Relationship between acoustic features and speech intelligibility

Akiko Amano-Kusumoto
M.S., Sophia University, 2002

A dissertation submitted to the faculty of the
Oregon Health & Sciences University
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Electrical Engineering

December 2010

The dissertation "Relationship between acoustic features and speech intelligibility" by Akiko Amano-Kusumoto has been examined and approved by the following Examination Committee:

John-Paul Hosom
Assistant Professor
Thesis Research Advisor

Alexander Kain
Assistant Professor

Jan P. H. van Santen
Professor

Marjorie R. Leek
Senior Research Career Scientist
Portland VA Medical Center

Acknowledgements

I thought I was running a 26.2 mile marathon, but turned out to be a lot longer than a marathon. It seemed it would never end. It would have never ended without all the support I received.

First of all, I would like to thank my advisor, John-Paul Hosom, who has been by my side all the time. He gave me advice and suggestion, but never forced me to do things. He was open to my questions, and asked important questions. He took his time to listen to my ideas. He never blamed me for the mistakes I made. He encouraged me when the results turned out badly by saying I was doing interesting research. Thank you for your guidance and endless support.

I would like to thank my thesis committee members for their valuable comments and suggestions for my thesis. Thank you to Alexander Kain for giving me insights into speech signal processing, advice, and stimulating discussions; Jan van Santen for his advice on statistical analyses; and Marjorie Leek for teaching me that I cannot solve all the problems.

I would like to thank my long-time officemates, Qi Miao and Kristy Hollingshead, for their support and encouragement. We sure liked our warm office, sharing the moments of ups and downs together.

I would like to thank my colleagues at the Center for Spoken Language Understanding (CSLU) for their support and friendship. Thank you to Esther Klabbers for her precious advice on F0 analysis; Xiaochuan Niu for his code on the initial formant contour model; Taniya Mishra for coming up the idea of setting a goal every day; Nate Bodenstab for his constructive advice on my presentations; Pete Jacobs for his difficult questions; Emily Tucker Prud'hommeaux for offering her voice as a female speaker; Maider Lehr for her encouragement; and Meg Mitchell for her expertise on phonetic labels and formant correction. Special thanks to Raychel Moldover for proofreading my papers, asking questions,

giving suggestions, and introducing me to Toastmasters where we practice public speaking and leadership skills.

The National Science Foundation, with its financial support (NSF Grants: BCS-0826654 and IIS-0915754), made it possible to continue my research at OHSU.

I would like to thank all the subjects who participated in my perceptual experiments, as well as those who helped me with recruiting subjects. I tested a total of 145 people who were recruited from the Elsie Stuhr Center, Hillsboro Senior Center, Portland State University, Department of Biomedical Engineering at OHSU, CSLU, and my personal connections.

Last, but not least, I would like to thank my family for their love and encouragement. Thanks to my parents and parents-in-law for allowing me to stay in the United States. Thank you to my sister, Yoko, for her honest advice. Thank you to my husband, Yuichi, for his sacrifice, continuous support, and unconditional love.

Contents

List of Tables

List of Figures

Abstract

Relationship between acoustic features and speech intelligibility

Akiko Amano-Kusumoto

Supervising Professor: John-Paul Hosom

A number of studies have shown that the intelligibility of speech spoken deliberately clearly, referred to as "clear speech" or CLR speech, is higher than that of speech spoken during typical communication, referred to as "conversational speech" or CNV speech. Significant changes in the acoustic features of CLR speech, as compared to those of CNV speech, have been found in previous studies. However, little is known about the relationship between speech intelligibility and the individual sets of acoustic features that are typical in CLR speech. Our long-term goal is to better understand and model those features that contribute to speech intelligibility for different groups of normal-hearing listeners. One objective of this thesis is to identify acoustic features that contribute to the increased intelligibility of CLR speech over CNV speech, which we refer to as "relevant features" for normal-hearing listeners. Our hypothesis is that some acoustic features are more relevant to increased speech intelligibility than others. We have proposed a "hybridization" algorithm that replaces a single feature, or a combination of features, of CNV speech with those of CLR speech, in order to examine the relative contribution of the features to intelligibility. "Hybridized" (HYB) speech is the synthesized speech whose features consist of both CNV and CLR features.

In perceptual experiments, we confirmed that it is possible to obtain intelligibility of

HYB speech that is higher than the intelligibility of CNV speech. In particular, replacing a combination of the acoustic features of duration and spectrum from CNV speech with the corresponding features extracted from CLR speech yielded higher intelligibility than the intelligibility of the baseline CNV speech. In addition, replacing formant frequencies of CNV speech with those of CLR speech was also effective to improve vowel intelligibility of /iː/ and /ei/ without changing the phoneme duration. On the other hand, we found that a combination of energy, fundamental frequency (F0), phoneme sequence, and non-speech (pause) patterns is not a contributing factor to improved intelligibility of CLR speech for the speaker we studied.

To further understand the relationship between phoneme duration and spectrum (i. e. formant frequencies), a formant contour model was developed by decomposing a formant contour into formant targets and coarticulation functions. The effect of speaking style on formant targets was examined using estimated model parameters. The results of model fitting experiments using consonant-vowel-consonant (CVC) words from two speakers showed low fitting error rate with both style-independent (error: 0.3090–0.3834 Bark) and style-dependent (error: 0.2780–0.3712 Bark) targets. At last, we examined methods to reduce formant-tracking errors and to improve the accuracy in extracting F2 slope as applications of this formant contour model.

Chapter 1

Introduction

1.1 Specific aims of the thesis

A number of studies have shown that the intelligibility of speech spoken deliberately clearly, referred to as clear (or CLR) speech, is higher than that of speech spoken during typical communication, referred to as conversational (or CNV) speech. Significant changes in the acoustic features of CLR speech, as compared to those of CNV speech, have been found in previous studies. However, little is known about the cause of increased speech intelligibility and the relationship between speech intelligibility and *sets* of acoustic features that are typical in CLR speech. In this thesis, we examine which acoustic features contribute to speech intelligibility.

One objective in this thesis is to determine which acoustic features of CLR speech are relevant to increased intelligibility. Up until this point, no specific acoustic features have been found which directly contribute to the increased intelligibility of CLR speech. The hypothesis is that some acoustic features are more important to increased intelligibility than others (*relevant* features). We examine differences in intelligibility of the relevant features with young listeners (age: 18–40) with normal hearing. To analyze the relative contributions of a variety of acoustic features to speech intelligibility, we propose a "hybridization" (HYB) algorithm that replaces a single feature, or a combination of features, of CNV speech with those extracted from the CLR speech, using parallel recordings of sentences. Upon waveform synthesis of these "hybrid" (HYB) features, we examine the resulting HYB speech in human perceptual experiments to evaluate relative speech intelligibility levels.

Our specific aims were:

Specific Aim 1: To identify high-level acoustic features that are relevant to increased intelligibility of CLR speech. By using either sentences or consonant-vowel-consonant (CVC) words for the test stimuli (one male speaker), depending on the acoustic features to be evaluated, we replaced a single feature or a combination of features of CNV speech with those of CLR speech using the HYB algorithm, and synthesize HYB speech. We assessed the intelligibility of the HYB speech through perceptual experiments using young listeners (age: 18–40) with normal hearing. In this study, we focused on a single feature or a combination of features such as phoneme duration, energy, fundamental frequency (F0), and spectrum (more precisely, formant frequencies). If the HYB speech had significantly higher intelligibility than the baseline CNV speech, we defined the features that were extracted from CLR speech as relevant features.

Specific Aim 2: To develop a model of relevant features that have been identified as a result of Specific Aim 1. A combination of formant contour and phoneme duration have been identified as relevant features as a result of Specific Aim 1. We develop a model of the formant contour with a linear combination of formant target values and coarticulation functions. We estimated model parameters by minimizing the error between the modeled formant contour and observed data on limited context /w/–/V/–/l/ (one male speaker) as well as general CVC words (one male and one female speakers). We characterized differences in formant target values and coarticulation functions based on speaking style and speaking rate. The style-independent estimated formant targets yielded the first known report of unvoiced consonant targets with a data-driven approach, as opposed to a rule-based approach.

Specific Aim 3: To develop applications of the formant contour model. The automatic extraction of formant frequencies is an important process for speech analysis and formant-based speech synthesis. Especially, extracting F2 transition (or slope) information accurately is important, since F2 transition is used to diagnose speech-related disorder (e. g. dysarthric speech). As an application of the contour model, we have proposed a method to apply above-mentioned formant contour model to reduce formant-tracking errors made by existing formant-tracking software, and to extract F2 slope from the model.

First, we examined whether we can estimate model parameters using the automatically-extract formant contour, which might contain formant-tracking errors. Then, we reduced the formant-tracking errors using the estimated contour model, and detected tokens that had formant-tracking errors. Finally, the extracted F2 slope from the model was compared with that of manually corrected formant contours.

1.2 Organization of the thesis

Chapter 2 provides a literature review of the studies that relate to speech intelligibility and acoustic features. We list our expectation of which acoustic features may or may not be relevant to increased intelligibility of CLR speech. Also, existing signal processing algorithms to improve intelligibility are included in this chapter. In Chapter 3, we identify high-level acoustic features that are relevant to increased intelligibility of CLR speech (Specific Aim 1). We propose a hybridization (HYB) algorithm to identify acoustic features that are relevant to increased intelligibility. We examine a single feature or combinations of high-level acoustic features (i. e. prosodic and spectral features) to verify that (1) we can improve speech intelligibility by hybridizing acoustic features of CLR speech over the baseline speech and (2) we can synthesize the HYB speech without introducing major artifacts. In Chapter 4, we continue to identify relevant features on limited /w/–/*V*/–/l/ word material (Specific Aim 1). Specifically, we describe the characteristics of a formant contour as part of the spectral feature and its relationship with phoneme durations. We determine whether vowel intelligibility can be improved by modifying formant steady-state and formant transitions independent of the phoneme duration, and whether the combination of formant frequencies and phoneme durations can be modified to maximize intelligibility. The results of a series of perceptual experiments using young listeners are discussed in both Chapters 3 and 4.

Chapter 5 describes the methodology of the formant contour model and preliminary results (Specific Aim 2). With limited data (/w/–/*V*/–/l/ word), we further characterize the effect of speaking style and speaking rate on formant contour by examining model parameters. In Chapter 6, we expand the number of phonemes in a database, so that the

context-independent formant target values are estimated using the formant contour model. The dependency of speaking style on formant targets, estimation of data-driven consonant targets, and speaker differences are discussed in this chapter. We elucidate whether a speaker aims at either different positions of the articulators based on the speaking style, or global positions regardless of the speaking style using modeled parameters.

Chapter 7 presents an application of the formant contour model to reduce formant-tracking errors (Specific Aim 3). The first step is to extract formant model parameters using an automatically-extracted formant contour. Next, the method to analyze the error and binary classification of detecting tracking errors are described. Finally, the results from the error analysis, error detection threshold, and F2 slope distribution of various sets of the data are discussed.

Finally, Chapter 8 concludes this thesis by providing the contributions of the thesis, possible applications, limitations of our findings, and future directions in this area of research.

Chapter 2

Background

2.1 Introduction

A number of studies have shown that the intelligibility of speech spoken deliberately clearly, referred to as clear (or CLR) speech, is higher than that of speech spoken during typical communication, referred to as conversational (or CNV) speech. Significant changes in the acoustic features of CLR speech, as compared to those of CNV speech, have been found in previous studies. However, little is known about the cause of increased speech intelligibility and the relationship between speech intelligibility and *sets* of acoustic features that are typical in CLR speech. In this chapter, we provided a review of those studies regarding the importance of acoustic features to speech intelligibility.

In the next section "Conversational and clear speech" (Section 2.2), we summarize prior work that studied differences between CNV and CLR speech, for a variety of speakers, in terms of speech intelligibility and in terms of acoustic characteristics. The speech intelligibility levels measured in these studies include phoneme, word, and sentence intelligibility, with and without semantic context. While intelligibility at one level (i. e. sentence) can not be used to reliably predict intelligibility at a different level (i. e. word), there is a dependency between phoneme, syllable, word, and sentence intelligibility levels [12]. For example, while the identity of an unclear phoneme may be recovered from a larger (e. g. word) context, phoneme intelligibility does impact word intelligibility. Therefore, we do not restrict this review of prior work to only one level, but we include results from all levels. General hypotheses about the importance of acoustic features that result from this review of prior work then have to be tested under more controlled circumstances, such as

with specific test materials and noise conditions.

In the section "Relationship between acoustic features and speech intelligibility" (Section 2.3), we list a variety of acoustic features, grouped by prosodic and spectral features, and their correlation with speech intelligibility. We suggest whether each feature could contribute to increased intelligibility. The studies included in this section are not limited to studies of the intelligibility of CLR speech, but include studies that investigate speech intelligibility more generally. Finally, in the section "Digital signal processing of speech to increase intelligibility" (Section 2.4), we summarize existing signal-processing techniques to improve speech intelligibility.

2.2 Conversational (CNV) and Clear (CLR) Speech

Conversational speech is the speech elicited when speakers are instructed to "speak in the same manner as you would during an ordinary conversation" [79]. On the other hand, clear speech, which is the outcome of a clear speaking style, is obtained by instructing speakers to "speak clearly, as you would when talking to hearing-impaired listeners" [79]. These two types of speech, defined here as CNV and CLR respectively, have commonly been referred to as "conversational" and "clear" [e. g. 80]. However, these labels are ultimately problematic since the term "clear" might imply intelligibility of the perceived speech, whereas the term "conversational" might imply speech produced as part of a dialogue. It is also possible that "conversational" speech has, in some cases, intelligibility equal to that of "clear" speech. Therefore, in this thesis, we avoid confusion by using the terms CNV and CLR to specifically refer to only the *style* of speech generated in response to the instructions mentioned.

An increase in loudness when people speak in the presence of background noise is called the Lombard voice reflex [59]. Lombard speech is the outcome of the Lombard reflex, which is produced when the noise is introduced to the *speaker*, while CLR speech is produced when the speaker believes that the *listeners* have hearing difficulty. Both speaking styles have an effect on intelligibility, so that the resulting speech can be more easily understood by listeners (or speakers themselves). A number of studies have investigated the effects

of Lombard speech on speech intelligibility, as well as the acoustic features of Lombard speech [83, 24, 47, 100]. However, it is not known how the acoustic properties of CLR speech and Lombard speech differ. In this thesis, the intelligibility and acoustic features of only CLR speech are investigated. In this section, we review prior studies that have examined the intelligibility of CNV and CLR speech under different conditions. Acoustic differences between CNV and CLR speech are discussed, as well as the effects of different speakers.

2.2.1 Intelligibility of CNV and CLR speech

Many studies have tested intelligibility of CNV and CLR speech under various conditions [e. g. 79, 30, 29, 54, 65, 15, 67]. The increased average intelligibility of CLR speech, as compared to CNV speech, has been shown in studies of different types of listeners, including normal-hearing listeners aged 18–32 [54, 65, 29, 67], normal-hearing listeners aged 61–88 [40], hearing-impaired listeners aged 60–89 [79, 89, 105], simulated hearing-impaired listeners aged 19–33 [67], cochlear-implant users [65], and school-age children with and without learning disabilities [15].

In almost all cases, the average intelligibility of CLR speech was higher than that of CNV speech, across speakers. For example, intelligibility differences for hearing-impaired listeners evaluating nonsense sentences spoken in the CNV and CLR speaking styles were found to be significant, with a difference of about 17 percentage points [79]. Picheny *et al.* [79] concluded that the CLR speech advantage was independent of listener, presentation level, and frequency-gain characteristics. On the other hand, Ferguson and Kewley-Port later reported that the CLR speech benefit can be listener specific [30]. Their results showed that vowel intelligibility of CLR speech was not higher than CNV speech for elderly hearing-impaired listeners when identifying the front vowels. The authors speculated that this result may have been because increased F2 values were in a region where the hearing-impaired listeners had a sloping hearing loss (i. e. 2000–2500 Hz). Similarly simulated hearing-impaired listeners performed better only for distinguishing sibilants in /s/-/ʃ/, and /z/-/ʒ/ pairs with a CLR speaking style, and had decreased intelligibility for voiceless non-sibilants in /f/-/θ/ pairs in perception of fricative consonants [67]. These studies

showed that the benefit from a CLR speaking style was dependent on the characteristics of a group of listeners, namely whether they have hearing loss and type of hearing loss [30, 67].

Intelligibility differences between CNV and CLR speech have been demonstrated on a variety of speech materials: vowels in /b/–/V/–/d/ context [30], VCV syllables identifying fricatives [67], nonsense sentences [e. g. 79, 78], and meaningful sentences [14, 89]. Sentence level materials are more similar to everyday communication environments, while word level material with nonsense syllables provides more control over the phonemes that are evaluated. The availability of semantic cues, which are present in meaningful sentences, is an important factor that influences speech intelligibility [37]. Listeners can compensate for deteriorated speech by using semantic cues. Elderly listeners are relatively better at utilizing semantic cues than younger listeners, since linguistic knowledge is well preserved with age, although aging can result in deteriorated speech understanding in the presence of background noise [36, 82, 96, 108]. It is important to note that speech intelligibility at the phoneme level should not be used to predict sentence-level intelligibility.

As shown in [30], the benefit from the CLR speaking style might be different for young listeners (age: 20–23 years) with normal hearing and elderly listeners (age: 60–89 years) with hearing impairment. Ferguson and Kewley-Port concluded that the intelligibility difference between CNV and CLR speech varies according to the listener's age and hearing status (with or without hearing impairment). In summary, speech intelligibility of CNV and CLR speech has been examined using various speech materials and various listener groups. In general, the average intelligibility of CLR speech is higher than that of CNV speech, although age and hearing status may affect the differences. In this thesis, we focus on testing young listeners aged 18–40.

2.2.2 Acoustic differences between CNV and CLR speech

Previous work has examined acoustic differences between CNV and CLR speech [15, 80, 55, 30]. Although these previous findings were not always in agreement because of variability in speakers, speech materials, and analysis methods, we review the results that investigated acoustic differences. Major findings can be grouped according to prosodic (a combination of fundamental frequency, energy, and phoneme duration), spectral (such as formant and

formant-normalized spectrum), and phonological features.

For prosodic features of CLR speech, the fundamental frequency (F0) generally showed a slight increase in mean and variability (or range) [80, 15, 55]. The consonant-vowel energy ratio (CVR) was increased in CLR speech, particularly for stops and fricatives (i. e. energies of the consonants had greater relative energy in CLR speech) [15]. In another study, an increased CVR was found only in affricates of CLR speech [55]. Picheny *et al.* [80] reported greater root-mean square (RMS) intensities for unvoiced stop consonants in CLR speech than in CNV speech. Phoneme durations were lengthened, especially in the tense vowels /iː/, /u/, /ɑ/, and /ɔ/ [80, 30, 31]. Pause durations were longer and their occurrence was more frequent. As a result of prolonged phoneme durations and increased pause durations, the speaking rate was significantly decreased from 160–200 words per minute (wpm) in CNV speech to 90–100 wpm in CLR speech [80, 55]. Increased amplitude modulation for low modulation frequencies (up to 3–4 Hz) of CLR speech was also found on a limited number of speakers [55]. We interpret this to mean that the depth of envelope of the CLR speech signal was amplified; therefore, syllables were better separated from each other. The duration between the time of the burst and the onset of the voicing has been defined as voice onset time (VOT). One of the talkers showed VOTs for voiceless stop consonants increased in CLR speech, which led to increased differences between voiced and voiceless stop consonants [55].

For spectral features of CLR speech, vowel formant frequencies showed an expanded vowel space [80, 30, 15]. Long-term average spectra had higher energies at higher frequencies [55], which can be interpreted as decreased spectral tilt. Although formant undershoot was observed with /w/–/ V /–/l/ contexts in both CNV and CLR speech, the amount of second formant displacement from the target was significantly less in CLR speech. The formant displacement was dependent on vowel duration more for the lax vowels for both CNV and CLR speech, similar to the findings from Picheny *et al.* [80] that the formant frequencies showed more variation in lax vowels [68].

For phonological features, studies have shown a number of differences between CNV and CLR speech. Vowel reduction (i. e. vowels becoming schwa-like), degemination (i. e. two similar phonemes merged into one sound), and alveolar flaps occurred more often

in CNV speech. On the other hand, bursts of the stop consonants in word final position tended to be released more often in CLR speech. Also, sound insertion of a schwa after a voiced consonant occurred more often in CLR speech [80, 55].

As a summary of acoustic differences, we conclude that prosodic, spectral, and phonological features show differences between CNV and CLR speech. The following are general trends among the studies: (1) F0 mean and range for CLR speech are increased relative to CNV speech, (2) phoneme durations are longer in CLR speech, (3) amplitude modulation is increased for CLR speech, (4) vowel spaces are increased in CLR speech, (5) there are higher energies at higher frequency regions in CLR speech, and (6) phoneme insertions (e. g. schwa) occur more often in CLR speech. Even though many features are reported as being different between the two speaking styles, different speakers in each study showed different acoustic-phonetic effects. In the next section, the effects of different speakers on speech intelligibility are discussed.

2.2.3 The effects of different speakers

A number of studies have looked at the effects of different speakers on the intelligibility of CLR speech [29, 89, 18]. It has been shown that both young and elderly speakers need minimal instruction and practice in producing CNV and CLR speech [89]. However, longer-term training in producing CLR speech can lead to changes in more types of speech parameters, more stable changes, and better understanding of speech by listeners with hearing loss [18]. Although simple instructions for eliciting CLR speech yielded increased intelligibility, training in producing CLR speech caused substantial changes in CLR speech features; people with hearing loss performed as well as normal hearing listeners when listening to a speaker who received CLR speech training [18].

Prior work that found different speakers employ different strategies to produce CLR speech [30] led to a study examining 41 speakers producing /b/–/V/–/d/ utterances [29]. Ferguson demonstrated significant speaker differences in vowel intelligibility for normal-hearing listeners [29], but the only factor that was found to be correlated with the increased intelligibility of CLR speech was gender, while other factors, such as speakers' age and communication experience with hearing-impaired listeners, were not. Also, Ferguson and

Kewley-Port [31] demonstrated that "atypical talkers" produced significant perceptual CLR speech benefit over the CNV speech, but did not present spectral and durational effects, and vice versa. Bond and Moore found that characteristics of these least-intelligible speakers revealed shorter word and vowel durations, the least differentiated vowel spaces, minimal cues for consonantal contrasts (e. g. the VOT in stop consonants), and the most varied amplitude of stressed vowels [11].

It is unclear how the acoustic features from different speakers affect speech intelligibility. The reason for less intelligible speech might not be the result of one "ambiguous" feature, but a combination or interaction of several features. In this thesis, we focus on examining relevant features from one speaker (Specific Aim 1). The speaker-dependency of the relevant features (whether a set of relevant features from one speaker is applicable to another speaker) will be examined in the future. It will be important to determine if the strategies that one speaker utilizes to produce increased intelligibility are effective for different speakers (Section 8.4.2).

2.3 Relationship between acoustic features and speech intelligibility

A number of studies have investigated the relationship between acoustic features and speech intelligibility by evaluating a number of talkers who produced a range of speech intelligibility levels. Some of these studies have computed the correlation between changes in a certain acoustic feature (e. g., fundamental frequency) and speech intelligibility [e. g. 38]. Other studies have investigated the relationship between stimulus variability and spoken word recognition [95]. In this section, we described acoustic features and their relationship to intelligibility, classified into prosodic and spectral features as described in Section 2.2.2. We summarized this section by listing all the acoustic features which may (or may not) contribute to an increase in speech intelligibility. The studies covered in this section are not limited to the intelligibility of CLR speech, but are more general studies of acoustic features and speech intelligibility [16, 38, 95].

2.3.1 Prosodic features

Fundamental Frequency (F0)

The relationship between F0 and speech intelligibility has been investigated in many studies [e. g. 16, 38, 94, 31]. Bradlow *et al.* [16] found that there was no correlation between mean F0 and sentence intelligibility when gender was taken into account. Manipulating F0 values of synthesized speech with a global 10 %, 20 %, or 30 % decrease, or a 10 %, 20 %, or 30 % increase did not have an impact on word identification rate [94]. According to the phonetic relevance hypothesis of Sommers and Barcroft [94], these results suggest that F0 mean is not relevant to intelligibility.

The range of F0 (the difference between maximum F0 and minimum F0) was found in one case to be significantly correlated with sentence intelligibility across a set of 20 speakers [16], while Hazan and Markham [38] did not find significant correlations between F0 range and word intelligibility. These two studies had several differences, including speech material (sentence versus word intelligibility), measurement of F0 values (logarithmic versus linear scale), and different speakers.

The question of whether F0 is an important cue for English phoneme identification still remains debatable. Klatt [53] stated that English vowels can be described in terms of the frequencies of the lowest three formants and formant transitions, regardless of F0 values. Also, according to O'Shaughnessy [76], in a nontonal language such as English, F0 is virtually independent of phoneme identity. A study by Hoemeke and Diehl [43], however, showed that perception of vowel height is influenced by the distance between F0 and F1. The relationship between F0 and F1 is not reflected in the F0 mean or range, and so the impact of F0 on intelligibility may be characterized by a more complex relationship. While CLR speech usually shows higher mean F0 values, as shown in Section 2.2.1 [e. g. 80], the relationship between F0 values and speech intelligibility is still unclear.

Energy

Two types of energy measurement have been considered in a number of studies: one is the relative energy ratio between consonants and neighboring vowels, or the consonant-vowel-ratio (CVR), and the other one is overall energy. Hazan and Markham [38] found

no significant correlation between word intelligibility and CVR for nasals, fricatives, or stop consonants in naturally-produced speech. Hazan and Simpson [39] did demonstrate that artificial enhancement of the CVR leads to improved intelligibility (on the order of 10 percentage points) at both the VCV word and nonsense sentence level. They concluded, however, that it is not straightforward to determine which consonants to amplify or the appropriate level of amplification. Furthermore, the amount of enhancement required to improve intelligibility is much greater than can be produced by the human speech production system. Even though the CVR was increased for stop release burst and fricatives in CLR speech as compared with CNV speech [15] (Section 2.2.1), we conclude from these studies of energy that the CVR may not contribute to increased speech intelligibility of naturally-spoken speech. However, artificial enhancement of the CVR could be an effective method to improve intelligibility.

The overall energy (intensity) of the test stimuli is one of the factors that significantly affects intelligibility [32]. Because overall intensity has a known impact on intelligibility and is typically not a feature of interest, most studies normalize the overall amplitude for both CNV and CLR speech to the same value. Despite this energy normalization, CLR speech is still generally more intelligible than CNV speech; therefore, we conclude that there are other features contributing to increased speech intelligibility.

Duration

A number of studies have looked at the effect of duration (at the phoneme, word and sentence level) on speech intelligibility [54, 16, 38, 105]. Monosyllabic word duration was found to be positively correlated with word intelligibility [38]. On the other hand, Bradlow *et al.* [16] did not find speaking rate to be correlated with sentence intelligibility. In this study, speaking rate was measured from overall sentence duration, which may be different from speaking rate measured from individual word duration. Varying speaking rate, both naturally and digitally, has resulted in impaired identification of spoken words, which supports the importance of speaking rate for intelligibility [94]. The degree of correlation, if any, may depend on the definition of speaking rate.

In a study from Krause and Braida [55], even when speaking rates were matched in

CNV and CLR speaking styles (with differences of no more than 25 wpm), CLR speech still had higher intelligibility. (For the normal rate, CNV speech was 45 % intelligible, while CLR speech was 59 % intelligible). They concluded that CLR speech has acoustic differences other than speaking rate that are inherently different from CNV speech.

Hillenbrand *et al.* [41] stated that duration is an important cue for vowel identity. Changing the vowel duration of /hVd/ syllables degraded vowel identity, and perception of the following vowel contrasts were significantly affected: (/ɑ/–/ɔ/–/ʌ/), and (/æ/–/ɛ/). Bradlow *et al.* [16] found a positive correlation between stop closure duration and rate of /d/ detection as in "walled town"). A long duration of /s/ relative to the surrounding vowels (as in "play seems") led to syllable affiliation ("place seems"). This study showed that inter-segmental timing is important for speech intelligibility.

It should be noted that the phoneme-level perceptions [e. g. 41] in nonsense syllables could be less important in the case of word- or sentence-level perception, but in a controlled experiment at the phoneme level we should pay attention to inter-segmental timing. We speculate that one reason why stretching phoneme duration uniformly does not improve speech intelligibility [81] is that the naturalness of inter-segmental timing is disrupted. It may be possible to reduce errors in the inter-segmental timing of synthetic or modified speech by taking the duration of each phoneme from naturally-spoken CLR speech [e. g. 105], as reported in Section 2.4.

In summary, even though many studies have been conducted in terms of phoneme, word, and sentence duration, the findings are not all in agreement. Therefore, it is difficult to conclude whether speaking rate alone, or duration alone, are acoustic features that are responsible for the increased speech intelligibility of CLR speech.

Pauses

A combination of phoneme duration and pauses determines speaking rate, which has been studied as one acoustic feature. Krause and Braida [55] found that when they controlled the speaking rate in CNV and CLR speech, the pause distribution (frequency and duration) was nearly equivalent in both CNV and CLR speech. They concluded that pause duration and frequency are not necessary components of increased speech intelligibility. Krause and

Braida [55], however, tested young (aged 18 to 29 years) listeners with normal hearing in their perceptual experiments. We speculate that people with hearing loss or elderly listeners (over 60 years) may be able to get benefit from frequent pauses and longer pause durations. The importance of pauses still needs to be investigated further in the future, based on the listeners' age and hearing status.

Amplitude Modulation

The temporal envelope may play an important role in speech intelligibility [25, 26]. Drullman *et al.* [25, 26] suggested that amplitude modulation in the range between 4 Hz and 16 Hz is the most important for sentence intelligibility, and that amplitude modulation as low as 2 Hz is important for phoneme identification. Liu and Zeng [66] examined the importance of the temporal envelope and fine temporal structure (the complementary set to the temporal envelope for representing amplitude) on speech perception in "auditory chimera" speech. Their conclusion was that the temporal envelope contributes more to the CLR speech advantage at high signal-to-noise ratios, while the temporal fine structure contributes more at low signal-to-noise ratios. However, the intelligibility of the auditory chimera was poorer than that of the original CNV speech. Therefore, it is important to mention that the presence of processing artifacts noted by the authors might have influenced their findings. Similarly, the results from Krause and Braida [56] showed that altering the temporal intensity envelope had detrimental effect on intelligibility due to processing artifacts, although increased modulation spectra of CLR speech was reported previously [55]. In conclusion, if the processing artifacts can be minimized by altering the temporal intensity envelope, the importance of amplitude modulation can be further investigated in the future.

2.3.2 Spectral features

Formant Frequencies

The importance of formant movements for speech intelligibility has been examined using naturally-produced speech, synthesized speech with original formants, and flat formants [42]. The synthesized signals with original formants had a higher vowel identification rate than signals with flat formants, indicating that formant movement plays an important role in vowel identification. Another study by Smits *et al.* [93] showed that the formant transitions associated with prevocalic voiced stops are more effective than the bursts of the same stops for stop identity, despite the fact that the relative importance of formant transitions was shown to be highly dependent on the vowel context.

Turner *et al.* [104] investigated the effect of lengthening formant transitions of the stop consonants on synthesized syllables with hearing-impaired listeners. As the formant transition region of synthetic speech was lengthened (with a minimum of 5 ms and a maximum of 160 ms), the stop identification rates increased rapidly and were close to perfect performance for transitions 20 ms and longer for normal hearing listeners, while not all hearing-impaired listeners showed improvement with lengthened formant transitions. They concluded that the more severe a listener's hearing loss, the less benefit they obtained from lengthened formant transitions [104].

Formant undershoot of the second formant frequency has been identified in the vowels /iː/, /ɪ/, /ɛ/, and /ei/ for different speaking styles, including CLR speech, but undershoot is less dramatic in the CLR speech style than in the CNV speaking style [68]. Chen [19] also reported that in his CVC materials, the first and second formant frequencies of tense vowels tended to reach their target frequencies and to have less variance in CLR than in CNV speech. However, Krause and Braida [55] showed that formant values extracted from the vowel midpoints were not closer to the formant target frequencies nor less variant in CLR speech spoken at CNV speaking rate (CLR/normal) than in CNV speech. Authors stated that the formant contour of CLR/normal speech might have reached to the formant target frequencies closer than CNV speech, and that measurement at only one time point might not be sufficient to capture the differences. They speculated that listeners might rely on

the entire formant movement throughout a vowel rather than focus on the midpoint of a vowel or the transition only. Studies on the size of the vowel space indicate that speakers with larger vowel spaces are more intelligible than speakers with reduced spaces [38, 16]. In particular, speakers who had a wide F1 range (defined as the difference between F1 for /iː/ as in "easy" and F1 for /ɑ/ in "pot") appeared to have higher intelligibility scores than speakers with a smaller F1 range. The F2 range (defined as the difference between F2 for /iː/ and F2 for /ɔ/) was found to be significantly correlated with word intelligibility [38], but was not found to be correlated with sentence intelligibility [16]. This difference may be due to differences in speech materials or in the methods used to elicit speech material. According to Ferguson and Kewley-Port [30], steady-state formant values for back vowels, dynamic formant movement, and duration for front vowels are the primary cues for the vowel identities with young normal-hearing listeners.

In summary, studies show that formant transitions of the stop consonants are an important feature [93]. From the study of Turner *et al.* [104], we expect that lengthening formant transitions of the synthetic speech might be an efficient technique to improve stop consonant identification for normal hearing listeners. On the other hand, lengthening formant transitions of natural speech might impact the vowel identity. The success of Turner's study on improving the stop identification rates may be because they used synthetic syllables. We do not expect that Turner's work on synthetic syllables can be generalized to natural speech sentences, because stop consonants appear in a wider variety of vocalic contexts, which leads to the importance of a more general coarticulation model. We speculate that the vowel space (F1 and F2 range) is a significant feature for increased speech intelligibility, and that formant transitions may also play an important role.

Spectral Balance

The spectral balance, as opposed to prosody, is another factor that may affect speech intelligibility. Speakers tend to increase vocal effort and raise their overall energy to make speech more intelligible. Liénard and DiBenedetto [63] found that increasing vocal effort is correlated with increased values of F0, and formant amplitudes of F1, F2 and F3. Also they found that the formant amplitudes in the higher frequency range increased

more than those in the lower range, thereby decreasing spectral tilt by increasing vocal effort. Formant bandwidths were significantly narrower in CLR speech than in CNV speech, both spoken at a normal rate (approximately 200 wpm), which indicates higher formant amplitudes in the short-term spectra of CLR speech as compared with CNV speech [55]. An increase in energy in the 1–3 kHz frequency range of the long-term average spectrum (LTAS) has been found to be significantly correlated with intelligibility [55, 38]. Hazan and Markham [38] also found that the *slope* of the LTAS was not correlated with intelligibility, which may contradict previous findings since an increase in energy in the high frequency range should be equivalent to a decreased slope of the LTAS. These two results (spectrum energy and slope) indicate that the increased energy between 1–3 kHz did not lead to a decrease of spectral slope. Amplifying the spectrum between 1000 and 3000 Hz yielded an intelligibility increase over CNV speech levels, but did not reach CLR speech levels [56]. These results suggest that the increased energy in the 1–3 kHz range of CLR speech is partially responsible for increased intelligibility of CLR speech. Therefore, we expect that an energy increase in the 1–3 kHz range and formant bandwidths may be a contributing factor, but spectral balance (or return phase of the glottal source) may not be a significant feature contributing to speech intelligibility.

Speaker Characteristics

It has been noted that speech quality has an effect on speech intelligibility. In particular, variation in speaking style or voice type (normal, nasalized, child-directed, whispered, excited, and elongated) within a presentation block reduced word intelligibility relative to presentation with a single speaking style [94]. According to the phonetic relevance hypothesis, this result indicates that the speech quality (resulting from different speaking styles) is relevant to word intelligibility. Findings from subjective ratings were that less-intelligible speakers were judged as sounding "mumbly, unpleasant, muffled, or weak", relative to the more intelligible speakers [38]. However, dimensions that are related to the quality of voiced excitation (harsh/smooth, creaky/non-creaky, husky/not-husky) were not found to be correlated with intelligibility [38]. Although subjective factors, such as mumbly or unpleasant, may lead to a speech intelligibility decrease, voice quality is difficult

to quantify and directly match to acoustic features other than glottal source parameters.

Another important aspect of speech intelligibility is the impact of gender difference. Female speakers are generally more intelligible than male speakers [38, 16]. It is possible that the wider F0 range of females is one of the characteristics that contributes to the higher intelligibility of female speech relative to male speech. In addition, female speakers tend to have a larger vowel space, more precise inter-segmental timing than male speakers, and less frequent alveolar flapping [16, 15]. Therefore, it is not clear whether the observed intelligibility difference, with 93.4 % for female speech and 81.1 % for male speech in the study from [16], was due to one factor or a combination of these factors.

2.3.3 Summary of acoustic features

The correlation results from prior studies do not necessarily imply that features that are highly correlated with speech intelligibility are the *cause* of an improvement in speech intelligibility. It is still not known which individual acoustic feature, or combination of features, make speech more intelligible, or their relative contribution to increased speech intelligibility. None of these prior studies have looked at the effect of a combination of features, which may interact with each other.

Sommers and Barcroft [95] and Sommers *et al.* [94] hypothesized that variation in features that are relevant to speech perception degrades speech intelligibility. It may be true that stimulus variability in speech can affect speech intelligibility, but there can also be other sources that degrade speech intelligibility. In some of their experiments, degraded intelligibility could have been due to synthesized speech with features that are different from what humans would normally produce, instead of the relevancy of the feature to speech perception. Although using synthesized speech is one way to control feature values (for example, F0 values), we speculate that the values used should be within the degree found in natural speech in order to test perceptual relevancy in realistic settings.

Based on this prior work, we expect that formant transitions, temporal envelope, F1 and F2 ranges, energy in 1–3 kHz, formant bandwidth, and VOT are the acoustic features that may be most responsible for increased speech intelligibility. F0 mean and CVR may not be as important to speech intelligibility. Other features, including duration (speaking

rate), F0 trajectories, F0 range, long-term energy, spectral balance (or glottal source characteristics), and pause duration are not conclusively significant for speech intelligibility, because of the unclear or contradicting results from prior studies.

2.4 Digital signal processing of speech to increase intelligibility

In this section, we introduce signal processing schemes that modify some of the acoustic features listed in Section 2.3, namely CV energy ratio, pause durations, phoneme durations, and temporal envelopes.

Elderly listeners have difficulty processing brief consonant cues such as the burst portions of stops [37]. One of the approaches to improve consonant identification has been to enhance the intensity of the consonants in consonant-vowel (CV) syllables. Successful improvement has been obtained by amplifying the consonant energy in consonant-vowel (CV) and vowel-consonant-vowel (VCV) sequences for normal hearing listeners [34, 39] and hearing-impaired listeners [35]. After adjusting the degree of amplification of the burst and aspiration, Hazan and Simpson also improved the intelligibility of nonsense sentence materials [39]. Similarly, in a study by Skowronski and Harris amplifying the CV ratio, what they called energy redistribution using voiced/unvoiced information (ERVU), as well as increasing the spectral energy center of gravity by high-pass filtering (HPF), increased monosyllabic word intelligibility over unmodified speech [92]. Another study showed that lengthening only consonant durations in CV syllables did not affect intelligibility, using normal-hearing listeners aged 65–72 [34]. In the same study, a combination of amplifying the consonant energy and lengthening consonant durations improved consonant identification, but the results did not improve relative to the stimuli with only consonant energy amplification [34].

Increasing pause durations in a sentence for people with hearing loss [105] or inserting pauses between words in a sentence for both young and old people with and without hearing loss [36] did not improve the intelligibility of meaningful sentences. We speculate that one of the reasons for this negative result could be that inserting pauses between

words disrupts the normal prosodic contour of a sentence, or pausing may simply not be an important component of speech intelligibility. Although Liu and Zeng [66] obtained a 13 % absolute improvement by inserting pauses between words to slow down the speaking rate, they did not exclude pauses when calculating root-mean-square (rms) values used in energy normalization, which resulted in increased energy in sentences with longer pauses. Therefore, the improvement they observed might have been a result of increased SNRs during level normalization.

Lengthening phoneme durations to decrease the speaking rate has also not improved intelligibility [73, 105]. Uchanski *et al.* analyzed the length of phoneme durations in CLR speech and non-uniformly lengthened the phoneme durations of CNV speech to match the durations in CLR speech, using a segment-by-segment time-scaling method [105]. Intelligibility of the slowed CNV speech was worse than the original CNV speech. The researchers cited degradations introduced by the signal processing (signal-processing artifacts) as the probable reason for the failure of this approach.

From prior studies on the importance of the temporal envelope for the intelligibility of manner of articulation and voicing [e. g. 25], Narne and Vanaja amplified the depth of modulation of the speech envelope by 15 dB [71]. The results showed the envelope-enhanced speech had improved consonant identification rates by listeners with auditory neuropathy. Another study by Kusumoto *et al.* [57] showed that modulation enhancement of the temporal envelope from 1 to 16 Hz improved consonant recognition rates by 6 percentage points in a reverberant condition with normal-hearing listeners. It is important to note that these improvements were either on CV syllables or on CVC words. As Hazan and Simpson [39] pointed out, a successful speech modification technique for word recognition does not necessarily transfer to improvements in sentence recognition.

In summary, a very limited number of studies [e. g. 39] have succeeded at modifying speech to improve intelligibility at the sentence level. One of the possible reasons for negative results may be that one aspect of the speech signal (e. g. duration or energy) was modified and other features remained intact. It may be necessary to modify sets of features that interact with each other. Also, the degree of modification may not have been natural in some studies. Finally, the speech modification process itself may have caused

unnaturalness. In this thesis, extracting feature(s) from naturally-spoken CLR speech and modifying the corresponding features of CNV speech in a controlled manner will allow evaluation of the impact of a set of features. A successful modification will have potential applications in novel signal processing algorithms for hearing aids and other assistive listening devices, for post-processing speech output from general-purpose communication devices, such as telephones and video playback devices, and for objective measures of speech intelligibility.

2.5 Conclusion

Our hypothesis is that if one (or a combination) of acoustic features contribute to increased speech intelligibility, appropriate modification of those features in CNV speech will improve speech intelligibility. Therefore, we have listed prior research on acoustic features that may contribute to speech intelligibility. To summarize the results of this prior work, it is plausible that formant transitions, temporal envelope, F1 and F2 ranges, formant bandwidth, and VOT contribute to increased speech intelligibility.

In this thesis, we examine (1) which specific acoustic features are responsible for the increased speech intelligibility of CLR speech, and (2) if it is possible to modify these features to improve speech intelligibility (Specific Aim 1). In the future, it will be worth investigating how the degree of contribution from those features might vary depending on a listener's age and hearing status (See 8.4.3). For example, for elderly listeners who might have temporal processing deficits, temporal (or prosodic) features may be required for increased intelligibility.

The reason why the majority of prior signal processing techniques have not been successful may be because of (1) signal-processing artifacts, which introduce clicks, noise, or distorted sounds, or (2) only one feature was modified even though one feature may have interacted with other features. Therefore, it is necessary to minimize the effects of signal-processing artifacts. To address the second reason, a combination of features were modified for an improved intelligibility. In this thesis features are taken directly from naturally spoken speech, either from CNV or CLR speech, to eliminate the possibility that the degree of modification exceeds that of typical human speech production.

Chapter 3

The importance of spectral and prosodic features to sentence intelligibility

3.1 Introduction

A "hybridizaton" algorithm is proposed to approach our specific aim 1, which is to determine which acoustic features of CLR speech are relevant to speech intelligibility. The hybridization algorithm replaces a single feature or a combination of features of CNV speech with those of CLR speech. "Hybridized" (HYB) speech is the synthesized speech whose features consist of both CNV and CLR features. By examining the intelligibility of HYB speech, it is possible to determine whether the specified CLR features contribute to improved intelligibility of CNV speech.

In this chapter[1], Experiments 3–1 through 3–3 are conducted to verify that (1) HYB speech can have improved intelligibility over the baseline (CNV speech) and (2) it is possible to create HYB speech without introducing major artifacts, by examining HYB speech quality. We build tree structures, which contain a set of CLR speech features that constitute part of the HYB feature set. The rest of the HYB feature set are obtained from CNV features. Features at each node are tested to determine if those CLR features are relevant or not. Figure 3.1 shows an example of a tree structure with acoustic feature types at each node. If one set of features (e. g. spectrum) at one node yields significant improvement over CNV speech, the feature set would be split further (e. g. formant and

[1]Part of this chapter was published in Kusumoto *et al.* and Kain *et al.* [58, 48].

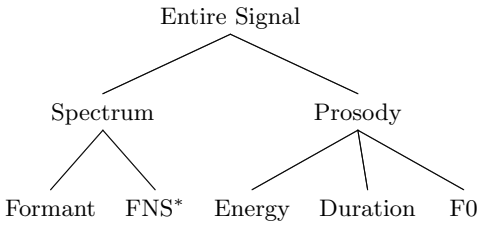


Figure 3.1: Suggested acoustic features and tree structure. (*FNS is formant-normalized spectrum.)

formant-normalized spectrum) to identify the specific aspects of that feature which contribute to improved intelligibility. If no improvement is observed, then the set of features would be repartitioned (e. g. spectrum plus duration).

In a preliminary experiment, we tested the importance of the prosodic group of features (**E**nergy, **F**0, **D**uration, **N**on-speech) and the spectral group (**S**pectrum, **P**honeme sequence) for elderly subjects (age 66–75) with hearing sensitivity less than 35 dBHL at frequency ranges from 250–4000 Hz. The intelligibility of CLR speech was shown to be significantly better than that of CNV speech, however the synthesized speech (or HYB speech) did not yield improvement with either feature set (prosodic or spectral). Our hypotheses for the negative results in HYB speech were (**H1**) phoneme duration in the prosodic feature group can not be separated from the spectral feature group without a negative impact on intelligibility and (**H2**) speech processing artifacts of the HYB speech degraded the speech signal. In this chapter, we focus on determining relevant features for young (18–40) subjects instead of elderly subjects. We address the two hypotheses in Experiments 3–1 through 3–3.

3.2 Text materials and recording

We used seventy syntactically and semantically valid sentences in our experiments [87] (i.e. "*His shirt was clean but one button was gone*") (listed in Appendix A). Each sentence contains five keywords for scoring (underlined in the example above). Four keywords out of five are monosyllables and one is a disyllable. Each sentence is relatively short, which reduces the possibility of testing the effect of memory. The manipulation of prosodic

cues in speech production is facilitated by using sentence material as opposed to isolated syllables and words.

One male, who is a native speaker of American English with no professional training in public speaking, was recruited as a speaker for our experiments. First, he recorded 70 sentences spoken in the CNV speaking style, followed by the same 70 sentences spoken in the CLR speaking style. For the CNV speech, he was instructed to recite the text materials in a way that he uses to communicate in his daily life. When recording CLR speech, he was instructed to speak clearly, as he would when communicating with elderly listeners or hearing-impaired listeners.

The recording was completed in a sound-treated booth (*Whisperroom*) located inside a control room. An administrator monitored the recording in the control room to ensure that the speaker pronounced each sentence correctly. Two display monitors were used, one inside the booth for the speaker and one in the control room for the administrator. The text of each sentence was displayed simultaneously on both monitors. Recordings were made using a head-mounted, close-talking microphone (AKG HSC200), positioned approximately $5cm$ and off-axis from the speakers mouth. The speaker recorded the materials at his own pace by clicking "record", "stop", and "next" buttons on the monitor with the mouse. The speech signals were recorded digitally at a sampling rate of 44.1 kHz with 16 bit resolution, and then downsampled to 22.05 kHz.

3.3 Hybridization algorithm

The hybridization algorithm is a digital signal processing technique that replaces certain acoustic features from CNV speech with those of CLR speech. In this series of experiments, the algorithm replaces the following acoustic features: fundamental frequency (F), long-term energy (E), phoneme duration (D), spectrum (S), phoneme sequence (P), and non-speech events such as pauses (N). A HYB "configuration" (described in Table 3.2) indicates the source of acoustic features in each kind of HYB speech (either from CNV or CLR speech).

Figure 3.2 represents the hybridization algorithm (as an example with the HYB-D condition), which consists of several stages: preparation of the database (Stages: 1–5), feature extraction (Stage: 6), HYB configuration (Stage: 7), and waveform synthesis (Stage: 8).

