment except we sort RMSEs based on the foot count divided by the utterance duration. In Figures 6.5c, as expected, the results of the CLR style are compressed in x-axis and shifted to the left side due to longer utterance durations compared to CNV speech, but still the lowest RMSEs in CLR (blue line) is higher than the lowest RMSEs in CNV (red line), across all numbers of feet that are adjusted by the utterance duration. This implies not only that more feet are present in the CLR style than in the CNV style, but also that this increase of the number of feet is independent of speaking rate and the F_0-range of the utterance. For instance, in Figure 6.5c, to archive a fit with 0.06 RMSE (on normalized scale) CLR style needs 1.5 feet while CNV style can be modeled with 1.2 feet with the same RMSE. To measure the overall difference between the two styles, we calculated the area under the lowest RMSEs curve in CLR ($S_{CLR} = 0.094$) with in CNV ($S_{CNV} = 0.074$) and used their ratio as the comparison measurement $Ratio = S_{CLR}/S_{CNV} = 1.28$. The higher ratio the strongest evidence that the two curves are separable. In Figure 6.6 the S_{CLR} and S_{CNV} are differentiated with two colors red and blue, respectively. Each datapoint represents an individual RMSE with a specific foot assignment. Solid lines represent the lowest value of RMSE in CLR and CNV.

From the above, can we conclude that the RMSE difference between the two styles is because of the foot structure differences? Yes, we can conclude that CLR speech can be characterized by an increase in the number of feet for the sentence "The fish twisted and turned on the bent hook." Also by minimizing the effect of utterance F_0-range and duration, we showed that this characteristic is caused by F_0-dynamic changes, and it is independent of F_0-range and speaking rate. We repeat this process for each sentence spoken in the CNV and CLR styles in both data sets H and A.

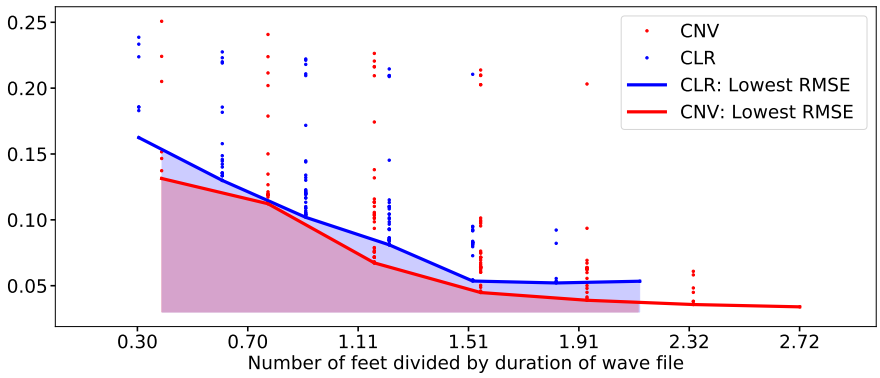


Figure 6.6: Each dot represents an individual RMSE with respect to GENE for this sentence "The fish twisted and turned on the bent hook." with a specific foot assignment. These RMSEs are sorted based on foot count divided by the utterance duration. Solid lines represent the lowest value of RMSE. CLR and CNV data are differentiated with two colors red and blue, respectively.

Figure 6.7 shows the distribution of the $Ratio = S_{CLR}/S_{CNV}$ for both data sets in three conditions: log, normalized, and normalized scale adjusted duration (Norm_Dur). In Log and normalized scale the mean of the ratio distribution is higher than 1.0 (black dashed line) for both data sets. It is interesting to note that the effect of minimizing F_0-range is more aggressive on database H compared to database A. The same conclusion can be made for the effect of minimizing duration: is more aggressive on database H compared to database A.

For database H, when minimizing the effect of utterance F_0-range and duration in normalized scale adjusted duration the mean of the ratio distribution is higher than 1.0. This reveals not only that more feet are present in the CLR style than in the CNV style, but also that this increase of the number of feet is independent of speaking rate and the F_0-range of the utterance. These results are consistent with those of Krause [76] that the speaking rate alone is not fully responsible for differentiating the two styles. Similar to Smiljanić et al [104, 142] we show that CLR speech can be characterized by enhancing prosodic structure, but also we show that this characteristic is caused by F_0-dynamic changes and not only by speaking rate. For database A, the mean of the ratio distribution is slightly above one. Therefore, the results for database A are not stronger than database H. In order to demonstrate whether these RMSE differences are significant or not for both data sets, we took the following steps. For each sentence, we calculated the $Ratio = S_{CLR}/S_{CNV}$. We used a one-sample one-tailed t-test to perform statistical comparisons of these values.[2] The Table 6.4 shows RMSE values are significantly higher in CLR style for all

[2]We use a one-tailed t-test, because we want to show that the number of feet is different in CLR and CNV styles, but also to show that in CLR style more feet are needed.

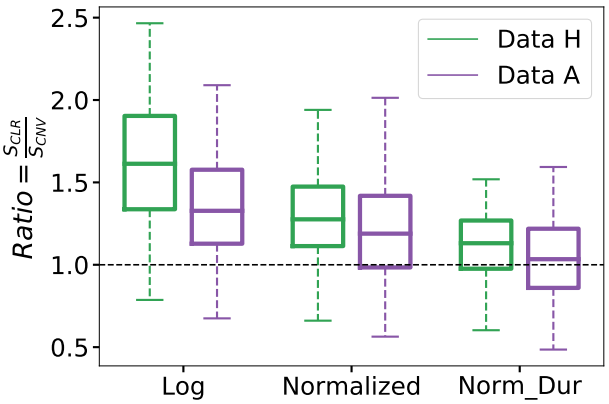


Figure 6.7: Side-by-side box plots showing the distribution of the ratio for both data in six conditions. The term "Norm_dur" stands for normalized scale, adjusted duration.

	$Ratio = S_{CLR}/S_{CNV}$		
	Log scale	Normalized scale	Normalized scale, adjusted duration
Data A	**	**	–
Data H	***	***	**

– p > 0.05 * p < 0.05 ** p < 0.01 *** p < 1.0e-10

Table 6.4: Results of one-sample one-tailed t-test between CLR and CNV speech in terms of foot count adjusted by utterance duration.

comparisons for database H. The sentences in database H are syntactically and semantically normal; now we focus on database A that consists of syntactically correct but semantically anomalous sentences.

After minimizing the effect of utterance duration in our experiment we did not find significant differences between the two styles on the normalized scale adjusted duration in database A (in Table 6.4, and also in Figure 6.7, the purple box plot under the Norm_Dur condition), while in Log and normalized scales the result of this condition was significant. This suggests that for a semantically anomalous sentence, the speaker uses similar prosodic structure (similar foot structure and similar number of feet) in both styles, and only talks slower in CLR style. These results are consistent with those of Kain [66]. For database A, they did not find any significant difference between CLR and CNV styles, and they suggested that this lack of significance may be caused by difficulty to understand semantically anomalous sentences.

6.4 Intonation Based Classifier

Earlier in this chapter, we showed that F_0 dynamic differences between two speaker groups can be used to differentiate one from another. Also, we showed that GENIE was better in capturing these F_0 dynamic differences than the baseline methods. Motivated by these findings, we propose a new intonation based approach to classify at least two speaker groups. Our assumption is that two speaker groups can be differentiated through their F_0 dynamic differences. We apply this speaker classification framework to intonation-based classification of dialects of English. A dialect of a language is defined as "a pattern of pronunciation and/or vocabulary of a language used by the community of native speakers belonging to some geographical region" [84]. Dialect classification is a special case of speaker (or speaker state) classification.[3]

During training, we apply the DRIFT training phase that is based on GENIE (described in Section 5.2.2) to F_0 contours in each dialect group. This allow us to produce a structured inventory of GENIE's component accent and phrase curves parameters for each dialect. During testing, for a given test F_0 contour and dialect, we determine which dialect's inventory of phrase and accent curves best account for a sentence by using Non-negative Matrix Factorization (NMF) and a sparsity measure.

In Section 6.4.1, we propose a group classifier that combines the NMF algorithm and a sparsity measure. Then in Section 6.4.2 we show how the DRIFT's training phase and the new classifier can combine to classify dialect groups using their F_0 dynamic differences. In Section 6.4.3, we compare out proposed classifier against DRIFT's testing phase on an F_0 reconstructing task and find out that the proposed method reduces the RMSE by 50%. Finally in Section 6.4.4, we compare our proposed classifier against a baseline on a dialect classification task.

6.4.1 Group Classification Using NMF and a Sparsity Measure

NMF is a popular algorithm for feature extraction in machine learning, computer vision, and signal processing. One reason for this popularity is that NMF naturally favors sparse representations. This inherit ability of NMF lead us to combine it with a sparsity measure in the context of group classification. In rest of the section, we first review the NMF algorithm. We then discuss how we combine NMF with a sparsity measure. Finally, we discuss our training and test procedures of the proposed classifier.

[3]In this section, we did not distinguish among dialects within a country. For example American eastern and western dialects are considered as American dialect.

