CROSS-LINGUAL VOICE CONVERSION

by

Oytun Türk

B.S., Electrical and Electronics Engineering, Bogaziçi University, 2000

M.S., Electrical and Electronics Engineering, Bogaziçi University, 2003

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Electrical and Electronics Engineering

Bogaziçi University

2007

ACKNOWLEDGEMENTS

I would like to dedicate this thesis to my lovely wife, Aylin. I would like to thank you in opening new perspectives, supporting me during the changes, and sharing my life. Without your love, support, and inspiration, this work would be far from being complete. I hope you all shinny and happy days in your quest for the "green fish".

Voice conversion involves interesting scientific research questions and fascinating practical applications. I was first introduced to this topic by my thesis advisor Prof. Dr. Levent Arslan during his lectures at Bogazici University in 1999. It was then things moved so fast and I started working as a research scientist in his young start-up company Sestek Inc. I am proud to be a part of this initiative which has now become the leading company in speech processing technology in Turkey. I would like to thank to Prof. Dr. Levent Arslan for guiding and supporting me through all these years.

I would like to express my gratitude to Dr. Engin Erzin (Koç University) and Dr. Murat Saraçlar who participated in the thesis progress committee and guided me through all this research. I would also like to thank to Dr. Hakan Erdogan (Sabanci University) and Dr. Kerem Harmanci for reading my thesis and participating in the final presentation committee.

I would like to thank to my ex-colleagues at Sestek and friends at the partner company GVZ for sharing a creative working environment while performing cutting-edge research in speech and audio processing technology. Mr. Fred Deutsch, Mr. Michael Hill, and people at Voxonic Inc, New York, deserve special thanks for their belief and support in voice conversion technology. I would like to thank to Mr. Arie Deutsch who I believe is the actual "president of entertainment".

Dr. Baris Bozkurt is a special friend for keeping me up-to-date in speech and audio processing technology as well as in music with his very useful discussions and inspirations. I would like to thank to Dr. Marc Schröder for introducing me into the state-of-the-art text-to-speech synthesis technology, and for great discussions on voice

modification and conversion research in TTS. All my colleagues at DFKI, Germany, also deserve special thanks. Special thanks go to Dr. Ismail Koçak for his great ideas in speech processing applications for speech therapy.

And our families, thanks for your loving and support: Beril and Alp, our mothers Sevil Türk and Gülcihan Çetin, and our fathers Osman Türk and Ahmet Çetin. My father, Osman Türk, was the first person who inspired me in scientific research by allowing me to join his geological explorations when I was a little child.

Great friends Nuri, Devrim, Aykan, Selen, Ece, Özgür, Sebnem, Deniz, Emre, Itir, Hürrem "the cat", and CPD: I wish you all the best. I would like to thank to my friends at Bogazici University and at BUSIM including Ömer Sayli, Ebru Arisoy, and Hazim K. Ekenel. I would also like to thank to everyone at the Bogazici University Electrical and Electronics Engineering Department.

All musician friends, it was a real pleasure playing with you: Hakan, Onur, Arkin (Tayyar); Ismail, Bora, Aysegül, Gülsüm, and Murat (Patron Band).

Finally to everyone who will read this thesis: I hope I was able to present relevant information and inspiring ideas in cross-lingual voice conversion research.

23.08.2007,
Berlin, Germany

ABSTRACT

CROSS-LINGUAL VOICE CONVERSION

Cross-lingual voice conversion refers to the automatic transformation of a source speaker's voice to a target speaker's voice in a language that the target speaker can not speak. It involves a set of statistical analysis, pattern recognition, machine learning, and signal processing techniques. This study focuses on the problems related to cross-lingual voice conversion by discussing open research questions, presenting new methods, and performing comparisons with the state-of-the-art techniques. In the training stage, a Phonetic Hidden Markov Model based automatic segmentation and alignment method is developed for cross-lingual applications which support text-independent and text-dependent modes. Vocal tract transformation function is estimated using weighted speech frame mapping in more detail. Adjusting the weights, similarity to target voice and output quality can be balanced depending on the requirements of the cross-lingual voice conversion application. A context-matching algorithm is developed to reduce the one-to-many mapping problems and enable non-parallel training. Another set of improvements are proposed for prosody transformation including stylistic modeling and transformation of pitch and the speaking rate. A high quality cross-lingual voice conversion database is designed for the evaluation of the proposed methods. The database consists of recordings from bilingual speakers of American English and Turkish. It is employed in objective and subjective evaluations, and in case studies for testing new ideas in cross-lingual voice conversion.

ÖZET

DILLER ARASINDA KONUSMACI DÖNÜSTÜRME

Diller arasinda konusmaci dönüstürmede amaç bir kisinin sesinin hedeflenen bir baska kisinin sesine, hedef konusmacinin konusamadigi bir dilde otomatik olarak dönüstürülmesidir. Dönüsüm için çesitli istatistiksel analiz, örüntü tanima, makine ögrenmesi ve sinyal isleme teknikleri kullanilmaktadir. Bu çalisma, diller arasinda konusmaci dönüstürme konusuna özel problemlere odaklanarak henüz çözülmemis arastirma konularinin belirlenmesini, yeni yöntemlerin gelistirilmesini ve halen kullanilan konusmaci dönüstürme yöntemleriyle karsilastirilmasini amaçlamaktadir. Egitim asamasinda diller arasinda dönüsüm için Fonetik Sakli Markov Modelleri'ne dayali, metinden bagimsiz ve metine bagimli çalisabilen bir otomatik bölütleme ve hizalama yöntemi gelistirilmistir. Agirlikli konusma çerçevelerine dayali esleme ile girtlak dönüsüm fonksiyonu detayli olarak kestirilmektedir. Agirliklarin ayarlanmasiyla girtlak dönüsüm islevindeki ani degisiklikler azaltilarak daha dogal çikti elde edilebilmektedir. Bire karsi çoklu eslestirme sorunlarinin azaltilmasini ve içerikleri farkli kaynak ve hedef egitim veri tabanlarinin kullanilabilmesini saglayan bir baglam kullanan eslestirme yöntemi gelistirilmistir. Gelistirilen diger yeni yöntemler bürünsel dönüsümde hedef sese benzerligin kalitede belirgin azalma olmayacak sekilde arttirilmasini amaçlamaktadir. Bu yöntemler ses perdesi egrisindeki hareketlerin ve konusma hizinin hedef konusmacinin tarzina uygun sekilde dönüstürülmesini kapsamaktadir. Gelistirilen yöntemlerin denenmesi için yüksek kaliteli bir diller arasinda konusmaci dönüstürme veri tabani tasarlanmistir. Veri tabani Amerikan aksanli Ingilizce ve Türkçe konusabilen kisilerden toplanmis, nesnel ve öznel deneylerde kullanilmistir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

a	Linear prediction coefficients vector
c	Cepstrum vector
C	Codebook
d	Distance
D	Duration
f	Frequency in Hertz
f0	Fundamental frequency
fs	Sampling rate
F(w)	Vocal tract transformation filter spectrum
H(w)	Signal spectrum
H(w)	Signal spectrum obtained after weighted averaging
P	Linear prediction order
p(t)	Pitch scaling ratio at time t
s	Subscript for source speaker
t	Subscript for target speaker
u	Line spectral frequency vector
$\hat{u}$	Weighted average of line spectral frequency vectors
v	Normalized weight
w	Angular frequency in radians
x(k)	Discrete-time sequence
?	Codebook entry weighting factor
μ	Mean value
s	Standard deviation
s^2	Variance
ANN	Artificial Neural Network
CD	Cepstral Distance
DTW	Dynamic Time Warping

FD-PSOLA	Frequency Domain Pitch Synchronous Overlap Add Algorithm
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LD	Inverse Harmonic Weighted Line Spectral Frequency Distance
LPC	Linear Prediction Coefficients
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
MFCC	Mel Frequency Cepstral Coefficients
PSOLA	Pitch Synchronous Overlap Add Algorithm
RBFN	Radial Basis Function Network
SD	Spectral Distortion
SD_{fw}	Frequency Weighted Spectral Distortion
SNR	Signal-to-Noise Ratio
SNR_{seg}	Segmental Signal-to-Noise Ratio
SOM	Self-Organizing Map
TD-PSOLA	Time Domain Pitch Synchronous Overlap Add Algorithm
TTS	Text-To-Speech Synthesis
VQ	Vector Quantization
WCeps	Weighted Linear Predictive Cepstral Distance

1. INTRODUCTION

1.1. Definitions

The aim of voice conversion is to transform a source speaker's voice characteristics using signal processing techniques such that the output is identified as the voice of a target speaker. It employs two common stages in general: Training and transformation. The voice conversion system gathers information from the source and target speaker's voices and automatically formulates voice conversion rules in the training stage. For this purpose, training databases from source and target speakers are acoustically analyzed and a mapping between the acoustic spaces of the two speakers is estimated. The transformation stage employs the mapping obtained in the training stage to modify the source voice signal in order to match the characteristics of the target voice. The modification is performed using a set of signal processing algorithms that modify the vocal tract and the prosody characteristics.

Depending on the languages in which the training and test data are available, voice conversion applications can be categorized in two groups. In monolingual voice conversion, the language in which the training data is available and the language in which the target speaker's voice will be generated are identical. The training data is collected in the common language and target speaker's voice is generated in that language. On the contrary, languages of the training and transformation data are different in cross-lingual voice conversion. Cross-lingual voice conversion can be further divided into two sub-categories. In the first sub-category, a bilingual source speaker is available. The training database is collected in the target speaker's language typically in the form of identical utterances from the source and the target speaker. The source speaker records a separate transformation set in the transformation language. The training is performed using the source and the target databases in the training language and the transformation language material is transformed by mapping the transformation data to the training data. Although this type of voice conversion has its own problems, it is generally an easier problem when compared to the second category.

In the second category, a bilingual source speaker is not available. In this case, the training databases are not identical both in terms of content and language.

Formal definitions of important terms related to voice conversion and cross-lingual voice conversion are as follows:

- **Target speaker:** The voice of the speaker that the voice conversion algorithm is aimed to produce at the output.
- **Source speaker:** The speaker whose voice is input to a voice conversion algorithm and is modified to obtain an output that would sound like the target speaker's voice.
- **Training language:** Language in which the training database is collected.
- **Transformation language:** Language of the source speaker's input recordings to be transformed.
- **Monolingual voice conversion:** Voice conversion application in which training and transformation languages are identical.
- **Cross-lingual voice conversion:** Voice conversion application in which training and transformation languages are different.
- **Parallel voice conversion:** Voice conversion application in which the source and the target training material are identical in text content.
- **Non-parallel voice conversion:** Voice conversion application in which the source and the target training material are not identical in text content.

The following definitions related to the proficiency of speakers in one or more languages are taken from http://en.wikipedia.org and will be used in the following sections:

- **L1:** Language acquired during childhood without formal education.
- **L2:** Language learnt at a later age.
- **Monolingual speaker:** Speaker with communicative skills in only one language.

- **Bilingual speaker:** Speaker with communicative skills in two languages.
- **Multilingual speaker:** Speaker with communicative skills in more than one language.
- **Bilingual competence:** Relative proficiency of a bilingual speaker in two languages. Linguists distinguished at least three levels of bilingual competence including coordinate bilingualism, compound bilingualism, and subordinate bilingualism as defined below.
- **Coordinate bilingualism:** Bilingualism in which the linguistic elements in the speaker's mind are all related to their own concepts. This type of bilingual speaker usually belongs to different cultural communities that do not frequently interact (i.e. French-English speaker in Quebec, Canada). The pronunciation patterns of the speaker significantly differ in the two languages.
- **Compound bilingualism:** Bilingualism in which the corresponding linguistic elements in the two languages are mostly attached to the same concept in the brain (i.e. fluent L2 speakers and speakers in minority communities).
- **Subordinate bilingualism:** Bilingualism in which the linguistic elements of one of the languages are only available through the elements of the other language (i.e. beginning level L2 learners).
- **Prosody:** Intonation, rhythm, and vocal stress in speech.
- **Pronunciation:** The way a word or a language is usually spoken or the manner in which someone utters a word.
- **Accent:** A method of pronouncing words common to a certain region. It can also refer to the stress on a certain syllable.

1.2. Applications

Voice conversion provides an efficient mechanism to analyze, model, store, and transform perceived characteristics of speech. It has a number of interesting applications in text-to-speech synthesis, voice quality analysis and transformation, emotion research, speech recognition, and speaker identification. Applications in dubbing, music, computer games, and healthcare industry have also emerged in the recent years. Depending on the application, voice conversion techniques can be used

for modifying or normalizing speaker identity as well as modifying a set of acoustic and prosodic characteristics.

Text-To-Speech Synthesis (TTS) quality has increased by the employment of large databases and unit-selection techniques (Hunt and Black, 1996; Dutoit, 1997). As voice conversion requires less training data (5-10 minutes of voice recordings), it is advantageous to employ voice conversion for creating new TTS voices out of the existing ones (Kain and Macon, 1998; Zhang, et. al., 2001). This approach can be used in both monolingual and multilingual frameworks with the ultimate goal of generating high-quality synthetic speech from any speaker's voice in any language.

Emotional text-to-speech techniques aim to generate speech in different emotional modes such as excited, happy, sad, or angry. The main goal of these techniques is not only to generate speech in a given emotional state but also to have control on the amount of the emotion to be generated. Voice conversion techniques can serve as a useful tool for both goals by transforming a given emotional state into another with control on the continuity and amount of modification in a parametric manner.

