High Resolution Voice Transformation

Alexander Blouke Kain
B.A. Computer Science and Mathematics, Rockford College, 1995

A dissertation submitted to the faculty of the
OGI School of Science & Engineering at
Oregon Health & Science University
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Engineering

October 2001

The dissertation "High Resolution Voice Transformation" by Alexander Blouke Kain has been examined and approved by the following Examination Committee:

Jan P. H. van Santen
Professor
Thesis Research Adviser

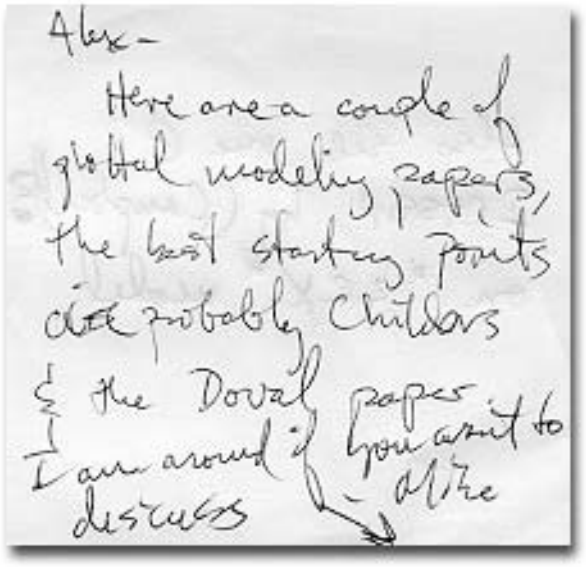
Eric A. Wan
Associate Professor

Xubo Song
Assistant Professor

Dominique Genoud
Nuance Communications

Dedication

This thesis is dedicated to my advisor Michael W. Macon.



Acknowledgements

This work was made possible by the advice, experience, and support of many people. I would like to thank the following faculty members: Jan van Santen, Erik Wan, Xubo Song, Todd Leen, Misha Pavel, Hynek Hermansky, Ron Cole, Jody House, and Ettienne Bernard. Further, I would like to thank the center and departmental staff: Jacques de Villier, Charlene Edayan, Terry Durham, and Hannah Hadfield.

I am very grateful for having had the opportunity to study among my colleagues: John-Paul Hosom, Johan Wouters, Andrew Cronk, Vincent Pagel, Rudolph van der Merwe, Ed Kaiser, Alex Nelson, Brian Mak, Carlos Avendano. Thank you for being there with me.

Finally, I want to acknowledge the never-ending support and infinite love from my parents Maria and Peter Kain, my wife Lauri, and her parents Morley and Kay Blouke.

As with anything of substance, the success of this work is a direct result of the extraordinary community around me.

Contents

List of Tables

List of Figures

Abstract

High Resolution
Voice Transformation

Alexander Blouke Kain

Supervising Professor: Jan P. H. van Santen

Speaker identity, the sound of a person's voice, plays an important role in human communication. With speech systems becoming more and more ubiquitous, Voice Transformation (VT), a technology that modifies a source speaker's speech utterance to sound as if a target speaker had spoken it, offers a number of useful applications. For example, a novice user can adapt a text-to-speech system to speak with a new voice quickly and inexpensively.

In this dissertation, we consider new approaches in both the design and the evaluation of VT techniques. We propose a new type of speech corpus that is especially suited to VT research and development by consisting of naturally time-aligned sentences. Consequently, removal of individual prosodic characteristics, such as fundamental pitch and durations, requires only very little processing and results in high-quality speech samples that only differ in their segmental properties, our focus of transformation. These "prosody-normalized" speech samples are used for training VT systems, as well as for evaluating their transformation performance objectively and subjectively.

Our baseline transformation system (SET) is based on transforming the spectral envelope as represented by the LPC spectrum, using a harmonic sinusoidal model for analysis and synthesis. The transformation function is implemented as a regressive, joint-density

Gaussian mixture model, trained on aligned LSF vectors by an expectation maximization algorithm. We improve upon the baseline by adding a residual prediction module, which predicts target LPC residuals from transformed LPC spectral envelopes, using a classifier and residual codebooks. The resulting high resolution transformation system (HRT) is capable of rendering transformed speech with a high degree of spectral detail.

Because of the severe shortcomings of evaluating VT performance objectively, we propose a subjective evaluation strategy, consisting of several listening tests. In a speaker discrimination test, the HRT system performed significantly better than the SET system. However, discrimination is below that of natural utterances. Similarly, listeners selected the HRT system over other systems in a system comparison test. Finally, listeners rated the speech quality of the HRT system as better than the SET system. However, the quality of natural utterances was considered better than that of transformed speech.

Chapter 1

Introduction

In this dissertation, we consider new approaches in both the design and the evaluation of a newly emerging speech technology called *voice transformation* (VT). The goal of VT is to modify a *source speaker's* speech utterance to sound as if a *target speaker* had spoken it. An effective VT system generates natural, intelligible speech that is clearly identifiable as spoken by the target speaker.

In this chapter, we first motivate the use of VT systems by a number of example applications, followed by a brief description of current voice transformation approaches. We then continue with presenting a summary of our proposed approach and our contributions. Finally, we outline the organization of the dissertation.

1.1 Motivation

The sound of a person's voice, also known as *speaker identity*, plays an important part in our daily communication. For example, speaker identity allows us to recognize family members and friends from their voices alone. Also, speaker identity makes it possible to differentiate between speakers in a conference call or on a radio program. Consequently, there are a number of useful applications for controlling the speaker identity by means of a VT system, especially when integrated into other speech systems with either synthetic or natural speech output.

An example application is the integration of a VT system with a text-to-speech (TTS) synthesizer. Today's state-of-the-art TTS systems are based on a concatenative synthesis method in which a system retrieves natural speech segments from a database and joins

them together to generate a new utterance. The synthesis database contains an organized collection of carefully recorded speech, and the speaker identity of the synthesis output bears resemblance to the original speaker identity of the database speaker. The creation of a synthesis database for a new synthesis voice is a significant recording and labeling effort, and requires a significant amount of computational resources. For example, a speaker may be required to talk in a constrained way for several hours to collect even a relatively small speech inventory of 2,500 diphones. The speech waveforms are stored on disk and processed, typically requiring on the order of hundreds of megabytes and several hours of CPU time. In addition, trained labelers can spend from 10–100 hours for every hour of recorded speech, depending on the complexity of the transcriptions [14].

Using VT technology, new synthesis voices can be created by novice users quickly and inexpensively by creating a "speaker model" from a small number of speech utterances produced by the desired target speaker. The speaker model describes the characteristics of the target speaker's voice. Using different speaker models, the synthesis system can generate speech signals with different speaker identities from a single speaker database (see Figure 1.1), which plays the role of the source speaker [57, 37, 38, 36]. This approach is very well suited for the development of a voice of a speaking-impaired person who is unable to sustain continuous speech or if the speech for a desired target speaker is limited to recordings, such as for a diseased or unavailable speaker. In another application, the speaker model can be in the form of a small attachment to an email message describing the sender's voice characteristics which can then used by a system or service to speak the message in the sender's voice.

Another application is in the area of very-low-bandwidth coding of speech. Speech coding systems that are designed to operate at 2400 bps or less do not preserve speaker identity during transmission [78]. For these systems, VT algorithms have the potential to render the decoded speech at the receiver so that it matches the speaker identity of the transmitting speaker.

Provided a sufficiently high level of VT quality is achieved, movies and TV-shows could be dubbed in the original actors' voices, and language interpreters may assume the voices of their clients [4, 3]. Researchers have also considered a VT system for rendering

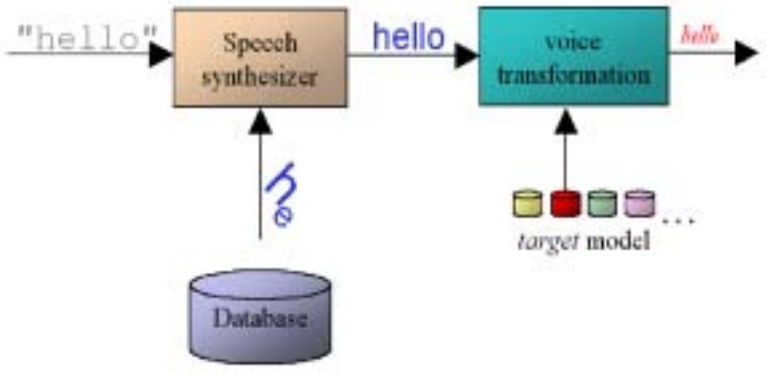


Figure 1.1: A text-to-speech synthesizer in conjunction with a voice transformation system. Fonts are used to represent speaker identity. The synthesizer retrieves chunks of speech from a database, according to an input text. The assembled synthetic speech is input to the voice transformation system, which uses a speaker model to render the final output speech to sound like a desired target speaker.

acoustically impaired speech more intelligible [3, 61, 85].

A logical extension of VT research is the control of a single speaker's voice quality. For example, the assessment and correction of voice quality is desirable in large speech databases for concatenative synthesizers, because the qualitative perception of a speaker's voice often changes over the course of a long recording session [82]. Another application is the modeling of voice quality changes with prosodic factors, such as spectral changes that occur with varying pitch [40].

1.2 Summary of current voice transformation approaches

VT systems modify speech by changing the parameters of an acoustic representation of the speech signal. Often, the acoustic parameters represent a model of the short-term spectrum, such as spectral envelopes and formant estimates. Before use, the VT system must be trained on examples of speech from the source and target speakers. A transformation function captures the relationship between speech model parameters of the source and target speech. Researchers have implemented the transformation function in many ways, including mapping codebooks, class-based functions, artificial neural networks and mixture models. While new research on VT systems is published continuously, transfor-

mation performance is difficult to assess because of several shortcomings in the evaluation process. Most notably, researchers use perceptual listening tests that are not well-suited for evaluating *speaker recognizability*, the recognizability of the transformed speaker as the desired target speaker. Often, these types of listening tests have yielded inconclusive results and included few speaker combinations, raising doubts about the generality of the algorithm. Despite the lack of conclusive test results, the performance of current VT systems falls short of expectations. For example, transformations within the same gender are problematic, due to an insufficiently coarse [2, 37, 61, 64, 91] or ineffective [5, 52] generation of the target speaker's speech spectrum.

1.3 Summary of proposed approach

In this work, we propose a *high resolution* VT algorithm that generates a *detailed* description of the transformed speech spectrum. We claim that it is ineffective and unnecessary to model and transform spectral details between the source and target speaker; instead, we propose to predict spectral details of the target from the transformed spectral envelope of the source. As a result, we show that speaker recognizability improves, as compared to other approaches. Further, we propose an evaluation framework designed for accurately measuring the speaker recognizability of the transformed speech. The framework consists of a custom-designed speech corpus and a series of speaker discrimination tests which include measurements of the natural ability of humans to distinguish and recognize the speakers of the speech corpus.

The following is a summary of our contributions:

- A novel VT algorithm which, for a given target speaker, predicts *LPC residuals* from *LPC spectral envelopes* (these terms are defined in Section 6.1.2), resulting in a high resolution spectral transformation. We show that our high resolution transformation approach leads to a significant increase in speaker recognizability as compared to other approaches.

- A subjective evaluation framework consisting of a special-purpose database, speaker

recognizability listening tests, and a measurement of the natural speaker recognizability of database speakers as a baseline for VT system performance.

1.4 Outline

The remainder of the dissertation is organized as follows:

Chapter 2 introduces some of the fundamental properties of the speech signal. We describe a physical and mathematical model of the speech production process and consider properties of the speech signal that are characteristic of the speaker.

Chapter 3 explains the modes and components of a VT system, followed by descriptions of previous approaches in the literature.

Chapter 4 identifies shortcomings in the previous approaches and presents our thesis and proposed approach.

Chapter 5 describes the design and recording of the special-purpose speech corpus.

Chapter 6 introduces the baseline transformation system. The system transforms the spectral envelope of speech by changing parameters of an all-pole model, using a transformation function implemented by a Gaussian mixture regression model.

Chapter 7 introduces the proposed residual prediction algorithm. After a detailed technical description, we integrate the algorithm into a new VT system and compare the results to several baselines.

Chapter 8 proposes a subjective evaluation methodology designed for measuring the speaker recognizability of the transformed speech signal. After a description of the design and administration of the perceptual test, we analyze and report the results.

Chapter 9 concludes our work and takes a look at the future.

Chapter 2

Basic Properties of the Speech Signal

In order to develop an effective VT system it is important to understand the fundamental properties of speech. The first section provides background on how speech sounds are produced and how their acoustics are modeled mathematically. The next section describes the speaker characteristics of a speech signal. Finally, the last section presents research on the recognition of speakers by humans.

2.1 A model of speech production

Human speech is produced by a part of the human anatomy called the *vocal tract*, which begins at the vocal cords, or *glottis*, and ends at the lips. The compression of the lungs induces a stream of air which flows through the windpipe and throat and escapes through the oral and nasal cavities. This airflow is the source of four types of sounds [88, 67]:

Aspiration noise The sound of air rushing through the entire vocal tract, similar to breathing through the mouth.

Frication noise The sound of turbulent flow at a point of narrow constriction, for example during the initial sound in "fair".

Plosion The sound of an air-burst, for example during the initial consonant in "ton".

Voicing A quasi-periodic vibration of the vocal cords or *glottis*, for example during the vowel in "key". The frequency of vibration is called the *fundamental frequency* or F_0 and is perceived as *pitch*.

The four types of sounds can occur in combination. For example, the initial sound in "vault" combines frication noise with voicing.

The sound-waves from these sound sources are further modified by the vocal tract shape, defined by the location and position of the tongue, jaw, lips, and velum (the soft part of the roof of the mouth). Different vocal tract shapes have different resonant frequencies, called *formants*, which are instrumental in developing the nature of the different speech sounds, called *phonemes*. Phonemes can be classified according to their manner of articulation, namely vowels ("beet"), nasals ("man"), plosives ("pod"), fricatives ("favor"), affricates ("church"), and approximants ("roll") [67].

It is useful to describe the acoustic properties of speech production under the assumptions of the *source-filter model* [56, 74]. In this model, a *source* or *excitation waveform* is input to a *time-varying filter*. This view of speech production is very powerful because it can explain the majority of speech phenomena. In the distinctions of the model, the excitation waveform accounts for the physiological sound sources listed above. For example, aspiration and frication noise can be modeled as random processes, plosion as a step-function, and voicing as a pulse train. A number of *glottal pulse models* have been proposed to describe the details of the pulse shape during voicing [76, 22, 44]. It is possible to classify the excitation waveform into an *unvoiced* and a *voiced* signal, which, in their simplest form, can be modeled as either a random signal or an impulse-train with varying F_0, respectively. Finally, the time-varying filter represents the contribution of the vocal tract shape by selectively attenuating certain frequencies of the excitation spectrum resulting in a speech spectrum with a particular *spectral envelope* and formant structure.

2.2 Speaker characteristics

The acoustic speech signal contains many types of information. Primarily, the signal carries information about the message (*what* was said), but also includes information about the speaker (*who* said it) and the environment (*where* it was said). Speaker characteristics describe the aspects of speech that are related to the person that produced it, independent of the message and the environment. The task of VT is thus to change the speaker

characteristics of a speech signal, while preserving other types of information.

The characteristics of a speaker are commonly divided into the following types of cues:

Segmental cues These describe the "sound" or "timbre" of the speaker's voice. Acoustic descriptors of segmental cues include formant locations and bandwidths, spectral tilt, F_0, and energy. Segmental cues depend mainly on the physiological and physical properties of the speech organs, but also on the speaker's emotional state [44].

Suprasegmental cues These describe the prosodic features related to the style of speaking, for example the duration of phonemes and the evolution of F_0 (intonation) and energy (stress) over an utterance. The average behavior of phoneme duration, F_0, and energy are perceived as *rate of speech*, *average pitch*, and *loudness*. These cues are influenced by social and psychological conditions [48].

Linguistic cues These include particular choices of words, dialects and accents. Linguistic cues are beyond the scope of this dissertation and will not be considered. At the same time, they are significantly reduced by the speaking style contained within the speech corpus of this work (see Section 5.2).

We will illustrate some of the segmental and suprasegmental cues by considering the differences between two different speakers in an example. Figure 2.1 shows the waveforms and spectrograms of a male and a female speaker uttering the sentence "Our plans right now are hazy". Examining the spectrograms, the differences in segmental cues can be observed in the different spectral realizations of the same phonemes. For instance, the formant bandwidths of the female speaker are wider and formant locations higher than that of the male speaker. It is generally assumed that some phonemes carry more speaker information than other phonemes. For example, a phoneme ranking based on automatic speaker verification scores resulted in vowels and nasals in first place, followed by fricatives, affricates and approximants, and plosives [19].

One of the differences in suprasegmental cues are manifested in the different duration lengths of the same individual phoneme groups between the different speakers. For instance, the duration of the initial word "our" is greater for the female speaker than

for the male speaker. Another discrepancy is the insertion of a small pause between the words "now" and "are" in the female example. Finally, an examination of the waveforms reveals a significantly higher F_0 and energy for the female speaker as compared to the male speaker.

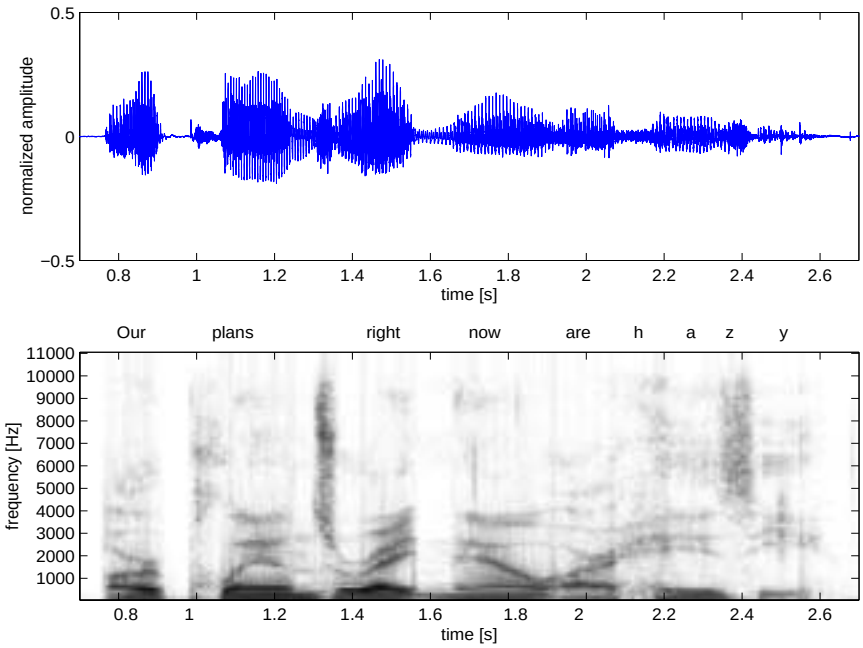
Suprasegmental cues can easily be changed at will. For example, it is easy for a speaker to slow his or her speech, lower the voice, or speak more softly. Segmental cues, however, are closely linked to the physiology of the speech production organs and can thus be considered as immutable. Indeed, impersonators predominantly mimick suprasegmental characteristics [48]. However, some segmental cues can be mimicked by impersonators who are especially skilled in changing some part of their vocal tract physically or in modifying the behavior of their glottal pulse. In this manner, even formant frequencies and bandwidths can be affected.

2.3 Speaker recognition and discrimination by humans

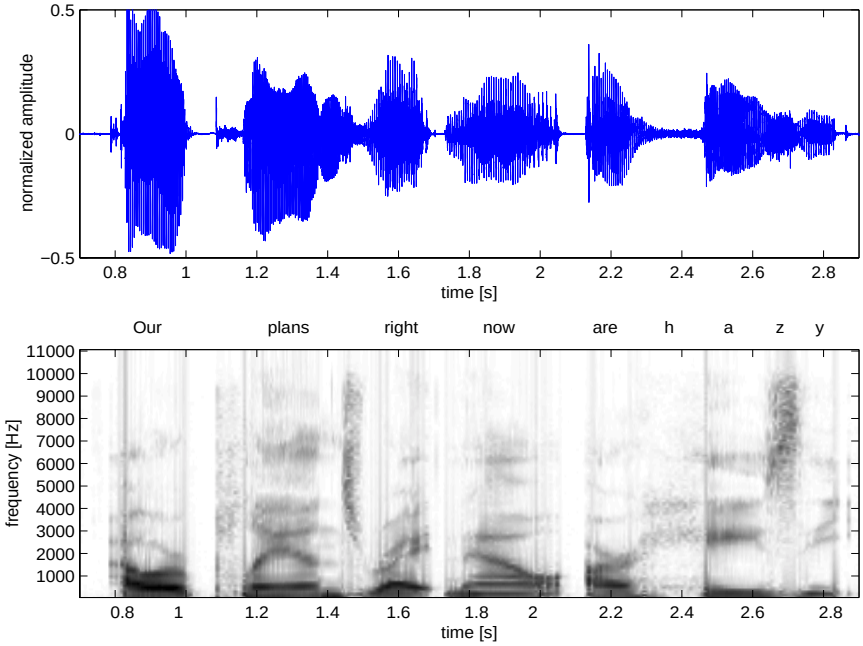
Human listeners are capable of identifying voices under various conditions and contexts with a fairly high degree of accuracy, especially when the voices are familiar to the listener. A perceptual experiment conducted by the Ladefogeds [49] measured the ability of one listener to recognize voices that were familiar to him, from a set that included 29 familiar and 24 unfamiliar voices. The experiment showed that 31% of the 29 familiar voices were correctly identified from the single word "hello", 66% from a single sentence, and only 83% from 30 s of speech.[1] Thus, human recognition is far from perfect, a fact we must consider during the evaluation of a VT system (see Chapter 8).

We now present three types of experiments from the literature that aim to uncover the perceptual significance of various acoustic cues on the identification of speakers: a voice rating test, correlation analysis on the discrimination of speakers, and correlation analysis on the recognition of artificially modified speech signals.

[1] The recognition of a speaker by a witness as evidence in a court of law is controversial [43, 11, 49].



(a) Male speaker.



(b) Female speaker.

Figure 2.1: Speech waveforms and pitch-synchronous magnitude spectrograms of a male and female speaker uttering the sentence "Our plans right now are hazy". Care has been taken to correctly display the pitch-synchronous spectrogram on the linear time-axis.

2.3.1 Voice rating

Voiers [96] classified speakers' voices using a multi-dimensional, perceptual taxonomy based on a set of English words. Starting with 550 potential voice descriptors, a more refined set of 48 bipolar (such as "fast" versus "slow") and 27 mono-polar (such as "scratchy") rating scales were developed by preliminary experimentation. Nine listeners characterized the voices of 80 young adult males after listening to one minute of speech. After a factor analysis using principal factors the author found that eight orthogonal dimensions were required to account for the systematic voice rating variance. He called the leading two dimensions "animation", which was highly correlated with the perceived rate of speech, and "perceived pitch". The last six dimensions had no direct correlation with any single test rating scale, but were correlated to a combination of scales.

2.3.2 Correlation analysis on natural speech

Matsumoto et al. [58] explored the correlation between the perceptual discrimination of speaker identity and the difference in elementary acoustical parameters of five sustained Japanese vowels. To measure speaker discrimination they employed a "same-different" listening test. In this type of test, listeners are presented with two speech samples in sequence and are then asked to indicate whether they thought the sentences were spoken by the same speaker or by two different speakers. Based on the results of this test, a psychological auditory space (PAS) was constructed using a multi-dimensional scaling procedure. The average F_0 was found to explain 55% of the total variance of the PAS. Adding the slope or *tilt* of the glottal source spectrum increased the explained variance by 16%; alternatively, adding the three lowest formant frequencies increased it by 26%. All together, all three acoustic features explained 84% of the total variance. In a related experiment, the authors studied hybrid voices produced by interchanging the approximated glottal source wave and vocal tract spectrum among speakers. The results suggest a relatively greater contribution of the vocal tract over the glottal source characteristics, other than the average F_0, to the ability of humans to discriminate speakers.

Similarly, Necioğlu [65] analyzed the TIMIT continuous speech corpus [25] with a

number of measures relevant to the discrimination of speaker characteristics. The following descriptors were found to have significant correlations with the perceptual dimensions of a multi-dimensional scaling of subjective speaker pair similarity judgments: median pitch, vocal tract length and other vocal tract features for males; median pitch, glottal tilt, and average duration of unvoiced segments for females. These results confirm that average pitch is the most identifying cue in discriminating between speakers, followed by segmental cues.

2.3.3 Correlation analysis on synthetic speech

Itoh et al. [34, 33] studied the effects of modifying acoustical parameters on the identification of speakers that were familiar to the listeners. The authors employed an ABX test, in which listeners are requested to judge whether the speaker of the stimulus labeled "X" was more likely to be speaker "A" or "B". Stimuli were created with a speech analysis-synthesis system capable of producing "hybrid" voices based on the interchange of the linear prediction coefficient (LPC) residual waveforms of two speakers, while keeping their respective LPC spectra constant. (LPC is defined in Section 6.1.2). The authors concluded from the results of the listening test that the LPC spectral envelope has a greater effect on speaker identification than the LPC residual.

In a thorough investigation by van Lancker et al. [92] on the recognition of familiar voices, recordings of famous voices (known to the listeners at the time) were presented in a test where the speech signal was played normally "forward", or "backward" by reversing the signal. Playing the speech forward as compared to playing it backward resulted in some voices being nearly unrecognizable, while others were recognized nearly as well. The authors concluded that people use different acoustic clues for the recognition of different voices and the set of critical parameters is not the same for all voices; instead, listeners select a subset from potential candidates. Further experiments with voices whose rate of speech was modified lead to the same conclusion [93].

A similar conclusion was drawn by Lavner et al. [51] on the identification of familiar voices from a single vowel. Stimuli were created by an analysis-synthesis system that modified parameters of a carefully estimated glottal waveform model, F_0, and formant

frequencies and bandwidths. Listeners were instructed to identify speech samples from a set of speakers. From the results of this test, the authors concluded that vocal tract features are more important to the identification process than glottal source features. Moreover, they found that changes to the same features affected the identification rate of speakers differently, suggesting that different sets of acoustic cues are used for identifying different speakers.

2.4 Summary and conclusion

Human speech is produced by a physiological process involving the lungs, vocal cords, and vocal tract. The resulting speech signal has measurable acoustic properties such as energy, F_0, and formant frequencies. The source-filter model is a simple yet powerful description of speech production. In this model, an excitation waveform, modeling sound sources such as frication noise and vocal cord vibrations during voicing, is input to a linear filter, describing the acoustic effects of the vocal tract shape.

Researchers have shown that both segmental and suprasegmental cues are perceptually significant for speaker recognition. Specifically, among the suprasegmental cues, the average value of F_0 and the rate of speech were found to contribute significantly to speaker recognition. However, it is an open question as to how much the exact behavior of prosodic movements, also known as microprosody, affects speaker recognition. Among the segmental cues, researchers have considered the spectral envelope and formant locations of major importance.

It is probable that the perception of speaker identity depends on *all* acoustic cues with varying degree. A VT approach taking into account a more comprehensive set of acoustic features is likely to outperform approaches with a simpler acoustic feature set.

Chapter 3

An Overview of Voice Transformation Systems

This chapter introduces published works in the area of VT research. The first section introduces the modes and components of a VT system in detail. Section 3.2 summarizes various evaluation methodologies and the obtained results.

3.1 Modes and components of voice transformation systems

There are two basic modes in a VT system:

Training In this mode, the system uses speech samples of a source and target speaker to estimate a transformation function.

Transformation After training has completed, the system transforms the source speaker's voice to sound like the target speaker.

Minimally, a VT system has the following components:

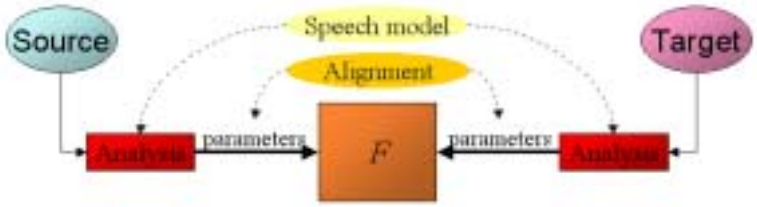
Speech corpus A collection of speech utterances that serve as training data during the training process and as test data during performance evaluations.

Speech model and features The speech model is a mathematical model of the speech signal. The type of model determines which aspects of the speech signal are modifiable by the system. The model parameters, or *features*, are obtained during a speech analysis step, both in training and transformation mode.