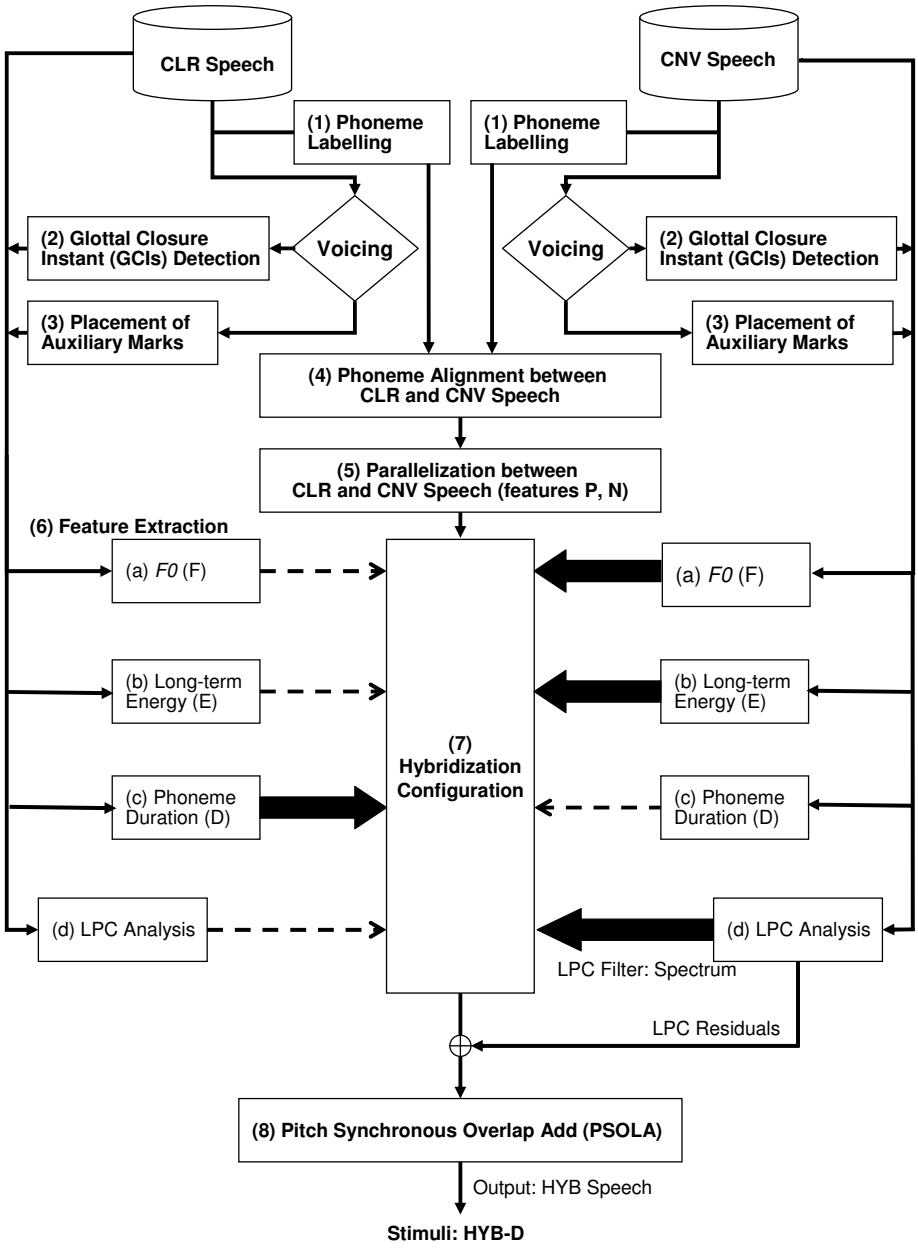


Figure 3.2: An example of hybridization algorithm.

3.3.1 Stage 1: Phoneme labeling

Initial phoneme identities and the time points at phoneme boundaries are obtained using an existing forced-alignment system [44]. The system works by converting the text material into phoneme pronunciations with choices of substitutions, insertions, and deletions (e. g. no word-final burst release /t/) using a custom pronunciation dictionary. An automatic speech recognition (ASR) system recognizes the phoneme sequence with the constraint of the given pronunciation of the sentence. The output of the ASR system provides phoneme identities and the time points of the phoneme boundaries. A trained labeler checks and adjusts the phoneme identities and boundaries manually.

3.3.2 Stage 2: Glottal closure instants (GCIs) detection

The time points when a speaker's vocal folds become fully closed, which are called glottal closure instants (GCIs), are necessary for the pitch-synchronous signal processing. The "Praat" software package is used to automatically detect GCIs [10]. Then, a trained labeler checks, adds, deletes, and adjusts GCIs to ensure maximum correctness of the locations of GCIs. The distance between two GCIs also yields the rate of vocal-fold vibration, known as fundamental frequency (F0).

3.3.3 Stage 3: Placement of auxiliary marks

The hybridization algorithm is a pitch-synchronous frame-by-frame processing algorithm. Analysis frames span three neighboring GCIs (in voiced regions) or auxiliary marks (in unvoiced regions). For the decision of whether a region is voiced or unvoiced, we use the distance between two consecutive GCIs. If the distance is greater than 16 ms (62.5 Hz), the region is considered unvoiced. The auxiliary marks are placed approximately every 10 ms over the entire unvoiced region, adjusting slightly so that two consecutive marks are not closer than 10 ms to each other.

3.3.4 Stage 4: Phoneme alignment between CNV and CLR speech

The phoneme sequences from CLR speech are often different from CNV speech, even when the speaker read the same text material. Most commonly observed differences are (1) the burst tends to be released at the end of the word in CLR speech (e. g. /t/); and (2) tense vowels are reduced in CNV speech (e. g. /iː/ to /ɪ/). The hybridization algorithm requires an identical phoneme sequence between CNV and CLR speech. Therefore, we insert, delete

Table 3.1: An example of the phoneme alignment operation and corresponding parallelization for a HYB-P configuration. The first two columns represent the phoneme sequence of CNV and CLR speech. The third column represents whether the phoneme or pause is inserted or deleted when the HYB configuration is Phoneme=CLR and Non-speech=CNV. In this example, while the plosive closure /d$^{(\urcorner)}$/ is an exact match, CLR plosive release /d$^{(\ulcorner)}$/ is inserted into the CNV speech.

CNV	**CLR**	**Operation**	**HYB**
b	b	-	b
ɪ	i	-	ɪ ; i
s	s	-	s
aɪ	aɪ	-	aɪ
d$^{(\urcorner)}$	d$^{(\urcorner)}$	-	d$^{(\urcorner)}$
-	d$^{(\ulcorner)}$	insertion	d$^{(\ulcorner)}$
-	(.)	deletion	-

or substitute appropriate phonemes into CNV or CLR speech according to the phoneme sequence (P) feature in the HYB configuration. At this stage, only a text operation is carried out, with no speech waveform manipulated.

We use the Dynamic Time Warping (DTW) algorithm to align the phoneme identities between CNV and CLR speech. In the DTW algorithm, the physiological aspects of each phoneme are represented using a phonetic feature vector, with the scale 1–10 for manner, 1–8 for place, 1–10 for height of articulation, and 1–5 for voicing (shown in Table B.1, Appendix B). After the Euclidean distance, in this four-dimensional feature space, between CNV and CLR feature vectors is obtained, the optimal alignment is determined by minimizing the cumulative Euclidean distance. As a result of the alignment, each phoneme in one speaking style is aligned with one of the phonemes in the other speaking style, either by exact match or substitution (one-to-one mapping). If no one-to-one mapping exists, then the phoneme is assigned to a deletion or insertion. A trained labeler checks and adjusts the final alignment to ensure maximum correctness.

The phoneme sequence (P) is one of the features specified in the HYB configuration (Stage 7). Non-speech events (N) are treated separately from the phoneme sequence. Based on the phoneme and non-speech sequence of a HYB sentence, an insertion or deletion operation in each speaking style may take place. The example of HYB-P operation on the word "beside" is aligned with CNV and CLR phoneme sequences in Table 3.1. In the table, phoneme sequences of the CNV and the CLR speech are shown in the first two columns. The operation (ins/del) in the third column is identified by aligning the two phoneme

sequences. The configuration HYB-P is the case where the feature P is taken from CLR and non-speech events (pause) from CNV speech, which results in HYB phoneme sequence as shown in the fourth column.

3.3.5 Stage 5: Parallelization of original waveforms of CNV and CLR speech

Given the HYB phoneme sequence from Stage 4, the "parallelization" of CNV and CLR speech waveforms is carried out by inserting or deleting the portion of waveform that contains the corresponding phoneme. For example, HYB-P shown in Table 3.1, the waveform of /d$^{(r)}$/ is copied from CLR speech, and inserted after the closure of the burst /d$^{(\neg)}$/, when the CLR-P configuration is targeted. In this stage, hybridization of phoneme sequence (P) and non-speech features (N) are completed.

3.3.6 Stage 6: Feature extraction

Acoustic features from CNV and CLR speech are extracted individually. The following six features, tested in Experiments 3–1 through 3–3, are extracted as follows: **(a) Long-term Energy** (E): After an A-weighted filter [46], root-mean-square (rmsA) values are calculated at each analysis frame. The energy contour is smoothed over time by first taking a median filter over 10 frames and then taking a Hanning window (5 frames) centered at the center pitch mark in each analysis frame. **(b) Fundamental frequency (F0)** (F): F0 values are calculated by inverting the distance between two consecutive GCIs, which exist only in voiced regions. No F0 values exist when the auxiliary marks are placed (unvoiced segment). **(c) Phoneme Duration** (D): The phoneme and pause durations are obtained from Stage 1 (phonemic labeling) as described above. **(d) Spectrum** (S): Linear Predictive Coding (LPC) is a low dimensional approximation of the speech signal that represents the spectral envelope associated with formant information. After pre-emphasis (factor of 0.98) on the waveform in one frame, the speech waveform is analyzed by LPC with order 24. The LPC residuals are used to excite the LPC filter in the process of waveform synthesis (Stage 8 described below). **(e) Phoneme sequence** (P) and **(d) Non-speech** (N): The phoneme sequence and non-speech features are determined at Stage 4. The waveform operation is completed at Stage 5.

Table 3.2: HYB configurations indicating the source of six acoustic features either from CLR or CNV speech. Experiments 3–1 through 3–3 are conducted testing eight HYB conditions. Original CNV and CLR speech are included in all experiments.

Exp.	Conditions	Energy	F0	Duration	Spectrum	Phoneme	Non-speech
	CNV	CNV	CNV	CNV	CNV	CNV	CNV
1.	HYB-EFN	CLR	CLR	CNV	CNV	CNV	CLR
1. & 2.	HYB-DSP	CNV	CNV	CLR	CLR	CLR	CNV
2.	HYB-D	CNV	CNV	CLR	CNV	CNV	CNV
2.	HYB-SP	CNV	CNV	CNV	CLR	CLR	CNV
3.	HYB-P	CNV	CNV	CNV	CNV	CLR	CNV
3.	HYB-S	CNV	CNV	CNV	CLR	CNV	CNV
3.	HYB-DS	CNV	CNV	CLR	CLR	CNV	CNV
3.	HYB-EFPN	CLR	CLR	CNV	CNV	CLR	CLR
	CLR	CLR	CLR	CLR	CLR	CLR	CLR

3.3.7 Stage 7: HYB configuration

In preparation for waveform synthesis, sets of HYB features are taken from particular subsets of CLR features and from the complementary subset of CNV features. Table 3.2 shows hybridization configurations used in the experiments and indicates the source of features in the eight conditions of HYB speech. The features include energy (E), F0 (F), duration (D), spectrum (S), phoneme sequence (P) and non-speech events such as pause or breath noise (N). The feature P determines insertion or deletion of a phoneme when the phoneme sequence of CLR speech does not match that of CNV speech. Likewise, the feature N determines insertion or deletion of non-speech events when CLR speech has pauses more or less often than in CNV speech, leading to insertion and deletion, respectively.

3.3.8 Stage 8: Feature replacement and waveform synthesis

In the final step of the hybridization algorithm, we replace a specified single feature, or a combination of features, of the CNV speech with the same type of features extracted from the CLR speech. A HYB speech waveform is synthesized by residual-excited, linear predictive coefficient (LPC) synthesis using pitch-synchronous, overlap add (PSOLA) similar to the work by Taylor *et al.* [101]. Waveform modifications, according to the specified configurations, are carried out at this stage as follows except for phoneme insertions and deletions (Stage 5).

For the energy modification, a gain factor contour is calculated by the ratio between the desired and the original energy contour. The gain contour is first filtered by a tenth-order median filter and then smoothed using a zero-phase low-pass filter. For the F0

modification, the distances between two consecutive frames are altered to meet the desired F0 values. For the duration modification, each frame is either repeated or deleted to meet the desired phoneme durations.

After applying LPC filters on the windowed LPC residuals (excitation signals of the LPC filters), each frame is overlapped by one pitch period and added together to construct speech waveforms using an asymmetric trapezoidal window.

3.4 Phonetic and acoustic characteristics

In this section, phonetic and acoustic characteristics of our speech corpus are described.

3.4.1 Phonetic characteristics

Phonetic characteristics in CLR speech include vowel modification and phoneme insertion (or deletion). The results of phoneme alignment between CNV and CLR speech (Stage 4 in Section 3.3.4) showed 69 labels from CLR speech (including 14 non-speech events (N) and 22 unvoiced burst releases) being inserted into CNV speech, while 7 labels from CNV speech (including 3 non-speech events) being inserted into CLR speech. The mean occurrence of non-speech events is 0.1714 and 0.3286 per sentence for CNV and CLR speech, respectively. Phoneme substitutions (e. g. /iː/ to /ɪ/) from CLR to CNV speech occurred 104 times out of 2175 labels in CNV and 2237 labels in CLR speech.

3.4.2 Acoustic characteristics

Table 3.3 shows a summary of the values of acoustic features (phoneme duration, fundamental frequency, energy, and long-term average spectrum). The mean of each feature over 70 sentences, the standard deviation, p values, and Cohen's d effect size are reported.

Phoneme duration: Total sentence duration, total vowel duration, duration of the last vowel in the sentence, total consonant duration, the longest vowel duration, the longest consonant duration, mean stop burst duration (/b/, /d/, /g/, /p/, /t/, and /k/), and total pause duration (non-speech event) are measured per sentence.

Fundamental frequency (F0): The F0 values (in Bark) are averaged over all vowel regions. The range is obtained by taking the difference between maximum and minimum F0 values over the entire sentence.

Table 3.3: Summary of the values of acoustic features of CLR and CNV speech. Mean, standard deviation (in parentheses) over 70 sentences, p values, and Cohen's d effect size are shown. Degree of freedom df are all equal to 69. Asterisks are shown for significance.

Num	Feature	CNV	CLR	p-values	Cohen's d
1	Total dur. $(sec)^*$	1.7720 (0.1793)	2.2997 (0.2585)	< 0.0001	2.3722
2	Total V dur. $(sec)^*$	0.6160 (0.1022)	0.8140 (0.1522)	< 0.0001	1.5274
3	Last V dur. (sec)	0.1185 (0.0475)	0.1329 (0.0564)	< 0.0001	0.2762
4	Total C dur. $(sec)^*$	1.1560 (0.1676)	1.4857 (0.2235)	< 0.0001	1.6691
5	Longest V dur. $(sec)^*$	0.1460 (0.0355)	0.1872 (0.0489)	< 0.0001	0.9642
6	Longest C dur. $(sec)^*$	0.1319 (0.0290)	0.1607 (0.0343)	< 0.0001	0.9068
7	Stop burst dur. $(sec)^*$	0.0282 (0.0100)	0.0335 (0.0124)	< 0.0001	0.4705
8	Burst count	3.9571 (1.6805)	4.4571 (1.7749)	< 0.0001	0.2893
9	Total pause dur. $(sec)^*$	0.0054 (0.0148)	0.0165 (0.0301)	0.0016	0.4680
10	Pause count*	0.1714 (0.4160)	0.3286 (0.5028)	0.0038	0.3407
11	Vowel F0 $(Bark)^*$	1.0250 (0.0548)	1.0469 (0.0494)	0.0009	0.4198
12	Vowel F0 range $(Bark)^*$	0.5067 (0.1211)	0.6835 (0.1952)	< 0.0001	1.0885
13	Vowel energy range (rms)	0.0935 (0.0194)	0.1002 (0.0216)	0.0164	0.3264
14	Cons. energy range (rms)	0.1126 (0.0201)	0.1058 (0.0240)	0.0048	0.3072
15	CV energy ratio (dB)	-10.8283 (3.1076)	-11.7212 (2.8478)	0.0168	0.2996
16	LTAS 500−3000 Hz (dB)	70.8501 (4.3108)	71.7552 (5.2317)	0.0969	–
17	LTAS 1000−2000 Hz (dB)	54.5146 (3.5854)	55.0141 (4.3958)	0.2746	–
18	LTAS 500−2000 Hz (dB)	69.1079 (4.3845)	69.8638 (5.3574)	0.1650	–
19	LTAS 1000−3000 Hz (dB)	58.0040 (3.6378)	58.9866 (4.3979)	0.0356	0.2435

Energy: The root-mean square (RMS) energy for vowels and consonants in energy-normalized sentences are examined. The RMS energy range is computed by taking the difference between maximum and minimum RMS energy values. The (CV) energy ratios are calculated by dividing the RMS energy of the consonants /b/, /d/, /g/, /p/, /t/, /k/, /f/, /v/, /s/, /z/, /m/, and /n/ by the RMS energy of the following vowel and converting to the dB scale [39].

Spectrum: The long-term average spectrum (LTAS) of each energy-normalized sentence is calculated in four frequency bands (500–3000 Hz; 1000–2000 Hz; 500–2000 Hz; 1000–3000 Hz), measured in dB [38].

Formant frequency: First and second formant trajectories and formant bandwidths are extracted using the Snack Sound Toolkit (`http://www.speech.kth.se/snack` [91]) in the vowel regions. The formant values are taken from the middle of the vowel. Formant information is measured for the following 37 features: the mean F1 and mean F2 values of seven vowels /iː/, /ɪ/, /u/, /ɛ/, /æ/, /ʌ/, and /ɑ/ and corresponding bandwidths (BW), and the mean distance between F1 and F2 frequencies of these seven vowels. The mean distance between F1 of /iː/ and F1 of /æ/ (for

Table 3.4: Summary of the values of formant frequencies of CNV and CLR speech. Formant frequencies are converted to Bark scale, while bandwidths are measured in Hz. Mean, standard deviation (in parentheses) over 70 sentences, p values (degrees of freedom), and Cohen's d effect size are shown. Asterisks are shown for significance.

Num	Feature	CNV	CLR	p-values (df)	Cohen's d
20	F1 range*	1.8663 (0.6062)	2.4190 (0.3942)	0.0007 (14)	1.0810
21	F2 range	3.5123 (2.2350)	3.1944 (1.7455)	0.5181(6)	–
22	F1−F2 dist. /iː/*	9.5385 (0.9352)	9.9292 (0.8162)	0.0060 (41)	0.4451
23	F1−F2 dist. /ɪ/	7.8363 (1.2866)	7.6693 (1.0688)	0.9363(40)	–
24	F1−F2 dist. /u/	7.4171 (1.9794)	7.0163 (1.7491)	0.7627(13)	–
25	F1−F2 dist. /ɛ/	6.5741 (1.1287)	6.1749 (1.0035)	0.0061 (35)	0.3738
26	F1−F2 dist. /ə/	5.5591 (1.2283)	5.4430 (1.2161)	0.1084(22)	–
27	F1−F2 dist. /ʌ/*	5.6557 (0.8945)	5.1884 (0.9628)	0.0002 (57)	0.5029
28	F1−F2 dist. /ɑ/*	3.8370 (0.6642)	3.1374 (0.5333)	< 0.0001 (35)	1.1615
29	F1 /iː/	3.4364 (0.3645)	3.4975 (0.4197)	0.3341(41)	–
30	F2 /iː/*	12.9749 (0.7366)	13.4267 (0.5277)	< 0.0001 (41)	0.7051
31	F1 /ɪ/*	3.7207 (0.3695)	4.0441 (0.3097)	< 0.0001 (40)	0.9486
32	F2 /ɪ/	11.5570 (1.1566)	11.7134 (1.0267)	0.0113 (40)	–
33	F1 /u/	3.3978 (0.2470)	3.4603 (0.1666)	0.3892(13)	–
34	F2 /u/	10.8149 (1.9994)	10.4766 (1.7975)	0.6782(13)	–
35	F1 /ɛ/*	4.3698 (0.6379)	4.9770 (0.5488)	< 0.0001 (35)	1.0205
36	F2 /ɛ/	10.9439 (0.8791)	11.1518 (0.7706)	0.0131 (35)	0.2515
37	F1 /ə/*	5.2538 (0.8072)	5.7584 (0.5256)	0.0011 (22)	0.7408
38	F2 /ə/	10.8129 (0.9509)	11.2014 (1.0150)	0.0411 (22)	0.3950
39	F1 /ʌ/*	4.0082 (0.5535)	4.2845 (0.3871)	0.0002 (57)	0.5785
40	F2 /ʌ/	9.6640 (0.7035)	9.4730 (0.8073)	0.0725(57)	–
41	F1 /ɑ/*	4.6753 (0.5759)	5.3405 (0.4087)	< 0.0001 (35)	1.3321
42	F2 /ɑ/	8.5123 (0.4450)	8.4778 (0.5744)	0.6917(35)	–
43	F1 BW /iː/	53.8484 (39.5396)	51.0692 (52.3017)	0.4989(41)	–
44	F2 BW /iː/	355.1191 (284.6750)	368.3416 (378.6380)	0.7362(41)	–
45	F1 BW /ɪ/	86.7463 (55.6818)	77.2378 (52.6412)	0.0774(40)	–
46	F2 BW /ɪ/	399.1492 (181.7969)	422.0937 (343.5215)	0.9147(40)	–
47	F1 BW /u/	45.6365 (26.0220)	49.6726 (22.1960)	0.7583(13)	–
48	F2 BW /u/	404.3236 (137.1914)	517.7733 (201.7040)	0.1028(13)	–
49	F1 BW /ɛ/	123.4645 (77.2469)	137.9733 (96.1892)	0.5785(35)	–
50	F2 BW /ɛ/	436.4626 (181.7730)	431.3602 (433.6370)	0.9155(35)	–
51	F1 BW /ə/	222.2775 (86.9371)	228.6250 (106.2508)	0.1558(22)	–
52	F2 BW /ə/	458.0514 (162.3406)	470.3792 (343.2922)	0.8346(22)	–
53	F1 BW /ʌ/	122.9606 (84.3485)	96.8572 (65.4114)	0.1561(57)	–
54	F2 BW /ʌ/	470.4471 (452.3613)	405.9840 (255.6977)	0.3822(57)	–
55	F1 BW /ɑ/	208.7103 (141.2040)	177.3777 (90.9007)	0.1201(35)	–
56	F2 BW /ɑ/	207.5404 (71.4986)	187.4356 (152.2285)	0.3702(35)	–

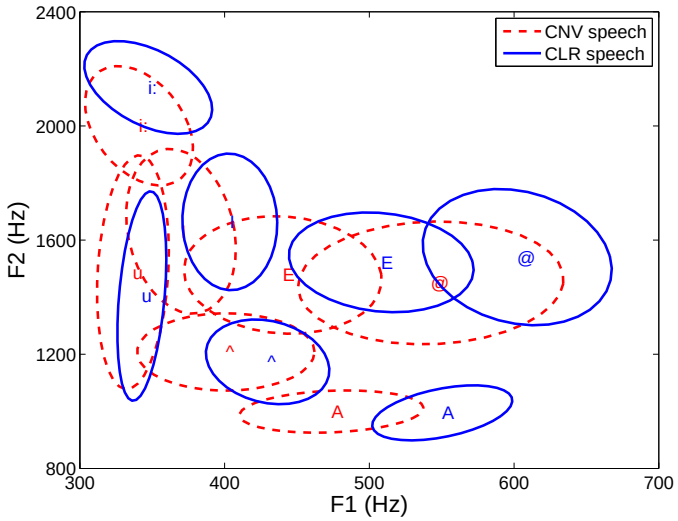


Figure 3.3: Formant frequency of 8 vowels in CNV and CLR speaking styles with one standard deviation in F1–F2 space.

the F1 range), and the mean distance between F2 of /u/ and F2 of /iː/ (for the F2 range) were included in order to estimate the vowel space. Figure 3.3 shows the formant frequencies of each vowel with ±1 standard deviation in F1–F2 space. Formant values were not manually corrected, and may contain formant tracking errors. Table 3.4 shows the summary of formant frequencies of CLR and CNV speech. The mean of each feature over 70 sentences, standard deviation, p values (degree of freedom), and Cohen's d effect size are reported.

This particular speaker successfully made both phonetic and prosodic changes (phoneme duration, F0, and energy) in his CLR speech production. However, spectral differences were focused on particular vowels (i. e. /ɛ/, /ɑ/), and formant bandwidths do not appear to be different between CLR and CNV speech. LTAS had increased energies at higher frequencies (1000–3000 Hz), which was consistent with the findings of Krause and Braida's study [55].

3.5 Perceptual experiment

3.5.1 Normalizing energy of speech and noise

It is important to eliminate the possibility that the increased intelligibility of CLR speech is solely due to an overall energy increase as compared to CNV speech. Therefore, in our speech recordings, the long-term energy of all sentences in the recordings was normalized

digitally. We calculated root-mean-square values of non-pausal portions of the sentences after A-weighted filtering (rmsA) [46]. We equalized the rmsA values digitally for all sentences by applying individual gain factors to each sentence, while keeping the global peak value within a threshold value (80 % of the quantization limit) to avoid peak clipping. In our experiments, the stimuli were presented in a 12-talker babble noise, which is an uncorrelated accumulation of speech from 12 talkers to simulate an everyday listening environment. The 12-talker babble noise has similar frequency characteristics as our test stimuli, and is widely used in tests that measure speech intelligibility. It was originally developed for the test of speech perception in noise (SPIN) [9]. The level of the noise relative to the level of the signal is defined as the signal-to-noise ratio (SNR).

3.5.2 Obtaining SNR–50 level

The word recognition test was carried out under the condition of added noise. The level of noise is set at the SNR–50 threshold level for each subject, to minimize between-subject variability. SNR–50 is that signal-to-noise ratio (SNR) level at which the subject can identify test sentences in the presence of noise 50 % of the time. The administrator controls a program, written in Matlab, for the SNR–50 test. The subject's task is to repeat the sentence aloud to the administrator immediately after they listen to a sentence. The administrator determines whether the subject has a positive or negative response for each sentence. The response is counted as positive when the subject can repeat four or more keywords (out of five) correctly. The response is counted as negative when the subject can repeat less than four keywords.

In the adaptive procedure [62], the SNR level is set to -3 dB as a start point. The first sentence is repeated until the subject can obtain a positive response, by increasing SNR levels (the noise level decreases). After the first positive response, a different sentence is presented to the subject each time. The noise level is increased (SNR decreases) when the response is positive, and the noise level is decreased (SNR increases) when negative. When the SNR is increased or decreased based on the subject's response (e.g. a positive response followed by a negative response), the count of "reversal" is incremented. The increment or decrement of the SNR starts with a 2 dB step size, and after 3 reversals the step size is decreased to 1 dB. The test is continued until 8 reversals are accomplished. The final SNR–50 level is estimated by averaging the SNR from the 3rd reversal to 8th reversal points. Twenty-two sentences are available to obtain the SNR–50 for each subject. The pool of sentences used in SNR–50 level was different from the intelligibility experiments

Table 3.5: Average SNR–50 with standard deviations in parentheses obtained in Experiments 3–1, 3–2, and 3–3.

Experiments	**Subjects' age**	**SNR–50 in dB**
Verification–CNV	22–33	−2.52 (0.84)
Verification–CLR	22–33	−4.26 (1.25)
3–1. CNV	21–39	0.58 (0.88)
3–2. CNV	19–39	−0.24 (1.11)
3–3. CNV	18–35	0.22 (1.44)

(shown in Appendix A).

3.5.3 Speech corpus verification

The goal of the first verification experiment was to verify that all CNV sentences are intelligible in the absence of background noise (e. g. for example, to find any mispronunciation). Five young listeners (mean age: 28.5) listened to all 70 sentences in the CNV style. If the subject responded with five out of five keywords correctly, the sentence was recored as positive. The mean sentence intelligibility rate was 98.57 %. It confirmed that the CNV sentences are intelligible. It was assumed that the intelligibility of CLR sentences would be even higher than CNV sentence intelligibility.

In the second verification experiment, we compared the intelligibility differences between CNV and CLR speech under the noise condition. The purpose of this experiment was to examine whether CNV and CLR speech have inherent intelligibility differences. Eight young listeners (mean age: 25.1) listened to sentences both in CNV and CLR speaking styles. SNR–50 values were obtained in both styles (Section 3.5.2). None of the sentences were presented more than once to the same subject. A sentence was marked correct if the subject correctly identified four out of five key words. The results from this verification experiment show SNR–50 levels were −2.52 (0.84) and −4.26 (1.25) for CNV and CLR speech, respectively. There were significant (α =0.05) main effects of speaking style ($F(1, 28)$ =7.32, p =0.012). This confirmed that the speech corpus reflects inherent differences between CNV and CLR speech.

3.6 Experiment 3–1: The effects of duration and spectral features from CLR speech

As discussed in Section 3.1, we conducted a preliminary experiment with elderly listeners and "prosodic" and "spectral" feature groups. Results from this experiment did not show any improvement over the baseline CNV speech. We developed two hypotheses to explain these results. (**H1**) phoneme duration in the prosodic feature group can not be separated from the spectral feature group without a negative impact on intelligibility and (**H2**) speech processing artifacts of the HYB speech degraded the speech signal. In Experiment 3–1, we examined (**H1**); whether "phoneme duration, in the prosodic feature group, can not be separated from the spectral feature group" by evaluating the speech intelligibility of HYB speech taking both phoneme duration and spectral features from CLR speech. Eight subjects (2 females and 6 males) who were in the age range 23–40 (mean: 28.38) participated in Experiment 3–1. All of them had normal hearing (self-reported) and were native speakers of American English.

3.6.1 Procedures and apparatus

The experiments were carried out on a portable PC (VIA Samuel 2, 599 MHz, 224 MB of RAM) and the sound was produced by a high-quality sound device (M-Audio USB Duo). Each subject listened to the stimuli binaurally through a pair of circumaural headphones (Sennheiser 280 Pro) in a quiet room. For the SNR–50 test, the energy of the noise was adjusted according to the desired SNR level. For the intelligibility experiments, the noise level was set at each subject's SNR–50 level. The energy of the sentence was acoustically calibrated at 65 dBA throughout the experiments. The administrator proceeded with the experiment, considering the subjects pace, by controlling a program written in Matlab to present the stimuli. After obtaining the SNR–50 level, the intelligibility experiments were carried out. Subjects were instructed to repeat the sentence aloud as best as they could after they listened to each sentence. They were informed that they were going to listen to semantically valid sentences, with given sample sentences. They were encouraged to guess when unsure or when they could not make out the meaning of the sentence.

Forty-eight sentences out of 70 were presented in a Latin square design in the intelligibility experiments. Each subject heard twelve sentences per condition, and none of the sentences were repeated. All four conditions were counterbalanced for each subject so that each condition was heard an equal number of times by the end of the experiment.

The order of the conditions was randomized, while the sentence order was kept same for all of the subjects.

3.6.2 Stimuli

The SNR–50 noise level was estimated for each subject to account for between-subject differences, and the estimated noise level was used for the intelligibility experiments (Section 3.5.2). For setting SNR–50, the stimuli were the original CNV speech without any modification. A total of 22 sentences were available for SNR–50 testing.

For the intelligibility experiments, four conditions of stimuli were tested in Experiment 3–1. The conditions included original CNV and CLR speech, and two HYB speech configurations, namely HYB-EFN and HYB-DSP, in order to examine the importance of matching the source of phoneme duration and the source of spectral features (Table 3.2). The first HYB speech configuration, HYB-EFN (**E**nergy, **F**0, **N**on-speech) replaced energy, F0, and non-speech features from CNV speech with those from CLR speech. The second HYB speech configuration, HYB-DSP (**D**uration, **S**pectrum, **P**honeme) replaced phoneme duration, spectrum and phoneme sequence from CNV speech with those from CLR speech.

3.6.3 Results and discussions

In Experiment 3–1, the average SNR–50 level was 0.58 dB (standard deviation: 0.88) (Table 3.5). Figure 3.4 shows the mean intelligibility rates in percent and standard deviations for the 4 conditions. The mean intelligibility of CNV was 68.75 % (12.4), while that of CLR was 93.75 % (11.57). The HYB speech had 61.46 % (19.89) and 91.67 % (8.91) for HYB-EFN and HYB-DSP, respectively.

The results from the 8 subjects were analyzed for statistical significance using the arc-sine transformation [6] given by the equation,

$$x = arcsin\sqrt{\frac{r + 3/8}{n + 3/4}} \tag{3.1}$$

where r represents the number of sentences a subject identified correctly, and n represents the number of sentences presented. The planned $t-$test showed that CLR speech was significantly better than CNV speech ($t(7) = 5.5795$, $p = 0.0004$) for this particular speaker with young subjects.

The hybridized speech HYB-DSP, consisting of a combination of phoneme duration and spectrum features from CLR speech, yielded significant improvement over the baseline

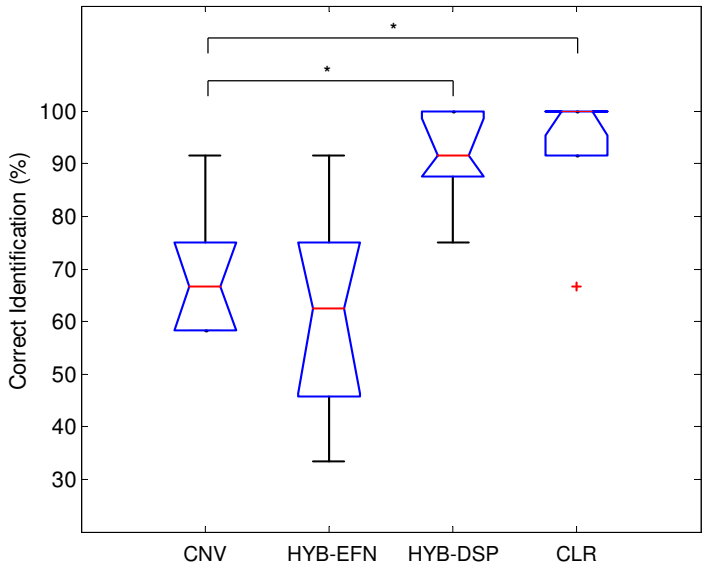


Figure 3.4: Intelligibility rates (in percent) in Experiment 3–1. Significant differences are shown with asterisks (*: $p < 0.05$).

CNV speech ($t(7) = 4.3467$, $p = 0.0017$). Therefore, it supports hypothesis **H1** which is that the source of phoneme duration and spectrum should not be different for improved intelligibility. However, the hypothesis is not confirmed because HYB speech with phoneme duration only and spectrum only were not tested on young subjects. On the other hand, F0 and energy from CLR speech did not help to improve intelligibility ($p = 0.1146$). In Experiment 3–2, the effect of individual features (D and S) is further examined.

3.7 Experiment 3–2: The effects of individual features versus combined features and signal processing artifacts

As a result of Experiment 3–1, a combination of phoneme duration and spectrum from CLR speech yielded a significant improvement in intelligibility over CNV speech. To further test hypothesis **H1**, which is that phoneme durations should not be separated from spectral features in order to maintain high intelligibility, phoneme duration only, spectral

features only, and a combination of both phoneme duration and spectral features from CLR speech were tested. To test hypothesis **H2**, which is that the hybridization algorithm introduces artifacts that cause decreased intelligibility, possible sources of artifacts in the hybridization algorithm were addressed and a new implementation was developed, which is referred to as Implementation 2. The intelligibility and the quality of the HYB speech produced by Implementation 2 compared with HYB speech from Implementation 1.

Twelve subjects (11 females and 1 male) who were in the age range 19–40 (mean: 29.17) participated in Experiment 3–2. For the subject recruitment, the same criteria as in Experiment 3–1 were used.

3.7.1 Implementation 2

We addressed four possible sources of artifacts and approaches to reduce the artifacts for improved speech quality.

1. **Insertion of auxiliary marks in voiced phonemes can cause unnecessary duplication of frames.**
   Auxiliary marks were placed when the distance between GCIs was greater than a minimum threshold (16 ms, corresponding to 62.5 Hz). However, the voiced region may have lower fundamental frequency than this minimum threshold if the speech sound is glottalized. Glottalization is a phenomenon that occurs when the vocal folds vibrate irregularly or at very low frequency (e.g. less than 70 Hz). In addition, errors in the judgment of voicing or misplacement of GCIs can be problematic. Therefore, phoneme identity was used to force voiced phonemes to be excluded from having auxiliary marks. This allowed GCIs located farther apart than 16 ms. When the speech is hybridized, glottalized portions of the phonemes sound glottalized, as in the original.

2. **Duplicating frames that contain bursts, as found in affricates, plosives and flaps, can cause audible artifacts.**
   Duration or F0 modification requires the duplication of frames. Because of the impulse-like changes in energy that occur in bursts, duplicating those regions of phonemes can be problematic. Therefore, the duplication of frames was prevented in bursts such as affricates (/ʧ/), /ʤ/), plosives (/p/, /t/, /k/, /b/, /d/, /k/), and flaps (/ɾt/, /ɾd/, /ɾn/). To simplify the implementation, duration or F0 were not modified in those phonemes.

3. **Duplicating frames in unvoiced-to-voiced transitions may deteriorate the naturalness of phoneme transitions.**
   These transitory frames are unique in terms of energy, pitch period, and formant frequency locations. The duplication of these frames during duration modification or F0 modification may cause unnaturalness in HYB speech waveforms. Therefore, duplication in frames of unvoiced-to-voiced transitions was prevented, and duplication was only performed at phoneme centers.

4. **Phoneme insertions and deletions can cause signal discontinuities.**
   In the step of parallelization (Stage 5 in Section 3.3.5), the waveform of the phoneme was inserted or deleted according to feature P, which can cause signal discontinuity. The signal discontinuity can lead to audible clicks at the concatenation points, which possibly reduces intelligibility and naturalness. Therefore, to smoothly fade in and fade out required waveforms using linearly weighted windows with one pitch period at concatenation points during phoneme insertion and deletion operations.

3.7.2 Quality experiments

In order to test the quality of the signals generated by Implementation 2, perceptual comparisons were made on pairs of segments using comparison-mean-opinion scores (CMOS). CMOS is a scoring scale with five discrete choices that indicate whether the first presented stimulus was (2) much better, (1) slightly better, (0) about the same, (-1) slightly worse, or (-2) much worse. The subjects listened to each pair of stimuli three times in sequence, and were allowed to repeat the stimuli any number of additional times. The administrator conducted the experiment by controlling a program written in Matlab to present stimuli.

3.7.3 Stimuli

For the SNR–50 test, the stimuli were the original CNV speech without any modification (the same as in Experiment 3–1 (Section 3.6.2)).

For the intelligibility experiments, six conditions were tested in Experiment 3–2. The conditions included original CNV and CLR speech, and four HYB speech configurations, namely HYB-DSP$_2$, HYB-D$_2$, HYB-SP$_2$, and HYB-SP (subscript 2 indicates that the signal was processed by Implementation 2). Three out of four sets of the HYB speech were processed using Implementation 2. The first set of HYB speech examined the combined effect of phoneme duration, spectral features, and phoneme sequence from CLR speech

(HYB-DSP$_2$: **D**uration, **S**pectrum, **P**honeme). The second set of HYB speech examined the effects of only phoneme duration from CLR speech (HYB-D$_2$: **D**uration). The third set of HYB speech examined the combination of spectral features and phoneme sequence from CLR speech (HYB-SP$_2$: **S**pectrum, **P**honeme). Lastly, the fourth set of hybrid speech was the same configuration as HYB-SP$_2$, but it was processed using Implementation 1 (HYB-SP). In Experiment 3–2, 48 sentences were tested with a Latin square design (8 sentences per condition).

For the quality experiment, to examine the speech quality with Implementations 1 and 2, fragments of the sentences, such as syllables, words, or phrases, were chosen as stimuli. First, all the segments that were affected by any four modifications in Implementation 2 were selected (Table 3.6). The detected segments included (1) the phonemes (vowels, nasals, flaps, and approximants) that are allowed to have very low F0 due to glottalization in voiced sounds by having pitch-epoch marks farther than the minimum threshold (16 ms), (2) the frames that contain, bursts as found in affricates, plosives and flaps, which are not duplicated in the process of duration or F0 modification, (3) the frames that contain unvoiced-to-voiced transitions, which are not be duplicated in the process of duration or F0 modification, (4) concatenation points where the phonemes are inserted or deleted by using linearly weighted windows during phoneme insertion and deletion operations.

By visual inspection of the spectrogram of the segments and by listening to segments that were processed by Implementation 1, 24 segments were manually selected. Then, the segments processed by both Implementation 1 and 2 were presented as stimuli.

3.7.4 Results and discussions

Intelligibility experiments

The average SNR–50 level was -0.24 dB (1.11) (Table 3.5). Figure 3.5 shows the mean intelligibility rates in percent and standard deviations for the 6 conditions. The intelligibility rates of CNV and CLR were 65.63 % (19.31) and 89.58 % (sd: 10.44), respectively. The intelligibility rates of HYB-DSP$_2$, HYB-D$_2$, HYB-SP$_2$, and HYB-SP were 81.25 %(15.54), 75.00 %(15.99), 76.04 % (12.45), and 58.33 % (24.03), respectively.

The results from the 12 subjects were analyzed for statistical significance using the arcsine transformation as described in Experiment 3–1 (Section 3.8.2). The intelligibility of CLR speech was significantly better than the intelligibility of CNV speech ($t(11) = 5.6336$, $p < 0.0001$).

A paired $t-$test comparison between HYB-DSP$_2$ and CNV ($t(11) = 3.1278$, $p = 0.0048$),

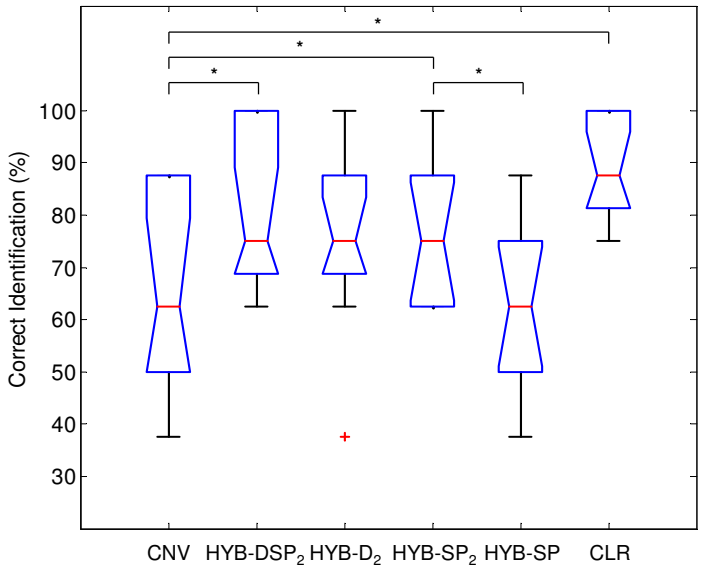


Figure 3.5: Intelligibility rates (in percent) in Experiment 3–2. Significant differences are shown with asterisks (*: $p < 0.05$).

as well as HYB-SP$_2$ and CNV speech ($t(11) = 2.2658$, $p = 0.0223$) showed a significant improvement in intelligibility for both configurations. The other HYB condition, HYB-D$_2$ was not significantly different from CNV speech ($p = 0.139$). A comparison of the Implementations 1 and 2 for one condition, namely between HYB-SP$_2$ and HYB-SP ($t(11) = 3.2895$, $p = 0.0036$), indicated a significant difference in intelligibility with Implementation 2.

The first hypothesis (**H1**) discussed in Section 3.6.3 was that phoneme durations were not independent of spectral features; the source of these two features can not be separated while maintaining intelligibility. The HYB feature set that consists of a combination of features DSP from CLR speech had higher intelligibility than the HYB feature set that consists of features D or SP only, which supports hypothesis **H1**. However, the combination of features SP also yielded significant improvement over CNV speech; therefore, hypothesis **H1** was not completely confirmed. The second hypothesis (**H2**) was that the signal-processing artifacts in HYB speech might have decreased intelligibility. The current results

Table 3.6: Comparison Mean Opinion Score (CMOS) results comparing Implementations 1 and 2. Asterisks are shown for significance (α= 0.05).

Category	Num. of samples	HYB configuration	CMOS
1: No auxiliary marks	6	EFDN	-0.53 (0.82)*
2: No duplication of stops	6	DP	0.56 (1.12)*
3: Unvoiced to voiced transition	6	EFDN	-0.06 (0.80)
4: Phoneme insertions	6	DP	0.40 (0.99)*

showed the intelligibility of HYB-SP$_2$ (processed by Implementation 2) was significantly better than HYB-SP (processed by Implementation 1), which confirms hypothesis **H2**.

In Experiment 2, the HYB-D, HYB-SP, as well as HYB-DSP conditions were evaluated to examine individual effects of duration and spectrum. In the case of HYB-SP and HYB-DSP, it is still not clear whether the improvement of HYB-DSP was more due to a large number of phoneme insertions, which is a characteristic of CLR speech, or the effects of duration and spectrum. To determine the contribution of phoneme insertions, the next experiment was planned for the comparison between HYB-P, HYB-S, HYB-DS, and HYB-EFPN individually.

Quality experiments

Table 3.6 shows the average CMOS results in each category over 12 subjects. The $t-$test was used to examine whether the average CMOS in each category is greater than 0 (which means that Implementation 2 improved the speech quality), or equal to or less than 0 (which means that Implementation 2 did not improve the speech quality).

Category 1 The mean of -0.53 was significantly *lower* than 0 ($t(71) = 5.4498$, $p = 3.45 \times 10^{-7}$). Not allowing very low F0 due to glottalization in the voiced regions led to de-glottalization of the phonemes that were glottalized in the original. On the other hand, allowing very low F0 in the voiced regions could lead to hybridization of glottalized phonemes and non-glottalized phonemes accurately. Even though the de-glottalized phonemes were perceived as sounding slightly better, it is not clear that this preference relates to intelligibility. Using Implementation 2, the glottalized vowels in HYB speech were better controlled in the hybridization algorithm.

Category 2 The mean of 0.56 was significantly higher than 0 ($t(71) = 4.1911$, $p = 3.45 \times 10^{-5}$). Restricting the number of duplicates of the frames that contain bursts in the processing of F0 and duration modifications led to significantly improved speech quality.

Category 3 The mean of -0.06 was not significantly different from 0 ($p = 0.28$). The results showed that restricting the number of duplicates of the frames that contain unvoiced-to-voiced/voiced-to-unvoiced transitions did not make any difference for perceptual preference.

Category 4 The mean of 0.40 was significantly higher than 0 ($t(71) = 3.4589$, $p = 4.60 \times 10^{-4}$). By applying a ramping window in Implementation 2, audible discontinuities were reduced, as compared with Implementation 1.

Overall, in two out of four categories the speech quality was improved in Implementation 2, though one category resulted in reduced quality.

3.8 Experiment 3–3: The effects of phoneme insertions from CLR speech

One of the characteristics of CLR speech is that bursts of stop consonants in a word final position are often released. Inserting /ə/ after voiced consonants is also found in CLR speech [80, 55]. In Experiment 3–3, the contribution of the CLR phoneme sequence was tested. Our hypothesis was that the improvement of HYB-DSP in Experiment 2 was due to the contributions of either duration or spectral features (or both), but not the phoneme insertions, and that the occurrence of phoneme insertions was not frequent enough to provide significant improvement.

Eighteen subjects (10 females and 8 male) who were in the age range 18–35 (mean: 24.06) participated in Experiment 3–3. For the subject recruitment, the same criteria as in Experiments 1 and 2 were used.

3.8.1 Stimuli

For SNR–50, the stimuli were the original CNV speech without any modification, the same as in previous experiments (Section 3.6.2). For the intelligibility experiments, six conditions of stimuli were tested in Experiment 3–3. The conditions included original CNV and CLR speech, and four HYB speech configurations, namely HYB-P_2 (**P**honeme), HYB-S_2 (**S**pectrum), HYB-DS_2 (**D**uration and **S**pectrum), HYB-$EFPN_2$ (**E**nergy, **F**0, **P**honeme, and **N**on-speech).

Previously, Experiment 2 showed the intelligibility of HYB-DSP_2 and HYB-SP_2 were significantly higher than baseline CNV speech. In order to examine the contributions of

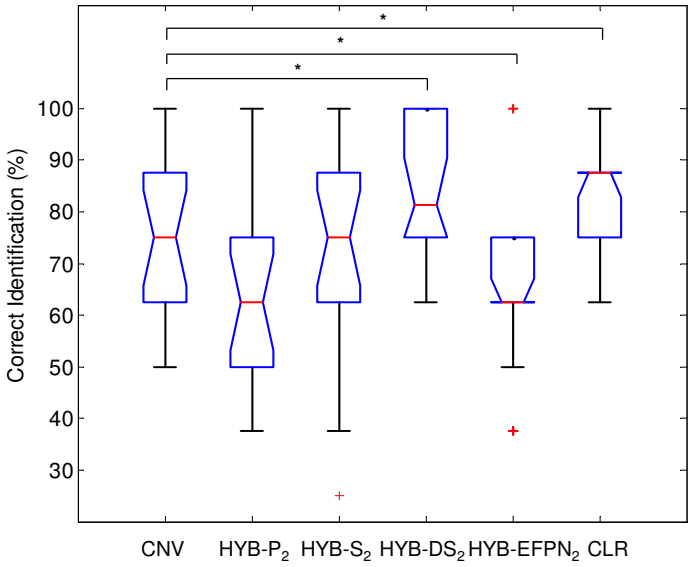


Figure 3.6: Intelligibility rates (in percent) in Experiment 3–3. Significant differences are shown with asterisks (*: $p < 0.05$).

phoneme sequence, the individual features of spectrum and phoneme were tested in HYB-P_2 and HYB-S_2. The combination of HYB-DS_2 was also tested. Experiment 1 showed the combination of features EFN (**E**nergy, **F**0, and **N**on-speech) were not contributing features to improved intelligibility of CLR speech. We again tested this condition by adding phoneme features (P) to the features EFN, to determine whether the phoneme sequence improves HYB-EFN speech intelligibility. For the experiment, 48 sentences were tested with a Latin square design (8 sentences per condition).

3.8.2 Results and discussions

The average SNR–50 was 0.22 dB (1.44) (Table 3.5). Figure 3.6 shows the mean intelligibility rates in percentage and standard deviations for 6 conditions. The mean intelligibility of CNV speech was 72.22 % (14.57), while that of CLR was 84.72 % (sd: 11.79). The intelligibilities of HYB speech were 65.97 % (19.56), 72.92 % (19.76), 81.94 % (14.36), and 64.58 % (15.01) for HYB-P_2, HYB-S_2, HYB-DS_2, HYB-$EFPN_2$, respectively.