As defined in Biology-Online.org, voice quality refers to the component of speech which gives the primary distinction to a given speaker's voice when pitch and loudness are excluded. Some of the descriptions of voice quality are harshness, breathiness, and nasality. In a similar fashion to emotional text-to-speech synthesis, voice conversion techniques can be used in voice quality modification in a controlled manner. Primary results of applying voice conversion techniques in voice quality control in emotional text-to-speech synthesis are discussed in one of the author's publications where voice conversion techniques are employed to interpolate between soft-modal, and modal-loud voice qualities (Turk, et. al. 2005). Modification of different acoustic and prosodic characteristics is also important for emotion research in which voice conversion techniques can be used as a tool to modify the speech signal with least amount of processing distortion. A comparison of emotional prosody in a multilingual setting is presented in (Burkhardt, et. al., 2006) in which we have applied parametric prosody transformation of pitch, duration, and jitter, and performed a listening test to investigate

the perception of different emotion related prosodic states in different languages in the context of emotional text-to-speech synthesis. The results indicated that parametric modification of prosodic parameters can produce part of the intended emotional states in text-to-speech synthesis independent of the language.

Robustness to speaker variations is an important issue for speech and speaker recognition systems. In speech recognition, voice conversion techniques can be used in modifying the speaker identity to match the trained models in a better manner similar to speaker adaptation. In speaker recognition, voice conversion can be used to build a reliable automatic performance testing tool. It can be employed for simulating attacks to the system by transforming the attacker's voice to one of the speaker's voice for which the system is trained to recognize.

With the development of high-quality voice conversion systems, many other applications can be implemented some of which were demonstrated in our previous work. We have reported a demonstration for dubbing movies by employing only several dubbers, generating the voice of famous actresses/actors in a foreign language which they can not speak, and generating the voices of actresses/actors who are not alive (Turk and Arslan, 2002, 2003). Other dubbing applications might be to regenerate the voices of actresses/actors who have lost their voice characteristics due to old age and to perform cross-lingual dubbing for radio broadcasts.

Voice conversion techniques can be used for singing voice transformation, singing voice synthesis, and Karaoke applications in the music industry. In our previous work, we have applied voice conversion to generate rap singers' voices who were originally American English speakers in Spanish and French. This approach can be integrated with singing voice synthesis to synthesize singing voice in different popular voices. In (Turajlic, et. al., 2003), the authors applied formant modification techniques successfully to Karaoke in which the users voice is modified to match the target voice in terms of average spectral characteristics in real-time. Another application field is video games in which it is required to generate voices of virtual characters which can be achieved by modifying existing voices. This may help voice design become an

adaptable part of the game scenario. It will enable the user to create different synthetic voices or even participate with her/his own voice in the game.

Accent transformation and accent normalization can be achieved by employing cross-lingual voice conversion to modify accent characteristics. For accent transformation, it is sufficient to use a source speaker who is native in the transformation language. Therefore, it is possible, for example, to make a native American English speaker speak Turkish in native Turkish accent. For this purpose, a bilingual source speaker who is native in Turkish but can also speak American English is required. However, when such a bilingual source speaker is not available, the problem gets more difficult. In this case, modeling of prosodic and acoustic characteristics in the transformation language from native speakers and applying modification to match those characteristics are required which can be achieved by cross-lingual voice conversion.

1.3. Literature Review

Voice conversion has been a popular topic in speech processing research for the last two decades (Abe, et. al., 1988; Arslan and Talkin, 1997, Arslan, 1999; Moulines and Sagisaka, 1995; Stylianou, et. al., 1998). There are two main stages in voice conversion: training and transformation. The flowcharts in Figure 1.1 and Figure 1.2 show the steps involved in both stages and a non-exhaustive listing of common methods employed.

The training stage involves three steps in general: Acoustic modeling, segmentation and alignment, and acoustic mapping. In the acoustic modeling stage, speaker-specific parameters are extracted from the speech waveform. These parameters describe the short-term and long-term characteristics of the source and target voices. Vocal tract, glottal source (pitch, spectral tilt, open/closed quotient), duration, and energy characteristics convey important speaker-specific information (Furui, 1986; Itoh and Saito, 1982; Kuwabara and Sagisaka, 1995; Matsumoto, et. al., 1973; Necioglu, et. al., 1998). Linear Prediction Coefficients (LPCs) (Makhoul, 1975), Line Spectral

Frequencies (LSFs) (Itakura, 1975a), Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980), formant frequencies and bandwidths (Holmes, et. al., 1990), and Sinusoidal Transform Coding (STC) parameters (McAulay and Quatieri, 1995) can be used for modeling the vocal tract characteristics. There has been considerable amount of work on the analysis, modeling and modification of glottal source characteristics in voice quality research (Childers and Lee, 1991; Childers, 1995; Fant, et. al., 1985). Pitch is one of the most important speaker-specific dimensions among the glottal source characteristics. It can be estimated using the autocorrelation function, average magnitude difference function, Fourier Transform, and harmonic analysis (Rabiner and Schafer, 1978). Dynamic programming is a popular method employed to avoid discontinuities and hence improve the robustness of the pitch detection algorithm (Talkin, 1995).

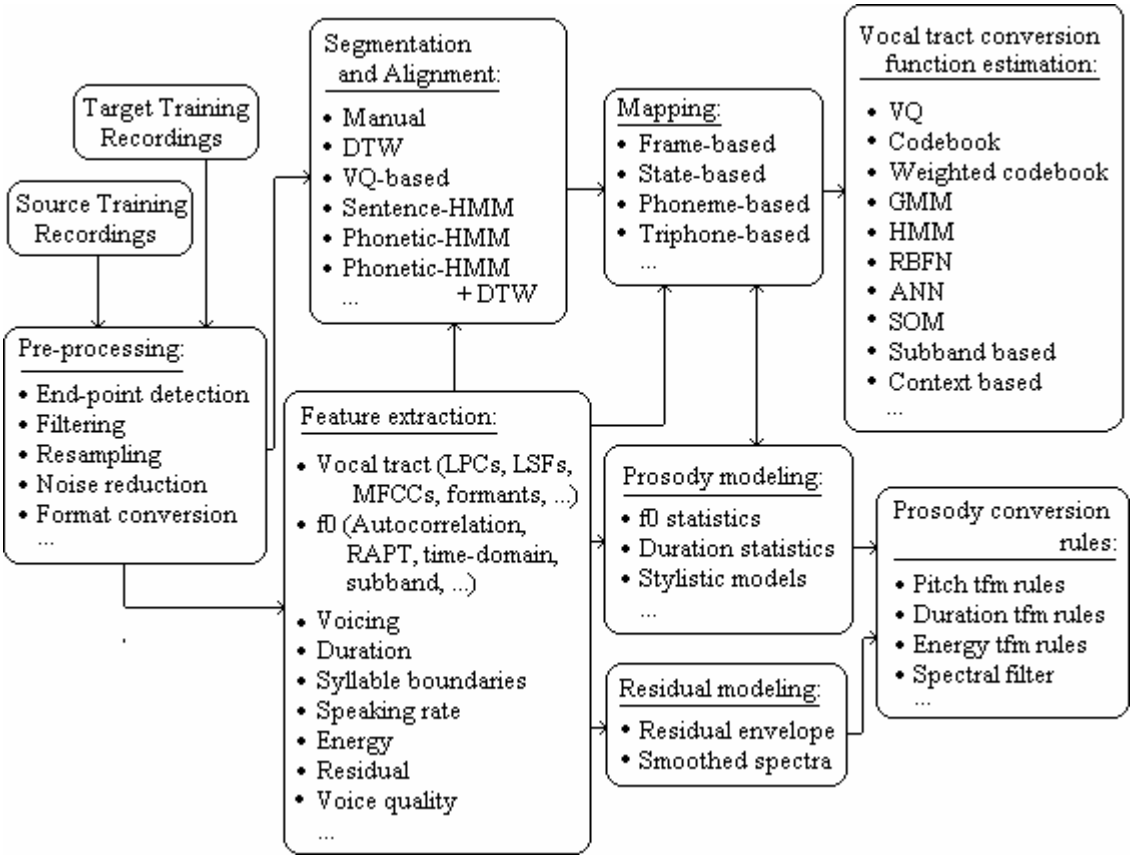

Figure 1.1. General flowchart for voice conversion training

The second step, alignment, is necessary to determine corresponding units in the source and target voices. This is due to the fact that the durations of sound units (i.e. phonemes or sub-phonemes) can be quite different among speakers. It is preferable to employ automatic alignment techniques like Dynamic Time Warping (DTW) (Itakura, 1975b), and Hidden Markov Models (HMMs) (Rabiner, 1989) because manual alignment is time consuming.

The final training step is the estimation of the acoustic mapping function between the source and the target speaker's acoustic spaces using machine learning techniques like vector clustering/quantization (Abe et. al., 1988), codebook mapping (Acero, 1993), weighted codebook mapping (Arslan and Talkin 1997, Arslan 1999), GMMs (Stylianou, et. al., 1998), Radial Basis Function Networks (RBFNs) (Drioli, 1999), Artificial Neural Networks (ANNs) (Narendranath, et. al., 1995), and Self Organizing Maps (SOMs) (Knohl and Rinscheid, 1993). The main distinction between the earlier methods (Abe et. al., 1988 and Acero, 1993) and more recent methods (Arslan and Talkin 1997, Arslan 1999, Stylianou, et. al., 1998) are that smoothing among the mapping units is performed to reduce distortion at frame boundaries. Another distinction of more recent methods is the employment of text and language independent automatic techniques for alignment such as Sentence-HMM and Dynamic Time Warping (DTW).

The transformation stage employs acoustic analysis techniques similar to the acoustic modeling step in training. Once the parameters of the input waveform are determined, voice conversion rules are employed to obtain the corresponding target parameters. Necessary modifications are performed on the input waveform to match the target speaker characteristics. The modifications include transformation of the vocal tract, glottal source, duration, and energy characteristics. The vocal tract characteristics can be transformed using formant modification (Mizuno and Abe, 1995), interpolation of the line spectral frequencies (Arslan, 1999), and sinusoidal modeling techniques (Laroche, et. al., 1993). There exist several methods for pitch modification: Time-Domain Pitch Synchronous Overlap-Add Algorithm (TD-PSOLA) (Moulines and Charpentier, 1990), Frequency-Domain Pitch Synchronous Overlap-Add Algorithm

(FD-PSOLA) (Moulines and Verhelst, 1995), sinusoidal synthesis (Quatieri and McAulay, 1992), and phase vocoding (Flanagan and Golden, 1966).

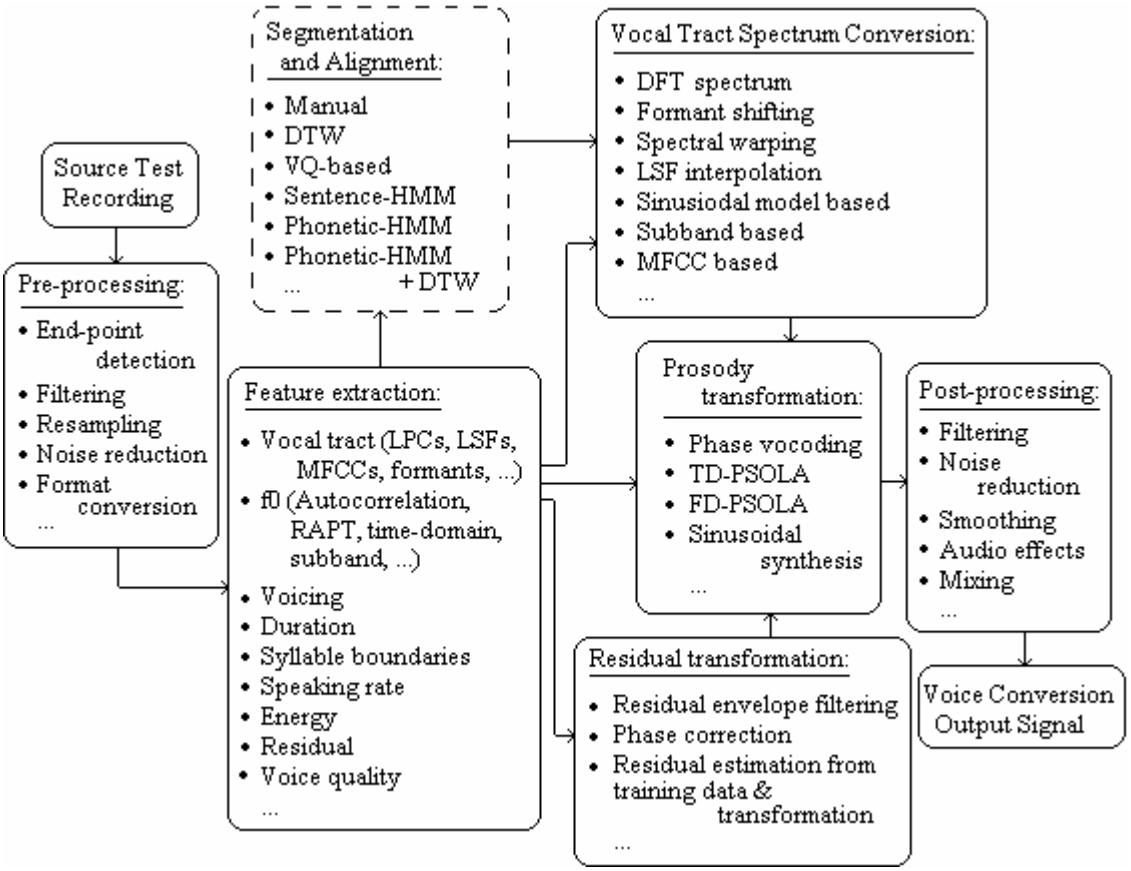

Figure 1.2. General flowchart for voice conversion transformation

Cross-lingual voice conversion is a fairly new topic for voice conversion research. In our previous studies, we have obtained successful results between different languages including English, French, German, Hebrew, Italian, Japanese, Russian, Spanish, and Turkish. Black and Lenzo discuss the possibility to adapt TTS engines to new languages and new voices without recording new databases or recording only a minor amount of new training data (Black and Lenzo, 2004). Latorre, et. al. proposed a new multilingual TTS technique that combines data from multiple monolingual speakers in different languages for creating an average voice (Latorre, et. al., 2005). This average voice is then used for synthesis and transformation to any target speaker's voice. Suendermann and Ney proposed a vocal tract length normalization scheme for

cross-lingual voice conversion applications (Suendermann and Ney, 2003). In (Suendermann and Ney, 2003) and (Duxans and Bonafonte, 2003) the authors focused on the development of voice conversion systems that do not require source and target speakers speaking identical utterances in the training set. Duxans, et. al., proposed two new methods to integrate dynamic and phonetic information in voice conversion and showed that including dynamic information does not improve voice conversion performance significantly as opposed to including phonetic information (Duxans, et. al., 2004). Mashimo, et. al. used GMM-based voice conversion for cross-lingual voice conversion and showed that the performance is comparable to the case of monolingual voice conversion (Mashimo, et. al., 2001).