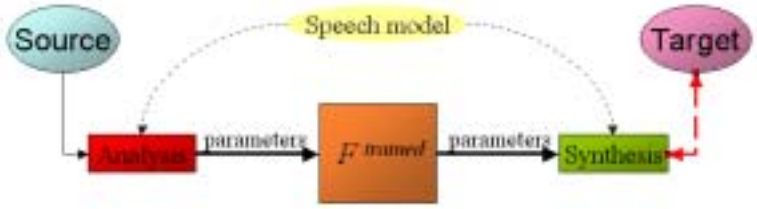
Transformation function The purpose of the transformation function is to map acoustic features of the source speaker to a new set of features that approximate those of the target speaker.

In training mode, the system analyzes speech utterances from a source and target speaker under the assumptions of a particular speech model (see top of Figure 3.1). Commonly used speech models are based on variants of a linear prediction technique, resulting in spectral envelope parameters [2, 37, 85] or formant estimates [61]. Recently, researchers have proposed systems that attempt to go beyond a spectral envelope transformation by modeling and transforming a detailed spectral representation [5, 52]. Every VT algorithm has a speech analysis stage, in which parameters of the speech model are extracted. After the analysis stage, the training process first constructs training data by grouping source and target features corresponding to the same underlying speech sounds. This feature association is typically achieved by applying a time-alignment or classification procedure such as dynamic time warping [2, 85], unsupervised hidden Markov modeling [5], or forced-alignment speech recognition [5]. These training data are then used to estimate a transformation function. The goal of this function is to capture the statistical relationship between the source and target features. The transformation function has been implemented in a variety of ways, including mapping codebooks [2], discrete transformation functions [61, 91], neural networks [64], and Gaussian mixture models [85, 38].

In transformation mode, the "trained" transformation function predicts target speech features from newly analyzed source speech features (see bottom of Figure 3.1). Finally, the predicted features are used to produce the final, transformed speech signal at the synthesis stage. Additionally, prosodic features such as F_0 contour, energy contour, and speaking rate of the source speaker are often trivially adjusted to match the target speaker's average prosody. The reason for not modeling suprasegmental cues *in detail* (for example intonation) is the difficult extraction and manipulation of higher level information (such as pitch-tones [77]) with present speech technologies [48]. While some progress has been made already regarding models of duration [80, 28], models of intonation are, at the mo-



(a) Training.



(b) Transformation.

Figure 3.1: VT system in training and transformation mode. The ovals labeled "source" and "target" represent speech data, the blocks represent system components, and the arrows describe the flow of information. The dashed line in the bottom panel between "synthesis" and "target" illustrates an optional evaluation process, in which transformed speech is compared to the target speech.

ment, notoriously difficult, inaccurate, and controversial [62, 45]. In almost all cases, the acquisition of knowledge about prosody specific to a speaker involves a significant manual effort, which currently makes transforming prosodic details unsuitable for automatic VT systems. For these reasons, we will focus on transforming only the segmental properties of the speech signal.

We now take a closer look at the various components of a VT system and their method of implementation in previously published research.

3.1.1 Speech corpus

The purpose of a speech corpus is to provide the necessary speech data for training the transformation function and for testing the performance of the VT system using objective and subjective evaluation measures. The size and contents of speech corpora in previous approaches vary widely. For example, a speech corpus may contain as little as five vowels [64, 12], a set of words [2, 91], short sentences [7], or one hour of read speech [5]. Alternatively, researchers have also used speech databases created for text-to-speech synthesizers [84, 37]. The optimal size of the speech corpus depends on the application, which may limit the amount of available speech data by design, for instance in the case of adapting a text-to-speech synthesizer to a user's voice [36].

Another aspect of a speech corpus is the number of speakers it contains. Speech corpora in published works have included at least two and at most six different speakers. A larger number of available speakers is advantageous for the evaluation of a VT system, because a larger sample size better represents the general speaker population.

There are many other factors that characterize a speech corpus. We will discuss them in Section 5, where we describe the process of designing and recording a speech corpus for VT system training and testing.

3.1.2 Speech model and features

To manipulate a speech signal by computer, it must be represented as the parameters of a speech model. In general, selecting an appropriate speech model depends on the application. For example, a speech recording and storage system may be designed to

store the signal as a digitized, sampled waveform. In this case, the model is simple with few assumptions on the form of the signal itself; however, the number of model parameters is high. The opposite is true for speech models used in the transmission of speech over low-bandwidth communication channels. In this case, the speech model is complex and contains a number of assumptions about the form or generating process of the signal. However, the number of parameters to be transmitted is low, as required by this application.

In the realm of VT systems, the optimal speech model is capable of producing a wide variety of speech that is intelligible, as well as natural and accurate with respect to speaker recognizability. These goals call for a speech model with many degrees of freedom and hence numerous parameters. At the same time, the transformation function is often trained from a limited amount of training data, favoring a low-dimensional parameter set. Because of these conflicting requirements, a judicious selection of a well-matched speech model and transformation function is essential.

In Section 2.3, we have shown that voice individuality is manifested in all acoustic cues with varying degrees. However, researchers have found evidence that segmental features (in the form of a description of the short-term spectrum) and the average behavior of suprasegmental features (mainly the rate of speech and average F_0) are sufficient for a high degree of speaker discrimination by humans. Moreover, studies in the related field of automatic speaker identification have demonstrated that the spectral envelope alone contains a great deal of information to identify speakers with the help of a computer [24, 23]. Specifically, systems based on a classification of the short-term spectral envelope can identify 16 speakers from a population of 49 speakers with an accuracy of as high as 94.5% using 5 seconds of clean speech [75]. Thus it is not surprising that VT systems have focused on transforming a representation of the short-term spectral envelope, while adjusting a source speaker's F_0, energy, and rate of speech to match those of the target speaker on average. The speech processing in VT systems is usually performed on small sections of speech at a time, also known as *frame-based* processing (see 6.2 for a more detailed description). An interesting exception to frame-based processing is a system by Abe that transforms entire phonetic units [1].

One very successful representation of the short-term spectrum in VT systems is the source-filter speech model (see Section 2.1), which approximates the vocal tract as a slowly varying filter by fitting a spectral envelope to the magnitude spectrum of a short segment of speech. Often, the model parameters are obtained by linear prediction (defined in Section 6.1.2), and the filter coefficients are called linear prediction coefficients (LPC). These parameters are usually converted to a number of alternative representations with more desirable properties, such as the ability to interpolate between parameters. For example, researchers have used cepstral coefficients [85], line spectral frequencies (LSF) [5, 38], and log area ratios [2]. Further analysis of the LPC spectrum can yield formant frequencies and bandwidths, derived either automatically [61] or manually [64].

An LPC residual is obtained by inverse filtering a speech segment with its corresponding LPC filters. Since the filter approximates the vocal tract, an inverse filtering removes the vocal tract contribution of the speech signal. Thus, the LPC residual approximates the glottal excitation waveform. It is possible to utilize the LPC residual of the source speaker without any spectral modifications during synthesis of a transformed utterance [91, 38]. The result is a more natural sounding speech signal. However, the residual also contains a certain degree of speaker information. To address this, several authors have proposed ways of improving VT algorithms by transforming either the original speech spectrum directly or the LPC residual in addition to the LPC spectrum. Because these approaches generate a detailed description of the transformed speech spectrum, or *spectral detail*, we will refer to them as *high resolution voice transformation* approaches. For example, Valbret et al. proposed a dynamic frequency warping (DFW) operating directly on the magnitude spectrum [91]. DFW is a technique that aims at obtaining an optimal, nonlinear warping function of the frequency axis to simulate the changes of speaker characteristics. However, the authors found DFW to be inferior to a more traditional spectral envelope mapping. Arslan et al. [6, 7, 5] formulated a codebook-based transformation of LPC residuals using a weighted combination of "excitation" filters, one for each speech class of a spectral envelope transformation. The "excitation" filters were derived from the average source and target residual spectra within one class. This approach can be viewed as a "two-level" spectral transformation, because both the original speech spectrum en-

velope and the LPC residual spectrum are transformed based on a single classification. This method seems problematic, because the authors found it necessary to incorporate a bandwidth modification method for improving the quality of transformed speech. Finally, in an altogether different approach, a long delay neural network predictor was trained to predict the excitation waveform. During transformation, the network weights were transformed along with spectral envelope parameters [52]. Unfortunately, a formal evaluation of this approach was neglected and thus its performance remains speculative.

3.1.3 Transformation function

The purpose of the transformation function is to capture the differences between source and target features that are due to the difference in speaker characteristics. Naturally, the durations of linguistic units (e.g. phonemes, diphones) differ between speakers, even when producing the same utterances. Therefore, the stream of features stemming from both speakers must be linguistically grouped or aligned in time with each other before training the transformation function. The time-alignment of a feature stream or the grouping of features of similar classes yields the necessary feature vector associations that ensure the preservation of linguistic content. Time-alignment has been implemented by a dynamic time warping (DTW) algorithm [74] in most of the previous approaches [2, 91, 61, 84, 52]. However, it is also possible to use a form of linguistic labeling, as obtained from the states of an unsupervised hidden Markov model (HMM) [7, 5], by forced-alignment speech recognition [36], or by a phonetic classifier [6, 5]. We now present three different methods for implementing a transformation function.

Mapping codebooks

One of the earliest works in the field of VT used a transformation technique called *mapping codebooks* [2]. In this implementation, the codevectors of a source codebook have a one-to-one correspondence to the codevectors of a target codebook. To generate these mapping codebooks, a vector quantization (VQ) algorithm first partitions the source and target feature spaces. Then, a DTW algorithm associates source and target vectors with each other and generates a two-dimensional histogram of their codevector correspondences. The

final target codebook is defined as the linear combination of the target codevectors, using the histogram as a weighting function. A fundamental problem with this technique is the fact that only a discrete set of target features are possible, which results in discontinuities in the speech signal. To overcome the shortcomings of the simple VQ approach, several researchers proposed a technique called weighted-VQ or fuzzy-VQ [48, 5]. This technique expresses the input vector as a combination of the neighboring codevectors, not as the nearest codevector. As a result, discontinuities in the feature stream disappear and the quality of the speech signal improves.

Discrete transformation functions

Several researchers have proposed to use individual transformation functions for each kind, or *class*, of speech sound. Each transformation function is representative of the relationship between source and target features of one class, also referred to as a *local* function. For example, Valbret et al. [91] employed two types of local transformation approaches: linear regression and dynamic frequency warping (DFW). For each class, an algorithm calculated the optimal transformations for both linear regression and DFW during the training process. Similarly, Mizuno et al. [61] calculated a set of linear transformation rules that depended on the input class. Discrete transformation functions are capable of producing an infinite number of target features. However, discontinuities can still occur in the output due to the discrete nature of selecting a single local transformation function.

Continuous transformation functions

An example of a continuous transformation function is an artificial neural network (ANN). It is well known that, theoretically, an ANN with a nonlinear hidden layer can approximate any arbitrary mapping [29, page 142]. Capitalizing on this, Narendranath et al. [64] transformed formant frequencies with the help of an ANN, trained by a back-propagation algorithm. They found that the network generalized properly to unseen data.

Using a *probabilistic* approach, several researchers proposed using Gaussian mixture models (GMM) to describe and map the source and target feature distributions. Stylianou et al. [85] performed a "soft" classification of the source feature space by constructing

a GMM that modeled the source feature distribution. Then, they estimated parameters of a mixture of locally linear transformation functions by solving normal equations for a least-squares problem based on the correspondence between source and target features. They demonstrated empirically that a GMM is more efficient and robust than a VQ-based technique, which is actually a simplified case of a GMM-based approach [41]. In a comparative experiment, the performance of a GMM was found to be as good as or better than other transformation function implementations, specifically approaches involving ANNs, standard VQ, fuzzy VQ, and linear regression [10].

An alternative method of implementing a probabilistic, locally linear transformation function using a GMM was introduced by Kain and Macon [37, 38, 36], drawing on research studying the use of GMMs for regression [26, 41]. In this approach, a GMM is estimated on the joint density of source and target features, and a subsequent regression yields the final transformation function (this approach is described in detail in Section 6.3.2). Modeling the joint density rather than only the source density can lead to a more judicious allocation of mixture components and avoids certain numerical problems when inverting large and possibly poorly conditioned matrices.

3.2 Evaluation and results

Researchers have used many different objective and subjective measures to gauge transformation performance. An objective evaluation can be indicative of transformation performance and is useful in comparing algorithmic alternatives within the same system framework. However, the output of a VT system is a speech signal intended to be heard by humans, and thus perceptual testing is the ultimate performance measure. The following are common objective and subjective VT performance evaluation measures.

3.2.1 Objective evaluation

A commonly used error measure in the field of speech research is the spectral distortion (SD) between two speech signals. In VT research, the average SD is measured between the source, transformed, and target utterances. For example, Abe et al. [2] measured

the ratio of SD between the transformed and target speech and the source and target speech $R = SD\,(transformed, target)\,/\,SD\,(source, target)$. They reported the value of R to range between 0.27 and 0.66, and concluded that the transformed speech was more similar to the target speech than the source speech. Similar ratios have been reported by other researchers [5, 38]. Stylianou et al. [84] used a SD measure to demonstrate that a VQ transformation scheme with 512 codevectors produces a 17% higher average SD than their proposed system with 64 Gaussian components. Finally, Abe showed that a simple mapping codebook was superior to a transformation based on phonetic units, in terms of a SD ratio [1]. However, this result was later contradicted by a perceptual experiment, emphasizing the weak correlation between most objective measures and human perception.

Another avenue is to use transformed speech as input to a speaker identification system and determine the likelihood of the identification of the target speaker. For example, Arslan measured the log-likelihood ratio of target speech to that of the source and transformed speech [5]. In all instances, the ratio increased significantly after the transformation process. However, some speaker combinations were transformed less successfully than others.

3.2.2 Subjective evaluation

The perceptual evaluation of a transformed speech signal has three dimensions of interest: intelligibility, naturalness, and speaker recognizability. An example of a test aimed at measuring speaker recognizability is the ABX test. In this test, participants listen to three stimuli **A**, **B**, and **X**, and are asked to decide whether stimulus **A** or **B** is closer to **X** in terms of speaker identity. **X** is typically the transformed voice, and **A** or **B** the source and target voices. Abe et al. [2] carried out such an ABX test and found that between 57% and 65% of transformed utterances were identified as being closer to the target speaker (12 listeners judging 40 words from 3 male speakers). Kain and Macon [37] researched the application of a VT system in conjunction with a TTS system. Using synthetic sentences, they found that male→female transformations were identified as closer to the target speech 97.5% of the time and those of male→male transformations 52% of the time (20 listeners judging 20 sentences). The latter score of 52% indicates that listeners were guessing, and

indeed interviews after the test revealed that listeners identified the transformed speaker as a third speaker, similar to neither the source nor the target speaker. Similarly, Arslan [5] reported a result of 100% for male→female transformations, and 78% in a male→male transformation (3 listeners judging ten 2–3 word phrases). Stylianou et al. [85] found that a spectral transformation resulted in scores up to 97% (20 listeners judging 3 sentences).

Although widely used, it is important to understand the fundamental flaw of testing speaker recognizability with an ABX test. While a score of 100% indicates that listeners thought the transformed speech was *closer* to the target speaker in terms of speaker identity, the test does not determine whether the transformed speaker is indistinguishable from the target speaker. In actuality, the transformed speech may *not* be recognizable as being spoken by the target speaker.

An improvement over the ABX test is the *pair-comparison* or *similarity test*. In this type of test, participants first listen to a stimulus-pair (of differing linguistic content, for example two different words) and then rate the similarity of the speakers on a rating scale. Using multi-dimensional scaling techniques, results can be projected onto a two-dimensional plane, representing the relative perceptual distances between stimuli. For example, Abe et al. [2] showed that transformed speech is "closest" to the target speech, as compared to partial transformations and the source speech. Stylianou et al. [85] compared statistics of listener ratings on several different types of stimulus-pairs on a scale from zero ("identical") to nine ("very different"). On average, source-source and target-target pairs were rated 0.5 and 1.5. Compared to the target, spectrally transformed stimuli were rated 2.0, and prosodic-only transformations were rated 7.9, almost the same as source-target pairs, which were rated at 8.0 (20 listeners judging 3 sentences). As a result, the authors concluded that purely prosodic modifications of the source speaker's speech did not significantly reduce the perceived dissimilarity between the source and target speaker. However, it is possible that the source and target speakers were already prosodically similar.

In this dissertation, we will focus on the subjective evaluation of speaker recognizability (see Chapter 8). Other important aspects of speech quality include the *naturalness* and the *intelligibility* of the speech signal. For example, Kain and Macon [37] measured the

naturalness of the transformed speech signal by carrying out a mean opinion score (MOS) test [89], a standard test for characterizing the quality of a speech signal with ratings 1 to 5 ("bad", "poor", "fair", "good", and "excellent"). Listeners scored the naturalness of transformed speech signals as 4.2 and 2.7 for a male→male and male→female transformation, respectively. In a second example, Arslan [5] measured the intelligibility of his system by analyzing transcriptions of transformed nonsense sentences. He found that the phone accuracy of the transformed speech was similar to that of the source speaker's speech.

Chapter 4

Thesis and Proposed Approach

In this chapter, we analyze the shortcomings of previous approaches and formulate the problem we address. We then present the thesis of this dissertation, which will be explored in the following chapters.

4.1 Problems of previous approaches

We identify two major shortcomings in the area of VT research: transformation performance and the methods by which this performance is evaluated. We define transformation performance as a measure that combines the degree of intelligibility, naturalness, and speaker recognizability of the transformed speech output. The evaluation of transformation performance incorporates the selection of objective and subjective measures, as well as a suitable speech corpus. We will now describe the problems in these two areas.

4.1.1 Transformation Performance

It is difficult to gauge the success of published VT approaches because of problems in the evaluation of VT system performance (addressed below). However, from the results of Section 3.2, it is clear that the state-of-the art is still short of satisfactory performance. For example, Arslan [5] used a formant bandwidth modification method that post-processes transformed speech in order to cope with bandwidth expansion problems. Stylianou et al. reported that listeners consider transformed speech to be "rather natural" [85, page 141], but sometimes a muffling effect was heard. In a study by Kain and Macon that focused on transformation of TTS utterances [37], listeners judged the naturalness of transformed

speech to be below that of the original synthetic speech. Moreover, listeners in their study reported that the transformed speech sometimes sounded like a third speaker, distinct from source and target speakers, though similar to both of them.

4.1.2 Evaluation

The following are frequently occurring shortcomings in evaluating transformation performance, drawing on information presented in Section 3.2:

- Perceptual listening tests are often not carried out, even though objective measures alone are inadequate for judging perceptual performance.
- Listening tests are informal, or are small-scale either in terms of the quantity or length of presented stimuli, or in terms of the number of listeners.
- Listening tests contain few source-target speaker combinations, due to the small number of available speakers in the researchers' speech corpora. Consequently, it is difficult to judge the generality of test results with respect to a larger speaker population.
- The widely administered ABX test does not adequately test for the recognizability of the transformed speaker.
- The lack of a standard VT speech corpus and standard format for evaluating transformation performance in perceptual listening tests hinders comparisons of results between different approaches.

4.2 Thesis and proposed approach

In this dissertation, we will advance the state of the art in the area of transformation performance and its evaluation. We claim that in order to mimic a speaker precisely and naturally, a VT system must produce transformed speech with a high spectral resolution. Problems with past VT performance can be traced to either the absence or inappropriate modeling of spectral details. Specifically, approaches that rely on transforming the spectral

properties of speech based on modeling the spectral envelope alone are low in spectral resolution and thus less effective (such as the SET system of Chapter 6). Previous VT systems that consider spectral details beyond the spectral envelope attempt to model and transform these spectral details of the source and target speaker. However, it is our hypothesis that *it is ineffective and unnecessary to model or transform spectral details of the source speaker; instead, we propose to predict spectral details of the target spectrum from the transformed spectral envelope.* This is motivated by the realization that, for a particular speaker, spectral details are correlated with speech sounds and can thus be described adequately using a finite number of classes. Chapter 7 describes our new transformation approach in detail.

Further, we propose a new evaluation strategy for measuring transformation performance with a focus on speaker recognizability, described in Chapter 8. As part of the evaluation, we measure the natural ability of humans to distinguish and recognize the speakers of the speech corpus. These measurements serve as a baseline against which a system's transformation performance can be compared.

Finally, we propose a new type of speech corpus, specifically designed for the task of training and evaluating VT systems. This speech corpus, the subject of the next chapter, is the exclusive source of speech data for all experiments described in this dissertation.

Chapter 5

Speech Corpus

A speech corpus, or speech database, is a collection of recorded speech data in the form of an organized hierarchy of waveforms and supporting files. The purpose of a speech corpus is to provide the necessary data for the design, training, and testing of speech systems. For VT systems, the speech corpus must satisfy particular requirements. During training, an adequate amount of data must be available for estimation of a transformation function. During evaluation, a sufficiently large number of sentences and speakers must be available for perceptual testing.

Four major issues concern the designer of a speech corpus for supporting research on VT systems:

Database size This refers to the amount of data that is available for each speaker of the corpus.

Phonetic coverage This measure describes how effectively the speech utterances of a speaker "span" the space of possible speech sounds, such as phonemes or *diphones*, the transition from the center of one phoneme to the center of the next phoneme.

Number of speakers The speakers of the speech corpus are a small sample of the total speaker population. Results obtained by testing a great number of source-target speaker combinations are more indicative of general performance; therefore, a larger pool of speakers is preferable.

Time-alignment During training of a VT system, source and target features of equivalent linguistic character must be associated. A very successful way of providing this

association is the time-alignment of source and target features of sentences with an identical phonetic transcription, that is, the same sentence was spoken by the source and the target speaker.

Our goal is to design and create a speech database that contains a "phonetically rich" set of sentences produced by multiple speakers. The use of identical sentences maximizes the probability of a consistent transcription across speakers, including the effects of vowel reduction and coarticulation. Additionally, the recording procedure was designed to result in a natural time-alignment between identical sentences produces by different speakers. This was achieved by using a "mimicking" approach. The "built-in" time-alignment serves two purposes: On one hand, it allows us to factor out some of the prosodic cues of speaker identity, and, on the other hand, it ensures an accurate time-alignment of the training data with only a minimum of additional signal processing.

5.1 Text material

A speech corpus for VT must have adequate phonetic coverage to effectively describe the different speech sounds of source and target speakers. One way to achieve phonetic coverage is to select the speech material in a specific, careful way. For example, one can manually create a list of words to cover all phonemic sounds, or cover a subset of the most common diphones. Unless the list is small, this process is labor-intensive. Fortunately, automatic procedures for selecting text with desirable features from a larger body of text exist. A widely used algorithm for this task is a *greedy search* [94]. The objective of this algorithm is to find a unit (e.g., word, sentence) in the text corpus that contains the largest number of features (e.g., diphones, phones) that are not yet covered by previously selected units, and then moving this unit from the text corpus to the list of selected units. The algorithm can also use a weighting to indicate preference of certain features. For example, if the weighting equals the inverse frequency of feature occurrence, then the algorithm will select rare units first with regular ones as a by-product, resulting in a shorter list.

In our work, we ran a greedy algorithm on a list of phonetic transcriptions of 1170 sentences, taken from the TIMIT [25] and Harvard Psychoacoustic Sentences [20] databases.

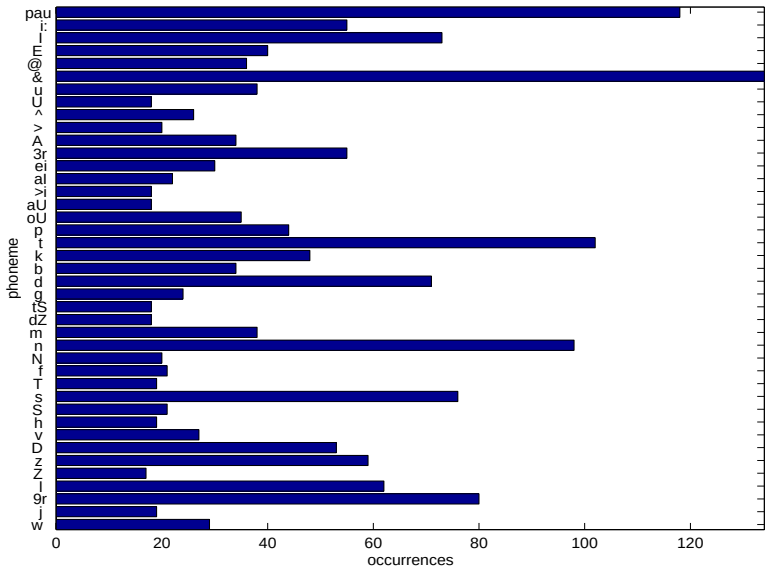


Figure 5.1: Histogram of phoneme content of the final 50 sentences selected by the greedy search algorithm.

Our selection criterion maximized the number of occurrences of rare phonemes, while including as many unique diphones as possible. Targeting one hour of recording time for each speaker, we limited the list of selected sentences to 50 sentences. In the final selection, each phoneme was represented on average 41 and at least 17 times (see Figure 5.1 for a phoneme histogram). The number of unique diphones was 693 out of the 1253 available unique diphones in the phonetic transcription. Every phoneme was included at least once by the third sentence.

5.2 Recording

First, we recorded a *template speaker* reading the 50 selected sentences. For each sentence, the text was displayed on a screen, and 3 "count-in" beeps were played before starting the recording. The template speaker aligned the first stressed syllable of the sentence with the imaginary 4^{th} beat. He spoke slowly and with a flat intonation. After acquiring speech from the template speaker, we recorded 5 male and 5 female *corpus speakers* (see Table 5.1). American English was the first and primary language for the template and corpus speakers.

Identifier	Gender	Age	Origin
F1	female	24	Northwest
F2	female	29	Midwest
F3	female	24	Northwest
F4	female	23	Northwest
F5	female	25	Northwest
M1	male	29	Northwest
M2	male	35	Northeast
M3	male	28	Midwest
M4	male	25	Midwest
M5	male	21	Northwest
template	male	32	Midwest

Table 5.1: Identifier, gender, age, and origin of corpus speakers.

The recording of a corpus speaker was divided into 3 contiguous tasks for each of the 50 sentences. All pertinent information was displayed on a screen, including the current sentence to be read. The tasks were:

Task 1 Speakers were instructed to "read the sentence on the screen naturally". This resulted in recordings that were unconstrained in terms of timing and intonation.

Task 2 The instructions were: "Listen and mouth along quietly with the template sentence, then mimic the sentence on your own." First, the template sentence was played, and immediately afterwards a recording of the speaker was taken.

Task 3 Speakers were instructed to "listen and speak along with the template sentence, then mimic the sentence on your own." Similar to task 2, the template sentence was played, and immediately afterwards a recording of the speaker was taken.

During Task 2 and Task 3, the original three "count-in" beeps were heard before the template speech and three beeps were also heard for the mimick sentences. In this manner, it was easy to begin mouthing or speaking along with the template, as the first stressed syllable fell on the imagined 4^{th} beat. In Task 2, we asked speakers to mouth along quietly to allow for hearing the template sentence clearly, while already providing the opportunity for practice. By Task 3, speakers were able to speak along with the template speaker easily. The mouthing or speaking along speech was not recorded. The reason for recording

two "mimick" sentences is to provide the possibility of establishing a measure of intra-speaker variance (see Section 6.6). Additionally, mimick performance usually improved the second time; therefore, only speech from Task 3 was used for estimating transformation functions. Speakers were told to mimic the timing, stress, and intonation patterns, but not the average pitch or voice quality. During the entire recording, an operator cued each task/sentence and assured satisfactory quality in the areas of recording levels, mimicking timing and intonation, as well as phonetic accuracy.

We recorded speech and laryngograph signals at a sampling frequency of 22 kHz, using a 16 bit encoding. Speakers were located in a professional sound-booth and wore a high-quality headphone/headset with a condenser microphone. The use of a headset ensured a consistent distance to the microphone. Beeps and template speech were played over the closed-type headphones. Additionally, speakers were connected to a laryngograph, which was recorded in parallel for subsequent pitch estimation. The final speech corpus contained approximately 5 minutes of speech for each speaker, excluding pauses.

We included two additional types of information with each speech waveform: *time marks* and *pitch marks*. To create time marks, every speech utterance was force-aligned using the CSLU Speech Toolkit [31, 87]. Time marks were defined as at the beginning of a new HMM state (up to 3 per phoneme). In other words, time marks divided and labeled the recorded speech into linguistic units. This information is necessary for the time-alignment process during transformation function estimation (see Section 6.3.2) or perceptual test stimulus creation (see Section 8.1.1). The second type of information, pitch marks, indicate the instant of glottal closure and allow for a pitch-synchronous analysis of the speech signal (see Section 6.2). We ran an algorithm included in the OGIresLPC package [53] that analyzed the laryngograph signal and created pitch marks. We verified time and pitch marks manually on sentences 41–50, the test set, to ensure a high degree of accuracy. Approximately one to five pitch mark corrections (usually additions) were necessary per sentence.