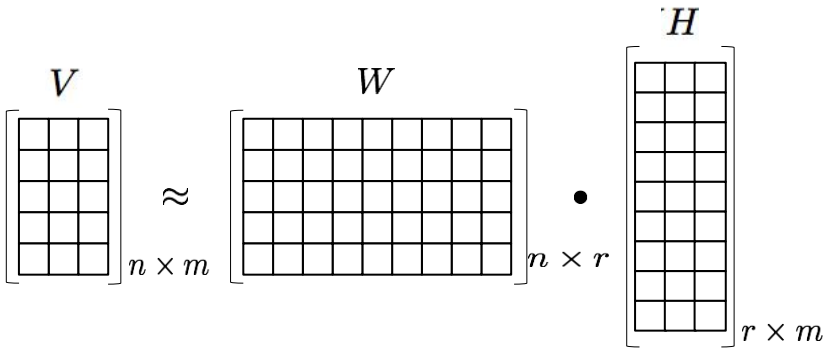


Figure 6.8: NMF schema

6.4.1.1 Non-negative Matrix Factorization

NMF is a well-known source decomposition algorithm [19]. Given a non-negative matrix V, this algorithm [82] factorizes the matrix $V_{n\times m}$ (n and m are the sample size and number of data samples, respectively) into two non-negative matrices $W_{n\times r}$ and $H_{r\times m}$ (r is equal to number of chosen basis vectors, Figure 6.8). This non-negativity has two advantages. First, it makes the matrices W and H easier to understand. Second, it is suitable for cases that non-negativity is inherent to the data (e.g., intonation analysis).

$$V \approx \hat{V} = WH \qquad W, H \geqslant 0 \tag{6.4}$$

The matrix W (we also refer to it as a *dictionary*) and the matrix H (we also refer to it as a *weight* vector) can be considered as a collection of basis vectors and a stack of weights corresponding to each basis vector, respectively. If the basis in W were optimized to approximate V linearly, then each column in Equation 6.4 can be rewritten as $v \approx \hat{v} = Wh$. This equation suggests that each data sample $v_{n\times 1}$ in V is a linear approximation of columns of W weighted by $h_{r\times 1}$. It has been shown that a good approximation of v is guaranteed if W consists of the hidden structure of the data [82]. The standard NMF algorithm consists of the following steps: initializing the matrices W and H, and updating them in each iteration until there is no significant improvement in the cost function (Equation 6.5).

$$Cost(v, \hat{v}) = \sum_{k=1}^{n} (v_k log \frac{v_k}{\hat{v}_k} - v_k + \hat{v}_k) \tag{6.5}$$

Randomly generating the matrix W and the vector h is a simple way of initializing them; however, advanced initialization strategies can be developed for improving fitness of the NMF

Algorithm 6.1 Non-negative Matrix Factorization (NMF) algorithm

INPUT:

v ▷ A given non negative vector

1: $W \leftarrow$ Dictionary

2: $W \leftarrow$ Normalizing W based on v ▷ optional step

3: $h \leftarrow$ Random vector (r) ▷ r is equal to number of bases in dictionary

4: **while** there is improvement in $cost(v, \hat{v})$ **do**

5: **for** α in h **do**

6: $h_\alpha \leftarrow h_\alpha \frac{\sum_i \frac{W_{i\alpha} v_i}{W_{i\alpha} h}}{\sum_k W_{k\alpha}}$

7: $\hat{v} \leftarrow Wh$

algorithm with fewer iterations or for extracting sparse features.

NMF does not provide a straightforward control over W in its update process. In some use cases, one might need to force the basis vectors to follow certain shapes or patterns based on domain knowledge. Imposing the constraints in the NMF algorithm is not straightforward. A modification of the NMF algorithm is to keep W constant and only update h. The W matrix can be filled with any method, which suits the use case. Also, updating only one factor (vector h) instead of two factors (matrix W, and vector h) reduces the algorithm complexity. The steps of the algorithm are shown in Algorithm 6.1.

6.4.1.2 Combining NMF with a Sparsity Measure

NMF has an inherent linear property; It makes it possible to approximate each column in V by columns of W weighted by rows in H. This property makes NMF a suitable algorithm for sparse coding that models a signal (e.g., column v in V) as linear combination of a few basis vectors [53, 95, 96, 51]. More specifically, the approximation of v ($\approx \hat{v} = Wh$) is achievable by finding few columns in W associated with high weight that minimize error between v and $\hat{v}$. Therefore if W consists of the hidden structure of v, then inspection of learned h may show that it is sparse. This motivated us to use the NMF algorithm for a group classification task; having a sparse learned h suggests that v is quite similar to at least one basis vector in a given group, and thus – by analogy to nearest-neighbor clustering – may in fact be a member of that group.

We propose to use the Gini coefficient to measure the sparseness of vector h^i. Although the Gini coefficient is mainly used for measuring income inequality, it also has been used as a measure of sparsity [130] in other tasks. It has been proven that the Gini coefficient was able to satisfy all reasonable sparsity criteria (like cloning, which does not affect sparsity) [58]. The Gini coefficient, computed in Equation 6.6, has a value between 0 and 1, with zero representing total equality between all values (i.e., high density), and 1 representing highest inequality (i.e., high sparsity).

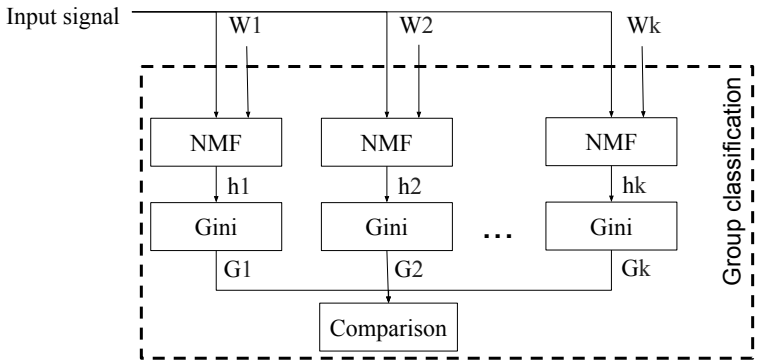


Figure 6.9: Proposed test schema of the speaker group classification using NMF and Gini

$$G_i = \frac{\sum_{p=1}^{r} \sum_{q=1}^{r} |h_p^i - h_q^i|}{2r \sum_{p=1}^{r} h_p^i} \tag{6.6}$$

6.4.1.3 Training and test procedures

To use NMF with the Gini sparsity measure as a classifier, we propose the following. During training, a dictionary is built for each group, i, which contains basis "templates" that represented the individual training data items. A user can fill these dictionaries by any method that suits the users task, such as chunks of real signal, syntacticly generated templates, or parametrically learned templates from training data. At test time as shown in Figure 6.9, for each given input, the NMF algorithm is used to decompose the input signal using the respective dictionaries of each group (W^i), resulting in weight vectors (h^i). Then, the Gini coefficients are computed for each h^i vector. At the end for classification, the group with the largest Gini coefficient is chosen, which is the weight vector that is the sparsest.

6.4.2 Using NMF and Sparsity for Intonation Based Dialect Classifier

In this section, we show how the classifier proposed in the previous section can combine with the DRIFT's training phase for dialect classification.

During training we follow the same methodology that we used in Section 5.2.2.2 for DRIFT's training phase (this step does not involve NMF algorithm). For each utterance in each dialect group we do the following. First, we run Festival to generate accent labels, syllable labels, and phrase boundaries. Second, we derive the foot structure. Third, we apply GENIE to compute the component curves. Fourth, we store *component curve parameter vectors* where each vector is associated with prosodic labels and component curve parameters. Specifically, phrase curve vectors

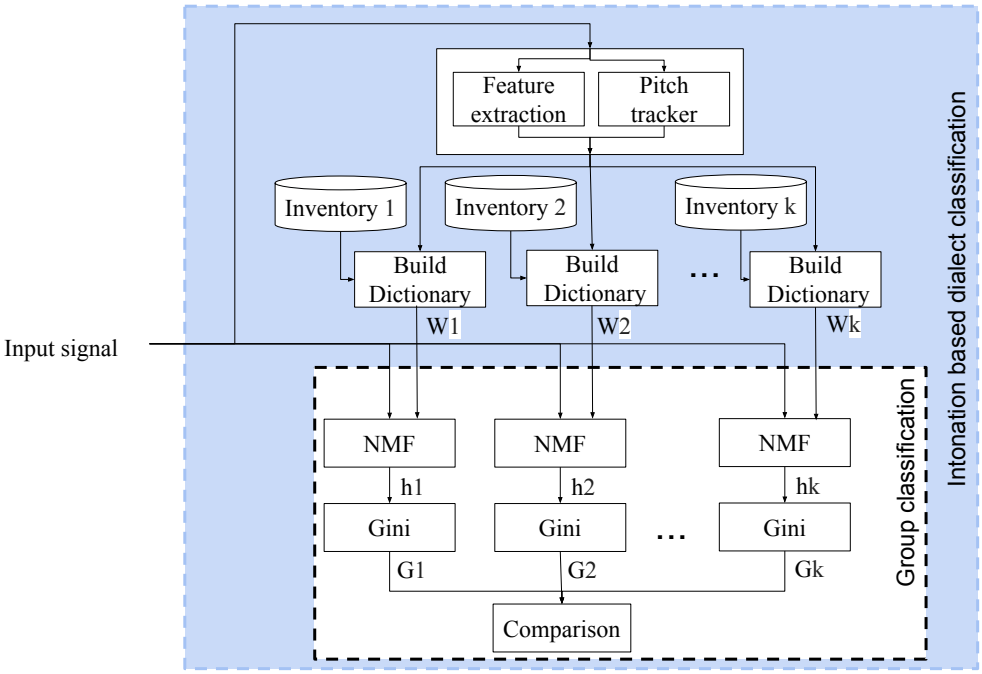


Figure 6.10: Proposed test schema of combination of the DRIFT method and the proposed speaker group classification.

are labeled in terms of the number of feet in the phrase and accent curve vectors in terms of position of the foot in the phrase and the number of syllables in the foot. Therefore, for each dialect, we produce a structured inventory of GENIE's component accent and phrase curves parameters.