Comparison of voice conversion performance for monolingual and cross-lingual voice conversion performance is an interesting research question. This question is partly addressed in different studies (Abe, et. al., 1990), (Suendermann, et. al., 2004). Abe and his colleagues report that voice conversion performance is lower when the training and test languages are different (Abe, et. al., 1990). In (Suendermann, et. al., 2004), it was shown that monolingual voice conversion was rated to be better in terms of both similarity to the target voice and quality for both English and Spanish.

1.4. Thesis Outline

This study focuses on the problem of cross-lingual voice conversion. Chapter 2 points out important problems in cross-lingual voice conversion research. The contributions of this study are summarized in correspondence with the common problems. In Chapter 3, the proposed cross-lingual voice conversion framework is introduced. First, the baseline voice conversion algorithm based on weighted codebook mapping is summarized. Then, proposed improvements for cross-lingual voice conversion are described. Chapter 4 focuses on the alignment of cross-lingual voice conversion databases. Alignment is one of the most important pre-processing steps of voice conversion training. A Phonetic-HMM based alignment method is developed and tested. The Phonetic-HMM method can handle both parallel and non-parallel training data in different languages. In Chapter 5, a detailed vocal tract transformation function

estimation procedure is described in order to search for the best matching speech frames in the source and the target training databases to estimate the vocal tract transformation filter in a robust and controllable manner. A context-matching algorithm that is directly linked with the Phonetic-HMM based alignment is developed. Context based search is performed using phonetic labels to reduce one-to-many mapping problems and to enable using non-parallel training databases. Finally, a statistical evaluation of different objective distance measures in the assessment of vocal tract transformation performance is performed. Chapter 6 focuses on the problem of prosody transformation in cross-lingual voice conversion. Stylistic prosody transformation methods are developed and integrated with the conventional prosody transformation algorithms to perform more detailed prosody transformation while keeping the additional distortion at an acceptable level. Algorithms for stylistic transformation of pitch contours and speaking rate are developed. In Chapter 7, the database designed for cross-lingual voice conversion research is described. A subjective test is performed to determine the dependence of cross-lingual voice conversion performance on source speaker proficiency in the training and transformation languages. Then, the performance of the proposed cross-lingual voice conversion algorithm is compared with the baseline algorithm in another subjective test. Finally, different objective measures are compared for the evaluation of vocal tract transformation performance. We show that inverse harmonic weighting based LSF distance is an appropriate choice. Chapter 8 presents a discussion of the results and the future work.

2. PROBLEM STATEMENT AND CONTRIBUTIONS

2.1. Open Problems in Cross-Lingual Voice Conversion Research

Cross-lingual voice conversion share common stages with the monolingual counterpart. These stages include the main training and transformation steps as discussed in detail in Chapter 3. However, there are significant differences considering the content of the training and transformation databases and the language backgrounds of the source and the target speakers. As a result of these differences, new methods should be carefully designed and evaluated for to improve performance in cross-lingual applications.

The performance of voice conversion is generally dependent on the match between the source training and transformation recordings. When there are differences between the two in terms of recording conditions, prosody, articulation or voice quality, the mapping of the source and target training data may be problematic. These problems result in the incorrect estimation of the transformation parameters. Although this problem is common to monolingual voice conversion, it is likely to be more severe for cross-lingual voice conversion. The main reason is the differences between the training and transformation languages. These differences are likely to increase the variation in prosody and articulation as well as in voice quality. It is also harder to find a good match for a given target voice when a bilingual source speaker is needed for cross-lingual voice conversion.

State-of-the-art vocal tract transformation methods may face excessive smoothing and over-fitting problems. A typical example is GMM based voice conversion (Meshabi, et. al., 2007). Excessive smoothing reduces similarity to the target voice and naturalness of the output. On the contrary, over-fitting may result in abrupt changes in the output vocal tract spectrum in successive speech frames resulting in distortion. In case of any problems, the training should be repeated by adjusting GMM parameters. Depending on the amount of available training data, it might not even be possible to estimate the GMM parameters in a robust manner for a sufficiently detailed mapping

from the source acoustic space to the target acoustic space. The main advantage of weighted codebook mapping based approaches is the possibility of handling these problems automatically during transformation by simply adjusting weights and number units used in the weighted estimation process.

In cross-lingual voice conversion, the target does not speak the transformation language at all. In the case of not having a bilingual source speaker, methods are required to map non-parallel training data in a robust and reliable manner. In some cases, it is not possible to have direct access to sufficient amount of good quality target data. An example is performing voice conversion to generate voices of celebrities. In this case, a method is required to estimate the mapping between the source and the target acoustic spaces in a robust and reliable manner.

For practical applications, it might be a good idea to select from an available set of source speakers to ensure better results. However, it might not be possible to have access to sufficient amount of source training material from all source speaker candidates to fully train and test a voice conversion algorithm. In this case, a robust and reliable method of comparing the performance when different source speakers are used is required in an objective manner. There are currently no well-known objective measures that relate well with subjective test results. In the case of cross-lingual voice conversion, objective or subjective testing could only be applied if the target material is available in the transformation language which requires the employment of bilingual speakers in tests. This complicates both the testing procedure and finding subjects in the transformation language.

The ultimate goal of cross-lingual voice conversion is to provide an algorithm for non-speech experts to generate the voice of any target speaker in any language in a fast and reliable manner. A flexible cross-lingual voice conversion tool should be integrated with a robust cross-lingual voice conversion algorithm that can generate the output with minimum amount of manual work. There is generally a trade-off between voice conversion quality and similarity to the target speaker's voice. If the speech signal is transformed with abruptly changing filters over time, the output quality might be so low

that it can not be used for any practical purposes. However, for some applications distortion might be tolerated to increase the similarity to the target voice. For example, in singing voice transformations, part of the distortion in the voice conversion output becomes inaudible when mixed with music. Therefore, the similarity versus quality trade-off should be easily controllable by adjusting a few parameters of the cross-lingual voice conversion algorithm.

Detailed estimation and transformation of acoustic features is required to get sufficiently close to the target speaker's voice. However, signal processing distortion may limit the applicability of severe modifications. For example, increasing or decreasing the pitch in a large amount, i.e. doubling it or halving it, usually results in distortion. In cross-lingual voice conversion, the variation in prosody is likely to be larger. The amount of processing required for pitch transformation increases with increased processing distortion at the output.

Another speaker specific prosodic characteristic is the speaking rate. It may change significantly in different languages. Modeling and transforming the speaking rate might be useful in making the voice conversion output sound closer to the target speaker's voice in cross-lingual applications.

2.2. Contributions

The original contributions of this study can be summarized as follows:

- Robust automatic alignment: An HMM-based robust phonetic aligner is integrated in cross-lingual voice conversion to handle both parallel and non-parallel training database cases. This Phonetic-HMM technique provides reliable alignment in cross-lingual voice conversion. It also enables the weighted speech frame mapping technique as an alternative to codebook mapping in cross-lingual voice conversion to estimate the vocal tract transformation function in more detail.

- Context matching based algorithm for parallel-data: A context-matching based vocal tract transformation function estimation algorithm is developed to reduce the one-to-many mapping problems in the source and target training databases. It uses context information to distinguish among several target candidates for a source phoneme to be converted. This corresponds to extracting information on target speaker's accent and using this information in the transformation step.
- Context matching based algorithm for non-parallel data: The context matching based algorithm developed can also be used for estimating the vocal tract transformation filter when the source and the target training data are not identical in content.
- Weighted frame mapping: A method is developed for directly matching speech frames of the transformation utterance in the training database. This method enables detailed estimation of the vocal tract transformation function. It has the advantage of avoiding excessive smoothing of the vocal tract transfer function which is a typical problem of conventional voice conversion methods. In order to reduce discontinuity at the output, weights of the speech frames can be adjusted parametrically. The trade-off between detailed estimation of the transformation function and continuity can be easily balanced in the transformation stage.
- Stylistic prosody transformation: New methods are developed that enable more detailed prosody transformation while keeping the added distortion at minimum level. The style of the target speaker is modeled in terms of pitch contour movements and speaking rate. The source prosody is transformed using a method that reduces additional distortion due to detailed prosody modification. The contribution of the new prosody transformation techniques is demonstrated in cross-lingual voice conversion examples.
- Donor selection in cross-lingual voice conversion: Tests are performed for investigating the dependence of cross-lingual voice conversion performance on the proficiency of the source speakers in the training and transformation languages.

- Collection of a cross-lingual voice conversion database: A cross-lingual database is designed which consists of phonetically balanced training material in English and transformation material in Turkish. The database is collected from bilingual speakers of American English and Turkish having different levels of proficiency in the two languages: native American English and L2 Turkish speakers, native Turkish and L2 American English speakers, and compound bilingual speakers.
- Evaluation of cross-lingual voice conversion performance: The proposed methods are compared with the baseline weighted codebook mapping based algorithm in objective and subjective tests.

3. CROSS-LINGUAL VOICE CONVERSION

3.1. Introduction

Although part of the monolingual voice conversion techniques can be readily applied to cross-lingual voice conversion, cross-lingual voice conversion possesses its own problems. The differences between the training and transformation languages as well as the increased possibility of having accent differences between the source and the target speakers make cross-lingual voice conversion a more difficult task in general.

In order to develop a robust cross-lingual voice conversion system, four main components are necessary:

- A robust automatic aligner to segment the source and the target training data that possibly have accent differences or even language differences
- A robust mapping method to find the mapping between the acoustic spaces of the source and the target speakers who may have different levels of proficiency in the training language
- A robust transformation function estimator to map the source transformation data in one language to source training data in another language
- A robust acoustic feature transformer capable of generating a speaker's accent, prosody, and style in one language using speech recordings in another language

The automatic aligner handles the segmentation of the source and the target material as well as matching between the transformation and the training material. In the case of cross-lingual voice conversion, both the segmentation and the mapping stages are likely to be more problematic. First of all, if an algorithm that requires a bilingual source speaker is used, the performance depends significantly on the source speaker's proficiency in the training and transformation languages as we show in Chapter 7. If there are significant accent differences between the source and the target

speakers, reliable automatic alignment becomes more difficult to achieve. On the other hand, if an algorithm that can use a non-parallel voice conversion database in two different languages, the problem gets even harder since mapping between the phonemes of the two languages would be necessary.

Estimating the mapping between the source and the target acoustic spaces in a reliable manner is a harder problem in cross-lingual applications. The first reason is the requirement of a bilingual source speaker. Depending on the proficiency of the source speaker in the training and test languages, poor mapping estimates between the source and the target training databases should be eliminated. On the contrary, when reliable mapping is obtained, the voice conversion algorithm should be able to use this information and perform more detailed transformation of the acoustic parameters.

The transformation stage requires mapping of the source material to be transformed with the source training material so that the corresponding target features can be estimated. When the training and the transformation languages are different, mapping becomes more prone to errors. Another problem in cross-lingual applications is the difference between the source and target speaker accents that needs to be compensated for employing robust vocal tract and prosody modification algorithms.

The transformation module should be able to transform the given source acoustic parameters in a robust and reliable manner. There is a trade-off between similarity to target voice and quality, i.e. more aggressive transformations require modification of the source data in a discontinuous manner which in turn reduces the quality. On the other hand, with smoother and higher quality transformations, the similarity to the target speaker might not be sufficient. In the case of cross-lingual voice conversion the prosody in one language may not match the prosody in the other and careful modification of the prosodic features might be required.

3.2. Baseline Voice Conversion Algorithm STASC

Before proceeding with the details of the proposed cross-lingual voice conversion algorithm, it will be useful to summarize the baseline voice conversion algorithm and how it can be used for cross-lingual voice conversion. We have used the "Speaker Transformation Algorithm using Segmental Codebooks – STASC" as the baseline method (Arslan, 1999). STASC is a two-stage codebook mapping method for voice conversion. In the training stage, it determines the corresponding acoustic parameters of the source and target speakers automatically and collects them in codebooks. In the transformation stage, the source speaker acoustic parameters are matched with the source speaker codebook on a frame-by-frame basis and the corresponding target parameters are determined. The transformed utterance is obtained by applying a time-varying filter on the source speaker utterance to match the target speaker's acoustic characteristics. Sections 3.2.1 and 3.2.2 describe the training and transformation stages briefly.