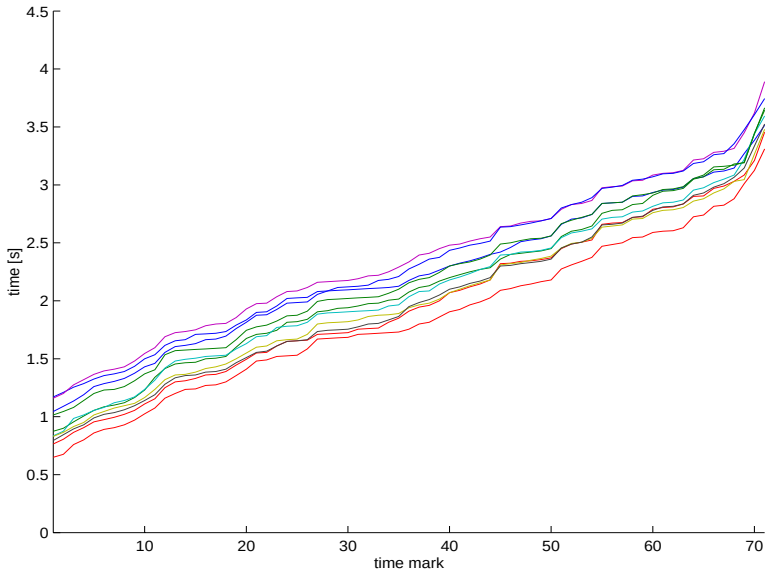


Figure 5.2: Time mark evolutions of a sentence of 10 speakers (solid lines) when mimicking the template speaker (dashed line).

5.3 Mimick performance

We studied the timing and intonation of the corpus speaker sentences in relation to the template speaker sentences to answer the question of how well corpus speakers performed the task of mimicking the template speaker. We measured timing accuracy by examining the evolution of time marks produced by the force-alignment process. Figure 5.2 shows the time evolution of an example sentence of all ten corpus speakers and the template speaker. One can observe that the beginning of the speech utterance is quite variable (the largest difference between speakers is more than a second), but that the lines are mostly parallel, indicating a similar time evolution across speakers. We can take a more informative view by considering the position of the time marks *relative* to those of the template speaker. Such a plot is given in Figure 5.3, in which the beginnings of the utterances have been normalized by a global shift. The average differences between mimicking speakers and the template speaker, computed as an absolute value (i.e. the drifting away from the template speaker in either direction) over all sentences, are speaker-dependent. Figure 5.4 shows the values which range from 54–99 ms. Some speakers can be observed to mimic more accurately than other speakers, on average.Conversely, Figure 5.5 shows the average

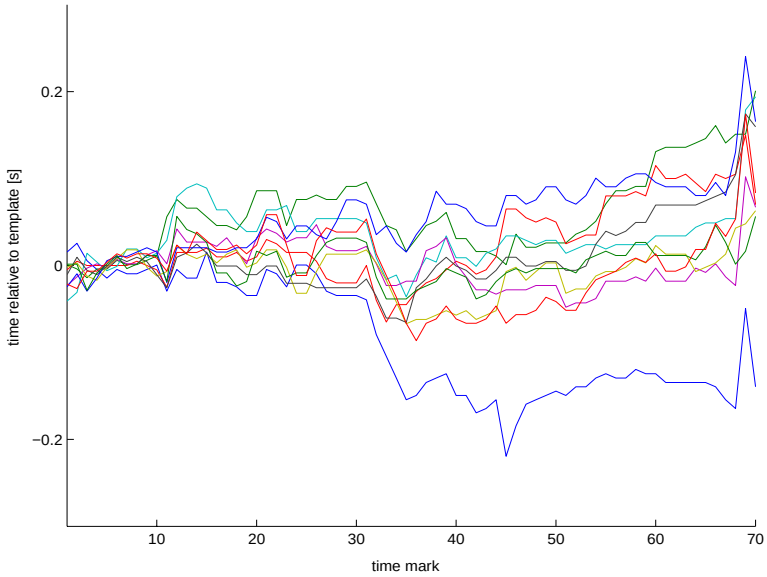


Figure 5.3: Time mark evolutions of 10 speakers relative to the template speakers. A negative value indicates that the speaker is ahead of the template speaker, while a positive value indicates drifting behind. In this sentence, most speakers drifted slightly behind the template speaker. One speaker, however, is ahead by 230 ms at one point.

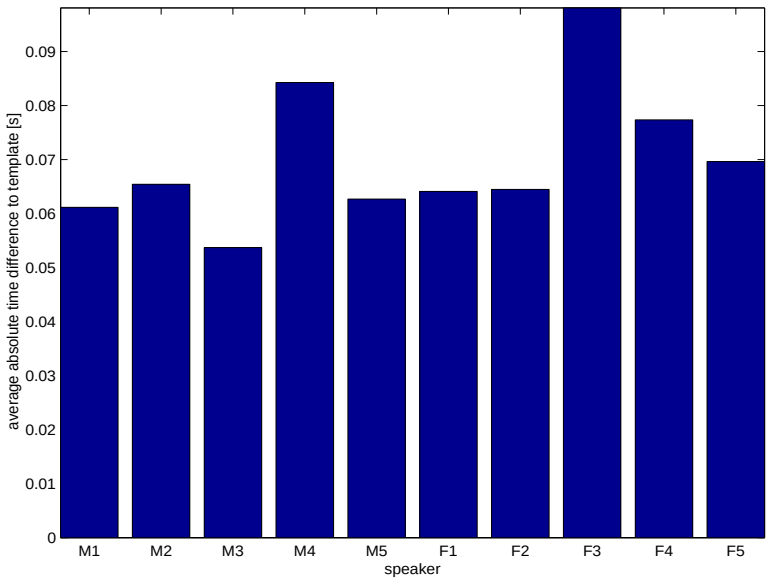


Figure 5.4: Average absolute differences to template speaker grouped by speakers.

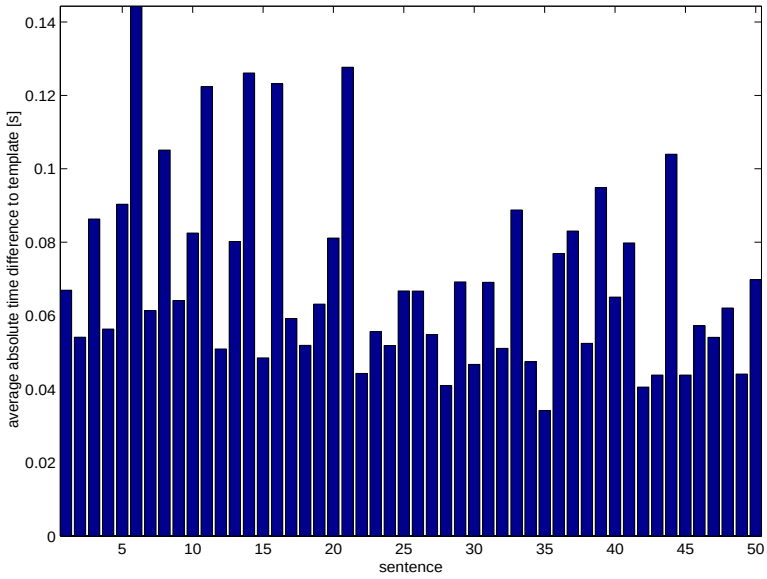


Figure 5.5: Average absolute differences to template speaker grouped by sentences.

absolute differences over all speakers on a sentence basis. A small improvement in timing accuracy towards the end of the recordings can be observed. Overall, the average drift over all speakers and all sentences was 70 ms, less than 2% of the average sentence length.

It is difficult to quantify the accuracy by which intonation has been mimicked. We limit ourselves to presenting Figure 5.6, which shows time-aligned, sentence-final F_0 curves of five male speakers. We observe that, generally, all speakers share the same pitch accents.

We conclude that we were able to achieve a high degree of natural time alignment and a reasonably similar pitch contour across speakers in the database by using a mimicking approach during recording of the speech corpus. In this manner, we minimize the degree of signal processing required for time-alignment tasks during training and testing of the VT systems under study.

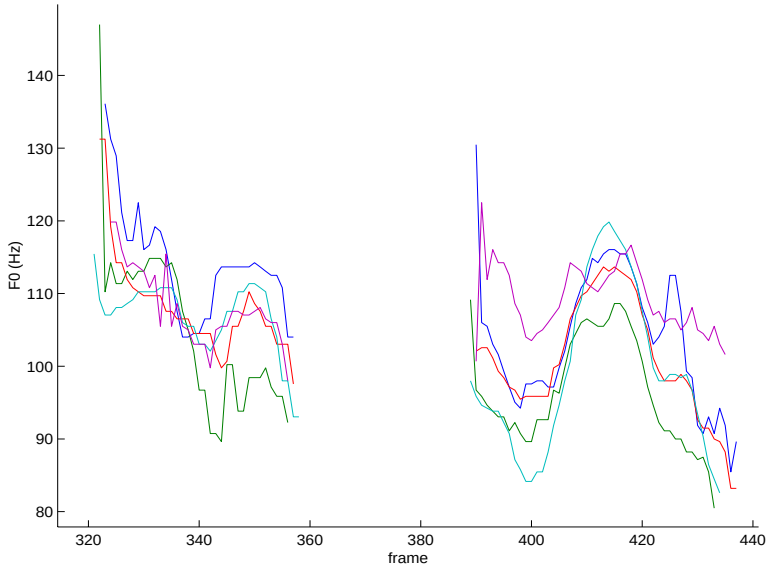


Figure 5.6: Time-aligned, sentence-final F_0 curves of five male speakers. The gap corresponds to an unvoiced segment.

Chapter 6

Transforming the Spectral Envelope

This chapter introduces a VT system which was implemented as representative of the state-of-the-art among spectral envelope transformation algorithms. It serves as a tool for measuring the baseline performance, which can be compared to the performance of innovations in later chapters. The system is designed to transform the spectral envelope of speech by changing parameters of an all-pole model, using a transformation function implemented by a Gaussian mixture regression model. For the sake of clarity in later chapters, we will refer to this system as the SET system.

Figure 6.1 shows an overview block diagram for the SET system in transformation mode. The system encompasses an analysis, transformation, and synthesis stage. During analysis, the system extracts LPC spectral envelope parameters from a sinusoidal representation of the source speech signal. Given the source parameters, a transformation function, trained on data from the source and target speaker, generates a new set of spectral envelope parameters that approximate the target speaker's parameters. Finally, the resulting transformed speech signal is synthesized using a simple impulse/noise excitation approach.

Figure 6.1: Block diagram for the spectral envelope transformation (SET) system.

6.1 Speech model

In this section, we introduce the mathematical representations of the speech signal in the SET system. Specifically, we describe the harmonic sinusoidal model and its coding by linear prediction coefficients, as well as a perceptually motivated warping of the linear frequency scale. Solutions to model equations and feature extraction is deferred to Section 6.2.

6.1.1 Harmonic sinusoidal model

We use a harmonic sinusoidal model [60, 83, 66] to represent the speech signal, a special case of the general sinusoidal model [59, 73]. In this model, a short segment of speech $s(n)$ is approximated as a sum of L sinusoids with amplitudes A and phases ϕ, at integer multiples of the fundamental frequency in radians w_0

$$s_{sin}(n) = \sum_{l=1}^{L} A_l \cos(nlw_0 + \phi_l) \tag{6.1}$$

$$= \sum_{l=-L}^{L} H_{sin}(l)\, e^{j(nlw_0)} \tag{6.2}$$

where

$s_{sin}(n)$	sinusoidal speech
$H_{sin}(l)$	complex amplitude of the l^{th} harmonic
$w_0 = \frac{2\pi F_0}{F_s}$	fundamental frequency in radians
$F_0 = \frac{w_0 F_s}{2\pi}$	fundamental frequency in Hz
F_s	sampling frequency in Hz
L	number of sinusoids, no greater than $\frac{F_s/2}{F_0}$

The motivation behind representing speech in this manner is the observation that voiced speech consists mostly of harmonics of a fundamental frequency. In fact, the sinusoidal model parameters correspond to the harmonic samples of the short-term Fourier-transform of a perfectly periodic signal. Given such a signal, the amplitudes and phases of the sinusoids are given by $A_l = |S(lw_0)|$ and $\phi_l = \angle S(lw_0)$ or, equivalently, $H_{sin}(l) =$

$S(lw_0)$, where

$$S(w) = \sum_{n=-N/2}^{N/2} s(n)\, e^{-jwn} \tag{6.3}$$

In the more common case in which the speech signal is not perfectly periodic, the parameters correspond to the peaks of the averaged spectrogram or *periodogram*. McAulay et al. have shown that the sinusoidal representation is valid even for unvoiced speech under certain assumptions [59]. Even though it is possible to compute parameters of the sinusoidal model via Equation 6.3, we will introduce a different solution algorithm in Section 6.2.

One of the great strengths of the harmonic sinusoidal model is that $s_{sin}(n)$ is perceptually almost indistinguishable from the original speech signal $s(n)$. Equally important, it has been shown to be capable of high-quality speech processing, particularly in the areas of speech coding and pitch and time-scale modifications in speech synthesis [54, 83]. It is further possible to change the magnitude and the phase of the speech spectrum independently and directly by synthesizing speech with altered model parameters. However, a voice transformation of H_{sin} is problematic because its dimensionality is high. For example, the analysis of a short segment of speech with $F_0 = 100\,Hz$ and $F_s = 22,050\,Hz$ requires $L = 110$ complex sinusoids or 220 real-valued parameters.

6.1.2 Sinusoidal parameter coding by a minimum phase all-pole model

By exploiting more speech production properties, an adequate lower-dimensional representation of H_{sin} can be found. In 2.1 we described how a linear filter can model the resonances of the vocal tract by approximating the spectral envelope of speech. If one constrains this filter to be all-pole and minimum phase, its transfer function can be written as

$$H_{lpc}(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{6.4}$$

where a_k are the coefficients for a filter of order p [68]. The filter coefficients are called *linear prediction coefficients* (LPC) [56, 74]. We will optimize the filter coefficients such that

$$H_{lpc}\left(e^{j\left(W^{-1}(lw_0)\right)}\right) \cong H_{sin}(l) \tag{6.5}$$

where $W^{-1}(\cdot)$ is related to an inverse warping function described below. The spectrum of the *LPC residual* is

$$H_{res} = \frac{H_{sin}}{H_{lpc}} \tag{6.6}$$

Equation 6.5 demonstrates the advantage of fitting a LPC model in the frequency-domain: Using a warping function, we are effectively able to vary the LPC modeling power selectively within frequency regions of the spectrum.

The acoustical basis of a minimum phase, all-pole filter is a simple physical model of the vocal tract. If one approximates the vocal tract by a sequence of rigid tubes, each of constant diameter, and neglects losses due to surface viscosity and other effects, then the transfer function of such a lossless tube model results in an all-pole transfer function [74]. The advantages of such a filter are its relatively low number of parameters (typically between 10 and 20) and their well-studied estimation using linear prediction techniques. Despite the simplifying assumptions of this highly constrained model, the quality of speech processing using a LPC model can be high [60, 97].

6.1.3 Spectral warping

An improvement of the perceptual quality of a speech analysis/synthesis system employing minimum phase, all-pole filters for coding sinusoidal parameters can be achieved by considering the non-uniform frequency sensitivity of the human ear. Particularly, the frequency resolution of the ear has been shown to be greater at low frequencies than at high frequencies [63]. One possible scale that approximates this property analytically is the bark scale. We can describe the relationship between the perceptual scale f' and the frequency scale f as $f' = b(f)$, where b represents the bark-scale warping. We use

$$f' = b(f) = 6 \cdot \log\left(\frac{f}{1200} + \sqrt{\left(\frac{f}{1200}\right)^2 + 1}\right) \tag{6.7}$$

and

$$f = b^{-1}(f') = 600 \cdot \left(e^{\frac{f'}{6}} - \frac{1}{e^{\frac{f'}{6}}}\right) \tag{6.8}$$

where the units of f are Hz, and the units of f' are bark. Figure 6.2 shows a frequency conversion graph between the linear and bark scale.

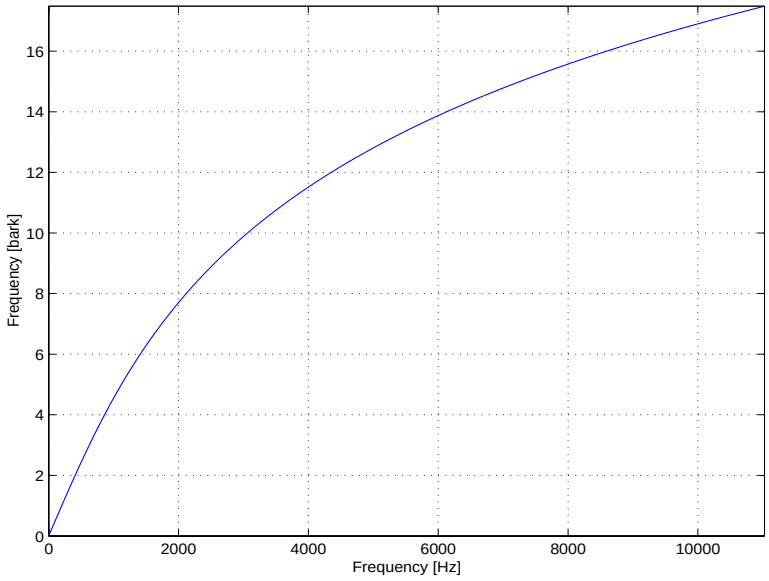


Figure 6.2: Frequency conversion between linear and bark scale

An all-pole filter that is fitted to a warped representation of the spectrum will model details of the spectral shape more closely in the lower frequencies, with a loss of detail at higher frequencies. This approach has been successfully used to reduce the dimensionality of model parameters in a speech coding and a spectral modification task [60, 97].

6.2 Analysis

This section describes how the speech features, in this case parameters of a spectral envelope model, are calculated from the speech signal. The speech signals are taken from the speech corpus described in 5. Each waveform is sampled at 22 kHz, with 16 bit resolution, and contains pitch marks (see Figure 6.3) and other additional information.

We perform the analysis, processing, and synthesis of speech by considering a small section of speech at a time. Therefore, the original speech waveform is apportioned into small, overlapping *frames* $s^m(n)$, thus the system is said to be *frame-based*. This operation is performed synchronously with F_0 (also called *pitch-synchronously*). Each frame contains the speech signal of two pitch periods, centered around the current pitch mark. In unvoiced sections of speech a constant frame-rate of 125 Hz is used in place of pitch mark and F_0

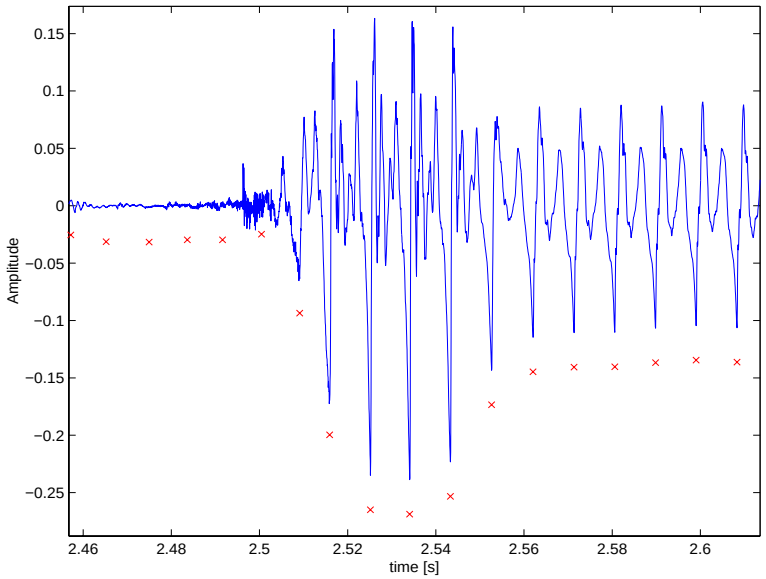


Figure 6.3: Speech waveform, transitioning from an unvoiced pause to a voiced segment (the word "the"). Pitch marks are indicated by the symbols "x" below the waveform in the voiced segment on the right. In the unvoiced segment on the left, the symbols display a constant frame-rate of 125 Hz. Pitch marks represent the instant of glottal closure and are used to define a pitch period.

information.

The sinusoidal parameters for frame m are calculated by minimizing a weighted time-domain least-squares criterion [50, 86, 83]

$$E_{sin} = \sum_{n=-T_0^{m-1}}^{T_0^m} w_a^2(n)\left(s^m(n) - s_{sin}^m(n)\right)^2 \tag{6.9}$$

where, dropping the frame notation,

$s(n)$	original speech frame
$s_{sin}(n)$	sinuisoidal speech frame from Equation 6.1
$w_a(n)$	analysis window
$T_0 = \frac{1}{F_0}$	instantanous fundamental period

We use a complex regression to minimize E_{sin} and obtain L sinusoidal spectrum parameters $H_{sin}(l)$. The number of complex sinusoids differs from frame to frame, since L is limited to $\frac{F_s/2}{F_0}$. The advantage of this approach versus a sampling of a periodogram (derived by Equation 6.3 or similar) is the concentration of minimum error on the center

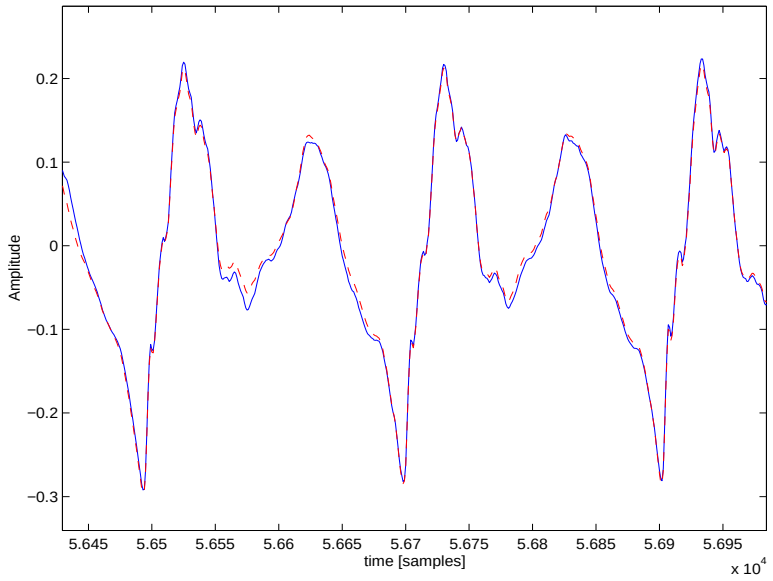


Figure 6.4: Speech waveform (solid line) and resynthesized sinusoidal waveform (dashed line).

of the speech frame, which is the beginning of a new pitch period. Since the frames are overlapping, errors at the beginning or at the end of the speech frame do not play a significant role. Further, the estimation is performed on a time-domain signal, which allows for the short time-frame of two pitch-periods, as compared to Fourier-transform-based methods which would necessitate longer frames with a constant F_0.

Figure 6.4 shows an example region of the original speech waveform and a resynthesized sinusoidal waveform. The differences between the two waveforms are due to the fact that Equation 6.9 involves two similar, but not identical, pitch periods of the speech signal.

Next, we resample the magnitude spectrum $|H_{sin}(l)|$ non-uniformly in accordance to the bark scale warping of the original frequencies, using cubic interpolation [90] (see Figure 6.5). Then, we compute the LPC filter coefficients a_k by an application of the Levinson-Durbin algorithm on the autocorrelation sequence of the warped power spectrum $|H_{sin}(W(l))|^2$[60], where

$$W(l) = L \cdot \frac{b(l \cdot F_0)}{b(L \cdot F_0)} \tag{6.10}$$

The model fit is displayed in Figures 6.6 and 6.7, compared to the original warped and unwarped spectra respectively.

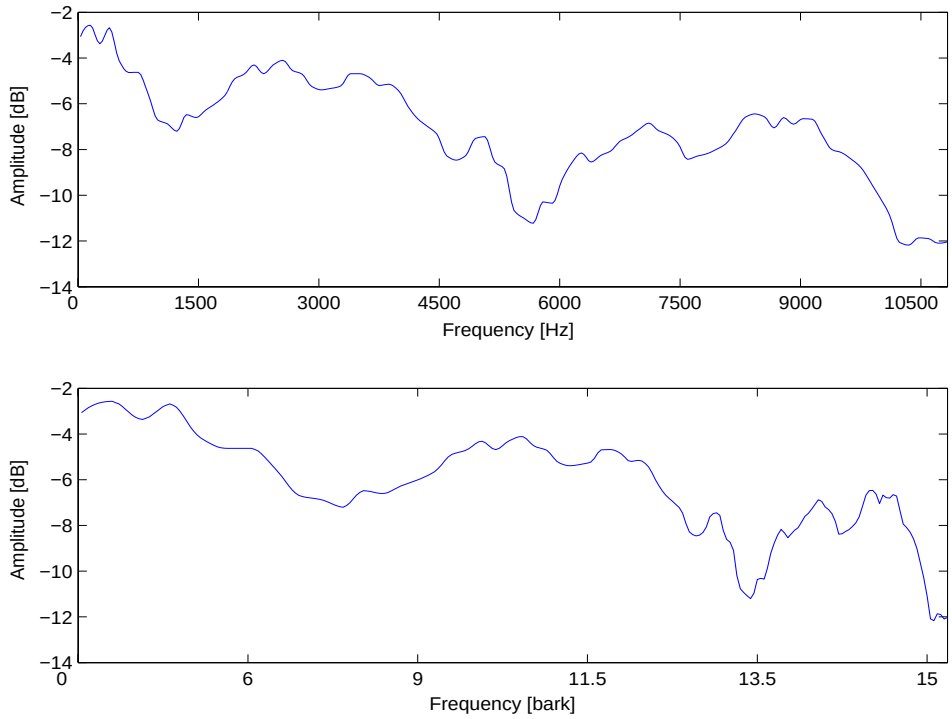


Figure 6.5: Sinusoidal magnitude spectrum before (top panel) and after bark warping (bottom panel).

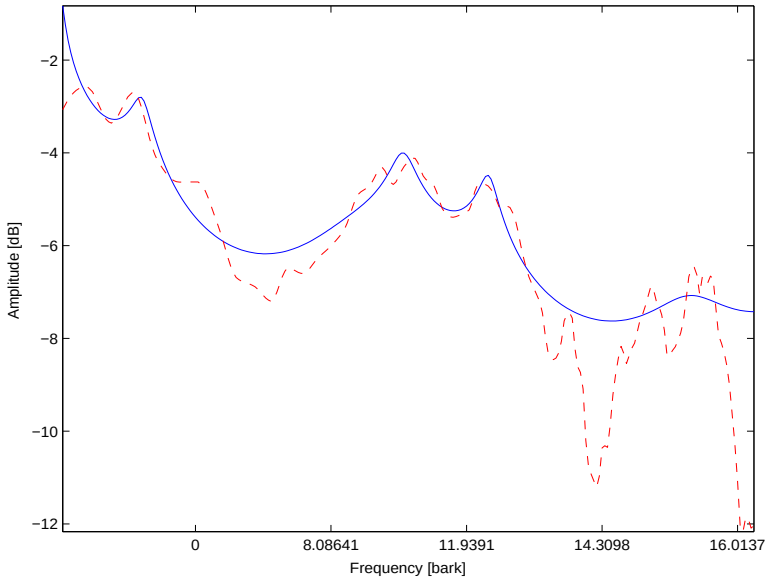


Figure 6.6: All-pole model fit (solid line) to the discrete, warped magnitude spectrum (dashed line).