During testing we need to take into account that test speech signals vary in duration and number of feet. Therefore, we have to build a dictionary with templates that match the duration of the input F_0 contour, and whose prosodic labels match those of the test signal. First, we run Festival to generate accent labels, syllable labels, and phrase boundaries. Second, we derive the foot structure and determine prosodic labels: accent type, foot position in the phrase, and number of syllables for each foot. Third, we use these prosodic labels and inventory from dialect group i, to construct an matrix W^i whose columns correspond to vectors in the inventory for dialect group i that have the same prosodic labels as the input signal, and consist of generated accent and phrase curves from these parameter vectors (we discuss details of this step next). After generating W for each group, we follow the steps described in Section 6.4.1 to classify the input signal (see Figure 6.10).

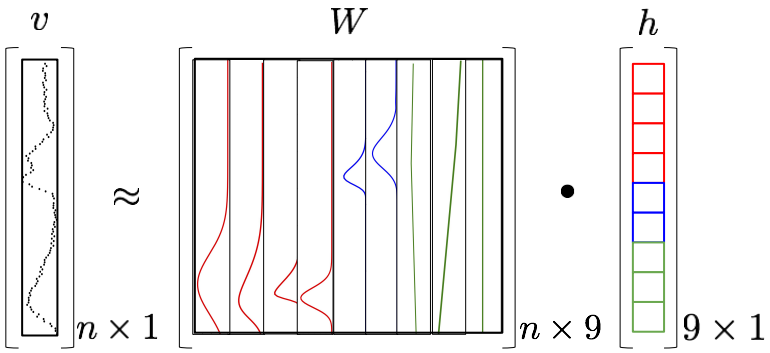


Figure 6.11: Proposed NMF schema. Vector v is the F_0 contour of a phrase unit with two feet.

6.4.2.1 Building the Dictionary

As mentioned above, the size of both dictionary W and weight vector h are changed based on the F_0 contour length and feet numbers of each test sample. Please note that we use the same inventories that are created using the DRIFT method during training, but different size of dictionaries. For example, for a test sample with three feet and a test sample with two feet, the total number of generated templates are different for each dialect. The j^{th} test sample and i^{th} dialect is represented as W^{ij}, and h^{ij}, but for ease of notation, we drop the index j, unless explicitly specified.

Algorithm 6.2 shows the required steps to build the dictionary (W), given a vector v along with its prosodic labels l and inventory of dialect group i ($Inventory_i$). We now give an example to illustrate how the algorithm works. Referring to Figure 6.11, we assume that the given vector v has two feet and there are nine component curve parameters vectors stored in $Inventory_i$ (Algorithm 6.2, input), which are extracted by applying DRIFT on the training F_0 contours. Four accent curves that appear only in the first foot within a phrase unit, two accent curves that appear only in the last foot of a phrase unit, and finally three phrase curves. Since in $Inventory_i$, we only stored the normalized parameters, we need to regenerate the component curves w.r.t the given vector v and its prosodic labels l. For each foot in v, we generate all relevant accent curves using $Inventory_i$ (Algorithm 6.2, steps 4 to 8). For example, for the first foot, four accent curves with length n are generated (red curves in Figure 6.11). Similarly, for the second foot, two accent curves with length n are generated (blue curves in Figure 6.11). Note that regenerating a stored accent curve in a length different from its original length does not affect shape and peak location of the generated accent curve. Finally (Algorithm 6.2, steps 9 to 11), we generate all of the relevant phrase curves using $Inventory_i$, and store them in matrix W^i (green curves in Figure 6.11).

Algorithm 6.2 Building the dictionary

INPUT:
$Inventory_i$ ▷ Inventory of component curves parameters of group i
v ▷ Input F_0 contour
l ▷ Prosodic labels of v
OUTPUT:
W^i ▷ Dictionary of group i

$n \leftarrow$ Length(v)
$W^i \leftarrow$ Empty matrix (n, r)
$k \leftarrow 0$
$Feet \leftarrow$ Extract all feet of v using l
for f in $Feet$ **do**
    $A \leftarrow$ Generate curves ($Inventory_i$, f, l, 'accent', length=n)
    for acc in A **do**
        $W^i[:, k] \leftarrow acc$, $k \leftarrow k+1$
$P \leftarrow$ Generate curves ($Inventory_i$, v, l, 'phrase', length=n)
for phr in P **do**
    $W^i[:, k] \leftarrow phr$, $k \leftarrow k+1$

6.4.3 Experiment 1: Validating the Use of NMF

As we discussed in Section 6.4.1, NMF favors a sparse representation of the input signal if W consists of the hidden structure of the input data. In this section, we investigate how well NMF can generate an approximation of input F_0 contour by finding few columns in W associated with high weight. In order to do so, we use NMF to decompose an input F_0 contour using pre-computed templates in matrix W, which are generated using DRIFT's training phase. We choose DRIFT's training phase to fill the matrix W (as in 6.4.2.1) since we have shown in the last few chapters that GENIE's component curves can capture underling intonational patterns.

We compare our proposed method against DRIFT's testing phase on an F_0 reconstruction task. We choose this task to show that : 1) our proposed method results in a better fit than DRIFT's testing phase, 2) our proposed method favors a sparse representation, and 3) the retrieved template curves follow the principle of GENIE's estimated component curves.

6.4.3.1 Corpus

For this experiment, we use the CMU Arctic database [75] as in Section 5.2.4.1. We use speaker SLT. Utterances were recorded in a sound proof room. This corpus contains 1132 utterances; we randomly chose 90% of the utterances for training and the remaining 10% for testing.

Methods	RMSE (Hz)	Sparseness
Proposed	5.92	0.82
DRIFT	11.12	–

Table 6.5: Comparison between two methods: the proposed method and DRIFT on CMU Arctic data. The second column shows the average RMSE between the predicted F_0 contour with each method and the original F_0 contour. The third column shows the average sparseness of vector h using the Gini coefficient .

6.4.3.2 Baseline

We used the DRIFT method from section . During training we generate an inventory of estimated accent and phrase curves parameters.

During testing, given an unseen F_0 contour of an utterance from the test set, we run Festival to generate the required textual information and with that information we determine the prosodic event units (phrase and foot). For each prosodic event textual features are extracted from text data. We search for stored, fitted accent curves associated with feet that optimally match to-be-synthesized feet in the feature space, while minimizing differences between successive accent curve heights. Similar searches are done to find suitable phrase curve parameters in the phrase inventory.

6.4.3.3 Using NMF as a F_0 contour generator

For using NMF as a F_0 contour generator, we follow the same methodology as discussed in Section 6.4.2. Note that the training part is identical to DRIFT's training phase. During test, after deriving the weight vector h, instead of using it for classification, we use it to retrieve component curves from the matrix W associated with weight higher than 0.9. Then, the estimated F_0 contour is calculated by adding the retrieved component curves.

6.4.3.4 Results

We used 10% of the CMU Arctic database for test purposes. We reconstruct each F_0 contour using two methods: the proposed method and the baseline method.

As shown in Table 6.5, our proposed method reconstructs the F_0 contours better than DRIFT in terms of average RMSE. Our proposed method achieved a fitting error of about half the magnitude as the DRIFT method. Then we calculate the sparseness of the weight vector h using Gini for each utterance and compute the average, resulting in a score of 0.82 (last column in Table 6.5). This high sparseness suggests that our proposed method chooses few templates to estimate an input F_0.

Going further, we investigate how much our proposed method's retrieved component curves are correlated to GENIE's estimated component curves. For each test sample, after we derive the foot structure, we do the followings. First, we apply GENIE to compute the component accent curve (A_i) for each foot f_i and the phrase curve (P). Second for each f_i, we consider all accent curves from the matrix W that are associated with f_i and have a weight higher than 0.9. Third, we add up those curves together to produce an accent curve ($A_i^{'}$). Fourth, we retrieve all phrase curves from the matrix W associated with weight higher than 0.9 and add them together to produce a phrase curve ($P^{'}$). Fifth, we calculate the correlation between GENIE's component curves (A_i and P) and our proposed method's component curves ($A_i^{'}$ and $P^{'}$). This gives us an average of 0.81. This high correlation suggests that the component curves estimated using the proposed method have very similar patterns compared to their corresponding curves using GENIE.

These results show that our proposed method is able to estimate F_0 contours better the DRIFT method.

6.4.4 Experiment 2: Pairwise Dialect Classification

The purpose of this experiment is to distinguish between speaker groups, which are speaking different dialects, using the proposed method. We report the results of pairwise comparison individually.

6.4.4.1 Corpus

There are two commonly used corpora for dialect classification. The National Institute for Standard in Technology (NIST) 2008 Speaker Recognition Evaluation Series (SRE; [97]) database consists of 12 languages with 14 dialects. The NIST 2008 SRE a suitable database for foreign accent recognition and language recognition studies. However, there are only two dialects for English (General American and Indian English) and it was unspecified whether the speakers were native or not, which makes it less suitable for dialect classification.

VoxForge [47] is a publicly available speech corpus[4] and consists of recordings in 18 languages and multiple dialects for some of the languages. Furthermore, each speaker is identified as a native or non-native speaker. Unfortunately, as we discussed in Section 6.1, it has been shown that applying accent classification to only one second of silence from utterances belonging to the VoxForge database without using any speech information can already reach a high accuracy [12]. This suggests that since VoxForge did not have control over the recording conditions (most participants used their own headsets), the quality and channel properties might differ between speaker groups,

[4]http://voxforge.org/

		Train data		Test data	
		#Speakers	#Feet	#Speakers	#Feet
Dialect groups	American	712	29959	73	3048
	Australian	224	9595	25	1016
	British	607	23878	69	2744
	Canadian	278	10870	31	1240
	Indian	206	8636	23	785
	Total	2027	82938	221	8833

Table 6.6: The total number of speakers and feet in each dialect group for train and test data.

which may cause the classification algorithms to learn channel characteristics. In our study since no spectra-related features are used, this should not affect our results – in fact, this makes the case for using just prosodic information.