3.2.1. Training

STASC uses the recordings of a set of identical phrases from source and target speakers in the training stage. A left-to-right Hidden Markov Model (HMM) with no skip is trained for each source speaker utterance and both the source and the target speaker utterances are force-aligned with this HMM. The number of states for each utterance is directly proportional to the duration of the utterance. For every 40 milliseconds, a new state is added to the HMM topology. With this model, neither the text nor the language of the utterance needs to be known. This automatic alignment procedure is called as the Sentence-HMM method.

Figure 3.1 shows the flowchart of the STASC training algorithm. In the acoustic feature extraction step, MFCCs are calculated for the source and target speaker utterances. Seven cepstral coefficients derived from a mel-frequency filterbank of 14 bands, log energy, and probability of voicing are combined to form the acoustic feature vector for each frame when the sampling rate is 16 KHz. Delta coefficients are also appended to the feature vector to model temporal variations in the speech signal.

Therefore, the final acoustic feature vector has 18 dimensions. An HMM is initialized using the segmental K-means algorithm and trained using the Baum-Welch algorithm for each source speaker utterance using the acoustic feature vectors obtained. Next, source and target speaker utterances are force-aligned with the corresponding source HMM using the Viterbi algorithm. One may also consider using speaker independent models, such that the sentence HMM is trained from both the source and target utterances at the same time. After Sentence-HMM based alignment, LSF vectors, fundamental frequency values, durations and energy values are calculated in the corresponding source and target HMM states. The state arithmetic means of those acoustic features are computed and stored in source and target speaker codebooks.

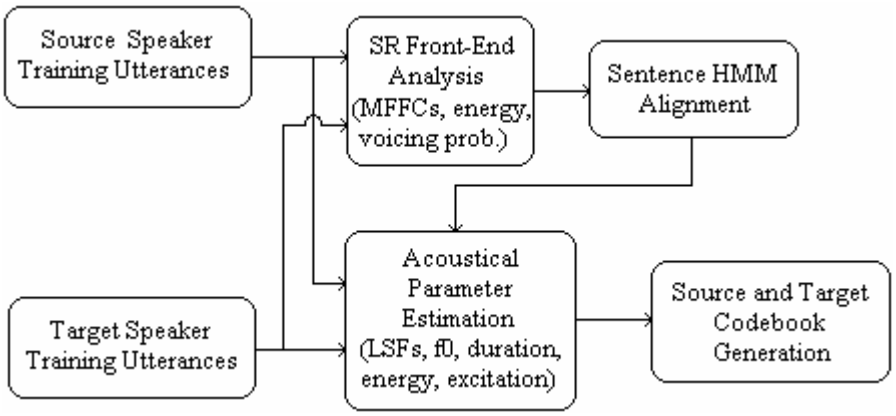

Figure 3.1. Flowchart of the STASC training algorithm.

There are several problems in the training stage of STASC. First of all, Sentence-HMM based alignment is not a robust method when there are differences in prosody, accent, or recording conditions of the source and target training data. The main reason behind this problem is the determination of HMM parameters using only a single source speaker utterance and then force-aligning the corresponding target utterance to that HMM. It is well known that speaker dependent HMMs work better for the specific speaker's voice they are trained for. However, this is not the case when the speaker dependent HMM is used to segment another speaker's recording. The alignment mismatches in the Sentence-HMM method may lead to distortion, reduced similarity to target voice. Even replacement of phonemes with "mutant" phonemes

which are a mixture of two or more phonemes in terms of spectrum can be observed (Turk and Arslan, 2006).

The second problem with Sentence-HMM based training is that it does not convey information on identity of phonemes in the training data and the context in which they exist. Therefore, it is not possible to make use of any phonetic or linguistic information in the transformation stage to improve the match between a given source speech frame to be transformed and the training material. In Chapter 4, we describe a speaker and language independent method to perform alignment in a robust manner using HMMs to cope with this problem.

The third disadvantage of using Sentence-HMM based alignment is the requirement for a parallel training database for the source and the target speaker. Although using a carefully recorded parallel database improves voice conversion performance significantly (Kain, 2001), it is not always possible to employ such a database for cross-lingual voice conversion. A typical example is the case when a monolingual speaker's voice will be transformed to a target speaker's voice who does not speak the language of the source speaker.

Another disadvantage of STASC training is in the modeling and transformation of prosody characteristics. For example, only the mean and the variance of the source speaker f0 are transformed to match the target pitch characteristics. This limits the capabilities of the prosody transformation module in matching the target prosody in terms of style in the transformation stage.

3.2.2. Transformation

Figure 3.2 shows the flowchart for the STASC transformation algorithm. The vocal tract and residual spectra are modified separately. First, linear prediction (LP) analysis for the input frame is performed pitch-synchronously. Next, LP parameters are converted to LSFs. The distance between the source input LSF vector and each LSF vector in the source codebook is computed using Equations 1 and 2:

$$d(m) = \sum_{n=1}^{P} \beta(n)|u(n) - C_s(m,n)| \ for \ m = 1, ..., M \tag{1}$$

$$\beta(n) = \begin{cases} \dfrac{1}{|u(2) - u(1)|} \ for \ n = 1, \\ \dfrac{1}{min(|u(n) - u(n-1)|, |u(n) - u(n+1)|)} \ for \ n = 2, ..., P-1, \\ \dfrac{1}{|u(P) - u(P-1)|} \ for \ n = P \end{cases} \tag{2}$$

where m is the codebook entry index, M is the codebook size, n is the index of LSF vector entries, P is the dimension of LSF vectors (order of LP analysis), u_n is the n^{th} entry of the LSF vector for the input source frame, $C_s(m,n)$ is the n^{th} entry of the m^{th} source codebook LSF vector, $d(m)$ is the weighted distance between the input source frame LSF vector u and the m^{th} source codebook LSF vector. LSF weights, $ß(n)$, are estimated using Equation 2. LSFs with closer values are assigned higher weights since closely spaced LSFs are more likely to correspond to formant frequency locations (Crosmer, 1985). Normalized codebook weights, v_m, are obtained by Equation 3 where using g=1.0 works well in practice.

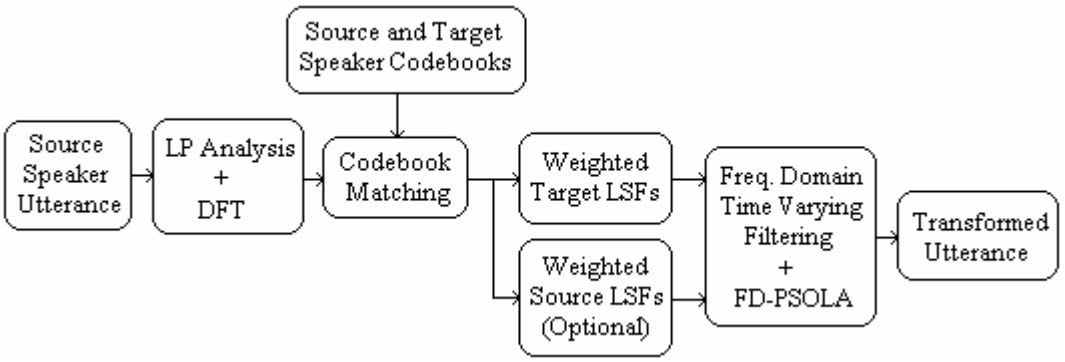

Figure 3.2. Flowchart of the STASC transformation algorithm.

$$v(m) = \frac{e^{-\gamma d(m)}}{\sum_{l=1}^{M} e^{-\gamma d(l)}} \tag{3}$$

Target speaker's vocal tract spectrum is estimated using Equations 4 and 5 where *y(n)* is the n^{th} entry of the estimated target LSF vector. In our notation here and onwards, circumflex represents that the feature is obtained by weighted averaging of codebook entries. In Equation 4, $C_t(m,n)$ is the n^{th} entry of the m^{th} target codebook LSF vector. The estimated target LSF vector *y* is converted to target LP coefficients, $\hat{a}_t$. Target vocal tract spectrum, $H_t(w)$, is estimated using Equation 5 where *w* is the angular frequency in radians and $\hat{a}_t(n)$ is the n^{th} entry of the target LP coefficients vector $\hat{a}_t$.

$$\hat{y}(n) = \sum_{m=1}^{M} v(m) C_t(m,n) \;\; for \;\; n = 1, ..., P \tag{4}$$

$$\hat{H}_t(w) = \left| \frac{1}{1 - \sum_{n=1}^{P} \hat{a}_t(n) e^{-jnw}} \right| \tag{5}$$

$F(w)$, the frequency response of the time varying vocal tract filter for the current frame, is computed using Equation 6.

$$F(w) = \frac{\hat{H}_t(w)}{H_s(w)} \quad \text{or} \quad F(w) = \frac{\hat{H}_t(w)}{\hat{H}_s(w)} \tag{6}$$

Note that the source vocal tract spectrum can be obtained in two different ways to give two different versions of the time varying vocal tract filter:

- Using the input speech frame using the original LP coefficients, $a_s(n)$, as in Equation 7.
- Using the LP coefficients $\hat{a}_s(n)$ that are obtained from $\hat{u}(n)$'s that are estimated by weighted averaging of the source codebook LSF vectors as in Equation 9.

In the latter case, the estimate of the source input frame LSF vector, $\hat{u}$, is obtained as a weighted average of the source codebook LSF vectors using Equation 8. LSF entries, $\hat{u}(n)$, are converted to LP coefficients, $\hat{a}_s(n)$'s and the source speaker vocal tract spectrum, $H_s(w)$, is estimated using Equation 9. In our simulations, we observed that using Equation 9 resulted in more natural and higher quality transformation output. This was mainly because the same type of averaging in both the numerator and denominator of the filter transfer function resulted in a smoother and balanced filter function across frames. However, there has been slight similarity degradation from the target speaker since $H_s(w)$, in this case was not able to filter out all the effects of the source vocal tract.

$$H_s(w) = \left|\frac{1}{1 - \sum_{n=1}^{P} a_s(n)e^{-jnw}}\right| \tag{7}$$

$$\hat{u}(n) = \sum_{m=1}^{M} v(m)C_s(m,n) \;\; for \;\; n = 1, ..., P \tag{8}$$

$$\hat{H}_s(w) = \left|\frac{1}{1 - \sum_{n=1}^{P} \hat{a}_s(n)e^{-jnw}}\right| \tag{9}$$

Prosodic modifications are performed on the excitation signal to match the target characteristics using the FD-PSOLA algorithm (Moulines and Verhelst, 1995). FD-PSOLA algorithm operates on a pitch-synchronous manner and first removes the vocal tract estimate from the spectrum and then applies necessary pitch modifications on the magnitude of the excitation spectrum either by compression or expansion in the frequency domain. Finally, it overlays the original spectrum on top of the modified excitation magnitude spectrum and leaves the original phase spectrum unchanged.

The transformation algorithm of STASC employs a full codebook search strategy when estimating the vocal tract transformation filter from the training database. When there are alignment mismatches in the training data, the whole codebook search strategy is likely to result in one-to-many mapping problems. A typical example might be the case when two source states that are acoustically similar are mapped into target states that are acoustically different in the training stage. In this case, two significantly different target state LSF vectors will be mixed up in the weighted target LSF estimation procedure. Depending on the difference between the selected target states, severe degradation in both output quality and similarity to target voice are likely to occur. This problem can be significantly reduced by using a robust aligner as well as context information to narrow the search space. Another disadvantage of using full codebook search is the memory and processor requirements when the training databases are large. For a full codebook search, the whole data extracted during the training should be loaded to the memory. Then, parameters extracted from each source speech frame to be transformed should be compared with a large number training parameters to find the best matches. For example, when an algorithm that uses parameters extracted from all available source and target training speech frames is used, each codebook can be as large as 500 Megabytes. Therefore, it is required to load part of the training data into memory and search for the best match. However, the search method used in STASC transformation does not directly enable partial codebook search. Another disadvantage of using STASC transformation is that it does not enable detailed prosody transformation while minimizing the additional distortion.

3.3. Cross-Lingual Voice Conversion Algorithm

The proposed cross-lingual voice conversion algorithm consists of two stages as in the baseline case: training and transformation. In the following subsections, we present an outline of the two stages. Implementation details are presented in the corresponding chapters.

3.3.1. Training

Figure 3.3 shows the flowchart of the proposed training algorithm. The training stage starts with the extraction of acoustic parameters from the source and the target speaker training recordings. The vocal tract spectrum is represented in two forms: Mel Frequency Cepstral Coefficients (MFCCs) for the alignment stage and line spectral frequencies (LSFs) for the transformation stage. A fixed window size of 20 ms is used with a skip size of 10 ms for MFCC and LSF analyses. Pitch contours are extracted using the RAPT algorithm (Talkin, 1995). Voicing, f0 statistics (mean and variance), energy as well as stylistic pitch and speaking rate transformation parameters as described in Chapter 6 are computed. Then, all source and target recordings are segmented using Phonetic-HMM based segmentation. The segmentation can be performed in two ways: Text-independent and text-dependent. If the text transcription of the training material is not available, phoneme recognition is performed on the source recordings first. If the training database is parallel, the target recordings are force-aligned to the corresponding source phoneme sequences. Otherwise, target recordings are segmented using phoneme recognition. After the alignment, additional features to be used in transformation are extracted including phoneme durations and stylistic duration transformation parameters. If the training databases are parallel, the acoustic features extracted are paired on a frame-by-frame basis using the alignment information. The phonetic context of each aligned acoustic feature pair is determined from the labels. The resulting acoustic feature vectors and context information are saved into two binary speaker model files for the source and the target speaker separately. For non-parallel training databases, no pairing is performed and all extracted parameters and information on their phonetic context are saved. In this case, the mapping of the source and the target acoustic spaces is performed in the transformation stage by context matching.