The results from the 18 subjects were analyzed for statistical significance using the arc-sine transformation as previously described in Section 3.6.3. The intelligibility of CLR speech was significantly better than the intelligibility of CNV speech ($t(17) = 3.524$, $p < 0.003$).

A paired $t-$test comparison between CNV speech and HYB-P_2, HYB-S_2 revealed neither having a significant difference ($p = 0.40$, and 0.83, respectively). On the other hand, HYB-DS_2 yielded a significant improvement over CNV speech ($p = 0.05$). These results indicate that the contribution of phoneme sequence was not significant. Obtaining a significant improvement in the HYB-DS_2 condition suggested that the previous results from HYB-DSP_2 were most likely due to the contribution of DS, not due to the phoneme sequence. In the case of HYB-$EFPN_2$, the intelligibility was significantly *decreased* from CNV speech ($p < 0.01$). From the results in Experiment 1, which did not show a significant difference for HYB-EFN, by adding the P feature the intelligibility was decreased, despite the improvements of Implementation 2. It may be possible that the phoneme insertion (or deletion) operations from CLR to CNV caused unnaturalness in the phoneme sequence.

These results suggest that for this speaker the combination of DS and DSP had significant effects on intelligibility, but not the individual features D, S, or P. In fact, having feature P caused an intelligibility decrease from CNV speech.

3.8.3 Phoneme Confusions

To further investigate effects of features from CLR speech at the phoneme level, the responses from 18 subjects were transcribed and phoneme confusions were analyzed in each condition. First, correct word sequences were aligned with subjects' response at word level.

For example,

(1) Smoky fires LACK FLAME AND # heat

(2) Smoky fires # FROM # THE heat

The sequence (1) is the correct response, and the sequence (2) is an example of the subject's response. The words "LACK" and "AND" were deleted, while the word "THE" was inserted, "FLAME" was substituted with "FROM". In this analysis, word deletions and insertions were disregarded. Then, substituted words (in this example "FLAME" and "FROM") were aligned at phoneme level,

(1) /f/ /l/ /ei/ /m/

(2) /f/ /ɹ/ /ʌ/ /m/

Table 3.7: The error patterns, in percent, for substitution errors (voicing, manner, place, and height) and insertion/deletion errors at the phoneme level.

Total Count	voice	manner	place	height	Total Count	Ins.	Del.
Condition: CNV							
V (6)	0.0	33.3	50.0	83.3	V (3)	33.3	66.7
C (6)	83.3	50.0	66.7	50.0	C (10)	20.0	80.0
Condition: HYB-P							
V (4)	0.0	25.0	100.0	50.0	V (7)	85.7	14.3
C (7)	42.9	42.9	85.7	57.1	C (14)	85.7	14.3
Condition: HYB-S							
V (5)	0.0	0.0	60.0	80.0	V (2)	0.0	100.0
C (15)	60.0	40.0	86.7	20.0	C (17)	47.1	52.9
Condition: HYB-DS							
V (5)	0.0	40.0	40.0	100.0	V (2)	0.0	100.0
C (10)	20.0	10.0	90.0	40.0	C (11)	54.5	45.5
Condition: HYB-EFPN							
V (8)	0.0	0.0	37.5	100.0	V (5)	20.0	80.0
C (5)	0.0	20.0	100.0	20.0	C (15)	60.0	40.0
Condition: CLR							
V (1)	0.0	0.0	0.0	100.0	V (3)	66.7	33.3
C (2)	50.0	50.0	50.0	50.0	C (16)	62.5	37.5

In this case, /l/ and /ei/ were substituted with /ɻ/ and /ʌ/, respectively. For the substitution errors, each phoneme is characterized in terms of voicing, manner, place, and height.

The error patterns, in percent, for substitution errors (voicing, manner, place, and height) and insertion/deletion errors are shown in Table 3.7. The total numbers of errors for vowels and consonants are indicated in parentheses. The percentage of the error types (count in each type of errors divided by the total number of errors) is indicated in each row. One substitution error can consist of more than one error type, therefore the sum of error percentages can be greater than 100 %. The consonant confusions showed that major substitution errors were the place of articulation in all HYB conditions. The insertion errors of consonants were more common than deletion errors in most of the HYB conditions and CLR speech, while the opposite was true for CNV speech. Though vowel confusion at the phoneme level was less frequent, vowel substitution errors were mostly height errors.

3.9 Conclusions

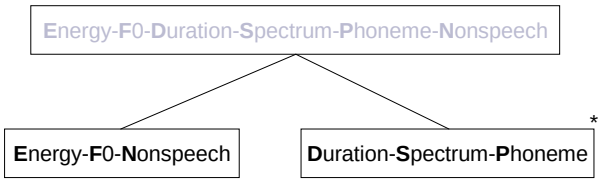
From these experiments, it was confirmed that the intelligibility of CNV speech can be improved by replacing features from CNV speech with those of CLR speech. These results present the first known case in which sentence-level intelligibility of CNV speech has been improved by modification of the speech signal using CLR speech features. We confirmed the hypothesis that the earlier version of the hybridization algorithm could be improved by 4 major modifications to the algorithm in Implementation 2.

The hypothesis that the source of phoneme durations should be matched with that of spectral features was not completely supported, since significant improvement was obtained in one case with spectrum modification but without duration modification. However, the third experiments suggest that the interaction between spectrum and duration may still be worth investigating.

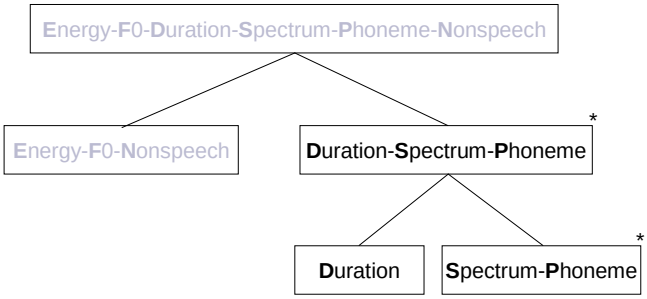
In a series of experiments, tree structures were built with nodes in the tree specifying features from CLR speech to determine relevant features. Figure 3.7 represents the tree structures obtained from Experiments 1 through 3. In Experiment 1, the feature combination HYB-EFN and HYB-DSP were tested. Because HYB-DSP had improved intelligibility over CNV speech, features DSP were split into D and SP in Experiment 2. As a result of Experiment 2, HYB-SP as well as HYB-DSP conditions were effective. In Experiment 3–3, on the other hand, features DSP were again repartitioned into DS and P to examine the contribution of the phoneme sequences separately from the spectral feature. In these experiments, the speech corpus from only one speaker was utilized, hence it may not be straightforward to generalize the results for different speakers. One study showed, for example, that different speakers employ different strategies to produce CLR speech [30].

Our other work [5] (not reported here) examining the features that are relevant to classify CNV and CLR speaking style using machine learning techniques showed that only about 9 features (mostly prosodic features) out of 56 features are needed to capture the most predictive power. The results from this chapter indicate that the features that are important for speaking style classification and for intelligibility may be different.

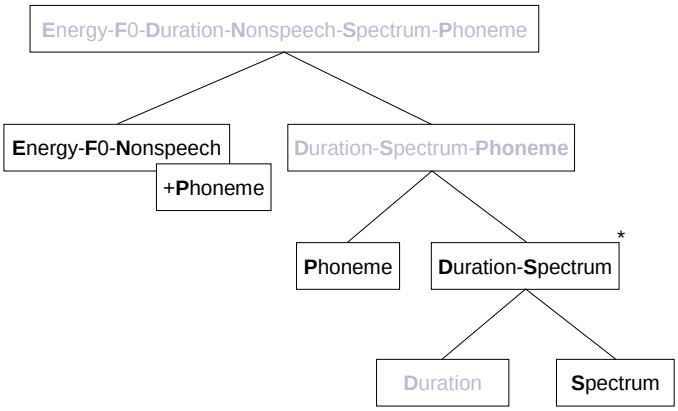
In conclusion, a combination of the features spectrum (S) and duration (D) was sufficient to improve intelligibility of CNV speech for this speaker and this sentence material, while F0, energy, phoneme sequence, and pause information were not. Therefore, the following chapters focus on the spectral and duration features. In the next chapter, we describe the effect of formants and duration on vowel intelligibility and a method to modify formant frequencies.



(a) In Experiment 3–1, a combination of duration (D), spectrum (S) and phoneme sequence (P) improved intelligibility of CNV speech.



(b) In Experiment 3–2, a combination of spectrum (S) and phoneme (P) as well as duration (D), spectrum (S) and phoneme (P) improved intelligibility of CNV speech.



(c) In Experiment 3–3, a combination of duration (D) and spectrum (S) improved intelligibility of CNV speech.

Figure 3.7: Tree structures obtained from Experiments 3–1 through 3–3. Significance as compared with original CNV speech was shown with the asterisks ($p < 0.05$).

Chapter 4

The effect of formant contours and phoneme durations on vowel intelligibility

4.1 Introduction

From the series of experiments discussed in Chapter 3, the combination of phoneme duration and spectral features from CLR speech was effective to improve sentence intelligibility of CNV speech. Spectral features can be considered as a combination of formant frequencies and formant-normalized spectrum (glottal source and nasal resonance information). As a result of an unpublished pilot study, differences in the formant-normalized spectrum were not observed between CNV and CLR speech for the particular speaker in the speech corpus used in Chapter 3. Therefore, the relationship between formant frequencies as a spectral feature and phoneme durations is examined in this chapter[1]. Although meaningful sentences with 5 keywords were used previously, meaningful consonant-vowel-consonant (CVC) words were used in this chapter to focus on understanding formant dynamics.

It is known that the degree of formant undershoot depends on speaking style, word stress, vowel duration, and neighboring consonants [64]. The study in Furui [33] showed that the formant transition region, where the slope is the greatest, is the most important region for syllable (consonant-vowel-consonant) identification. Speaking styles have significant effects on the formant transition at the phoneme boundaries [68]. However, it is not clear how much the observed formant undershoot and the formant transitions of CNV speech are detrimental to vowel intelligibility, as compared with the CLR speech.

The objectives of the study in this chapter are as a continuation of Specific Aim 1, (1) to characterize formant steady-state values and formant transitions with different speaking

[1]Part of this chapter was published in Amano-Kusumoto and Hosom [2].

rates and styles, (2) to determine whether vowel intelligibility can be improved by modifying formant steady state values and formant transitions independently of phoneme duration, and (3) to determine whether the combination of formant frequencies and phoneme durations can be modified to maximize intelligibility.

In this chapter, the study of Moon and Lindblom [68] was extended by testing CNV and CLR speech spoken at different speaking rates. The contributions of speaking style, speaking rate, and vowel identity to intelligibility levels were examined. Next, we examined whether acoustic features of formant steady-state (SS) values (and transitions) contribute to the improved intelligibility of CLR speech, by creating hybrid (HYB) speech. HYB speech is, in this case, a synthetic speech signal that contains acoustic features from both CNV and CLR speech, which is similar to the method discussed in Chapter 3 (Section 3.3). The formant contours of CNV speech were modified to match the SS and transitions of CLR speech.

4.2 Text materials: CVC words

As an extension of Moon's study [68], the four front vowels (/iː/, /ɪ/, /ɛ/ and /ei/) surrounded by the consonants /w/–/l/ at two speaking rates (SLOW and FAST) were recorded in this study. The /w/–/*V*/–/l/ context with front vowels provides large second formant (F2) movement between consonants and the vowel due to coarticulation.

4.2.1 Speech Material

Four test words (*wheel, will, well*, and *wail*) in a carrier sentence were repeated 16 times each. The carrier sentence "it's easy to tell the size of a *WORD*" was used to facilitate the use of prosodic manipulation upon the elicitation of CNV and CLR speech at different speaking rates. The word of interest was equally stressed. The total of 64 sentences was randomized and the order of sentences was kept the same for each speaking style. Speech materials were spoken in four speaking styles (CNV/SLOW, CNV, CLR, and CLR/FAST).

4.2.2 Recordings

The speech signals were recorded digitally at a sampling rate of 16 kHz with 16 bit resolution. One male speaker, a native speaker of North-American English, recorded the speech materials in four recording sessions. The recording of CNV speech was followed by CLR speech in the first and second recording sessions, using the speaker's own distinction

Table 4.1: Formant SS values of four vowels in four speaking conditions. Standard deviations are shown in parentheses.

	Conditions	/iː/	/ɪ/	/ɛ/	/ei/
F1	CNV/SLOW	316.48 (16.29)	441.85 (10.35)	635.10 (39.81)	482.63 (30.34)
	CNV	375.59 (26.46)	457.96 (11.72)	581.24 (51.43)	498.34 (33.58)
	CLR	319.29 (15.67)	439.29 (13.61)	685.48 (24.09)	414.77 (22.31)
	CLR/FAST	332.97 (12.85)	472.32 (26.45)	664.21 (17.03)	498.97 (26.01)
F2	CNV/SLOW	2163.11 (76.86)	1526.42 (45.47)	1374.37 (49.37)	1773.94 (50.32)
	CNV	1830.81 (105.50)	1304.65 (54.24)	1215.76 (43.07)	1601.06 (47.71)
	CLR	2439.31 (43.59)	1724.17 (70.38)	1547.69 (76.45)	2113.96 (50.34)
	CLR/FAST	2273.69 (71.17)	1527.14 (83.17)	1468.93 (50.30)	1963.60 (56.89)

between CNV and CLR speech production. For the third session, CNV speech was spoken at a deliberately slow rate of the speaker's choice. For the fourth session, CLR speech was recorded at a fast speaking rate. A speaking rate other than natural is indicated after the speaking style, e. g. CLR/FAST and CNV/SLOW. It was not the goal of this study to match the speaking rate of CLR/FAST speech with CNV speech, or to match the rate of CNV/SLOW speech with CLR speech. The purpose was to have variety of speaking rates with CNV and CLR speaking styles. The average speaking rates, measured excluding pause durations, were 149 wpm, 365 wpm, 179 wpm, and 289 wpm for CNV/SLOW, CNV, CLR, and CLR/FAST, respectively.

4.3 Acoustic analysis of speech materials

In this section, we analyze the acoustic characteristics of our speech corpus. All words were annotated and segmented automatically using forced alignment [44]. Formant contours of the word of interest were extracted using the Snack Sound Toolkit [91]. A trained transcriber manually corrected phoneme boundaries and formant contours. Acoustic analysis included measurement of vowel steady-state (SS) frequency, formant slope, and the relationship with vowel duration.

4.3.1 Vowel steady-state values

Formant SS values were extracted at the midpoints of each vowel for each sample. The average of F2 SS values over 16 samples in each vowel (/iː/, /ɪ/, /ɛ/, /ei/) in four conditions (CNV/SLOW, CNV, CLR, CLR/FAST) are shown in Table 4.1. SS values of F1 and F2 were submitted to one-way analysis of variance (ANOVA) examining the effect of speaking style. Prior to the analysis, F1 values were converted to the distance from the center

Table 4.2: F2 slope (Hz/ms) at vowel onset and offset, for four vowels in four speaking conditions. Standard deviations are shown in parentheses.

		/iː/	/ɪ/	/ɛ/	/ei/
Vowel onset	CNV/SLOW	17.50 (3.61)	10.27 (2.68)	7.30 (2.12)	10.34 (2.50)
	CNV	16.50 (4.83)	8.13 (2.12)	6.74 (1.45)	11.96 (2.23)
	CLR	32.35 (6.85)	18.56 (5.63)	12.87 (2.04)	20.59 (2.16)
	CLR/FAST	35.53 (4.73)	15.74 (2.19)	12.49 (1.67)	19.16 (2.12)
Vowel offset	CNV/SLOW	−6.53 (2.33)	−4.93 (1.66)	−2.31 (1.57)	−6.26 (1.55)
	CNV	−10.56 (2.46)	−5.89 (2.51)	−3.94 (1.46)	−10.06 (2.81)
	CLR	−15.83 (2.73)	−6.79 (1.40)	−4.93 (1.49)	−12.62 (2.31)
	CLR/FAST	−14.81 (2.51)	−5.61 (2.28)	−4.49 (0.90)	−11.46 (3.64)

(500Hz). Both F1 and F2 SS frequencies, the main effect of speaking style was significant ($p = 0.0011$, $p = 3.8378 \times 10^{-11}$, respectively). Multiple comparison showed that both F1 and F2 SS values of CLR speech are significantly higher than those of CNV speech.

The effect of speaking rate on formant frequencies was tested using a pairwise, two-tailed $t-$test comparing CNV/SLOW and CNV speech, as well as CLR/FAST and CLR speech, for each vowel. All vowels in comparison pairs for F1 and F2, except for the F1 of the vowel /ei/ in CNV conditions and F1 of the vowel /ɪ/ and /ei/ in CLR conditions, were significantly different ($p < 0.001$).

Speaking style caused a significant effect on both F1 and F2 SS values with /w/–/V/–/l/ contexts, even with different speaking rates combined. This indicates that for CLR speech front vowels, the vowel space is expanded along both F1 and F2 dimension. Speaking rate (CNV/SLOW and CNV, or CLR and CLR/FAST) also had a significant effect on F2 SS values, as well as F1 SS values for some vowels.

4.3.2 F2 slope

Since speaking style had a significant effect on F2 SS only, F2 slope is further analyzed in this section. The F2 slope was measured over the 20 ms at phoneme boundaries by fitting a straight line to the observed data. Table 4.2 shows the average F2 slope for four vowels in four conditions. F2 slopes at the transition from pre- and post-vocalic consonant of the vowel for each speaking style were submitted to a one-way ANOVA. The main effect of speaking style was significant both at onset and offset transitions ($p < 0.0001$ and $p = 2.13 \times 10^{-9}$). The post-hoc analysis showed that, as expected, F2 slopes of CLR speech are steeper than those of CNV speech both at onset and offset transitions.

A significant effect was not shown for the effect of speaking rate on F2 slope at the

Table 4.3: Vowel durations (ms) of four vowels in four speaking conditions. Standard deviations are shown in parentheses.

	/iː/	/ɪ/	/ɛ/	/ei/
CNV/SLOW	268.13 (49.83)	154.38 (30.54)	174.38 (33.66)	267.50 (36.06)
CNV	98.75 (8.85)	78.13 (10.47)	88.75 (12.58)	118.75 (15.00)
CLR	229.38 (18.79)	141.88 (23.44)	151.88 (24.82)	236.25 (15.86)
CLR/FAST	151.88 (21.98)	100.63 (12.37)	120.00 (13.17)	172.50 (15.28)

vowel onset transitions for all vowels, based on a pairwise $t-$test comparing CNV and CNV/SLOW speech ($p = 0.1927, 0.0904, 0.5547, 0.0797$) as well as CLR and CLR/FAST speech ($p = 0.5236, 0.0165, 0.4364, 0.0327$), while all vowels showed a significant effect at the vowel offset transitions ($p < 0.01$) except for /ɪ/ ($p = 0.2663$) and /ɛ/ ($p = 0.0131$) in CNV comparisons.

The results suggest that, in general, F2 slopes are determined by the speaking style, independent of speaking rate at vowel onset. Consistent with the study by Moon [68] (in which F2 slope was measured from two points, 25 % down from the F2 peak and 25 % up from middle of /w/), these results showed the movement of articulators producing CLR speech was faster than in CNV speech. On the other hand, the duration of the vowel was not the key determinant of the rate of moving the articulators. It is also an important finding that differences in speaking rate caused changes in F2 slopes at the vowel offset for most of the vowels.

4.3.3 The relationship between F2 steady-state frequencies and vowel durations

Vowel duration (ms) was measured from the beginning to the end of the vowel. The average vowel duration is shown in Table 4.3 for the four vowels in the four conditions. Figure 4.1 shows the F2 SS frequencies as a function of vowel duration, where outliers in terms of duration were detected using a modified z-score test (outliers are shown with stars in the figure) [45].

$$Z_i = \left| \frac{0.6745(x_i - \bar{x})}{median(x_i - \bar{x})} \right| \tag{4.1}$$

where x_i represents each sample, and $\bar{x}$ is the average over 16 samples of the vowel per condition. The criteria for detecting an outlier was $Z_i > 3.5$.

According to the study from Lindblom [64], the degree to which a vowel reaches its target has an exponential relationship with the vowel duration. Therefore, an exponential

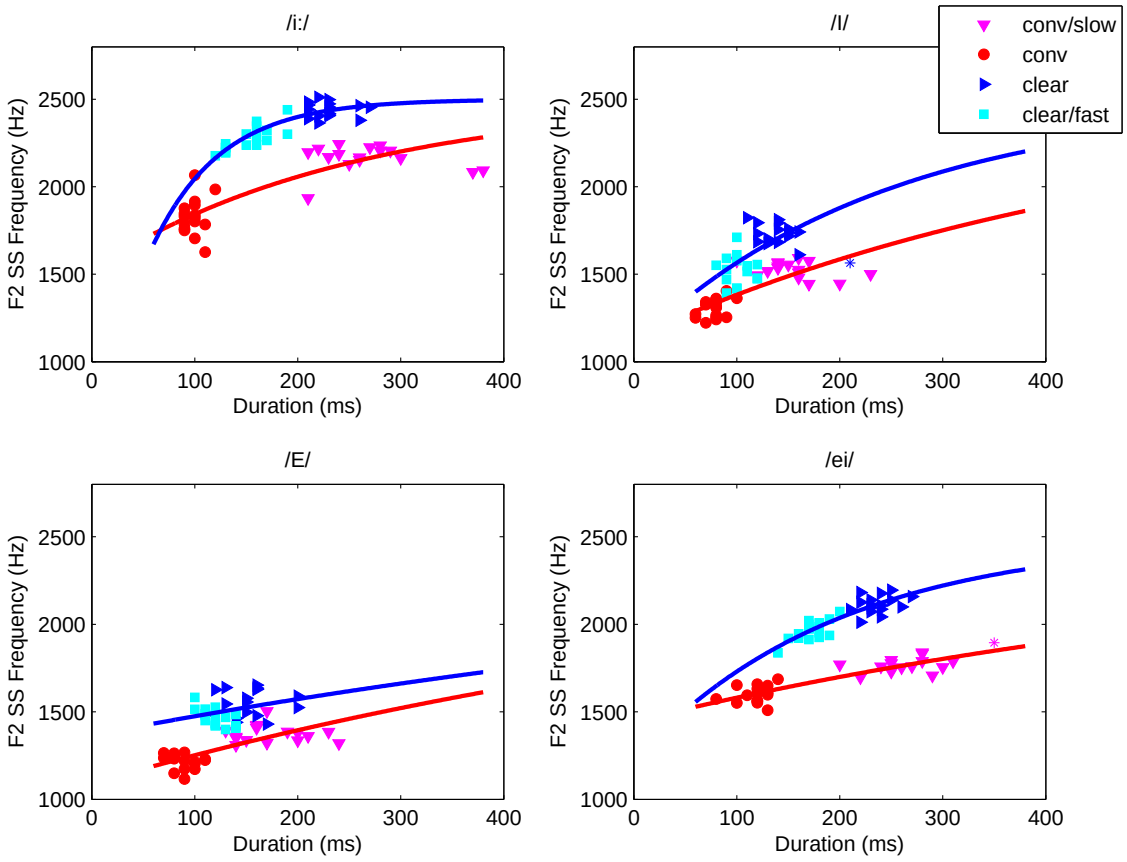


Figure 4.1: Formant frequencies as a function of vowel duration in four conditions. Outliers in terms of vowel duration are represented with asterisks. Two lines show the fitted exponential curves in CLR and CNV speaking styles separately.

function ($Ae^{-\lambda t}$) was fitted in each group of speaking styles by varying parameters A and λ to minimize the sum of squared errors.

As a result of permutation test in multiple regression [75, 20], the effect of speaking style regardless of speaking rate was significantly different for all vowels ($p < 0.0001$). Pairwise two-tailed $t-$tests showed that CNV/SLOW has longer phoneme duration than CLR speech for all vowels (p=0.0034, 0.0309 0.0198 and 0.0024, respectively). The results showed that despite the duration of CLR speech being less than that of CNV/SLOW speech, F2 values of CLR speech for all vowels are higher than those of CNV/SLOW speech.

Picheny *et al.* reported that the differences in duration between CNV and CLR speech are greater for tense vowels than for lax vowels [80]. Similarly in our study, we found that the duration differences in lax vowels were much less than in tense vowels. On the other hand, in the study by Moon and Lindblom [68] investigating duration dependencies of F2

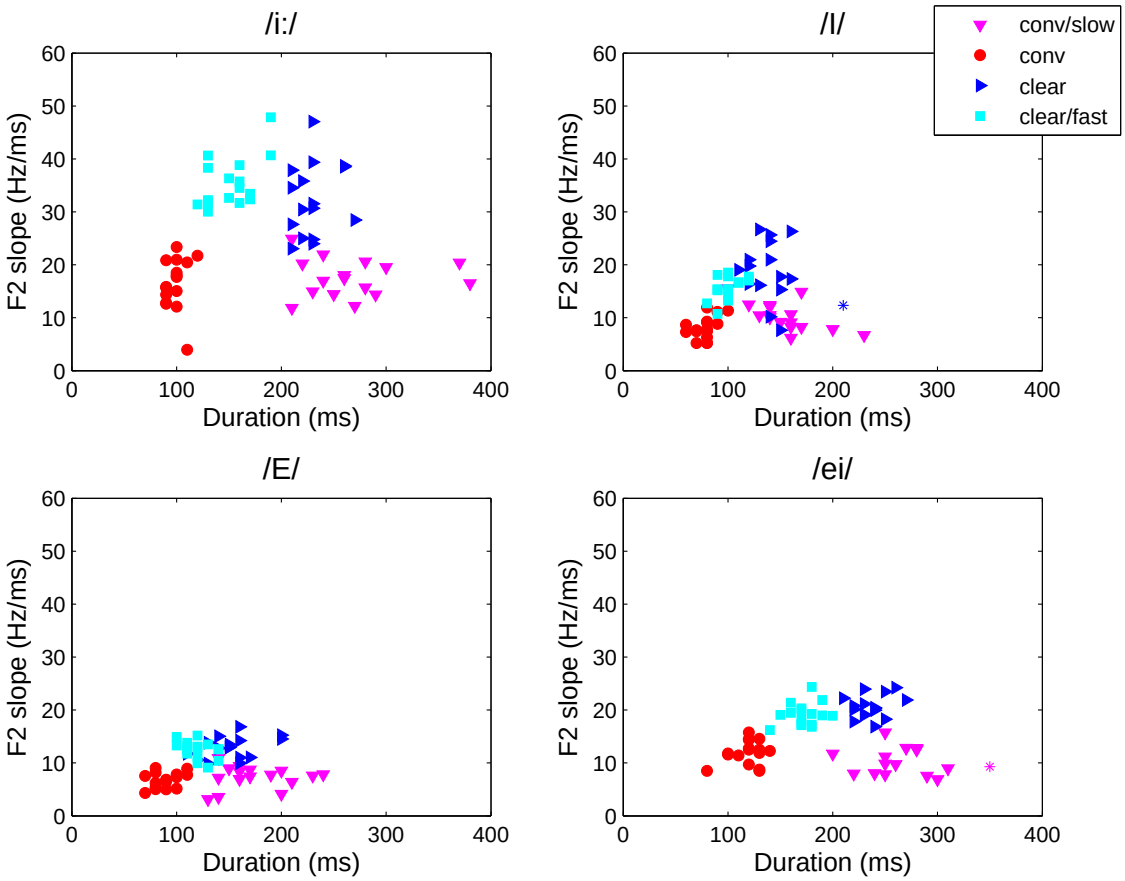


Figure 4.2: F2 slope (Hz/ms) frequencies as a function of vowel duration with four conditions. Outliers in terms of vowel duration are represented as asterisks.

frequencies with clear and citation form of speech, F2 changed little with longer vowel duration for tense vowels (especially for the vowel /iː/), while lax vowels showed longer duration resulting in higher F2 values. In our study, duration dependencies on F2 values showed that longer duration for *both* tense and lax vowels caused higher F2 values, unlike Moon's study [68].

4.3.4 The relationship between F2 slope and vowel durations

Similar to Figure 4.1 in the previous section, Figure 4.2 shows the F2 onset slope as a function of vowel duration, where duration outliers were detected using a modified z-score test (outliers are shown with stars in the figure) [45]. Unlike F2 SS values, the F2 slope does not appear to change as the vowel duration becomes longer, with the possible exception of the vowel /ɪ/ in the CLR style.

As a result of acoustic analysis of our speech corpus, vowel F2 steady-state values have a relationship with both speaking style and speaking rate, while F2 slopes at phoneme boundaries vary based only on speaking style for vowel onset and both on speaking style and rate for vowel offset. F1 SS values were not different based on speaking style.

4.4 Experiment 4–1: Intelligibility of naturally spoken CNV and CLR speech at different speaking rates

A perceptual experiment was conducted to examine whether the effects of CLR speech vary based on speaking rate and vowel identity. Krause and Braida [54] concluded that fast spoken CLR speech from speakers with professional public speaking has an intelligibility advantage over CNV speech. We investigate if their finding holds in our speech corpus, and also look at vowel identity as a factor in intelligibility. Four vowels (/iː/, /ɪ/, /ɛ/ and /ei/) with one phoneme context (/w/–/V/–/l/) were tested, in four speaking styles (CNV/SLOW, CNV, CLR, and CLR/FAST).

Ten adults, aged between 19 and 38 years, were recruited for Experiment 4–1. All listeners were native speakers of North-American English with self-reported normal hearing.

4.4.1 Normalizing loudness

It is important to keep the loudness of vowels constant in test words, since loudness plays an important role for speech intelligibility. First, the root mean square value of the vowel in a test word (RMS_v) was calculated. The gain factor (G_v) was then obtained for the vowel to have a normalized RMS_v value. Finally, the energy of the test word was multiplied by G_v.

4.4.2 Normalizing F0 contour

The different speaking styles (CLR and CNV) resulted in differences in F0, which is consistent with previous work [80, 55]. Since it is unknown whether increased F0 values in CLR speech contribute to its increased intelligibility, in order to evaluate the importance of only formant values and phoneme duration, the F0 contours were normalized over the four conditions.

First, an F0 contour model was derived from CLR/FAST based on the F0 onset values of each phoneme (/w/, /V/, /l/), the F0 offset values of the /l/, and the maximum F0 value in /V/. These five points in the F0 contour model were the average values of CLR/FAST

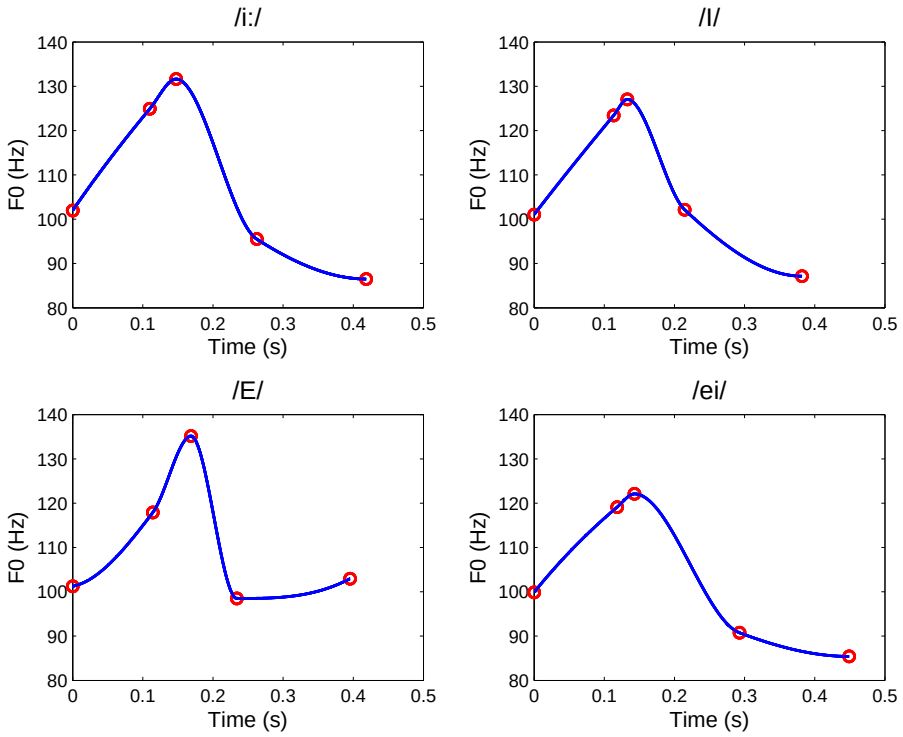


Figure 4.3: F0 contour model used to normalized F0 values for the four conditions. Red circles from the left- to right-hand side indicate (1) onset of /w/, (2) onset of /V/, (3) maximum point of /V/, (4) onset of /l/, and (5) offset of /l/.

Table 4.4: F0 values of CLR/FAST for the F0 contour model.

	/iː/	/ɪ/	/ɛ/	/ei/
p1 (onset of /w/)	101.97	101.04	101.25	99.88
p2 (onset of /V/)	124.91	123.43	117.91	119.11
p3 (Max F0)	131.63	127.06	135.17	122.10
p4 (onset of /l/)	95.51	102.13	98.49	90.76
p5 (offset of /l/)	86.50	87.14	102.97	85.41
Mean F0	107.72	105.69	104.14	103.10
Word duration (sec)	0.42	0.38	0.40	0.45

speech over 16 repetitions for each vowel. An F0 contour was then derived by interpolating these five points with cubic spline interpolation. The CLR/FAST speech was chosen for the F0 contour model, because the resulting F0 modification in each condition did not require lowering F0 values, which may lead to more noticeable signal artifacts.

Figure 4.3 shows the F0 contour model with values shown in Table 4.4. The F0 contour model was stretched to match the duration of each phoneme with the observed data in each condition. In the F0 modification stage, the F0 values of each sample in all four conditions were modified to the values of F0 contour model.

4.4.3 Procedures and apparatus

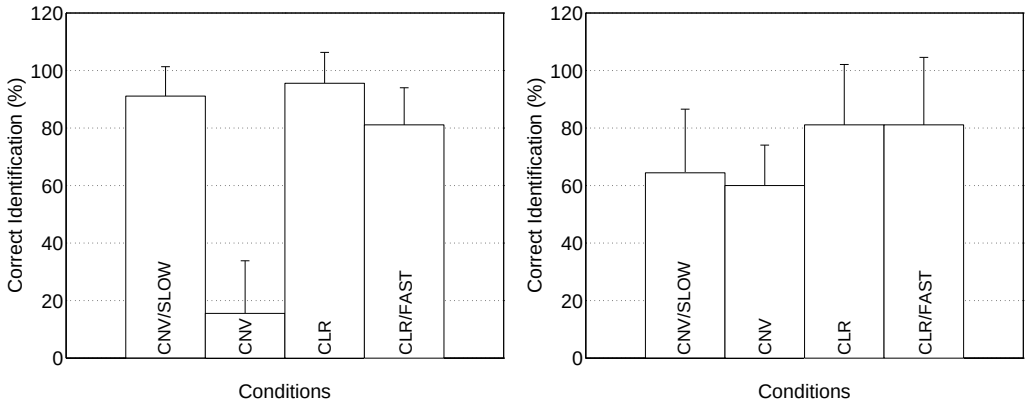
The perceptual experiment took place individually for each listener in a perceptual testing booth (*Whisper Room,* SE2000 series). A listener was seated in front of a computer monitor, listening to stimuli through circumaural headphones (Sennheiser HD 280 Pro), binaurally. A forced-choice test was used with four buttons, corresponding to the four choices ("*wheel*", "*will*", "*well*", "*whale*"), appearing on the user-interface screen.

12-talker babble noise was used to simulate a noisy environment. The energy of the noise was adjusted to meet the desired SNR–50 level for each listener. The SNR–50 level refers to the signal-to-noise (SNR) ratio at which a listener can correctly identify the stimuli in the CNV speaking style 50 % of the time. The SNR–50 level was obtained for each listener using the up-down adaptive procedure described in Section 3.5.2 [62].

The listeners were tested in three sessions: the first two sessions were used for obtaining the listener's SNR–50 level, and the third session was for the vowel identification experiment. SNR–50 values from the second session were used for the vowel identification experiments, while the values from the first session were disregarded. The total of 144 stimuli (4 /w/–/V/–/l/ words × 4 speaking styles × 9 repetitions) were tested using a Latin Square design.

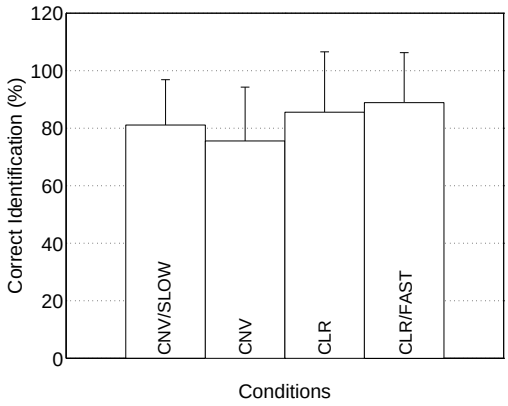
4.4.4 Results and discussions

The results of Experiment 4–1 are shown in Figure 4.4, representing percent correct rates for each vowel identity in all four conditions, averaged over 10 listeners. The average noise level (SNR–50) was −1.08 dB (std: 2.11). Percent correct rates were converted to rationalized arcsine units (RAUs) prior to statistical analysis [97]. The effect of vowels (/iː/, /ɪ/, /ɛ/ and /ei/) and speaking styles (CNV/SLOW, CNV, CLR and CLR/FAST) were submitted to a two-way repeated-measures analysis of variance (ANOVA).

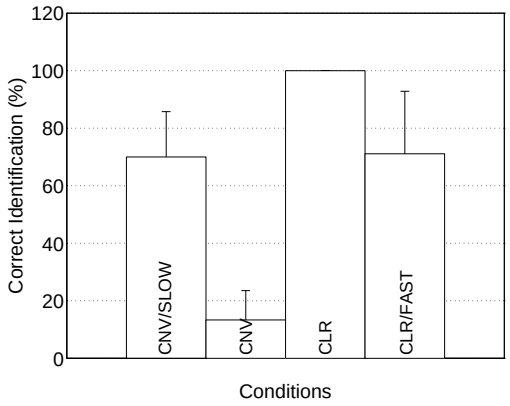


(a) Vowel /iː/

(b) Vowel /ɪ/



(c) Vowel /ɛ/

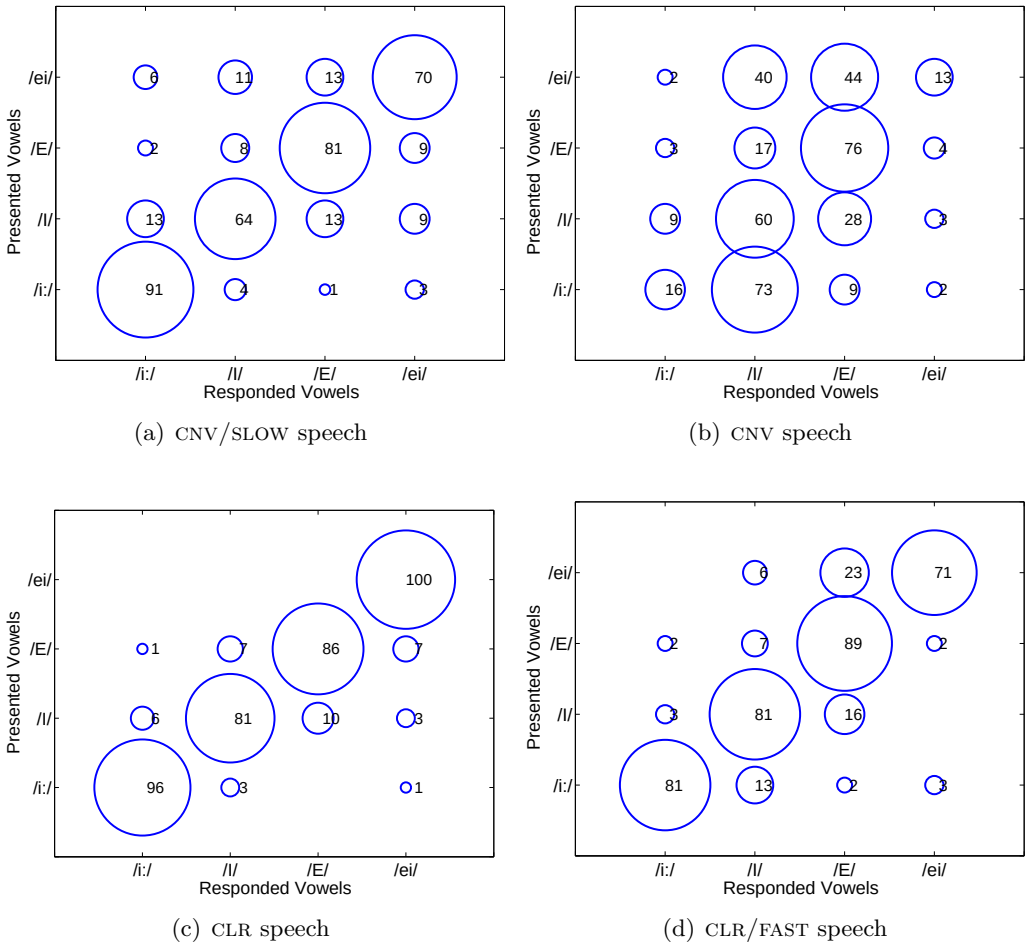


(d) Vowel /ei/

Figure 4.4: Percent correct rates for four vowels in four conditions (two speaking styles and two speaking rates).

The results of two-way ANOVA (vowel identities × speaking styles) show that the main effects of vowels and speaking styles were significant ($p = 0.001$ and $p < 0.0001$). For the tense vowels (/iː/ and /ei/), CLR speech was significantly more intelligible than CNV speech (both $p < 0.01$). On the other hand, for lax vowels (/ɪ/ and /ɛ/) the effects of speaking style were not significant (both $p > 0.05$).

The confusion matrices are shown in Figure 4.5, representing responded and presented vowels on the horizontal and vertical axes, respectively. The numbers represent the percentage of responses, with a maximum of 100 % per vowel. The confusion pattern in CNV speech, which had the least intelligibility compared with any other speaking style, showed that listeners tended to perceive the vowel /iː/ as /ɪ/ ("*wheel*" as "*will*") (73.33 %), and



(a) CNV/SLOW speech

(b) CNV speech

(c) CLR speech

(d) CLR/FAST speech

Figure 4.5: Confusion matrices representing responded and presented vowels on the horizontal and vertical axes, respectively. The diagonal responses are the correct answers; the percentage is shown at the center of each circle.

the vowel /ei/ as /ɛ/ (44.44 %) and /ɪ/ (40.00 %) ("*whale*" as "*well*") and "*will*"). In general, with 16-talker-babble noise at the SNR–50 level, tense vowels with short vowel durations were more often perceived as lax vowels, while long tense vowels tended to be identified correctly.

The speaking rate affected intelligibility, showing that CLR/FAST speech is less intelligible than CLR speech, and that CNV speech is less intelligible than CNV/SLOW speech. This indicates that the faster speaking rates resulted in less intelligible speech. Figure 4.6 shows percent correct rates based on the vowel duration of the stimulus. It is clearly

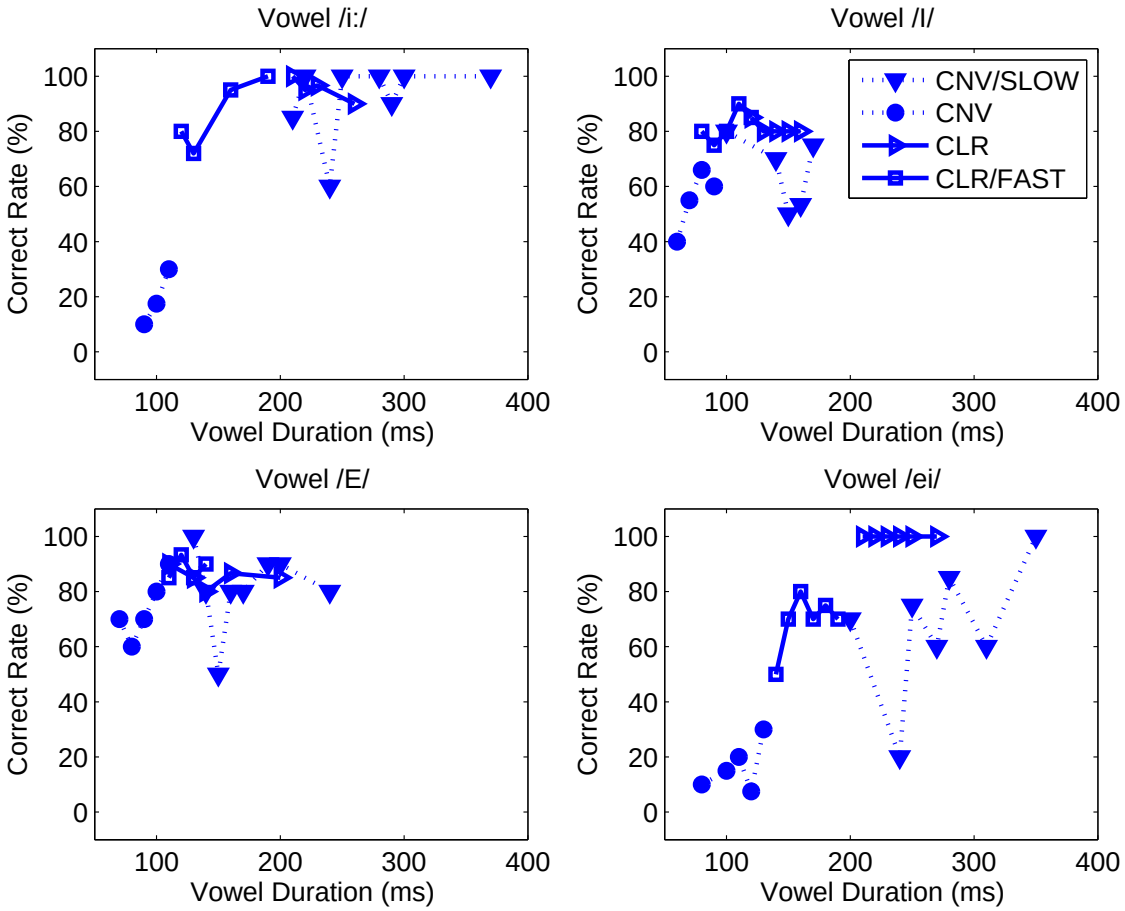


Figure 4.6: Percent correct rates as a function of vowel durations for each vowel in four conditions in Experiment 4–1.

shown that the shorter vowel durations have less percent correct rates. The one exception is in the CNV/SLOW data, which shows a notch at one duration per vowel. This seems to be an anomaly that is not indicative of the underlying trend. For CNV speech, it is uncertain whether the cause of less intelligible tense vowels is due to short vowel duration or the large amount of formant undershoot.

Krause and Braida's study showed that the benefit of CLR speech is extended to normal speaking rates (approximately 200 wpm), but the fastest speaking rate with a CLR speech benefit was observed at 218 wpm. Similarly, our results also show that it may not be possible to obtain a level of intelligibility equal to that of CLR speech at a fast speaking rate (289 wpm) for naturally−produced speech. In Experiment 4–2, we examine whether it is possible to improve vowel intelligibility by modifying formant frequencies with and

without modifying word duration.

4.5 Hybridization algorithm

The hybridization (HYB) algorithm proposed here is a signal processing technique that modifies certain acoustic features of CNV speech to match those of CLR speech, similar to the one described in Chapter 3 [48]. Four front vowels (/iː/, /ɪ/, /ɛ/ and /ei/) were examined to test whether it is possible to improve vowel intelligibility by modifying formant contours in a word.

Results from acoustic analysis (Section 4.3) revealed that CNV and CLR speech had inherently different F2 SS values and different F2 slopes at phoneme boundaries. Experiment 4–1 showed that the vowel intelligibility for /iː/ and /ei/ was higher for CLR speech than CNV speech, and that short durations negatively impacted the intelligibility of tense vowels. These results motivated us to examine whether intelligibility could be improved by reducing the degree of formant undershoot, or whether CNV speech with modified formant contours is inherently less intelligible than CLR speech because of the short duration of CNV speech. The formant transition region is known to be important for vowel perception [33], therefore the formant transition region in addition to F2 SS value was tested in a perceptual experiment. The stimuli consisted of enhanced formant SS values only, or both enhanced SS values and formant transitions in hybrid speech (Section 4.5). Our hypothesis was that formant SS values (and possibly transitions) of CLR speech contribute to improved intelligibility, independent of duration.

The first step of the hybridization algorithm was to extract target formant SS and/or formant transition values from CLR speech. Then, the HYB formant contours with these target values were designed. The third step was to modify formant values of CNV speech by analysis and synthesis methods to match the target formant contours. Although acoustic characteristics revealed differences in only F2 SS values and F2 slopes, in formant modification it was necessary to modify F1 through F4 frequencies to prevent, for instance, F2 being raised higher than F3.