We used a total of 2248 male English speakers from the VoxForge corpus with five different dialects: American (Am), Australian (Au), British (Br), Canadian (Ca), and Indian (In). In Table 6.6, we have summarized the details of the corpus in terms of the number of speakers, and the number of feet per group. There is no overlap between train and test speakers. We labeled the database with the phoneme transcription via Kaldi [127] trained on the Librispeech corpus [118].

Since acoustic features extracted from the VoxForge corpus are not reliable for classification, it is our concern to check the reliability of prosodic features (F_0 contours). After extracting the F_0 contours, we used the robust standard deviation method to filter out corrupted F_0 contours. We removed a sample if any point of its F_0 contours (voiced segments) was outside of its F_0 mean$\pm$20% of its F_0 standard deviation.

6.4.4.2 Baseline

We need to select a baseline system that uses intonational features but not spectra-related features as we discussed above.

We first tried to used a number of simple F_0 features: mean, standard deviation, and skewness across the utterance. We calculated these features from all the F_0 contours in the training data. We then trained various classifiers (e.g., SVM and logistic regression) on these features and used them for pairwise classification. The resulting classifiers performed at chance level, showing that these commonly used features were not effective for distinguishing between dialects; we attribute this to the failure of these features to capture local F_0 dynamic changes. Generally, there are two ways to address this issue. The first one is capturing underling intonational patterns, as our proposed method does. The second one is to use features that represents local F_0 dynamics better

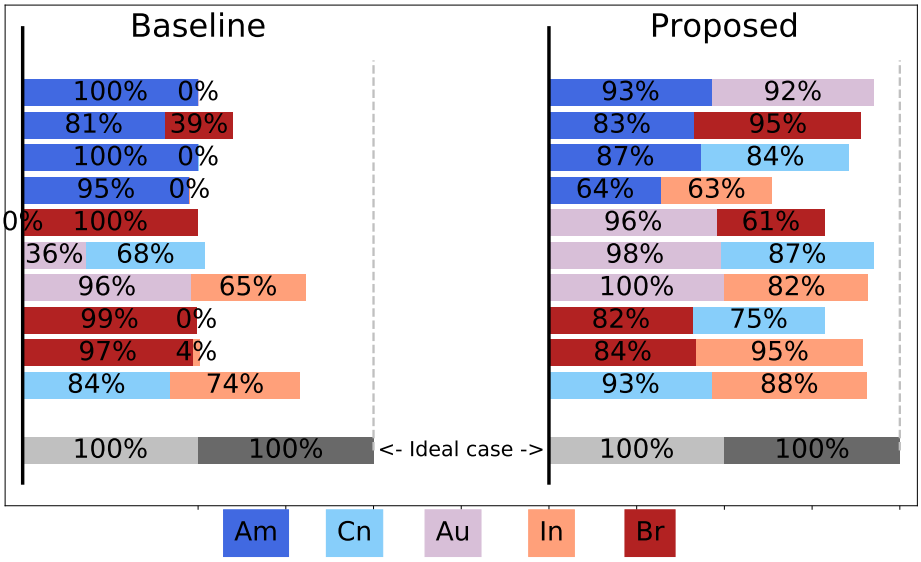


Figure 6.12: Pairwise accuracy plot. Each row shows side-by-side the average detection accuracy per speaker for both methods for a dialect pair.

than the simple F_0 features. Although no one have done this for dialect classification, Ma et al. [94] used such features for emotion recognition. Hence, we are using it as our baseline.

In Ma's method, for a pair of emotions, logistic bayesian regression is used to learn a binary classifier. In the original article, a combination of spectral and prosodic features from the INTERSPEECH 2010 feature set was used. To make this method comparable with our proposed method, we only used prosody related features from the INTERSPEECH 2010 feature set.

6.4.4.3 Results

For each pair of dialect, we trained two classifiers using both the baseline and our method with the same training data. Then we test the two methods using the same test data. For each speaker in test data, we used all of their utterances to decide which dialect the speaker belongs.

Figure 6.12 shows the pairwise comparison results for all ten dialect pairs between our classifier and baseline method. Each row is an individual group pair with the baseline results displayed in the left side and our classifier's results in the right side. The average accuracy of correct detections per speaker for each pair is reported on each row. From the figure it is easy to see that our classifier does much better than the baseline. We achieved an average accuracy of 85%, while the baseline achieved an average accuracy of 55% which is almost at chance. These finding are interesting in the following manners. 1) Even though the features set used in the baseline consists of a variety of prosodic features (e.g., jitter and shimmer) that are meant to represent the F_0 dynamics in both

short-term and long-term intervals, this large features set is ineffective in distinguishing dialects. This ineffectiveness indicates the difficulty of the dialect classification task. 2) Achieving high accuracy using our classifier validates our assumption of using NMF with sparsity measure. 3) Since W consists of templates that are based on GENIE's component curves, we can conclude that GENIE can capture F_0 dynamic differences between different English dialects. 4) English dialects can be distinguished by their F_0 dynamic differences. These results are consistent with our finding in clear vs. conversational and individuals with dysarthria vs. neurotypical individuals studies. There are probably many other classification tasks that can make use of F_0 dynamic differences.

Looking at individual pairs, our classifier performs worst for American English vs. Indian English, 64% of American speakers were correctly identified and 63% for the Indian speakers. One explanation is that Indian English speakers have shown frequent use of continuation rises or question rises in statements [103] instead of rise-fall patterns as in American English. This behavior violates GENIE's fourth assumption in Chapter 3. In this case, our classifier, which uses templates that are based on GENIE, attempts to fit a rise-fall accent curve into a raising intonational pattern in Indian dialect which results in inaccurate estimation during training and testing.

Even though the accuracy is not very high for American English vs. Indian English pair, our classifier does not have a bias unlike the baseline. In this pair, the baseline method was able to correctly detect 95% of all American English speakers, while it failed to correctly detect a single Indian English speaker. Therefore by misclassifying all the *In* speakers as American English speaker, the baseline method showed strong bias towards the American English dialect. In order to visually represent the classification bias between the proposed and baseline methods, we calculated the total number of prediction for each dialect in each pairs. The results are illustrated in Figure 6.13, where the ideal case represents the results of a perfect classification between two groups. Therefore, the closer the boundary between the light cyan bar and dark cyan bar is to the red dashed line, the fairer the classifier is. For example in the top row of Figure 6.13, our classifier fairly distinguishes between the two groups American English and Australian while the baseline is biased to classify all speakers in this pair as American English. The baseline has a strong bias to choose the American and British groups over other groups. Our classifier does not have a strong (or any) bias in any classification, and shows a small amount of bias towards Australian English in the (Australian, British) pair.

The above results suggest that extracting large sets of prosodic features (which used by most studies in this area of research) not only leads to inaccurate classification, but also may result in a strong bias in classification. In order to improve the performance of classification systems is it essential to use a more effective way to represent the prosodic characteristics of dialects. Achieving

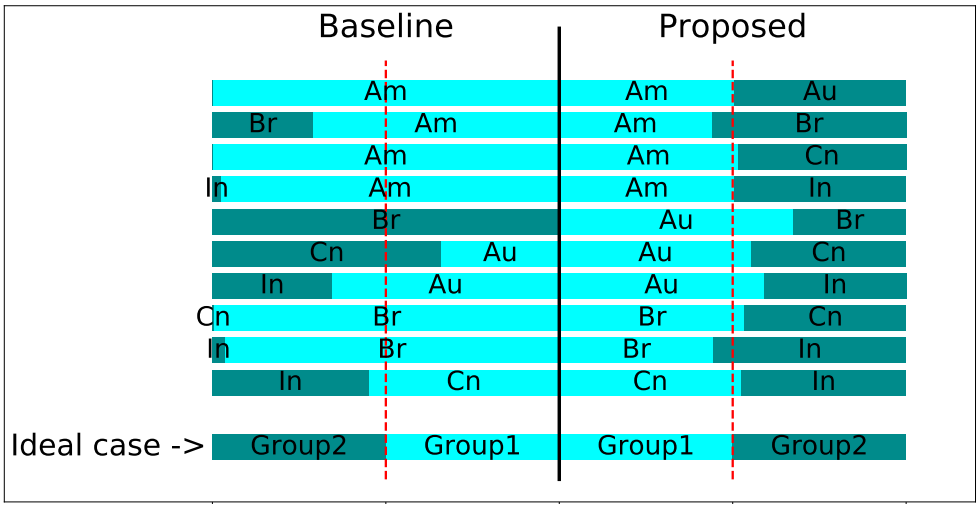


Figure 6.13: Pairwise bias plot. Each row shows side-by-side the classification bias for both methods for a dialect pair.

high prediction with low bias indicates that our classifier can differentiate between two speaker groups by only capturing their prosodic characteristics differences.

6.5 Conclusion

In this chapter we wanted to determine whether we can use F_0 dynamics differences between two speaker groups to differentiate one from another. In order to make this determination we performed an investigation on F_0 dynamics differences between two pairs of speaker groups:

F_0 dynamics in hypokinetic dysarthria: The results suggest that modest levels of classification accuracy are obtained with a model based approach. Even if the accuracy is definitely too low for any practical use, the results are statistically highly significant. This is both important and surprising, given that the groups did not differ in speech intelligibility. Importantly, *GENIE accent curve* results were better than the conventional baseline method, which uses global statistics – mean and standard deviation, or coefficient of variability. In addition, *GENIE accent curve* results were also better than those of less sophisticated methods, such as the raw accent method. Very broadly speaking, this could mean that it is in the fine details of F_0 dynamics that very mild forms of dysarthria first become visible.