A comparison of the baseline and the proposed training algorithms highlight the following important differences:

- Phonetic-HMM based alignment is used instead of Sentence-HMM based alignment. This results in better alignment since the HMM parameters are estimated from a large speech corpora from many speakers. It also enables the employment of phonetic context information and non-parallel databases for cross-lingual voice conversion.
- It is possible to perform forced-alignment to text if the text transcription of the training databases is available.
- The vocal tract spectrum parameters are kept on a frame-by-frame basis instead of using the state averaging method of the baseline algorithm. This helps to perform more detailed transformation of the vocal tract spectrum.
- Stylistic prosody parameters are extracted along with the average prosody transformation parameters used in the baseline method.

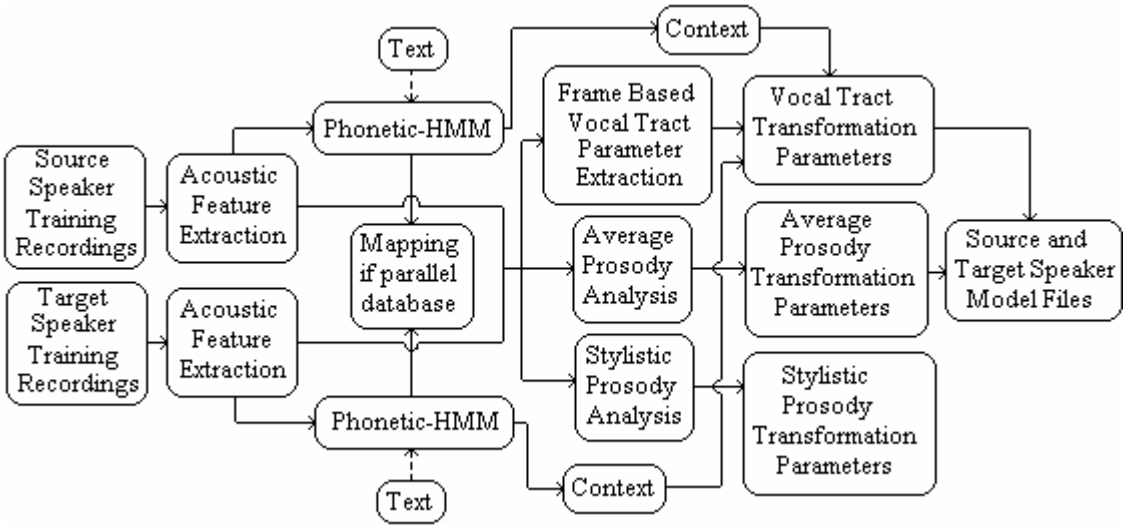

Figure 3.3. Flowchart of the cross-lingual training algorithm

3.3.2. Transformation

Figure 3.4 shows the flowchart of the transformation algorithm. First, a set of acoustic features are extracted including MFCCs, LSFs, f0, and energy. Phonetic-HMM based alignment is performed either in text-independent mode using phoneme

recognition or in text-dependent mode using forced-alignment to a given phonetic transcription. Statistical averages of pitch (mean and variance) and stylistic pitch, duration and speaking rate are computed. The source LSF vectors are matched with the source training LSF vectors using inverse LSF weighting as described in Section 3.2.2. The matching is performed either on a subset of the source training LSF vectors by considering the context-match as described in Chapter 5 or on the full set as described in Section 3.2.2. For cross-lingual voice conversion applications with parallel databases, we prefer the second method as it results in better quality output. However, for the non-parallel case, the only choice is to employ the context-matching based technique. The matching procedure outputs a set of target LSF vectors and corresponding weights in order to estimate a target vocal tract spectrum for the current source input speech frame. The prosody transformation module take the mean and the variance of the source and the target f0 values, average source and target phoneme durations for the current context as well as stylistic prosody features to transform the prosody using the FD-PSOLA algorithm. The vocal tract is transformed by filtering the source speaker's residual spectrum with the estimated target vocal tract spectrum. The vocal tract and prosody transformation can be simultaneously performed using the FD-PSOLA algorithm which eventually produces the cross-lingual voice conversion output signal.

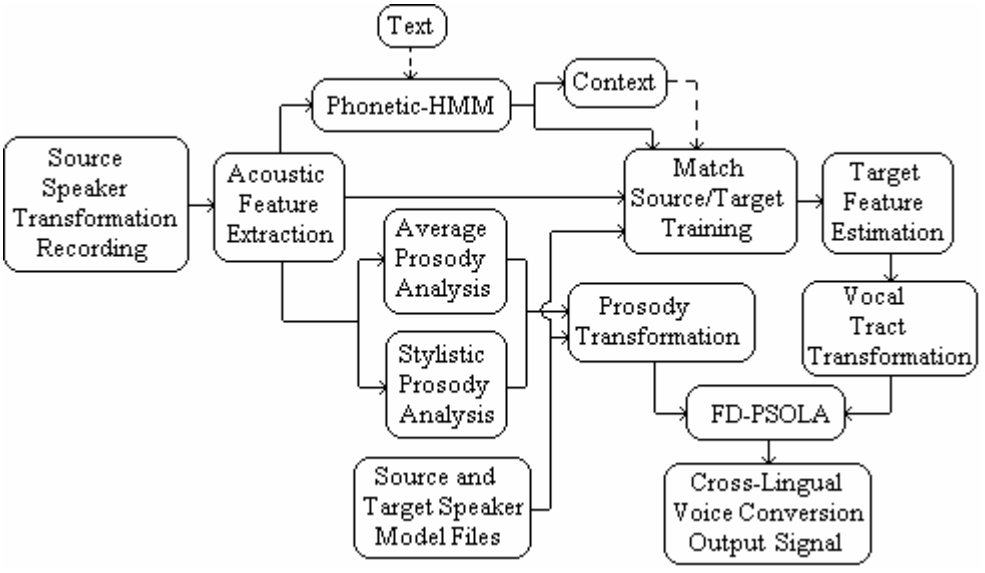

Figure 3.4. Flowchart of the cross-lingual transformation algorithm

The major differences between the proposed transformation algorithm and the baseline algorithm are:

- The vocal tract transformation function is estimated directly from the speech frames in the source and target training database instead of state averages as in STASC. This helps to perform more detailed vocal tract transformation.
- Non-parallel databases can be used for estimating the target vocal tract spectrum in the transformation stage.
- Using the context information search space can be restricted in both parallel and non-parallel databases to reduce one-to-many mapping problems as well as to minimize memory and processor requirements.
- In addition to transforming the mean and the variance of f0 and average speaking rate to match the target characteristics, stylistic prosody characteristics can be transformed including movements of pitch contours, global and local speaking rate.

4. SEGMENTATION AND ALIGNMENT

4.1. Introduction

The aim of speech signal segmentation is to find boundaries of acoustic or phonetic events along the time axis. The resulting boundaries include but are not restricted to sentence, word, phoneme, or sub-phoneme boundaries. In voice conversion applications, the estimated boundaries should be on the level of phonemes or even smaller units in order to enable the estimation of a voice conversion function in sufficient detail. Alignment is a special kind of segmentation in which the segmentation module is given more detailed information about the nature of the acoustic and phonetic events that one is interested in. For example, alignment of a speech signal to a given phoneme sequence involves a segmentation stage to search for the boundaries of the given phonemes within the signal. In the case of voice conversion with parallel training databases, segmentation is required in the training stage to segment either the source or the target utterance and then align the other speaker's utterance using the segmented acoustic events. Non-parallel training requires segmentation of the source and the target training data. Alignment is then performed by clustering the segments and finding a mapping among them.

Our previous experience from a large number of monolingual and cross-lingual transformations shows that problems in the alignment may result in non-reliable and poor quality voice conversion output. Conventional voice conversion algorithms perform alignment using techniques that rely on only speaker dependent information. As an example, Dynamic Time Warping (DTW) was commonly used in codebook mapping (Abe, et. al., 1988) and GMM based voice conversion algorithms (Stylianou, et. al., 1998). DTW finds a minimum error alignment path given a set of acoustic features and a distance measure between those acoustic features. As the optimal path is constructed by using only information from a single utterance, the alignment performance is significantly dependent on variations in prosody, accent, voice quality, and recording conditions. Sentence-HMM based alignment is another alternative for aligning the source and the target training utterances automatically (Arslan, 1999). It is

also non-robust to differences in the source and the target recordings as the model parameters are extracted from only the pair of source and target utterances to be aligned.

When cross-lingual applications are considered, segmentation and alignment becomes a more difficult task in general. First of all, the source and the target training material are likely to contain more accent variation when a bilingual source speaker is used. Non-parallel training techniques enable using monolingual source speakers for cross-lingual applications. In this case, the source and target training materials are collected in different languages. Therefore, a segmentation and alignment module that can handle both languages is required. In the transformation stage, source transformation recording in one language should be aligned with the source training recordings in another language. Therefore the phonemes in one language should be mapped to those in another language to estimate the transformation parameters from the training material. It is preferred to employ an automatic mapping process. Otherwise, it will be significantly difficult to specify the mapping of phonemes among the new languages and existing ones manually.

In order to improve alignment performance and to enable the employment of phonetic information in the cross-lingual voice conversion process, we propose to use Phonetic-HMMs. Phonetic-HMM based segmentation and alignment uses models trained from a large speaker independent database. It has the following advantages over conventional alignment methods:

- Phonetic-HMM parameters are estimated from a large number of speakers having different accent and prosody characteristics. Therefore, the models cover a significantly wider range of prosody, accent, and voice quality characteristics as compared to using a single utterance pair for estimating the HMM parameters.
- Phonetic-HMM can be trained in a robust manner using well-known techniques from speech recognition research (Rabiner, 1989), (Woodland, et. al. 1994).

- Phonetic-HMM can be used both for phoneme recognition and forced-alignment to a given text transcription.
- Using silence models trained from a large amount of acoustic data, end-point detection can be performed in a more robust manner. This helps to improve alignment performance especially when the silence in the beginning or at the end of the source and target training files are significantly different.
- Employment of Phonetic-HMM for phoneme recognition enables using non-parallel cross-lingual voice conversion training databases.
- Databases in different languages can be combined to estimate the parameters of a multilingual Phonetic-HMM which might improve segmentation and alignment performance in cross-lingual voice conversion. This property also solves the problem of mapping the phonemes in one language to another language in cross-lingual voice conversion.

The cross-lingual voice conversion database collected in this study consists of native American English target speakers and bilingual Turkish source speakers. Therefore, an overview of the phoneme inventories of both languages is necessary at this point. There are different phonetic alphabets designed for American English including TIMIT (Garofolo, et. al., 1990) and SAMPA (Wells, 1997). SAMPA is also available for a large number of languages including Turkish which makes it a natural choice for our study. A commonly used version of the American English SAMPA set consists of 44 phonemes (24 consonants, 17 vowels, and 3 silence/pause symbols). The Turkish SAMPA set includes 37 phonemes (26 consonants, 8 vowels, 3 silence/pause symbols). 28 phonemes are common in the two languages (22 consonants, 4 vowels, and 3 silence/pause symbols). There are 15 distinct American English phonemes (2 consonants and 13 vowels) that do not exist in the Turkish phoneme set. There are 8 distinct Turkish phonemes (4 consonants and 4 vowels) that do not exist in the American English phoneme set. The tables in Appendix B show lists of common and distinct phonemes of American English and Turkish SAMPA phoneme sets along with exemplar words and transcriptions. We have used the TIMIT phoneme set in training Phonetic-HMMs in American English. In the case of multi-lingual Phonetic-HMM training, a Phonetic-HMM was trained in one language, the database in the second

language is segmented by phoneme recognition using the initial HMM, and the HMM models are updated using the data from the second language.

4.2. Method

Phonetic-HMMs are trained using two large multi-speaker databases. The first database consists of recordings from 200 native Turkish speakers (95 female, 105 male). It was collected at Sabanci University for large-vocabulary speech recognition purposes (Erdogan, et. al., 2005). For each speaker, approximately 100 utterances are recorded where the utterances are selected from a phonetically balanced set. The second database is the training set of the American English TIMIT corpus (Garofolo, et. al., 1990). Both databases were recorded in 16 KHz, 16 bits PCM format. The Hidden Markov Toolkit (HTK) is used for training and performing segmentation and alignment (Woodland, et. al., 1994). The acoustic feature vectors used in HMM training were 26-dimensional: 12 MFCCs, energy, and the corresponding delta parameters. Each phoneme was modeled using a 3-state HMM with a number of Gaussian mixture components for each state. Different numbers of Gaussian mixture components in the range four to twelve were tested. The number of mixtures that resulted in best alignment performance was used in the final evaluations. Phoneme recognition using the Viterbi algorithm is used for segmentation. Viterbi algorithm is also used for the alignment of a speech signal to a given phoneme sequence. Each HMM was appended an entry and an exit state to enable transitions from one model to another using the Viterbi algorithm.