4.5.1 Hybridization conditions

Three HYB conditions were evaluated to test the effect of SS values, transitions, and vowel duration (HYB-M, HYB-MT, and HYB-CD). The term HYB-M indicates that formant SS

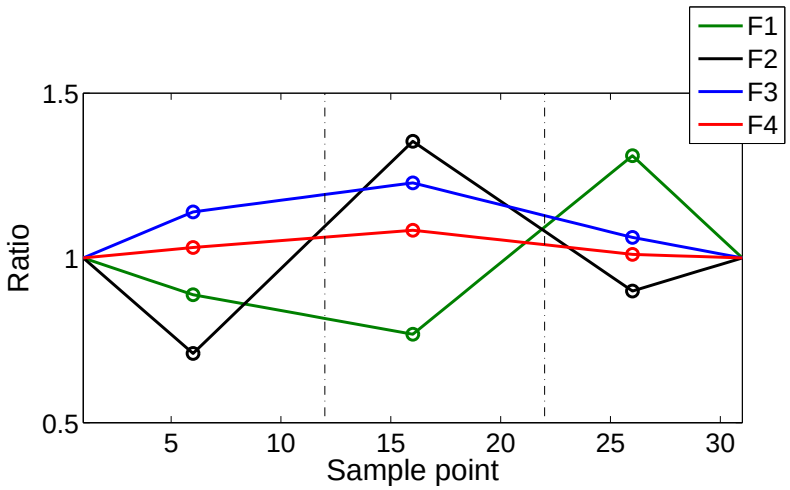
values of HYB speech at midpoints are those of CLR speech. Similarly, the term HYB-MT indicates that SS values at phoneme midpoints and formant transitions at phoneme boundaries of HYB speech are those of CLR speech. Finally, the third condition HYB-CD indicates that the entire formant contour (not only SS values and transitions) and phoneme durations of HYB speech are those of CLR speech. The reason to include phoneme durations in HYB-CD is because our previous study showed that changing the combination of short-term spectra and phoneme durations improved the sentence intelligibility over that of CNV speech [48]. This condition goes one step further, examining only the formant contour (not the formant-normalized spectrum) and phoneme duration. Also, HYB-CD provides a test of the quality of our hybridization method with formant modification.

HYB-M: CLR steady-state values at midpoints

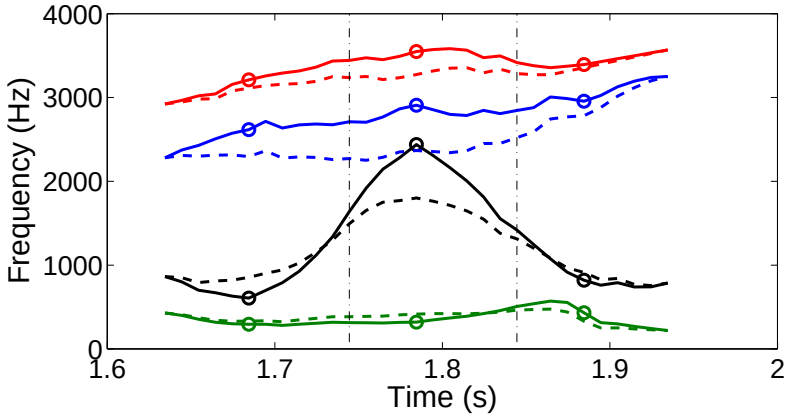
The target SS values were extracted at the midpoints of each phoneme /w/, /V/ and /l/ of CLR speech, and averaged over 16 samples per word. The process of designing a HYB-M formant contour required a weighting function for each formant contour (F1 through F4) shown in Figure 4.7(a). The weighting function was designed to be linearly ascending or descending to describe the ratio between the target SS values and the CNV SS values at the midpoint of each phoneme, using phoneme duration from CNV speech. The points at the beginning and ending of the weighting function were set to a ratio of 1.0 to avoid any discontinuities from the unmodified preceding and following waveform. Then, the original CNV formant contour was multiplied by the weighting function to obtain the HYB-M formant contours. The resulting HYB-M formant contour has target SS values from CLR speech. Figure 4.7(b) shows the original CNV (dashed line) and HYB-M formant contour (solid line) after the weighting function was applied.

HYB-MT: CLR steady-state values at midpoints and formant transitions

In addition to the target SS values, target transitions at phoneme boundaries over a 20 ms range were extracted and averaged over 16 samples of CLR speech per word. In designing the weighting function to include phoneme transitions, the ratio was calculated between target values and CNV formant values (F1 through F4) at midpoints of the phonemes, and at three points near the phoneme boundaries for both /w/ to /V/ and /V/ to /l/. Similar to the previous condition HYB-M, the weighting function for each formant was designed to be linearly ascending or descending to describe the desired ratio, as shown in Figure 4.8(a). Then, the original CNV formant contour was multiplied by the weighting



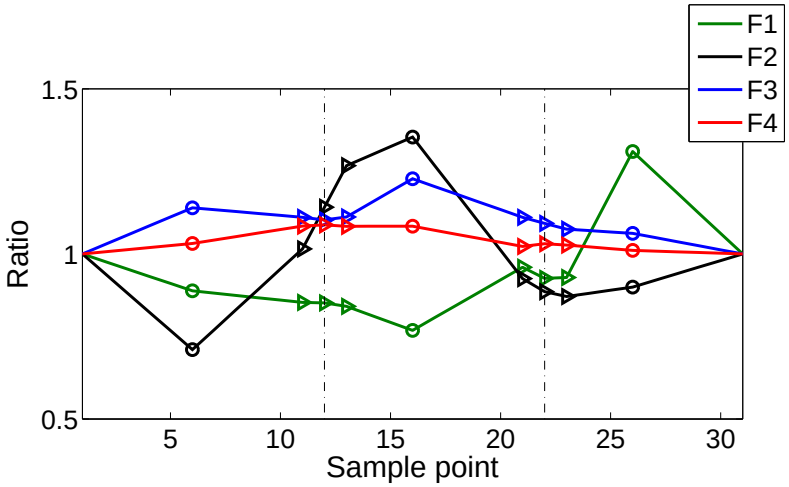
(a) Weighting functions applied to the CNV formant contours



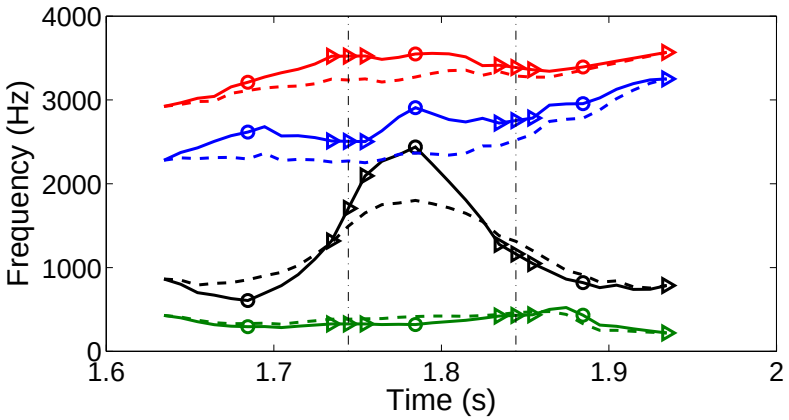
(b) Formant contours (F1 through F4) of the word "*wheel*". Dotted lines are formant contours of CNV speech, and solid lines are the modified contours in HYB speech. The circles on the formant contours indicate the SS values of CLR speech.

Figure 4.7: HYB-M condition. The HYB-M contour was obtained by multiplying original CNV contours with weighting functions (a). Vertical dashed lines in (a) and (b) represent phoneme boundaries.

function. In this way, the formant contours of HYB-MT were guaranteed to have SS values and transitions at the phoneme boundaries that were identical with target values. The formant contours of HYB-MT (solid line) and the original CNV speech (dashed line) are shown in Figure 4.8(b).



(a) Weighting functions applied to the CNV formant contours



(b) Formant contours (F1 through F4) of the word "*wheel*". Dotted lines are CNV formant contours, and solid lines are the modified contours HYB-MT. The circles and triangles on the formant contours indicate the SS values and the transitions of CLR speech, respectively.

Figure 4.8: HYB-MT condition. The HYB-MT contour was obtained by multiplying original CNV contours with weight functions (a). Vertical dashed lines in (a) and (b) represent phoneme boundaries.

HYB-CD: CLR formant contours with phoneme durations

Unlike HYB-M and HYB-MT conditions, the process of designing HYB-CD formant contours did not require a weighting function. Because phoneme durations were modified to match CLR speech at the synthesis stage, formant contours from CLR speech were copied as HYB-CD formant contours. As shown in Figure 4.9, the formant contours of HYB-CD (solid line)

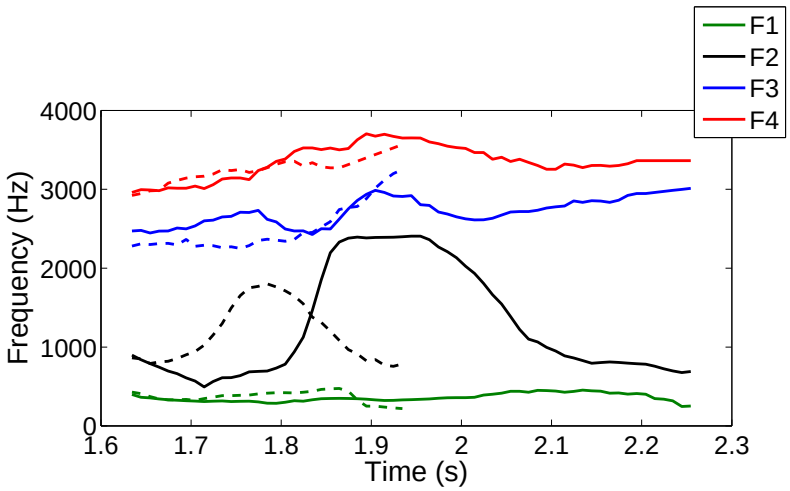


Figure 4.9: HYB-CD condition. Formant contours (F1 through F4) of the word "*wheel*" are shown. Dotted lines are CNV formant contours, and solid lines are the modified contours. The duration of each phoneme is stretched to match that of CLR speech.

have longer phoneme durations than the original CNV speech (dashed line). This condition still has the glottal source of CNV speech.

4.5.2 Speech synthesis with HYB formant contours

For all three HYB conditions, after new formant contours were designed, the original CNV speech was analyzed, the existing formants were removed by inverse filtering, and the HYB speech was synthesized with new formant contours.

First, the speech waveform was analyzed with a pitch-synchronous frame that spans two pitch periods with a one pitch period overlap. The speech signal $S(z)$ (z-transformation) can be represented as $S(z) = Q(z) \times V(z)$, where $Q(z)$ is the residual signal, and $V(z)$ is a vocal tract transfer function. $V(z)$ can be modeled as complex pole pairs represented as

$$V(z) = \frac{1}{\prod_{k=1}^{4}(1 - r^{(k)}e^{j\theta^{(k)}}z^{-1})(1 - r^{(k)}e^{-j\theta^{(k)}}z^{-1})} \tag{4.2}$$

$$r^{(k)} = e^{-\pi f_b^{(k)}/Fs} \tag{4.3}$$

$$\theta^{(k)} = \frac{2\pi f_f^{(k)}}{F_s} \tag{4.4}$$

where $f_f^{(k)}$ is the k-th formant frequency, $f_b^{(k)}$ is its bandwidth, F_s is the sampling frequency.

Four resonant frequencies (F1 through F4) in one frame were removed by applying an inverse filter ($V_{CNV}(z)$) that was designed with formant frequency values from CNV speech. The residual signal ($Q_{CNV}(z)$) from inverse filtering contained primarily the glottal source and higher formants spoken in the CNV style.

The new HYB formant contours (F1 through F4) were used to design all-pole digital filters acting as vocal tract filters ($V_{HYB}(z)$). The bandwidths of each filter were unchanged from those of the original CNV speech. The speech waveform in each frame was obtained by applying the all-pole digital filters ($V_{HYB}(z)$) to the residual signal ($Q(z)$).

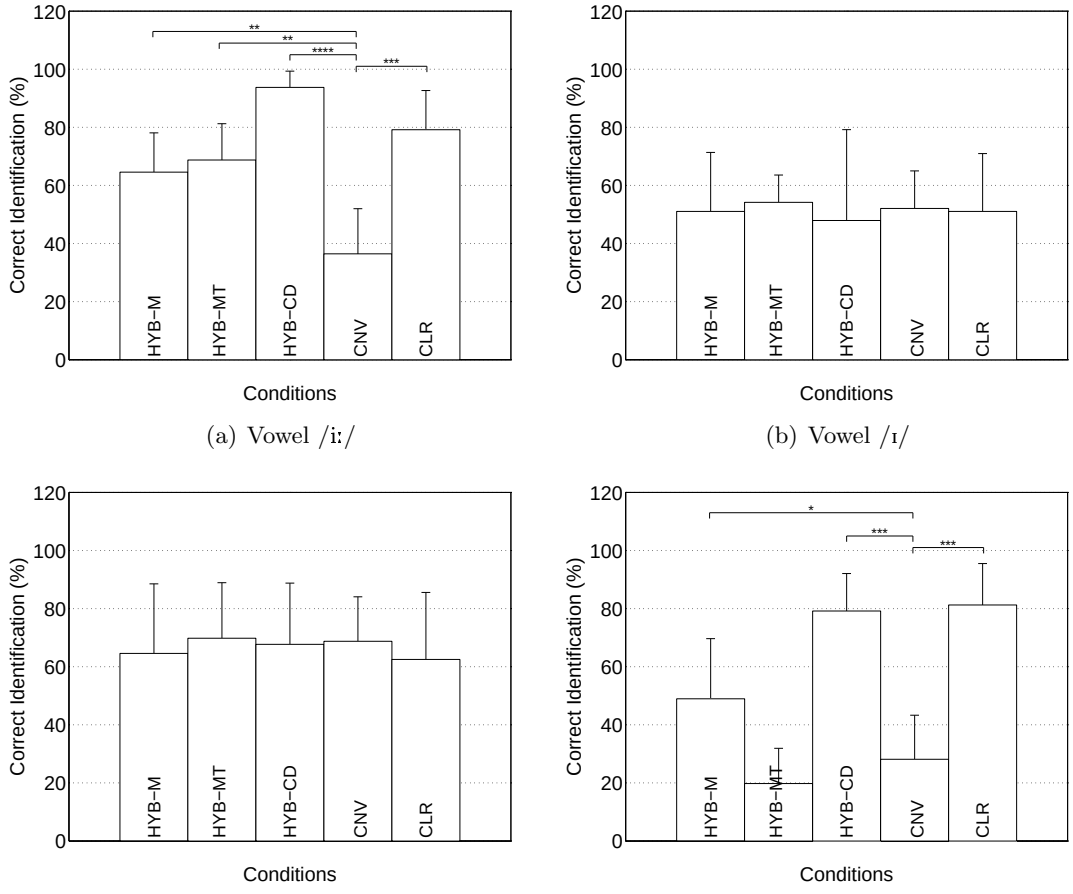
For HYB-CD, it was required to stretch the phoneme duration to match that of CLR speech. At the synthesis stage, residual signals were repeated as necessary to obtain the desired CLR phoneme durations. The HYB-M and HYB-MT conditions had no duration modification. In all cases, the pitch-synchronous overlap-add method was used to create the final waveform.

4.6 Experiment 4–2: The effects of formant contours and phoneme durations on vowel intelligibility

A perceptual experiment was conducted to examine whether HYB speech with CLR speech formant values with and without stretching phoneme durations is more intelligible than CNV speech. Six adults, aged between 19 and 34 years, participated in Experiment 4–2.

4.6.1 Procedures and apparatus

Four vowels (/iː/, /ɪ/, /ɛ/ and /ei/) with /w/–/V/–/l/ contexts were tested in 12-talker-babble noise at each listener's SNR–50 level. The stimuli used in Experiment 4–2 were the three types of HYB speech (HYB-M, HYB-MT, and HYB-CD) and the original CNV and CLR speech as a baseline. All stimuli had a normalized energy value and F0 contour as described in Sections 4.4.1 and 4.4.2. The experimental procedure was the same as in Experiment 4–1 (Section 4.4.3). A total of 320 stimuli (4 /w/–/V/–/l/ words $\times$ 5 conditions $\times$ 16 repetitions) were presented to each listener.



(a) Vowel /iː/ (b) Vowel /ɪ/

(c) Vowel /ɛ/ (d) Vowel /ei/

Figure 4.10: Percent correct rates for five conditions. Significant differences are shown with asterisks (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$).

4.6.2 Results and discussions

The results of Experiment 4–2 are shown in Figure 4.10. The average SNR–50 used for the background noise was $-4.44\,\mathrm{dB}$ (std: 0.97). Percent correct rates were converted to the rationalized arcsine units (RAUs) prior to statistical analysis [97]. The planned $t-$test revealed that the difference in vowel intelligibility between CNV and CLR speech was significant for tense vowels (both $p < 0.001$). For the vowel /iː/, the intelligibility differences between CNV and all three HYB conditions were significant (HYB-M, $p = 0.0043$; HYB-MT, $p = 0.0016$; HYB-CD, $p < 0.0001$). For the vowel /ei/, only HYB-M and HYB-CD speech improved vowel intelligibility over CNV speech ($p = 0.0388$, $p < 0.001$). The

difference in results between HYB-M and HYB-MT for the vowel /ei/ suggests the importance of the overall formant trajectory, especially for diphthongs. For example, the variability of the F2 contour of diphthongs was studied by Weismer and Berry [107]. Their results showed the effect of speaking rate was not a simple compression or expansion of some prototype formant contour. From the results of our experiments, it is not yet clear how the contour should be modified to maximize intelligibility for diphthong (/ei/). Because of the lack of intelligibility difference between CNV and CLR for lax vowels, there was no room for the formant modification to improve the intelligibility of CNV speech.

The confusion patterns (Figure 4.11) showed that "*wheel*" in CNV speech was perceived as "*will*" 39.58 % of the time, while the correct response occurred 36.46 % of the time. The "*wheel*"–"*will*" confusion was improved to 26.04 %, 22.92 %, and 4.17 % for HYB-M, HYB-MT, and HYB-CD, respectively. The vowel /ei/ in "*whale*" was often confused with /ɪ/ in "*will*", at 40.63 %, 25.00 %, and 32.29 % in CNV, HYB-M and HYB-MT conditions. However, "*whale*" was perceived as "*wheel*" in HYB-MT (30.21 %) much more than in HYB-M (12.50 %). Similar to the results in Experiment 4–1, the short tense vowels were more often perceived as lax vowels. It is worth noting that the vowel /ei/ was confused with /iː/ in HYB conditions, even at short durations.

In summary, for the vowels /iː/ and /ei/, even with short durations, the formant modification targeting CLR SS values was effective to significantly improve vowel intelligibility. The HYB-CD results confirm that the hybridization algorithm can yield high-quality and highly intelligible speech when modifying formant frequencies. It can also be concluded that spectral tilt and formant bandwidth were not important contributions to the improved intelligibility of CLR speech for our normal-hearing listeners and this speaker.

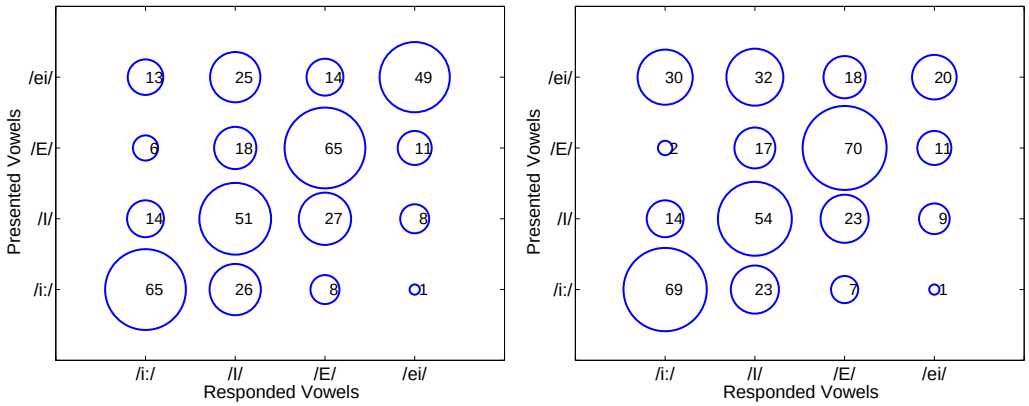
4.7 Conclusions

In this chapter, we examined the effect of formant contour and phoneme durations, and developed a technique to improve vowel intelligibility. The results of acoustic analysis showed that F2 SS values were determined based on speaking style and rate, while F2 slopes at the phoneme boundary vary based only on speaking style and F1 SS values were not different regardless of speaking style and rate.

The results of the first perceptual experiment showed that it may not be possible to obtain a level of intelligibility equal to that of CLR speech in naturally-spoken fast speech. The shorter tense vowel was often identified as lax vowel.

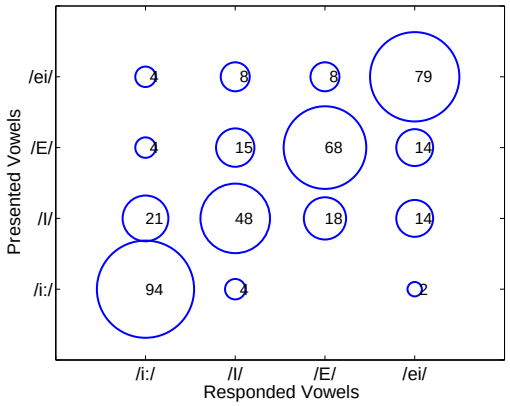
We proposed a HYB algorithm to modify formant frequencies to match those of CLR speech, as well as to lengthen the phoneme durations of CNV speech to those of CLR speech. The second perceptual experiment showed that (1) the HYB algorithm can successfully increase the intelligibility of CNV speech to CLR intelligibility levels by formant and duration modification (HYB-CD), (2) vowels with short durations can have significantly improved intelligibility by formant modification, and (3) spectral tilt and formant bandwidths did not contribute to improving intelligibility.

Similar to the results in Chapter 3, the results in this study are also limited to one male speaker, which cannot be easily generalized to different speakers. To further understand the relationship between formant contour and phoneme duration, we parameterize the characteristics of the formant contour using a formant contour model (Chapter 5). In addition, we increase the number of consonants and vowels in the CVC format, and number of speakers (Chapter 6).

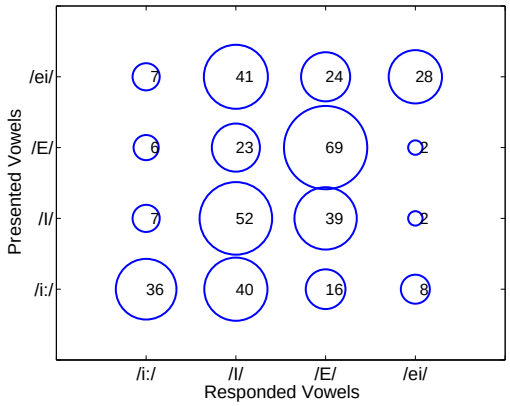


(a) HYB-M speech

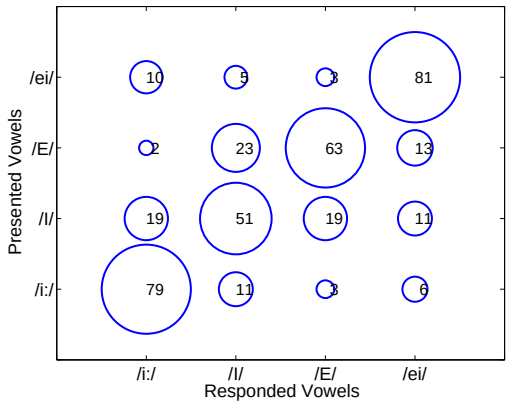
(b) HYB-MT speech



(c) HYB-CD speech



(d) CNV speech



(e) CLR speech

Figure 4.11: Confusion matrices representing responded and presented vowels on the horizontal and vertical axes, respectively. The diagonal responses are the correct answers; the percentage is shown at the center of each circle.

Chapter 5

Effect of speaking style and speaking rate on formant contours with limited phoneme contexts

5.1 Introduction

In Chapter 4, vowel intelligibility of CNV speech was significantly improved by modifying formant contours to resemble those of CLR speech. The condition where formant steady-state values of CNV speech were matched to those of CLR speech was effective, without modifying phoneme durations. However, synthesizing the formant contour with CLR speech steady-state values required a method to systematically modify the formant transitions. In this chapter[1], we examine the formant transition characteristics of CLR speech. The effect of speaking style and speaking rate is investigated within the context of a formant contour model.

Weismer and Berry [107] showed that the effect of speaking rate on F2 cannot be captured by linear compression or expansion of some prototype contour. By using the proposed formant contour model, the effect of speaking rate and style on formant contours is characterized with a formant coarticulation coefficient and formant target values.

In a study by Broad and Clermont [17], formant contours of vowels in C_1VC_2 contexts (/b/, /d/, /g/) were decomposed into coarticulation functions (exponential functions) and vowel target values. In this study, by extending Broad and Clermont's work, we modeled the formant contours of C_1VC_2 words for the first four formants using sigmoid functions for coarticulation. The results of the formant contour model can be applied, for example, to synthesize formant contours in text-to-speech synthesis. In this chapter,

[1]Part of this chapter was published in Amano-Kusumoto and Hosom [3].

the method to model formant contours (Section 5.2), the results of the formant model on vowels in /w/–/V/–/l/ and /t/–/V/–/l/ contexts (Section 5.3), analysis of model parameters (Section 5.4), and the relationship between model slope parameter s and F2 slope (Section 5.4.3) are discussed.

A formant contour model is employed (1) to characterize the effect of speaking style and speaking rate on formant contours, and (2) to develop a method to synthesize natural formant contours.

5.2 Method: Modeling formant contours

The formant contour is modeled as a linear combination of target formant values of C_1, V, and C_2. The equation of the formant contour model is given as

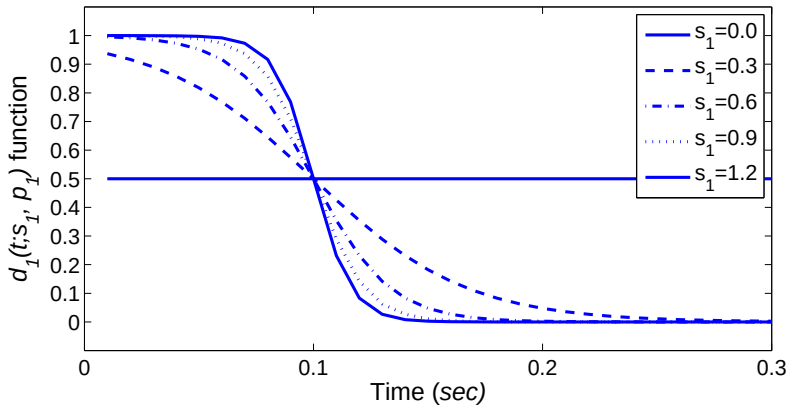
$$\begin{aligned}\widehat{\mathbf{F}}(t) &= \mathcal{d}_1(t; s_1, p_1)\mathbf{T}_{C_1} + (1 - \mathcal{d}_1(t; s_1, p_1) - \mathcal{d}_2(t; -s_2, p_2))\mathbf{T}_V + \mathcal{d}_2(t; -s_2, p_2) \cdot \mathbf{T}_{C_2} \\ &= \mathcal{d}_1(t; s_1, p_1) \cdot (\mathbf{T}_{C_1} - \mathbf{T}_V) + \mathbf{T}_V + \mathcal{d}_2(t; -s_2, p_2) \cdot (\mathbf{T}_{C_2} - \mathbf{T}_V) \qquad (5.1)\end{aligned}$$

where $\widehat{\mathbf{F}}(t)$ is the formant contour of a CVC word as a function of time t, as in a study by Niu and van Santen [74]. $\mathbf{T}_{C_1}$, $\mathbf{T}_V$, and $\mathbf{T}_{C_2}$ are the target formant vectors of the prevocalic consonant (C_1), vowel (V) and postvocalic consonant (C_2), respectively. Each target formant vector consists of the first four formant values (dimension 4×1). The function $\mathcal{d}(t; s, p)$ represents the degree of coarticulation of C_1 and C_2, which is proportional to the differences in target formant values. The exponential curve for the degree of coarticulation in Broad and Clerment [17] was changed to the sigmoid function

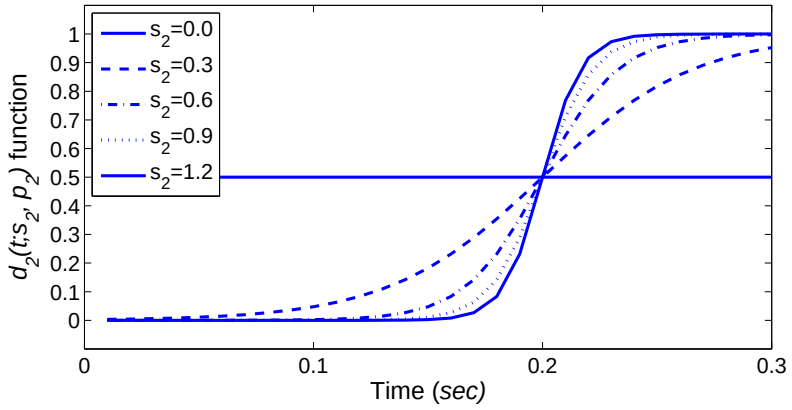
$$\mathcal{d}(t; s, p) = \frac{1}{1 + e^{s(t-p)}} \qquad (5.2)$$

This sigmoid function restricts the coarticulatory effects to be smoothly and monotonically decreasing or increasing. The coarticulation function $\mathcal{d}(t; s, p)$ is characterized with two coefficients s (slope) and p (slope position) of the sigmoid functions. While the formant contour is defined in Broad's study [17] from the onset to the offset of the vowel in C_1VC_2 contexts, we model the formant contour from the beginning of C_1 to the end of C_2 for sonorant consonants.

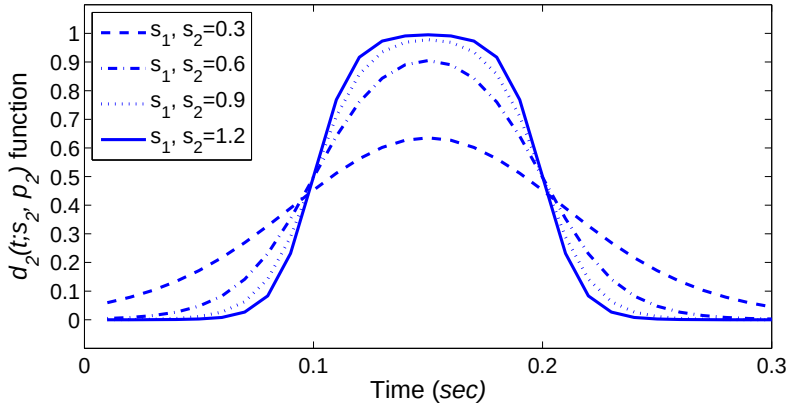
Figure 5.1 shows examples of coarticulation functions $\mathcal{d}_1$ and $\mathcal{d}_2$. Five values of s (from 0.0 to 1.2) for each function with fixed p value are shown where total word length equals to 0.3 (sec). The greater s value is associated with steeper slope, while the function values have limits from 0 to 1.



(a) $d_1(t; s_1, p_1)$ with 5 values of s and $p_1 = 0.1\,(\text{sec})$.



(b) $d_2(t; -s_2, p_2)$ with 5 values of s and $p_2 = 0.2\,(\text{sec})$.



(c) $(1 - d_1(t; -s_1, p_1) - d_2(t; -s_2, p_2))$ with 4 values of s, $p_1 = 0.1\,(\text{sec})$ and $p_2 = 0.2\,(\text{sec})$.

Figure 5.1: Examples of coarticulation function with fixed p for each function where total word length equals to $0.3\,(\text{sec})$.

5.2.1 Estimating model parameters

Four coefficients in $\mathscr{d}(t;s,p)$ are estimated by minimizing the error function $Err1$, while a total of 12 target values from F1 to F4 ($\mathbf{T}_{C_1}$, $\mathbf{T}_V$, and $\mathbf{T}_{C_2}$) are estimated by minimizing $Err2$.

$$Err1^{(k)} = \frac{\mathbf{w}\sum_{t=T_1^{(k)}}^{T_2^{(k)}}\left(\widehat{\mathbf{F}}(t)^{(k)} - \mathbf{F}(t)^{(k)}\right)^2}{N^{(k)}} \tag{5.3}$$

$$Err2 = \sum_{k=1}^{K} Err1^{(k)} \tag{5.4}$$

where $\mathbf{F}(t)^{(k)}$ and $\widehat{\mathbf{F}}(t)^{(k)}$ are the observed and estimated (with given parameters) formant contours in the k-th word, respectively. $N^{(k)}$ is the number of frames from $T_1^{(k)}$ to $T_2^{(k)}$. K is the number of words ($K = 4$) used in the training set for a particular style. The contribution to the error from F1 and F2 are weighted more than F3 and F4, represented as $\mathbf{w} = \begin{bmatrix}1 & 1 & 0.25 & 0.01\end{bmatrix}$. The error is not divided by the sum of the weights, because it does not affect the ranking. Formant frequencies are converted from Hertz to Bark frequency scale, which ranges from 1 to 24 and corresponds to the first 24 critical bands of hearing [103]. $Err2$ is approximately K times greater than $Err1$.

The time t ranges from T_1 to T_2, which are determined from the consonant's manner of articulation. When C_1 (or C_2) is an approximant (i. e. /w/ or /l/), T_1 (or T_2) is located at the middle of C_1 (or C_2), respectively. The time t does not start at $t = 0$ and end at the end of the word, because of the coarticulation effects on C_1 from the phoneme preceded by C_1 and the effect of weak energy at the end of the word on C_2. When C_1 (or C_2) is an unvoiced consonant (i. e. /t/), T_1 (or T_2) equals the onset (or offset) of V, respectively.

While coarticulation functions $\mathscr{d}(t;s,p)$ are estimated for each token (k), formant target values, $\mathbf{T}_{C_1}$, $\mathbf{T}_V$, and $\mathbf{T}_{C_2}$, are estimated in two ways. One is to estimate one set of target values for each speaking style (style-dependent target), and the other is to estimate one set for all speaking styles (global target). The difference is whether a speaker is thought to have a distinct formant target depending on the speaking style, or an identical formant target regardless of speaking style. For global-target estimation, Equation 5.4 becomes

$$Err2' = \sum_{style=1}^{S}\sum_{k=1}^{K} Err1^{(k)} \tag{5.5}$$

where S is the number of speaking styles ($S = 4$).

For better estimation, parameters are constrained as follows:

- $d_1(t; s_1, p_1) + d_2(t; -s_2, p_2) \leq 1.0$.
- $s > 0$ in Equation 5.2.
- p in Equation 5.2 is within 50 ms of the corresponding phoneme boundary.
- $200 < F1 < 900$, $400 < F2 < 3000$, $1800 < F3 < 3700$, $2500 < F4 < 5000$.
- $F2 - F1 > 200$, $F3 - F4 > 200$, $F4 - F3 > 200$.

Model parameters are estimated using a hill-climbing approach. First, coefficients s in $d(t; s, p)$ are initialized with $s = 0.7$ for both functions, while coefficients p are initialized to be the corresponding phoneme boundaries. For style-dependent target estimation, formant target values for /w/, /V/ and /l/ are initialized with the observed values at the middle of each phoneme for the approximants, while the targets for the consonant /t/ are initialized with values from the table provided by Allen *et al.* shown in Table C.1 [1]. For global-target estimation, all target values are initialized with the values provided by Allen *et al.*. The initial step size is 0.1 for s, 1.0 frame for p, and 50 Hz for the formant target. We evaluate one parameter at a time. The order of evaluating parameters is s_1, p_1, s_2, p_2 for $Err1$, and $F1_{C_1}$, $F1_V$, $F1_{C_2}$, $F2_{C_1}$, $F2_V$, $F2_{C_2}$, and so on for $Err2$. One iteration is completed when all four (or twelve) parameters are evaluated in $Err1$ (or $Err2$). For each parameter, every time $Err1$ (or $Err2$) increases, the direction of the change is switched. After each parameter is evaluated in both directions, we switch to the next parameter. When the error reduction is less than 10 % of the previous error value at the end of an iteration, the step size is halved. The iterative process continues until the error reduction becomes less than a threshold ($\epsilon = 10 \times 1.0^{-5}$), or the maximum iteration (30) is reached. The step size at the exit must also be less than 0.001 for s and 0.01 for p, otherwise the iteration continues.

The outputs from this operation are s and p for the $d(t; s, p)$, and the target formant matrix $[\mathbf{T}_{C_1}, \mathbf{T}_V, \mathbf{T}_{C_2}]$ (4×3).

5.3 Results of formant contour model

We now describe the results of the formant contour model on the corpus from Chapter 4. For the experiments, we used four /w/–/V/–/l/ words (*wheel, will, well*, and *wail*)

and one /t/–/V/–/l/ word (*tool*). Each word was repeated 16 times in CNV, CNV/SLOW, CLR/FAST and CLR styles; intelligibility was previously examined in Chapter 4.4. The total of 320 (5 words × 4 styles × 16 repetitions) samples were analyzed.

We compare the modeled formant contour with the observed data in training or test sets. In Equation 5.3, the number of words used (K) is 4 in the training set and 12 in the test set. $\widehat{\mathbf{F}}(t)^{(k)}$ in Equation 5.1 is computed with the specified s_1 and s_2 (i. e. values shown in Table 5.1) in $\acute{d}(t; s, p)$, and target values ($\mathbf{T}_{C_1}$, $\mathbf{T}_V$, $\mathbf{T}_{C_2}$) estimated with either style-dependent or globally estimated formant targets. The slope location represented by p is not our point of interest. Therefore, p_1 and p_2 are adjusted to provide the best fit by minimizing $Err1$ in Equation. 5.3 given s_1, s_2, and the target values. A more detailed description is provided below in Section 5.3.1.

5.3.1 Goodness of fit

The normalized sum of least squares was calculated as,

$$E^{(k)}_{s,target} = \frac{\mathbf{w} \sum_{t=T_1^{(k)}}^{T_2^{(k)}} \left(\widehat{\mathbf{F}}(t)^{(k)} - \mathbf{F}(t)^{(k)} \right)^2}{N^{(k)}} \tag{5.6}$$

where $\mathbf{F}(t)^{(k)}$ and $\widehat{\mathbf{F}}(t)^{(k)}$ are the observed and estimated formant contours in the k-th word, respectively. The weight ($\mathbf{w}$) was set to $\begin{bmatrix} 1 & 1 & 0.25 & 0.01 \end{bmatrix}$ as in Section 5.2.1. The error is not divided by the sum of the weights, because it does not affect the ranking. $N^{(k)}$ is the number of frames from $T_1^{(k)}$ to $T_2^{(k)}$, which contribute to the error. $E_{s,target}$ is calculated for each word k, and its subscripts $s, target$ indicate that the error depends on the variables s and formant target values. Mean $E_{s,target}$ values over the samples in the training and test sets are reported.

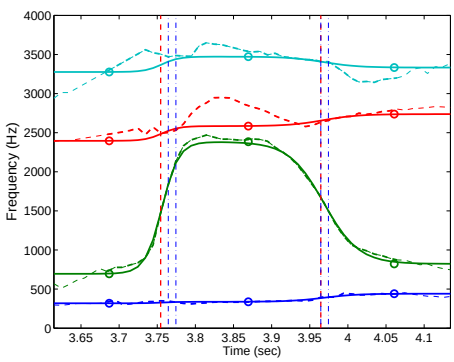
Figures 5.2(a) and 5.2(c) show examples of estimated formant contours in the test set with Average$_{vowel,style}$ value for s and style-independent target values (the word "*wheel*" in CLR and CNV speaking style). Solid lines are estimated formant contours (F1 through F4), while the observed data are shown in a thick dotted line during their contribution to the error (from $T_1^{(k)}$ to $T_2^{(k)}$). The estimated style-independent target values are represented with circles, which are common in both speaking styles (Figures 5.2(a) and 5.2(c)). Blue vertical lines represent beginnings and endings of the phoneme, while red vertical lines represent the slope location (p), indicating steepest point of the slope, which was adjusted to best fit the given data. Figures 5.2(b) and 5.2(d) show examples of estimated

Table 5.1: Experiment in goodness of fit with different configurations and error rates (in Bark squared). Cfg. 8 shown in bold font is used for further analysis in Section 5.4.

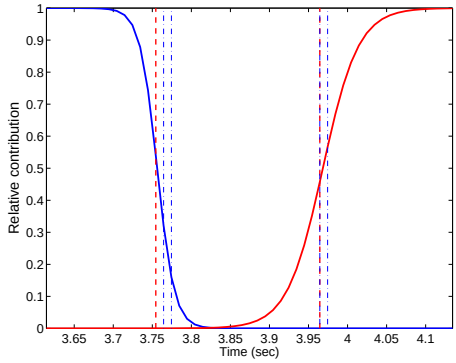
Cfg.	Coefficient s_1 and s_2	Target Values	$\mathbf{F}(t)^{(k)}$	Mean $E_{s,target}$
1	Estimated	Style-dependent	Training set	0.0426
2	Average$_{vowel,style}$	Style-dependent	Training set	0.0591
3	Average$_{vowel,style}$	Style-dependent	Test set	0.0764
4	Average$_{vowel}$	Style-dependent	Test set	0.0879
5	Average$_{style}$	Style-dependent	Test set	0.0861
6	Global mean	Style-dependent	Test set	0.0943
7	Random value	Style-dependent	Test set	0.3155
8	**Estimated** (estimated with global target)	**Globally estimated**	Training set	0.0656
9	Average$_{vowel,style}$	Globally estimated	Training set	0.0732
10	Average$_{vowel,style}$	Globally estimated	Test set	0.0833
11	Average$_{vowel}$	Globally estimated	Test set	0.0993
12	Average$_{style}$	Globally estimated	Test set	0.0913
13	Global mean	Globally estimated	Test set	0.1056
14	Random value	Globally estimated	Test set	0.3676
15	Estimated (estimated with generic target)	Generic values	Training set	0.4233
16	Average$_{vowel,style}$	Generic values	Training set	0.4305
17	Average$_{vowel,style}$	Generic values	Test set	0.4322
18	Average$_{vowel}$	Generic values	Test set	0.4415
19	Average$_{style}$	Generic values	Test set	0.4446
20	Global mean	Generic values	Test set	0.4556
21	Random value	Generic values	Test set	0.6716
22	Random value	Random values	Test set	5.0718

$\acute{d}_1(t; s_1, p_1)$ and $\acute{d}_2(t; -s_2, p_2)$ with Average$_{vowel,style}$ value for s_1, s_2 and best-fit values for corresponding p_1 and p_2.

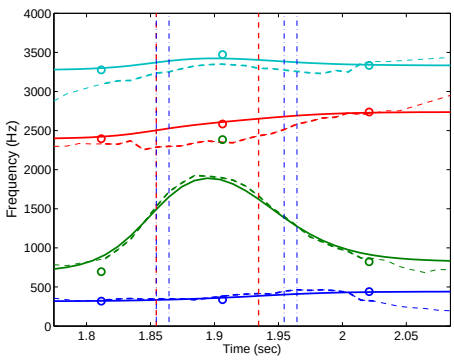
In order to evaluate the error range for this model, we model formant contours with six different values of coefficient s for $\acute{d}(t; s, p)$, and four sets of target values. For the coefficient s, the following six values are evaluated: (1) estimated s for each sample in training (estimated per token), (2) estimated s averaged over four samples in the training set for each speaking style and vowel (Average$_{style,vowel}$), (3) estimated s averaged over 16 (4 samples $\times$ 4 styles) samples (Average$_{vowel}$), (4) estimated s averaged over 20 (4 samples $\times$ 5 vowels) samples (Average$_{style}$), (5) estimated s averaged over 80 (4 samples $\times$ 4 styles $\times$ 5 vowels) samples (global mean), and (6) pseudo-random values from a uniform



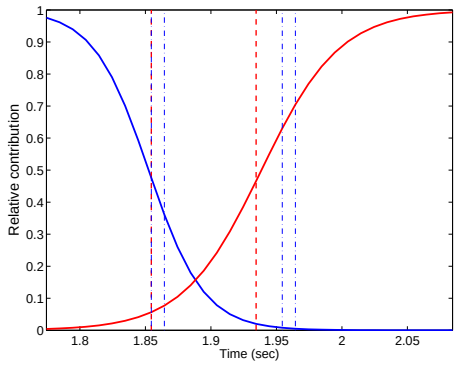
(a) Estimated (solid lines) and observed (dotted lines) formant contour spoken in CLR are shown (circle).



(b) Estimated $\acute{d}_1(t; s_1, p_1)$ (solid line) and $\acute{d}_2(t; -s_2, p_2)$ (dashed line) for CLR speech style in the test set. The estimated formant with Average$_{vowel,style}$ s values ($s_1 = 0.9096$ target values, which are optimized globally, and $s_2 = 0.4390$).



(c) Estimated (solid lines) and observed (dotted lines) formant contour spoken in CNV style in the test set. The estimated formant target values, which are optimized globally, are shown (circle).



(d) Estimated $\acute{d}_2(t; -s_2, p_2)$ (solid line) and $\acute{d}(t; s_2, p_2)$ (dashed line) for CNV speech with Average$_{vowel,style}$ s values ($s_1 = 0.4740$ and $s_2 = 0.3352$).

Figure 5.2: The results of formant contour model (Cfg. 10) for the word "*wheel*" in two speaking styles (CLR and CNV). In all cases, blue vertical dash-dot lines show the phoneme boundaries, while vertical red dashed lines represent p_1 and p_2.

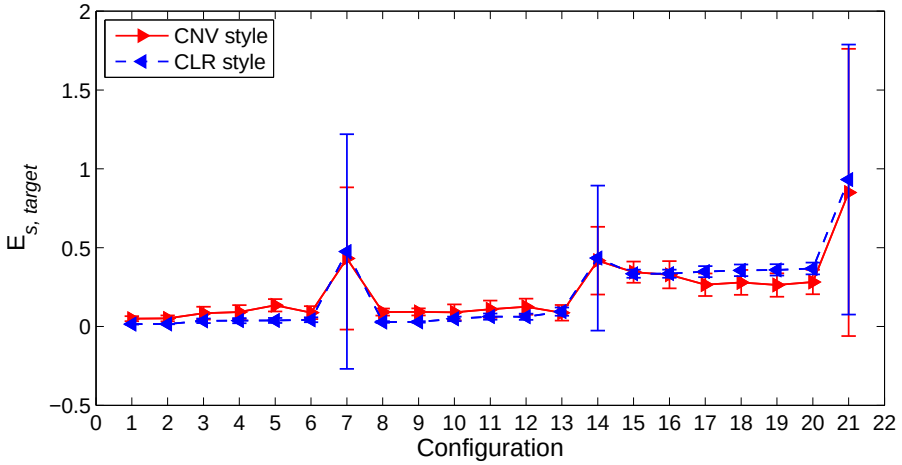
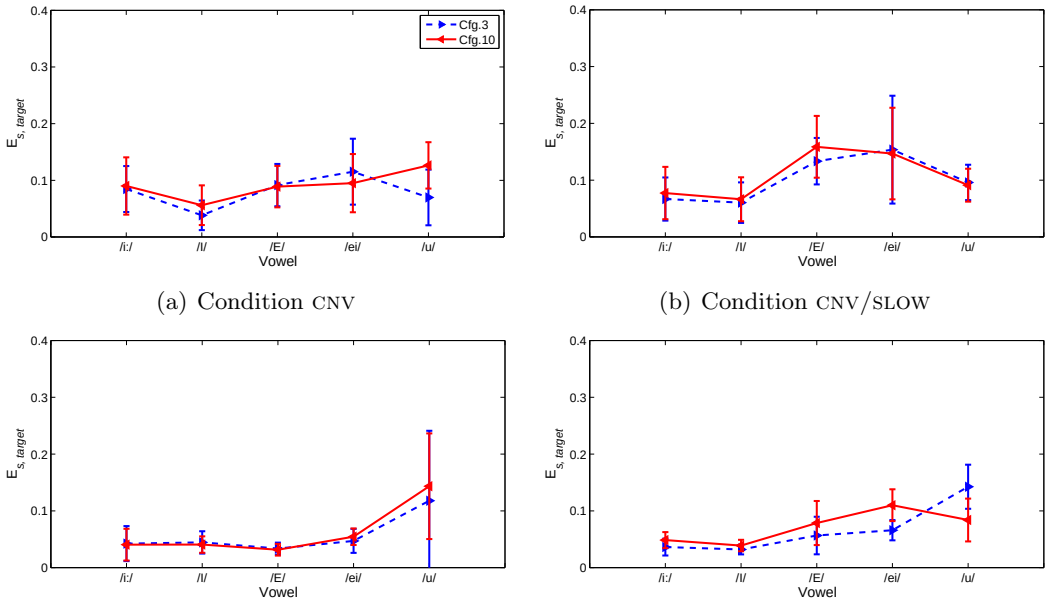


Figure 5.3: Mean $E_{s,target}$ values and standard deviations for 21 configurations for the vowel /iː/. Two conditions are CNV and CLR styles.



(a) Condition CNV

(b) Condition CNV/SLOW

(c) Condition CLR/FAST

(d) Condition CLR

Figure 5.4: Mean $E_{s,target}$ value for each vowel in four conditions.

distribution between 0 and the maximum estimated value of coefficient s (random value). For the target values, the following four sets of values are evaluated: (1) estimated targets for each speaking style (style-dependent target), (2) estimated targets for all speaking styles (globally estimated), (3) generic values provided by Allen *et al.* [1] independent of the speaking style (generic values shown in Table C.1 in Appendix C), and (4) random values with a uniform distribution with the restrictions of $200 < F1 < 900$, $400 < F2 < 3000$, $1800 < F3 < 3700$, $2500 < F4 < 5000$, $F2 - F1 > 200$, $F3 - F4 > 200$, $F4 - F3 > 200$ (random values). Note when the target values are changed, $\mathcal{d}(t; s, p)$ functions are estimated with given target values. For example, for Cfg. 9 in Table 5.1, coefficients s_1 and s_2 for $\mathcal{d}(t; s, p)$ are estimated with global target values in the training set, and averaged over four samples. The errors are evaluated as compared with the observed formant contours ($\mathbf{F}(t)^{(k)}$) in either the training or the test sets. Twenty two configurations in total were designed for the error analysis. For the configurations involving random values (Cfg. 7, 14, 21 and 22), the process of obtaining $E_{s,target}$ was repeated ten times and averaged over all samples.