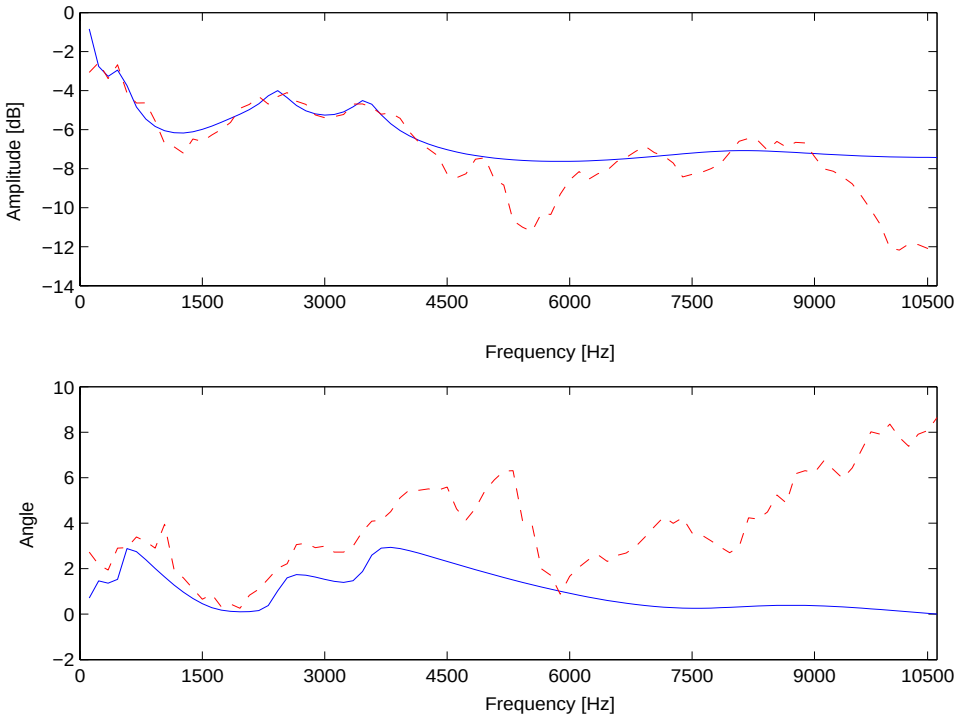


Figure 6.7: All-pole model fit (solid line) to the discrete, linear spectrum (dashed line). The magnitude (top panel) and unwrapped phase (bottom panel) spectra are shown.

Finally, the filter coefficients a_k of $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ are converted to line spectral frequencies (LSF) [32]. In this alternative representation

$$A(z) = \frac{1}{2}\left(P(z) + Q(z)\right) \tag{6.11}$$

where

$$P(z) = A(z) - z^{-(p+1)} A\left(z^{-1}\right) \tag{6.12}$$

$$Q(z) = A(z) + z^{-(p+1)} A\left(z^{-1}\right) \tag{6.13}$$

The LSF parameters are the complex zeros, or roots, of polynomials $P(z)$ and $Q(z)$. Two important properties of LSF are that all zeros of $P(z)$ and $Q(z)$ are on the unit circle and that their zeros are interlaced with each other. Consequently, $A(z)$ can be expressed by a sorted and interleaved list of angles L of the complex zeros of $P(z)$ and $Q(z)$. The advantage of converting to L is the fact that all p angle values can easily be modified, while the stability and minimum-phase property of the filter is guaranteed [81]. For this and other reasons, LSF parameters have been shown to posses superior interpolation properties compared to other linear prediction representations [69]. Good interpolation properties are critical for our transformation function (see 6.3.2), which uses a weighted sum of linear transformations to approximate target features seen during training. LSF are used extensively in speech coding [42, 15, 70] and speech compression [81].

Figure 6.8 shows the evolution of LSF parameters over an example sentence. A comparison with the LPC spectrogram reveals a degree of similarity between LPC pole movements and LSF trajectories. For instance, two closely spaced LSF parameters correspond to a spectral peak with a narrow bandwidth, which is likely to be a formant.

6.3 Training

The purpose of the training stage is to estimate parameters of a transformation function so that it can predict target speaker features Y from source speaker features X. In our frame-based system, the features of one frame describe only a small portion of speech and thus a sequence of features, or *feature stream*, represents an entire utterance. Because

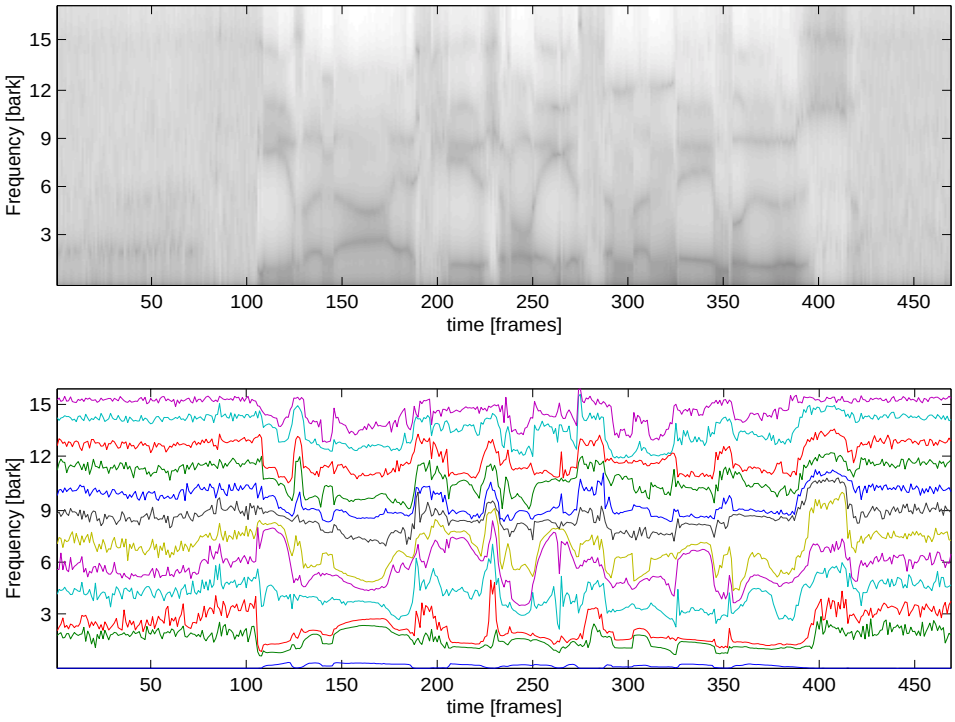


Figure 6.8: Bark-warped LPC spectrogram (top panel) and corresponding LSF trajectories (bottom panel).

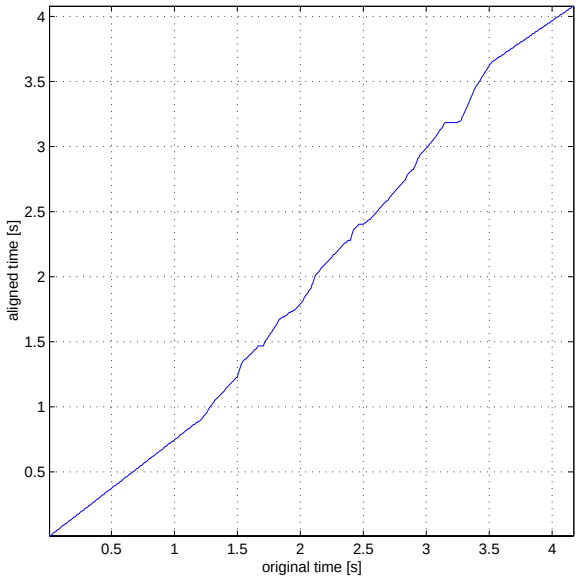


Figure 6.9: Example of an alignment path.

of natural variations in the durations of linguistic units between different speakers, the feature streams X and Y must be time-aligned, allowing the transformation function to learn the relationship between source and target features of equal phonetic content.

6.3.1 Time-alignment

Time-alignment is performed for each sentence of each source/target speaker pair. The goal of time-alignment is to modify the source and target speech feature stream in such a way that the resulting feature streams can be thought of as describing the same phonetic content frame by frame. We achieve alignment by selectively deleting or repeating frames from the target speaker feature stream to match the number of source frames within phonetically equivalent regions defined by two time marks, as defined in Section 5. Alternatively, we can avoid deleting any frames altogether by stretching the shorter region of one speaker to the length of the longer one of the other speaker. In practice, the choice of alignment policy was unimportant, due to the already highly aligned sentences of the speech corpus. An example alignment path is shown in Figure 6.9.

After alignment, we collect aligned LSF feature vectors into N frames of source data

$$X_{p\times N} = \left[L^1_{source}\, L^2_{source}\, L^3_{source}\, \ldots\, L^N_{source}\right] \tag{6.14}$$

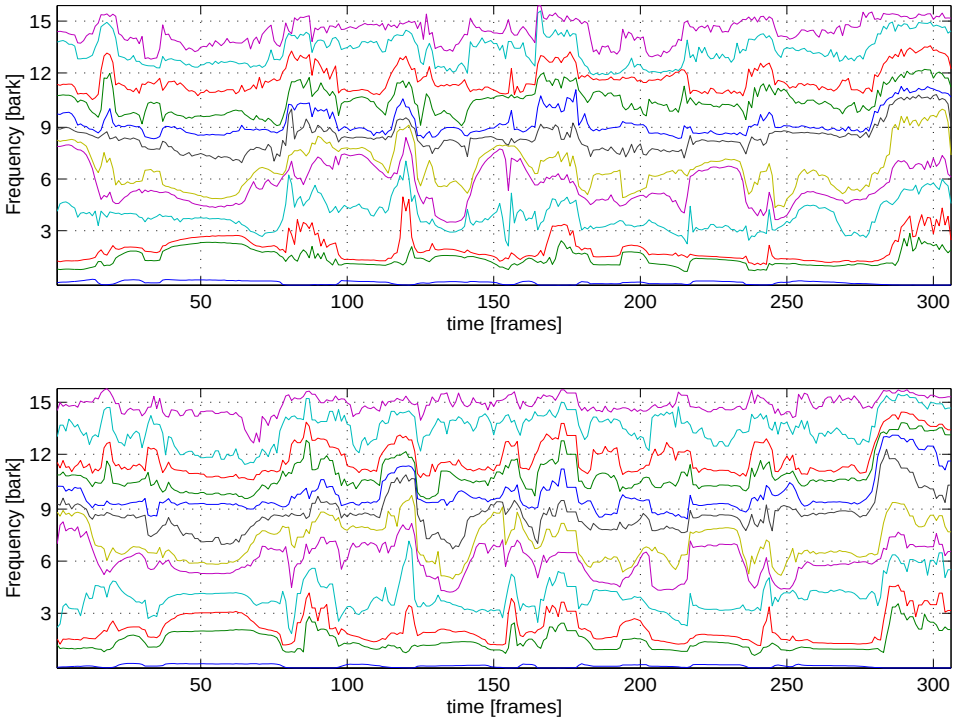


Figure 6.10: Two aligned LSF feature streams.

and, respectively, target data

$$Y_{p\times N} = \left[L^1_{target}\, L^2_{target}\, L^3_{target}\, \ldots\, L^N_{target}\right] \tag{6.15}$$

Beginning and ending silences are not included in the training data sets. An example of a single sentence of two aligned LSF feature streams is shown in Figure 6.10.

The value of N depends on the amount of available training data from both source and target speaker. In our experiments, $N > 10,000$.

6.3.2 Estimation of the transformation function

The purpose of the transformation function is to map the source speaker's speech features X to an estimate of the corresponding target speaker's speech features Y. We choose a Gaussian mixture model (GMM) approach to implement a locally linear, probabilistic transformation function, similar to Stylianou et al.[85] The advantage of using a mixture of locally simple models, such as Gaussians in a GMM, is the fast and accurate fitting of the few model parameters, as compared to linear and non-linear global models (for

example, principal components analysis and neural networks) [41]. Further, a GMM is particularly well suited to the task of VT [10] and is also used successfully in the related field of speaker identification [75].

In contrast to previous VT approaches utilizing a GMM, we obtain the desired transformation function by estimating the joint density $P(X, Y)$ and then evaluating $\hat{Y} = E[Y|X]$, the *expectation* of Y given X [26]. This approach uses the estimate of the data density to cope with incomplete data, and is thus especially well-suited to the task of VT, as it is likely that the acoustical feature spaces of the source and target speaker have been sampled sparsely. We now introduce a GMM formally and show how it can be used in the role of a transformation function.

Gaussian mixture model (GMM)

A mixture model allows the probability distribution of x to be modeled as the sum or *mixture* of Q *components*, also referred to as classes. In the case of a Gaussian mixture mode, the components are normal distributions (Gaussians), and the probability distribution is given by

$$P_{GMM}(x; \alpha, \mu, \Sigma) = \sum_{q=1}^{Q} \alpha_q N(x; \mu_q, \Sigma_q), \quad \sum_{q=1}^{Q} \alpha_q = 1, \quad \alpha_q \geq 0 \tag{6.16}$$

where α_q denotes the prior probabilities of x having been generated by component q, and $N(x; \mu, \Sigma)$ denotes the n-dimensional normal distribution with mean vector μ and covariance matrix Σ given by

$$N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{6.17}$$

Contrary to classification schemes with "hard" class boundaries, data points have varying degrees of "membership" to all local models; this is referred to as "soft" partitioning. The conditional probability of a GMM class q given x is derived by direct application of Bayes' rule

$$p(c_q|x) = \frac{\alpha_q N(x; \mu_q, \Sigma_q)}{\sum_{p=1}^{Q} \alpha_p N(x; \mu_p, \Sigma_p)} \tag{6.18}$$

The GMM parameters $\{\alpha, \mu, \Sigma\}$ are estimated by application of an Expectation-Maximization (EM) algorithm [75], an iterative method for computing the maximum

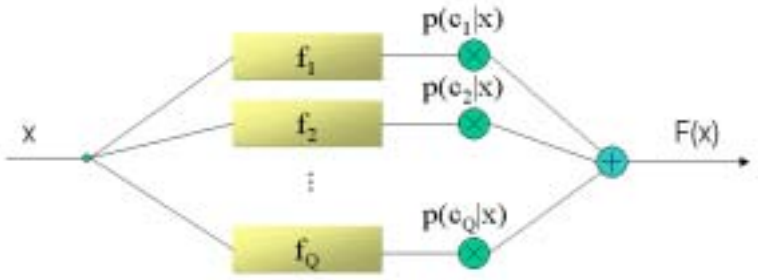


Figure 6.11: Illustration of transformation function implemented by a mixture of locally linear transformation, weighted by a posterior probability.

likelihood parameter estimates. Initially, we set $\alpha = 1/Q$, μ equal to the Q codevectors generated by a vector quantization algorithm, and covariances Σ equal to the identity matrix. Then, the EM algorithm is run until the likelihood $P_{GMM}\left(x; \alpha, \mu, \Sigma\right)$ is maximized or the number of iterations exceeds a threshold of ten iterations, at which point we have observed the likelihood to increase only marginally.

During the EM procedure, it is important to guard against any of the covariance matrices becoming close to singular. We avoid singularities by adding a constant diagonal matrix $\epsilon \cdot I$ to the covariance matrices after each iteration, effectively constraining the volume of the covariance matrices to a lower bound. This technique also regularizes the mixture density. We use an empirically determined value $\epsilon = 0.001$.

Gaussian mixture model for regression

The goal of regression analysis is to predict output data from given input data. We fit a GMM to the joint probability density of inputs and outputs and estimate the parameters of a regression function [26]. The regression is formulated as a weighted sum of linear models, where the weights correspond to the posterior probability of a given input belonging to a particular class, described by

$$\widehat{y} = F\left(x\right) = \sum_{q=1}^{Q} \left(W_q x + b_q\right) \cdot p\left(c_q|x\right) \tag{6.19}$$

where W_q is a transformation matrix and b_q a bias vector of class q. Figure 6.11 is an illustration of Equation 6.19.

To estimate parameters of the transformation function F, we begin by estimating a GMM of the joint density of source *and* target features $p(Z) = p(X, Y)$ where

$$Z_{2p\times N} = \begin{bmatrix} X_{p\times N} \\ Y_{p\times N} \end{bmatrix} \tag{6.20}$$

and X and Y are the aligned source and target features streams of length N and dimension p. The expected value of a feature vector y given feature vector x is the regression

$$F(x) = E[y|x] = \int y \cdot p(y|x)\, dy \tag{6.21}$$

$$= \sum_{q=1}^{Q} \left(\mu_q^Y + \Sigma_q^{YX}\left(\Sigma_q^{XX}\right)^{-1}\left(x - \mu_q^X\right)\right) \cdot p(c_q|x) \tag{6.22}$$

where

$$\Sigma_q = \begin{bmatrix} \Sigma_q^{XX} & \Sigma_q^{XY} \\ \Sigma_q^{YX} & \Sigma_q^{YY} \end{bmatrix} \tag{6.23}$$

$$\mu_q = \begin{bmatrix} \mu_q^X \\ \mu_q^Y \end{bmatrix} \tag{6.24}$$

and Equation 6.18 becomes

$$p(c_q|x) = \frac{\alpha_q N\left(x; \mu_q^X, \Sigma_q^{XX}\right)}{\sum_{p=1}^{Q} \alpha_p N\left(x; \mu_p^X, \Sigma_p^{XX}\right)} \tag{6.25}$$

as shown in Kambhatla's work on Gaussian mixture models for statistical data processing [41]. From this, it follows that

$$W_q = \Sigma_q^{YX}\left(\Sigma_q^{XX}\right)^{-1} \tag{6.26}$$

and

$$b_q = \mu_q^Y - \Sigma_q^{YX}\left(\Sigma_q^{XX}\right)^{-1}\mu_q^X \tag{6.27}$$

The joint density (JD) approach estimates mixture components based on observations of both the source and target feature vectors, which may lead to a more judicious allocation of the component Gaussians than is possible by a clustering of the source feature vectors alone [37].

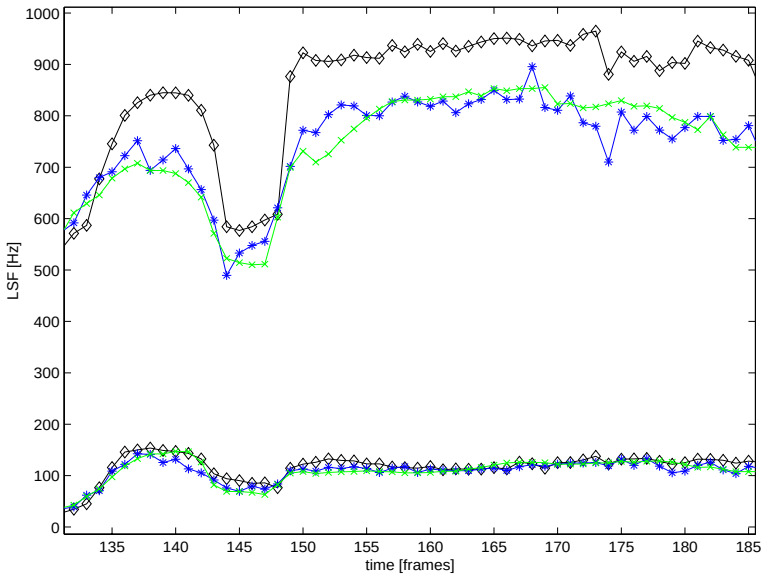


Figure 6.12: Trajectories of the first and second LSF component of the source (diamond symbols), transformed (star symbols), and target speaker ("X" symbols).

6.4 Transformation

In transformation mode, the system analyzes a source speech file and transforms the extracted features X to $\widehat{Y}$, an estimate of the target speaker's LSF parameters. Figure 6.12 shows an example of source, transformed, and target values of a single LSF component.

For each frame, we calculate the transformed spectral envelope by converting the predicted LSF parameters $\widehat{L}$ back to LPC filter coefficients via Equation 6.11 and evaluate the resulting LPC system function $\widehat{H}_{lpc}\left(e^{j\left(W^{-1}(w)\right)}\right) = 1/\widehat{A}\left(e^{j\left(W^{-1}(w)\right)}\right)$ non-uniformly on the unit circle. Figure 6.13 shows the unwarped, LPC log-magnitude spectrograms of the source, transformed, and target speakers. Finally, we set the transformed sinusoidal spectrum

$$\widehat{H}_{sin}\left(l\right) = \widehat{H}_{lpc}\left(e^{j\left(W^{-1}(lw_0)\right)}\right) \tag{6.28}$$

the inverse warped LPC spectrum sampled at the harmonics. If the current frame is unvoiced, then we additionally replace the phase $\angle\widehat{H}_{sin}$ by a random phase vector. This heuristic is equivalent to using an impulse train for voiced, and white noise for unvoiced segments to excite an LPC filter in the time-domain.

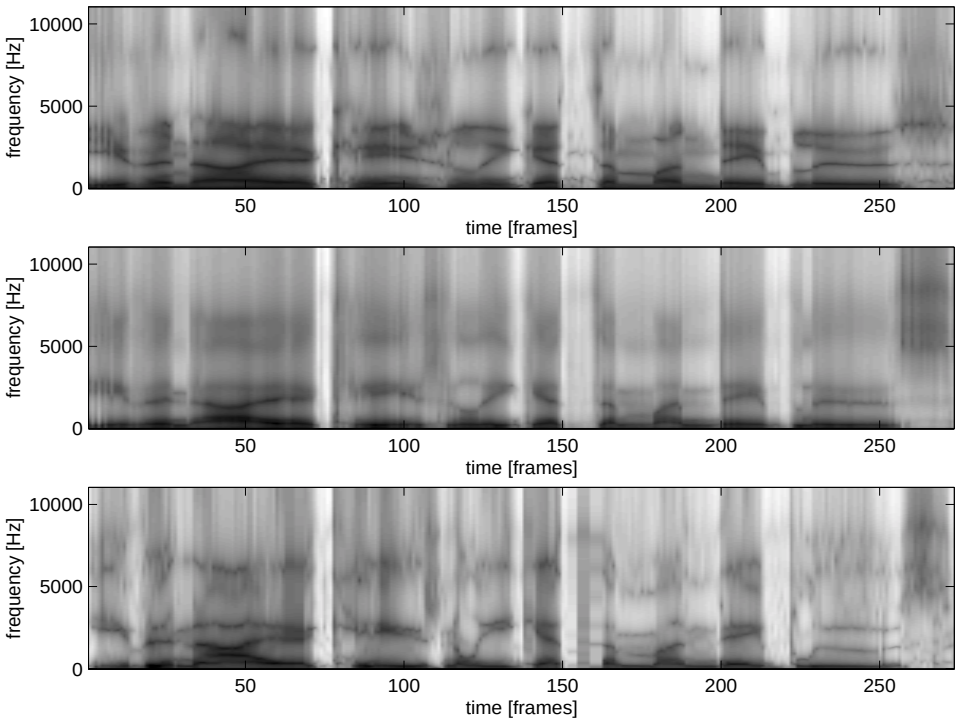


Figure 6.13: Unwarped, LPC log-magnitude spectrograms of the sentence "even I occasionally get the Monday blues", derived from the source (top panel), transformed (middle panel), and target speaker (bottom panel).

6.5 Synthesis

Given a transformed sinusoidal spectrum $\widehat{H}_{sin}$, we compute a frame of the speech signal by a weighted summation of harmonic sinusoids

$$\widehat{s}(n) = g \cdot \sum_{l=1}^{L} \widehat{H}_{sin}(l)\, e^{jlw_0 n} \tag{6.29}$$

where the scalar g is defined as

$$g = \frac{\sqrt{\sum \|H_{sin}(l)\|^2}}{\sqrt{\sum \left\|\widehat{H}_{sin}(l)\right\|^2}} \tag{6.30}$$

Equation 6.30 adjusts the energy of each transformed frame to equal the energy of its respective source frame.

Since we treat the sinusoidal parameters as constant within one frame of speech, discontinuities are avoided by an overlap-add (OLA) approach that eliminates the need to continuously vary the parameters to interpolate between sine-wave tracks. Moreover, OLA allows for a simple implementation of pitch and time-scale modifications [73], which will be discussed in Section 8.1.1. After all speech frames are computed, they are weighted, overlapped, and added, as given by

$$\widehat{s}(n) = w_s^{m-1}(n)\, \widehat{s}^{m-1}(n) + w_s^m(n - T_0^m)\, \widehat{s}^m(n - T_0^m) \tag{6.31}$$

where w_s^m is the complementary synthesis window for frame m obeying the constraint

$$w_s^{m-1}(n) + w_s^m(n - T_0^m) = 1 \tag{6.32}$$

and T_0^m is the fundamental pitch period of frame m in samples. We use an asymmetric trapezoidal window of the shape shown in Figure 6.14.

6.6 Objective evaluation

In this section, we explore the dependency of SET system performance on the choice of training parameters and on particular speaker combinations using an objective evaluation function. After introducing the selection method for speaker combinations, we define two error measures and two corresponding performance indices and report the resulting measurements.

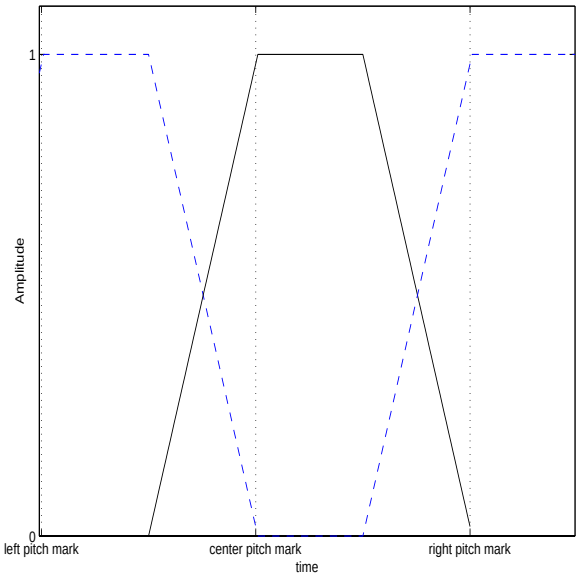


Figure 6.14: Asymmetric trapezoidal synthesis window.

6.6.1 Speech data

We use speech data from the speech corpus introduced in Chapter 5 to train the VT system under varying conditions. For each speaker, the database contains 50 sentences of read speech, resulting in approximately 5 minutes of speech and more than 10,000 feature vectors. Sentences 1–40 (80%) are used as training data and sentences 41–50 (20%) as test data.

Out of the $10^2 - 10 = 90$ possible speaker combinations, we choose 5 male→male, 5 male→female, 5 female→male, and 5 female→female speaker combinations, for a total of 20 transformations. Each speaker is used as a source and as a target twice. Table 6.1 shows the speaker combination matrix, using speaker identifiers introduced in Table 5.1.

6.6.2 Errors and performance indices

There exist two training parameters in training the SET system: the number of mixture components Q and the LPC order p. We varied $Q = 1, 2, 4, 8, 16, 32$ and $p = 12, 14, 16, 18, 20$. A transformation function was estimated using the training data for all 20 speaker combinations and all 30 training parameter combinations.

Three kinds of distances, or errors, are of particular interest in a voice transformation

source/target	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5
M1	-	X					X			
M2		-	X					X		
M3			-	X					X	
M4				-	X					X
M5	X				-	X				
F1		X				-	X			
F2			X				-	X		
F3				X				-	X	
F4					X				-	X
F5	X					X				-

Table 6.1: Speaker combination matrix. Source speakers are represented as rows, target speakers as columns. The symbol "X" indicates a speaker combination that was included in the objective test. For example, a transformation function exists with M1 as the source and M2 as the target speaker. The resulting transformation voice is identified as M1→M2.

system: the *transformation* error $E\left(t\left(n\right),\widehat{t}\left(n\right)\right)$, the *inter-speaker* error $E\left(t\left(n\right),s\left(n\right)\right)$, and the *intra-speaker* error $E\left(t\left(n\right),t_2\left(n\right)\right)$, where $t\left(n\right)$ represents the target speaker's speech, $s\left(n\right)$ the source speaker's speech, $\widehat{t}\left(n\right)$ the transformed speech, and $t_2\left(n\right)$ a second rendition of the target speaker's utterance. The inter-speaker error describes the degree of difference between the source and target speaker, the intra-speaker error gives a measure of how much variability is present from one rendition to the next of the same sentence, and the transformation error represents the difference between the target and the transformed speaker. All three errors are conceptual and cannot be measured directly, but can be approximated using objective and subjective evaluations.

To determine transformation performance objectively, we establish two error metrics we call *performance indices* that compare the transformed speech to the target speech, using the test data. The first metric involves the mean errors between the source and target LSF parameters, and the transformed and target LSF parameters. We define the mean LSF error between two feature streams A and B as

$$E_{LSF}\left(A,B\right)=\frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(L_A^{m,i}-L_B^{m,i}\right)^2} \tag{6.33}$$

where

M	number of frames in feature streams
p	LPC order
$L^{m,i}$	LSF vector component i in frame m

We define the LSF transformation performance index as

$$P_{LSF} = 1 - \frac{E_{LSF}\left(t\left(n\right), \widehat{t}\left(n\right)\right)}{E_{LSF}\left(t\left(n\right), s\left(n\right)\right)} \tag{6.34}$$

The reason for calculating a performance index P_{LSF} instead of working with the transformation error $E_{LSF}\left(t\left(n\right), \widehat{t}\left(n\right)\right)$ is the need for a normalization of errors across different speaker combinations and LPC orders. This is achieved by comparing the transformation error to the inter-speaker error $E_{LSF}\left(t\left(n\right), s\left(n\right)\right)$. As a result, P_{LSF} is zero if the transformation error equals the inter-speaker error, and less than zero if the transformation error is even larger. Conversely, P_{LSF} approaches one as the transformation error approaches zero. In practice, the transformation error is unlikely to approach zero, because there are many ways in which a speaker may render the same utterance.