Depending on whether the text transcription is available for given training and transformation databases, Phonetic-HMM based segmentation and alignment module can be used in three ways for cross-lingual voice conversion:

- Text-independent mode, parallel training: The source training utterances are segmented using phoneme recognition. The target training utterances are force-aligned to the corresponding source phoneme sequences. In transformation, if context-matching will be used, the source utterance to be

transformed is also segmented by phoneme recognition and the context-matching algorithm described in Chapter 5 is employed for estimating the target parameters. Otherwise, full search as in the case of STASC transformation is performed by considering all available source and target training data.

- Text-independent mode, non-parallel training: Both source and target training databases are segmented using phoneme recognition. In the transformation stage, the source recording is segmented by phoneme recognition and the closest matches in the source and the target training data are determined using the context matching algorithm described in Chapter 5.
- Text-dependent mode, parallel training: The source and the target training utterances are force-aligned to the corresponding phoneme sequences extracted from the corresponding text transcriptions. A decision tree based letter-to-phoneme module for American English is employed for converting the text into the corresponding phoneme sequence. The module was trained using the CMU Pronouncing Dictionary. The decision tree training module is available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict. As it was not possible to perform context matching using phonetic transcriptions in two different languages, full acoustic search as in STASC transformation is employed.

In fact, there was a fourth possibility to use text-dependent mode in non-parallel databases. However, it is not directly possible to perform automatic context-matching between phonemes of two languages that have distinct phonemes. Therefore, we excluded this possibility from the tests.

4.3. Evaluations

4.3.1. Alignment Performance

An objective distance measure is developed in order to compare the performance in the alignment of source and target speaker training utterances using different

Phonetic-HMM architectures. To compare an alignment for a given source-target utterance pair with a reference alignment, we need the corresponding source and target speech frame indices using the two alignments and a measure to evaluate how target indices differ in one alignment as compared to the other.

Let us perform an indexing of the source and target speech frames in an utterance recording as (1, 2, ..., i_s , ..., I_s) and (1, 2, ..., i_t, ..., I_t) respectively. Given the source speech frame index i_s, the corresponding target speech frame index $M(i_s)$ is determined by linear mapping:

$$M(i_s) = \frac{(n - a_s)}{(b_s - a_s + 1)}(b_t - a_t + 1) + a_t \tag{10}$$

where a_s and b_s are the first and the last speech frame indices in a source label respectively. a_t and b_t are the first and the last speech frame indices in the corresponding target label respectively. Similarly, $M_{ref}(i_s)$, the corresponding target speech frame indices for each source speech frame are found using the reference alignment. Figure 4.1 shows an example of the speech frame index mapping process.

The absolute difference between the target frame index correspondence using the alignment method and the reference alignment is used as a measure of the similarity between the two alignment patterns:

$$D(i_s) = |M_{ref}(i_s) - M(i_s)|.ss \tag{11}$$

where $M_{ref}(i_s)$ is the target speech frame index corresponding to source speech frame index n using the reference alignment and ss is the skip size in seconds. $d(i_s)$ is the mismatch in seconds in the aligned target frame according to the reference and the given alignments. When the given alignment matches the reference alignment perfectly, $d(i_s)$ will be zero. Otherwise, $d(i_s)$ will be a positive real number that increases as the mismatch between the two alignments increases. The mean of the alignment mismatch

score corresponds to the average shift of a given label boundary as compared to the reference alignment. The manual labels served as the reference alignments against which the outputs of alternative alignment methods are compared. Sentence-HMM based alignment results are also included in the tests for performance comparison.

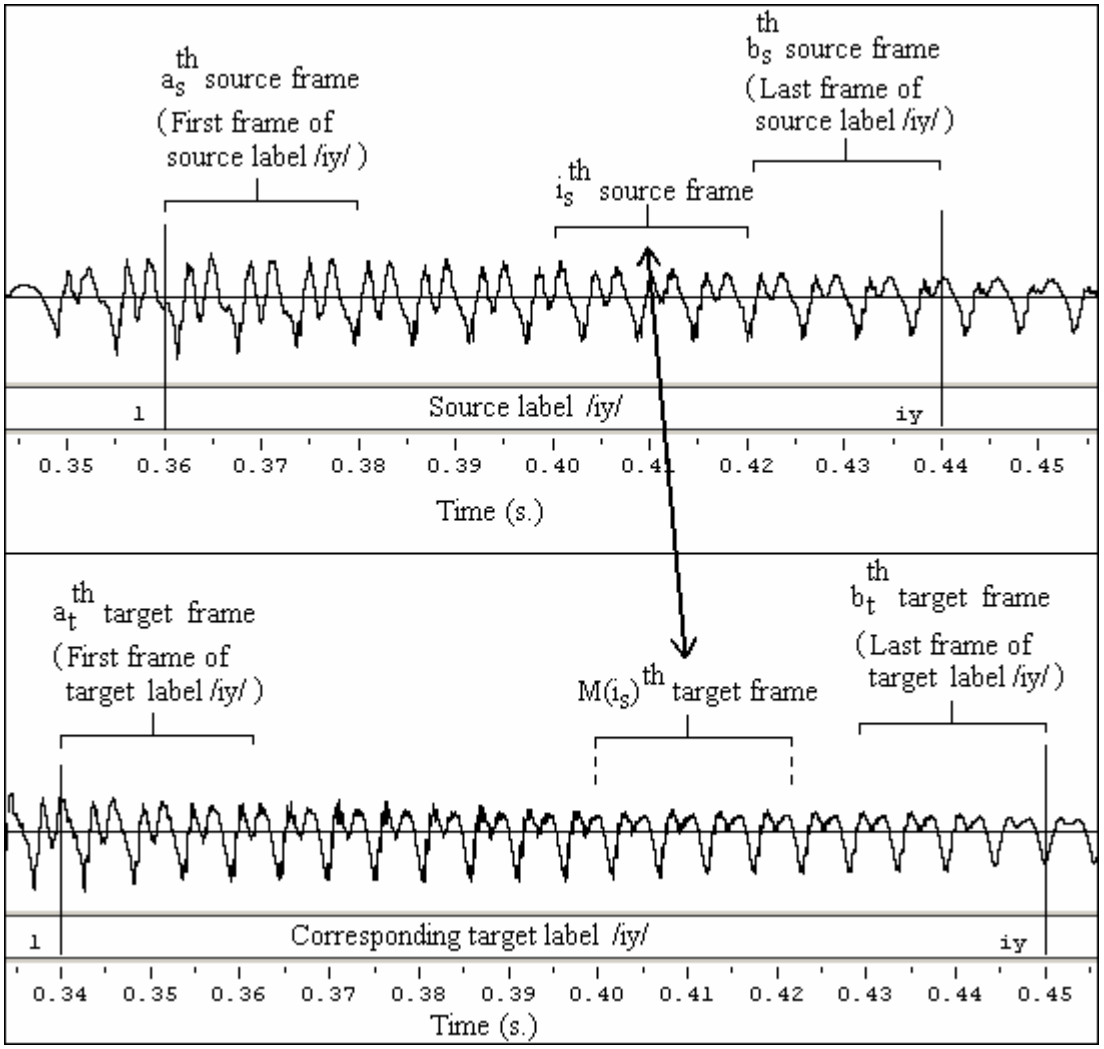

Figure 4.1. An example of the speech frame index mapping process between a source label and the corresponding target label.

For alignment comparison, identical utterances from the source and target speaker pairs are required. We have used the TIMIT utterances “sa1” and “sa2” from 40 speakers as the test set. All combinations of speaker pairs are considered. Therefore, we had 40x39 = 1560 source-target alignments for each method. For each source-target

speaker pair and for each Phonetic-HMM aligner, we first perform phoneme recognition with the corresponding Phonetic-HMM to label the source recording. The corresponding target recording is force-aligned with the recognized phoneme string using the Viterbi algorithm. Table 4.1 shows the contents of the training and test databases.

Type	**Language**	**# Speakers**	**# Duration**
Training	English	325 (135 female, 190 male)	2 hr 47 min
Training	Turkish	175 (81 female, 94 male)	6 hr 32 min
Test	English	40 (20 female, 20 male)	9 min

Table 4.1. Contents of training and test databases for alignment

We have trained Phonetic-HMMs using the TIMIT phoneme set and two sets of acoustic data. The first three rows of Table 4.2 show the different Phonetic-HMM architectures trained using different sets of acoustic data. The number of mixtures that resulted in best performance is also noted for each case. We have used the HTK Toolkit for training context-independent HMMs (Woodland, et. al., 1994). In the case when English and Turkish acoustic data were used together, we first trained base models using the data in English only, performed phoneme recognition on the Turkish data, and updated the models with additional iterations of the Baum-Welch algorithm. In our case, convergence was achieved in four additional iterations. As the baseline method, we have used Sentence-HMM alignment. In this case, for each source training utterance, an HMM was trained using one mixtures per state. The number of states was proportional to the duration of the utterance. A new state was added to the HMM architecture for every 40 milliseconds. The corresponding target utterance was force-aligned with the source HMM to obtain the final alignment.

Note that for HMM_ETV in Table 4.2, we have used different number of mixtures for different phonemes depending on the number of occurrences of phonemes in the manually aligned data. Otherwise, it was not possible to obtain reliable model parameters using more than eight mixtures with the given amount of acoustic data. We

have divided the TIMIT phonemes into four groups depending on the frequency of occurrence in the TIMIT training set as shown in Table 4.3. For this purpose, the histogram of total speech frames for each phoneme is computed as in Figure 4.2. The histogram is divided into four non-overlapping ranges. For each range, a variable number of mixtures were assigned to the corresponding Phonetic-HMM depending on the total acoustic data. The motivation was to use less number of mixtures for infrequent phonemes for which the acoustic data is limited to obtain more robust models. When the available data is large for a specific phoneme, the number of mixtures is increased to model the variability of acoustic data in a better fashion. We have chosen four ranges by examining the histogram and assigned the corresponding number of mixtures for each state as shown in the following table. There were a total of 177080 observations for 61 phonemes in the training set.

HMM	**Acoustic data**	**Number of mixtures per state**
HMM_E	English	6
HMM_ET	English+Turkish	8
HMM_ETV	English+Turkish	Variable (4, 6, 8, 12)
Sentence-HMM	Single training utterance pairs	1

Table 4.2. HMM architectures

Total occurrences in the training database	**Range in histogram *H*=normalized number of occurrences for a given phoneme**	**Mixtures per state**	**Phonemes**
<1771	$H<0.01$	4	aw, axh, ch, el, em, en, eng, epi, hh, hv, jh, ng, nx, oy, pau, th, uh, uw, y, zh
[1771, 3542)	$0.01<=H<0.02$	6	aa, ah, ao, axr, ay, b, bcl, dh, dx, er, ey, f, g, gcl, ow, p, pcl, sh, ux, v, w
[3542, 7083)	$0.02<=H<0.04$	8	ae, ax, d, dcl, eh, ih, iy, k, kcl, l, m, n, q, r, t, tcl, z
>7083	$H>=0.04$	10	h#, ix, s

Table 4.3. Choice of total mixtures per state for each phoneme

We have compared the average alignment mismatch scores for different alignment methods in statistical tests. For this purpose, alignment mismatch scores are computed for all methods for all source speech frames. Results are compared pair by pair using pair wise t-tests. The pair wise t-test is a statistical test to compare the mean values of two distributions where two set of samples come from (Kreyszig, 1970). The test returns a p-value for the probability of observing a specified result. As an example, it can be employed to evaluate the probability of the mean value of a set of samples being greater than that of another set of samples within a significance level.

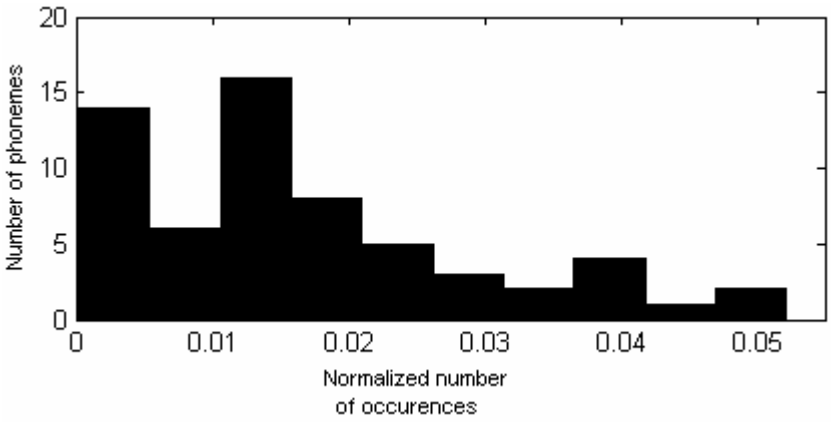

Figure 4.2. Normalized number of occurrences of phonemes in the TIMIT training corpus

	Sentence-HMM	**HMM_E**	**HMM_ET**	**HMM_ETV**
Sentence-HMM	x	x	x	x
HMM_E	0	x	x	x
HMM_ET	0	0.0028	x	x
HMM_ETV	0	1.0e-11	2.7e-5	x

Table 4.4. Pair wise comparison of mean alignment mismatch scores. For the underlined p-values, the corresponding aligner in the first column results in lower average mismatch score as compared to the aligner in the first row. Results are given for a confidence level of 99%

For each aligner pair, the following hypothesis was tested: The mean alignment score of the first aligner is significantly less than the second aligner. Since a significance level of 99% is used, p-values less than 0.01 show that the mean alignment mismatch score of the first method is significantly lower than that of the second method. We have used the TIMIT test corpus in the evaluations. For this purpose, 5082 source-target speaker utterance pairs were selected. For each pair, the alignment mismatch score between the manual alignment and alignment using one of the HMM architectures given in Table 4.2 are computed. Table 4.4 shows pair wise comparisons of different alignment methods.