F_0 dynamics in clear and conversational speech: By investigating F_0 dynamics differences between CLR and CNV speech, we showed that the speaker uses more feet (i.e., emphasizes more words) to increase the clarity of an utterance. We also showed that this increase of the

foot count is independent of theF_0 range and duration of the utterance. Results for database H were statistically significant, but not for database A. The lack of significance between F_0 dynamics differences of CLR and CNV for database A shows that the speaker uses the same number of feet in CNV speech as in CLR speech. Even though the features used in each study are different, our results are consistent with the results of the following studies : 1)[76] CLR and CNV speech can be differentiated independently of speaking rate. 2) [142] more prosodic units are present in CLR speech than CNV. 3) [65] a significant difference between CLR and CNV for database A was not shown. Therefore, we can conclude that for this speaker the proposed method is a robust method.

Due to success of the first two studies in differentiating two speaker groups through their F_0 dynamics differences, we conducted the third study. In section 6.4, we used the NMF algorithm and the Gini coefficient to perform a group classification. We applied this method to intonation based English dialect classification. In a pairwise comparison framework, we showed that our proposed classifier has less bias and more accurate results compared to a baseline method; the latter had a strong bias toward American and British dialects. These results suggest that the "templates" that are based on GENIE carry more prosodic characteristics of dialects than the baseline large feature set.

Chapter 7

Summary and Future Directions

7.1 Discussion of Contributions

The main focus of this dissertation was the development of a quantitative superpositional intonation model for American English to be used as an analysis and synthesis tool of intonational characteristics in a variety of speech processing applications. As discussed in the first chapter, the purpose of this thesis was to examine the performance of the proposed model in the following aspects:

1. Generating high-quality prediction of F_0 contour, while being linguistically descriptive using a limited number of variables.

2. Modeling real world variations, such as: differences in speaking style, intonational functions, speech data, etc.

In the third chapter, we proposed a quantitive superpositional-based intonation model to estimate F_0 contours using syllable stress, pitch accent, and prosodic phrase boundary labels. We explained what the shared assumptions were between the proposed model and GSM, and how it differed from GSM's other implementations. According to our model, the F_0 contour for a single-phrase utterance can be defined as the sum of a phrase curve and any number of accent curves, one for each foot. Two log-linear curves are used to model the phrase curves, and a combination of the skewed normal distribution and a sigmoid function is used to model three different types of accent curves. First, the skewed normal distribution is employed to model rise-fall accents that occur in non-phrase-final positions as well as, in statements, in utterance-final positions. Second, a sigmoid function is used to model the rise at the end of a yes-no question utterance. And, third, the sum of the skewed normal distribution and the sigmoid function is used to model continuation accents at the end of a non-utterance-final phrase. The parameters for all functions

were optimized simultaneously in one pass. Next, we presented the methodology of the proposed method in section 3.1. Finally, in section 3.3 we examined the proposed method potential to be used as both a synthesis and analysis tool for English intonation through several experiments. In the first part, we discussed that the proposed model follows the prosodic structure of English pronunciation. We showed that the model can capture all intonational patterns categorized by the ToBI system. Therefore the proposed model quantitatively decomposes any F_0 contour into its intonational patterns using a limited set of component curve classes, where each class corresponds to a phonological entity. In the second part, the model was used for objective testing to show it can produce accurate results in comparison with GSM's other implementation (PRISM). We refer to the proposed model "GENeralized Intonation model for English language" (GENIE). In view of the findings discussed here, we can formulate the following finding:

GENIE: is a generalized superpositional intonation model for the English language that has the potential to be used for synthesis and analysis use-cases

In this dissertation, we have provided several frameworks to evaluate the performance of GENIE in term of predictiveness. As proof-of-concept of predictiveness, in the third chapter, we showed that GENIE estimates the F_0 contours of synthetically generated data with very low fitting error which was imperceivable for human ear (overall RMSE was lower than two semitones). We achieved the same results (the overall RMSE lower than two semitones) for both all-sonorant speech data and kid emotional data. For showing that GENIE is capable of generating high-quality F_0 contour, in the fifth chapter, we provided two intonation generation methods: one data-driven based and another neural networks based. Both methods resulted in F_0 curves that were guaranteed to have the desired smooth suprasegmental shapes, and were well-suited to handle sparse training data as well. Perceptual results indicated superior performance of both methods compared to a frame-based approach. Therefore, we can conclude that GENIE can be used for high-quality intonation generation; however, as it was mentioned in the first chapter, in higher level of predictiveness, it is unavoidable for the model to not be linguistically descriptive as well. For showing that GENIE can be used to achieve high-quality prediction of F0 contour, we provided a intonational classifier in the sixth chapter. The assumption behind this classifier was that two speaker groups can be differentiated through their differences in F_0 dynamics. This assumption was examined through investigating the F_0 dynamics differences between two pair of speaker groups. Even though there was no speech intelligibility difference between the speaker groups in both pairs, GENIE was able to tell speaker groups apart in a statistically significant manner. We believe the reason behind this ability is that the component curves of GENIE are linguistically meaningful.

After showing that the F_0 dynamics differences between two speaker groups can be used to differentiate one from another, we proposed to use NMF algorithm and Gini coefficient to perform a speaker group classification. During training, the data-driven intonation generator was applied to F_0 contours in each speaker group, producing a library of parameter vectors characterizing the individual shapes of each component curves; these parameter vectors were labeled in terms some linguistic features. During test, for a given test F_0 contour, we created a "dictionary" of template curves by retrieving parameter vectors whose linguistic labels matched those of the test contour and used these vectors to generate curves with the same duration as that of the test contour. Then the NMF algorithm was used to decompose the input F_0 contours using the dictionary for each speaker group, producing a weight vector for each group. Classification was based on the largest Gini coefficient value of these weight vectors. It should be noted that this classifier does not use any common machine learning method for classification. We evaluated this classifier in a dialect classification framework. We concluded that GENIE can be used to predict a F_0 contour. In the fourth chapter, we showed how GENIE can be used to encode a given F_0 contour in terms of intonational event (e.g., pause-less phrase boundary), then we used this ability (of GENIE to encode F_0 contour) to distinguish between two speaker groups in sixth chapter (differentiating CLR vs CNV speech.) In view of the findings discussed here, we can formulate following finding:

GENIE: provides the high-quality prediction of F_0 contours with few free parameters, while being linguistically descriptive.

To evaluate GENIE's ability to model subtle intonational variation, several frameworks proposed. In the fifth chapter, we performed a perceptual test to examine the ability of GENIE to generate F_0 contour with specific intonational function using marked-up input. We showed that GENIE's ability to convey contrastive stress was comparable to that of natural speech. Also we provided a framework for intonation adaptation which uses GENIE to generate speaker-specific F_0 contour. We showed that the proposed intonation adaptation method shows promise as a way to capture the dynamics of the F_0 contours of a target speaker. Through this dissertation, we examined the performance of GENIE on a variety of data: from synthetically generated data, to more varied and spontaneous (hence less structured) data with a variety of speaker (state) cases. This leads us into the last finding:

GENIE: models real world variations, such as: differences in speaking style, intonational functions, speech data, etc.

7.2 Future Work of Thesis Contributions

As the contributions of this dissertation, we developed a generalized model for English intonation, and it can be broadly used in speech analysis and synthesis applications. However, there still remain some challenges – which have not been investigated in the scope of this study. Some future works for the contributions of the dissertation are listed below:

- Intonation annotation: Usage of GENIE requires detection of foot and intonational phrase boundaries. In chapter four, we showed that GENIE can be used for detection of pause-less phrase boundaries using the goodness-of-fit of the model. Also in the sixth chapter, we showed that this goodness-of-fit of the model can be used to detect the original foot structure of the speaker. Foot boundary detection depends on syllable stress and pitch accent labels. Syllable stress labels are predetermined in English; however pitch accent are variable and based on the speaker's style. Hence, we suggest exploring the ability of GENIE to detect pitch accent label in the same way it was used for tracking down the foot structure. One use-case of this approach could be in stylized speech data which speakers usually do not follow the pitch accent patterns of news-read data – such as emotion data, or patients with speaking disorders. In automatic speech recognition application, this approach could be incorporated with any stress detection method for pitch accent and intonational phrase boundary detection.

- Intonation generation and adaptation: In the fifth chapter, we proposed an intonation generator approach for TTS. This approach, via markup, can generate compelling contrastive stress (contrastive stress is an intonation function which is one type of focus). We generated this intonation function simply by multiplying the accent curve under focus with positive value (increasing the amplitude parameters); however as discussed in the first chapter, different types of focus result in different F_0 movements. For example a narrow focus can be better modeled by changing skewness and scale of the accent curve rather than only changing the magnitude. A possible next step is to study the possible relationship of intonational functions (e.g., focus) with component curve parameters. The use-cases of this approach are listed as follow:

  - Stylizing a TTS system which it is trained on news-read data
  - Increasing clarity of a TTS system by applying more emphasis on certain words
    * Useful for children learning
    * Useful for hearing-impaired listeners

  - Clarifying spoken utterance of a patient suffering from a speaking problem by transforming the perceived identity of the clarified TTS system (previous item) to the patient.

- Intonation classification: in the sixth chapter, we proposed an intonation-based classifier which uses the NMF algorithm for classification. In the proposed method we only used the NMF algorithm for test purposes and not for training (the matrix W, *dictionary*, was generated using the DRIFT method). However, we believe there is a possibility to use the NMF algorithm for training purposes (to build the *dictionary*) as well. The suggested steps are as follow: 1) initializing the matrix W with all possible component curves that can be generated by changing GENIE's parameters. 2) For each F_0 contour in training data, applying the NMF algorithm while only updating matrix H. 3) Applying a threshold [1] on matrix H to pick the most weighted component curves from the matrix W. Furthermore:

  - This approach can be used to filter out similar templates (component curves with very close dynamics) in the matrix W, if we have filled the matrix W using DRIFT's inventory in the initialization step. It is similar to implementing a decision-tree but instead of using a GMM as a representative of component curves in each leaf (last sub-inventory), we use the NMF algorithm to compress the matrix W as a representative of all the component curves in the inventory.