Table 5.1 shows the configuration number, the source of coefficients s, source of target values, formant contours to be evaluated, and the resulting mean $E_{s,target}$ values over all samples. These configurations, and their order were selected in order to evaluate the performance of the model. It was expected that the error within a set of target values (Cfg. 1–7, 8–14, and 15–21) would increase with each configuration, and that the error would increase from style-dependent to globally-estimated to random target values. While Cfg. 1 (0.0426) shows the minimum error, Cfg. 22 (5.0718) shows the maximum error expected in the modeling.

As a result of error analysis, the $E_{s,target}$ difference between Cfg. 2 (0.0591) and 3 (0.0764) shows the difference between the training and test sets with style-dependent target values. The $E_{s,target}$ difference between Cfg. 9 (0.0732) and 10 (0.0833) shows the difference between the training and test sets with global target values. In general, the errors with style-dependent target values are smaller than those with global target values. It indicates that style-dependent target values allow the formant contour model to fit the data better than the globally estimated target values. On the other hand, the difference in errors between Cfg. 9 and 10 is smaller than that of Cfg. 2 and 3. It shows that globally estimated targets provide estimations that generalize better than style-dependent targets. Relatively little over-fitting to the training set was observed in global-target values. Cfg. 5 or 12 $Average_{style}$ (0.0861, 0.0913) yielded smaller error than Cfg. 4 or 11

$Average_{vowel}$ (0.0879, 0.0993). This indicates that coefficients s_1 and s_2 are more sensitive to the speaking style than vowel identity, as the error increases with average over the style (Average$_{vowel}$). These results confirm our expectations of the model.

In order to understand the relationship between difference in errors and perceptible change, we computed the approximate just noticeable difference (JND) from previous work. Kewley-Port and Watson report that thresholds for formant-frequency discrimination are about 14 Hz in the F1 frequency range (below 800 Hz) and about 1.5% in the F2 frequency range [51]. According to their results, the thresholds for the neutral vowel (F1=500 Hz, F2=1500 Hz) are 14 Hz (1.2039 Bark squared) and 23 Hz (2.2350 Bark squared) for F1 and F2, respectively. Our approximated JND of the neutral vowel is 2.2350 Bark summed over all 4 formants. The errors in Cfg. 1–6 and Cfg. 8–13 (Table 5.1) are within the approximated JND, even though thresholds only for F1 and F2 are considered.

Figure 5.3 shows mean $E_{s,target}$ distribution for 21 configurations for the vowel /iː/ in two conditions, CNV and CLR styles (other two speaking styles are not shown). For the error difference between the CNV and CLR styles, in general the formant contour model fits better in CLR style except for Cfg. 17–20. This is because the F1 contour of CNV speech usually has an abrupt drop towards the end of the word as shown in Figure 5.2(c). The formant contour model does not fit well in this region, which causes higher error in the CNV style. In both speaking styles, the error gradually increases from Cfg. 1 to Cfg. 7, from 8 to 14, and from 15 to 21 as coefficient s changes from a specific to a more generalized value. The error in the CLR style is larger than CNV speech in Cfg. 17–20, because the generic values in those configurations are far from the observed CLR formant contour and the gap between the generic target and observed values was not compensated with the degree of coarticulation (s) only in Cfg. 17–20.

Figure 5.4 shows mean $E_{s,target}$ distribution in Cfg. 3 and 10 for each vowel in four conditions. The $E_{s,target}$ difference in Cfg. 3 and 10 is particularly large for /ei/ in CLR condition. We speculate that in Cfg. 10, with globally estimated targets, formant movements of CLR speech are too dynamic to model with only coefficients s_1 and s_2 in $\acute{d}(t; s, p)$ function, as opposed to Cfg. 3 which uses style-dependent target values.

5.4 Characterizing formant shapes in terms of speaking styles and speaking rates

In this section, we analyze the effect of speaking rates and styles on formant contours by examining coefficient s_1 and s_2 in the $\acute{d}(t;s,p)$ function and the estimated target values.

5.4.1 Estimated $\acute{d}(t;s,p)$ parameters

We examine whether coarticulation functions ($\acute{d}(t;s,p)$) depend only on speaking style or on the combination of speaking style and speaking rate. We use the results from Cfg. 8, in which target values are estimated globally. Setting the target values constant for all speaking styles (global target) allows us to examine the differences in $\acute{d}_1(t;s_1,p_1)$ and $\acute{d}_2(t;s_2,p_2)$ functions as a function of speaking style.

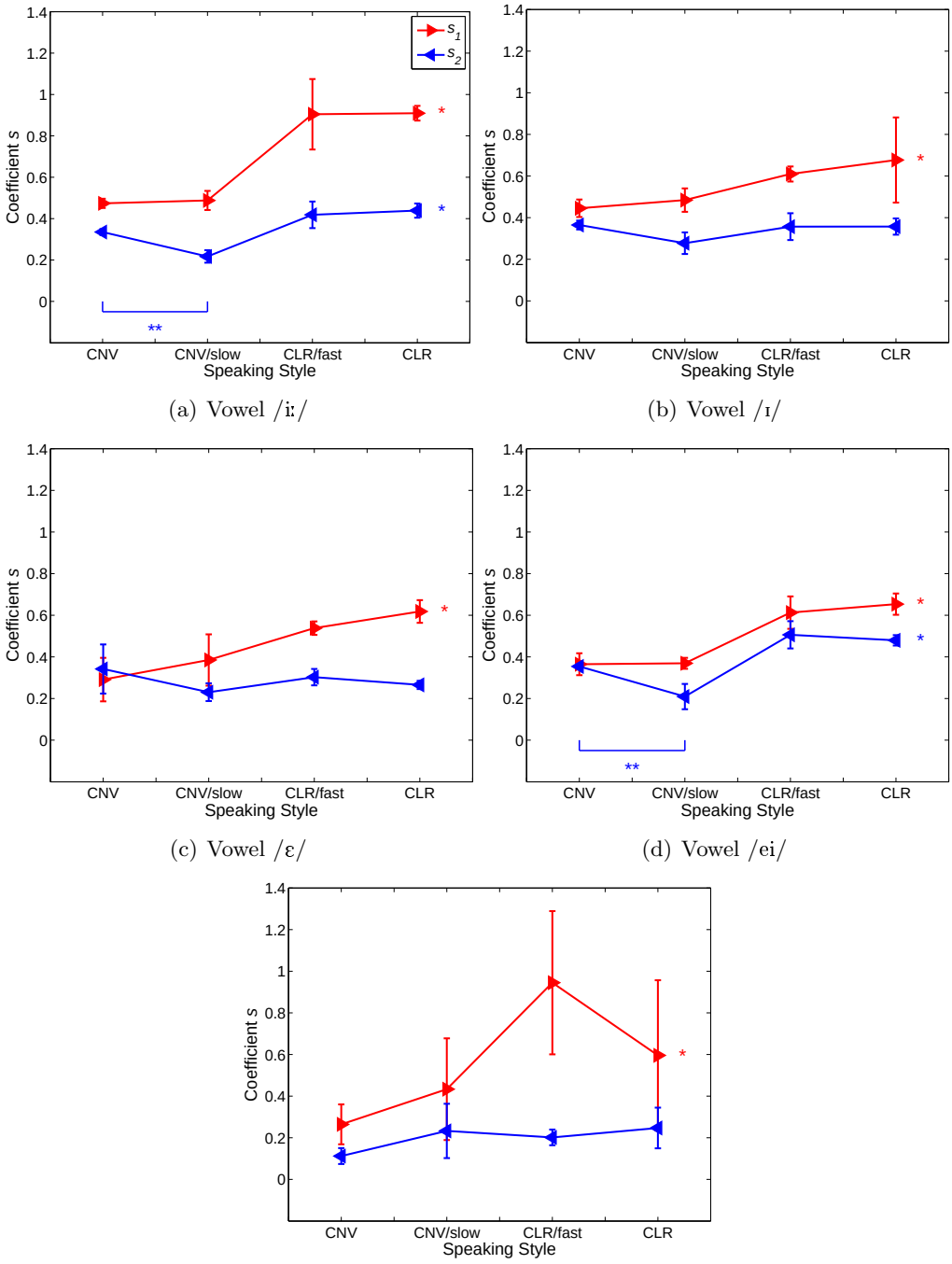
The characteristics of formant slope are represented by coefficient s_1 and s_2 in Equations 5.2. Figure 5.5 shows the estimated mean values of coefficient s_1 and s_2 in $\acute{d}(t;s,p)$ for the four conditions. A higher coefficient indicates steeper slope.

The effects of speaking style on the coefficient s in $\acute{d}(t;s,p)$ in each vowel were analyzed with a one-way analysis of variance (ANOVA). The results of s_1 (C_1 to V transition) showed that the main effects of speaking style were significant for all vowels, ($F_{/i:/}(1,14) = 102.5, p = 8.0137 \times 10^{-8}$, $F_{/I/}(1,14) = 11.43, p = 0.0045$, $F_{/E/}(1,14) = 26.65, p = 0.0001$, $F_{/ei/}(1,14) = 99.96, p = 9.3663 \times 10^{-8}$, $F_{/u/}(1,14) = 7.93, p = 0.0137$). Post-hoc tests (HSD) for speaking style effects by all vowels showed that the slope was steeper for the CLR speech than the CNV speech ($p < 0.05$).

The results of s_2 (V to C_2 transition) showed that the main effects of speaking style were significant for two vowels /iː/ and /ei/, ($F_{/i:/}(1,14) = 27.17, p = 0.0001$, $F_{/ei/}(1,14) = 35.87, p = 3.31562 \times 10^{-5}$), but not the vowels /ɪ/, /ɛ/, and /u/ ($p = 0.2071$, $p = 0.9572$, $p = 0.2839$, respectively), Post-hoc tests (HSD) for speaking style effects by two vowels (/iː/, and /ei/) showed that the slope was steeper for the CLR speech than the CNV speech ($p < 0.05$).

A two-sample t-test was performed comparing the differences between CNV and CNV/SLOW, and between CLR and CLR/FAST. None of the vowels showed a significant difference in s_1 due to speaking rate ($p > 0.05$). In terms of s_2, the difference between CNV and CNV/SLOW was significant for the two vowels (/iː/ and /ei/) shown with asterisks (** in Figure 5.5), while none of the vowels showed a difference between CLR and CLR/FAST (all $p > 0.05$).

The question of whether coarticulation functions ($\acute{d}(t;s,p)$) depend only on speaking



(a) Vowel /iː/

(b) Vowel /ɪ/

(c) Vowel /ɛ/

(d) Vowel /ei/

(e) Vowel /u/

Figure 5.5: Average values of coefficient s_1 in $d_1(t; s_1, p_1)$ (red) and $d_2(t; s_2, p_2)$ (blue) functions for four conditions in each vowel. Asterisks (*) on the right-hand side indicate significant main effect of the speaking style ($p < 0.05$). The significant differences between speaking rates (t-test, $p < 0.01$) are indicated with asterisks (**).

style, or the combination of speaking style and speaking rate, can be answered as follows: In general, slopes are determined independent of the speaking rate, but are dependent on the speaking style. The effect of speaking rate, when observed, is often observed at the offset of the vowel. It is not yet clear if the dependency of the slope on the speaking rate (CNV and CNV/SLOW difference) is perceptually important. The slopes s_1 and s_2 in CLR speech are generally steeper than those of CNV speech.

5.4.2 Estimated formant target values

Figure 5.6(a) shows the results of estimated formant targets (Cfg. 1–7). The observed data are shown in blue bold (CLR style) and in red italics (CNV style) in all three figures. Style-dependent target values for CLR speech (shown in red) and for CNV speech (shown in blue) are estimated separately and shown in boxes. Figure 5.6(b) shows the results of globally estimated formant targets (Cfg. 8–14). Figure 5.6(c) shows the generic formant targets (Cfg. 15–21). The estimated values of the front vowels tend to be located at the most extreme F1 and F2 values, regardless of local or global estimation. This is due to a characteristic of the formant model, in which the formant values cannot be estimated in the direction of overshoot (which is not observed in normal speech), but only in the direction of undershoot (which is common in normal speech).

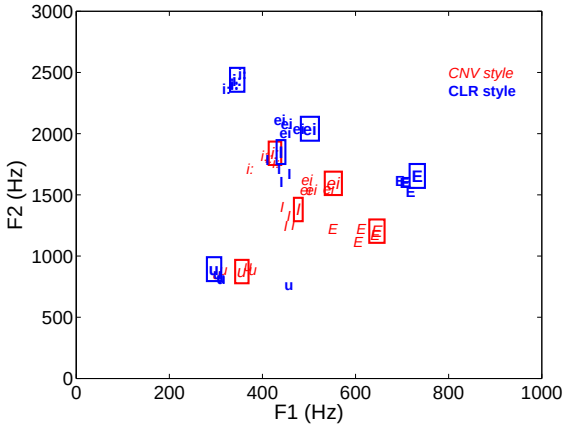
Although the estimated target values are far away from the observed CNV formant values, the mean values in error between CLR and CNV speech are reasonably close (Figure 5.3). This indicates that we can model formant contours with formant target values, that do not depend on speaking style.

Our hypothesis in estimating formant target values is that a speaker sets either different formant target values for different speaking styles, or one global target value per vowel regardless of speaking style. With limited data from one CVC context, it is difficult to develop a reasonable answer to this hypothesis. We address this hypothesis when we increase the number of words with a variety of phoneme contexts (Chapter 6).

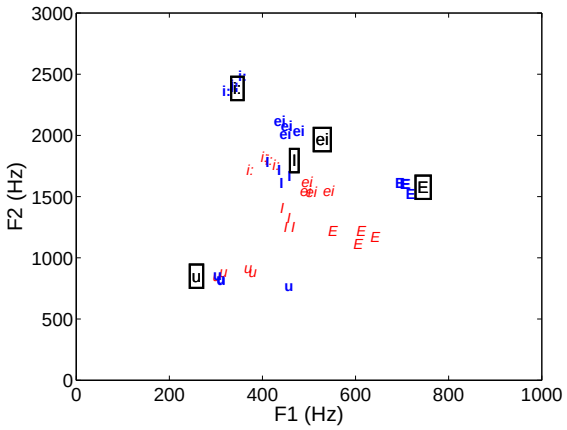
5.4.3 Relationship between model parameters and F2 slope

The F2 slope at the phoneme boundary represents one parameterization of the dynamics of the formant contour. We examine whether the direct measure of F2 slope and the coarticulation coefficient s are correlated.

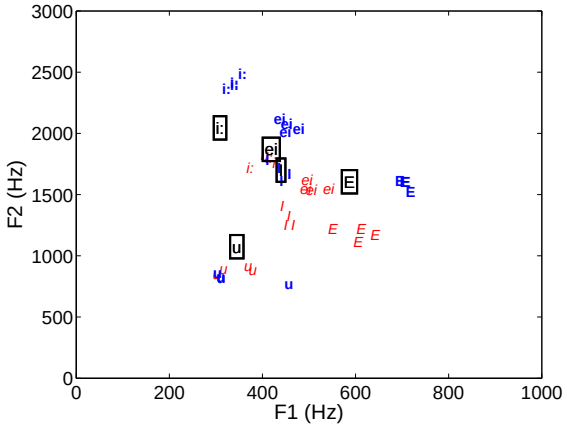
The F2 slope was measured over the 20 ms at the phoneme boundary by fitting a



(a) Style-dependent formant target values (Cfg. 1–7).



(b) Globally estimated formant target values (Cfg. 8–14).



(c) Generic target values (Cfg. 15–21).

Figure 5.6: Three sets of formant target values (style-dependent, globally estimated, and generic target values). The observed data are shown in blue bold (CLR style) and in red italics (CNV style) in all three figures.

straight line to the observed data at C_1V and VC_2 transitions. The vowel /u/ in /t/–/V/–/l/ is excluded from this analysis because its formant shape is monotonically decreasing, while those of front vowels in a /w/–/V/–/l/ context have a concave shape. The mean over the 16 repetitions and four front vowels of F2 slope for CLR and CNV styles were 20.91 Hz/ms, 11.09 Hz/ms at vowel onset and 9.57 Hz/ms, 6.31 Hz/ms at vowel offset, respectively (described in Chapter 4.3.2). Similar to the results in Section 5.4.1, the main effect of speaking style, both at C_1V and VC_2 transitions, was significant (both $p < 0.05$). Speaking rate had a significant effect only at the vowel offset using paired t-tests between CNV and CNV/SLOW ($p < 0.05$), and CLR and CLR/FAST ($p < 0.05$). Being consistent with the study by Moon and Lindblom [68], this indicates that the movement of articulators producing CLR speech was faster than in CNV speech. On the other hand, the speaking rate was not a key determinant of the rate of movement of articulators at the vowel onset transition.

Figure 5.7 shows the relationship between the direct measurements of F2 slope and coefficient s_1 and s_2 in the $\mathcal{d}(t; s, p)$ function with four speaking conditions (CNV, CNV/SLOW, CLR, and CLR/FAST). The estimated coefficients s_1 and s_2, which represent a coarticulation measure of F1 through F4, are strongly correlated with the F2 slope (Hz/ms) (Pearson's correlation coefficient $r = 0.8527$, $p < 0.05$). If the direct measure of F2 slope is good enough to predict coarticulation effects in the formant contour model, the parameter estimation may be simplified in the future, requiring only target estimation.

5.5 Conclusions

In this chapter, we discussed a method to model the formant contour with coarticulation functions ($\mathcal{d}(t; s, p)$) and target values. We presented preliminary results on five vowels with /w/–/V/–/l/ and /t/–/V/–/l/ contexts, and discussed the characteristics of formant shapes by examining estimated coefficients s_1 and s_2 in $\mathcal{d}(t; s, p)$. The advantages of modeling the formant contour are the following.

- It allows us to analyze formant targets separately from the coarticulation effects.
- It allows us to parameterize the coarticulation effects.
- It helps to synthesize natural formant contours with given phoneme durations.

Our results from analysis of modeled formant contours indicate: (1) the slope depends on speaking style for the vowels at onset, and depends on speaking style and the speaking

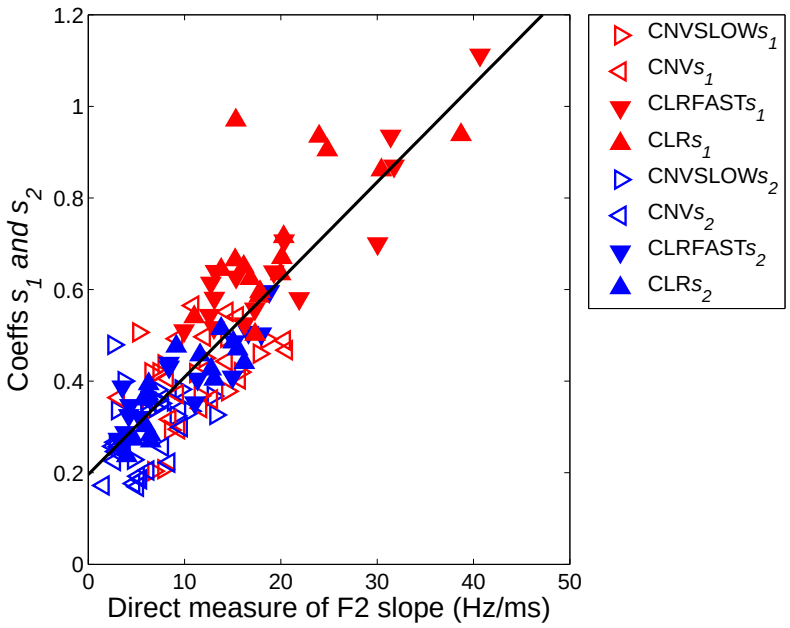


Figure 5.7: Relationship between direct measure of F2 slope and coefficients s_1 and s_2 in $\acute{d}_1(t; s_1, p_1)$ (vowel onset) and $\acute{d}_2(t; -s_2, p_2)$ (vowel offset) functions. All four speaking styles were combined.

rate for some vowels at offset, and (2) slopes at the vowel onset are steeper for CLR speech than for CNV speech for the front vowels. The direct measurement of F2 slope was strongly correlated with coarticulation coefficients, which may lead to less parameters to estimate.

The results from analyzing goodness of fit of the model indicate that the errors in the vowel /ei/ in general are higher. The vowel /ei/ may require two targets, as it is considered a diphthong. We may also need to investigate different basis functions, other than the sigmoid function, for coarticulation functions $\acute{d}(t; s, p)$. One idea is to use a non-symmetric sigmoid function which is characterized with two coefficients for the slope. Estimating the coarticulation function $\acute{d}(t; s, p)$ per formant also leads to a better fit, although it requires more parameters to estimate. There is a trade off between fewer parameters to estimate and higher error in estimated formant contours. The selection of the better model depends on whether the difference between the complex and simple model makes any perceptual difference. Due to the scope of this thesis, we do not further investigate such models. When we obtain a variety of CVC words, we expand the formant contour model into context-independent formant contour model. Then, we determine whether the speakers have different target values for CLR and CNV speech (Chapter 6).

Chapter 6

Effect of speaking style on formant contours with a variety of phoneme contexts

6.1 Introduction

Previously, we successfully improved intelligibility of CNV speech to CLR speech levels by modifying formant contours and durations to resemble those of CLR speech (Chapter 4). However, synthesizing a formant contour based on CLR speech steady-state values (or targets) and CNV speech durations requires a method to appropriately control the formant transitions, since formant transitions are an important perceptual cue for intelligibility [33], as well as the entire contour for diphthongs.

In order to capture the difference in formant contour shapes as a function of duration and speaking style, we presented a method to model formant contours and results of model parameter estimation, with a limited number of consonants and vowels (/w/–/V/–/l/ and /t/–/V/–/l/ words) (Chapter 5). The importance of this model is that we can synthesize realistic formant contours with a given phoneme duration, as opposed to linear compression or expansion of existing contours. Analyzing the model parameters also allows us to characterize the difference in speaking style. The contour model fitted well to the observed data (error as low as 0.2062 Bark[1]) with a linear combination of style-dependent target values and coarticulation functions.

In Chapter 5, we raised an important question in estimating formant target values; whether a speaker sets different formant target values for different speaking styles, or

[1]Note that the error rate reported in this chapter was changed to Bark from Bark squared in previous chapter.

Table 6.1: Number of occurrences (percentage) of the vowels in our speech corpus and CMU dictionary.

Vowel	Our speech corpus		CMU dictionary	
iː	37	(15.29 %)	574	(16.01 %)
ɪ	37	(15.29 %)	482	(13.44 %)
ɛ	31	(12.81 %)	450	(12.55 %)
æ	36	(14.88 %)	460	(12.83 %)
u	24	(9.92 %)	333	(9.29 %)
ʊ	6	(2.48 %)	73	(2.04 %)
ʌ	25	(10.33 %)	335	(9.34 %)
ɑ	46	(19.01 %)	878	(24.49 %)
SUM	242		3585	

one global target value per vowel regardless of speaking style. To investigate the style dependency of the target values, we estimate formant targets by two methods, style-independent and style-dependent estimation on a variety of *CVC* words in this chapter[2]. The mean error rate is examined to compare the performance from the two methods.

In this chapter, we present a newly developed speech corpus (consonant-vowel-consonant words) with two speaking styles (Section 6.2), and acoustic analyses of the speech corpus (Section 6.3). We briefly describe the previously developed formant contour model (Section 6.4). In Experiment 6–1, we discuss the parameter estimation procedure (style-independent and dependent), the goodness of fit, and analysis of estimated model parameters (Section 6.5). Experiment 6–2, we focus on style-independent target estimation with an increased amount of training data. We report on the data-driven consonant target values, as opposed to a rule-based approach (Section 6.6). Finally, we discuss the similarities and dissimilarities between two speakers in terms of speech production and the behavior of the contour model (Section 6.7).

6.2 Text material and recording (CVC words)

6.2.1 Creating CVC words

242 *CVC* meaningful words, consisting of 23 initial and final consonants and 8 vowels, are created. The vowels in the list include both front (/iː/, /ɪ/, /ɛ/, /æ/) and back (/u/, /ʊ/, /ʌ/, /ɑ/) vowels, excluding diphthongs. The vowel /ɔ/ (as in *caught*) is combined with

[2]Part of this chapter was published in Amano-Kusumoto *et al.* [4].

Table 6.2: Number of occurrences (percentage) of the consonants in our speech corpus and CMU dictionary.

	Our speech corpus				CMU dictionary			
Consonant	Prevocalic (C_1)		Postvocalic (C_2)		Prevocalic (C_1)		Postvocalic (C_2)	
p	12	(4.96 %)	19	(7.85 %)	179	(4.99 %)	198	(5.52 %)
t	13	(5.37 %)	24	(9.92 %)	171	(4.77 %)	243	(6.78 %)
k	14	(5.79 %)	23	(9.50 %)	279	(7.78 %)	338	(9.43 %)
b	20	(8.26 %)	8	(3.31 %)	256	(7.14 %)	117	(3.26 %)
d	15	(6.20 %)	18	(7.44 %)	212	(5.91 %)	179	(4.99 %)
g	9	(3.72 %)	9	(3.72 %)	133	(3.71 %)	125	(3.49 %)
s	15	(6.20 %)	19	(7.85 %)	201	(5.61 %)	233	(6.50 %)
ʃ	7	(2.89 %)	7	(2.89 %)	175	(4.88 %)	119	(3.32 %)
f	11	(4.55 %)	7	(2.89 %)	139	(3.88 %)	120	(3.35 %)
v	2	(0.83 %)	3	(1.24 %)	68	(1.90 %)	42	(1.17 %)
θ	3	(1.24 %)	3	(1.24 %)	41	(1.14 %)	84	(2.34 %)
ð	2	(0.83 %)	0	(0.00 %)	16	(0.45 %)	8	(0.22 %)
z	2	(0.83 %)	11	(4.55 %)	78	(2.18 %)	248	(6.92 %)
ʧ	6	(2.48 %)	10	(4.13 %)	74	(2.06 %)	131	(3.65 %)
ʤ	10	(4.13 %)	5	(2.07 %)	126	(3.51 %)	41	(1.14 %)
l	21	(8.68 %)	21	(8.68 %)	257	(7.17 %)	358	(9.99 %)
ɻ	17	(7.02 %)	7	(2.89 %)	286	(7.98 %)	290	(8.09 %)
j	7	(2.89 %)	0	(0.00 %)	81	(2.26 %)	0	(0.00 %)
w	14	(5.79 %)	0	(0.00 %)	170	(4.74 %)	3	(0.08 %)
m	18	(7.44 %)	17	(7.02 %)	214	(5.97 %)	200	(5.58 %)
n	9	(3.72 %)	22	(9.09 %)	189	(5.27 %)	375	(10.46 %)
ŋ	0	(0.00 %)	9	(3.72 %)	0	(0.00 %)	126	(3.51 %)
h	15	(6.20 %)	0	(0.00 %)	234	(6.53 %)	4	(0.11 %)
SUM	242		242		3585		3585	

Table 6.3: List of places of articulation and groupings.

	Place of Articulation	Consonant
1	Labial (Lbl)	p, b, m
2	Labio-Dental (L-D)	f, v
3	Dental (Dtl)	θ, ð
4	Alveolar (Alv)	t, d, s, z, n
5	Palato-Alveolar (P-A)	ʃ, ʧ, ʤ
6	Palatal (Plt)	j
7	Velar (Vlr)	k, g, ŋ
8	Glottal (Glt)	h
9	Lateral	l
10	Rhotic	ɻ
11	Labialized velar	w

Table 6.4: Number of occurrences of C_1 (left column)–V (top row) combinations in our speech corpus. C_1 is grouped by the place of articulation.

		iː	ɪ	ɛ	æ	u	ʊ	ʌ	ɑ	sum
1	(Lbl)	7	6	5	12	3	2	5	10	50
2	(L-D)	1	0	2	3	2	0	3	2	13
3	(Dtl)	0	3	0	1	0	0	1	0	5
4	(Alv)	12	9	7	4	7	1	7	7	54
5	(P-A)	3	3	5	3	2	1	1	5	23
6	(Plt)	0	0	2	1	2	0	1	1	7
7	(Vlr)	2	4	1	5	3	2	1	5	23
8	(Glt)	2	2	2	3	1	0	1	4	15
9	(l)	6	2	2	3	2	0	1	5	21
10	(ɻ)	2	4	1	2	2	0	3	3	17
11	(w)	2	4	4	0	0	0	1	3	14
sum		37	37	31	37	24	6	25	45	242

Table 6.5: Number of occurrence of V(left column)–C_2 (top row) combinations in our speech corpus. C_2 is grouped by the place of articulation. The transitions from V to 6 (Plt), 8 (Glt) and 11 (w) are rare in English and not available in our corpus.

	1 (Lbl)	2 (L-D)	3 (Dtl)	4 (Alv)	5 (P-A)	6 (Plt)	7 (Vlr)	8 (Glt)	9 (l)	10 (ɻ)	11 (w)	sum
iː	5	4	1	12	2	0	3	0	7	3	0	37
ɪ	7	0	0	14	6	0	8	0	2	0	0	37
ɛ	1	2	0	16	3	0	4	0	5	0	0	31
æ	10	1	0	16	4	0	6	0	0	0	0	37
u	5	2	2	11	0	0	1	0	3	0	0	24
ʊ	0	0	0	2	1	0	3	0	0	0	0	6
ʌ	3	1	0	12	4	0	5	0	0	0	0	25
ɑ	13	0	0	11	2	0	11	0	4	4	0	45
sum	44	10	3	94	22	0	41	0	21	7	0	242

/ɑ/, since the vowel /ɔ/ is often pronounced as /ɑ/ in West-Coast American English.

The consonants include stops, fricatives, affricates, approximants, nasals, and aspiration. The number of phoneme occurrences (percentage in parentheses) are listed in the middle column in Table 6.1 for vowels and Table 6.2 for consonants. The phoneme distribution in our speech corpus is matched with the distribution found in the CMU dictionary in order to make the test material reflect, as closely as possible, typical patterns found in English. The right column in Tables 6.1 and 6.2 shows the number of phoneme occurrences (percentage in parentheses) of the vowels and of the consonants in the CMU dictionary. All words are listed in Table D.1, Appendix D.

Consonants can be grouped by the place of articulation, which influencies the formant contour shape. Places of articulation include bilabial (/p/, /b/, /m/), labio-dental (/f/, /v/), dental (/θ/, /ð/), alveolar (/t/, /d/, /s/, /z/, /n/), /palato-alveolar (/ʃ/, /ʧ/, /ʤ/), palatal (/j/), velar (/k/, /g/, /ŋ/), glottal (/h/), and individual consonants (/l/, /ɻ/, /w/). The place of articulation is summarized in Table 6.3. The number of occurences of C_1 to V transition, and of V to C_2 transition, is listed in Tables 6.4 and 6.5, respectively.

The CVC words are pronounced in carrier sentences, which facilitates speakers to manipulate prosodic information upon the elicitation of CNV and CLR speech. The following five sentences are used:

- I know the meaning of the word *WORD*.
- Make a sentence using the word *WORD*.
- Use a dictionary to look up the word *WORD*.
- Her last name sounds like the word *WORD*.
- I'm tired of hearing the word *WORD*.

These sentences provide neutral meaning and have a consistent phoneme /d/ followed by the test word in a sentence-final context. The stress is not placed on the test word.

6.2.2 Recordings

Two speakers (one male: JPH and one female: ETH) are recorded in order to examine the effect of different speakers, with two repetitions per word. Speaking styles include CNV and CLR styles with the following instructions, based on previous literature [18]:

1. We introduce speakers to the concept of CNV and CLR speech and discuss differences in speaking rate, articulation, F0 fluctuation, and pausing [18]. Also, we encourage speakers to "enunciate consonants more carefully and with greater effort than in CNV speech and avoid slurring words together" [40].

2. We have speakers listen to audio samples of CNV and CLR speech.

3. As practice sessions, we record 5 sentences of CNV and CLR speech.

4. We listen to the recorded speech, and we listen to audio samples of corresponding CNV and CLR speech.

5. We discuss potential improvements in terms of acoustic differences of CNV and CLR speech.

6. We record CNV speech in two sessions, and CLR speech in two sessions. Each session takes place over four consecutive days. At the beginning of each session, we review the above instructions.

6.2.3 Feature extraction

The initial estimates of the following features were extracted using existing software: phoneme boundaries (forced-alignment) [44], glottal-closure instants (CSLU toolkit) and formant contours (Snack Sound Toolkit (http://www.speech.kth.se/snack) [91]).

Then, trained labelers manually corrected all features by visually inspecting the waveform, spectrogram, and phoneme identity. To ensure the best accuracy of formant tracking, the following screenings were completed: (1) any samples that had neighboring formants closer than 200 Hz were detected and manually corrected, and (2) any samples that were 3 standard deviations away from the vowel mean value for each speaking style were detected and manually verified or corrected. The results of formant synthesized speech described below (Section 6.2.4), were utilized to identify incorrect phoneme boundaries, GCIs, and formant frequencies.

6.2.4 Perceptual validation

The accuracy of formant frequencies were verified using a formant analysis-by-synthesis procedure. The assumption was that if formant frequency values in the voiced segment were incorrect, the listeners might incorrectly identify the formant-synthesized word.

Vowel regions as well as voiced consonants (only liquid, approximant and glide) were synthesized with existing formant contours and an artificial glottal source (OQ = 0.6, SK = 0.75 with the Rosenberg glottal source model [86]). For the initial consonant /h/, the formant values were excited with white noise. Other features, such as vowel duration, F0 values, and energy remained the same as in the original sample. For the words starting with consonants /ʃ/, /θ/, and /h/, a ramping window was used at the beginning of the consonant. The formant bandwidths from the original speech were smoothed, so that neighboring frames had differences less than 20 Hz. The other consonants were copied from the original waveform. As the result of perceptual testing using four listeners, any words that were identified incorrectly were detected, then those formant frequency values and/or phoneme boundaries were manually corrected.

6.3 Acoustic analyses of CVC words

In this section, we analyze the acoustic features of the CVC words in our speech corpus. The analyses include visual inspection of formant contour shape and quantitative analyses of formant steady state values, formant transitions, and phoneme durations. We examine differences in CNV and CLR speaking styles and speaker characteristics.

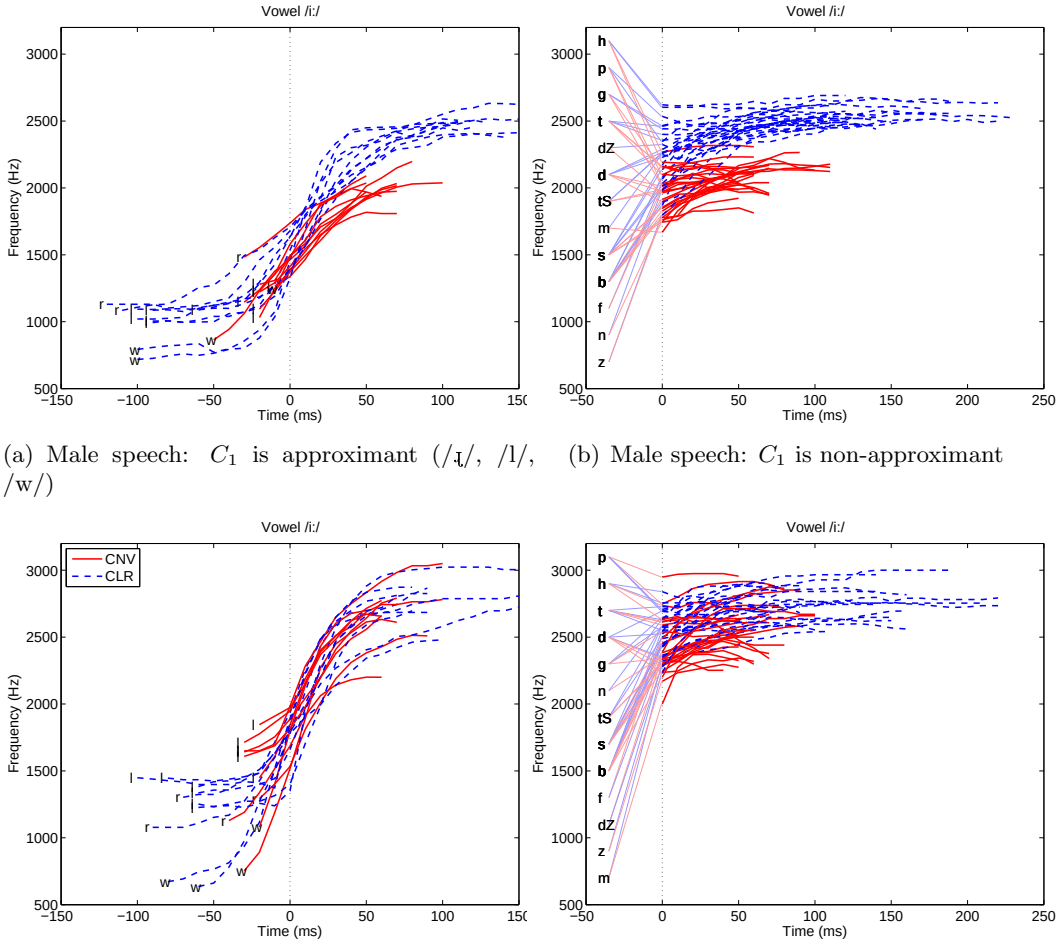
6.3.1 Formant contour shape

Figures 6.1(a)–6.1(d) show the F2 contour shape from middle of C_1 to the middle of V, without normalizing the time scale of both CNV and CLR speech for two speakers. The vowel is restricted to /iː/, while C_1 is grouped as approximants (/ɻ/, /l/, /w/) in Figures 6.1(a) and 6.1(c) and non-approximants in 6.1(b) and 6.1(d). Each contour is the average of two tokens centered at the vowel onset ($t = 0$ ms).

For both male and female speech, the CNV duration is shorter than CLR duration in every case. The contour values for the male speech are more clustered than the female speech. For the male speech, the formant contour reaches higher formant frequencies in CLR speech than CNV speech, and the vowel onset values for non-approximant tokens are higher in CLR speech than CNV speech.

6.3.2 Formant steady state values

Formant steady states (SS) were measured at the middle of the vowel. Figure 6.2 shows the mean of each vowel (with ±1 standard deviation) in F1–F2 space for CNV (solid red



(a) Male speech: C_1 is approximant (/ɻ/, /l/, /w/)

(b) Male speech: C_1 is non-approximant

(c) Female speech: C_1 is approximant (/ɻ/, /l/, /w/)

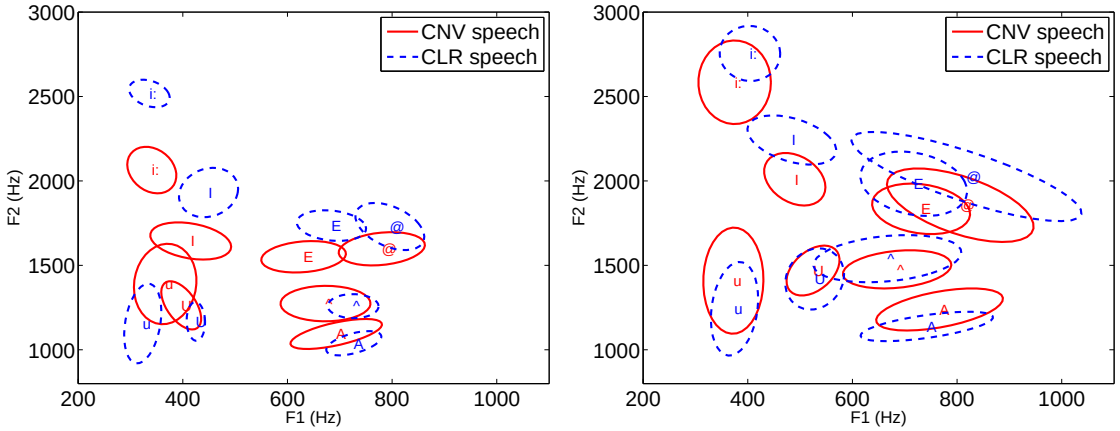
(d) Female speech: C_1 is non-approximant

Figure 6.1: F2 contour shape from middle of C_1 to middle of V (/iː/) for CNV (red solid lines) and CLR (blue dashed lines). All contours are centered at the $C_1 - V$ boundary (0 ms).

lines) and CLR (dashed blue lines) speech. The figure shows that the CLR speech vowel space is expanded, with lower /u/ F2 and higher /iː/ F2, and that CNV vowels have greater variability. Comparing male and female speech, female speech has a larger vowel space and more overlap between CNV and CLR speech.

6.3.3 Formant transition

Previous studies have suggested that the formant transition is an important perceptual cue for intelligibility [33]. As shown in the study by Krause and Braida [55], the analysis



(a) Male speech (b) Female speech

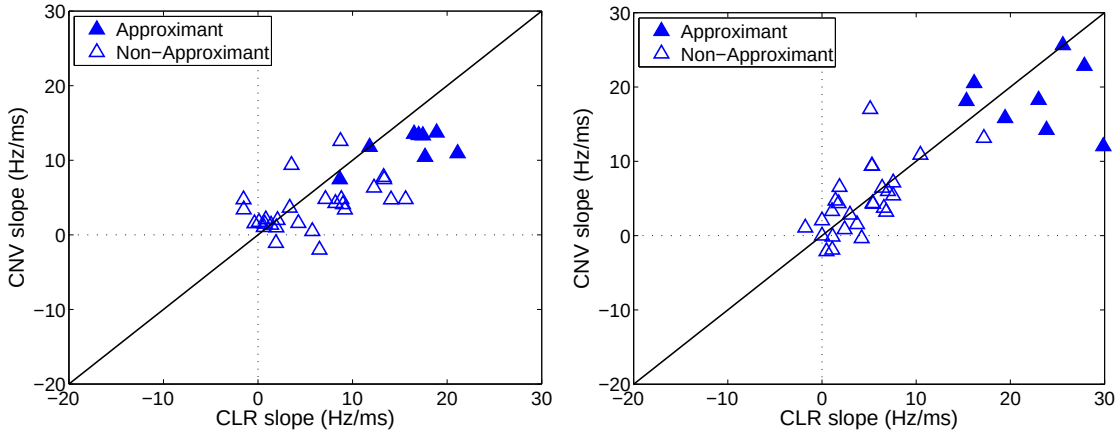
Figure 6.2: Formant steady-state values (with ±1 standard deviation) in F1/F2 space for CNV and CLR speech. The phonemes shown with dashed blue lines are for CLR speech, and solid red lines are for CNV speech.

of the formant transition is often difficult since a different place of articulation in the consonant leads to a different shape of contour, and there are not so many of the same CV transition pairs available in the dataset. In this study, regardless of the contour shape, the formant transition is analyzed at ±20 ms (+20 ms or −20 ms when C_1 or C_2 is unvoiced). The formant transition is measured using the following equation, based on delta values that are used as a spectral change measure in automatic speech recognition (ASR) [33],

$$\delta[t] = \frac{\sum_{w=-W}^{W} w \cdot y[t+w]}{\sum_{w=-W}^{W} w^2} \tag{6.1}$$

where w is a window size ($W = 2$), in 10 ms frames. The time point t is given either at vowel onset (T_1) or the vowel offset (T_2).

In order to compare style differences in CNV and CLR speech, the F2 slope values are plotted in Figures 6.3(a) (male speech) and 6.3(b) (female speech) for the vowel /iː/. In this figure, the slope values ($\delta[t]$) of CLR speech are averaged over 2 tokens (x-axis), and slope values of CNV speech are also averaged over two tokens (y-axis). The black diagonal lines represent no difference between CNV and CLR speech. For the positive slope values, any points below the black line (and for the negative slope values, any points above the black line) indicate that CLR slope is steeper than CNV slope. These cases are found at the vowel onset when prevocalic consonants are approximants (/ɻ/, /l/, /j/, /w/) in both



(a) Male speech (b) Female speech

Figure 6.3: F2 onset slope difference of the vowel /iː/ between CLR and CNV speaking style.

Table 6.6: Mean vowel duration (ms) of 8 vowels (standard deviation) in two speaking styles.

	Male speech			Female speech		
Vowel	CNV speech	CLR speech	Ratio	CNV speech	CLR speech	Ratio
/iː/	146.39 (39.81)	308.62 (101.82)	2.10	145.86 (39.21)	275.22 (124.41)	1.89
/ɪ/	101.17 (18.55)	173.58 (44.14)	1.72	130.66 (28.67)	180.92 (62.93)	1.38
/ɛ/	123.07 (22.05)	216.61 (44.82)	1.76	154.75 (27.18)	234.39 (85.78)	1.51
/æ/	173.54 (27.70)	372.23 (99.23)	2.13	209.67 (41.61)	352.64 (114.18)	1.68
/u/	151.99 (43.53)	311.51 (98.18)	2.05	153.54 (36.33)	292.88 (132.62)	1.91
/ʊ/	106.38 (15.34)	196.82 (63.99)	1.85	118.83 (25.17)	190.76 (72.59)	1.61
/ʌ/	117.83 (23.07)	214.22 (45.98)	1.82	151.96 (27.68)	250.74 (84.87)	1.65
/ɑ/	156.26 (27.40)	322.76 (93.57)	2.07	178.61 (40.63)	294.85 (110.29)	1.65

male and female speech. However, these patterns were not found in vowel offset slopes.

6.3.4 Phoneme duration

Speaking rates, excluding pause duration longer than 10 ms, were 309 wpm (CNV) and 126 wpm (CLR) for male speech, and 291 wpm (CNV) and 120 wpm (CLR) for female speech. Vowel durations of the word of interest were measured to reflect speaking rate. Table 6.6 shows the mean (standard deviation) for the two speaking styles, and the ratios between mean CLR and CNV durations for each speaker.

Table 6.7: Average F0 values of CNV and CLR speech at the phoneme boundary. Only voiced consonants (approximants) are averaged over 484 samples per speaking style. Peak F0 values are not necessarily within the vowel.

	Male speech		Female speech	
Value point	CNV speech	CLR speech	CNV speech	CLR speech
Word onset	88.55	110.50	158.41	179.40
Vowel onset	91.11	143.72	161.44	200.39
Vowel offset	80.89	90.50	146.50	160.11
Word offset	76.96	77.58	151.53	161.11
Maximum value	93.36	149.14	166.14	208.04
Mean value	86.43	121.00	155.04	181.74

The results of a two-sample t-test show that CLR speech durations were significantly longer than those of CNV speech for all vowels ($p < 0.0001$, $p < 0.01$ for male and female speech, respectively). Male speech had more dramatic changes (average 1.93 ratio) in CLR speech production than the female speech (average 1.62 ratio). For both speakers, the duration of tense vowels was more susceptible to the speaking style than lax vowels, being consistent with the previous result from Picheny *et al.* [80]. Both prevocalic and postvocalic consonants were also longer in CLR speech than in CNV speech. Approximants (/ɹ/, /l/ /j/, /w/) of CLR speech are 2.35–3.29 times as long as those of CNV speech for male speech, and 1.97–2.73 times longer for female speech. The smallest changes between CNV and CLR speech observed in consonants (voiced stops) are ratios of 0.95–1.37 for male and 0.98–1.53 for female speech. The consonant position also makes a difference in CLR speech production. For example, consonant /l/ ratios were 3.29 (C_1) and 2.09 (C_2) for male, and 2.73 (C_1) and 2.45 (C_2) for female. In general, prevocalic consonants C_1 are more prolonged than postvocalic consonants (C_2) for both speakers.

6.3.5 Fundamental frequency (F0) contours

The fundamental frequency (F0) values were calculated by inversing the distance between GCIs during the voiced segment (described in Section 6.2.3). Table 6.7 shows the average F0 values of CNV and CLR speech at five points (word onset/offset, vowel onset/offset, max value) and mean values. CLR speech production resulted in higher F0 mean values and variation as compared with CNV speech. The F0 values were averaged over 484 samples per speaking style with only available points (no F0 values within unvoiced segments). For both speakers, the difference between CNV and CLR speech is larger at word onset (and at the F0 peak) than at word offset, indicating that CLR speech has a more dynamic F0

contour.

In summary, the features described above (formant contours, formant steady-state values, phoneme durations, F0 contours) indicate differences in speaking styles, which are consistent with previous studies [80, 55]. The results show that our speech corpus demonstrates CLR and CNV speech characteristics, and that the male speaker had larger acoustic differences in CLR speech than the female speaker.

6.4 Formant contour model

As described in Section 5.2, the equation of the formant contour model is given as

$$\begin{aligned}\widehat{\mathbf{F}}(t) &= \acute{d}_1(t; s_1, p_1)\mathbf{T}_{C_1} + (1 - \acute{d}_1(t; s_1, p_1) - \acute{d}_2(t; -s_2, p_2))\mathbf{T}_V + \acute{d}_2(t; -s_2, p_2) \cdot \mathbf{T}_{C_2} \\ &= \acute{d}_1(t; s_1, p_1) \cdot (\mathbf{T}_{C_1} - \mathbf{T}_V) + \mathbf{T}_V + \acute{d}_2(t; -s_2, p_2) \cdot (\mathbf{T}_{C_2} - \mathbf{T}_V) \qquad (6.2)\end{aligned}$$

where $\widehat{\mathbf{F}}(t)$ is the estimated formant contour of a CVC word as a function of time t, as in a study by Niu and van Santen [74]. $\mathbf{T}_{C_1}$, $\mathbf{T}_V$, and $\mathbf{T}_{C_2}$ are the target formant vectors of the prevocalic consonant (C_1), vowel (V) and postvocalic consonant (C_2), respectively. The target formant vectors consist of the first four formant values (4×1 dimension). The function $\acute{d}(t; s, p)$ represents the degree of coarticulation of C_1 and C_2, and is proportional to the differences in target formant values.