	Sentence-HMM	**HMM_E**	**HMM_ET**	**HMM_ETV**
Score (ms)	78.6	34.0	33.3	33.0

Table 4.5. Mean alignment mismatch score in milliseconds using different HMM architectures

Table 4.5 shows the mean alignment mismatch scores for different alignment methods. We observe that the Phonetic-HMM based alignment mismatch scores are fairly low as compared to the Sentence-HMM case. Even the worst Phonetic-HMM based aligner, HMM_E, resulted in significantly lower mean alignment mismatch score as compared to the Sentence-HMM based aligner. In the best case, the alignment mismatch score was 33.0 milliseconds. We observe that when acoustic data is extended using the Turkish database, a larger number of mixtures per state are required as expected. Another observation is that when monolingual HMMs are extended with data from another language, comparable alignment performance can be obtained by carefully adjusting the number of mixtures per state in HMM training.

4.3.2. Voice Conversion Performance

In order to compare the voice conversion performance using Phonetic-HMM for segmentation and alignment in cross-lingual voice conversion, we designed an

objective voice conversion test. Two methods were compared with the Sentence-HMM based method and vocal tract transplantations:

- Text-independent mode, parallel training (TIP)
- Text-dependent mode, parallel training (TDP)

A compound bilingual male speaker of American English and Turkish from the voice conversion database described in Chapter 7 is used as the target. A male Turkish speaker from the same database is used as the source. The training set consisted of 80 utterances in English and the test set consisted of 10 utterances in Turkish. We have used the LSF distance measure to rate the objective similarity of the transformation outputs to target speaker's reference recordings in Turkish. LSF distance resulted in better performance as compared to a number of objective distance measures for evaluating vocal tract transformation performance. Section 5.4 presents the details on objective measure selection.

The Phonetic-HMM architecture HMM_ETV was used for segmenting and aligning the utterances. Note that for comparison, manual alignment of the transformed and target reference utterances was employed. In transformations, no smoothing was applied as a post-processing step in order to compare direct target frame reconstruction performance. State-averaging method was used in training for comparing the effect of alignments only.

Figure 4.3 shows the results. The original average LSF distance between the source and the target speaker was 4.81. Sentence-HMM resulted in a distance of 4.20 while Phonetic-HMM in text-independent and text-dependent modes reduced the distance to 3.89 and 3.84 respectively. For comparison, we have also computed the average LSF distance between the target speaker test recordings and vocal tract transplantation outputs which turned out to be 2.94. Note that transplantation requires the target speaker recordings to be available for test utterances. It is a copy-paste method in which the source vocal tract spectrum is directly replaced by the corresponding target vocal tract spectrum using the alignment information. Therefore, it

corresponds to an ideal vocal tract transformation. The Phonetic-HMM based methods result in a reduction of 0.31 and 0.36 in the average LSF distance to the target speaker as compared to the Sentence-HMM based alignment.

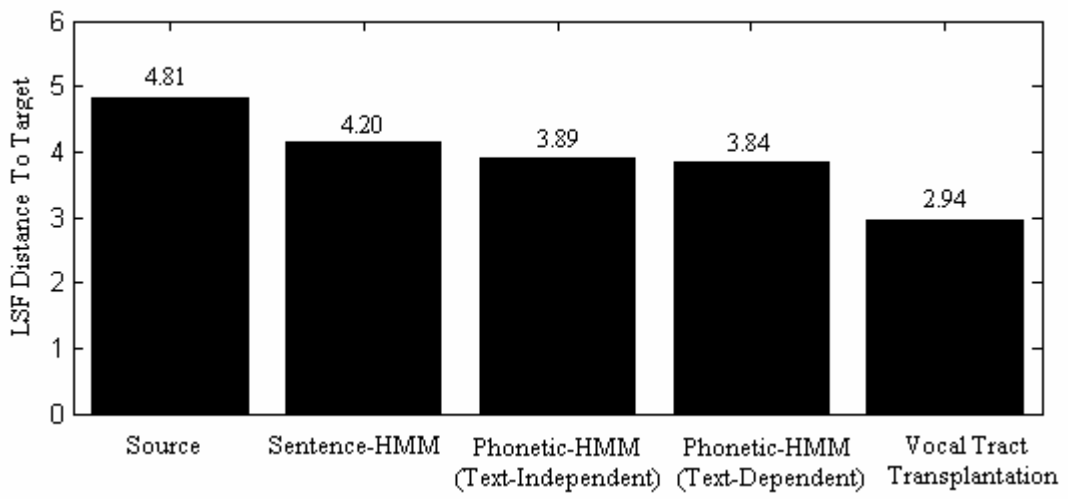

Figure 4.3. Objective comparison of voice conversion performance using different alignment and segmentation methods

5. VOCAL TRACT TRANSFORMATION USING WEIGHTED FRAME MAPPING

5.1. Introduction

Transplantation of the vocal tract spectrum refers to replacing the vocal tract spectrum of the source speaker with that of the target speaker by time-varying filtering techniques. Our previous research on vocal tract transplantations show that the output is significantly close to the target voice provided that average prosody characteristics are also modified to match that of the target speaker's (Turk, 2003). It is not possible to use a transplantation technique directly in voice conversion since this would require every possible transformation utterance to be recorded from the target speaker. However, it is possible to estimate the vocal tract transformation function in more detail to make the voice conversion output closer to the transplantation results.

Inspired by the closeness of vocal tract transplantations to target speaker voices, a more detailed vocal tract transformation algorithm may help to improve cross-lingual voice conversion performance. In order to perform detailed transformation with sufficient quality, the following requirements should be satisfied:

- The vocal tract transformation function should be estimated directly from the training speech frames as in the case of vocal tract transplantation.
- The alignment between the source and the target training utterances should be performed in a robust manner such that no large misalignments are present. Otherwise, the detailed vocal tract transformation may be estimated incorrectly for some speech frames resulting in distortion and lower similarity to target voice.
- Estimation of vocal tract transformation functions directly from the speech frames increases the memory and processor requirements in general. In order to be able to use this technique in larger voice conversion databases, i.e. on the order of hundreds of utterances, a searching strategy is required.

- Even in the case of robust and reliable alignments, depending on the content of the material to be transformed, the vocal tract transformation function estimated directly from the speech frames may contain discontinuities among consecutive speech frames. Appropriate weighting of speech frames is required in order to reduce the discontinuities.
- A reliable and robust objective distance measure is required for the evaluation of vocal tract transformation performance.

The conventional training method in STASC uses a state-averaged version of the LSF parameters to represent the vocal tract transformation function. The aim of this pre-smoothing step is to reduce discontinuities in the resulting transformation function by smoothing the source and the target spectrum estimates for each state. This kind of pre-smoothing has an important disadvantage. The smoothing is directly performed on the source and target LSF vectors which result in a reduction in the detail of vocal tract transformation. It is not possible to recover the detail information since it is performed in the training stage. Even if re-training is possible, the Sentence-HMM based method does not provide sufficiently good alignments to estimate the transformation function from the individual source and target speech frames. On the contrary, the Phonetic-HMM based method described in Chapter 4 provides a reliable framework for obtaining the alignment. Therefore, it can be used for detailed vocal tract transformation function estimation on a frame-by-frame basis as we describe in this chapter.

The estimation of the vocal tract transformation function in more detail has the disadvantage of increasing the possibility of observing discontinuities at the output. In order to reduce the discontinuities in a controllable manner, weighting of speech frames as used in STASC transformation should be employed. Combined with the detailed frame mapping and weighting, this technique provides the framework for more detailed vocal tract transformation. In addition to these techniques, we also propose a context matching algorithm which can be used to limit the search space in the case of large training databases.

The proposed technique has several common ideas with unit selection text-to-speech algorithms. As in the case of unit selection synthesis, the primary goal is to be able to use the training data as much as possible. An important distinction is in the amount of available data. For unit selection synthesis hours of speech data along with prosodic and linguistic information are available. Therefore, unit selection databases can be designed to cover a large amount of contextual and prosodic variation. The synthesis algorithm can make use of all information in the database to minimize target and concatenation costs to generate speech. On the contrary, voice conversion databases are generally limited to tens to hundreds of utterances, i.e. one minute to ten minutes of speech. It is not possible to use a unit selection algorithm directly for searching for the closest source and target matches since the coverage is low for a voice conversion database. The algorithm described in this chapter can be classified as a unit selection algorithm that uses a target cost function consisting of purely acoustic features.

The proposed method has the following advantages for cross-lingual voice conversion:

- Making use of more training data in order to perform more detailed transformation
- Reducing the one-to-many mapping problems by successfully constraining the acoustic matching process
- Employing non-parallel databases for training especially when the source and the target do not speak the same language
- Improving the memory and processor load constraints when large training databases are used

The general flowchart of the proposed vocal tract transformation function estimation algorithm is given in Figure 5.1. The training starts with the mapping of the source and the target speech frames in the case of parallel training databases. Context information is also extracted for context-matching. In the case of non-parallel training databases, context information is the only tool for mapping the source and the target

acoustic spaces. The acoustic features extracted from the speech frames are saved in binary files along with detailed context information. Mapping between the source and target speech frames is also saved in this binary file in case of parallel training.

In the transformation, a detailed vocal tract transformation function is estimated for each source speech frame to be transformed using weighted frame mapping. Context-information can be used for matching the source input speech frame with the training source speech frames and finding the corresponding target features. In the case of parallel training databases, full search is another option since the correspondence between the source and target training speech frames are known. After weighted frame mapping the vocal tract transformation function is estimated as a time-varying frequency domain filter.

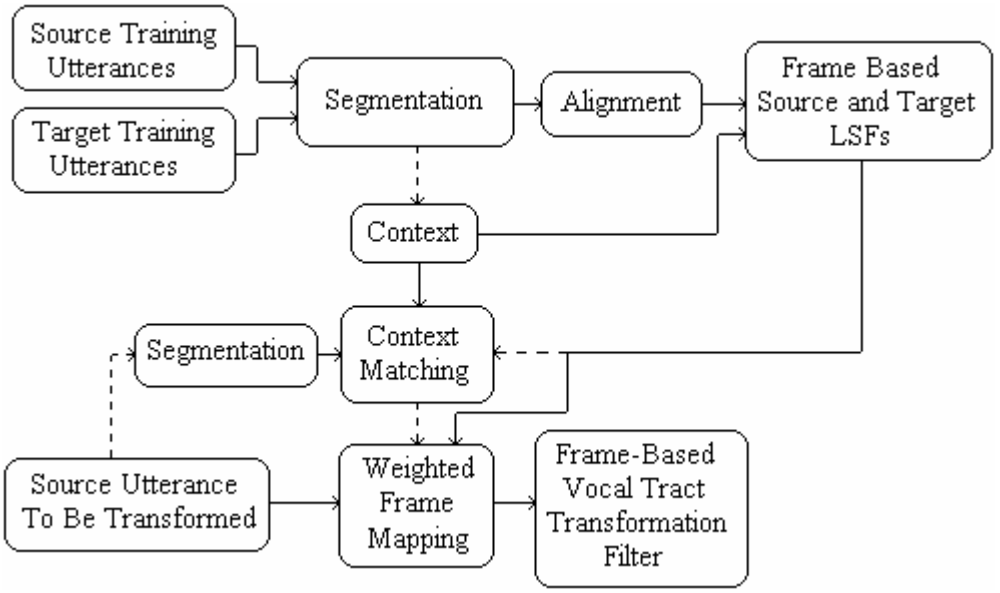

Figure 5.1. Proposed vocal tract transformation function estimation framework

5.2. Weighted Frame Mapping

In weighted frame mapping, the aim is to estimate the vocal tract transformation function from the source and the target training speech frames directly. For this purpose, the source and the target training databases are aligned using a Phonetic-HMM

aligner. In the case of a parallel training database, for each source speech frame, the corresponding target speech frame is determined by linear mapping of speech frames as described in Section 4.3.1. In the case of non-parallel training, all source and target speech frame acoustic parameters are saved in binary files and the transformation filter estimation is performed using the context-matching method as described in the next sub-section.

During transformation, the N-closest source speech frames in the training data are determined using the LSF distance based method described in Section 3.2.2 for each source input speech frame. *N* is set to a smaller number as compared to STASC transformation, typically to 3 whereas *N*=6 to 10 is commonly used in the baseline algorithm. This helps to reduce excessive smoothing.

5.3. Context-Matching

Figure 5.1 shows a flowchart of the context-matching based algorithm. In the training stage, the source and target utterances are segmented using the Phonetic-HMM method. It is possible to use text transcriptions at this stage as described in Chapter 4. For each source and target speech frame in the training database, the context information and the acoustic features (LSFs, f0, voicing, energy) are recorded. The first step in the context-matching based algorithm is the extraction of context information in the training phase. For each speech frame in the training database, the phonetic context is recorded up to 20 previous and 20 next phonemes. Using such wide context information allows the transformation algorithm to use longer target training patterns during transformation. When there is no match in the training database, the context is reduced and the search process is repeated. The normalized location, l_{norm}, of the speech frame inside the current phoneme is also calculated using:

$$l_{norm} = \frac{(ss-1)i + 0.5ws}{I} \tag{13}$$

where *ss* and *ws* are the skip size and window size in seconds, *I* is the total number of speech frames in the current phoneme, and i is the index of the speech frame in the

current phoneme ($1 = i = I$). Note that l_{norm} takes values in the range [0.0, 1.0]. It is appropriately normalized prior to integrating it with the context match scores as described below.