  - This approach can be used for investigating the similarity between two speaker groups. Hypothesized steps are as follow: 1) Generating the matrix W from the first speaker group using DRIFT's inventory. 2) Compressing the matrix W of the first speaker group (as described above). 3) Applying the NMF algorithm on the second speaker group's data, while using the compressed matrix W of the first speaker group. 4) In the end applying a threshold on matrix H to pick the most similar component curves shared between two speaker groups.

- Database issue: In some of the frameworks proposed in this dissertation, we used databases containing limited amount of data. Specially in the case of the clear vs conversational speech classification which recordings of a single male speaker were used. Therefore, for being confident about generalization it is important to use more diverse databases.

---

[1]

  - Defining a threshold might be as easy as determining a fixed value (e.g., all value above 0.5 are eligible), or a more complex measure (e.g., Gini measure).

- Implementation improvement: In the third chapter, for modeling the continuation accent at the end of a non-utterance-final phrase, we proposed using the sum of the skewed normal distribution and the sigmoid function (Equation 3.6) which has seven parameters. The two parameter sets $\{C, \omega, \xi, \alpha\}$ and $\{D, \beta, \gamma\}$ indicates {amplitudes, scale, location, skewness} of the skewed normal distribution, and {amplitudes, slope, location} of the sigmoid function. It is possible to decrease GENIE's degree of the freedom by one point. This can be done by defining the slope (β) of the sigmoid function in terms of the scale and skewness of the skewed normal distribution. One solution is using skew normal cumulative distribution function (Equation 7.1). However, it should be investigated that this decrease in the degree of the freedom does not effect the flexibility of GENIE on capturing various accent curve types.

$$g(t) = D(\Phi((\frac{t-\gamma}{\omega}) - 2T(\frac{t-\gamma}{\omega}, \alpha)) \ , \qquad \text{where } T \text{ is Owen's function} \tag{7.1}$$

Bibliography

[1] David Abercrombie et al. *Elements of general phonetics*, volume 203. Edinburgh University Press Edinburgh, 1967.

[2] Andre G Adami, Radu Mihaescu, Douglas A Reynolds, and John J Godfrey. Modeling prosodic dynamics for speaker recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 4, pages IV–788. IEEE, 2003.

[3] Archana Agarwal, Anurag Jain, Nupur Prakash, and SS Agrawal. Word based emotion conversion in hindi language. In *2010 3rd International Conference on Computer Science and Information Technology*, volume 9, pages 419–423. IEEE, 2010.

[4] Starlet Ben Alex, Ben P Babu, and Leena Mary. Utterance and syllable level prosodic features for automatic emotion recognition. In *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 31–35. IEEE, 2018.

[5] Akiko Amano-Kusumoto, John-Paul Hosom, Alexander Kain, and Justin M Aronoff. Determining the relevance of different aspects of formant contours to intelligibility. *Speech Communication*, 59:1–9, 2014.

[6] Gopala Krishna Anumanchipalli. *Intra-lingual and cross-lingual prosody modelling.* PhD thesis, PhD thesis, Carnegie Mellon University, 2013. Cited in: 2.5, 5.1, 7, 7.1, 2013.

[7] Gopala Krishna Anumanchipalli, Luis C Oliveira, and Alan W Black. A statistical phrase/accent model for intonation modeling. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[8] A Batliner, E Nöth, B Möbius, G Möhler, et al. Prosodic models and speech recognition: towards the common ground. *Proc. Prosody-2000 (Krakow, Poland)*, 2000.

[9] Alireza Bayestehtashk and Izhak Shafran. Parsimonious multivariate copula model for density estimation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5750–5754. IEEE, 2013.

[10] Alan Black, Paul Taylor, Richard Caley, Rob Clark, Korin Richmond, Simon King, Volker Strom, and Heiga Zen. The festival speech synthesis system, version 1.4. 2. *Unpublished document available via http://www. cstr. ed. ac. uk/projects/festival. html*, 6:365–377, 2001.

[11] Alan W Black and Paul A Taylor. Assigning phrase breaks from part-of-speech sequences. 1997.

[12] Benjamin Bock and Lior Shamir. Assessing the efficacy of benchmarks for automatic speech accent recognition. In *Proceedings of the 8th International Conference on Mobile Multimedia Communications*, pages 133–136. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015.

[13] Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

[14] Dwight Bolinger. Intonation and its parts. *Language*, pages 505–533, 1982.

[15] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk — a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, January 2011.

[16] Gerda Martina Cambier-Langeveld et al. *Temporal marking of accents and boundaries*. Thesus, 2000.

[17] Michael J Carey, Eluned S Parris, Harvey Lloyd-Thomas, and Stephen Bennett. Robust prosodic features for speaker identification. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1800–1803. IEEE, 1996.

[18] David T Chappell and John HL Hansen. Speaker-specific pitch contour modeling and modification. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 885–888. IEEE, 1998.

[19] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2006.

[20] Cynthia G Clopper. Frequency of stress patterns in english: A computational analysis. *IULC Working Papers Online*, 2(2):1–9, 2002.

[21] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[22] Alan Cruttenden. *Rises in english*. Routledge, 1995.

[23] David Crystal. *Prosodic systems and intonation in English*, volume 1. CUP Archive, 1969.

[24] Anne Cutler and David M Carter. The predominance of strong initial syllables in the english vocabulary. *Computer Speech & Language*, 2(3-4):133–142, 1987.

[25] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[26] Najim Dehak, Pierre Dumouchel, and Patrick Kenny. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103, 2007.

[27] Pierre Dumouchel, Najim Dehak, Yazid Attabi, Reda Dehak, and Narjes Boufaden. Cepstral and long-term features for emotion recognition. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[28] Mahsa Sadat Elyasi Langarani and Jan Van Santen. Modeling fundamental frequency dynamics in hypokinetic dysarthria. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 272–276. IEEE, 2014.

[29] Mahsa Sadat Elyasi Langarani and Jan Van Santen. Speaker intonation adaptation for transforming text-to-speech synthesis speaker identity. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 116–123. IEEE, 2015.

[30] Mahsa Sadat Elyasi Langarani and Jan van Santen. Automatic, model-based detection of pause-less phrase boundaries from fundamental frequency and duration features. In *SSW*, pages 1–6, 2016.

[31] Mahsa Sadat Elyasi Langarani and Jan van Santen. Foot-based intonation for text-to-speech synthesis using neural networks. *Speech Prosody 2016*, pages 1009–1013, 2016.

[32] Mahsa Sadat Elyasi Langarani and Jan van Santen. Recurrent convolutional networks for classification of speaker groups based on prosodic information. *Women in Machine learning Workshop (WiML)*, 2017.

[33] Mahsa Sadat Elyasi Langarani, Esther Klabbers, and Jan Van Santen. A novel pitch decomposition method for the generalized linear alignment model. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2584–2588. IEEE, 2014.

[34] Mahsa Sadat Elyasi Langarani, Jan van Santen, Seyed Hamidreza Mohammadi, and Alexander Kain. Data-driven foot-based intonation generator for text-to-speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[35] Taoufik En-Najjary, Olivier Rosec, and Thierry Chonavel. A new method for pitch prediction from spectral envelope and its application in voice conversion. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[36] Raul Fernandez, Asaf Rendel, Bhuvana Ramabhadran, and Ron Hoory. F0 contour prediction with a deep belief network-gaussian process hybrid model. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6885–6889. IEEE, 2013.

[37] Hiroya Fons-ant and Shigeo Naessentialr. A model for the synthesis of pitch contours of connected speech. *annual report of the engineering research institute*, 23, 1969.

[38] GJ Freij and Fralik Fallside. Lexical stress recognition using hidden markov models. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 135–138. IEEE, 1988.

[39] Hiroya Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech*, pages 39–55. Springer, 1983.

[40] Hiroya Fujisaki. Information, prosody, and modeling-with emphasis on tonal features of speech. In *Speech Prosody 2004, International Conference*, 2004.

[41] Hiroya Fujisaki and Keikichi Hirose. 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, 1982.

[42] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org, 2017.

[43] Stephanie Gillespie, Yash-Yee Logan, Elliot Moore, Jacqueline Laures-Gore, Scott Russell, and Rupal Patel. Cross-database models for the classification of dysarthria presence. *Proc. Interspeech 2017*, pages 3127–3131, 2017.

[44] Ben Gillett and Simon King. Transforming f0 contours. 2003.

[45] Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg, and Sachin Kajarekar. Combining prosodic lexical and cepstral systems for deceptive speech detection. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

[46] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword third edition ldc2007t07. In *Web Download. Philadelphia: Linguistic Data Consortium*, 2007.

[47] Alexander Gruenstein, Ian McGraw, and Andrew M Sutherland. A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *SLaTE*, pages 109–112, 2009.

[48] John HL Hansen and Gang Liu. Unsupervised accent classification for deep data fusion of accent and language information. *Speech Communication*, 78:19–33, 2016.

[49] Zdeněk Hanzlíček and Jindrich Matoušek. F0 transformation within the voice conversion framework. 2007.

[50] Peter Hawkins. *Introducing phonology*, volume 7. Routledge, 2018.

[51] Matthias Heiler and Christoph Schnörr. Learning sparse representations by non-negative matrix factorization and sequential cone programming. *Journal of Machine Learning Research*, 7(Jul):1385–1407, 2006.

[52] Elina E Helander and Jani Nurminen. A novel method for prosody prediction in voice conversion. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–509. IEEE, 2007.