To estimate the magnitude of acceptable renditions, we measure the target intra-speaker error $E_{LSF}\left(t\left(n\right), t_2\left(n\right)\right)$, calculated from time-aligned utterances from Task 2 and Task 3 recordings (see Section 5.2). Thus, the intra-speaker error approximates the lower bound of an achievable transformation error and, consequently, the optimal value of P_{LSF} will be less than one. Similarly, the transformation error is expected to be below the inter-speaker error for an effective VT system, with $P_{LSF} > 0$.

P_{LSF} operates on parameters that are inputs and outputs of the transformation function and thus directly describes the performance of the transformation function. However, the mean LSF error E_{LSF} is not a standard measure of error between two speech signals. To include a widely used error measure, we also test our result by calculating the spectral distortion (SD) in dB [71, page 443] between the spectra of two signals, H_A and H_B, defined as

$$E_{SD}\left(A, B\right) = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(20\cdot\log_{10}H_A^m\left(e^{j2\pi\frac{N}{n}}\right) - 20\cdot\log_{10}H_B^m\left(e^{j2\pi\frac{N}{n}}\right)\right)^2} \tag{6.35}$$

with $N = 256$. Similar to Equation 6.34, we define the SD transformation performance index as

$$P_{SD} = 1 - \frac{E_{SD}\left(t(n), \widehat{t}(n)\right)}{E_{SD}\left(t(n), s(n)\right)} \tag{6.36}$$

The SD performance index represents a variation of a commonly used measure of VT system performance [2, 5, 84]. The characteristics of P_{SD} are equivalent to those of P_{LSF} discussed above.

6.6.3 Results

Figures 6.15 and 6.16 show the inter-speaker errors $E_{LSF}\left(t(n), s(n)\right)$, transformation errors $E_{LSF}\left(t(n), \widehat{t}(n)\right)$, and intra-speaker errors $E_{LSF}\left(t(n), t_2(n)\right)$ for one source-target speaker combination. $E_{LSF}\left(t(n), \widehat{t}(n)\right)$ is shown for all values of Q, the number of mixture components. All errors are grouped by values of p, the LPC order.

Examining Figure 6.15, one can observe a decrease of the training error $E_{LSF}\left(t(n), \widehat{t}(n)\right)$ with an increase in Q, for any particular choice of p. This is to be expected, because an increase in the number of mixture components results in a more accurate modeling of the probability densities (see Section 6.11). In the theoretical limit, using as many mixture components as there are feature vectors would result in a degenerate table-lookup algorithm with perfect prediction (assuming a one-to-one function). It can also be observed that the errors at higher values of Q are relatively close to the lower bound approximated by $E_{LSF}\left(t(n), t_2(n)\right)$, indicating a good fit of the transformation function to the training data.

Figure 6.16 shows the same types of errors evaluated on the test data. As expected, test errors are generally higher than their corresponding training errors. Initially, the error decreases with an increase in Q, but then increases. This indicates an over-fitting of the estimation function to the training data for values of $Q > 16$ for $p \leq 18$, but $Q > 8$ for $p > 18$. Thus, the optimal number of mixture components depends on the choice of LPC order.

Figure 6.17 shows the resulting LSF transformation performance index P_{LSF} for the same speaker combination. In this representation, we can compare the effects of p on

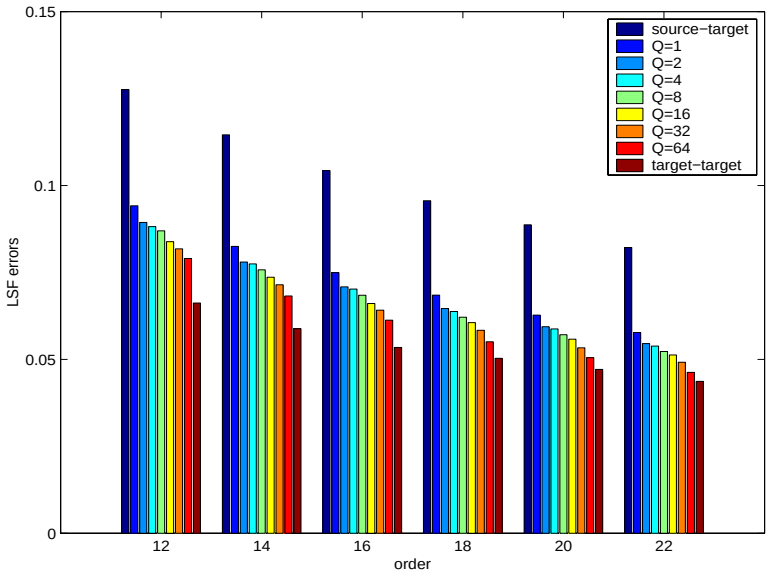


Figure 6.15: Inter-speaker, transformation, and target intra-speaker LSF training errors for the speaker combination M1→F2.

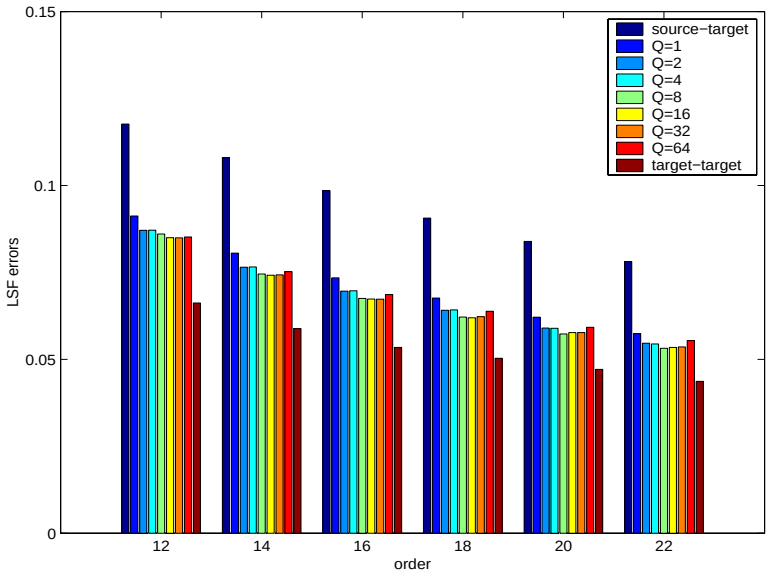
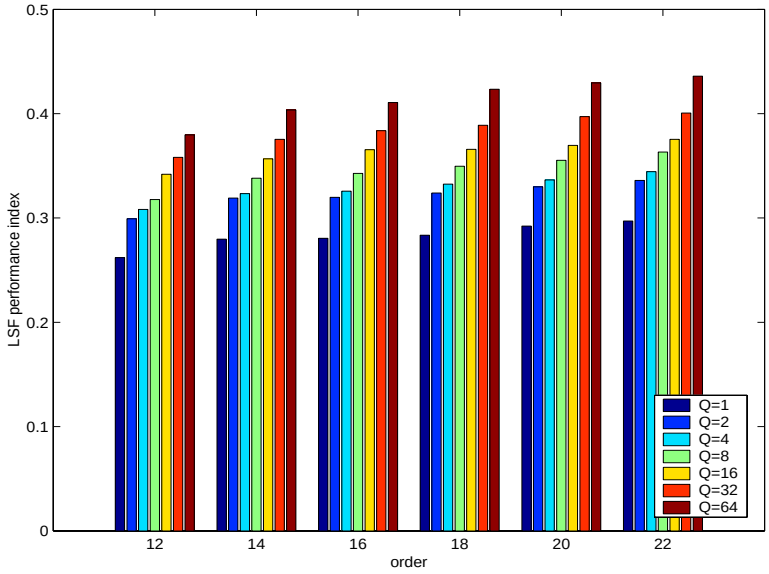
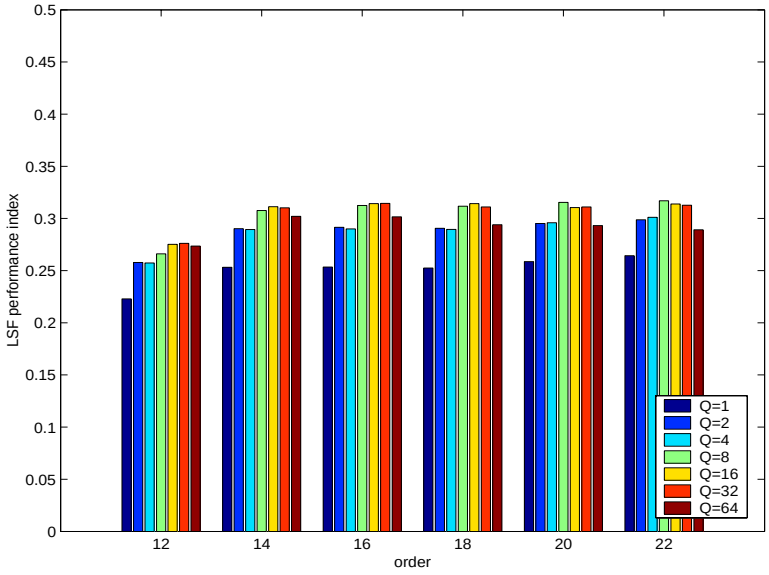


Figure 6.16: Inter-speaker, transformation, and target intra-speaker LSF test errors for the speaker combination M1→F2.



(a) Training.



(b) Testing.

Figure 6.17: LSF transformation performance index P_{LSF} for the speaker combination M1→F2 for both training and test data.

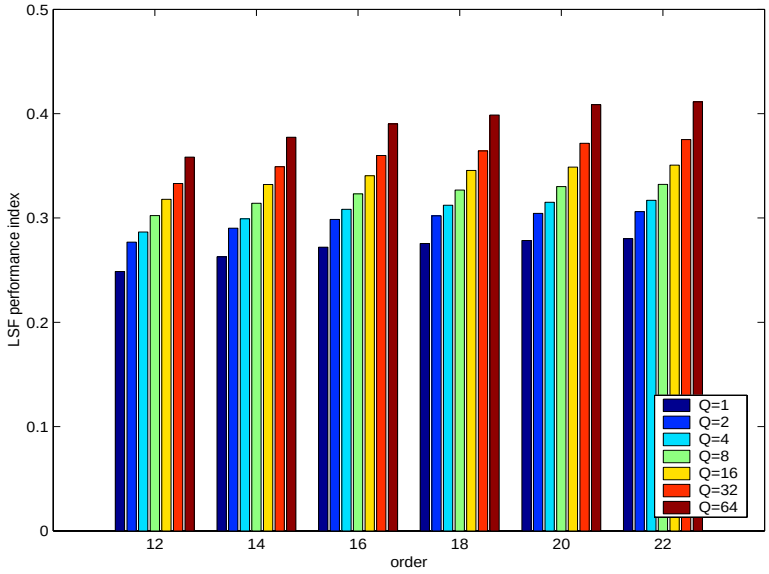
transformation performance. While a continuous increase in P_{LSF} with p for a fixed Q is apparent on the training data, we observe only marginal improvements for $Q < 8$ and no improvements and even small losses at $Q > 8$ for $p > 14$. Our interpretation is that once p is large enough to model the major events in the spectrum, the LSF parameters are a good description of the acoustic "state" of the speaker, and thus a further increase in p does not benefit transformation performance.

We have compared P_{LSF} between several speaker combinations, as well as between averages of male→male, male→female, female→male, and female→female transformations. All results were approximately the same, and no dependencies on gender or other factors could be identified. Figure 6.18 shows P_{LSF}, averaged over *all* speaker combinations. Indeed, the result is quite similar to the single speaker combination in Figure 6.17, except that the values now show the general trends (for example, over-fitting) more clearly.

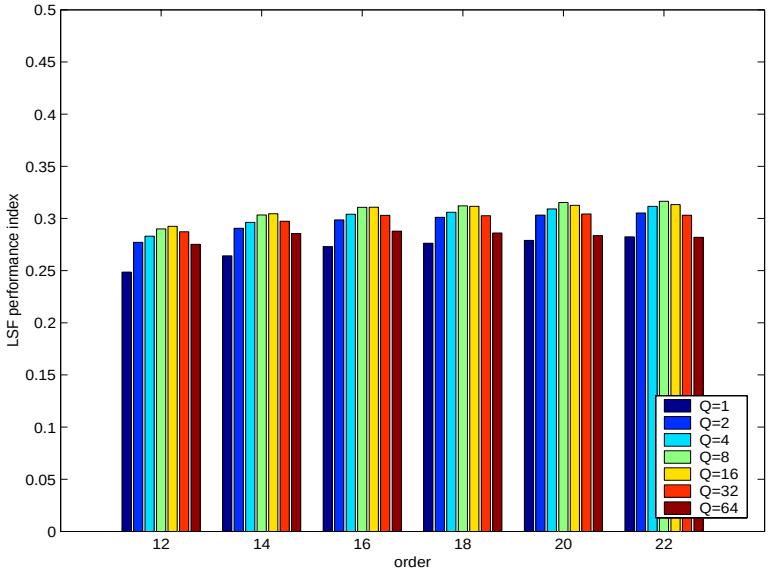
Figure 6.19 shows P_{SD} averaged over all speaker combinations. We observe that P_{SD} is similar to P_{LSF}. In contrast to P_{LSF}, however, P_{SD} increases slightly with an increase in p, for any fixed Q. This is due to an increase of the total number of parameters of the transformation function, an effect that was canceled in calculating P_{LSF} through the presence of the $1/p$ term in Equation 6.34.

Finally, we report the actual values of the spectral distortion errors on the test set. On average, $E_{SD}\left(t\left(n\right), s\left(n\right)\right)$ was 9.3 dB (ranging from 7 dB to 12.7 dB), $E_{LSF}\left(t\left(n\right), \widehat{t}\left(n\right)\right)$ was 7.5 dB (ranging from from 5.5 dB to 11.2 dB), and the intra-speaker error $E_{SD}\left(t\left(n\right), t_2\left(n\right)\right)$ was 5.3 dB (ranging from 4 dB to 6.1 dB). For comparison, it is commonly thought that values of less than 1 dB are not humanly detectable [95], whereas a value of 3 dB is considered unacceptable for a LSF-based, 10 bits/frame speech coding system [71].

At first, it may seem contradictory that $E_{SD}\left(t\left(n\right), t_2\left(n\right)\right)$, which is much larger than 3 dB, should be considered as an unacceptable result. After all, a second rendition of the same sentence by the same speaker, aligned in both time and pitch, is the best approximation to an original rendition! This situation is a reminder of the limited usefulness of an objective function to assess the quality of speech. Rather, they can be useful in comparing the performance of varying parameters within the same system, as we have done, and the error values of $E_{LSF}\left(t\left(n\right), \widehat{t}\left(n\right)\right)$ should always be thought to be bounded from below by

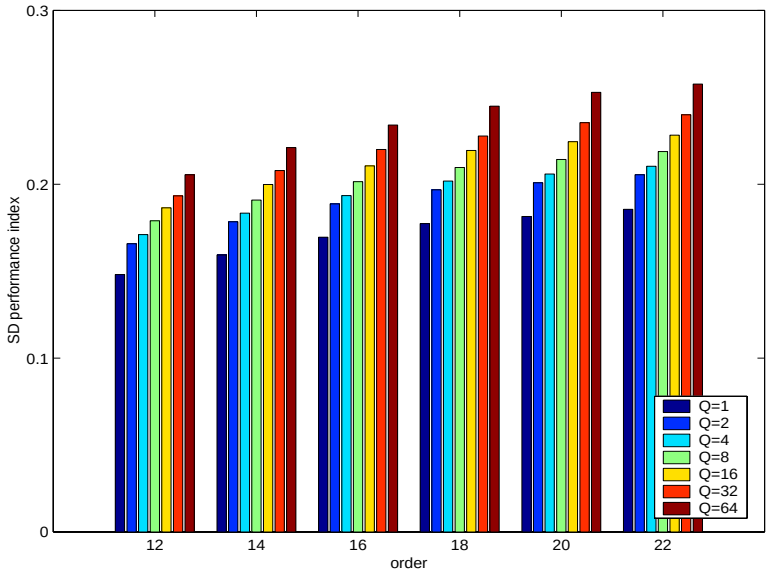


(a) Training.

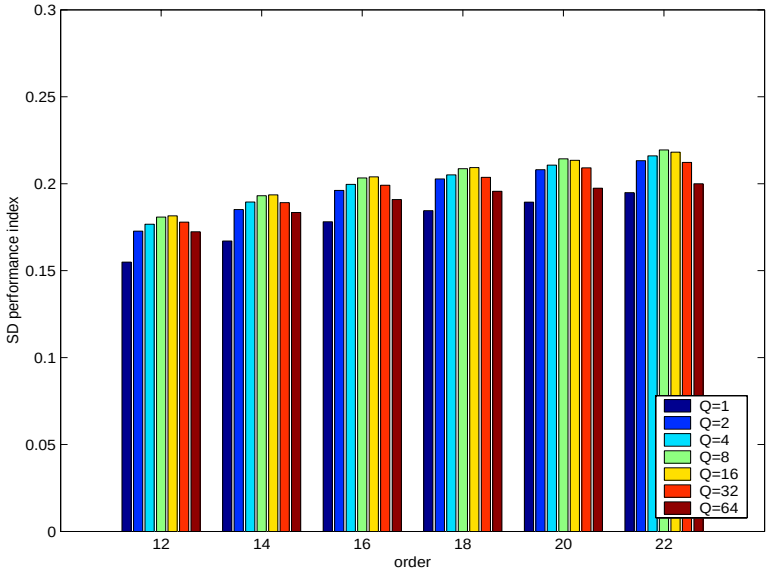


(b) Testing.

Figure 6.18: LSF transformation performance index P_{LSF} averaged over all speaker combinations, for both training and test data.



(a) Training.



(b) Testing.

Figure 6.19: SD transformation performance index P_{SD} averaged over all speaker combinations, for both training and test data.

the value of $E_{SD}\left(t\left(n\right), t_2\left(n\right)\right)$. In Chapter 8, we will describe a method that allows for a subjective evaluation of the transformation performance of the SET system.

Chapter 7

High Resolution Voice Transformation

This chapter proposes a new approach to voice transformation. Using the SET system described in Chapter 6 as the baseline system, we design and implement a *residual prediction* system, resulting in high resolution transformations. In the first section, we motivate the advantages of synthesizing the transformed output with a high degree of spectral detail and give an overview of the algorithm. Section 7.2 describes the technical details of the implementation of the residual prediction system. Finally, Section 7.3 presents the results of several objective evaluations.

7.1 Motivation and design overview

We motivate the generation of a detailed transformation spectrum with an experimental result from a study by Kain and Macon [37]. In this study, the authors tested a VT system that modified the LPC spectral envelope of the source signal while leaving the LPC residuals of the source speaker unmodified. In an ABX test using synthetic speech samples, listeners judged a male→male transformation to be closer to the target speech with an average score of 52%. From this result and from interviews with listeners, the authors concluded that the transformed speech was perceived as neither the source nor the target speaker, but instead as a new, third speaker. In a second test, the authors modified the VT system to replace the source LPC residuals with the corresponding target LPC residuals from the target speaker's synthesis database. After the modification, the ABX scores were 100%. The authors concluded that the LPC residual contains a significant amount of speaker information and that an effective VT system must go beyond a spectral

envelope representation of the speech signal.

The LPC residual can contain several types of information about the speech signal. Since the residual is the "error" of the LPC analysis (as calculated by Equation 6.6) it can be considered to contain all the effects of the speech signal which are not accounted for by the assumptions of the LPC model described in Section 6.4. Thus, the LPC residual may contain effects of one or more of the following:

- Nonlinear interaction of the vocal tract with the sub-glottal area.
- Wave absorption at the muscle tissues.
- Resonances not modeled by the LPC filter, if the LPC order was chosen too small.
- Spectral zeros during nasalization.
- Details of the glottal pulse shape, such as secondary glottal pulses (diplophony).

We claim that these effects, manifested in *spectral details* not captured in the LPC spectral envelopes, contain important speaker information. This claim is supported by a second perceptual experiment conducted by Kain and Macon [39]. In this test, 16 listeners judged 120 pairs of sentences to be either spoken by the same speaker or by two different speakers. The results of two conditions, separated by gender, are reproduced in Table 7.1. The first condition consisted of natural utterances which were normalized in regard to their pitch, time, and energy. The second condition was equivalent to the first condition, with the exception of replacing the short-term spectrum by its corresponding spectral envelope described by the LPC spectrum, using a sinusoidal analysis-resynthesis scheme. Thus, the only difference between the two conditions was the presence and the absence of the LPC residual in the first and second condition, respectively. The difference in listeners' discrimination abilities was significant for the set of male speakers. It follows that a VT system that includes LPC residual effects in the transformed speech signal has increased performance in the recognizability of some transformed speakers as the target speaker, as we will show explicitly in Chapter 8.[1]

[1] It has also been shown that the addition of spectral detail can improve the accuracy of an automatic speaker identification system by up to 8% [35].

Condition	males	females
prosody normalized	83 (79–87)	89 (85–92)
prosody normalized and LPC coded	71 (65–76)	88 (84–91)

Table 7.1: A subset of perceptual listening test results in a study by Kain and Macon [39]. Shown are the percentages of correct discrimination of speakers, averaged over all responses and listeners. The 95% confidence interval is in parentheses.

There exist two previous approaches which attempt to render the transformed speech spectrum with the target speaker's spectral details. The systems of Arslan et al. [6, 7, 5] and Lee et al. [52] are based on a spectral envelope transformation system similar to Chapter 6, but in addition, the acoustic feature set includes a parameterization of the source and target speakers' LPC residuals, as described in Section 3.1.2. The authors of both approaches attribute the improved performance to the integration of the LPC residual. However, the representation of the LPC residual was limited in both cases, either to a magnitude spectrum representation (Arslan et al.) or to the output of a nonlinear predictor (Lee et al.).

In contrast to *transforming* the LPC residual, we propose a method in which we *predict* the target LPC residual from the transformed LPC spectral envelopes during voiced speech. The underlying assumption of this approach is that for a particular speaker and within some phonetically-similar class the residuals are similar and predictable. Specifically, the residual's magnitude spectrum contains the systematic errors made by the spectral envelope fit that are particular to a certain speech sound (e.g., zeros during a nasal), and the phase spectrum contains important information about the natural phase dispersion of the signal, as opposed to the minimum phase assumption of the LPC model.

Given a speaker's LPC parameters of an utterance, the proposed residual prediction system is able to approximate the original speech waveform more accurately than a simple LPC coding by adding spectral details to the LPC spectrum. There is an interesting parallel to a recent publication by Etxebarria et al.[21], in which the authors aim to improve the naturalness of a particular TTS system, MBROLA [17, 18]. The MBROLA system uses a single fixed phase vector for the entire speaker database. The study suggests that naturalness improves when different phase vectors are used for different phonemes,

as shown by informal perceptual tests. The residual prediction system is similar to this approach, except it is based on an unsupervised training method, and includes both phase and magnitude spectrum information.

It must be noted that the proposition of predicting the residual from the envelope seems counter to the prevailing notion that the residual is "white", that is, completely uncorrelated with the spectral envelope. This assumption is made in classical source-filter theory (see Section 2.1) and for many speech systems it is fairly good approximation. However, when only one speaker is considered, we will show that the residual is correlated with the spectral envelope, making prediction possible. Now the process of VT can be viewed as a type of speaker-dependent speech coding, where the transformed spectral envelope parameters are the transmission parameters, describing the state of the speaker. At the receiver, residual prediction is used (after a training phase) to go from the low-dimensional representation of speech to the high-dimensional representation, in our case the short term spectrum. We will show that this strategy is effective, in terms of both objective and subjective evaluation measures.

7.2 Implementation

This section describes the implementation of the residual prediction (RP) system. In the first two subsections, we will consider the simple case of coding the waveform of an individual speaker through LPC parameters. Section 7.2.3 will integrate the RP system into the final transformation system.

The RP system consists of a LPC parameter classifier and a LPC residual codebook. Before use, the system must be trained on speech data of the speaker whose LPC residuals are to be predicted. We will now describe the training process and the operation of the RP system in detail.

7.2.1 Training

Two tasks must be carried out during training: building a classifier and constructing a codebook. Each class of the classifier is associated with an entry in the codebook. Two

data sets are necessary for training the RP system, the set of LPC parameters of voiced frames (unvoiced frames will be discussed below) and the collection of associated LPC residuals.

To start, we calculate LPC parameters a_k using the approach described in 6.2, and then convert them to a cepstral representation c_k via the recursion [8]

$$c_1 = a_1 \tag{7.1}$$

$$c_n = \sum_{k=1}^{n-1} \left(\frac{k}{n} - 1 \right) a_k c_{n-k} - a_n \tag{7.2}$$

Let $C_{train\ p \times N}$ represent the LPC cepstra of all voiced frames in the training data, where p is the LPC order and N is the number of voiced frames.

In a second step, The LPC residual magnitude and phase spectra are calculated via

$$R_m(l) = 20 \cdot \log_{10} H_{sin}(l) - 20 \cdot \log_{10} H_{lpc}\left(e^{j\left(W^{-1}(lw_0)\right)}\right) \tag{7.3}$$

and similarly

$$R_p(l) = \angle H_{sin}(l) - \angle H_{lpc}\left(e^{j\left(W^{-1}(lw_0)\right)}\right) \tag{7.4}$$

where $W^{-1}(\cdot)$ is an inverse warping function described in Section 6.1.3. We resample the spectra R_m and R_p to a length of exactly 100 points, because their spectral resolution differs from frame to frame, depending on F_0, due to our pitch-synchronous analysis. Given a resolution of 100 points, residual spectra can be stored without any losses for any F_0 above approximately 110 Hz. Resampling the spectrum is implemented by interpolation. We assume a cubic function for the residual magnitude spectrum R_m. For the residual phase spectrum R_p, however, we use a nearest-neighbor interpolation, because of discontinuities that arise from the modulo 2π representation of phase. Let $M_{100 \times N}$ and $P_{100 \times N}$ represent the frequency-normalized residual magnitude and phase spectra of all voiced frames.

Classifier

The goal of the classifier is to assign degrees of class-membership to an incoming LPC parameter vector. To this end, we estimate a GMM (see Section 6.3.2) with Q mixture

components on C_{train}

$$P_{GMM}\left(C_{train};\alpha,\mu,\Sigma\right)=\sum_{q=1}^{Q}\alpha_q N\left(C_{train};\mu_q,\Sigma_q\right),\ \sum_{q=1}^{Q}\alpha_q=1,\ \alpha_q\geq 0 \tag{7.5}$$

where $N\left(x;\mu,\Sigma\right)$ denotes the p-dimensional normal distribution with mean vector μ and covariance matrix Σ (see Equation 6.17).

The estimation of the GMM is achieved by the EM algorithm described in 6.3.2. Since training a GMM minimizes the local mean squared error, the distance metric during classification is equivalent to a log magnitude spectral distance because of the use of a cepstral representation.

Residual codebooks

Once the classifier is established, the residual codebook can be populated. The procedure for this differs between magnitude and phase spectra.

We use the following probabilistic approach for calculating the magnitude of codebook spectra. We first calculate the GMM posterior probabilities of C_{train} for each class q and frame i

$$h_{q,i}=p\left(c_q|C^i\right)=\frac{\alpha_q N\left(C^i;\mu_q,\Sigma_q\right)}{\sum_{p=1}^{Q}\alpha_p N\left(C^i;\mu_p,\Sigma_p\right)} \tag{7.6}$$

Then, the magnitude of codebook entry q is

$$m_q=\sum_{i=1}^{N}M_i\cdot\frac{h_{q,i}}{\sum_{j=1}^{N}h_{q,j}} \tag{7.7}$$

Thus, each entry is the normalized, weighted sum of all residual magnitude spectra, where the weights correspond to the degree of membership to that particular class. This approach works well because it is in effect an averaging and smoothing operation, resulting in codebook entries that are representative of the spectral trends in their class.