6.4.1 Coarticulation function

The exponential curve for the transition function in Broad's study [17] has been changed to a sigmoid function, which is called a coarticulation function in this study:

$$\acute{d}(t; s, p) = \frac{1}{1 + e^{s(t-p)}} \qquad (6.3)$$

It is characterized by two coefficients, s (slope) and p (slope position) of the sigmoid function. Examples of coarticulation functions are shown in Figure 5.1 in Chapter 5. While the formant contour is defined in Broad's study [17] from the onset to the offset of the vowel in C_1VC_2 contexts, we model the formant contour from the middle of C_1 to the middle of C_2 in cases of approximants, as shown in the study by Amano-Kusumoto and Hosom [3].

6.4.2 Constraints on parameters

Parameters are very gently constrained for a better model:

- $d_1(t; s_1, p_1) + d_2(t; -s_2, p_2) \leq 1.0$.
- $s \geq 0$ in Equation 6.3.
- p_1 (or p_2) in Equation 6.3 may range from middle of C_1 (or V) to the middle of V (or C_2).
- $200 < F1 < 1000$, $400 < F2 < 3000$, $900 < F3 < 3700$, $2500 < F4 < 5000$.
- $F2 - F1 > 200$, $F3 - F4 > 200$, $F4 - F3 > 200$.

When the value s_1 or s_2 approaches zero, the slope of coarticulation function d becomes shallow. When this happens, the constraint $d_1(t; s_1, p_1) + d_2(t; -s_2, p_2) \leq 1.0$ must be enforced, since a shallow slope results in an unrealistically high contribution of the consonant throughout the utterance.

6.5 Experiment 6–1: Speaking style dependencies of target formants

We fitted the formant contour model to our speech corpus of two different speaking styles for both male and female speech. The parameters were learned on training sets and tested on test sets with a jackknife procedure. First, the total of 968 tokens in a data set were randomly partitioned into 20 test-set groups. For the style-independent target estimation, each test-set group contains 48 (or 49) tokens of two speaking styles. On the other hand, for the style-dependent target estimation, each test-set group contains 48 (or 49) tokens of only one speaking style. The training dataset consists of tokens that are not present in the test-set group. The number of training data should be same for style-independent and style-dependent target estimation. Therefore, for the style-independent target estimation, each training group contains 436 (or 435) tokens from both speaking styles, while for style-dependent target estimation, each training group contains 436 (or 435) tokens of only one speaking style. We ensured that each training group contains at least two tokens of each phoneme.

6.5.1 Estimating model parameters: Style-independent targets

We estimate model parameters (s_1, p_1, s_2, p_2) per token, and formant target values $\mathbf{T}$ for each phoneme for both speaking styles. (This is called context-independent and style-independent target estimation.)

The error function to estimate parameters (s_1, p_1, s_2, p_2) is

$$Err1^{(k)} = \sqrt{\frac{\mathbf{w} \sum_{t=T_1^{(k)}}^{T_2^{(k)}} \left(\widehat{\mathbf{F}}(t)^{(k)} - \mathbf{F}(t)^{(k)} \right)^2}{N^{(k)}}} \tag{6.4}$$

where $\mathbf{F}(t)^{(k)}$ and $\widehat{\mathbf{F}}(t)^{(k)}$ are the observed and estimated formant contours for the k-th word, converted to Bark scale. $N^{(k)}$ is the number of frames from $T_1^{(k)}$ to $T_2^{(k)}$. The contribution to the error from F1 and F2 are weighted more than F3 and F4, represented as $\mathbf{w} = \begin{bmatrix} 1.0 & 1.0 & 0.25 & 0.01 \end{bmatrix}$. The error is not divided by the sum of the weights, because it does not affect the ranking. $T_1^{(k)}$ (or $T_2^{(k)}$) is located at the middle of C_1 (or C_2) when C_1 (or C_2) is an approximant. Otherwise, $T_1^{(k)}$ (or $T_2^{(k)}$) is located at the C_1V (or VC_2) boundary.

Formant targets are estimated by minimizing,

$$Err2 = \sum_{style=1}^{S} \sum_{k:phn \in k}^{K} Err1_{style}^{(k)} \tag{6.5}$$

$Err1^{(k)}$ is summed over all the words (k) that have a particular phoneme (phn) and over all speaking styles. K is the number of words ($K = 218$) used in the training set per style. A total of 436 (or 435) words for both styles were used. The $Err2$ value is comparable with both $Err3$ values from each speaking style (below in Section 6.5.2). S is the number of speaking styles ($S = 2$). Model parameters are estimated using a hill-climbing approach, the same as in Chapter 5.2.1.

6.5.2 Estimating model parameters: Style-dependent targets

We estimate model parameters (s_1, p_1, s_2, p_2) in $\mathcal{d}(t; s, p)$ per token, and formant target values $\mathbf{T}$ for each phoneme in each speaking style. (This is called context-independent and style-dependent target estimation, where one set of formant values is estimated per phoneme for each speaking style.)

The same error function (Equation 6.4) is used to estimate parameters (s_1, p_1, s_2, p_2), while Equation 6.5 is changed to

$$Err3_{style} = \sum_{k:phn \in k}^{K} Err1_{style}^{(k)} \tag{6.6}$$

$Err1^{(k)}$ is summed over all the words (k) that have a particular phoneme (phn) for each speaking style. K is the number of words ($K = 436$) used in the training set. The initial

Table 6.8: Mean error $E_{s,target}$ (in Bark) and standard deviation in parentheses in training and test sets. Style-independent II is the result of an increased amount of training data (described in Section 6.6).

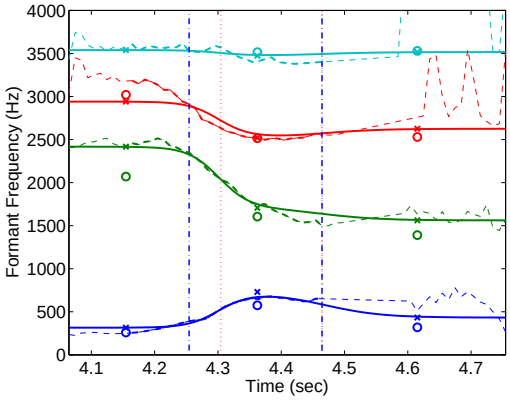
		Training set		Test set		
		CNV	CLR	CNV	CLR	Average
Male speech	Style-Indep.	0.2740 (0.1246)	0.2975 (0.1091)	0.3076 (0.1572)	0.3104 (0.1164)	0.3090 (0.1383)
	Style-Dep.	0.2592 (0.1242)	0.2747 (0.1041)	0.2780 (0.1362)	0.2851 (0.1107)	0.2815 (0.1241)
	Style-Indep. II.	0.2881 (0.1323)	0.3005 (0.1121)	0.3020 (0.1398)	0.3079 (0.1182)	0.3049 (0.1294)
Female speech	Style-Indep.	0.3404 (0.1517)	0.3727 (0.1770)	0.3679 (0.1607)	0.3990 (0.1925)	0.3834 (0.1779)
	Style-Dep.	0.3345 (0.1573)	0.3547 (0.1685)	0.3568 (0.1650)	0.3857 (0.1893)	0.3712 (0.1781)
	Style-Indep. II.	0.3532 (0.1529)	0.3790 (0.1798)	0.3666 (0.1614)	0.3941 (0.1900)	0.3804 (0.1767)
Male speech	Model validation	0.1063 (0.0516)	0.1010 (0.0412)	0.1014 (0.0521)	0.0929 (0.0311)	0.0972 (0.0416)

values and the hill-climbing approach for the optimization method are the same as in style-independent target estimation.

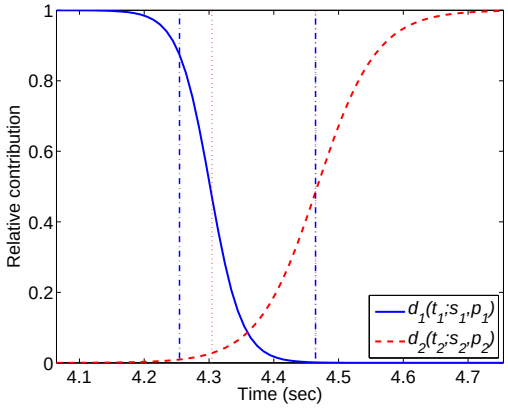
6.5.3 Results: Goodness of fit

The root mean square error $E^{(k)}_{s,target}$ (in Bark) is calculated for each word k as in Equation 6.4. The subscripts s, $target$ indicate that the error depends on the variables s and $target$, while the value p for slope position is adjusted to the best fit in each token. Mean $E^{(k)}_{s,target}$ values over the samples in the training and test sets are reported for each style in Table 6.8. The error difference between training set and test set is relatively small, indicating little over-fitting to the training set. The error rate was successfully reduced with style-dependent target estimation both in training and test sets.

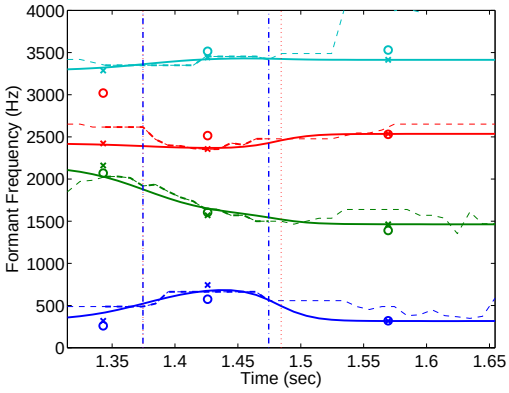
Figures 6.4(a) and 6.4(c) represent the modeled formant contour of the word "*yes*" (/j ɛ s/) (solid lines) as well as observed contour (dotted lines) in both CNV and CLR styles of the male speech. The initial values (circles in Figures 6.4(a) and 6.4(c)) are taken from Allen *et al.* [1]. The estimated target values (crosses) are the result of style-dependent targets for each speaking style. Figures 6.4(b) and 6.4(d) show the corresponding coarticulation functions ($d_1(t_1; s_1, p_1)$ and $d_2(t_2; s_2, p_2)$). The slope of the coarticulation functions show



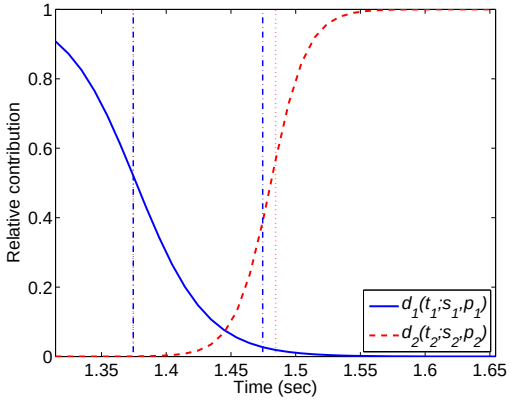
(a) Modeled formant contour (solid lines) and observed contour of CLR style (dotted lines). The $E^{(k)}_{s,target}$ was measured as 0.1803 Bark.



(b) Coarticulation functions $d(t;s,p)$ for CLR style. $s_1 = 0.4102, s_2 = 0.2188$.



(c) Modeled formant contour (solid lines) and observed contour of CNV style (dotted lines). The $E^{(k)}_{s,target}$ was measured as 0.1964 Bark.

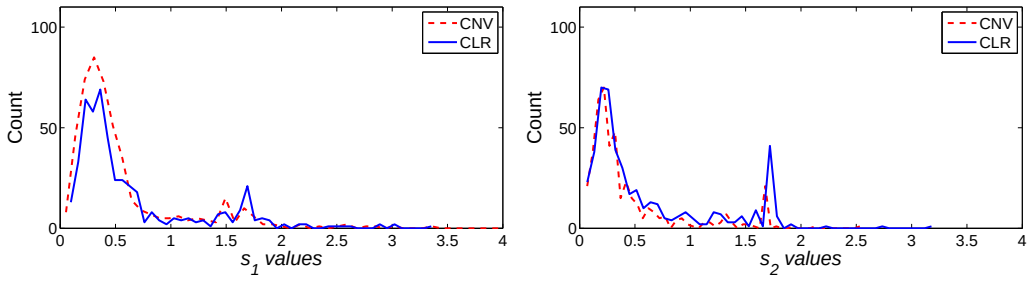


(d) Coarticulation functions $d(t;s,p)$ for CNV style. $s_1 = 0.3672, s_2 = 0.7203$.

Figure 6.4: The results of formant contour model for the word "*yes* (/j ɛ s/)" of male speech. Circles (same in (a) and (c)) are the initial values, while crosses are estimated values with style-dependent estimation (different in (a) and (c)). The estimated target values are the average of 10 groups per speaking style from the training set. The coarticulation parameters (s and p) are adjusted to minimize the error with given target values per token. Blue vertical dash-dot lines show the phoneme boundaries, while vertical red dashed lines represent p_1 (left) and p_2 (right).

how fast the contour moves from C_1 to V or V to C_2 targets. The $E^{(k)}_{s,target}$ values were 0.1803 Bark (Figure 6.4(a)) and 0.1964 (Figure 6.4(c)) in CLR and CNV styles, respectively.

The test-set results were submitted to four way analysis of variance (ANOVA) (2 speakers × 2 methods × 2 speaking styles × 8 vowel identities). All of the main effects were significant ($F(1,3870) = 206.74$, $p < 0.0001$, $F(1,3870) = 10.5$, $p = 0.0012$, $F(1,3870) =$



(a) Male s_1 values

(b) Male s_2 values

(c) Female s_1 values

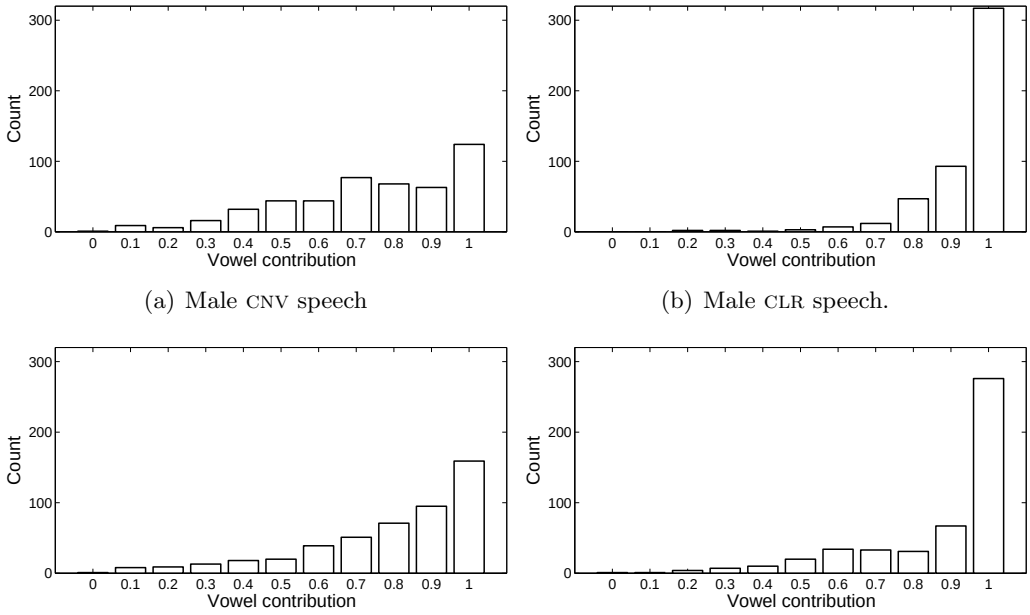
(d) Female s_2 values

Figure 6.5: Histograms of s values for each speaker.

9.65, $p = 0.0019$, $F(7, 3864) = 5.85$, $p < 0.0001$, respectively). The post-hoc analysis (HSD) showed that the model fitted better with the male speech than female speech. The style-dependent estimation had slightly lower error rate, but significant. CNV speech was better fitted to the model than CLR speech. The error rate on words with the vowel /u/ was significantly different from vowels (/iː/, /ɪ/ and /ɛ/) and the vowel /iː/ was different from /æ/ at $\alpha = 0.0018$ (0.05/28) level, but no other combinations were significantly different.

The reason for the error difference between male and female speech might be solely due to the fact that female speech has a higher standard deviation in formant steady state values, as shown in Figure 6.2(b). Particularly, the female speaker pronounced the vowel /æ/ in CLR style as /j–æ/ (as a diphthong), which makes the error increase. Modeling diphthongs may require two sets of vowel targets.

The main effect of speaking style and significant interaction between speaker and speaking style ($F(1, 483) = 6.38$, $p = 0.0116$) indicates that CLR speech had a higher error than CNV speech for the female speech. In general, words with the vowel /u/ had a higher error rate, perhaps because the contour /u/ has high variability due to neighboring consonants. Although this variability can be captured by the model in theory, the wide



(a) Male CNV speech

(b) Male CLR speech.

(c) Female CNV speech

(d) Female CLR speech

Figure 6.6: Histograms of vowel contribution for each speaking style and each speaker.

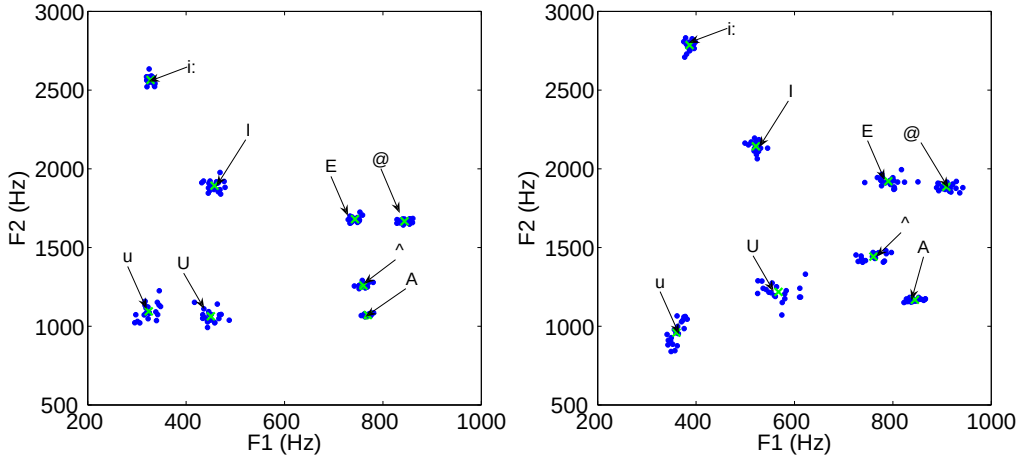
difference in observed values may make parameter estimation less robust.

At last, the parameters s and p in the coarticulation functions were adjusted per token, given the mean estimated target values shown in crosses in Figure 6.7(a) and 6.8(a), which resulted in mean error rates of 0.1017 (SD: 0.1141) (CNV) and 0.1024 (0.0791) (CLR) for the male speech, and 0.1437 (SD: 0.1433) (CNV) and 0.1767 (0.2044) (CLR) for the female speech, respectively. The resulting parameters s and p were used to synthesize the formant contour (male only in Section 6.5.4) and to characterize the parameter s (Section 6.5.5).

6.5.4 Formant model validation

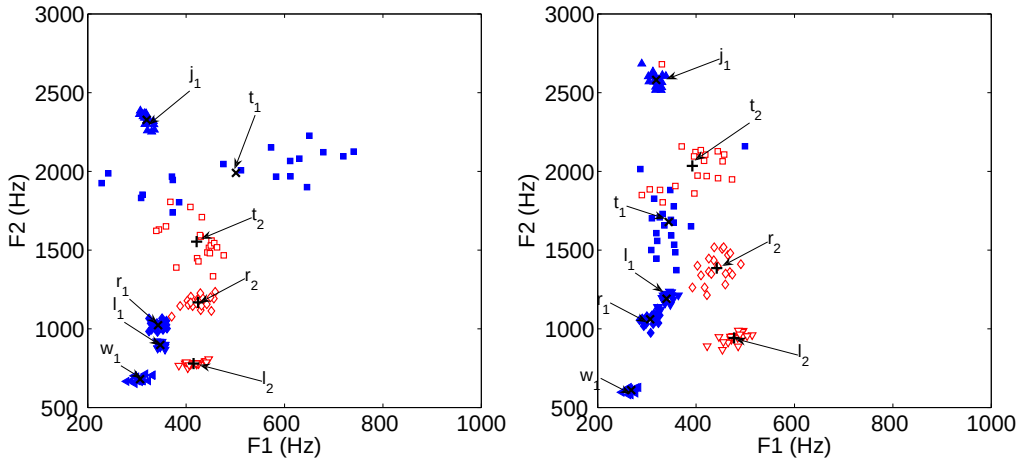
In order to test the robustness of parameter estimation, we re-estimated parameters using synthetic formant contours as the observed values. The formant contour was synthesized with previously estimated formant target values (style-independent targets, the mean of 20 values) and coarticulation functions $d(t; s, p)$ from male speech plus noise. The noise was created by uniformly distributed random values with zero mean. The standard deviation of the noise was 0.4467 Bark, which is the mean error rate corresponding to F1 only from the male speaker, style-independent target estimation in Table 6.8.

As a result, re-estimated target values were close to the originally estimated targets.



(a) Male speech (b) Female speech

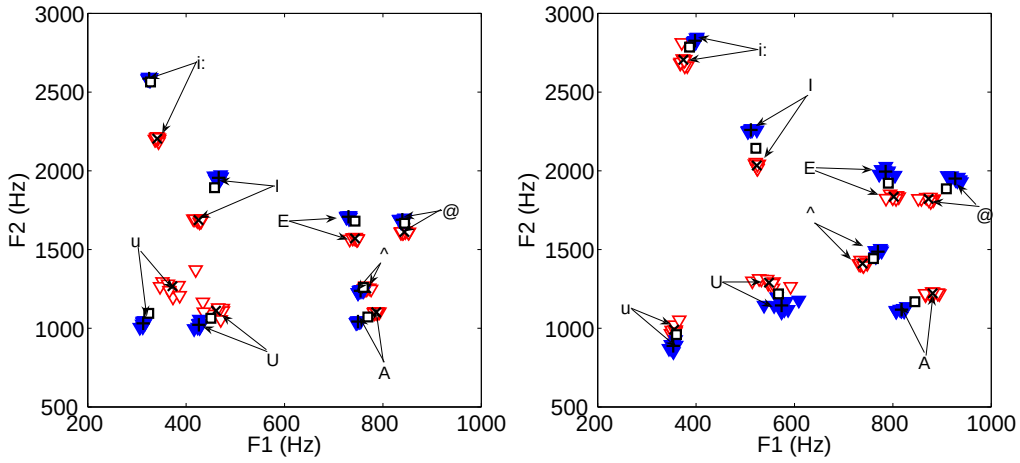
Figure 6.7: Estimated style-independent vowel target values in F1−F2 space. The means of each phoneme from style-independent target estimation are shown in black crosses.



(a) Male speech (b) Female speech

Figure 6.8: Estimated style-independent consonant target values in F1−F2 space. Selected consonants are shown for C_1 (filled blue: /t/, /ɹ/, /l/, /j/ and /w/) and C_2 (open red: /t/, /ɹ/ and /l/).

The difference between initially estimated formant targets (shown in Figure 6.7(a)) and re-estimated formant targets with synthetic formant contours was, on average, 3.64, 7.75, 9.26, and 13.27 Hz for F1 through F4, respectively. The error was calculated between re-estimated formant contours and synthetic formant contours (shown in Table 6.8). Very small error rates in the training and test sets confirms that the parameter estimation



(a) Male speech

(b) Female speech

Figure 6.9: Estimated style-dependent vowel formant target values in F1−F2 space for CNV style (open red) and CLR style (filled blue). The means of each phoneme from style-dependent target estimation are shown in black crosses, while the means from style-independent target estimation are shown in black squares.

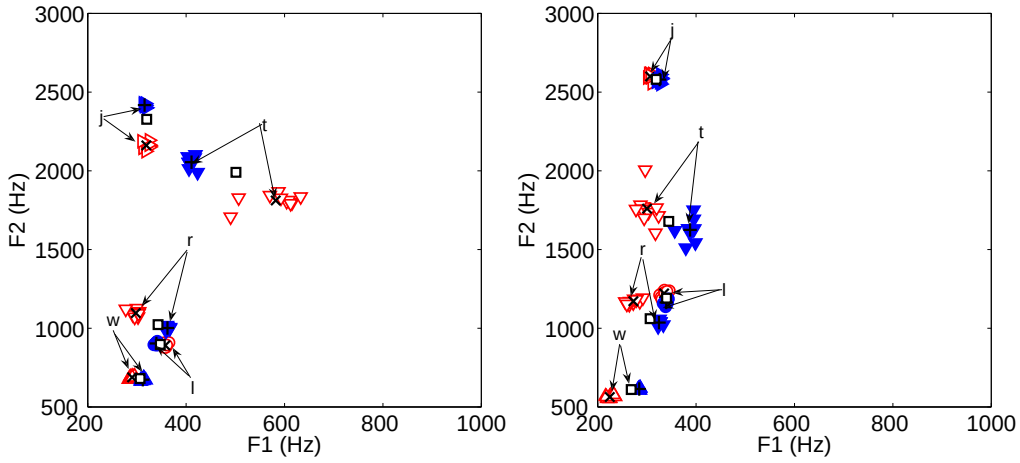
process is robust.

6.5.5 Estimated $d(t; s, p)$ parameters

Parameters s_1 and s_2 in the coarticulation functions tell us the transition slope regardless of the target value. We analyze s_1 and s_2 parameters comparing CNV and CLR speech, which were estimated with style-independent targets. The style-independent targets were used to characterize coarticulation parameters so that the effect of speaking style was observed in the coarticulation parameters.

Our previous study with style-independent targets (Chapter 5) showed that the slopes at the vowel onset are steeper for CLR speech than for CNV speech for the front vowels, while slopes at the vowel offset are steeper for the vowel /iː/ and /ei/for male speech. This effect may depend on the phoneme context and the speaker. Therefore, we first re-evaluated these earlier results by analyzing s in only the words with C_1 =/w/ and V = front vowels for s_1 (vowel onset transition), and V = /iː/ and C_2 = /l/ for s_2 (vowel offset transition). These pairs are the cases where significant differences between CNV and CLR speech were reported in Section 5.4.1.

Prior to the analysis, outliers in terms of the root mean square error $E^{(k)}_{s,target}$ in Equation 6.7 were excluded using a modified $z-$score test. The criteria for detecting an outlier



(a) Male speech (b) Female speech

Figure 6.10: Estimated style-dependent C_1 target values: Selected consonants (/t/, /ɹ/, /l/, /j/ and /w/) are shown in F1−F2 space for CNV style (open red) and CLR style (filled blue). The means of each phoneme from style-dependent target estimation are shown in black crosses, while the means from style-independent target estimation are shown in black squares.

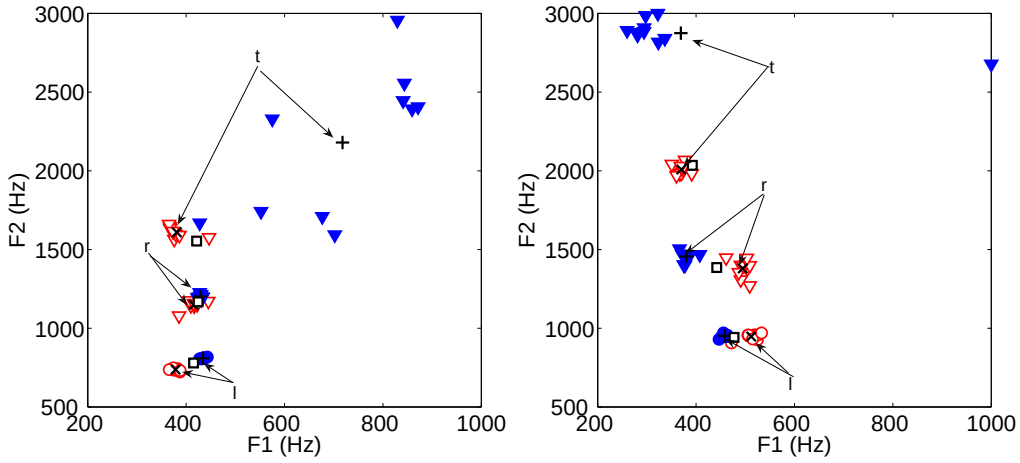
was $Z_i > 3.5$.

$$Z_i = \left| \frac{0.6745(E^{(k)}_{s,target} - \overline{E}_{s,target})}{median(E^{(k)}_{s,target} - \overline{E}_{s,target})} \right| \tag{6.7}$$

63 tokens for male speech ($E^{(k)}_{s,target} = 0.2755 - 0.4364$ Bark) and 54 tokens for female speech ($E^{(k)}_{s,target} = 0.4772 - 2.0356$ Bark) out of 968 tokens per speaker were excluded.

For male speech, the mean s_1 values for C_1 = /w/ to V = front vowels transitions were 0.4313 (SD: 0.0848) and 0.5447 (0.1284) for CNV and CLR speech, respectively. A two-sample $t-$test showed that the mean of CLR s_1 was higher than the mean of CNV s_1 ($df = 37, p = 0.0012$), which confirms the earlier result. On the other hand, mean s_2 values from V = /iː/ to C_2 = /l/ in our corpus were 0.2585 (SD: 0.0324) and 0.2348 (0.0291) for CNV and CLR speech, respectively. The mean s_2 values of CNV were greater than those of CLR speech ($df = 23$, $p = 0.0336$), which is opposite from the prior results. The prior study's significant difference in s_2 may have been due to a bias caused by a smaller set of phonetic contexts.

For female speech, the mean s_1 values for C_1 = /w/ to V = front vowel transitions were 0.5542 (SD: 0.1351) and 0.5875 (0.1706) for CNV and CLR speech, respectively. The difference was not significant ($df = 38$, $p = 0.2488$). The mean s_2 values from V =



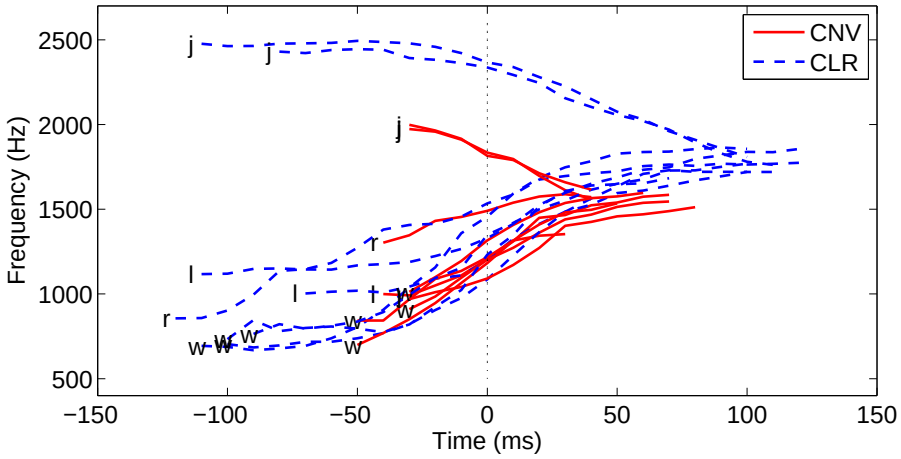
(a) Male speech (b) Female speech

Figure 6.11: Estimated style-dependent C_2 target values: Selected consonants (/t/, /ɹ/ and /l/) are shown in F1−F2 space for CNV style (open red) and CLR style (filled blue). The means of each phoneme from style-dependent target estimation are shown in black crosses, while the means from style-independent target estimation are shown in black squares.

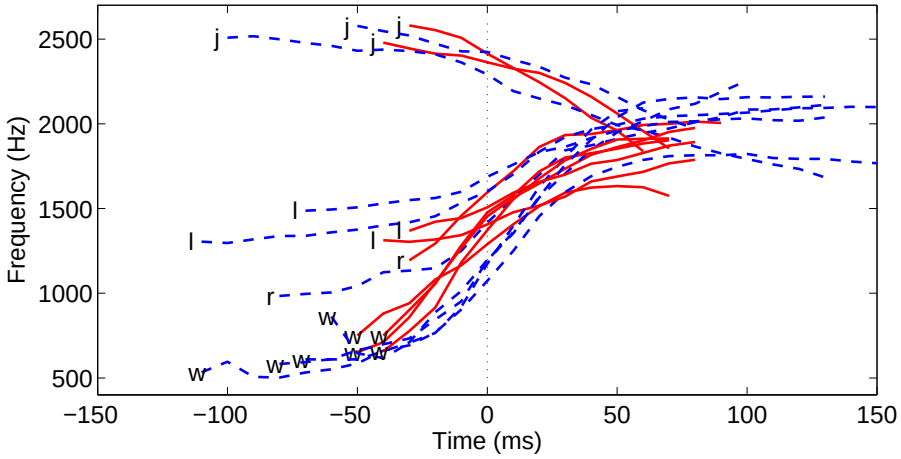
/iː/ to C_2 = /l/ were 0.2990 (SD: 0.0326) and 0.2028 (0.0502) for CNV and CLR speech, respectively. Similar to the results of male speech, CNV parameter (s_2) had steeper values than those of CLR speech ($df = 26$, $p = 1.1915 \times 10^{-6}$) at the vowel offset.

Then, we analyzed s parameters on all words excluding the above-mentioned outliers. The average s_1 values for male speech were 0.5506 (SD: 0.5255) and 0.6381 (0.5747) for CNV and CLR speech, respectively, and the average s_2 values were 0.4437 (SD: 0.4418) and 0.6013 (0.5590), respectively. For female speech, the average s_1 values were 0.6008 (SD: 0.5010) and 0.6607 (0.5996) for CNV and CLR speech, and the average s_2 values were 0.4886 (SD: 0.4319) and 0.6121 (0.5905), respectively.

For both speakers, the histograms of s_1 and s_2 show bimodal distributions where a second peak exists at approximately 1.6–1.7 (male) and 3.3–3.4 (female), for both s_1 and s_2 values (Figures 6.5). The second peak is observed when the maximum slope (p value) is located in the non-approximant consonant, in which case the slope does not have much impact on the error. For those tokens, the shape of the formant contour is flat and there is not much movement in the voiced region, which lead to higher s_1 and s_2 values. Further research is required to understand the effect of speaking style on s values (i. e. investigating further constraints on p in $d(t; s, p)$ values).



(a) Male speech



(b) Female speech

Figure 6.12: Observed F2 contours for C_1= approximant, V= /ɛ/ in both CNV (red solid lines) and CLR (blue dashed lines) speech. All contours are centered at the $C_1{-}V$ boundary (0 ms).

6.5.6 Contribution of the vowel target

In addition to the parameters s and p, the combination of d_1 and d_2 provides us with important information about the contribution of the vowel. The vowel contribution is shown as $C(t) = 1 - (d_1(t; s_1, p_1) + d_2(t; -s_2, p_2))$ in Equation 6.2, expressing the contribution of the vowel target and formant undershoot. The values of C range between 0.0 and 1.0. The value 0.0 means no influence of the vowel, and 1.0 shows the fully-articulated vowel. Figure 6.6 shows the histogram of maximum vowel contribution ($\max_{t=T_1 \cdots T_2} C(t)$) with

style-independent targets. For both male and female speech, approximately 60 % of CLR speech tokens reach close to 1.0 within the vowel region, while 25 to 30 % of CNV speech tokens reach close to 1.0. The shallow slope in s and/or p locations that are close to each other can result in lower vowel contribution. The difference between CNV and CLR was not always present in the value of s (Section 6.5.5), however the combination of s and p in vowel contribution revealed a difference in formant undershoot.

6.5.7 Estimated formant target values

The parameter estimation process resulted in 20 sets of formant target values with style-independent target estimation, and 10 sets each for two speaking styles with style-dependent target estimation. Figures 6.7 ($\mathbf{T_V}$) and 6.8 ($\mathbf{T_{C_1}}$ and $\mathbf{T_{C_2}}$) show estimated style-independent target values from the 20 different training sets.

Figures 6.9 ($\mathbf{T_V}$), 6.10 ($\mathbf{T_{C_1}}$) and 6.11 ($\mathbf{T_{C_2}}$) show estimated style-dependent target values from the 10 different training sets, while mean values from style-independent targets are shown in squares as a comparison. Selected consonants (/t/, /ɻ/, /l/, /j/ and /w/) are shown.

The estimated vowel target values ($\mathbf{T_V}$) are tightly clustered with different data partitions both for style-independent and style-dependent target estimation. The estimated vowel target values from style-independent estimation (shown in black squares in Figure 6.9) are closer to the CLR targets (filled blue triangles) than those from the style-dependent estimation.

For the voiced consonants ($\mathbf{T_{C_1}}$ and $\mathbf{T_{C_2}}$) /w/, /j/, /ɻ/ and /l/, where formant values are available, the estimated target values are tightly clustered, similar to the vowels. The consonant /ɻ/ in the C_1 position, in particular, shows two distinct target classes for CNV and CLR speaking styles. For the voiced stop consonants /b/, /d/, /g/ (not shown in these figures), even though formant values are not available, the estimated targets are closely clustered, while the unvoiced consonants (only /t/ is shown in Figures 6.8, 6.10 and 6.11) show more scattered estimated values.

Previously, the formant contour model was fitted to words with a limited context (/w/–/V/–/l/ and /t/–/V/–/l/) in Chapter 5. Results showed that the estimated target values tended to be located at the most extreme F1 and F2 values. Unlike this previous study in which we had limited phoneme contexts, the current style-independent target estimation yields higher error because the model cannot reduce the error simply through extreme target values.

Figures 6.12 show observed F2 contours for C_1 approximants and the vowel /ɛ/, for both speaking styles and for both speakers. The CLR formant contour reaches F2 of vowel /ɛ/ at a higher frequency, regardless of context, compared with CNV speech. Because the model works well in the direction of undershoot, in the style-independent target estimation, the model tends to put target /ɛ/ at the extreme position (high F2 value) for /w, l, ɹ/–/ɛ/ transition. On the other hand, for /j/-/ɛ/ transition, the model works well when the target is at a lower F2 value. Therefore, the style-independent estimation yielded a larger error.

In summary, based on the lower test-set error from model fitting, the style-dependent target is more likely to reflect human speech production. However, it may not be a fair comparison between the two estimations, because of the number of parameters and amount of training data we used in this experiment. We further examine the style-dependency of the targets in next section.

6.6 Experiment 6–2: Data-driven consonant target

The formant contour model allows us to estimate formant targets in unvoiced consonants using only available formant data. It is possible because of the symmetric characteristics of the sigmoid function as a coarticulation function, in which unvoiced-consonant targets are projected using the formant transition information. The sigmoid function fits well to sonorant vowel data, and so it is natural to extend this function to unvoiced consonants. In this section, we examine the estimated consonant target values using this data-driven approach, as opposed a rule-based approach.

Model parameters (s_1, p_1, s_2, p_2 and $\mathbf{T}$) were estimated in 20 groups of training sets with style-independent target estimation. The amount of training data was increased from the previous estimation (Section 6.5.1). 460 tokens (K) were used in the training set per style, which comes to the total of 920 training tokens from both styles in Equation 6.5. The additional constraint of vowel contribution was added, so that $d_1(t; s_1, p_1) + d_2(t; -s_2, p_2) \geq 0.40$ was met at least one time. The hill-climbing approach was used for the model parameter estimation.

As the result of style-independent estimation, Figures 6.13(a) and 6.13(b) show the estimated consonant formant targets for the male and female speaker, respectively, where only bilabials and alveolars are shown. Each phoneme is plotted with 20 target values from each training group, which is more tightly clustered than in the case of the smaller

set of training data (Figure 6.8). With the increase in training data, the formant targets, especially for the unvoiced consonants, became more consistent estimates.

The effects of place of articulation were observed in F2 for both speakers. Mainly, the speakers produce bilabial and alveolar in different F2 target positions. With little variation, F2 of alveolars are located between 1500 Hz and 2000 Hz. A difference in F2 targets based on consonant position was also observed, where C_2 targets were usually higher than C_1 targets. It should be noted that the estimated targets depend on the speaker for some consonants, as well. As shown in Figures 6.13(a) and 6.13(b), F1 of prevocalic /t/ has a mean of 511 Hz for the male speaker and 390 Hz for the female speaker. The estimated /p/ targets for the female speaker are not as consistent as other consonant targets. In order for the unvoiced consonant to be consistently estimated, the model requires formant movement within the vowel (mainly at the formant transition). The words with C_1 =/p/ for female speech show a flat formant contour, which might have led to non-consistent /p/ targets. The mean values of estimated targets for all consonants are reported for two speakers along with generic values by Allen *et al.* [1] in Table E.1, Appendix E.

Mean $E^{(k)}_{s,target}$ values over the samples in the training and test sets are reported for each style in Table 6.8 (Style-Indep. II. condition). The mean error rate in the training sets was slightly higher than the previous results described in Section 6.5.3, but slightly lower in the test sets. The smaller error difference between training and test sets indicates the style-independent targets with an increased amount of training data are better estimates of the model than previous style-independent estimation. The error rate from the style-independent estimation was still slightly higher than the style-dependent case, which suggests the style-dependent estimate is still a slightly better model of formant contours.

The test-set results were submitted to three way analysis of variance (ANOVA) (2 speakers $\times$ 2 speaking styles $\times$ 8 vowel identities). All of the main effects were significant ($F(1, 1934) = 121.04$, $p < 0.0001$, $F(1, 1934) = 5.95$, $p = 0.0148$, $F(7, 1928) = 15.23$, $p < 0.0001$, respectively). Similar to the previous results (Section 6.5.3), the post-hoc analysis (HSD) showed that the model fitted better with the male speech than female speech. CNV speech was better fitted to the model than CLR speech. The error rate on words with the vowel /u/ was significantly different from vowels (/iː/, /ɪ/, /ɛ/, /ʊ/, /ʌ/, and /ɑ/), the vowel /iː/ was different from /æ/ and /ɑ/, and the vowel /ɪ/ was different from /æ/ at $\alpha = 0.0018$ (0.05/28) level. The error rate varies based on the vowel identities more than we found in the previous estimation (Section 6.5.3).

In summary, we estimated consonant targets consistently by increasing the amount of training data with style-independent target estimation, even for non-approximants. The estimated targets demonstrate the expected characteristics of place of articulation, and also show differences in consonant position and speaker differences. These results are the first known report of a data-driven approach to estimating consonant targets.

6.7 Discussion: Speaker dependency

In this chapter, the number of speakers was increased to two (male and female) to examine speaker differences. Since one speaker per gender was recorded, we focus on speaker differences rather than gender differences. First of all, both speakers were able to produce CLR speech significantly differently from CNV speech, in terms of formant steady-state values, formant transitions, phoneme durations, and F0 (Section 6.3). However, the formant contour plots and formant steady-state values showed that the female's speech has more variability between CNV and CLR speech in the vowel space, which resulted in a higher fitting error. The contribution of the vowel for both speakers demonstrated a similar effect, namely that the CLR style has more likely fully-articulated vowels than the CNV style.

When the amount of training data was matched per training group with style-independent and style-dependent estimation, the estimated formant target values showed similar results for both male and female speakers (Section 6.5.7). That is, the estimated target values are well clustered for the vowels and approximants, and not as well clustered for non-approximants. Although the estimated values for unvoiced consonants might not be similar between the two speakers, the estimated target values are better clustered with style-dependent than style-independent target estimation for both speakers.

When the amount of training data was increased per training group, for style-independent estimation, well-clustered consonant targets were obtained even for non-approximants (Section 6.6). Both speakers demonstrate a difference in F2 based on the place of articulation.

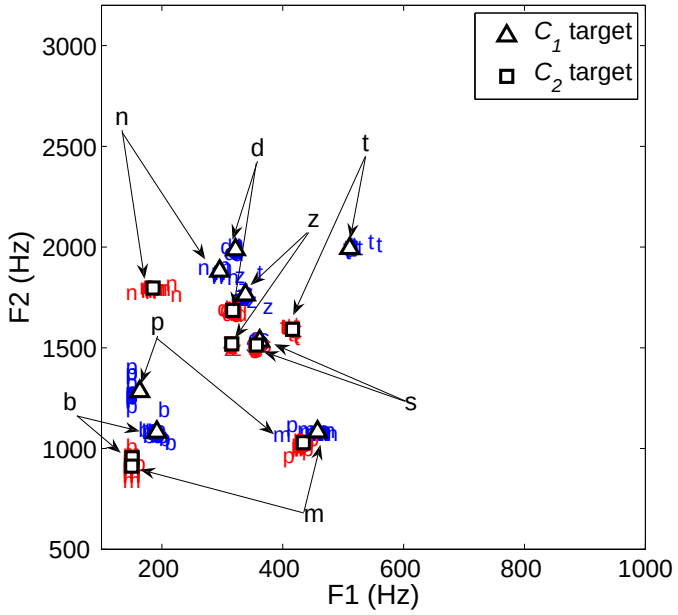
6.8 Conclusions

In this study, we presented a newly-developed speech corpus of CVC words with CNV and CLR speaking styles, and analyzed the acoustic characteristics of this corpus. A previously developed formant contour model (Chapter 5) was verified on this greater variety of phoneme contexts with a linear combination of coarticulation functions and

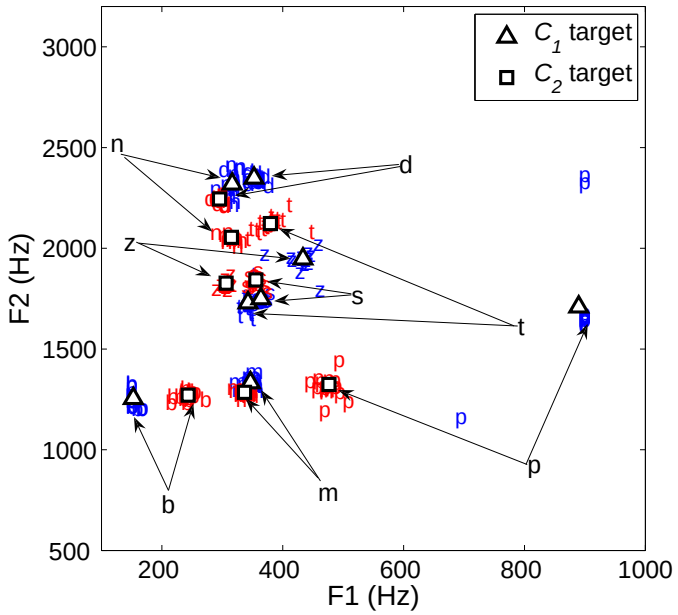
style-independent formant target values. The model demonstrated relatively low error rates on training and test data, well-clustered formant target values for approximants from the training dataset, and a small error rate from model validation. A larger variance in the female speech resulted in a higher fitting error rate than for the male speech. The characteristics of CLR formant dynamics are partially found in the parameters of the coarticulation functions (/w/–/V/–/l/ context only), as well as in the contribution of the vowel. These characteristics may be useful for formant-based speech synthesis in the future.

We examined whether style-dependent targets would yield a better model of human speech production, compared with style-independent targets. The mean test-set error rate was reduced with style-dependent target estimation, while both CNV and CLR formant contours were fitted equally well. On the other hand, with the doubled size of training data with style-independent estimation, the estimated consonant targets are found to be much more tightly clustered. The estimated targets might depend on the size of training data. Due to the limited amount of data in our CVC speech corpus, it is not possible to increase the size of training data for style-dependent estimation. We speculate that with a larger amount of training data, the clustering of unvoiced consonants in style-dependent estimation could be as good as, or better than, style-independent estimation. Although it is not conclusive about the style dependency of the target, because of the reduced test-set error, the style-dependent target is more likely.

The perceptual effects of the modeled formant contour are not yet confirmed. In the future, the formant contour model will need to be validated using a re-synthesized signal and perceptual testing. When the original and re-synthesized signals are not perceptually different, we will be able to use model-based formant synthesis to create a speech signal as natural as the original signal. This future plan for perceptual evaluation is discussed in Section 8.4.1.



(a) Male speech



(b) Female speech

Figure 6.13: Estimated style-independent consonant target values (showing only bilabials and alveolars) in F1−F2 space with larger amount of training data. The blue and red fonts represent C_1 and C_2 values, respectively. The means of each phoneme from style-independent target estimation are shown in triangles (C_1) and squares (C_2).

Chapter 7

Applications of the formant contour model

In previous chapters (Chapters 3 and 4), the spectrum, and formant frequencies in particular, is shown to be a relevant feature for the increased intelligibility of CLR speech. Other researchers have investigated abnormalities in formant frequencies or F2 rate in dysarthric speech [69, 52], which may contribute to the reduced intelligibility of dysarthric speech. In addition, other studies have showed that formant information, especially transition information, can be used to diagnose learning disability [13] and stuttering as distinct from non fluency [85]. Also, modifying the formants has been shown to be an effective way to improve vowel intelligibility of dysarthric speech [49]. In these studies, extracting formant values is an important process for analysis and synthesis. To obtain accurate formant information, the annotators usually correct formant-tracking errors manually, by visually inspecting the spectrogram and using phonemic information. Manual correction of formant frequencies is not an easy task and is labor intensive, requiring knowledge of the speech signal in cases of ambiguous or invisible resonant frequencies. Therefore, errors in formant tracking prevent researchers from examining a larger number of data samples.

In this chapter, we investigate methods to reduce or detect formant contour errors made by existing software, as one potential application of the formant contour model. We apply the formant contour model (Chapter 6) using automatically-extracted formant contours without manual correction. We examined (1) whether the modeled contour can reduce formant-tracking errors, (2) whether we can detect tokens that have formant-tracking errors using only automatically-extracted formant information, and (3) whether we can improve the accuracy in extracting F2 slope.