[53] Romain Hennequin, Roland Badeau, and Bertrand David. Nmf with time–frequency activations to model nonstationary audio events. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):744–753, 2011.

[54] Daniel Hirst and Albert Di Cristo. A survey of intonation systems. *Intonation systems: A survey of twenty languages*, pages 1–44, 1998.

[55] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

[56] Pierre-Edouard Honnet, Branislav Gerazov, and Philip N Garner. Atom decomposition-based intonation modelling. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4744–4748. IEEE, 2015.

[57] John-Paul Hosom. *Automatic time alignment of phonemes using acoustic-phonetic information*. Oregon Graduate Institute of Science and Technology, 2000.

[58] Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[59] Zeynep Inanoglu and Steve Young. A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[60] Zeynep Inanoglu and Steve Young. Emotion conversion using f0 segment selection. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[61] Zeynep Inanoglu and Steve Young. Data-driven emotion conversion in spoken english. *Speech Communication*, 51(3):268–283, 2009.

[62] U Jensen, Roger K Moore, Paul Dalsgaard, and Børge Lindberg. Modelling intonation contours at the phrase level using continuous density hidden markov models. *Computer Speech & Language*, 8(3):247–260, 1994.

[63] Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 285–288. IEEE, 1998.

[64] Alexander Kain and Jan Van Santen. Using speech transformation to increase speech intelligibility for the hearing-and speaking-impaired. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3605–3608. IEEE, 2009.

[65] Alexander Kain, Akiko Amano-Kusumoto, and John-Paul Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *The Journal of the Acoustical Society of America*, 124(4):2308–2319, 2008.

[66] Alexander Kain, Akiko Amano-Kusumoto, and John-Paul Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *The Journal of the Acoustical Society of America*, 124(4):2308–2319, 2008.

[67] Sachin Kajarekar, Luciana Ferrer, Kemal Sönmez, Jing Zheng, Elizabeth Shriberg, and Andreas Stolcke. Modeling nerfs for speaker recognition. In *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[68] Shiyin Kang, Xiaojun Qian, and Helen Meng. Multi-distribution deep belief network for speech synthesis. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8012–8016. IEEE, 2013.

[69] Tomi Kinnunen and R González-Hautamäki. Long-term f0 modeling for text-independent speaker recognition. In *Proceedings of the 10th International Conference Speech and Computer (SPECOM), Patras, Greece*, pages 567–570, 2005.

[70] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010.

[71] Esther Klabbers. Text-to-speech synthesis. In *Foundations in Sound Design for Embedded Media*, pages 297–317. Routledge, 2019.

[72] Esther Klabbers, Taniya Mishra, and Jan PH van Santen. Analysis of affective speech recordings using the superpositional intonation model. In *SSW*, pages 339–344, 2007.

[73] Dennis H Klatt. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *The Journal of the Acoustical Society of America*, 53(1):8–16, 1973.

[74] Dennis H Klatt. Vowel lengthening is syntactically determined in a connected discourse. *Journal of phonetics*, 3(3):129–140, 1975.

[75] John Kominek and Alan W Black. The cmu arctic speech databases. In *Fifth ISCA workshop on speech synthesis*, 2004.

[76] Jean C Krause and Louis D Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1):362–378, 2004.

[77] Akiko Kusumoto, Alexander B Kain, John-Paul Hosom, and Jan PH van Santen. Hybridizing conversational and clear speech. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[78] D Robert Ladd. *Intonational phonology*. Cambridge University Press, 2008.

[79] Tanya Lambert, Norbert Braunschweiler, and Sabine Buchholz. How (not) to select your voice corpus: random selection vs. phonologically balanced. In *SSW*, pages 264–269, 2007.

[80] Javier Latorre and Masami Akamine. Multilevel parametric-base f0 model for speech synthesis. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[81] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9):1162–1171, 2011.

[82] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[83] Younggun Lee and Taesu Kim. Robust and fine-grained prosody control of end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915. IEEE, 2019.

[84] Yun Lei and John HL Hansen. Dialect classification via text-independent training and testing for arabic, spanish, and chinese. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):85–96, 2011.

[85] Philip Lieberman. Intonation, perception, and language. *MIT Research Monograph*, 1967.

[86] Zhen-Hua Ling, Li Deng, and Dong Yu. 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, 2013.

[87] Zhen-Hua Ling, Li Deng, and Dong Yu. Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7825–7829. IEEE, 2013.

[88] Fang Liu, Yi Xu, Santitham Prom-on, and Alan CL Yu. Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling. *Journal of Speech Sciences*, 3 (1):85–140, 2013.

[89] Sheng Liu and Fan-Gang Zeng. Temporal properties in clear speech perception. *The Journal of the Acoustical Society of America*, 120(1):424–432, 2006.

[90] Damien Lolive, Nelly Barbot, and Olivier Boeffard. Pitch and duration transformation with non-parallel data. *Speech Prosody 2008*, pages 111–114, 2008.

[91] Heng Lu, Simon King, and Oliver Watts. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. In *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[92] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. Emotional voice conversion with adaptive scales f0 based on wavelet transform using limited amount of emotional data. In *INTERSPEECH*, pages 3399–3403, 2017.

[93] Jianchun Ma and Wenju Liu. Voice conversion based on joint pitch and spectral transformation with component group-gmm. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 199–203. IEEE, 2005.

[94] Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai. Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space. *Proc. Interspeech 2017*, pages 1238–1242, 2017.

[95] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

[96] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[97] Alvin Martin and Mark Przybocki. The nist 1999 speaker recognition evaluation?an overview. *Digital signal processing*, 10(1-3):1–18, 2000.

[98] Leena Mary. Extraction and representation of prosody for speaker, language, emotion, and speech recognition. In *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*, pages 23–43. Springer, 2019.

[99] Leena Mary and B Yegnanarayana. Prosodic features for speaker verification. In *Ninth International Conference on Spoken Language Processing*, 2006.

[100] Leena Mary and Bayya Yegnanarayana. Extraction and representation of prosodic features for language and speaker recognition. *Speech communication*, 50(10):782–796, 2008.

[101] T Masuko. Pitch pattern generation using multi-space probability distribution hmm. *IEICE Trans. Inf. & Syst.,(Japanese Edition), D-II*, 85(7):1600–1609, 2000.

[102] Takashi Masuko, Keiichi Tokuda, Noboru Miyazaki, and Takao Kobayashi. Pitch pattern generation using multispace probability distribution hmm. *Systems and Computers in Japan*, 33(6):62–72, 2002.

[103] Olga Maxwell. *The intonational phonology of Indian English: An autosegmental-metrical analysis based on Bengali and Kannada English*. PhD thesis, 2014.

[104] Joanne L Miller, Kerry P Green, and Adam Reeves. Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, 43(1-3):106–115, 1986.

[105] Huaiping Ming, Dongyan Huang, Lei Xie, Shaofei Zhang, Minghui Dong, and Haizhou Li. Exemplar-based sparse representation of timbre and prosody for voice conversion. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5175–5179. IEEE, 2016.

[106] Taniya Mishra. *Decomposition of fundamental frequency contours in the general superpositional intonation model*. PhD thesis, Oregon Health & Science University, Department of Science & Engineering, 2008.

[107] Seyed Hamidreza Mohammadi and Alexander Kain. Voice conversion using deep neural networks with speaker-independent pre-training. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 19–23. IEEE, 2014.

[108] Seyed Hamidreza Mohammadi and Alexander Kain. Semi-supervised training of a voice conversion mapping function using a joint-autoencoder. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[109] Seyed Hamidreza Mohammadi, Alexander Kain, and Jan PH van Santen. Making conversational vowels more clear. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[110] Eric Morley, Esther Klabbers, Jan PH van Santen, Alexander Kain, and Seyed Hamidreza Mohammadi. Synthetic f0 can effectively convey speaker id in delexicalized speech. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[111] Matthew Newville, Till Stensitzki, Daniel B Allen, Michal Rawlik, Antonino Ingargiola, and Andrew Nelson. Lmfit: Non-linear least-square minimization and curve-fitting for python. *Astrophysics Source Code Library*, 2016.

[112] Raymond WM Ng, Cheung-Chi Leung, Ville Hautamäki, Tan Lee, Bin Ma, and Haizhou Li. Towards long-range prosodic attribute modeling for language recognition. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[113] Raymond WM Ng, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li. Prosodic attribute model for spoken language identification. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5022–5025. IEEE, 2010.

[114] Raymond WM Ng, Tan Lee, Cheung-Chi Leung, Bin Ma, and Haizhou Li. Spoken language recognition with prosodic features. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1841–1853, 2013.

[115] Brenda Nicodemus. *Prosodic markers and utterance boundaries in American Sign Language interpretation*. Gallaudet University Press, 2009.

[116] Oliver Niebuhr, Mariapaola d'Imperio, Barbara Gili Fivela, and Francesco Cangemi. Are there" shapers" and" aligners"? individual differences in signalling pitch accent category. In *ICPhS*, pages 120–123, 2011.

[117] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[118] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE, 2015.

[119] David John Patterson. Linguistic approach to pitch range modelling. 2000.

[120] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[121] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al.

Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.

[122] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.

[123] Janet Pierrehumbert. Synthesizing intonation. *The Journal of the Acoustical Society of America*, 70(4):985–995, 1981.

[124] Janet Pierrehumbert. Tonal elements and their alignment. In *Prosody: Theory and experiment*, pages 11–36. Springer, 2000.

[125] Janet Breckenridge Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, 1980.

[126] John F Pitrelli, Raimo Bakis, Ellen M Eide, Raul Fernandez, Wael Hamza, and Michael A Picheny. The ibm expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1099–1108, 2006.