An example of the magnitude spectra in a complete codebook is shown in Figure 7.1. We observe that there are some global trends shared by all entries. For example, an attenuation at the extremely low frequencies, as well as a significant roll-off at the higher frequencies. Otherwise, the codebook entries have quite different peaks and valleys,

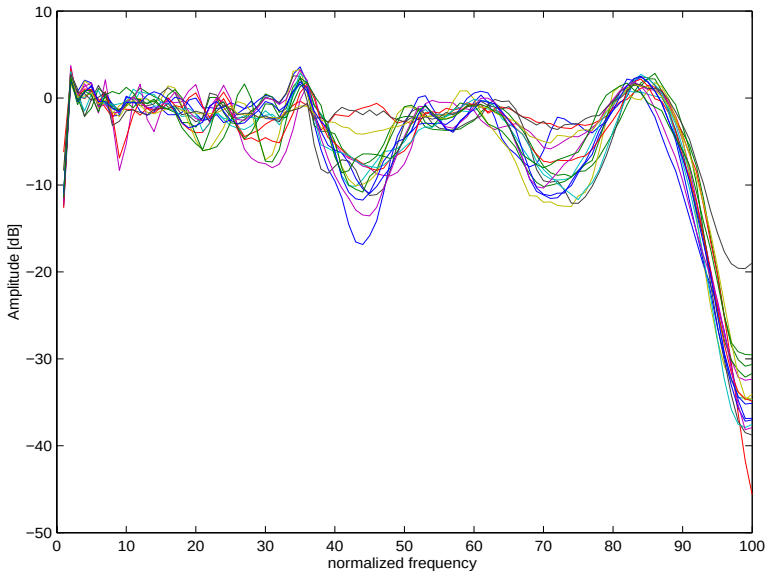


Figure 7.1: Magnitude spectra in a 16-entry residual codebook.

supporting our hypothesis of a dependence between residuals and the spectral envelope of a speaker. This is confirmed numerically in Section 7.3.1.

Unfortunately, a probabilistic approach can not be used for determining the phase of codebook spectra. The values of phase spectra are given modulo 2π, and thus any summation operation is ill-defined. Using the unwrapped phase spectrum may work for lower frequencies, where the phases can be assumed to be slowly changing with frequency, but at higher frequencies phases become more chaotic and the previous assumption is invalid. For these reasons, we choose the residual phase vector of the centroid of each class for the phase of the codebook entry for that class by letting

$$p_q = P_{\arg_i \max h_{q,i}} \tag{7.8}$$

where residual phase vector with the maximum likelihood. A disadvantage of this method is that it is possible to choose a residual phase spectrum that is not representative of the general trends within that class. Vectors m and p together constitute the final codebook.

7.2.2 Residual prediction

Once the RP system is trained, it can predict LPC residuals from the LPC spectral envelope of one speaker using the following approach. First, we calculate the GMM posterior probabilities from a set of new cepstral vectors C_{test} for all frames i and classes q using Equation 7.6. Then, the frequency-normalized residual magnitude spectrum of all voiced frames is given by a weighted sum of magnitude codebook entries

$$\widehat{M}_i = \sum_{q=1}^{Q} m_q \cdot h_{q,i} \tag{7.9}$$

and the phase spectrum is given by the most likely phase codebook entry

$$\widehat{P}_i = p_{\arg_q \max h_{q,i}} \tag{7.10}$$

Similar to Section 7.2.1, it is possible to perform a "soft prediction" on magnitude spectra, but not on phase spectra, which do not allow addition. However, the "hard" prediction of residual phase vectors introduces an audible degradation of the final speech, most often perceived as "roughness". This can happen whenever a switch between phase codebook entries occurs. To alleviate the problem phase discontinuity, we first unwrap the trajectories of each harmonic phase stored in matrix $\widehat{P}$ over all frames. Then, we smooth the trajectories of all voiced regions by zero-phase filtering with an eight-point Hanning window.

In order to accommodate the various F_0 values of different frames, the lengths of $\widehat{M}$ and $\widehat{P}$ are adjusted to match the lengths of their corresponding synthesis frames. This is accomplished through a second resampling by cubic interpolation for the magnitudes and a nearest-neighbor interpolation for the phases, resulting in $\widehat{R}_m$ and $\widehat{R}_p$, respectively.

Finally, the discrete complex spectrum of a synthesis frame is calculated via

$$\widehat{H}_{sin}(l) = 10^{\left(\left(20\cdot\log_{10}\left(H_{lpc}\left(e^{j\left(W^{-1}(l\,w_0)\right)}\right)\right)+\widehat{R}_m(l)\right)/20\right)} \cdot e^{j\left(\angle H_{lpc}\left(e^{j\left(W^{-1}(l\,w_0)\right)}\right)+\widehat{R}_p(l)\right)} \tag{7.11}$$

7.2.3 Transformation

We combine the RP system with the SET system of Chapter 6 resulting in a high resolution transformation (HRT) system. An overview block diagram of the HRT system is shown in

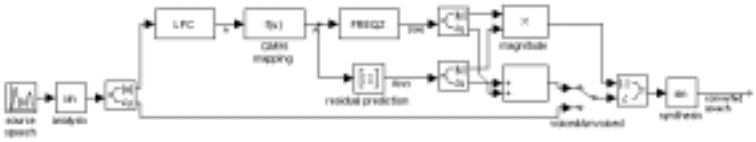


Figure 7.2: Block diagram of the HRT system.

Figure 7.2. The analysis and transformation stages are unmodified, but the synthesis stage now incorporates the RP system, which predicts the target speaker's LPC residuals from the *transformed* LPC spectral envelope parameters. During unvoiced speech, the target residual spectra are resampled versions of the source speaker's residual spectra. Using Equation 7.11, the HRT system integrates both the transformed LPC spectral envelope and the predicted LPC residuals into the transformed speech signal waveform using the synthesis approach described in Section 6.5.

7.3 Objective evaluation

In this section, we will test several aspects of the RP system, using objective measures. We use the speech database described in Section 5 and the analysis methods described in Section 6.2 to construct training and test data for each of the ten speakers. There are 40 sentences and approximately 8,000 voiced frames available for training, and 10 sentences or approximately 2,000 voiced frames for testing.

7.3.1 Codebook validation

In a first experiment, we validate the effectiveness of magnitude codebook entries by comparing them to residual magnitude spectra that are either within the entry's class or out of its class. We use a spectral distortion similar to Equation 6.35 and define the within-class error

$$SD_{in}\left(q\right) = \frac{1}{N_{in}} \sum_{i=\arg_i \max h_{q,i}} \sqrt{\frac{1}{100} \sum_{j=1}^{100} \left(m_{j,q} - M_{j,i}\right)^2}$$

and the out-of-class error

$$SD_{out}\left(q\right) = \frac{1}{N_{out}} \sum_{o\neq\arg_i \max h_{q,i}} \sqrt{\frac{1}{100} \sum_{j=1}^{100} \left(m_{j,q} - M_{j,o}\right)^2}$$

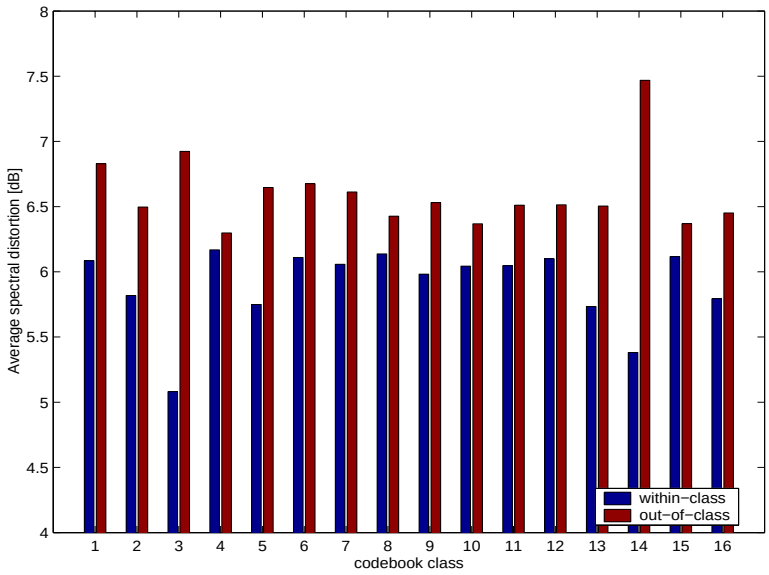


Figure 7.3: Values of the within-class error SD_{in} and the out-of-class error SD_{out} for an example codebook with 16 entries.

where $h_{q,i}$ is given by Equation 7.6. The average values of SD_{in} and SD_{out} in dB, measured on the training data, are shown in Figure 7.3 for an example codebook with 16 entries. We observe that the within-class SD error is always below that of the out-of-class SD for any class.

Comparing the values for SD_{in} among different classes, we observe that some codebook entries have a much lower within-class error. We conclude that some classes have a lower variance in the residual magnitude spectrum than others. A codebook containing classes with a high variance may benefit from an increase in the number of classes. This will be the subject of the next experiment.

7.3.2 Speech coding performance

The goal of a second experiment is to characterize the relationship between the performance of the RP system and its training parameters p, the LPC order, and Q, the number of classes that are used in the classifier and the number of codebook entries, using a speech coding task. To measure the speech coding performance of the RP system, we measure the signal-to-noise ratio (SNR) of the coded speech signal. The signal-to-noise ratio (SNR)

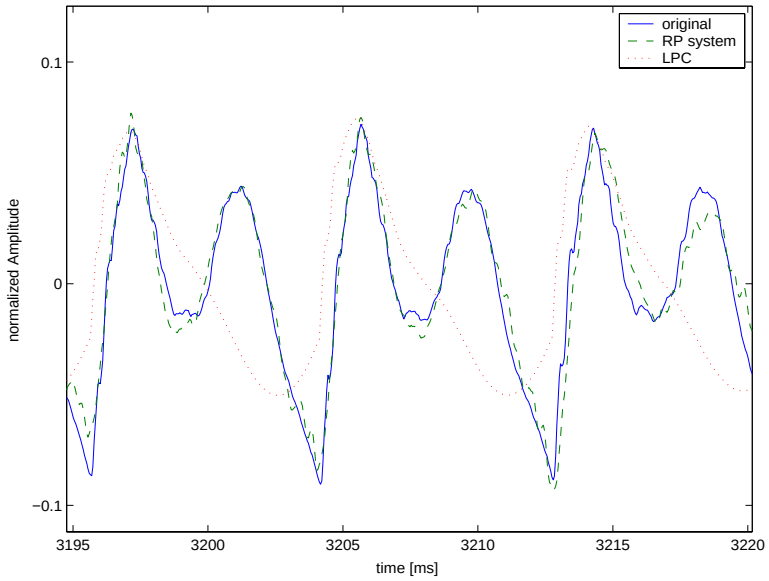


Figure 7.4: A small segment of an original speech signal, the output of the RP system, and the LPC coded signal.

is defined as the ratio of the signal energy to the quantization noise energy. Assuming zero-mean signals,

$$SNR\left(s\left(n\right),\widehat{s}\left(n\right)\right)=10\cdot\log_{10}\frac{\sum s\left(n\right)^{2}}{\sum\left(\widehat{s}\left(n\right)-s\left(n\right)\right)^{2}} \tag{7.12}$$

on a dB scale, where $s\left(n\right)$ represents the original speech signal and $\widehat{s}\left(n\right)$ its coded form. We calculate the SNR value of an utterance by averaging SNR values for small segments of approximately 20 ms duration. Perceptual quality is better reflected in a segmental SNR value than in a SNR calculated on an entire utterance, because errors in low- and high-energy portions of speech are computed separately [47, 9].

Figure 7.4 shows a small segment of three speech signals: the original signal, the output of the RP system, and the LPC coded signal. The LPC parameters for the latter are identical to the input of the RP system. We observe that the RP system is capable of approximating the original waveform much closer than the simple LPC coding.

We calculate $SNR\left(s\left(n\right),\widehat{s}\left(n\right)\right)$, where $s\left(n\right)$ represents an original waveform of any one of ten speakers, and $\widehat{s}\left(n\right)$ its coded form, either by a LPC coding, or the output of the RP system with all possible configurations of $p=12,14,16,18,20,22$ and $Q=16,32,64,128$. Figure 7.5 shows the results, averaged over 10 test sentences as well as over all male

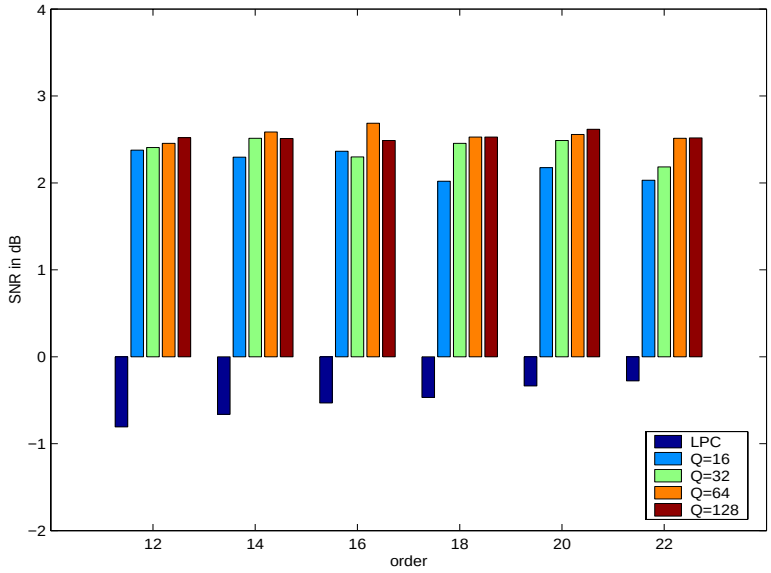


Figure 7.5: Average SNR values between the original speech signal and various coded forms, for male speakers.

speakers. As we can see, the RP system has a consistently higher SNR value than the LPC coding. The optimal value of Q depends on p, though not systematically. While the SNR values for the LPC coded signal increase with an increase in p, the performance of the RP system is approximately independent of p. SNR values averaged over female speakers are shown in Figure 7.6. In contrast to male speakers, we see an increase in SNR values as Q increases, for any p. However, the RP system still results in higher SNR values than the LPC coded signal, in any configuration. Thus, we have shown that the RP system is able to approximate the original waveform more closely than a LPC coding in a speech coding task, using identical LPC parameters.

7.3.3 Transformation performance

In a third experiment, we compare the transformation performance of the baseline SET system with the HRT system and another recently published high resolution voice transformation approach, called STASC, proposed by Arslan et al. [6, 7, 5]. Because we did not have access to this system, we implemented its fundamental ideas into the existing analysis-synthesis framework of Section 6.1, resulting in an alternative (ALT) transforma-

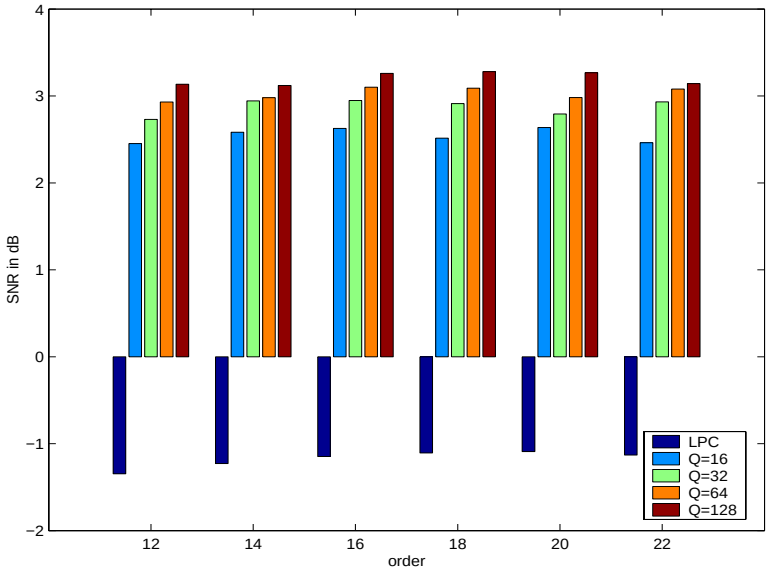


Figure 7.6: Average SNR values between the original speech signal and various coded forms, for female speakers.

tion system.

In the ALT system, a transformed speech spectrum corresponds to

$$\widehat{S}(w) = F_{exc}(w) \cdot F_{tract}(w) \cdot S(w) \tag{7.13}$$

where $F_{exc}(w)$ represents an excitation filter, $F_{tract}(w)$ a vocal tract filter, and $S(w)$ and $\widehat{S}(w)$ denote the source and transformed speech spectrum, respectively. In order to calculate $F_{tract}(w)$ and $F_{exc}(w)$, the algorithm first generates codebooks of phonetically associated LSF from speech of both the source and the target speaker during training. During transformation, a codebook weight estimation method approximates the LSF vectors X of the source speaker as

$$\widehat{X} = \sum_{q=1}^{Q} v_q \cdot x_q \tag{7.14}$$

where Q is the codebook size, x_q the q^{th} source LSF codeword, and v_q its weight. Given v, the excitation filter is the weighted combination of codeword excitation filters

$$F_{exc}(w) = \sum_{q=1}^{Q} v_q \frac{T^q_{exc}(w)}{S^q_{exc}(w)} \tag{7.15}$$

where $S^q_{exc}(w)$ and $T^q_{exc}(w)$ represent average source and target LPC residual magnitude spectra of the q^{th} codeword. Similar to Equation 7.14, the LSF vectors $\widehat{Y}$ of the transformed speech are approximated as

$$\widehat{Y} = \sum_{q=1}^{Q} v_q \cdot y_q \tag{7.16}$$

where y_q is the q^{th} target LSF codeword. Finally, $F_{tract}(w)$ is the LPC filter calculated from $\widehat{Y}$.

The key difference between the HRT and the ALT system is that the ALT system produces the target envelope by mixing a fixed set of target envelopes, whereas the HRT system applies a transformation function to the source envelope. Further, only the magnitude of the short-term spectrum is modified, whereas the HRT system operates on the complex spectrum. Another difference is the use of identical weights for both the vocal tract and the excitation filters in the ALT system. This limits the number and types of transformations to the LPC residual. In comparison, the HRT system completely separates the process of generating transformed LPC envelopes and LPC residuals.

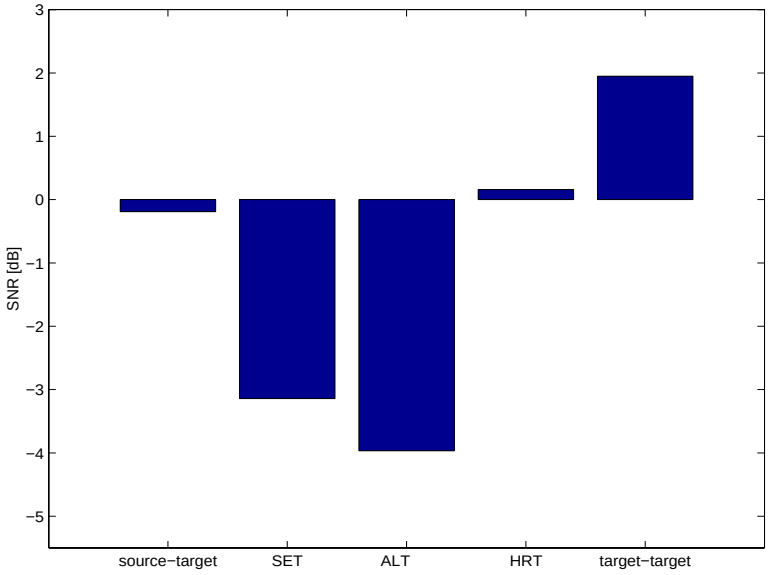
We selected 5 male→male, 5 male→female, 5 female→male, and 5 female→female speaker combinations, for a total of 20 transformations, to reduce the complexity of 90 possible speaker combinations, similar to the evaluation of the SET system in Section 6.6. Each speaker is used as a source and as a target twice. Table 6.1 shows the speaker combination matrix, using speaker identifiers from Table 5.1.

We measured the following segmental signal-to-noise ratios: $SNR(t(n), s(n))$, $SNR\left(t(n), \widehat{t}(n)\right)$, and $SNR(t(n), t_2(n))$, where $t(n)$ represents the target speech signal, $s(n)$ the source speech signal, $\widehat{t}(n)$ the transformed speech signal, and $t_2(n)$ a second realization of $t(n)$ by the target speaker. All speech signals were time-aligned appropriately. Thus, the first measure corresponds to the inter-speaker distance, the second to the distance between the target and the transformed speech, and the last to the intra-speaker distance. Transformed speech was generated by the SET, ALT, and HRT system for comparison.

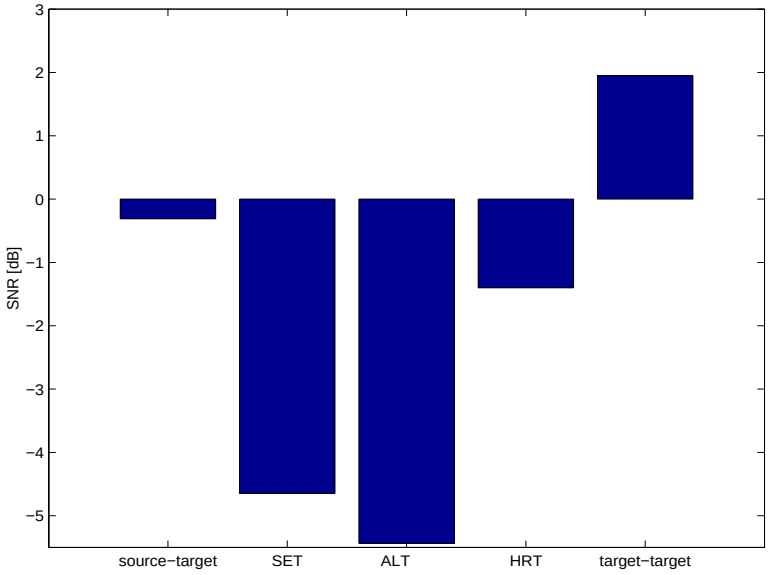
Figures 7.7 and 7.8 show results for a male and a female target speaker. Each figure displays SNR values that are averaged over the 10 test sentences, and over 5 male or 5 female source speaker combinations. Comparing the three transformation systems, we

observe that the output of the HRT system has the best performance of all three, with the SET system in second place. In the case of female→male transformations, the performance of all three systems deteriorates. This is due to the degrading effect of the pitch modification that is necessary for alignment, which in this case decreases F_0 by a factor of approximately 2. We further observe that the ALT system is consistently below that of the two other systems, even the SET system. We speculate that this is due to the difference in paradigms: whereas SET and HRT generate a mixture of mappings acting on the input vector, ALT generates a new target feature vector by mixing target feature codewords.

Interestingly, SNR values for both the SET and the ALT systems were lower than the inter-speaker distances. However, one must bear in mind the limits of an objective evaluation: as we will see, a human is unlikely to rate $s(n)$ as closer to $t(n)$ than $\widehat{t}(n)$. While an objective measure allows as a first glimpse of the situation, only a subjective evaluation can uncover the perceptual significances between the various systems. Such a subjective evaluation is described in the next chapter.

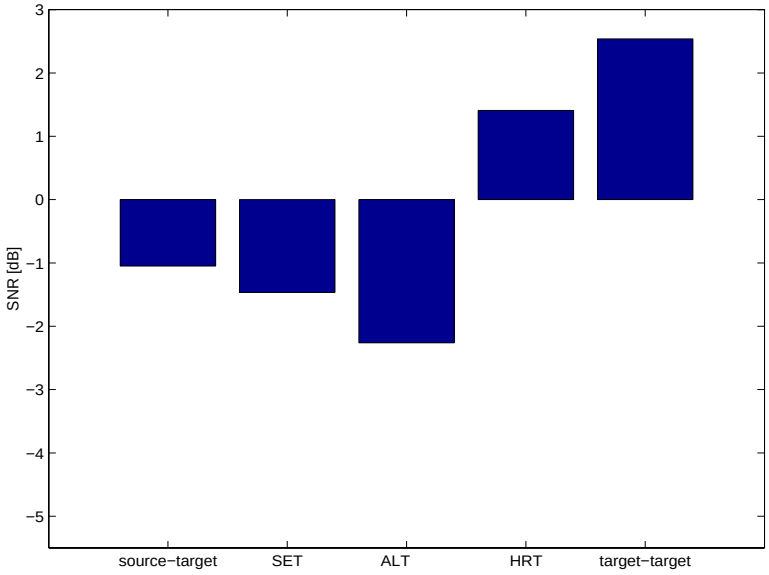


(a) Male to male (intra-gender).

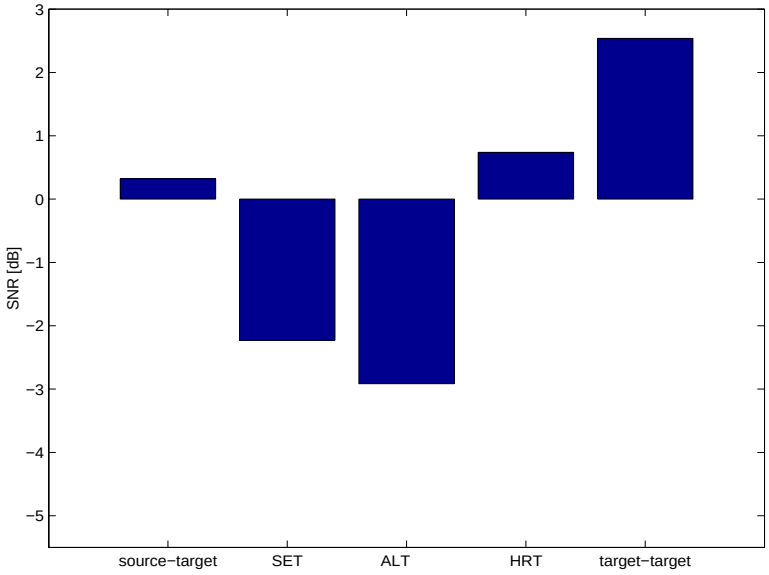


(b) Female to male (inter-gender).

Figure 7.7: Results with male target.



(a) Male to female (intergender).



(b) Female to female (intra-gender).

Figure 7.8: Results with female target.

Chapter 8

Subjective Evaluation

In this chapter we propose a new strategy for the evaluation of VT system performance, especially with regard to the speaker recognizability of the transformed speech signal. Our evaluation consists of three types of perceptual tests: a speaker discrimination test, a system comparison test, and a speech quality comparison test. We first describe the design and administration of the perceptual tests in Section 8.1 and then analyze and discuss the test results in Section 8.2.

8.1 Perceptual test design

The output of a VT system can be evaluated along three major dimensions: Intelligibility, naturalness, and speaker recognizability. Evaluation of the first two areas is commonplace for many speech systems, for example speech coders and TTS systems, and several test standards have been developed [47]. The area of speaker recognizability, however, has received relatively little attention [78]. In particular, the evaluation of speaker recognizability of transformed speech signals has been limited to the strategies described in Section 3.2.

We propose a new approach to measuring the speaker recognizability of a transformed speech signal. Using two speaker discrimination tests, we establish the degree by which listeners can recognize the speaker identity of a transformed speech signal as that of the target speaker. The performance of a listener is estimated using speech from different transformation systems as well as untransformed speech. In addition, we describe two tests that address the issue of intelligibility and naturalness, namely an ABX system

comparison test and a comparison category rating (CCR) test of speech quality. We will now describe the test stimuli, the three test designs, and the administration of the tests to listeners.

8.1.1 Stimuli

The following four conditions were used for audio stimuli that are presented during listener testing:

NAT speech from originally recorded sentences, without any spectral transformation.

SET transformed speech output produced by the SET system described in Chapter 6.

HRT transformed speech output produced by the HRT system proposed in Chapter 7.

ALT transformed speech output produced by the ALT system, a previous transformation approach described in Section 7.3.3.

All VT systems were trained on the 40 training sentences (about 5 minutes of speech) of the speech corpus (described in Chapter 5). The transformed speech output sentences were obtained by using the 10 test sentences as input. We used 5 male→male, 5 male→female, 5 female→male, and 5 female→female speaker combinations. Table 6.1 shows the 20 speaker combinations, using speaker identifiers from Table 5.1. Thus, each condition produced 200 sentences, for a total of 800 available test stimuli.

Clearly, the four conditions differ in their spectral contents, given the same sentence and the same target speaker (or original speaker in the case of NAT). With regards to the prosodic content, we aimed at exploring two possibilities. In the first one, we transplant the prosody of the target speaker onto the transformed speech, thus simulating the "optimal" prosodic transformation by an idealized system. This scenario has the benefits of comparing the impact of a joint spectral and prosodic transformation to a prosodic-only transformation ($\text{NAT}_{\text{target}}$), and, at the same time, using unmodified sentences directly from the speech corpus for the NAT condition ($\text{NAT}_{\text{original}}$).