7.1 Experiment 7–1: Reducing formant-tracking errors

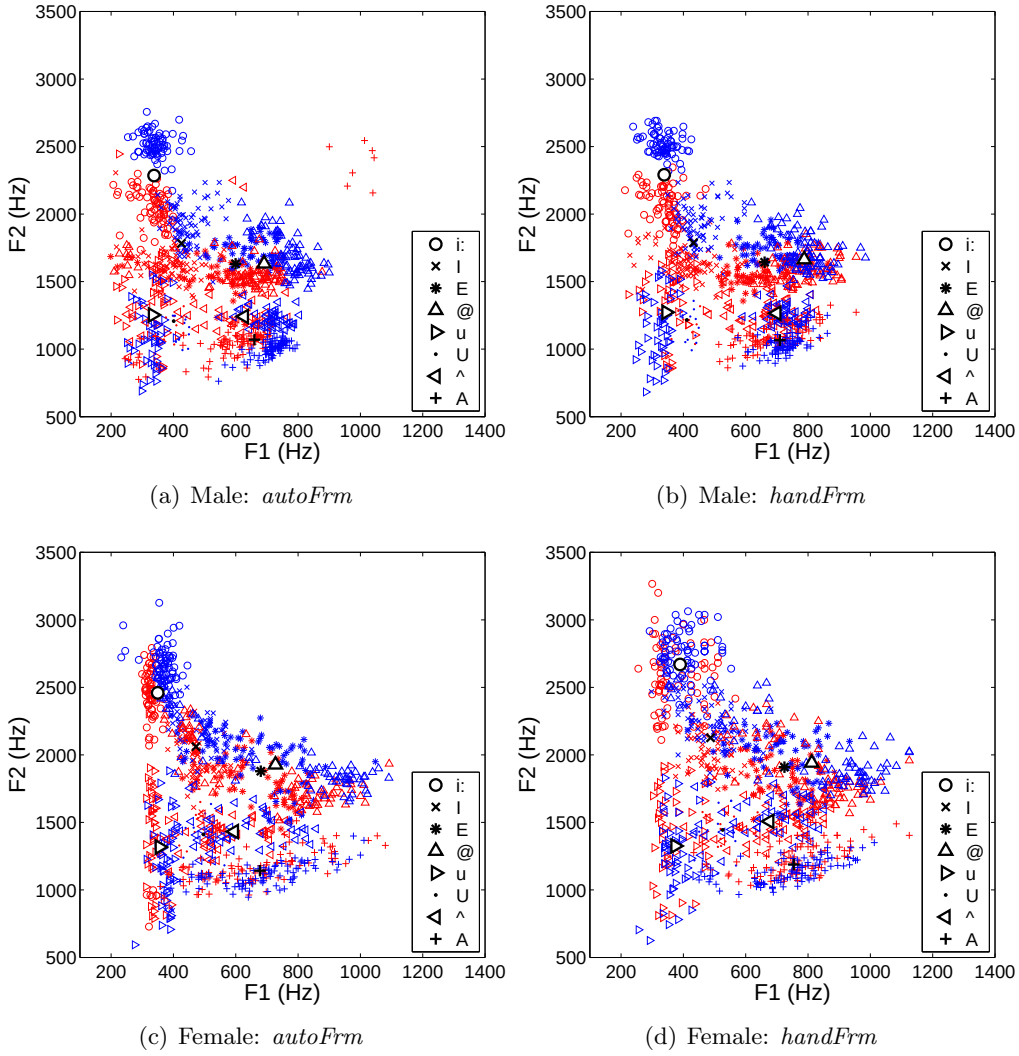
We examined whether we can model formant contours using automatically-extracted formant values instead of manually-corrected formants. Automatically-extracted values may contain formant-tracking errors, which may influence the robustness of parameter estimation. Then, we tested how much we can reduce formant-tracking errors. We used the same 242 CVC corpus described in Section 6.2, from the same two speakers.

The formant contour was initially extracted with the Snack Sound Toolkit [91, 90], which uses ESPS $Waves+$ libraries [28]. Without manual correction, it contains formant-tracking errors. We refer to automatically-extracted formants as *autoFrm*, and manually-corrected formants as *handFrm*. Figures 7.1(a)–7.1(d) show the observed steady-state values of *autoFrm* and *handFrm* for two speakers, taken at the vowel midpoints.

7.1.1 Formant target estimation

The *autoFrm* contours were fitted to the formant contour model by adjusting parameters (s and p per token, and global targets $\mathbf{T}$). The method to estimate model parameters was identical with Section 6.6. 20 groups of training sets, which have 920 tokens for both speaking styles per training group, were used in style-independent target estimation. The style-independent target estimation was used because it has more applicable situations for future use than style-dependent target estimation. The errors in Equations 6.4 and 6.5 were minimized iteratively using the hill-climbing method.

Estimated vowel targets from *autoFrm* and *handFrm* are shown in Figures 7.2(a)–7.2(b). 20 values per vowel from 20 training groups are shown in blue (*autoFrm*) and red (*handFrm*), while mean values of 20 estimated targets are shown in black downward triangles (*autoFrm*) and squares (*handFrm*). For all cases, the estimated vowel targets are tightly clustered as seen in previous results (Section 6.6). For male speech, the vowel F1 values with *handFrm* are higher than those with *autoFrm*; especially 63 Hz, 75 Hz and 54 Hz higher in vowels /ɛ/, /æ/ and /ʌ/ than *autoFrm* cases, respectively. Larger differences were found in female speech. Not only F1 differences in F1 of /ɛ/, /æ/, and /ʌ/, but also F1 of /ɑ/ and F2 of /iː/ and /ɪ/ were different by 60 Hz, 206 Hz and 92 Hz, respectively. The target values using *handFrm* were at more extreme positions in the F1–F2 space.



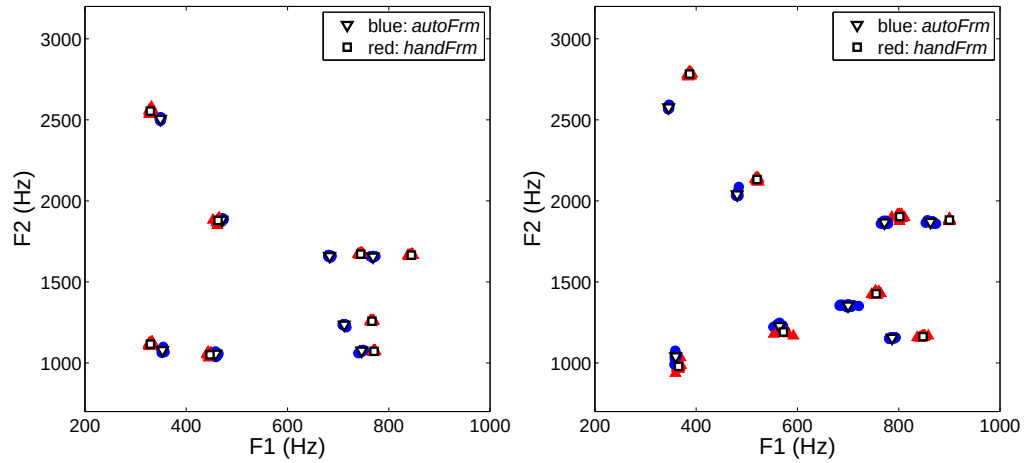
(a) Male: *autoFrm* (b) Male: *handFrm*

(c) Female: *autoFrm* (d) Female: *handFrm*

Figure 7.1: The steady-state values of *autoFrm* and *handFrm* for two speakers. CNV and CLR tokens are shown in red and blue, while mean values are shown in black.

7.1.2 Re-estimating coarticulation parameters

With a given estimated target, the modeled formant contour fitted well to the *autoFrm* contour, because of the flexibility in coarticulation functions, even if *autoFrm* contained formant-tracking error(s). An example of *autoFrm* with a tracking error and the corresponding fitted model are shown in Figure 7.3(a). In this example, a formant-tracking error in the second half of the vowel /ʌ/ was observed. Since p_2 was located in the middle of the vowel, the modeled contour was away from the *handFrm* contour, but fitted the *autoFrm* contour well. We restricted the p range to prevent cases like this, and increased



(a) Male speech (b) Female speech

Figure 7.2: Estimated vowel target from the training data using *autoFrm* (blue) and *handFrm* (red). Formant targets are style-independently estimated. The black markers show the mean of 20 values.

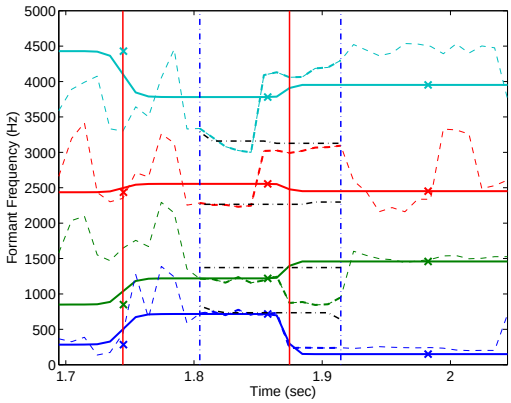
the minimum amount of vowel contribution ($C(t) = 1 - (d_1(t; s_1, p_1) + d_2(t; -s_2, p_2))$).

The coarticulation parameters (s and p) were re-estimated given mean estimated target values with these extra constraints. The constraints were:

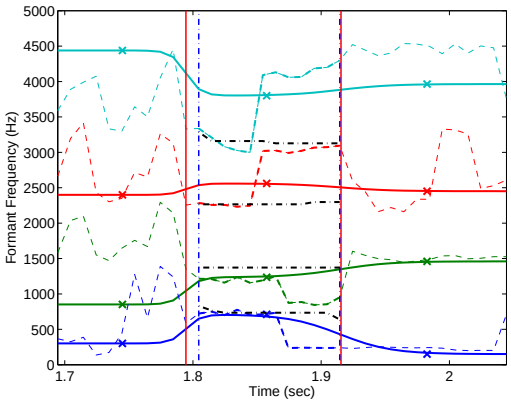
- $C(t) \geq 0.75$ was met at least one time.
- p_1 (or p_2) in Equation 6.3 was placed either at the end of C_1 or beginning of V (or the end of V or beginning of C_2).

The modeled contour is shown in Figure 7.3(b) with re-estimated s and p values. We refer to this modeled contour as $autoFrmModel$. With extra constraints, p values are located close to the phoneme boundaries, and $autoFrmModel$ is closer to $handFrm$. The model was also fitted to the $handFrm$ data with the same constraints, which we refer to as $handFrmModel$.

When coarticulation parameters (s and p) were estimated, the target values of /h/, /k/, /g/, and /ŋ/ were changed, because these consonants are largely influenced by their neighboring vowel. The target of /h/ was set to the observed vowel values at the CV boundary. For /k/, /g/, and /ŋ/ in a front vowel context (determined by F2 value $\geq$ 1500 Hz), F2 was set to $\frac{F2+F3}{2}$, while F1, F3 and F4 were set to the estimated target values for corresponding consonant. For /k/, /g/, and /ŋ/ in a back vowel context (determined by F2 value $<$ 1500 Hz), all four formant targets were set to the estimated target values



(a) Modeled formant contour (solid lines) without the extra constraint and observed contour (dotted lines). Black dash-dot lines are the *handFrm* contour. The model contour was fitted well to the *autoFrm* contour containing formant-tracking errors.



(b) Modeled formant contour (solid lines) with the extra constraint and observed contour (dotted lines). Red lines (p_1 and p_2) are restricted to be close to C_1V/VC_2 boundaries. Black dash-dot lines are the *handFrm* contour.

Figure 7.3: An example of the contour model before and after re-estimatation of the coarticulation parameters. The word is "*fun* (/f ʌ n/)" (male speech) in CNV style. Blue vertical dash-dot lines show the phoneme boundaries, while vertical red dashed lines represent p_1 (left) and p_2 (right).

for the corresponding consonant, same as other consonants. Each formant contour was then synthesized with mean formant target values and re-estimated s and p values.

7.1.3 Error analysis

The error was calculated as

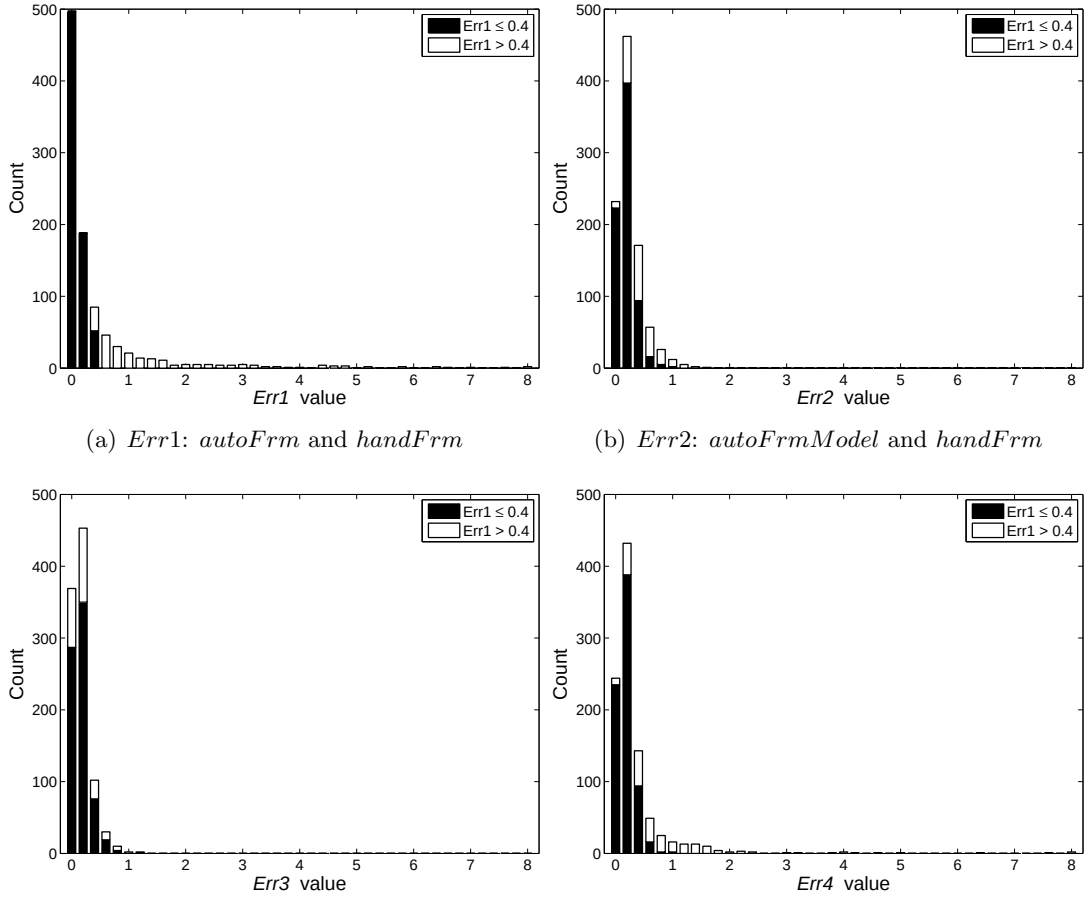
$$Err^{(k)} = \sqrt{\frac{\mathbf{w}\sum_{t=T_1^{(k)}}^{T_2^{(k)}}\left(\mathbf{F_1}(t)^{(k)} - \mathbf{F_2}(t)^{(k)}\right)^2}{N^{(k)}}} \tag{7.1}$$

where $\mathbf{w} = \begin{bmatrix}1.0 & 1.0 & 0.25 & 0.01\end{bmatrix}$, and $N^{(k)}$ is the number of frames from $T_1^{(k)}$ to $T_2^{(k)}$ for a particular token k. The error is not divided by the sum of the weights, because it does not affect the ranking. $T_1^{(k)}$ (or $T_2^{(k)}$) is located at the middle of C_1 (or C_2) when C_1 (or C_2) is an approximant. Otherwise, $T_1^{(k)}$ (or $T_2^{(k)}$) is located at the C_1V (or VC_2) boundary.

$\mathbf{F_1}(t)^{(k)}$ and $\mathbf{F_2}(t)^{(k)}$ were set as follows:

$Err1$: $\mathbf{F_1}(t)^{(k)}$=*autoFrm* and $\mathbf{F_2}(t)^{(k)}$=*handFrm*.

$Err2$: $\mathbf{F_1}(t)^{(k)}$=*autoFrmModel* and $\mathbf{F_2}(t)^{(k)}$=*handFrm*.



(a) $Err1$: $autoFrm$ and $handFrm$

(b) $Err2$: $autoFrmModel$ and $handFrm$

(c) $Err3$: $handFrmModel$ and $handFrm$

(d) $Err4$: $autoFrmModel$ and $autoFrm$

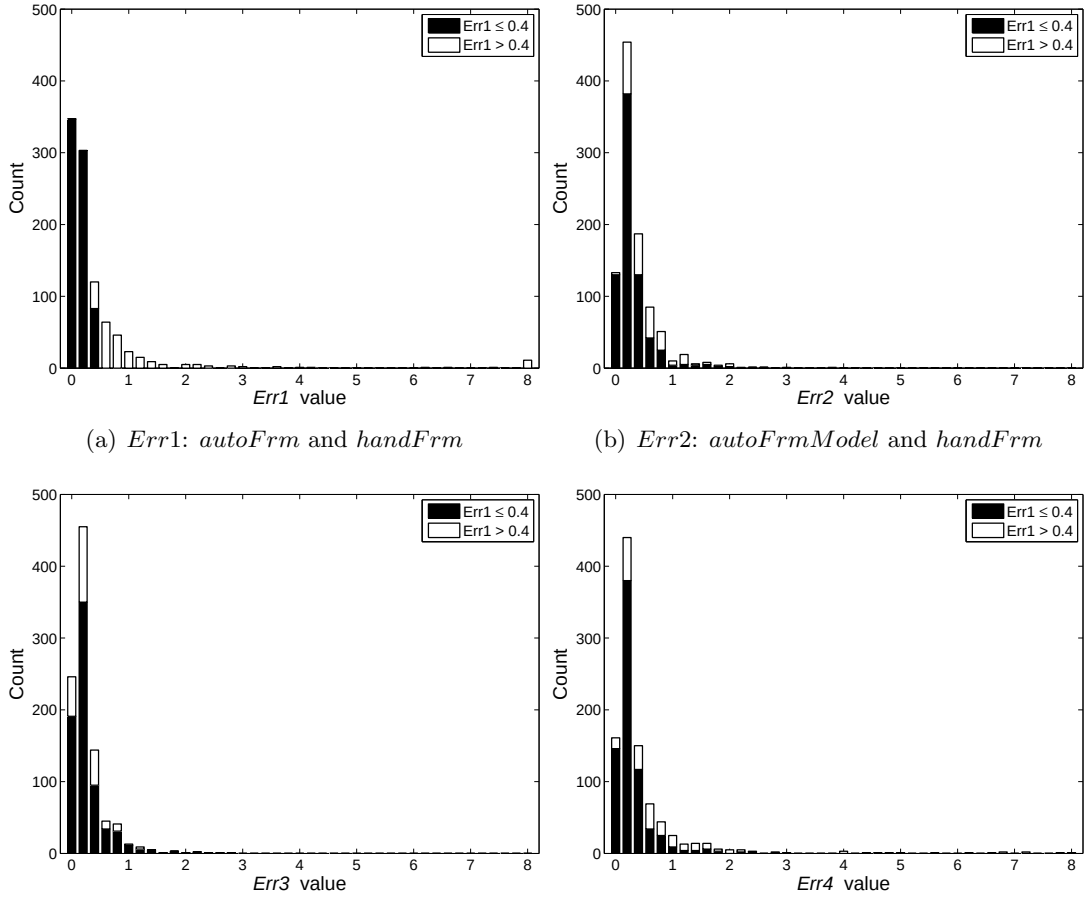
Figure 7.4: Histograms of $Err1$ through $Err4$ for the male speaker. Filled bars are the tokens that have $Err1 \leq 0.4$, and open bars are the tokens that have $Err1 > 0.4$.

$Err3$: $\mathbf{F_1}(t)^{(k)}$=$handFrmModel$ and $\mathbf{F_2}(t)^{(k)}$=$handFrm$.

$Err4$: $\mathbf{F_1}(t)^{(k)}$=$autoFrmModel$ and $\mathbf{F_2}(t)^{(k)}$=$autoFrm$.

The mean error rates of these four conditions are shown in Table 7.1. Since *handFrm* was based on *autoFrm*, $Err1$ can be zero when $autoFrm$ does not have a formant-tracking error, while $Err2$, $Err3$, and $Err4$ cannot be zero except in exceptionally rare cases. To focus on problematic tokens, a subset of tokens were selected based on the threshold of $Err1 > 0.4$. 184 tokens of CNV speech (38.02 %) and 47 tokens of CLR speech (9.71 %) were selected for the male speaker, and 140 tokens of CNV speech (28.93 %) and 95 tokens of CLR speech (19.63 %) were selected for the female speaker.

For the male speech, the mean error was reduced from 1.6000 in $Err1$ to 0.4562 in



(a) $Err1$: $autoFrm$ and $handFrm$

(b) $Err2$: $autoFrmModel$ and $handFrm$

(c) $Err3$: $handFrmModel$l and $handFrm$

(d) $Err4$: $autoFrmModel$ and $autoFrm$

Figure 7.5: Histograms of $Err1$ through $Err4$ for the female speaker. Filled bars are the tokens that have $Err1 \leq 0.4$, and open bars are the tokens that have $Err1 > 0.4$.

$Err2$, showing that $autoFrmModel$ was closer to the "correct" formant contours. The standard deviation was also reduced from 1.5852 to 0.2858. Similarly for female speech, the mean error was reduced from 1.4940 in $Err1$ to 0.5785 in $Err2$. The standard deviation was also reduced from 2.3366 to 0.4864.

Figures 7.4(a)–7.4(d) show the histograms of $Err1$ through $Err4$ for tokens with $Err1 \leq 0.4$ and for tokens with $Err1 > 0.4$ for the male speaker. The histograms show that 56 outliers in $Err1$ ($Err1 > 2.0$) were removed in $Err2$ (maximum value of $Err2$ was 1.6502). 121 out of 231 tokens (52.8 %) had $Err2 < 0.4$. In terms of $Err3$ ($handFrmModel$ as compared with the $handFrm$), 199 out of 231 tokens (86.15 %) had $Err3 < 0.4$. For those tokens (initially had $Err1 > 0.4$), the model worked well at reducing the error from $autoFrm$. On the other hand, 45 tokens (49 tokens) out of 231 had

Table 7.1: The mean error rate (standard deviation) in four conditions. 184 tokens (or 140 tokens) of CNV speech and 47 tokens (or 95 tokens) of CLR speech for male (or female) are selected based on the threshold of $Err1 > 0.4$.

		All tokens			Selected tokens ($Err1 > 0.4$)		
		CNV	CLR	Total	CNV	CLR	Total
Male speech	$Err1$	0.7560	0.1449	0.4505	1.8071	0.7894	1.6000
		(1.3363)	(0.2651)	(1.0102)	(1.7027)	(0.4323)	(1.5852)
	$Err2$	0.3255	0.1709	0.2482	0.4984	0.2930	0.4562
		(0.2609)	(0.1173)	(0.2165)	(0.2995)	(0.1333)	(0.2858)
	$Err3$	0.2085	0.1466	0.1776	0.2118	0.1788	0.2050
		(0.1825)	(0.1092)	(0.1535)	(0.2141)	(0.0982)	(0.1964)
	$Err4$	0.5153	0.1989	0.3571	1.0103	0.6218	0.9312
		(0.8954)	(0.2345)	(0.6730)	(1.2963)	(0.5197)	(1.1898)
Female speech	$Err1$	0.4987	0.4292	0.4640	1.3731	1.6722	1.4940
		(1.2737)	(1.3115)	(1.2925)	(2.1265)	(2.6173)	(2.3366)
	$Err2$	0.3223	0.3927	0.3575	0.4828	0.7195	0.5785
		(0.2847)	(0.4372)	(0.3704)	(0.3196)	(0.6357)	(0.4864)
	$Err3$	0.2498	0.2932	0.2715	0.2350	0.3014	0.2619
		(0.2392)	(0.3552)	(0.3035)	(0.1963)	(0.2450)	(0.2193)
	$Err4$	0.4626	0.4707	0.4666	0.9653	1.1192	1.0275
		(0.9433)	(0.7627)	(0.8574)	(1.6002)	(1.3851)	(1.5157)

$Err2 > 0.4$ (or $Err3 > 0.4$). Those tokens show that the model contour was still different from the "correct" formant contours.

For female speech, the histograms show a similar trend to that of male speech (Figures 7.5(a)–7.5(d)). The maximum $Err1$ value is 16.5160, while the maximum $Err2$ value went down to 3.7358. Similar to the male speech, 105 out of 235 tokens (44.68 %) had $Err2 < 0.4$. 195 out of 235 tokens (82.98 %) had $Err3 < 0.4$. The 31 outliers in which $Err1 > 2.0$ were reduced to 8 cases for $Err2$ ($Err2 > 2.0$), and to 6 cases for $Err3$ ($Err3 > 2.0$). On the other hand, the number of tokens which had $Err2 > 0.4$ among those for which $Err1 \leq 0.4$ was 135 (57.45 %) for $Err2$ and 127 (54.04 %) for $Err3$. Those are the cases when the model did not work well.

The *autoFrm* contours tend to have more formant tracking errors in CNV speech than in CLR speech, because of the low energy of resonant frequencies in CNV speech. Many errors in $Err1$ were often seen on the words with /h/, /ɻ/, nasals, and unvoiced stops, or the vowels /æ/ and /ɑ/ for the male speaker. In addition to these phonemes, the words with /iː/ often had a tracking error for the female speaker. When the model did not work well ($Err1 \leq 0.4$ and $Err2 > 0.4$ (or $Err3 > 0.4$)), many of the words were associated

Table 7.2: The performance rate (%) with several detection thresholds ($\theta_1 = 0.4$).

		$\theta_2 = 0.2$ Detected as		$\theta_2 = 0.3$ Detected as		$\theta_2 = 0.4$ Detected as	
	True class	Positive	Negative	Positive	Negative	Positive	Negative
Male	Positive	89.18	10.82	77.06	22.94	66.23	33.77
	Negative	32.43	67.57	15.47	84.53	6.11	93.89
Female	Positive	82.13	17.87	68.09	31.91	62.55	37.45
	Negative	47.48	52.52	28.24	71.76	18.69	81.31

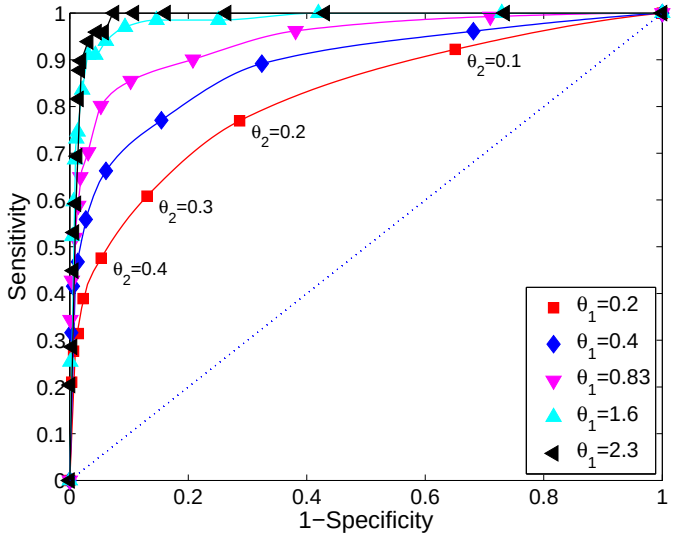
with the consonants /s/ or /l/. In particular, as described in Section 6.5.3, the model had high error rates on words with the vowel /u/.

For both speakers, the $Err4$ results are included in Table 7.1, because the next experiment examines the accuracy of *detecting* tokens that have formant-tracking errors. (This detection can not be done by referencing the $handFrm$ data.) In general, $Err4$ distributions have similar outliers ($Err4 > 2.0$) as $Err1$ (18 outliers for male and 31 for female).

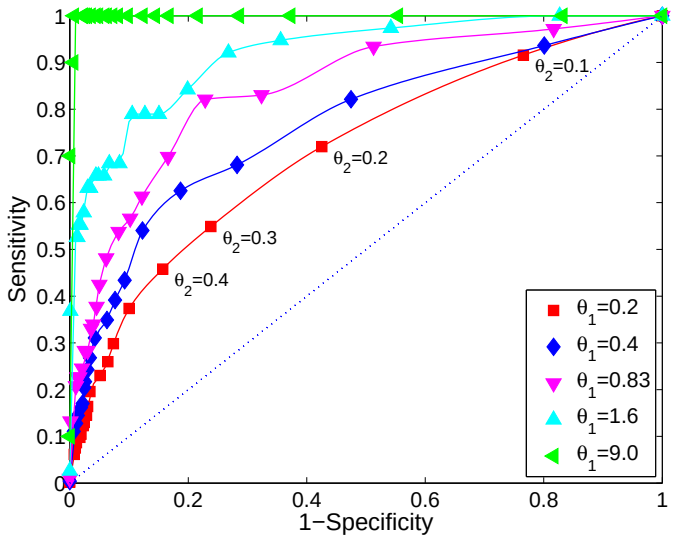
Our method to reduce formant-tracking errors is different from other algorithms such as Lee's approach [60]. One advantage of our method is that we set the vowel contribution ($C(t)$) less than one, which allows us to model coarticulation functions that do not reach the target. In contrast, Lee's algorithm could miss correct formant values due to strong coarticulation, since the correct formant values may exist outside of the search range.

A disadvantage of our method is that even when automatically-extracted formant contours do not have a tracking error (equal to $handFrm$), the model yields errors to some extent. Our results showed that the model yields a mean error rate as low as 0.1776 for male speech and 0.2715 for female speech ($Err3$ for all tokens). ($Err3$ is when the model is fitted to the manually corrected-formant contour.) The reason for this error is partly because the model does not fit well when the estimated target is below the observed contour (if contour is convex) or above (if concave).

In summary, we have been able to estimate model parameters using automatically-extracted formant contours. The estimated vowel formant targets using $autoFrm$ were close to those of $handFrm$, except for the F1 values of /ɛ/, /æ/, and /ʌ/ for the male speech. The female speech showed a larger F1–F2 space of $handFrm$ targets than $autoFrm$ targets. The estimated vowel targets were well clustered across training groups for both speakers. The mean error rate for selected tokens in $autoFrm$ was reduced using $autoFrmModel$. In the next experiment, we examine the possibility of detecting



(a) Male speech



(b) Female speech

Figure 7.6: ROC curve for two speakers.

formant-tracking errors using the error between *autoFrm* and *autoFrmModel*.

7.2 Experiment 7–2: Detecting formant-tracking errors

In this section, we examine how well we can detect tokens that have formant-tracking errors using *autoFrmModel* data. Our "correct" answer as to whether a particular token

has an error is determined by $Err1$. Previously (Section 7.1.3), a total of 231 tokens out of 968 were considered to have a formant-tracking error based on the threshold of $Err1 > 0.4$ (a threshold $\theta_1 = 0.4$). When we set the threshold for $Err4$ to be 0.2 ($\theta_2 = 0.2$) given $\theta_1 = 0.4$, we correctly flagged 206 tokens out of 231 (89.18 %) as having a formant-tracking error, and correctly identified 498 tokens out of 737 (67.57 %) as no error. When we set θ_2 to 0.3, the number of correctly identified as an error was decreased to 178 (77.06 %), and the number of correctly identified as no error was increased to 623 (84.53 %). The results of sensitivity and specificity with $\theta_2 = 0.2$ through 0.4 are shown for both speakers in Table 7.2.

Then, we changed the threshold of $Err1$ (or θ_1) from 0.2 to 2.3 for the male speaker (0.2 to 9.0 for the female speaker) to characterize the error detection performance. Receiver operating characteristic (ROC) curves are shown in Figures 7.6(a) and 7.6(b). As θ_2 was varied from low to high, more tokens were classified as no error (decreasing sensitivity), and also less false negatives were detected (increasing specificity). When a higher θ_1 was used, corresponding to more significant errors, classification performance increased.

One study on formant-tracking algorithms showed that a conventional formant-tracking algorithm has a 13.00 % error for male speech and 15.82 % error for female speech [60]. We assume that the ESPS formant-tracking algorithm has a similar error rate. The error rates 13.00 % and 15.82 % were converted to $\theta_1 = 0.83$ for the male and $\theta_1 = 1.6$ for the female. The area under the ROC curve was 0.9435 ($\theta_1 = 0.83$, ROC curve shown in magenta) and 0.9213 ($\theta_1 = 1.6$, ROC curve shown in cyan) for male and female speech, respectively.

Lee *et al.* improved their formant-tracking error rate to 5.03 % (male) and 3.73 % (female) using phonemic information [60]. We took another measurement for the case assuming that the error rate was similar to that of Lee's algorithm. The $\theta_1 = 2.3$ was equivalent to 5.03 % error (49 tokens) for the male speaker, while $\theta_1 = 9.0$ was equivalent to 3.73 % error (10 tokens) for the female speaker. The area under the ROC curve was 0.9894 ($\theta_1 = 2.3$, ROC curve shown in black) and 0.9989 ($\theta_1 = 9.0$, ROC curve shown in green) for male and female speech, respectively.

For this type of problem, detecting formant-tracking errors, false positives (missing a token with an error) are more critical than false negatives (detecting a correct token as an error). This is true whether the user is looking for reliable formant data and removing bad tokens or looking for tokens to correct manually. The user should choose θ_2 that yields maximum sensitivity, while specificity is reasonably high. The actual performance, however, will depend on θ_1, or how much of a difference is necessary to qualify as an error.

Table 7.3: The mean error rate of the F2 slope (Hz/ms) at the vowel onset and offset positions. The F2 values from *autoFrm* ($Err1$), *autoFrmModel* ($Err2$), and *handFromModel* ($Err3$) are compared with those of *handFrm* per token, averaged over $C_1 - V$ and $V - C_2$ results.

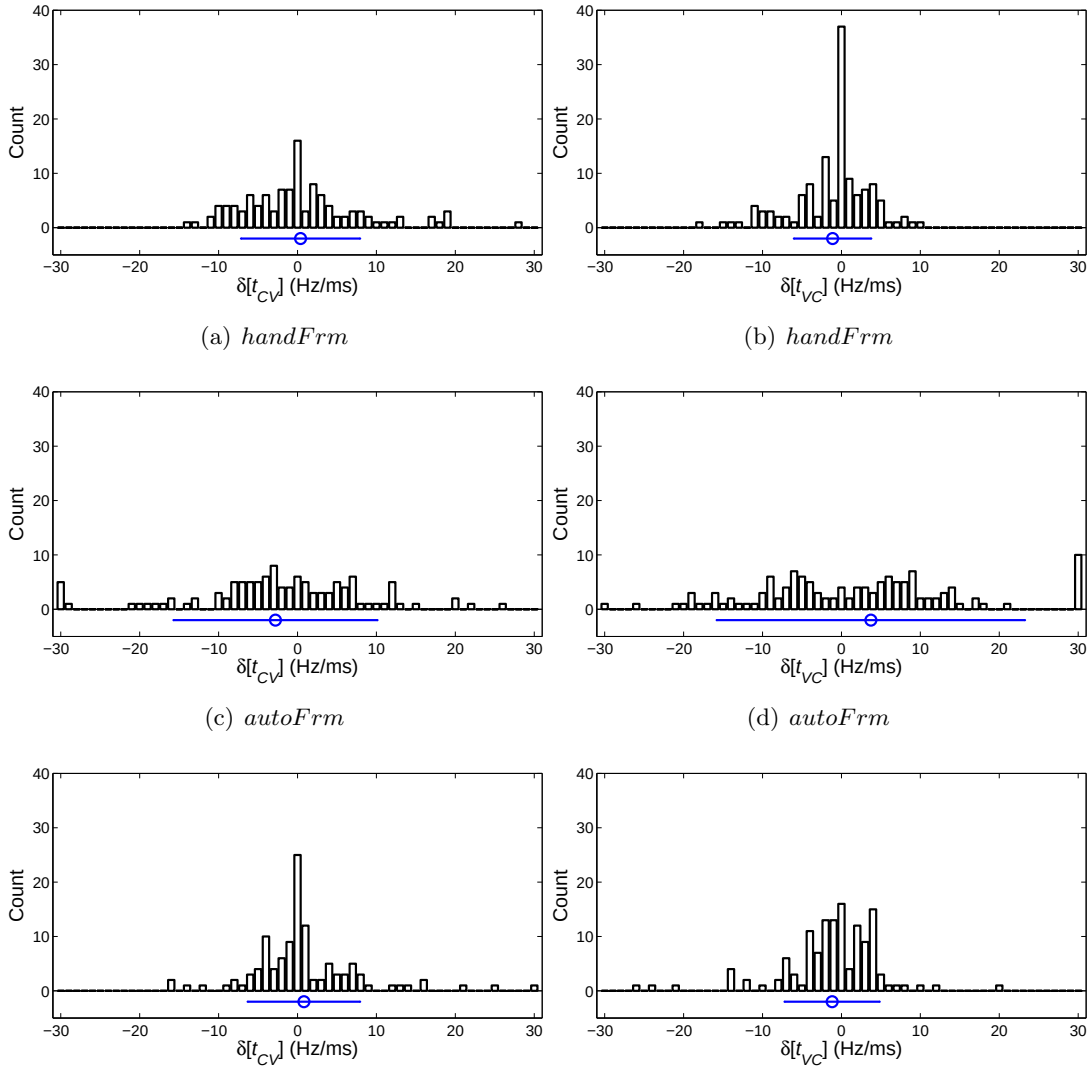
		$C_1 - V$ slope			$V - C_2$ slope		
		CNV	CLR	Total	CNV	CLR	Total
Male	$Err1$	29.2111	13.7493	21.4802	91.0437	21.0100	56.0269
		(225.3951)	(55.0472)	(164.1602)	(539.9426)	(246.3304)	(420.8963)
	$Err2$	15.5952	14.1357	14.8654	16.3280	12.7018	14.5149
		(26.7646)	(29.3005)	(28.0562)	(32.7154)	(40.4759)	(36.8264)
	$Err3$	15.5631	14.1083	14.8357	13.6895	9.4296	11.5595
		(26.7463)	(29.2764)	(28.0349)	(23.4201)	(15.4325)	(19.9366)
Female	$Err1$	51.2548	22.5451	36.8999	112.2834	94.4392	103.3613
		(282.2140)	(74.3125)	(206.7506)	(602.1887)	(481.6179)	(545.0374)
	$Err2$	28.3735	29.6626	29.0181	20.7707	24.7869	22.7788
		(64.4309)	(67.6788)	(66.0438)	(44.3864)	(67.6797)	(57.2363)
	$Err3$	28.3278	29.7524	29.0401	17.8760	12.1120	14.9940
		(64.3721)	(67.6377)	(65.9949)	(35.4843)	(22.3340)	(29.7721)

In summary, we examined how well we can detect formant-tracking errors by using the error between automatic formant values and the model contour estimated from automatic formant values. The performance rates were shown to be well above chance level for both speakers, which shows that this application can be a useful tool for such speech analysis.

7.3 Experiment 7–3: Extracting F2 slope

The third application is to extract F2 slope using the formant contour model. The formant transition contains important information for speech intelligibility [33]. Many studies have shown F2 slope (or delta, transition) to be a useful acoustic measure for analysis of the speech of ALS patients [50, 69], for stuttered speech [85], and for dysarthric speech [52]. We found that the formant tracking algorithm often makes an error at a voicing transition (unvoiced to voiced or voiced to unvoiced). This is the point of interest, where the F2 slope is measured. In this study, we extract the F2 delta values using three sets of data: *autoFrm*, *autoFrmModel* and *handFrm*. If the results of *autoFrmModel* are similar to those of *handFrm*, F2 slope analysis becomes a less labor intensive process.

The F2 delta values ($\delta[t_{cv}]$ and $\delta[t_{vc}]$ (Hz/ms)) were extracted by fitting a line over the range of $t\pm20$ ms of the CV/VC boundary. When formant values were not available in consonants, only values from t=0 ms to 20 ms at the CV boundary (or t=−20 ms to



(a) $handFrm$ (b) $handFrm$

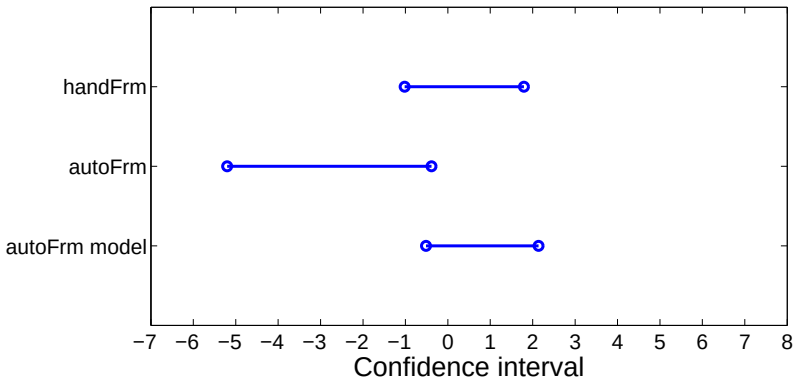
(c) $autoFrm$ (d) $autoFrm$

(e) $autoFrmModel$ (f) $autoFrmModel$

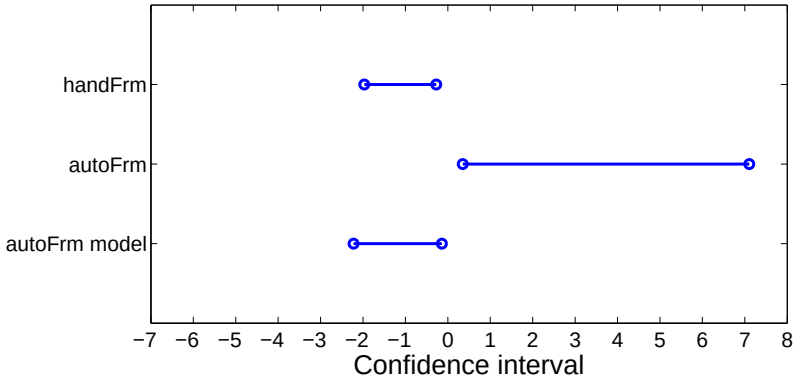
Figure 7.7: Histograms of $\delta[t_{cv}]$ and $\delta[t_{vc}]$ for the male speaker for the selected tokens ($Err1 > 25\,\mathrm{Hz/ms}$).

0 ms at the VC boundary) were computed.

The error was analyzed in terms of the difference between extracted F2 slope of $autoFrm$ and $handFrm$, $autoFrmModel$ and $handFrm$, and $handFrmModel$ and $handFrm$, on a per token basis. The squared error was reported for three datasets: $Err1 : autoFrm$, $Err2 : autoFrmModel$, and $Err3 : handFrmModel$, averaged over all $C_1 - V$ transitions (Err_{cv}) and $V - C_2$ (Err_{vc}) transitions. The results are shown in Table 7.3 for both speakers. Overall, $Err1$ slope was reduced from 56.0269 to 14.5149 in $Err2$ for male speech,



(a) CV boundary

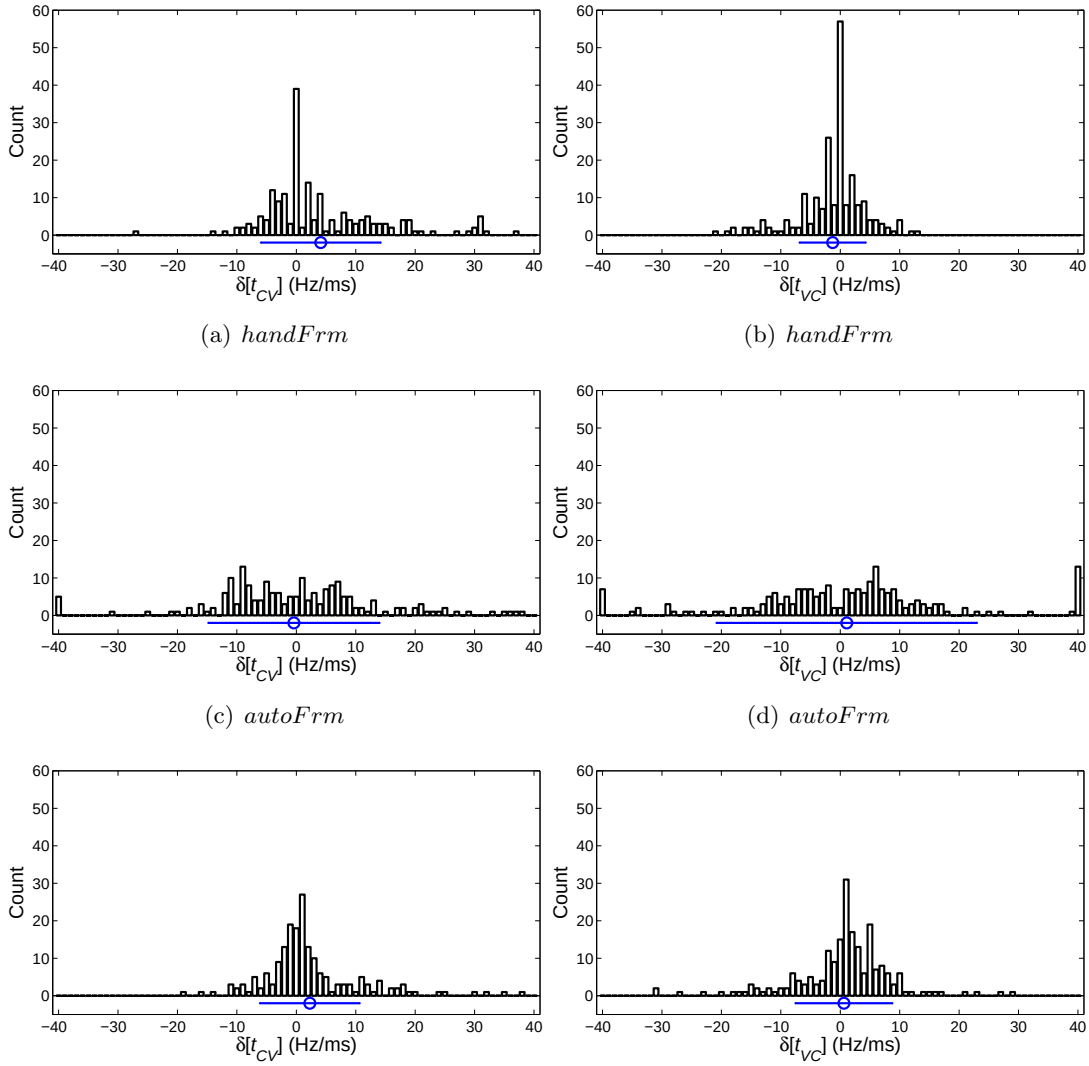


(b) VC boundary

Figure 7.8: Confidence interval of the $\delta[t_{cv}]$ and $\delta[t_{vc}]$ for the male speaker.

and from 103.3613 to 22.7788 for female speech. The standard deviation was also reduced from 420.8963 in $Err1$ to 36.8264 in $Err2$ (male), and from 545.0374 to 57.2363 (female). The $V - C_2$ slope had higher $Err1$ values than $C_1 - V$ slope.

Then, tokens that have $Err1 > 25.0$ (Hz/ms) were further analyzed. A total of 113 tokens of $\delta[t_{cv}]$ and 131 tokens of $\delta[t_{vc}]$ were selected from the male speaker, while 187 of $\delta[t_{cv}]$ and 208 of $\delta[t_{vc}]$ were from the female speaker. The $\delta[t_{cv}]$ and $\delta[t_{vc}]$ distributions are shown in Figures 7.7(a)–7.7(f) (male) and Figures 7.9(a)–7.9(f) (female). At the bottom of each histogram, the mean (circles) and standard deviation (lines) of each dataset are shown. Both speakers show that the F2 slope of $autoFrm$ has many outliers, most of which were removed in $autoFrmModel$ cases. The histograms of $autoFrm$ show flat distributions, while those of $handFrm$ and $autoFrmModel$ have peaks around 0 Hz/ms. The confidence intervals of the three distributions are shown in Figures 7.8 (male) and 7.10 (female). The distributions of $autoFrmModel$ overlap with those of $handFrm$ well,



(a) $handFrm$ (b) $handFrm$

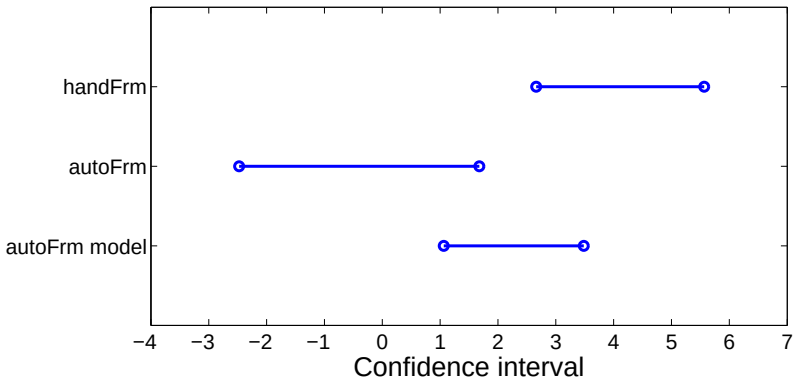
(c) $autoFrm$ (d) $autoFrm$

(e) $autoFrmModel$ (f) $autoFrmModel$

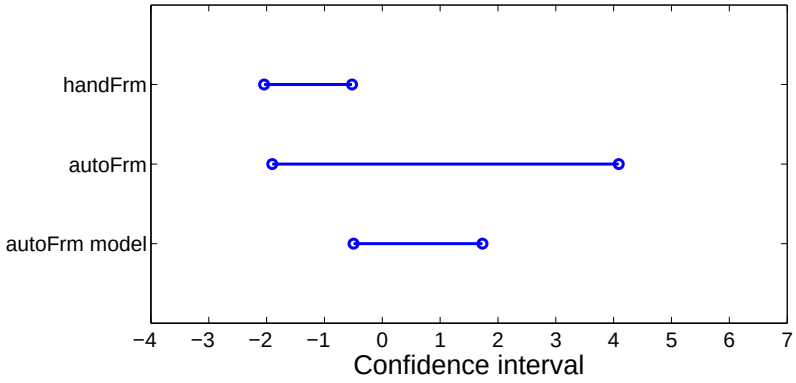
Figure 7.9: Histograms of $\delta[t_{cv}]$ and $\delta[t_{vc}]$ for the female speaker for the selected tokens ($Err1 > 25\,\mathrm{Hz/ms}$).

except for the female at VC boundary.