[127] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[128] Yao Qian, Zhizheng Wu, Boyang Gao, and Frank K Soong. Improved prosody generation by maximizing joint probability of state and longer units. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1702–1710, 2010.

[129] Manuel Sam Ribeiro and Robert AJ Clark. A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4909–4913. IEEE, 2015.

[130] Scott Rickard and Maurice Fallon. The gini index of speech. In *Proceedings of the 38th Conference on Information Science and Systems (CISS?04)*, 2004.

[131] Srikanth Ronanki, Gustav Eje Henter, Zhizheng Wu, and Simon King. A template-based approach for speech synthesis intonation generation using lstms. In *INTERSPEECH*, pages 2463–2467, 2016.

[132] Andrew Rosenberg. *Automatic detection and classification of prosodic events*. Columbia University, 2009.

[133] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silberger, G. E. Urbanek, and M. Weinstock. Ieee recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246, 1969.

[134] WJ Ryan and KW Burk. Perceptual and acoustic correlates of aging in the speech of males. *Journal of communication disorders*, 7(2):181–192, 1974.

[135] Jan PH van Santen and Adam L Buchsbaum. Methods for optimal text selection. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[136] Jan PH van Santen and Julia Hirschberg. Segmental effects on timing and height of pitch contours. In *Third International Conference on Spoken Language Processing*, 1994.

[137] Jan PH van Santen, Taniya Mishra, and Esther Klabbers. Estimating phrase curves in the general superpositional intonation model. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[138] Bastian Schnell and Philip N Garner. A neural model to predict parameters for a generalized command response model of intonation. In *Interspeech*, pages 3147–3151, 2018.

[139] Stefanie Shattuck-Hufnagel and Alice E Turk. A prosody tutorial for investigators of auditory sentence processing. *Journal of psycholinguistic research*, 25(2):193–247, 1996.

[140] Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Anand Venkataraman, and Andreas Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3-4):455–472, 2005.

[141] Kim Silverman, Mary Beckman, John Pitrelli, Mori Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. Tobi: A standard for labeling english prosody. In *Second international conference on spoken language processing*, 1992.

[142] Rajka Smiljanić and Ann R Bradlow. Temporal organization of english clear and conversational speech. *The Journal of the Acoustical Society of America*, 124(5):3171–3182, 2008.

[143] Kemal Sönmez, Elizabeth Shriberg, Larry Heck, and Mitchel Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Fifth International Conference on Spoken Language Processing*, 1998.

[144] M Kemal Sönmez, Larry Heck, Mitchel Weintraub, and Elizabeth Shriberg. A lognormal tied mixture model of pitch for prosody based speaker recognition. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[145] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

[146] Richard W Sproat. *Multilingual text-to-speech synthesis*. KLUWER academic publishers, 1997.

[147] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller. Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5688–5691. IEEE, 2011.

[148] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing*, 6(2):131–142, 1998.

[149] Xuejing Sun. The determination, analysis, and synthesis of fundamental frequency. *Unpublished doctoral dissertation, Northwestern University*, 2002.

[150] Johan Sundberg. Maximum speed of pitch changes in singers and untrained subjects. *J. Phonetics*, 7(2):71–79, 1979.

[151] Antti Santeri Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, Martti Vainio, et al. Wavelets for intonation modeling in hmm speech synthesis. In *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*. ISCA, 2013.

[152] Johan t Hart. Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, 69(3):811–821, 1981.

[153] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788. IEEE, 2018.

[154] Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. A postfilter to modify the modulation spectrum in hmm-based speech synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 290–294. IEEE, 2014.

[155] Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 805–808. IEEE, 2001.

[156] Jianhua Tao, Yongguo Kang, and Aijun Li. Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1145–1154, 2006.

[157] Paul Taylor. The tilt intonation model. In *Fifth International Conference on Spoken Language Processing*, 1998.

[158] Paul Taylor. Analysis and synthesis of intonation using the tilt model. *The Journal of the acoustical society of America*, 107(3):1697–1714, 2000.

[159] Jonathan Teutenberg, Catherine Watson, and Patricia Riddle. Modelling and synthesising f0 contours with the discrete cosine transform. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3973–3976. IEEE, 2008.

[160] Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Multi-space probability distribution hmm. *IEICE TRANSACTIONS on Information and Systems*, 85 (3):455–464, 2002.

[161] Alice E Turk and James R Sawusch. The domain of accentual lengthening in american english. *Journal of Phonetics*, 25(1):25–41, 1997.

[162] Oytun Turk and Levent M Arslan. Voice conversion methods for vocal tract and pitch contour modification. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[163] Jacqueline Vaissière. 10 perception of intonation. *The handbook of speech perception*, page 236, 2008.

[164] Jan van Santen and Bernd Möbius. A model of fundamental frequency contour alignment, 1999.

[165] Jan PH Van Santen. Contextual effects on vowel duration. *Speech communication*, 11(6): 513–546, 1992.

[166] Jan PH van Santen. Quantitative modeling of pitch accent alignment. In *Proceedings of Speech Prosody 2002*, pages 107–112, 2002.

[167] Jan PH van Santen and Adam L Buchsbaum. Methods for optimal text selection. In *EuroSpeech*, 1997.

[168] Jan PH van Santen and Bernd Möbius. Modeling pitch accent curves. In *Intonation: Theory, Models and Applications*, 1997.

[169] Jan PH Van Santen and Bernd Möbius. A quantitative model of f o generation and alignment. In *Intonation*, pages 269–288. Springer, 2000.

[170] JPH van Santen, ET Prud?hommeaux, LM Black, and M Mitchell. Computational prosodic markers for autism. *Autism*, 14(3):215–236, 2010.

[171] Nancy Vaughan, Daniel Storzbach, and Izumi Furukawa. Sequencing versus nonsequencing working memory in understanding of rapid speech by older listeners. *Journal of the American Academy of Audiology*, 17(7):506–518, 2006.

[172] Christophe Veaux and Xavier Rodet. Intonation conversion from neutral to expressive speech. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[173] Stefanie Veilleux, Nanette Shattuck-Hufnagel and Alejna Brugos. 6.911 transcribing prosodic structure of spoken utterances with tobi. URL `https://ocw.mit.edu`.

[174] Garima Vyas, Malay Kishore Dutta, Jiri Prinosil, and Pavol Harár. An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features. In *Telecommunications and Signal Processing (TSP), 2016 39th International Conference on*, pages 515–518. IEEE, 2016.

[175] Miaomiao Wang, Miaomiao Wen, Keikichi Hirose, and Nobuaki Minematsu. Emotional voice conversion for mandarin using tone nucleus model–small corpus and high efficiency. In *Speech Prosody 2012*, 2012.

[176] Mu Wang, Zhiyong Wu, Xixin Wu, Helen Meng, Shiyin Kang, Jia Jia, and Lianhong Cai. Emphatic speech synthesis and control based on characteristic transferring in end-to-end speech synthesis. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.

[177] Wenfu Wang, Shuang Xu, Bo Xu, et al. First step towards end-to-end parametric tts synthesis: Generating spectral parameters with neural attention. 2016.

[178] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Autoregressive neural f0 model for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8):1406–1419, 2018.

[179] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[180] Chung-Hsien Wu, Chi-Chun Hsia, Chung-Han Lee, and Mai-Chun Lin. Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1394–1405, 2009.

[181] Zhi-Zheng Wu, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li. Text-independent f0 transformation with non-parallel data for voice conversion. In *Eleventh annual conference of the international speech communication association*, 2010.

[182] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1506–1521, 2014.

[183] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication*, 66:130–153, 2015.

[184] Feng-Long Xie, Yao Qian, Frank K Soong, and Haifeng Li. Pitch transformation in neural network based voice conversion. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 197–200. IEEE, 2014.

[185] Yi Xu. Speech prosody: A methodological review. *Journal of Speech Sciences*, 1(1):85–115, 2011.

[186] Yi Xu and Xuejing Sun. Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America*, 111(3):1399–1413, 2002.

[187] Yi Xu and Ching X Xu. Phonetic realization of focus in english declarative intonation. *Journal of Phonetics*, 33(2):159–197, 2005.

[188] Yi Xu, Albert Lee, Santitham Prom-on, and Fang Liu. Explaining the penta model: a reply to arvaniti and ladd. *Phonology*, 32(3):505–535, 2015.

[189] Kathryn Yorkston, David Beukelman, and R Tice. Speech intelligibility test for windows. *Lincoln (NE): Tice Technology*, 1996.

[190] Takayoshi Yoshimura. Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems. *PhD diss, Nagoya Institute of Technology*, 2002.

[191] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Duration modeling for hmm-based speech synthesis. In *Fifth International Conference on Spoken Language Processing*, 1998.

[192] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology*, 1999.

[193] Steve J Young and Sj Young. *The HTK hidden Markov model toolkit: Design and philosophy.* University of Cambridge, Department of Engineering, 1993.

[194] Kai Yu and Steve Young. Continuous f0 modeling for hmm based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5): 1071–1079, 2010.

[195] Kai Yu, Tomoki Toda, Milica Gasic, Simon Keizer, Francois Mairesse, Blaise Thomson, and Steve Young. Probablistic modelling of f0 in unvoiced regions in hmm based speech synthesis. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3773–3776. IEEE, 2009.

[196] Kai Yu, François Mairesse, and Steve Young. Word-level emphasis modelling in hmm-based speech synthesis. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4238–4241. IEEE, 2010.

[197] Stephen A Zahorian and Hongbing Hu. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6):4559–4571, 2008.

[198] Heiga Ze, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE, 2013.

[199] Heiga Zen, Keiichiro Oura, Takashi Nose, Junichi Yamagishi, Shinji Sako, Tomoki Toda, Takashi Masuko, Alan W Black, and Keiichi Tokuda. Recent development of the hmm-based speech synthesis system (hts). 2009.