The second possibility avoids the issue of speaker-specific prosody by normalizing the prosodic information of all stimuli, as manifested in pitch, time, and energy. We "aver-

aged" F_0 values, time anchors (as described in Section 5.2), and speech frame energies of the same sentence from different speakers of the same gender to produce a *single* gender-specific evolution of these prosodic descriptors. Then, the speech signals of all speakers of that gender were modified to be in accordance with the "generic" prosodic descriptors. In this manner, pitch, duration, and loudness information contributing to speaker identification was completely removed, while preserving the naturalness of the speech signal. Figure 8.1 shows an example of speech waveforms of the same sentence, spoken by five different male speakers, modified to have generic and identical F_0 values, durations, and energies. Given such sentences, listeners could only use the short-term spectrum to discriminate among speakers. It must be noted that the sentences of the NAT condition ($\mathrm{NAT}_{\mathrm{norm}}$) had a slightly coded quality due to the normalization process. However, they were nearly indistinguishable in quality from the unmodified speech utterances, because only small changes to time and pitch were required due to the special properties of the speech corpus (as described in Section 5.3) [16]. In fact, since care is taken that test listeners are unfamiliar with the speakers of the database, they accept the "normalized" speakers as regular speakers.

Whether stimuli have normalized prosody or target prosody will be clear from the context or indicated by a subscript. Note that the condition $\mathrm{NAT}_{\mathrm{original}}$ denotes the completely unmodified speech signal, as provided by the speech corpus.

8.1.2 Speaker discrimination test

A VT system has successfully transformed the speaker identity of the source speech if the system output can easily be recognized as speech from the target speaker. However, a problem with measuring *speaker recognition* is the assumption of prior knowledge of the voices to be recognized. This is troublesome because it is difficult to control familiarity and to collect data that matches speakers and listeners. It is possible to use unfamiliar voices, but training listeners to recognize a set of unfamiliar voices is subject to memory limitations and the results can be significantly affected by the specific composition of the speaker set [79].

The problems of measuring speaker recognition can be overcome by measuring *speaker*

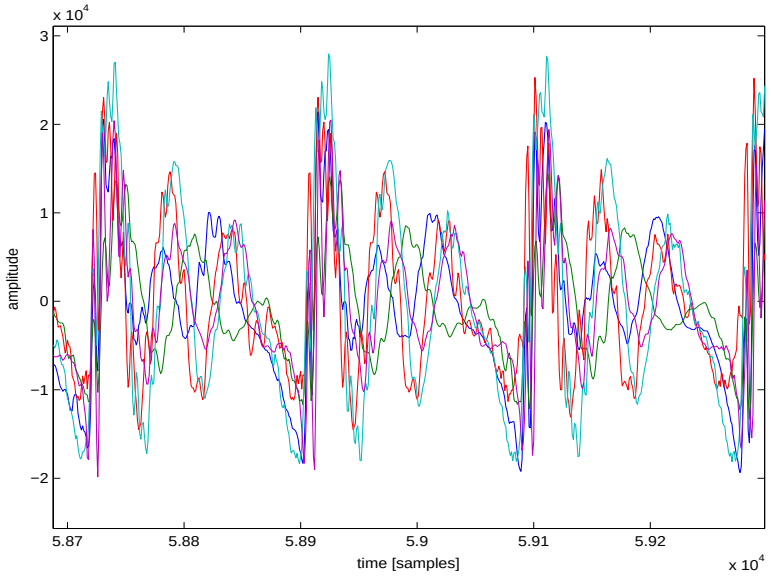


Figure 8.1: A close-up of speech waveforms of the same sentence spoken by five different male speakers. The waveforms were modified to have identical F_0 values, durations, and speech frame energies. Consequently, corresponding pitch epochs of all five speakers start and stop at the same time, and have the same energy.

discrimination or *speaker similarity* of unfamiliar voices [46]. These measures relate to recognition in that the process of listening to a set of test stimuli can be viewed as a very brief training period, followed immediately by a limited-domain recognition process. For example, an ABX test presents three stimuli **A**, **B**, and **X** and listeners are requested to decide whether stimulus **A** or **B** is closer to **X** in terms of speaker identity. As a result, the ABX test compares the speaker similarities or *speaker distances* between **A** and **X**, as well as **B** and **X**, measuring which speaker distance is perceptually smaller. Another type of test presents two stimuli **A** and **B** and requires listeners to indicate the speaker similarity on a numerical scale, in effect measuring the degree of perceived similarity between **A** and **B**. Both the ABX and the AB test have been used by researchers to test their proposed VT systems (see Section 3.2).

One useful property of the ABX test is the elimination of any response bias; in other words, people are equally likely to choose **A** or **B** if indeed **A**=**B**. However, a disadvantage of the ABX test is the fact that it involves 3 stimuli. The results may be affected by memory effects, such as forgetting the first stimulus by the time a participant listens to

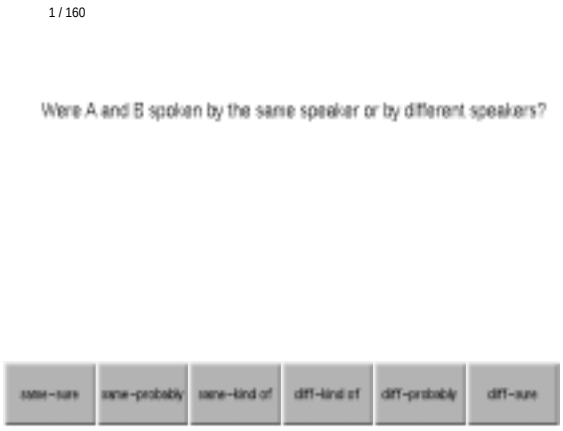


Figure 8.2: Interactive window of the speaker discrimination test.

the third stimulus, or by ordering effects if the test allows for interaction. Perhaps an even greater problem is its low resolution in assessing speaker similarity — a 100% correct test result does not imply that the transformed speaker is indiscriminable from the target speaker.

Because of these shortcomings of the ABX test for measuring speaker recognizability, we choose an AB test with a "same/different" task, similar to some previous approaches [58, 78]. Our perceptual test is in the form of an interactive program that presents listeners with two different sentences in sequence and requests listeners to indicate whether they think the sentences were spoken by either the *same* or by two *different* speakers. The two sentences are produced by either exclusively male or female speakers, because inter-gender confusions rarely occur, as was shown in a previous study [72]. We incorporate a measure of confidence into the same/different decision, resulting in the possible responses shown in Figure 8.2. The power of testing speaker discrimination lies in the fact that listeners are presented with an everyday task, as opposed to having to construct an artificial "similarity distance", as is the case with the ABX test.

In the case of prosody-normalized voices, the sentence pairs compare all four conditions (SET, HRT, ALT, NAT) to the natural stimuli of condition NAT. In this manner, we are able to estimate both transformation performance of conditions SET, HRT, and ALT, as well as discrimination performance of condition NAT (Figure 8.1 shows the four types of condition pairs). The latter approximates an upper bound of a listener's ability to

A/B	B/A (same)	B/A (diff.)	Measuring speaker discriminability of
$SET^{x\rightarrow y}_{norm}$	NAT^{y}_{norm}	NAT^{x}_{norm}	SET system
$ALT^{x\rightarrow y}_{norm}$	NAT^{y}_{norm}	NAT^{x}_{norm}	ALT system
$HRT^{x\rightarrow y}_{norm}$	NAT^{y}_{norm}	NAT^{x}_{norm}	HRT system
NAT^{y}_{norm}	NAT^{y}_{norm}	NAT^{x}_{norm}	untransformed speech

Table 8.1: Four ways of pairing conditions within a stimulus pair and their resulting measurements, using stimuli with normalized prosody. The subscript denotes the source of prosodic content, while the superscript denotes the source of short-term spectral content, with x and y representing source and target speakers, respectively.

A/B	B/A (same)	B/A (diff.)	Measuring of speaker discriminability of
$SET^{x\rightarrow y}_{y}$	NAT^{y}_{y}	NAT^{x}_{x}	SET system
$HRT^{x\rightarrow y}_{y}$	NAT^{y}_{y}	NAT^{x}_{x}	HRT system
NAT^{x}_{y}	NAT^{y}_{y}	NAT^{x}_{x}	prosody-only transformation system
NAT^{y}_{y}	NAT^{y}_{y}	NAT^{x}_{x}	original speech

Table 8.2: Four ways of pairing conditions within a stimulus pair and their resulting measurements, using stimuli with target or original prosody. The subscript denotes the source of prosodic content, while the superscript denotes the source of short-term spectral content, with x and y representing source and target speakers, respectively.

discriminate among the unfamiliar voices of the speech corpus, without being provided any speaker-specific prosodic cues.

In the case of stimuli with target prosody, the sentence pairs compare the original, unmodified sentences of the speech corpus $NAT_{original}$ to the transformation conditions SET_{target}, HRT_{target}, and NAT_{target}. The last condition, NAT_{target}, simulates a transformation system that operates on prosodic features (pitch, timing, energy) exclusively; however, it does so "optimally" by transplanting the desired target speaker's prosody onto the source speaker's. The four types of condition pairs are shown in Table 8.2. This test is designed to reveal differences between the baseline and the proposed system in the presence of speaker-specific prosodic cues. At the same time, we include the prosodic-only transformation for an estimation of the general significance of spectral transformations. In the interest of limiting the duration of the test administration we did not include the ALT system in this configuration.

Table 8.3 shows an example of all stimuli that involve a transformation to or from the

A/B	B/A	same/different
M5→M1	M1	same
F5→M1	M1	same
M1→M2	M1	different
M1→M2	M1	different

Table 8.3: Example of stimulus pairs presented together with speaker M1.

speaker M1 (see Table 5.1 for a description of speaker identifiers) while measuring the speaker discriminability of a transformed speech signal. When testing transformations that result in a speaker other than M1 we use M1 as the source speaker, evaluating whether the transformation was successful in removing the speaker identity of M1. We include all available speaker combinations involving M1, except for M1→F2, since we do not test across genders (see Table 6.1 for reference). However, we then have to include M1→M2 twice, in order to balance the the number of "same" and "different" pairs. This pairing is repeated for the four combinations of condition pairs (shown in Tables 8.1 and 8.2). Thus, testing of each of the 10 speakers requires four trials and each condition pair combination requires 40 trials, for a total of 160 trials for all four condition pairs. Both tests incorporated 5 short breaks and took approximately 30 minutes to complete.

While each listener is presented with the same stimuli over the course of the test, the sequence of stimuli pairs, the condition pairs, and the order of presenting **A** and **B** within trials are randomized. Furthermore, a speaker never repeats the same sentence during the entire test. However, some sentences are used more than once, spoken by different speakers. Finally, we interleave female and male stimuli from trial to trial in order to delay the learning of voice characteristics as much as possible.

8.1.3 System comparison test

The purpose of the system comparison test is to determine which transformation system (in conditions SET, ALT, and HRT) is capable of generating a transformed speech signal that is perceptually most similar to the desired target speaker's natural utterance (in condition NAT), given the *same* sentences spoken by the *same* target speaker. The test is realized as a forced-choice ABX test, where stimuli A and B represent the systems under

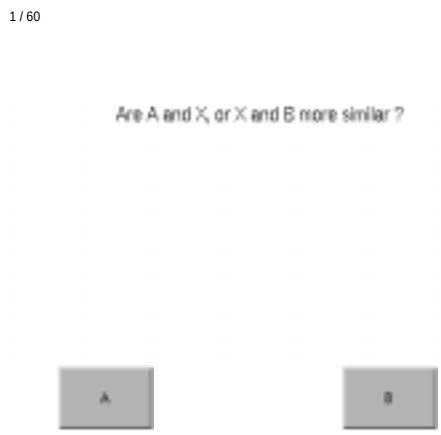


Figure 8.3: Interactive window of the system preference test.

comparison, and X represents the natural utterance. The key difference between this test and the test from Section 8.1.2 is that we are now comparing systems, not speakers.

In each trial, listeners were played stimuli in the order A, X, and then B. They were then asked to indicate whether they thought A or B was more similar to X, using labeled buttons (see Figure 8.3). There were 3 different combinations of systems, each of which was represented by 20 sentences, for a total of 60 trials. The test took 15 minutes to complete.

8.1.4 Speech quality comparison test

To assess the speech quality of the various VT systems in terms of both intelligibility and naturalness, we compared them against each other, as well as against the natural utterances. The test was implemented as a comparison category rating (CCR) test [89], in which listeners are asked to indicate the change in speech quality of two speech samples using a response scale, resulting in a comparison mean opinion score (CMOS).

In this test, all conditions are compared to each other, resulting in 6 different combinations. Each combination was represented by 10 sentences, for a total of 60 trials. In each trial, listeners were played stimuli A and then B. They were asked to select one of five buttons that best matched their response. The buttons were labeled “much worse”, “slightly worse”, “about the same”, “slightly better”, and “much better” (see Figure 8.4). The test took 12 minutes to complete.

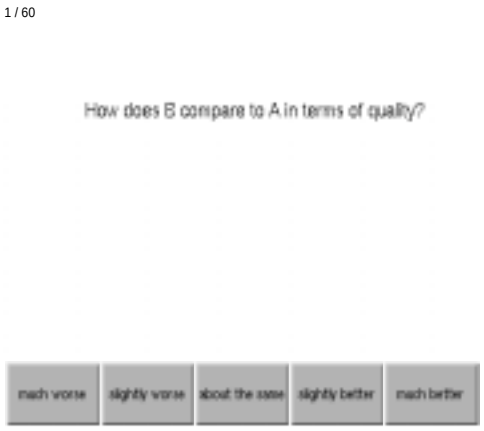


Figure 8.4: Interactive window of the speech quality comparison test.

8.1.5 Administration

Tests were given to listeners who were unfamiliar with the test speakers. Test stimuli were played over headphones. At the beginning of each test, listeners were first presented with two task familiarization trials, to get acquainted with the nature of the speech samples and the test interface. Responses given during familiarization were discarded. It is important to note that the familiarization phase was not in any way a training phase, since all tasks are designed to measure the "everyday" performance of listeners.

The relative loudnesses of stimuli using different sentences are equalized using a 'B' weighting curve [63, p. 56], removing the possibility of discrimination based on different energy levels. Further, we add a small amount of white noise to stimuli (average signal-to-noise ratio better than 50 dB) to mask slightly different noise-floors and other varying recording conditions.

We decided against the possibility of replaying stimuli. Presenting the stimuli only once has the advantages of a simpler interface and guards against effects of ordering and exhaustion of excessively careful listeners. Finally, in the interest of keeping the test short, all stimuli were faded out after 3 seconds, roughly equalizing their length of presentation.

8.2 Perceptual test results

Four non-identical, but overlapping groups of listeners participated as subjects in the listening tests. 16 listeners took part in the speaker discrimination test with normalized prosody, 16 listeners in the speaker discrimination test with target prosody, 10 listeners in the system preference test, and 10 listeners in the speech quality comparison test. We allowed for both native speakers of English and non-native speakers who were fluent in English to participate. We now present and discuss the test results.

8.2.1 Speaker discrimination test with normalized prosody

We assign a similarity score with each one of the possible responses, as shown in Table 8.4. The resulting averages of all listeners' responses are shown in Table 8.5 and their distributions are shown in Figure 8.5. We observe that the scores given to pairs of different speakers are relatively constant across the different conditions. However, the scores given to pairs of sentences from the same speaker are significantly different for each condition, as shown by sign tests and a one-way analysis of variance of the response scores (with $\alpha = 0.01$, H_0: the means of all condition pairs are the same) [30]. The NAT↔NAT condition pair resulted in the lowest score, followed by the HRT↔NAT, SET↔NAT, and ALT↔NAT pairs, in that order. We conclude that the HRT system allows listeners to discriminate between speakers significantly better than the baseline SET system or the alternative ALT system. Specifically, this improvement is realized by rendering a transformed speech signal that is very similar to the target speaker in comparison, so that "same" responses were given with high confidence. This, in turn, means that the speaker recognizability of transformed speech by the HRT system is superior to transformed speech produced by the SET and ALT systems. At the same time, the output of the HRT system is still more confusable than the corresponding natural utterance.

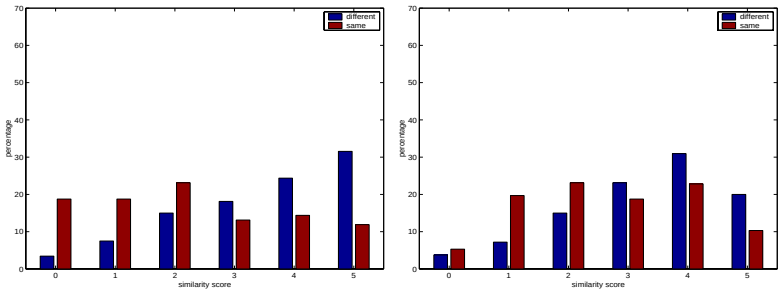
Analyzing the results further, we can interpret the similarity scores as distances between speakers. When distances between speakers are projected onto a lower-dimensional plane, clusters of speakers will form. These clusters indicate that its members are perceived as similar in terms of speaker identity. Using a technique called multi-dimensional scaling

Speakers	Confidence	Similarity score
same	sure	0
same	probably	1
same	kind of	2
different	kind of	3
different	probably	4
different	sure	5

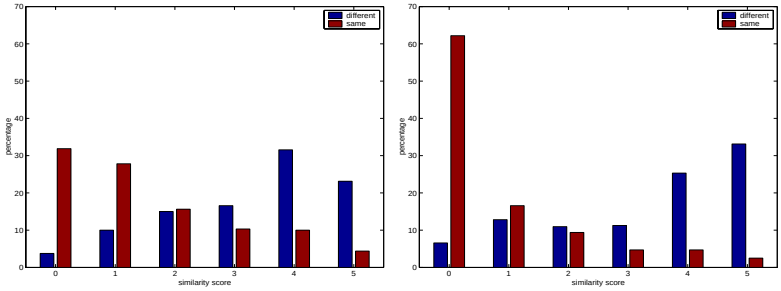
Table 8.4: Perceptual response scale and assigned similarity score.

Condition pairs	SET↔NAT	ALT↔NAT	HRT↔NAT	NAT↔NAT
same	2.21	2.65	1.52	0.81
different	3.47	3.30	3.32	3.35

Table 8.5: Average response scores of the four condition pairs, given that the stimulus pairs were spoken by the same or by two different speakers.



(a) SET-NAT (b) ALT-NAT



(c) HRT-NAT (d) NAT-NAT

Figure 8.5: Distribution of listeners' responses under the four test conditions.

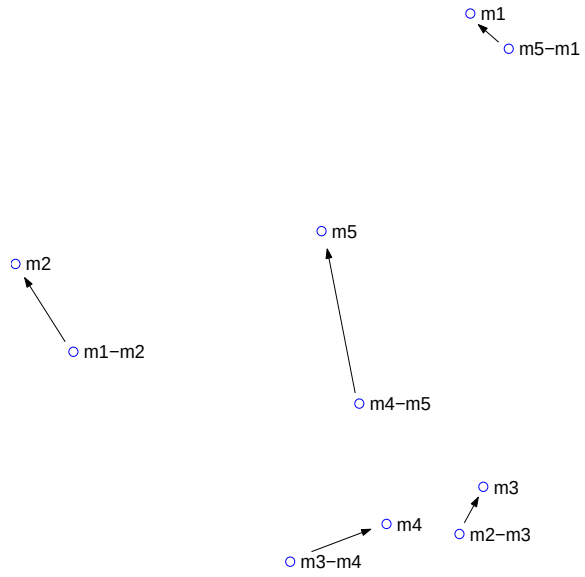


Figure 8.6: Similarity scores projected onto a two-dimensional plane, using a multi-dimensional scaling. For clarity, only stimuli from the condition pair HRT-NAT are displayed. The axes have no particular significance, as this is just one out of many possible configurations.

(MDS) [58], we transform the obtained similarity scores to a two-dimensional picture. Figure 8.6 shows a subset of the results, for clarity only male speakers and transformed speakers using the HRT system are included. Information about the distances between natural speakers are taken from the condition involving natural stimuli. We observe that transformed speakers are perceived to be close to their targets (M1→M2, M5→M1, M2→M3) or significantly closer to their target than their source speaker (M4→M5). The exception is the transformed speaker M3→M4, which is as far from M4 as M4 is from M3. In this case, the system transformed the speaker identity of M3, but not accurately to match that of M4.

Finally, we study the data using the receiver operating characteristic (ROC) curves [13] of the four condition combinations. The ROC curves, shown in Figure 8.7, are derived by plotting the cumulative sum of the percentage of false rejections versus that of correct acceptances, after collapsing the similarity scores to the binary categories "same speaker" or "different speaker". The curves are bounded from below by the straight diagonal $x = y$ which corresponds to listeners responding randomly. The more area is present under the

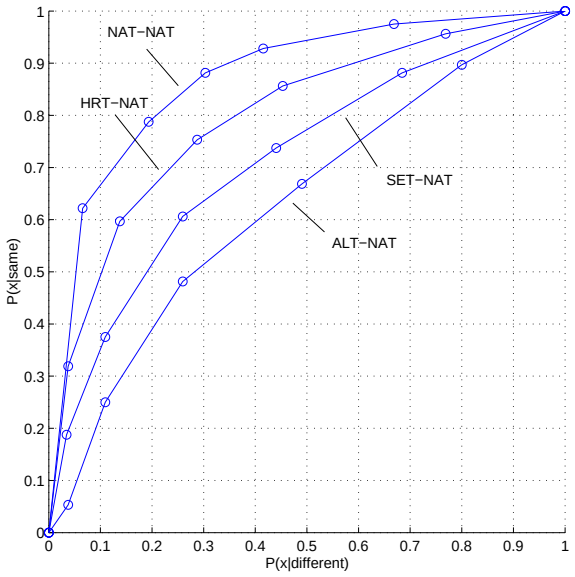


Figure 8.7: Receiver operating characteristic curves for the four condition combinations. The circles are datapoints from direct measurements, connected by straight lines.

ROC curve, the better the discriminability. We observe that the curves are ordered similar to previous results, with the natural condition as the most discriminant, followed by the HRT, SET, and ALT system.

In conclusion, we have shown in several ways that the HRT system significantly outperforms the baseline SET system, and the alternative ALT system. However, its speaker recognizability is still below that of natural speech.

8.2.2 Speaker discrimination test with target prosody

Table 8.6 shows the averages of all listeners' responses to the stimuli involving target or original prosody. Table 8.8 shows the distributions, and Figure 8.9 the corresponding ROC curves. As in the test with normalized prosody, we observe that the scores given to pairs of different speakers are similar across the condition pairs SET↔NAT, HRT↔NAT, and NAT↔NAT; however, the scores given to pairs involving the same speaker are significantly different (as shown by sign tests and a one-way analysis of variance of the response scores with $\alpha = 0.01$, H_0: the means of all condition pairs are the same). In fact, the proposed HRT system gave a significantly lower score than the baseline SET system. We conclude

Condition pairs	$NAT_{target} \leftrightarrow NAT_{original}$	SET↔NAT	HRT↔NAT	NAT↔NAT
same	3.46	2.26	1.40	0.52
different	1.43	3.77	3.73	3.95

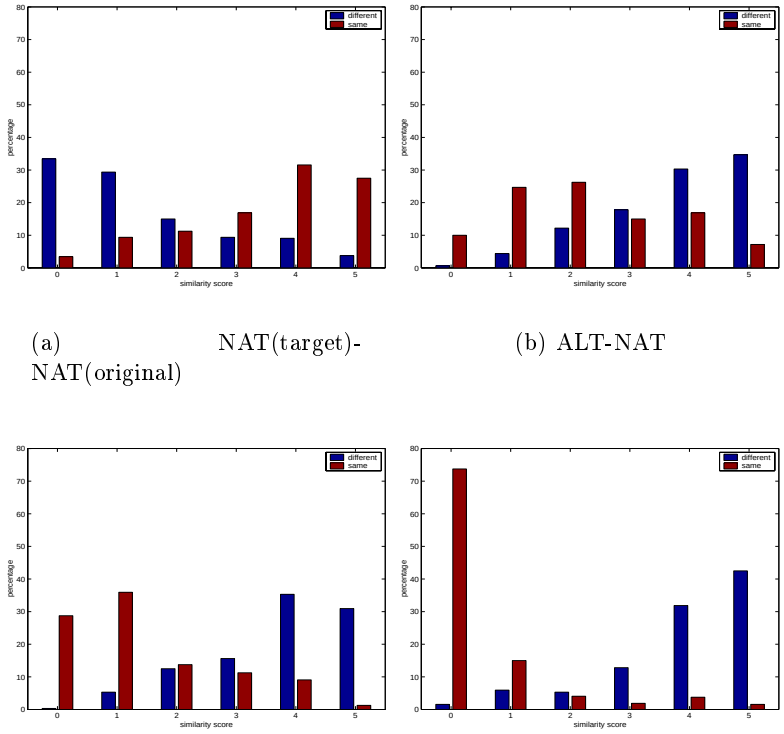
Table 8.6: Average response scores of the four condition pairs, given that the stimulus pairs were spoken by the same or by two different speakers.

that even in the presence of speaker-specific prosodic cues the HRT system allows listeners to discriminate between speakers significantly better than th SET system. This is realized not by making different speakers sound more different, but by rendering the transformed speech to be very similar to the target speaker, as seen by the higher confidence that was given to "same" responses. At the same time, the speaker discriminability of unmodified utterances is still significantly greater.

In the case of the $NAT_{target} \leftrightarrow NAT_{original}$ condition, we have the perverse situation that a high score results from the "same" responses and a low score results from "different" responses. This shows the ineffectiveness of a prosody-only transformation. In other words, modifying the prosody of a source speaker to that of a target speaker does not lead to a speech signal that is recognized as the target speaker. Rather, it is often perceived as a new, third speaker. Equally, modifying the prosody of a speaker does not necessarily lead to the perception of a different speaker from the original speech signal. However, this depends on the prosodic modification strength; for example, a "different" response is more likely for inter-gender transformations. At the same time, these scores also show us the strength of a spectral-only transformation, one only has to exchange the labels of "same" and "different". Of course, the spectral-only transformation is much more successful when performed intra-gender. While prosody may not be the strongest clue for the recognition of speakers in our corpus, prosodic information does help, as is evidenced by the larger spread between responses and generally lower "same" and higher "different" score for the SET↔NAT, HRT↔NAT, and NAT↔NAT condition pairs as compared to Table 8.5.

8.2.3 System comparison test

Figure 8.10 shows the results of the system comparison test, using stimuli with normalized prosody. Of all responses, 95% selected the HRT system over the ALT system, 94.5%



(a) NAT(target)-NAT(original)

(b) ALT-NAT

(c) HRT-NAT

(d) NAT-NAT

Figure 8.8: Distribution of listeners' responses under the four test conditions.

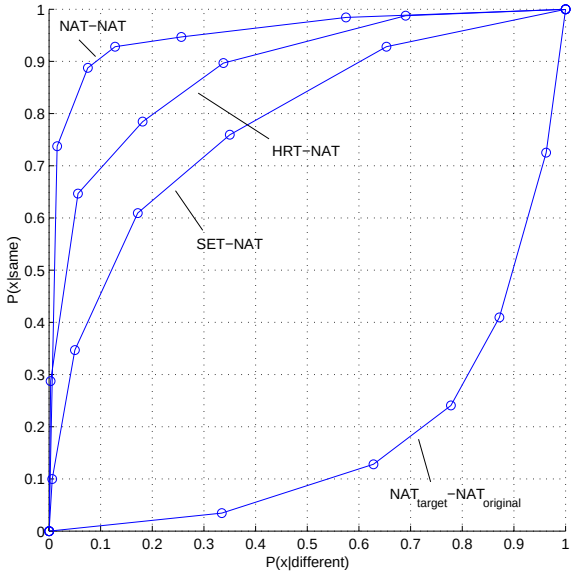


Figure 8.9: Receiver operating characteristic curves for the four condition combinations. The circles are datapoints from direct measurements, connected by straight lines.