The tokens that have high error rates ($Err1 > 25$) were often associated with the vowels /æ/ and /ɑ/. Regarding the consonants, the F2 slope of /h/ and /l/ had higher $Err1_{cv}$ and $Err1_{vc}$ rates, respectively. These F2 slope results are consistent with findings from previous results, where we analyzed the formant-tracking errors (described in Section 7.1.3). On the other hand, the words that have higher $Err2$ values were associated



(a) CV boundary



(b) VC boundary

Figure 7.10: Confidence interval of the $\delta[t_{cv}]$ and $\delta[t_{vc}]$ for the female speaker.

with vowels /iː/ and /ɪ/, and consonants /p/, /l/, /m/, and /b/. In these cases, the difference is large between vowel F2 target (high) and consonant F2 (low), regardless of the observed slope (the contour can be flat). We speculate that the model sometimes makes errors when extracting F2 slope in these phoneme contexts. In the future, when s and p values are estimated (Section 7.1.2), minimizing the error in the delta domain may lead to more accurate F2 slope extraction with the contour model.

7.4 Conclusions

In this chapter, we discussed three potential applications of the formant contour model: (1) reducing formant-tracking errors, (2) detecting tokens that have formant-tracking errors, and (3) better estimation of F2 slope values. Our results in the first experiment showed that the model parameters using automatically extracted formant values provide a good

fit to the data and can reduces tracking errors. The second experiment showed that error detection accuracy using the error between $autoFrm$ and $autoFrmModel$ was well above chance level. The area under the ROC curve was 0.9435 ($\theta_1 = 0.83$) and 0.9213 ($\theta_1 = 1.6$) for male and female speech, respectively.

Finally, we examined whether F2 slope values extracted from $autoFrmModel$ are more closely related with those of $handFrm$ F2 slope compared with $autoFrm$. The mean error rate of the F2 slope was reduced from 56.0269 ($autoFrm$) to 14.5149 ($autoFrmModel$). Also, the F2 slope histograms of $autoFrmModel$ had distributions more similar to that of $handFrm$ for both speakers.

From above results, the formant contour model and proposed parameter estimation method has demonstrated potential for use in speech analysis and synthesis tools. It may also be useful for low bit rate speech coding. It should be noted that manually corrected phoneme boundaries were used throughout this chapter, for model parameter estimation and F2 slope extraction. Still more work will be needed to fully automate the process.

Chapter 8

Conclusion

8.1 Contributions of the thesis

Specific Aim 1: To identify high-level acoustic features that are relevant to increased intelligibility of CLR speech. As the result of Experiments 3–1 through 3–3 in Chapter 3, we found that the combinations of (1) phoneme duration, spectrum and phoneme sequence, (2) spectrum and phoneme sequence, and (3) phoneme duration and spectrum were the relevant features. Relevant features were defined in Section 1.1 as the acoustic features that are important to increased intelligibility of CLR speech. These three combinations of features yielded significant improvement over the CNV speech of this particular male speaker. Furthermore, as the result of Experiments 4–1 and 4–2 in Chapter 4, front-vowel intelligibility can be significantly improved by modifying the formant contour, with and without changing phoneme durations. Spectral balance and formant bandwidth did not seem to have an effect on intelligibility as we obtained intelligibility matching CLR speech levels with those features remaining the same.

Throughout the perceptual experiments in Chapters 3 and 4, we developed algorithms to effectively modify certain acoustic features of CNV speech, so the resulting synthetic "hybridized" (HYB) speech has better intelligibility than that of CNV speech, potentially as good as that of CLR speech. In Chapter 3, a pitch-synchronous, residual-excited, linear predictive coefficient (LPC) analysis and synthesis method was used to modify spectral information, F0, and phoneme duration[1]. In Chapter 4, in addition to phoneme duration, formant contours were modified. First, formant contours were designed prior to formant modification so that the CNV formant contour obtained CLR steady-state and transition values. Formant values were then modified by removing existing formant values by inverse

[1]Both Alexander Kain and I were responsible for the implementation of this algorithm.

filtering, and by applying the target formant filter at each time frame. Perceptual results showed that HYB speech had better intelligibility with CLR formant steady-state values. These results indicate that we are able to modify acoustic features of CNV speech without introducing excessive signal-processing artifacts.

As described in the Background (Section 2.3.3), we predicted which features may (or may not) be responsible for increased intelligibility of CLR speech based on prior work that investigated correlations between CNV and CLR speech. From other research, we concluded that formant transitions, temporal envelope, F1 and F2 ranges, energy in the 1–3 kHz range, formant bandwidth, and voice onset time (VOT) are acoustic features that may be responsible for increased intelligibility of CLR speech. F0 mean and consonant-vowel-ratio (CVR) may not be as important to speech intelligibility. Other features, including phoneme duration (speaking rate), F0 trajectories, F0 range, long-term energy, spectral balance (or glottal source characteristics), and pause duration are not conclusively significant for speech intelligibility, because of the unclear or contradicting results from prior studies.

In this thesis, VOT, F1 and F2 range, CVR, and temporal envelope were not explicitly tested. Pause duration was not tested individually as studied in [66], but the results of Experiments 3–1 through 3–3 showed that the combination of pause (non-speech sequence), energy, and F0 did not help to improve intelligibility of CNV speech. Therefore, the pause information, energy, and F0 are not likely to be relevant features for this speaker. Without formant bandwidth modification and glottal source characteristics, we were able to improve intelligibility of CNV speech. Therefore, formant bandwidth and glottal source characteristics may also not be relevant features. The formant transitions were tested with formant steady-state values at the vowel mid point (Section 4.6). Experiment 4-2 showed improved vowel intelligibility with limited phoneme contexts. Therefore, the formant transition can be a relevant feature. Although our expectation was not clear about the impact of phoneme duration, because of the improved sentence and vowel intelligibility along with spectral (formant) information, phoneme duration (or speaking rate) is likely to be a relevant feature.

In this study, sentence materials (Chapter 3) and a limited numbers of words (Chapter 4) were used. The results cannot be easily generalized to other speakers without further testing. However, these results are a significant step forward to better understanding which acoustic-phonetic features *cause* clear speech to be more intelligible than ordinary conversational speech. Finally, one significant contribution of this work is demonstrating that

CNV speech intelligibility can be significantly improved through the use of CLR features.

Specific Aim 2: To develop a model of relevant features that have been identified as a result of Specific Aim 1. We developed a model of formant contours to characterize the relationship between formants and phoneme durations, which is the combination of the two most relevant features from Specific Aim 1 (described in Chapters 5 and 6). In the model, the formant contour was decomposed into formant targets and coarticulation functions. While formant targets were estimated globally (with style-independent or style-dependent targets), slope and slope location parameters in the coarticulation functions were estimated per token. During the parameter estimation process, the error was minimized between the observed and modeled formant contour. With three sets of formant target values (style-independent targets, style-dependent targets, and generic values) in a limited number of /w/–/V/–/l/ words, error values were as low as 0.2062 Bark. An analysis of the relationship between estimated parameters in the coarticulation functions (s_1 and s_2) and a direct measure of F2 slope showed a strong correlation (r =0.8527), indicating the potential to use the direct measure of F2 for computing coarticulation parameters (Chapter 5).

With a larger set of CVC words and an additional speaker, we successfully modeled the formant contours with global targets (either style-independent or style-dependent), slope, and slope location parameters, as described in Chapter 6. The formant contours of the male speaker (specific to this speaker) fit the model better than those of this female speaker, due to the high variance of the female speech formant values. One advantage of the formant contour model is that we can parameterize the coarticulation effects, analyze the formant targets separately from coarticulation effects, and synthesize formant contours with any specified phoneme duration.

A further experiment showed that the style-independent targets were found to be tightly clustered, even for the unvoiced consonants. The estimated consonant targets demonstrated the effect of place of articulation in F2 values. Although some differences between the two speakers were observed, the data-driven approach for the unvoiced-consonant targets is shown to be robust when we have a sufficient amount of training data. We include the complete list of consonant targets, which is the first known dataset obtained by a data-driven approach, in Appendix E.

Specific Aim 3: To develop applications of the formant contour model. We examined three possible applications of the contour model (Chapter 7). First, we

investigated whether we can reduce formant-tracking errors made by existing speech analysis software. We estimated the model parameters with style-independent target estimation using automatically extracted formant contours. Even with formant-tracking errors, the estimated vowel targets were found to be close to the targets estimated with manually corrected formant contours. The error, measuring the difference between manually-corrected formant contour and automatically-extracted formant contour, was reduced from 1.6091 Bark to 0.4562 Bark for the male speaker, and 1.4940 Bark to 0.5785 Bark for the female speaker, when evaluating tokens with high error (> 0.4 Bark).

The second application of the formant contour model considers the possibility of detecting tokens that have formant-tracking errors. For both speakers, the ROC curves showed that the error between modeled contour (estimated with $autoFrm$) and automatically extracted contour is a useful measure of detecting formant-tracking errors. When larger errors were made by the tracking software, better performance of this application was observed. This may be a useful application for speech analysis and synthesis. Using the model, the user can determine which tokens to select (throw away) or which tokens need be corrected manually.

Finally, the third application is to extract F2 slope values from the formant contour model using only automatically extracted formants. The F2 slope distributions showed similar results between manually corrected data and model data. Since F2 slope information is used to diagnose speech-related disorders, automatic F2 slope extraction may be very useful to medical applications.

8.2 Constraints and limitations

In the hybridization algorithm, prior to speech modification, we extracted phonetic features such as time-aligned phoneme labels. Features (glottal-closure instants, LPC analysis, and formant contours) were extracted offline. Manual correction of phoneme labels, glottal-closure instants, and formant frequencies were necessary for accurate analyses and good quality of synthetic speech. In Chapter 7, we showed that the parameters of the formant contour model can be estimated using automatically extracted formant contours. Likewise, an algorithm which does not require manual correction of features is one next step for future development. Such an algorithm will allow us to evaluate much larger datasets.

For the formant contour model, phoneme identity was an essential feature, since the

targets were estimated per phoneme. The model constraints were, in part, based on the manner of articulation. Also, the word sequence studied was always consonant-vowel-consonant (CVC), which makes the contour model more simple. It would be more useful for application to automatic speech recognition (ASR) if the model can be applied to any phoneme sequence or ultimately sentence.

The number of differences between CNV and CLR features depended on the speaker, as discussed in Chapter 6. In this study, the number of speakers was limited to one (male) in Specific Aim 1 and two (male and female) in Specific Aims 2 and 3. The results from these chapters were specific to these speakers. Therefore, these results may not be easily generalizable to other speakers. Investigating speaker-independent features is discussed below in future work (Chapter 8.4.2).

8.3 Applications

In this section, we list possible applications from the findings in Specific Aim 1.

8.3.1 Assistive listening devices

The digital signal processing techniques in current hearing devices includes linear or non-linear amplification in wide/narrow band channels, noise reduction, feedback cancellation, and sound source separation using a directional microphone. The results from this thesis might be applied to portable or wearable devices, including a hearing aid, that possibly transform the input speech (assumed to be CNV speech) into an approximation of CLR speech. Such a device might be tuned to meet an individual's needs to compensate for any hearing difficulties.

Different acoustic features can be incorporated depending on the application devices. For pre-recorded speech or video (e. g. DVD), real-time processing is not necessary. Therefore, more powerful processing and time warping techniques are possible. Whereas, for a hearing aid, reducing computational cost with real-time processing becomes a challenging topic. In particular, slowing down the speaking rate is not an optimal solution because of the potential asynchrony between the audio signal and visual representation. One possible processing scheme might be slowing down the steady-state portions (vowels and approximants) and speeding up the silence portion, as proposed in [72].

8.3.2 Objective measures

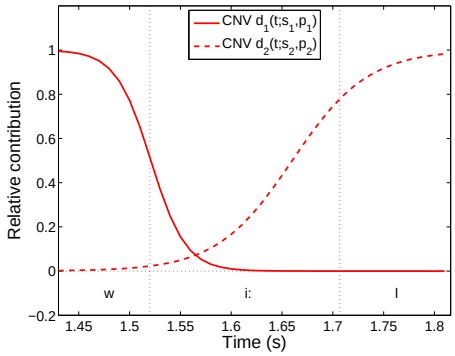
The articulation index (AI), later revised and standardized as the speech intelligibility index (SII), is an objective measurement to predict a listener's audibility, which is highly correlated with the intelligibility of the speech on a particular listener [7]. The SII is calculated with acoustic measurements (or estimates) of speech spectrum level, noise spectrum levels, and the listener's auditory threshold. Band importance functions (numerical value in each frequency band characterizing the relative importance of the frequency to speech intelligibility) are applied to each frequency band, and summed across all frequency bands into a single index. A variety of speech materials, such as the nonsense syllable test [77], CID-W22 [98], NU6 monosyllables [102, 99], Diagnostic Rhyme Test (DRT) [106, 27], short passages [21], and SPIN mono-syllables [8] are taken into consideration in the frequency importance function. The SII value ranges from 0.0 and 1.0, indicating low to high audibility. The speech spectrum is usually estimated with normal, raised, loud, and shouted vocal efforts. Due to the limitation of SII in non-stationary noise conditions, Rhebergen and Versfeld [88] have extended the SII for predicting speech intelligibility taking fluctuating noise, interrupted noise, and multiple-talker noise into consideration. The approach in their study measures the standard SII in a short-term window and averages over the entire speech segment, which yields an SII value for a particular condition.

As shown in the results of Specific Aim 1, speech intelligibility can be heavily affected by the speaking rate. Since the standard SII does not take any temporal features into account, speech intelligibility with different phoneme durations might end up in the same index value, when long-term spectral measures are equal. Similar to the extended SII, a direct measure of phoneme durations might better predict speech intelligibility. The results of this thesis can be applied to objective measures, such as SII, which incorporate a direct measure of speaking rate. More accurate prediction of speech intelligibility may result in more appropriate decisions for hearing aid fitting and counseling.

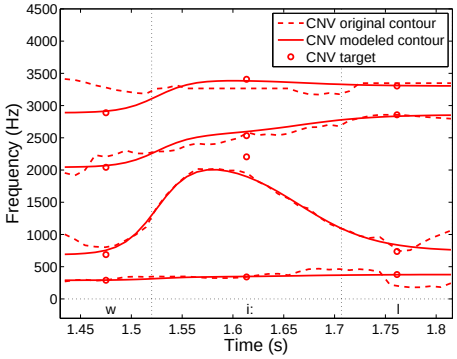
8.4 Future work

8.4.1 Perceptual effects of the formant contour model

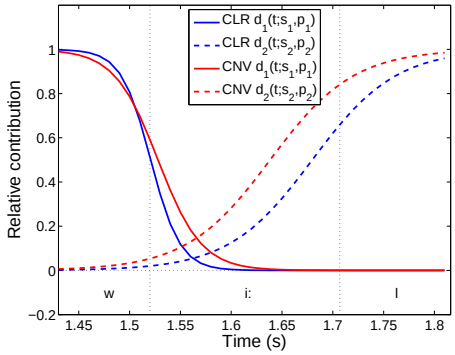
The formant contour model described in Chapter 6 resulted in an average fitting error of 0.2815–0.3834 Bark for style-independent/dependent target estimation for both male and female speech. Whether the fitted model within this error range makes a perceptual difference is not yet answered. Perceptual evaluation using synthetic speech with modeled



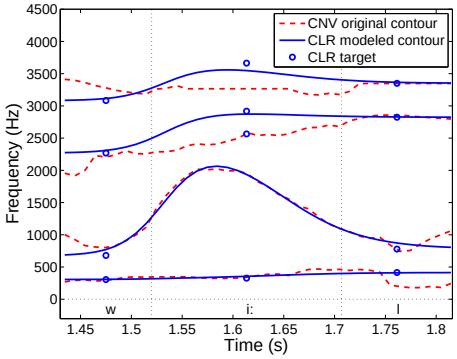
(a) Coarticulation function (d_1 and d_2) of CNV speech. Estimated parameters shown in this figure are $s_1 = 0.5836$, $s_2 = 0.2688$.



(b) Formant contour before (CNV original) and after (CNV modeled contour) the formant modification. CNV style-dependent target values are shown in circles.



(c) Coarticulation function (d_1 and d_2) of CLR speech with CNV duration. Estimated parameters shown in this figure (CLR) are $s_1 = 0.6570$, $s_2 = 0.2531$.



(d) Formant contour before (CNV original) and after (CLR modeled contour) the formant modification with CNV duration. CLR style-dependent target values are shown in circles.

Figure 8.1: Formant contours are modeled with style-dependent target values and coarticulation functions ($d(t; s, p)$), which are from either CLR or CNV speech with CNV duration. The color red is associated with CNV, while blue is with CLR speech. The word *wheel* is shown in this example.

contours is an important next step. Possible test conditions include two types of synthetic formant contours using estimated model parameters as shown in Figure 8.1. In the first condition, Figures 8.1(a) and 8.1(b) show the coarticulation functions with CNV parameters and the formant contour before and after modification with CNV target values. This condition is to validate the formant contour model of CNV speech, where CNV parameters of coarticulation functions and CNV target values are used. In the second condition, Figure 8.1(c) shows coarticulation functions with CNV and CLR parameters with CNV duration. Figure 8.1(d) shows the original CNV formant contour and synthetic contours with CLR target values but CNV duration. This condition is to evaluate whether it is possible to improve CNV intelligibility by modifying the formant contour using CLR model parameters but CNV duration. If the second condition is evaluated to have better intelligibility than the original CNV speech, there is the potential for real-time processing applications. These two examples of synthetic formant contours (and possibly other conditions) will be examined in perceptual experiments in the future.

8.4.2 Speaker dependency

Prior work that found different speakers employ different strategies to produce CLR speech [30] led to a study examining 41 speakers producing *CVC* utterances [29]. Ferguson demonstrated significant speaker differences in vowel intelligibility for normal-hearing listeners [29]. As described in Section 2.2.3, it will be necessary in the future to examine whether a set of relevant features from one speaker is valid for different speakers. To accomplish this goal, we introduce an inter-speaker hybridization method, where we take acoustic features from one speaker and synthesize HYB speech using the complementary features of a different speaker. The intelligibility of the resulting HYB speech will indicate which features are speaker independent.

8.4.3 Speech perception by elderly listeners

40–45 % of people over the age of 65, and about 83 % over the age of 70, experience significant sensorineural hearing loss [22]. In addition, listeners over 60, regardless of their hearing levels, often have difficulty understanding speech in the presence of background noise [61, 84] and reverberation [70]. Presbycusis, a hearing disorder associated with aging that is characterized by high-frequency hearing loss, is also characterized by reduced speech understanding ability. From a five-year longitudinal study, Divenyi *et al.* [23] found that decline in speech understanding ability accelerated significantly with aging, relative

to the decline of audiometric measures. Therefore, speech understanding involves the combination of auditory and cognitive functions.

The CLR speech benefit by elderly hearing-impaired listeners (aged 60–89) has been reported in [79, 89, 30, 105]. Moreover, studies have shown that the benefit of CLR speech depends on the characteristics of a group of listeners, namely whether they have hearing loss and the type of hearing loss [30, 67]. According to a study by Ferguson and Kewley-Port [30], talkers who produced front vowels with significantly higher F2 values were less intelligible due to the listeners' type of hearing loss. The authors speculated that this result was because increased F2 values were in a region where the hearing-impaired listeners had a sloping hearing loss (i. e. 2000–2500 Hz). Therefore, the acoustic features that we found relevant based on the perceptual results from young listeners (Specific Aim 1) might be different for different groups of listeners. For example, temporal features might be more important for elderly listeners who may have temporal processing deficits. In the future, we will need to evaluate what acoustic features are relevant for elderly listeners with or without hearing loss. We assume that in speech signal processing for elderly listeners, if an optimum acoustic signal is provided at the early stages of auditory processing, the cognitive load at later stages of processing will be minimized, which may lead to maximizing speech understanding.

Bibliography

[1] Allen, J., Hunnicutt, M. S., and Klatt, D., Eds. *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[2] Amano-Kusumoto, A., and Hosom, J.-P. The effect of formant trajectories and phoneme durations on vowel intelligibility. In *Proc. of ICASSP* (2009), pp. 4677–4680.

[3] Amano-Kusumoto, A., and Hosom, J.-P. Effect of speaking style and speaking rate on formant contours. In *Proc. of ICASSP* (2010), pp. 4202–4205.

[4] Amano-Kusumoto, A., Hosom, J.-P., and Kain, A. Speaking style dependency of formant targets. In *Proc. of InterSpeech* (2010).

[5] Amano-Kusumoto, A., Hosom, J.-P., and Shafran, I. Classifying clear and conversational speech based on acoustic features. In *Proc. of Interspeech* (2009), pp. 1735–1738.

[6] Anscombe, F. J. The transformation of poisson, binomial and negative-binomial data. *Biometrika 35*, 3 (1948), 246–254.

[7] ANSI-S3.5-1997. *American National Standard methods for the calculation of the speech intelligibility index*. American National Standard Institute (ANSI), 1997.

[8] Bell, S. T., Dirks, D. D., and Trine, T. D. Frequency-importance functions for words in high- and low-context sentences. *Journal of Speech and Hearing Research 35* (1992), 950–959.

[9] Bilger, R. C., and Nuetzel, J. M. Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research 27* (1984), 32–48.

[10] Boersma, P., and Weenink, D. Praat: Doing phonetics by computer (Ver. 4.5.08), 2005. Retrieved Dec. 2006 from `http://www.praat.org`.

[11] Bond, Z. S., and Moore, T. J. A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication 14*, 4 (1994), 325–337.

[12] Boothroyd, A., and Nittrouer, S. Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America 84*, 1 (1988), 101–114.

[13] Bradlow, A., Kraus, N., Nicol, T., McGee, T., Cunningham, J., Zecker, S., and Carrell, T. Effects of lengthened formant transition duration on discrimination and neural representation of synthetic cv syllables by normal and learning-disabled children. *Journal of Acoustical Society of America 106* (1999), 2086–2096.

[14] Bradlow, A. R., and Bent, T. The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America 112*, 1 (2002), 272–284.

[15] Bradlow, A. R., Krause, N., and Hayes, E. Speaking clearly for children with learning disabilities: sentence perception in noise. *Journal of Speech, Language, and Hearing Research 46* (2003), 80–97.

[16] Bradlow, A. R., Torretta, B. M., and Pisoni, D. B. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication 20* (1996), 255–272.

[17] Broad, D. J., and Clermont, F. A methodology for modeling vowel formant contours in CVC context. *Journal of the Acoustical Society of America 81*, 1 (1987), 155–165.

[18] Caissie, R., Campbell, M. M., Frenette, W. L., Scott, L., Howell, I., and Roy, A. Clear speech for adults with a hearing loss: Does intervention with communication partners make a difference. *Journal of American Academy of Audiology 16* (2005), 157–171.

[19] Chen, F. R. Acoustic characteristics and intelligibility of clear speech at the segmental level. *Master's project, Mass. Inst. Tech., Cambridge* (1980).

[20] Collins, M. A permutation test for planar regression. *Australian Journal of Statistics 29* (1987), 303–308.

[21] Cox, R. M., and McDaniel, D. M. Intelligibility rating of continuous discourse: application to hearing aid selection. *Journal of Acoustical Society of America 76* (1984), 758–766.

[22] Cruickshanks, K., Wiley, T., Tweed, B., Klein, B., Klein, R., Mares-Perlman, J., and Nondahl, D. The prevalence of hearing loss in older adults. *American Journal of Epidemiology 148* (1998), 879–885.

[23] Divenyi, P. L., Stark, P. B., and Haupt, K. M. Decline of speech understanding and auditory thresholds in the elderly. *Journal of the Acoustical Society of America 118*, 2 (2005), 1089–1100.

[24] Dreher, J. J., and OŃeill, J. J. Effects of ambient noise on speaker intelligibility for words and phrases. *Journal of the Acoustical Society of America 29* (1957), 1320–1323.

[25] Drullman, R., Festen, J. M., and Plomp, R. Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America 95*, 5 (1994), 2070–2680.

[26] Drullman, R., Festen, J. M., and Plomp, R. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America 95*, 2 (1994), 1053–1064.

[27] Duggirala, V., Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. Frequency importance functions for a feature recognition test material. *Journal of Acoustical Society of America 83* (1988), 2372–2382.

[28] Entropic Research Laboratory, I. *Entropic Signal Processing System (ESPS) waves+*, 1993. Version 5.0.

[29] Ferguson, S. H. Talker differences in clear and conversatoinal speech: Vowel intelligibility for normal-hearing listeners. *Journal of the Acoustical Society of America 116*, 4 (2004), 2365–2373.

[30] Ferguson, S. H., and Kewley-Port, D. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America 112*, 1 (2002), 259–271.

[31] Ferguson, S. H., and Kewley-Port, D. Talker differences in Clear and Conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research 50* (2007), 1241–1255.

[32] French, N. R., and Steinberg, J. C. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America 19* (1947), 90–119.

[33] Furui, S. On the role of spectral transition for speech perception. *Journal of Acoustical Society of America 80*, 4 (1986), 1016–1025.

[34] Gordon-Salant, S. Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing. *Journal of the Acoustical Society of America 82*, 6 (1986), 1599–1607.

[35] Gordon-Salant, S. Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. *Journal of the Acoustical Society of America 81*, 4 (1987), 1199–1202.

[36] Gordon-Salant, S., and Fitzgibbons, P. J. Selected cognitive factors and speech recognition performance among young and elderly listeners. *Journal of Speech and Hearing Research 40* (1997), 423–431.

[37] Gordon-Salant, S., and Fitzgibbons, P. J. Sources of age-related recognition difficulty for time-compressed speech. *Journal of Speech, Language, and Hearing Research 44* (2001), 709–719.

[38] Hazan, V., and Markham, D. Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the American Academy of Audiology 116*, 5 (2004), 3108–3118.

[39] Hazan, V., and Simpson, A. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication 24* (1998), 211–226.

[40] Helfer, K. S. Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal of the American Academy of Audiology 9* (1998), 234–242.

[41] Hillenbrand, J. M., and Clark, M. J. Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America 108*, 6 (2000), 3013–3022.

[42] Hillenbrand, J. M., and Nearey, T. M. Identification of resynthesized /hvd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America 406*, 6 (1999), 3509–3523.

[43] Hoemeke, K. A., and Diehl, R. L. Perception of vowel height: The role of F_1–F_0 distance. *Journal of the Acoustical Society of America 96*, 2 (1994), 661–674.

[44] Hosom, J. P. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication 51* (2009), 352–368.

[45] Iglewics, B., and Hoaglin, D. C. How to detect and handle outliers. *American Society for Quality Control 16* (1993).

[46] International Electrotechnical Commission. *Electroacoustics-sound level meters-part 1: Specifications, 61672*, 2002.

[47] Junqua, J. C. The lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of Acoustical Society of America 93*, 1 (1993), 510–524.

[48] Kain, A., Amano-Kusumoto, A., and Hosom, J.-P. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *Journal of the Acoustical Society of America 124*, 4 (2008), 2308–2319.

[49] Kain, A., Hosom, J.-P., Niu, X., van Santen, J., Fried-Oken, M., and Staehely, J. Improving the Intelligibility of Dysarthric Speech. *Speech Communication 49*, 9 (2007), 743–759.

[50] Kent, R. D., Kent, J. F., Weismer, G., Martin, R. E., Sufit, R. L., Brooks, B. R., and Rosenbek, J. C. Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics & phonetics 3*, 4 (1989), 347–358.

[51] Kewley-Port, D., and Watson, C. S. Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America 95*, 1 (1994), 485–496.

[52] Kim, Y., Weismer, G., Kent, R. D., and Duffy, J. R. Statistical models of f2 slope in relation to severity of dysarthria. *Folia Phoniatr Logop 61* (2009), 329–335.

[53] Klatt, D. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America 82*, 1 (1987), 737–793.

[54] Krause, J. C., and Braida, L. D. Investigation alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America 112*, 5 (2002), 2165–2172.

[55] Krause, J. C., and Braida, L. D. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America 15*, 1 (2004), 362–378.

[56] Krause, J. C., and Braida, L. D. Evaluationg the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *Journal of Acoustical Society of America 125*, 5 (2009), 3346–3357.

[57] Kusumoto, A., Arai, T., Kinoshita, K., Hodoshima, N., and Vaughan, N. Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Communication 45* (2005), 101–13.

[58] Kusumoto, A., Kain, A., Hosom, J.-P., and van Santen, J. Hybridizing Conversational and Clear Speech. *Proc. of Interspeech* (Aug. 2007), 370–373.

[59] Lane, H. L., and Tranel, B. The Lombard sign and the role on hearing in speech. *Journal of Speech and Hearing Research 14* (1971), 677–709.

[60] Lee, M., and van Santen, J. P. H. Formant tracking using context-dependent phonemic information. *IEEE Trans. on speech and audio processing 13*, 5 (2005), 741–750.

[61] Leshowitz, B., and Lindstrom, R. Masking and speech-to-noise ratio. *Audiology & hearing education 5* (1979), 5–8.

[62] Levitt, H. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America 49*, 2 (1971), 467–477.

[63] Liénard, J. S., and DiBenedetto, M. G. Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America 106*, 1 (1999), 411–422.

[64] Lindblom, B. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America 35*, 11 (1963), 1773–1781.

[65] Liu, S., Rio, E. D., Bradlow, A. R., and Zeng, F. G. Clear speech perception in acoustic and electric hearing. *Journal of the Acoustical Society of America 116*, 4 (2004), 2374–2383.

[66] Liu, S., and Zeng, F. G. Temporal properties in clear speech perception. *Journal of the Acoustical Society of America 120*, 1 (2006), 424–432.

[67] Maniwa, K., Jongman, A., and Wade, T. Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *Journal of the Acoustical Society of America 123*, 2 (2008), 1114–1125.

[68] Moon, S. J., and Lindblom, B. Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America 96*, 1 (1994), 40–55.

[69] Mulligan, M., Carpenter, J., Riddel, J., Delaney, M. K., Badger, G., and Tandan, P. K. R. Intelligibility and the acoustic characteristics of speech in amyotrophic lateral sclerosis (als). *Journal of Speech and Hearing Research 37* (1994), 496–503.

[70] Nábělek, A. K., and Robinson, P. Monaural and binaural speech perception in reverberation for various ages. *Journal of the Acoustical Society of America 71* (1982), 1242–1248.

[71] Narne, V. K., and Vanaja, C. S. Effect of envelop enhancement on speech perception in individuals with auditory neuropathy. *Ear & Hearing 29*, 1 (2008), 45–53.

[72] Nejime, Y., Aritsuka, T., Imamura, T., Ifukube, T., and Matsushima, J. A portable digital speech-rate converter for hearing impairment. *IEEE transactions on rehabilitation engineering 4*, 2 (1996), 73–83.

[73] Nejime, Y., and Moore, B. C. J. Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *Journal of the Acoustical Society of America 103*, 1 (1998), 572–576.

[74] Niu, X., and van Santen, J. P. H. A formant-trajectory model and its usage in comparing coarticulatory effects in dysarthric and normal speech. In *Proc. of the Third International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications* (Firenze, Italy, Dec. 2003).

[75] Oja, H. On permutation tests in multiple regression and analysis of covariance analysis problems. *Australian Journal of Statistics 29*, 1 (1987), 91–100.

[76] O'Shaughnessy, D. *Speech Communications: Human and Machine*, 2nd ed. IEEE Press, New York, 2000.

[77] Pavlovic, C. V., and Studebaker, G. A. An evaluation of some assumptions underlying the articulation index. *Journal of the Acoustical Society of America 75* (1984), 1606–1612.

[78] Payton, K. L., Uchanski, R. M., and Braida, L. D. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America 95*, 3 (1994), 1581–1592.

[79] Picheny, M. A., Durlach, N. I., and Braida, L. D. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research 28* (1985), 96–103.

[80] Picheny, M. A., Durlach, N. I., and Braida, L. D. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research 29* (1986), 434–446.

[81] Picheny, M. A., Durlach, N. I., and Braida, L. D. Speaking clearly for the hard of hearing. III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech Hearing Research 32* (1989), 600–603.

[82] Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America 97*, 1 (1995), 593–608.

[83] Pickett, J. M. Effects of vocal force on the intelligibility of speech sounds. *Journal of the Acoustical Society of America 28* (1956), 902–905.

[84] Plomp, R., and Mimpen, A. M. Speech-reception threshold for sentences as a function of age and noise level. *Journal of the Acoustical Society of America 66*, 5 (1979), 1333–1342.

[85] Prakash, B. Acoustic measures in the speech of children with stuttering and normal non-fluency - a key to differential diagnosis. In *Proc. of the workshop on Spken Language Processing* (2003), pp. 49–57.

[86] Rosenberg, A. E. Effect of glottal pulse shape on the quality of natural vowels. *Journal of Acoustical Society of America 49* (1970), 529–538.

[87] Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silberger, H. R., Urbanek, G. E., and Weinstock, M. IEEE Recommended practice for speech quality measurements. *IEEE Transactions on Audio Electroacoustics 17* (1969), 227–246.

[88] S., R. K., and J., V. N. A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *Journal of the Acoustical Society of America 117*, 4 (2005), 2181–2192.

[89] Schum, D. J. Intelligibility of clear and conversational speech of young and elderly talkers. *Journal of the American Academy of Audiology 7* (1996), 212–218.

[90] Sjölander. Recent developments regarding the WaveSurfer speech tool. In *Speech, Music and Hearing Quarterly Progress and Status Reports* (2002), vol. 44, pp. 53–55.

[91] Sjölander, K., and Beskow, J. WaveSurfer — an open source speech tool. In *Proc. of ICSLP* (2000), pp. 464–467.

[92] Skowronski, M. D., and Harris, J. G. Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication 48* (2006), 549–558.

[93] Smits, R., Bosch, L. T., and Collier, R. Eveluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. *Journal of the Acoustical Society of America 100*, 6 (1996), 3852–3864.

[94] Sommers, M. S., and Barcroft, J. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *Journal of the Acoustical Society of America 119*, 4 (2006), 2406–2416.

[95] Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America 96*, 3 (1994), 1314–1324.

[96] Stine, E. L., and Wingfield, A. Process and strategy in memory for speech among younger and older adults. *Psychology and Aging 2*, 3 (1987), 272–279.

[97] Studebaker, G. A. A "rationalized" arcsine transform. *Journal of Speech and Hearing Research 28* (1985), 455–462.

[98] Studebaker, G. A., and Sherbecoe, R. L. Frequency-importance and transfer functions for recorded cid w-22 word lists. *Journal of Speech and Hearing Research 34* (1991), 427–438.

[99] Studebaker, G. A., Sherbecoe, R. L., and Gilmore, C. Frequency-importance and transfer functions for the Auditec of Saint Louis recording of the NU-6 word test. *Journal of Speech and Hearing Research 36* (1993), 799–807.

[100] Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America 84*, 3 (1988), 917–928.

[101] Taylor, P., Black, A., and Caley, R. The Architecture of the Festival Speech Synthesis System. In *Proc. of the Third International Workshop on Speech Synthesis* (Sydney, Australia, Nov. 1998).

[102] Tillman, T. W., and Carhart, R. *An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6*, 2nd ed. Addison Wesley, Reading, Mass, 1981.

[103] Traunmuller, H. Analytical expressions for the tonotopic sensory scale. *Journal of Acoustical Society of America 88*, 1 (1990), 97–100.

[104] Turner, C. W., Smith, S. J., Aldridge, P. L., and Stewart, S. L. Formant transition duration and speech recognition in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America 101*, (5) (1997), 2822–2825.

[105] Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., and Durlach, N. I. Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research 39* (1996), 494–509.

[106] Voiers, W. D. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology 1* (1983), 30–39.

[107] Weismer, G., and Berry, J. Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *Journal of the Acoustical Society of America 113*, 6 (2003), 3362–3378.

[108] Wingfield, A., Aberdeen, J. S., and Stine, A. L. Word onset gating and linguistic context in spoken word recognition by young and elderly adults. *Journal of Gerontology 46*, 3 (1991), 127–129.

Appendix A

IEEE-Harvard sentences

Seventy sentences from IEEE-Harvard sentences [87] were used in experiments in Chapter 3. These sentences were grouped into two according to the informal perceptual experiments. In order to maximize the intelligibility difference between CNV and CLR speech, we identified 48 sentences with larger differences in acoustic characteristics between CNV and CLR speech. Remaining twenty-two sentences were used to set the noise level (SNR-50) prior to the intelligibility experiments.

	1.	The birch canoe slid on the smooth planks.
	2.	Glue the sheet to the dark blue background.
SNR-50	3.	It's easy to tell the depth of a well.
	4.	These days a chicken leg is a rare dish.
	5.	Rice is often served in round bowls.
	6.	The juice of lemons makes fine punch.
	7.	The box was thrown beside the parked truck.
	8.	The hogs were fed chopped corn and garbage.
	9.	Four hours of steady work faced us.
	10.	A large size in stockings is hard to sell.
	11.	The boy was there when the sun rose.
	12.	A rod is used to catch pink salmon.
	13.	The source of the huge river is the clear spring.
	14.	Kick the ball straight and follow through.
SNR-50	15.	Help the woman get back to her feet.
	16.	A pot of tea helps to pass the evening.
SNR-50	17.	Smoky fires lack flame and heat.
	18.	The soft cushion broke the man's fall.
	19.	The salt breeze came across from the sea.
SNR-50	20.	The girl at the booth sold fifty bonds.
	21.	The small pup gnawed a hole in the sock.
	22.	The fish twisted and turned on the bent hook.
	23.	Press the pants and sew a button on the vest.
	24.	The swan dive was far short of perfect.
	25.	The beauty of the view stunned the young boy.
	26.	Two blue fish swam in the tank.
	27.	Her purse was full of useless trash.
	28.	The colt reared and threw the tall rider.
	29.	It snowed, rained, and hailed the same morning.
	30.	Read verse out loud for pleasure.
	31.	Hoist the load to your left shoulder.
	32.	Take the winding path to reach the lake.
	33.	Note closely the size of the gas tank.
	34.	Wipe the grease off his dirty face.
	35.	Mend the coat before you go out.

SNR-50	36.	The wrist was badly strained and hung limp.
	37.	The stray cat gave birth to kittens.
SNR-50	38.	The young girl gave no clear response.
	39.	The meal was cooked before the bell rang.
	40.	What joy there is in living.
	41.	A king ruled the state in the early days.
	42.	The ship was torn apart on the sharp reef.
	43.	Sickness kept him home the third week.
SNR-50	44.	The wide road shimmered in the hot sun.
	45.	The lazy cow lay in the cool grass.
	46.	Lift the square stone over the fence.
	47.	The rope will bind the seven books at once.
	48.	Hop over the fence and plunge in.
	49.	The friendly gang left the drug store.
	50.	Mesh wire keeps chicks inside.
	51.	The frosty air passed through the coat.
	52.	The crooked maze failed to fool the mouse.
	53.	Adding fast leads to wrong sums.
SNR-50	54.	The show was a flop from the very start.
SNR-50	55.	A saw is a tool used for making boards.
SNR-50	56.	The wagon moved on well oiled wheels.
SNR-50	57.	March the soldiers past the next hill.
SNR-50	58.	A cup of sugar makes sweet fudge.
	59.	Place a rosebush near the porch steps.
SNR-50	60.	Both lost their lives in the raging storm.
SNR-50	61.	We talked of the side show in the circus.
SNR-50	62.	Use a pencil to write the first draft.
SNR-50	63.	He ran half way to the hardware store.
SNR-50	64.	The clock struck to mark the third period.
SNR-50	65.	A small creek cut across the field.
SNR-50	66.	Cars and buses stalled in snow drifts.
	67.	The set of china hit the floor with a crash.
SNR-50	68.	This is a grand season for hikes on the road.
SNR-50	69.	The dune rose from the edge of the water.
SNR-50	70.	Those words were the cue for the actor to leave.

Appendix B

Phonetic feature

Table B.1: Phonetic feature values.

Phoneme	voicing	manner	place	height
/iː/	1.0	2.0	4.0	4.0
/ɪ/	1.0	2.0	4.0	3.0
/ɛ/	1.0	2.0	4.0	2.0
/ə/	1.0	2.0	4.0	1.0
/ɨ/	1.0	2.0	5.0	3.0
/ə/	1.0	2.0	5.0	2.0
/ə̹/	2.0	2.0	5.0	2.0
/u/	1.0	2.0	6.0	4.0
/ʉ/	1.0	2.0	5.0	4.0
/ʊ/	1.0	2.0	6.0	3.0
/ʌ/	1.0	2.0	5.2	2.0
/ɔ/	1.0	2.0	6.0	0.5
/ɑ/	1.0	2.0	6.0	1.0
/ɝ/	1.0	3.0	5.0	2.0
/ɚ/	1.0	2.5	5.5	2.0
/ei/	1.0	1.0	4.0	2.5
/aɪ/	1.0	0.0	5.0	2.5
/ɔi/	1.0	0.0	5.0	2.0
/iʊ/	1.0	0.0	5.0	3.5
/aʊ/	1.0	1.0	6.0	2.5
/oʊ/	1.0	2.0	6.0	2.0
/p/	4.0	7.0	1.0	7.0
/t/	4.0	7.0	3.0	7.0
/k/	4.0	7.0	8.0	7.0
/b/	3.5	7.0	1.0	7.0
/d/	3.5	7.0	3.0	7.0
/g/	3.5	7.0	8.0	7.0
/ʧ/	4.0	6.5	3.5	6.0
/ʤ/	3.0	6.5	3.5	6.0
/m/	1.0	4.0	1.0	7.0
/n/	1.0	4.0	3.0	7.0
/ŋ/	1.0	4.0	8.0	7.0
/f/	4.0	6.0	1.0	6.0
/θ/	4.0	6.0	2.0	6.0
/s/	4.0	6.0	3.0	6.0
/ʃ/	4.0	6.0	4.0	6.0
/h/	2.0	2.0	5.0	2.5
/h̹/	1.0	2.0	5.0	2.5
/v/	3.0	6.0	1.0	6.0
/ð/	3.0	6.0	2.0	6.0
/z/	3.0	6.0	3.0	6.0
/ʒ/	3.0	6.0	4.0	6.0
/l/	1.0	3.0	3.0	6.0
/ɻ/	1.0	3.0	5.5	2.0
/j/	1.0	3.0	4.0	4.0
/w/	1.0	3.0	6.0	2.0
/l̥/	1.0	2.5	3.0	6.0
/m̥/	1.0	3.5	1.0	7.0
/n̥/	1.0	3.5	3.0	7.0
/ɾt/	2.5	7.0	3.0	7.0
/ɾd/	2.0	7.0	3.0	7.0
/ɾn/	2.0	4.0	3.0	7.0

Appendix C

Generic tables by Allen et. al.

Table C.1: Generic values (Hz) provided by Allen *et al.* [1]. F4 is given by $F3+1000$ (Hz).

Phoneme	F1	F2	F3	F4
/iː/	300	2045	2960	3960
/ɪ/	435	1700	2585	3585
/ɛ/	575	1605	2515	3515
/æ/	635	1575	2450	3450
/u/	335	1075	2200	3200
/ʊ/	475	1140	2370	3370
/ʌ/	620	1220	2550	3550
/ɑ/	700	1220	2600	3600
/p^{h}/	400	1100	2150	3150
/t^{h}/	400	1600	2600	3600
/k^{h}/	300	1990	2850	3850
/b/	200	1100	2150	3150
/d/	200	1600	2600	3600
/g/	200	1990	2850	3850
/ʧ/	350	1800	2820	3820
/ʤ/	260	1800	2820	3820
/m/	480	1270	2130	3130
/n/	480	1340	2470	3470
/ŋ/	480	1900	2800	3800
/f/	340	1100	2080	3080
/θ/	320	1290	2540	3540
/s/	320	1390	2530	3530
/ʃ/	300	1840	2750	3750
/h/	500	1500	2500	3500
/v/	220	1100	2080	3080
/ð/	270	1290	2540	3540
/z/	240	1390	2530	3530
/l/	310	1050	2880	3880
/ɻ/	310	1060	1380	2380
/j/	260	2070	3020	4020
/w/	290	610	2150	3150

Appendix D

CVC word list

Table D.1: List of 242 CVC words used in Chapters 6 and 7.

Word	Word	Word	Word	Word	Word	Word	Word
shop	leap	fall	lean	loop	van	red	moon
news	bees	tip	juice	use	bar	zoom	dog
hug	gas	whose	chief	dash	ten	pop	shook
room	heat	fun	goose	check	far	leak	sick
bus	loss	shell	log	tar	set	lock	seize
did	feet	pot	rock	dish	reed	leave	gem
pitt	rod	suit	tap	coop	lamb	neck	peas
ban	was	bed	nun	jam	nut	gym	ham
chef	leaf	run	rough	ridge	lid	led	sip
dip	fat	wig	ram	car	mass	mess	cod
rim	rich	shock	cheese	tongue	calm	took	will
mall	ship	kid	cat	loose	badge	hop	long
had	pen	gang	this	rob	wedge	big	job
mesh	lap	cot	tube	pub	bean	hiss	wheel
seem	thumb	fool	kin	beef	watch	mom	wash
bad	kneel	yell	fuss	teach	rear	thing	lack
kiss	well	den	jazz	pack	ring	sit	knit
fuzz	bag	which	food	hedge	wed	jean	gun
thin	gap	lease	teeth	shoot	move	moose	deep
pin	duke	youth	book	patch	hit	yam	lick
dock	hawk	hot	leg	sob	pitch	duck	back
yes	deer	roof	tool	lot	good	knob	pet
sing	such	miss	king	lung	bush	gear	ran
walk	map	vet	seek	have	weep	hen	mad
that	cab	ball	deal	batch	deck	mud	boss
yacht	sad	chalk	beach	sun	could	numb	wish
cool	top	heel	bun	man	mill	young	bomb
mob	dot	fan	fell	get	jog	team	rush
zeal	gel	chip	moss	much	judge	geek	meal
tooth	seal	piece	deaf	nag	says	hall	wet
chat	cop						

Appendix E

Mean estimated consonant target

Table E.1: Mean consonant target values described in Chapter 6 for both speakers.

	Male speech				Female speech				Generic values	
	C_1 target		C_2 target		C_1 target		C_2 target			
Phoneme	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
/p/	163	1283	434	1029	890	1709	476	1324	400	1100
/t/	511	1994	416	1592	342	1730	379	2123	400	1600
/k/	380	2281	358	1217	384	2588	269	2212	300	1990
/b/	192	1082	151	954	153	1255	243	1271	200	1100
/d/	322	1987	317	1686	352	2349	295	2245	200	1600
/ɡ/	204	2140	150	1274	263	2516	150	2030	200	1990
/ʧ/	205	1698	398	1612	423	1746	510	2371	350	1800
/ʤ/	345	1916	150	1643	330	2246	253	2136	260	1800
/m/	458	1083	150	912	346	1333	337	1285	480	1270
/n/	296	1881	185	1797	316	2320	315	2055	480	1340
/ŋ/	–	–	258	2524	–	–	269	2646	480	1900
/f/	415	1234	383	1385	385	1366	490	1692	340	1100
/θ/	158	1199	347	1638	274	2088	273	2072	320	1290
/s/	362	1534	356	1514	364	1750	355	1843	320	1390
/ʃ/	343	1813	398	1482	350	1982	619	2322	300	1840
/h/	583	1804	–	–	745	2484	–	–	500	1500
/v/	150	1402	324	737	151	2620	166	1013	220	1100
/ð/	391	1098	–	–	166	1737	–	–	270	1290
/z/	338	1763	316	1520	433	1948	306	1827	240	1390
/l/	348	898	411	777	338	1183	474	938	310	1050
/ɻ/	340	1025	440	1177	305	1060	447	1371	310	1060
/j/	315	2337	–	–	311	2615	–	–	260	2070
/w/	308	674	–	–	266	608	–	–	290	610

Biographical Note

Akiko Amano-Kusumoto was born on February 15, 1977 in Osaka, Japan. She received her bachelor and master of science degrees in Electrical Engineering from Sophia University in Japan. While she was in her Master's program, she spent one year as a visiting research scientist at the Oregon Graduate Institute of Science and Technology (OGI) in 2000. During her stay at OGI, she worked on a modulation enhancement algorithm to improve speech intelligibility in reverberant environment, which was later published in a *Speech Communication* journal article. After her Master's degree, she worked as a research engineer at the National Center for Rehabilitative Auditory Research (NCRAR), Portland VA Medical Center. After spending two and a half years at NCRAR, she joined the Ph.D. program at Oregon Health & Science University (OHSU) in the Fall of 2004. Her research work includes human speech perception, digital signal processing and speech analysis/synthesis. She also worked as a summer internship fellow at the Oregon Hearing Research Center at OHSU, Oral Dynamics Lab at the University of Toronto, and Human Communication Lab at the University of Toronto Mississauga. She is the author of 5 journal articles, 10 peer-reviewed conference papers, and 13 non-peer-reviewed conference presentations.