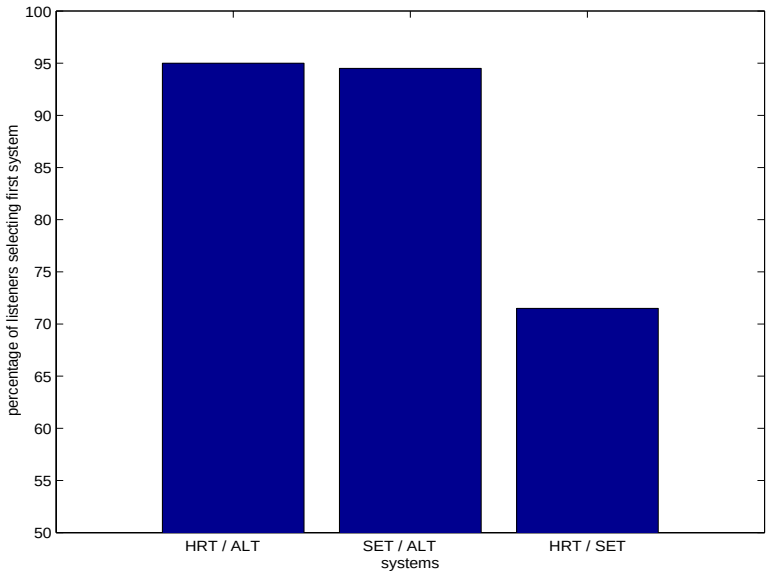


Figure 8.10: Average responses to the three conditions in the system comparison test.

selected the SET system over the ALT system, and 71.5% selected the HRT system over the SET system. In other words, the HRT system was found to be most similar to the natural target speaker utterance, followed by the SET system in second place and the ALT system in last place. The average preference of every listener agreed regarding the direction of preference. Thus, listeners were unanimous in their preference. This test confirms that the HRT system comes perceptually closest to the desired target utterance, compared to the SET and the ALT systems.

8.2.4 Speech quality comparison

Table 8.7 shows the listener response scale and its associated opinion scores. We used stimuli with normalized prosody in this test. The result of averaging the responses of all listeners results in the CMOS score, displayed in Figure 8.11. We observe that condition NAT was perceived to have the best quality, followed by the HRT, SET, and ALT systems. We note that the CMOS score between the HRT system and the natural condition was 0.97, which corresponds to the output of the HRT system as being rated "slightly worse", as compared to natural stimuli, whereas systems SET and ALT were rated at 1.3 and 1.9, respectively. Thus, we have several pieces of evidence, direct and indirect, that the HRT

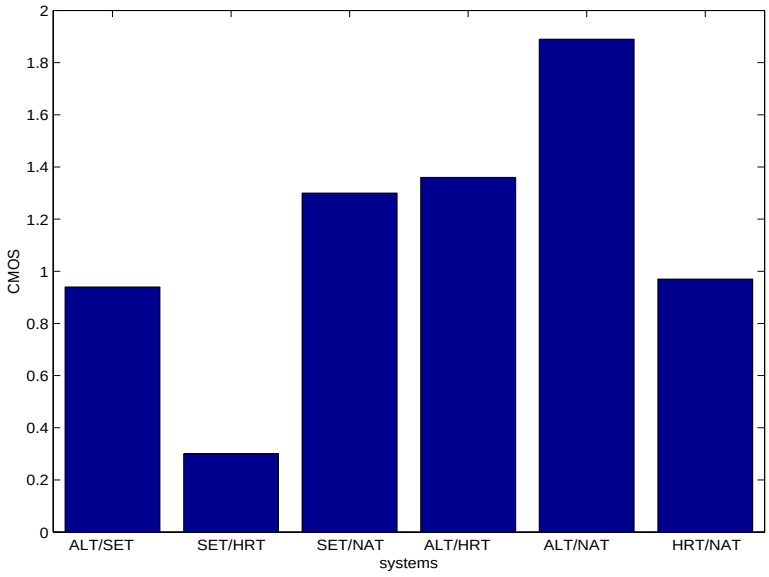
Response	Score
much worse	-2
slightly worse	-1
about the same	0
slightly better	1
much better	2

Table 8.7: Perceptual response scale and assigned opinion score.

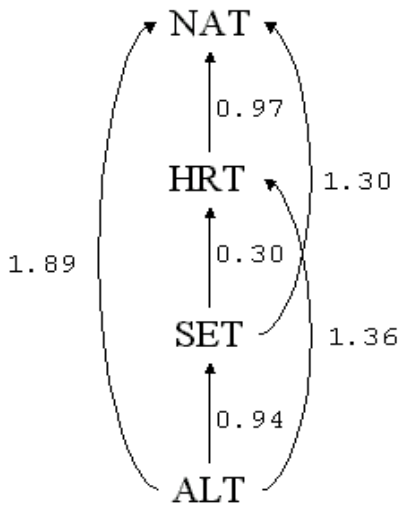
system produces the highest quality speech of the three transformation systems.

8.2.5 Conclusion

We employed four perceptual tests to subjectively evaluate different aspects of the output of three transformation systems. We conclude that not only does the HRT system produce speech that is more similar to the desired target speaker, but it does so with greater speech quality, as compared to our baseline SET system or an alternative approach, the ALT system. However, the output of the HRT system is still significantly below the speaker recognizability and speech quality of natural stimuli. Further, we have shown that spectral transformation is more significant than prosodic transformation, especially for intra-gender transformations. However, since the available speech data were of a constrained nature, we caution against a generalization of this result to other types of speech, for example spontaneous speech.



(a) Comparison mean opinion scores, averaged over all listeners.



(b) Ordering of systems. The arrows point in the direction of improvement.

Figure 8.11: Results of speech quality comparison test.

Chapter 9

Conclusion

9.1 Summary

Speaker identity, the sound of a person's voice, plays an important role in the communication between humans. With speech systems becoming more and more ubiquitous, VT technology offers a number of useful applications. For example, a novice user can adapt a TTS system to speak with a new voice quickly and inexpensively, requiring comparatively little training data. Other possibilities include the areas of speech coding, interpretive telephony, movie dubbing, and assistive technologies for the speaking-impaired. In this dissertation, we have considered new approaches in both the design and the evaluation of VT techniques.

In Chapter 5, we proposed and implemented a new type of speech corpus that is especially suited to research and development of VT systems by consisting of 50 phonetically balanced, naturally time-aligned sentences, spoken by 5 males and 5 females. Consequently, removal of individual prosodic characteristics, such as fundamental pitch and durations, requires only very little processing and results in high-quality speech samples that only differ in their segmental properties, our focus of transformation. These "prosody-normalized" speech samples are used for training VT systems, as well as for evaluating the transformation performance objectively and subjectively.

Chapter 6 introduced the spectral envelope transformation (SET) system, the baseline. The SET system is based on transforming the spectral envelope as represented by the LPC spectrum, using a harmonic sinusoidal model for analysis and synthesis. The transformation function is implemented as a regressive, joint-density Gaussian mixture

model, trained on aligned line spectral frequencies feature vectors by an expectation maximization algorithm. We evaluated the system with various training parameters, using objective measures.

In Chapter 7, we introduced the proposed high resolution transformation (HRT) system, a combination of the SET system and a residual prediction module, which can predict target LPC residuals from transformed LPC spectral envelopes. We described implementation details on training the classifier, constructing the residual codebooks, and synthesizing the transformed speech. In a series of objective evaluations, we tested the performance of the residual prediction module on a speech coding task and on a transformation task.

Because of the severe shortcomings of evaluating VT performance objectively, we propose a subjective evaluation strategy in Chapter 8. Using our speech corpus, we created stimuli that have exactly the same (gender-specific) F_0 values, frame energies, and phonetic durations. We also trained the SET, HRT, and an alternative VT system (ALT) with 20 different speaker combinations, adding their outputs to the pool of available stimuli. In speaker discrimination tests, the HRT system performed significantly better than the SET and ALT system. However, discrimination was not as good as with natural utterances. It was also found that spectral transformation was more significant than prosodic transformation for recognition of the voices in our speech corpus. Similarly, listeners overwhelmingly selected the HRT system over the other systems in a system comparison test. Finally, listeners rated the speech quality of the HRT system as slightly better than the SET system, and much better than the ALT system. However, the quality of natural utterances was considered better than that of any system's transformed speech.

9.2 Conclusion

The ideal VT system generates transformed speech that is of high quality and is easily recognizable as the target speaker. While we have advanced the state-of-the-art towards this goal, our system produces transformed speech with a quality slightly below that of natural speech, and it is still difficult at times to recognize the transformed speaker as the target speaker. However, there are many applications that can greatly benefit from VT

technologies, and we will no doubt see a continuation of research efforts.

Our contributions to the field are the improvement of transformation performance and its measurement. We succeeded by first developing a speech corpus that was specifically designed for training, testing, and evaluating VT systems. We then improved upon the baseline by predicting spectral details from the transformed spectral envelope, instead of modeling and transforming the source speaker's spectral detail. Finally, we were able to study the resulting transformation performance in a series of listening tests.

Most of our results are useful for other types of speech research as well. For example, the residual prediction module can be employed in a TTS system with limited acoustic inventory size (also called small-footprint TTS), since it can scalably compress the number of stored residuals. This compression may also prove useful in the area of speech coding, perhaps by continuously adapting a residual codebook at the receiver to store the current speaker's residuals.

Turning our proposed VT system into an "off-the-shelf" product would necessitate some additional components. Firstly, a highly accurate pitch tracking algorithm has to replace the laryngograph signal as the source of pitch and glottal closure instant information, since a laryngograph device is expensive and cumbersome in its use. Secondly, the source speaker's prosody must be matched to that of the target speaker. This can be achieved using simple scalings or parametric models to modify the average pitch, rate of speech, and loudness of the source speaker.

9.3 Future work

Further improvements in voice transformation performance can be achieved by addressing problems and extending existing solutions of the method described in this dissertation. We will briefly discuss three areas which will benefit greatly from further development:

Speech corpus The proposed method uses short sentences as speech material, which are read by a speaker during the recording of the speech corpus. Unfortunately, renditions of a whole sentence can be quite variable. For example, small pauses, glottal stops, and vowel reductions may occur. It may be that recording only single,

non-ambiguous words leads to better control over these variabilities, especially when automatic data acquisition is necessary.

Speech model A fundamental limitation of representing the short-term speech spectrum by LPC parameters is that the allocation of spectral peaks is performed independently for each frame. Consequently, the same LSF vector component may describe different acoustic events, such as first tracing one formant and then another, resulting in a discontinuous trajectory. For this reason, the synthesis of LSF vectors that have been interpolated or averaged may have problems with the naturalness of the speech signal. A solution to this problem may be found by employing more constrained methods that make use of the dependence between frames. Alternatively, one can use methods that extract information from the speech spectrum that is closely related to speech production, such as formant trajectories [55, 98]. These latter approaches allow for an even more compact representation of speaker identity.

Transformation function The GMM approach can be extended by either incorporating a dimension reduction algorithm, or by exploiting temporal dependencies in the feature stream. In the first case, a principal component analysis (PCA) can be used to first reduce the dimension of the feature space. A mixture representation of this is called a mixture of factor analyzers [27]. In the second case, a HMM can be used, allowing for context-dependent transformations. Since HMMs work well in recognition, they can be expected to work well in the context of VT [57].

Finally, another leap in transformation performance would be possible by modeling and transforming the detailed prosodic characteristics that are specific to a target speaker. This is a very difficult problem, although some beginnings have been made [80, 28].

Bibliography

[1] Abe, M. A segment-based approach to voice conversion. In *Proceedings of ICASSP '91* (Toronto, Canada, May 1991), vol. 2, pp. 765–768.

[2] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. Voice conversion through vector quantization. In *Proceedings of ICASSP '88* (New York, NY, 1988), pp. 655–658.

[3] Abe, M., and Shikano, K. Statistical analysis of bilingual speaker's speech for cross-language voice conversion. *Journal of the Acoustical Society of America vol. 90* , 1 (July 1991), 76–82.

[4] Abe, M., Shikano, K., and Kuwabara, H. Cross-language voice conversion. In *Proceedings of ICASSP '90* (Albuquerque, NM, April 1990), pp. 345–348.

[5] Arslan, L. M. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication vol. 28* , 3 (1999), 211–226.

[6] Arslan, L. M., and Talkin, D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *Proceedings of Eurospeech '97* (Rhodes, Greece, September 1997), vol. 3, pp. 1347–1350.

[7] Arslan, L. M., and Talkin, D. Speaker transformation using sentence HMM based alignments and detailed prosody modification. In *Proceedings of ICASSP '98* (Seattle, WA, May 1998), pp. 289–292.

[8] Atal, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America vol. 55* , 6 (June 1974), 1304–1312.

[9] Barnwell III, T. P., Nayebi, K., and Richardson, C. H. *Speech Coding: A Computer Laboratory Textbook* . Digital Signal Processing Laboratory Series. Georgia Tech, 1996.

[10] Baudoin, G., and Stylianou, Y. On the transformation of the speech spectrum for voice conversion. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), vol. 2, pp. 1405–1408.

[11] Bolt, R. H., Copper, F. S., David, Jr., E. E., Denes, P. B., Pickett, J. M., and Stevens, K. N. Speaker identification by speech spectrograms: some further observations. In *Speech Intelligibility and Speaker Recognition* , M. E. Hawley, Ed., vol. 11 of *Benchmark Papers in Acoustics* . Dowden, Hutchinson and Ross, Inc., New York, 1977, pp. 430–433. Reprinted from Acoust. Soc. Am. J. 54(2), 531–534 (1974).

[12] Childers, D. G. Glottal source modeling for voice conversion. *Speech Communication vol. 16* , 2 (1995), 127–138.

[13] Coombs, C. H., Dawes, R. M., and Tversky, A. *Mathematical Psychology - an elementary introduction* . Prentice-Hall, 1970, ch. 6.

[14] Cronk, A. Linguistic data services manager at the center for spoken language understanding (CSLU). Personal communication, March 2001.

[15] Crosmer, J. R., and Barnwell, T. P. I. A low bit rate segment vocoder based on line spectrum pairs. In *Proceedings of ICASSP '85* (Tempe, Florida, March 1985), pp. 240–243.

[16] Dutoit, T. *High Quality Text-to-Speech Synthesis of the French Language* . PhD thesis, Faculté Polytechnique de Mons, October 1993.

[17] Dutoit, T., and Leich, H. MBR-PSOLA : Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication vol. 13* , 3 (1993), 435–440.

[18] Dutoit, T., Pagel, V., Pierret, N., van der Vreken, O., and Bataille, F. The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996).

[19] Eatock, J. P., and Mason, J. S. A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Proceedings of ICASSP '94* (Adelaide, South Australia, April 1994), vol. 1, pp. 133–135.

[20] Egan, J. P. Articulation testing methods. *Laryngoscope vol. 58* (1948), 955–991.

[21] Etxebarria, B., Hernáez, I., Madariaga, I., Navas, E., Rodríguez, J. C., and Gándara, R. Improving quality in a speech synthesizer based on the MBROLA algorithm. In *Proceedings of Eurospeech '99* (Budapest, Hungary, September 1999), vol. 5, pp. 2299–2302.

[22] Fujisaki, H., and Ljungqvist, M. Proposal and evaluation of models for the glottal source waveform. In *Proceedings of ICASSP '86* (Tokyo, Japan, April 1986), pp. 1605–1608.

[23] Furui, S. Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication vol. 5* (1986), 183–197.

[24] Furui, S., and Akagi, M. Perception of voice individuality and physical correlates. *Trans. Committee on Hearing Res., Acoust. Soc. Japan vol. J66-A* (1985), 311–318.

[25] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgreen, N. L. DARPA TIMIT acoustic-phonetic continuous speech corpus. CD-ROM, National Institute of Standards and Technology, 1990.

[26] Ghahramani, Z., and Jordan, M. I. Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems* , J. D. Cowan, G. Tesauro, and J. Alspector, Eds. Morgan Kaufmann, San Mateo, California, 1994, ch. 6, pp. 120–127.

[27] Ghahramani, Z., and Roweis, S. Probabilistic models for unsupervised learning. In *Neural Information Processing Systems* (Denver, Colarado, December 1999), vol. 12 of *Tutorial* .

[28] Gu, W., Shih, C., and van Santen, P. H. An efficient speaker adaptation method for TTS duration model. In *Proceedings of Eurospeech '99* (Budapest, Hungary, September 1999).

[29] Hertz, J., Krogh, A., and Palmer, R. G. *Introduction to the Theory of Neural Computation* . Addison Wesley, 1991.

[30] Hogg, R. V., and Tanis, E. A. *Probability and Statistical Inference* , 5 ed. Prentice Hall, 1997.

[31] Hosom, J.-P. *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information* . PhD thesis, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Beaverton, OR, USA, May 2000. Published as Technical Report CSE-00-TH-002.

[32] Itakura, F. Line spectrum representation of linear predictive coefficients of speech signals. *Journal of the Acoustical Society of America vol. 57* , S35(A) (1975).

[33] Itoh, K. Perceptual analysis of speaker identity. In *Speech Science and Technology* , S. Saito, Ed. IOS press, 1992, ch. 2.6, pp. 133–145.

[34] Itoh, K., and Saito, S. Effects of acoustical feature parameters on perceptual speaker identity. *Review of the Electrical Communications Laboratories vol. 36* , 1 (1988), 135–141.

[35] Jankowski, Jr., C. R., Quatieri, T., and Reynolds, D. A. Measuring fine structure in speech: Application to speaker identification. In *Proceedings of ICASSP '95* (Detroit, MI, May 1995), vol. 1, pp. 325–328.

[36] Kain, A., and Macon, M. Personalizing a speech synthesizer by voice adaptation. In *Third ESCA/COCOSDA International Speech Synthesis Workshop* (November 1998), pp. 225–230.

[37] Kain, A., and Macon, M. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of ICASSP '98* (Seattle, WA, May 1998), vol. 1, pp. 285–288.

[38] Kain, A., and Macon, M. Text-to-speech voice adaptation from sparse training data. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 7, pp. 2847–2850.

[39] Kain, A., and Macon, M. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In *Proceedings of ICASSP '01* (Salt Lake City, UT, May 2001).

[40] Kain, A., and Stylianou, Y. Stochastic modeling of spectral adjustment for high quality pitch modification. In *Proceedings of ICASSP '00* (Istanbul, Turkey, May 2000).

[41] Kambhatla, N. *Local Models and Gaussian Mixture Models for Statistical Data Processing* . PhD thesis, Oregon Graduate Institute of Science and Technology, January 1996.

[42] Kang, G. S., and Fransen, L. J. Application of line-spectrum pairs to low-bit-rate speech encoders. In *Proceedings of ICASSP '85* (Tempe, Florida, March 1985), pp. 244–247.

[43] Kersta, L. G. Voiceprint identification. In *Speech Intelligibility and Speaker Recognition* , M. E. Hawley, Ed., vol. 11 of *Benchmark Papers in Acoustics* . Dowden, Hutchinson and Ross, Inc., New York, 1977, pp. 425–429. Reprinted from Nature 196(4851), 1253–1257 (1962).

[44] Klatt, D. H., and Klatt, L. C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America vol. 87* , 2 (1990), 820–857.

[45] Kohler, K. J. Parametric control of prosodic variables by symbolic input in TTS synthesis. In *Progress in Speech Synthesis* , J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds. Springer, 1996, ch. 37, pp. 459–476.

[46] Kreiman, J., and Papcun, G. Comparing discrimination and recognition of unfamiliar voices. *Speech Communication vol. 10* (1991), 265–275.

[47] Kroon, P. Evaluation of speech coders. In *Speech Coding and Synthesis* , W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, Amsterdam, Holland, 1995, ch. 13, pp. 467–494.

[48] Kuwabara, H., and Sagisaka, Y. Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communication vol. 16* , 2 (1995), 165–173.

[49] Ladefoged, P., and Ladefoged, J. The ability of listeners to identify voices. *UCLA Working Papers in Phonetics vol. 49* (1980), 43–51.

[50] Laroche, J., Stylianou, Y., and Moulines, E. HNS: Speech modification based on a harmonic+noise model. In *Proceedings of ICASSP '93* (Minneapolis, MN, April 1993), vol. 2, pp. 550–553.

[51] Lavner, Y., Gath, I., and Rosenhouse, J. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication vol. 30* , 1 (2000), 9–26.

[52] Lee, K. S., Youn, D. H., and Cha, I. W. A new voice transformation method based on both linear and nonlinear prediction analysis. In *Proceedings of ICSLP '96* (Philadelphia, PA, October 1996), vol. 3, pp. 1401–1404.

[53] Macon, M., Cronk, A., Wouters, J., and Kain, A. OGIresLPC: Diphone synthesizer using residual-excited linear prediction. Tech. rep., Department of Computer Science, Oregon Graduate Institute of Science and Technology, September 1997.

[54] Macon, M. W. *Speech synthesis based on sinusoidal modeling* . PhD thesis, Georgia Institute of Technology, Atlanta, Georgia, October 1996.

[55] Mannell, R. H. Formant diphone parameter extraction utilising a labelled single-speaker database. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), vol. 5, pp. 2003–2006.

[56] Markel, J. D., and Gray, A. H. *Linear Prediction of Speech* . Springer Verlag, Berlin, 1976.

[57] Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. Voice characteristics conversion for HMM-based speech synthesis system. In *Proceedings of ICASSP '97* (Munich, Germany, April 1997), pp. 1611–1614.

[58] Matsumoto, H., Hiki, S., Sone, T., and Nimura, T. Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics vol. 21* , 5 (October 1973), 428–436.

[59] McAulay, R. J., and Quatieri, T. F. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing vol. 34* , 4 (August 1986), 744–754.

[60] McAulay, R. J., and Quatieri, T. F. Sinusoidal coding. In *Speech Coding and Synthesis* , W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, Amsterdam, Holland, 1995, ch. 4, pp. 121–173.

[61] Mizuno, H., and Abe, M. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech Communication vol. 16* , 2 (1995), 153–164.

[62] Möbius, B. Synthesizing german intonation contours. In *Progress in Speech Synthesis* , J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds. Springer, 1996, ch. 32, pp. 401–415.

[63] Moore, B. C. J. *An Introduction to the Psychology of Hearing* , 4 ed. Academic Press, 1997.

[64] Narendranath, M., Murthy, H. A., Rajendran, S., and Yegnanarayana, B. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication vol. 16* , 2 (1995), 207–216.

[65] Necioğlu, Burhan, F., Clements, M. A., Barnwell III, T. P., and Schmidt-Nielsen, A. Perceptual relevance of objectively measured descriptors for speaker characterization. In *Proceedings of ICASSP '98* (Seattle, WA, May 1998), vol. 2, pp. 869–872.

[66] O'Brien, D., and Monaghan, A. I. C. Concatenative synthesis based on a harmonic model. *IEEE Transactions on Speech and Audio Processing vol. 9* , 1 (January 2001), 11–20.

[67] Olive, J. P., Greenwood, A., and Coleman, J. *Acoustics of American English Speech: A Dynamic Approach* . Springer-Verlag, 1993.

[68] Oppenheim, A. V., and Schafer, R. W. *Discrete-Time Signal Processing* . Signal Processing Series. Prentic Hall, 1989.

[69] Paliwal, K. K. Interpolation properties of linear prediction parametric representations. In *Proceedings of Eurospeech '95* (Madrid, Spain, September 1995), pp. 1029–1032.

[70] Paliwal, K. K., and Atal, B. S. Efficient vector quantization of lpc parameters at 24 bits/frame. *IEEE Transactions on Speech and Audio Processing vol. 1* , 1 (January 1993), 3–14.

[71] Paliwal, K. K., and Kleijn, W. B. Quantization of LPC parameters. In *Speech Coding and Synthesis* , W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, Amsterdam, Holland, 1995, ch. 12, pp. 433–466.

[72] Papamichalis, P. E., and Doddington, G. R. A speaker recognizibility test. In *Proceedings of ICASSP '84* (San Diego, California, 1984), vol. 2, p. 18B.6.

[73] Quatieri, T. F., and McAulay, R. J. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing vol. 40* , 3 (March 1992), 497–510.

[74] Rabiner, L. R., and Schafer, R. W. *Digital Processing of Speech Signals* . Signal Processing Series. Prentice-Hall, 1978.

[75] Reynolds, D. A., and Rose, R. C. Robust test-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing vol. 3* , 1 (January 1995), 72–83.

[76] Rosenberg, A. E. Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America vol. 49* , 2 (1971), 583–590.

[77] Ross, K., and Ostendorf, M. Shattuck-Hufnagel, S. Factors affecting pitch accent placement. In *Proceedings of ICSLP '92* (1992), vol. 1, pp. 365–368.

[78] Schmidt-Nielsen, A., and Brock, D. P. Speaker recognizability testing for voice coders. In *Proceedings of ICASSP '96* (Atlanta, GA, May 1996), vol. 2, pp. 1149–1152.

[79] Schmidt-Nielsen, A., and Stern, K. R. Recognition of previously unfamiliar speakers as a function of narrow-band processing and speaker selection. *Journal of the Acoustical Society of America vol. 79* , 4 (April 1986), 1174–1177.

[80] Shih, C., Gu, W., and van Santen, J. Efficient adaptation of TTS duration model to new speakers. In *Proceedings of ICSLP '98* (1998).

[81] Soong, F. K., and Juang, B.-H. Line spectrum pair (LSP) and speech data compression. In *Proceedings of ICASSP '83* (Boston, MA, 1983), pp. 1.10.1–1.10.4.

[82] Stylianou, Y. Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis. In *Proceedings of ICASSP '99* (1999).

[83] Stylianou, Y. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing vol. 9* , 1 (January 2001), 21–29.

[84] Stylianou, Y., Cappé, O., and Moulines, E. Statistical methods for voice quality transformation. In *Proceedings of Eurospeech '95* (Madrid, Spain, September 1995), pp. 447–450.

[85] Stylianou, Y., Cappé, O., and Moulines, E. Continous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing vol. 6* , 2 (March 1998), 131–142.

[86] Stylianou, Y., Laroche, J., and Moulines, E. High-quality speech modification based on a harmonic+noise model. In *Proceedings of Eurospeech '95* (Madrid, Spain, September 1995), vol. 1, pp. 451–454.

[87] Sutton, S., Cole, R., de Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, M., and Cohen, M. Universal speech tools: the CSLU toolkit. In *Proceedings of ICSLP '98* (Sydney, Australia, December 1998), pp. 3221–3224.

[88] Titze, I. R. *Principles of voice production* . Prentice-Hall, Inc., 1994.

[89] Union, I. T. ITU-T *Recommendation P.800: Methods for subjective determination of transmission quality* , August 1996.

[90] Unser, M., Adroubi, A., and Eden, M. B-spline signal processing. *IEEE Transactions on Speech and Audio Processing vol. 41* , 2 (February 1993), 821–833.

[91] Valbret, H., Moulines, E., and Tubach, J. P. Voice transformation using PSOLA technique. *Speech Communication vol. 11* , 2 (1992), 175–187.

[92] VAN LANCKER, D., KREIMAN, J., AND EMMOREY, K. Familiar voice recognition: patterns and parameters. part 1: Recognition of backward voices. *Journal of Phonetics vol. 13* (1985), 19–38.

[93] VAN LANCKER, D., KREIMAN, J., AND WICKENS, T. D. Familiar voice recognition: patterns and parameters. part 2x: Recognition of rate-altered voices. *Journal of Phonetics vol. 13* (1985), 39–52.

[94] VAN SANTEN, J. P. H., AND BUCHSBAUM, A. L. Methods for optimal text selection. In *Proceedings of Eurospeech '97* (Rhodes, Greece, September 1997), vol. 2, pp. 553–556.

[95] VELDHUIS, R., AND KOHLRAUSCH, A. Waveform coding and auditory masking. In *Speech Coding and Synthesis* , W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, Amsterdam, Holland, 1995, ch. 11, pp. 397–431.

[96] VOIERS, W. D. Toward the development of of practical methods of evaluating speaker recognizibility. In *Proceedings of ICASSP '79* (Washington, DC, 1979), pp. 793–796.

[97] WOUTERS, J. Control of spectral dynamics in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing vol. 9* , 1 (January 2001), 30–38.

[98] YEGNANARAYANA, B., AND VELDHUIS, R. N. J. Extraction of vocal-tract system characteristics from speech signals. *TrSAP vol. 6* , 4 (July 1998), 313–327.

Biographical Note

Alexander Blouke Kain was born on May 29, 1973 in West-Berlin, Germany. He received his Abitur from the John F. Kennedy School in 1992. He received his Bachelor of Arts degree in Computer Science and Mathematics from Rockford College, Illinois, in 1995, graduating summa cum laude. He entered the Oregon Graduate Institute of Science and Technology (OGI) in the fall of 1995, where he worked in the field of speech synthesis as a graduate research assistant. While at OGI, he served as teaching assistant at two short courses on speech synthesis and completed internships with Fluent Speech Technologies, Portland, OR, and AT&T, Florham Park, NJ. He is author on eight conference papers and one U.S. Patent.