In the transformation stage, the best matching codebook entries in terms of context are found using the following algorithm:

- Let the current phonetic context be:

$L_{tfm}(N\text{-}1)\ L_{tfm}(N\text{-}2)\ \ldots\ L_{tfm}(1)\ L_{tfm}(0)\ M_{tfm}\ R_{tfm}(0)\ R_{tfm}(1)\ \ldots\ R_{tfm}(N\text{-}2)\ R_{tfm}(N\text{-}1)$

where M_{tfm} is the label for the current phoneme, $L_{tfm}(i)$'s and $R_{tfm}(i)$'s are the labels of the N preceding and succeeding phonemes respectively in the source speaker utterance to be transformed.

- Let also the context in k^{th} source codebook entry be:

$L_k(N\text{-}1)\ L_k(N\text{-}2)\ \ldots\ L_k(1)\ L_k(0)\ M_k\ R_k(0)\ R_k(1)\ \ldots\ R_k(N\text{-}2)\ R_k(N\text{-}1)$

- Compute context matching score, *s*, by the following pseudo-code:

```
set s=0.0, w=10^N
if M_tfm = M_k then set s = s + w
for i=0 to N-1
    set w = 0.1w
    if L_tfm(i)=L_k(i) then s= s + w
    if R_tfm(i)=R_k(i) then s= s + w
    if L_tfm(i)≠L_k(i) and R_tfm(i)≠R_k(i) then break
set w = 0.1w
s = s + w.l_norm
```

Table 5.1. Pseudo-code for computing the context matching score

When the context is longer, it is required to perform the operations in the logarithm domain by replacing the last step with the following pseudo-code:

```
set s=0.0, w=MAX_CONTEXT (MAX_CONTEXT = 20 in our case)
if M_tfm = M_k then set s = logAdd(s,w)
for i=0 to N-1
    set w = w-1
    if L_tfm(i)=L_k(i) then s=logAdd(s,w)
    if R_tfm(i)=R_k(i) then s=logAdd(s,w)
    if L_tfm(i)≠L_k(i) and R_tfm(i)≠R_k(i) then break
set w = w-1
s = logAdd(s, w+l_norm)
```

Table 5.2. Pseudo-code for computing the context matching score in logarithmic domain

Note that logAdd(x,y) performs logarithmic addition in order to prevent overflow/underflow problems when large context is used. The pseudo-code of the logAdd function is shown in Table 5.3.

```
r = logAdd(logX, logY)
if logY>logX then switch the values of logX and logY.
if logX<-MIN_DOUBLE+1e-100 then set r = logX and return
if logY-logX<-20 then set r=logX and return.
otherwise set r=logX + log(1.0 + exp(logY-logX)) and return.
```

Table 5.3. Pseudo-code for addition in the logarithmic domain

The weighted LSF distances are then calculated only for the best matching entries instead of using all of the codebook entries. A weighted average of the LSFs are found similar to Equation 4 and used in transformation. The disadvantage of the context-based matching algorithm is that a significantly larger database is required to cover all possible contextual combinations. However, the algorithm can be used in two-modes

simultaneously by simply switching to the baseline method when sufficient amount of match cannot be found in the codebooks.

At this point, an analogy between the proposed target vocal tract estimation method and concatenative unit selection TTS techniques will be useful. In unit selection based TTS, target unit specifications are determined by prosodic and linguistic modules of the TTS system. Using a set of pre-selection trees, the search space for the units to concatenate is restricted. Then, the best set of units that will minimize a combination of the target and concatenation costs is determined. The target cost corresponds to the acoustic distance between a candidate unit and a target unit. The unit selection cost penalizes discontinuities at concatenation boundaries. The context matching algorithm combined with the weighted speech frame mapping can be considered as a special case of unit selection from the target training data with smoothing. The context matching step acts as a pre-selection stage using phonetic context similar to pre-selection using decision trees in unit selection based TTS (Black and Taylor, 1997). Since the amount of data for typical voice conversion training is restricted, only phonetic context is used in the proposed voice conversion method. Searching for the best candidates among the pre-selected speech frames using the LSF distance based acoustic distance measure corresponds to the employment of the target costs in unit selection TTS. The weighted estimation of the target vocal tract spectrum from the best acoustic matches provides a smoothing mechanism similar to the concatenation costs in unit selection TTS.

5.4. Objective Distance Measures for the Evaluation of Vocal Tract Similarity

It is rather difficult to design a subjective listening test in which the subjects will only focus on vocal tract transformation performance and ignore other acoustic clues on the target speaker's identity. This is due to the fact that acoustic features such as pitch and voice quality have relations with the vocal tract configuration (Kain and Stylianou, 2000), (d'Alessandro and Doval, 1998). Because of the difficulties in designing subjective listening tests to evaluate vocal tract transformation performance in an independent manner, this section aims to determine a suitable objective measure for vocal tract transformation performance assessment.

We start with a description of the desired properties of a reliable objective distance measure for vocal tract transformation evaluation. Based on these desired properties, a statistical comparison framework is designed. Then, the robustness and sensitivity of different objective measures in different speech signal modification scenarios are compared.

A reliable objective measure for evaluating vocal tract transformation performance in voice conversion applications should possess the following properties:

- Robustness to changes in the residual signal, i.e. pitch changes, algorithmic manipulations of the residual for pitch modification, etc.
- Robustness to noise and differences in recording conditions.
- Robustness to linguistic variations.

A large number of objective measures have been proposed in the speech coding and synthesis literature to compare reference signals with coding or synthesis outputs. The most popular measure is the signal-to-noise ratio (SNR):

$$SNR = 10log_{10}\frac{\sum_{n=0}^{M-1} s^2(n)}{\sum_{n=0}^{M-1}(s(n)-\hat{s}(n))^2} \tag{14}$$

where *M* is the window size, *s(n)* and *s(n)* are the speech signal samples. As SNR is not sufficiently sensitive to temporal changes in similarity of the reference and test signals, a locally estimated and averaged version called segmental SNR can be used instead:

$$SNR_{seg} = \frac{10}{L}\sum_{i=0}^{L-1} log_{10}\frac{\sum_{n=0}^{N-1} s^2(iN+n)}{\sum_{n=0}^{N-1}(s(iN+n)-\hat{s}(iN+n))^2} \tag{15}$$

where *L* is the total number of speech frames in the original and estimated signals and *N* is the window size. It is also possible to compare two speech signals based on

spectral distance measures. The most basic measure is spectral distortion (SD) as defined by:

$$SD = \frac{10}{L}\sum_{i=0}^{L-1}\int_{0}^{\frac{F_s}{2}}(log_{10}(A_i(f)) - log_{10}(\hat{A}_i(f)))^2 df \qquad (16)$$

where L is the total number of speech frames in the original and estimated signals, F_s is the sampling rate in Hz., and $A_i(f)$ and $\hat{A}_i(f)$ are the amplitude at frequency bin f of the spectrum of the i^{th} speech frame from the original and estimated signals respectively.

Spectrum-based objective measures enable perceptual weighting in the frequency domain similar to the human auditory system. As an example, frequency-weighted spectral distortion measure can be calculated using:

$$SD_{fw} = \frac{10}{L}\sum_{i=0}^{L-1}\sqrt{\frac{1}{W_0}\sum_{f=0}^{\frac{F_s}{2}}|W_B(f)|^2(log_{10}(\frac{|\hat{A}_i(f)|^2}{|A_i(f)|^2}))^2} \qquad (17)$$

where L is the total number of speech frames in the original and estimated signals, F_s is the sampling rate in Hz., and $A_i(f)$ and $\hat{A}_i(f)$ are the amplitude at frequency bin f of the spectrum of the i^{th} speech frame from the original and estimated signals respectively. $W_B(f)$ and W_0 are used for Bark-scale weighting of each frequency bin and normalization respectively. They can be computed using:

$$W_B(f) = \frac{1}{25 + 75[1 + 1.4\frac{f^2}{1000^2}]^{0.69}} \qquad W_0 = \sum_{f=0}^{\frac{F_s}{2}} W_B(f) \qquad (18)$$

Another possibility for comparing the perceptual similarity between two speech signals is to extract model parameters and to compute the distance between these parameters. Linear prediction analysis (Makhoul, 1975) and cepstral analysis (Oppenheim and Schafer, 1975) are two well-known modeling techniques. Linear

prediction derived line-spectral frequencies (LSFs) are commonly used for speech coding and voice conversion applications due to their good interpolation properties. Inverse harmonic weighting based LSF distance is a useful measure for computing the perceptual distance between two LSF vectors:

$$\beta(n) = \begin{cases} \dfrac{1}{|u(2)-u(1)|} \; for \; n=1, \\ \dfrac{1}{min(|u(n)-u(n-1)|, |u(n)-u(n+1)|)} \; for \; n=2,...,P-1, \\ \dfrac{1}{|u(P)-u(P-1)|} \; for \; n=P \end{cases} \quad (19)$$

$$LD = \sum_{n=1}^{P} \beta(n)|u_1(n) - u_2(n)| \quad (20)$$

where P is the linear prediction order, u_1 and u_2 are the P-dimensional LSF vectors, $ß(n)$'s are the inverse harmonic weights and LD is the LSF distance between the LSF vector u_1 and u_2.

Using cepstral analysis, it is possible to perform perceptual weighting in the cepstrum domain. Weighted cepstral distance can be computed using:

$$WCeps = \sum_{n=1}^{P} n^2 (c_1(n) - c_2(n))^2 \quad (21)$$

where c_1 and c_2 are the cepstrum vectors and P is the prediction order.

In order to compare the performance of this set of objective distance measures in evaluating vocal tract transformation, we used a set of original recordings and their processed versions. The processing was performed using different algorithms that may

or may not change the residual or the vocal tract spectrum. The aim is to determine the objective measure that possesses the following properties:

- Robustness to processing of the speech signal such that the vocal tract does not change significantly
- Sensitivity to processing of the speech signal such that the vocal tract changes significantly

Examples of processing algorithms for the first case include prosody modifications (pitch or duration scaling) and slight noise addition (i.e. 20 dB SNR). The algorithms for the second case include vocal tract scaling, filtering, and vocal tract transplantation. The original recording set contained a total of 100 sentence utterance recordings from 13 different speakers (seven female, six male). The following set of objective measures are computed among pairs of reference, input, and output recordings:

- Spectral Distortion (SD)
- Frequency Weighted Spectral Distortion (SD_{fw})
- Line Spectral Frequency Distance (LD)
- Weighted LP Cepstral Distance (WCeps)

For each processing algorithm and for each objective distance measure, we compute the distances between the following pairs of speech signals:

- Reference (Original Signal) – Output (Processed Signal): d_{ro}
- Reference (Original Signal) - Input (Another Original Signal): d_{ri}
- Input (Another Original Signal) – Output (Processed Signal): d_{io}

We applied pair wise t-tests to compare the expected values of d_{ro}, d_{ri}, and d_{io}. The pair wise t-test is a statistical test to compare the mean values of two distributions where two set of samples come from. The test returns a p-value for the probability of observing a specified result, i.e. the mean of distribution in which the first set of

samples comes from is greater than that of the second set, etc. The p-values corresponding to different cases can be used for comparing the performances of different objective measures in evaluating vocal tract transformation amount at a given significance level.

Considering the vocal tract similarity in the reference and input signals, we have two possibilities:

- Vocal tract spectrum is significantly different in the reference and input signals
- Vocal tract spectrum is not significantly different in the reference and input signals

When there is significant vocal tract difference between the reference and the input, we can have two cases depending on the processing algorithm used:

- Case 1: The processing algorithm modifies the input vocal tract spectrum significantly. An example for this case is vocal tract transplantation from speaker Y onto the residual of the speaker X. In this case, the distance pairs given above should have the following properties for a good objective distance measure:

  (i) $d_{ro} << d_{ri} \rightarrow r_1 = d_{ro}/d_{ri} << 1$

  (ii) $d_{ro} << d_{io} \rightarrow r_2 = d_{ro}/d_{io} << 1$

  (iii) $d_{io} \sim d_{ri} \rightarrow r_3 = d_{io}/d_{ri} \sim 1$

- Case 2: The processing algorithm does not modify the input vocal tract spectrum significantly. Examples of this case include pitch or time scale modification, slight noise addition, vocal tract transplantation with excessive amount of smoothing. The following properties should hold for a good objective distance measure:

(i) $d_{ro} \sim d_{ri} \rightarrow r_1 = d_{ro}/d_{ri} \sim 1$

(ii) $d_{ro} >> d_{io} \rightarrow r_2 = d_{ro}/d_{io} >> 1$

(iii) $d_{io} << d_{ri} \rightarrow r_3 = d_{io}/d_{ri} << 1$

When there is not significant vocal tract difference between the reference and the input, we can have two cases depending on the processing algorithm used:

- Case 3: The processing algorithm does not modify the input vocal tract spectrum significantly. As an example, the input signal can be a sentence utterance recording from speaker X, the reference signal can be the identical utterance recorded from the same speaker again, and the output can be pitch or time scaling of the same recording, or slight noise addition. In this case, a good objective distance measure should have the following properties:

  (i) $d_{ro} \sim d_{ri} \rightarrow r_1 = d_{ro}/d_{ri} \sim 1$

  (ii) $d_{ro} \sim d_{io} \rightarrow r_2 = d_{ro}/d_{io} \sim 1$

  (iii) $d_{io} \sim d_{ri} \rightarrow r_3 = d_{io}/d_{ri} \sim 1$

- Case 4: The processing algorithm modifies the input vocal tract spectrum significantly. Examples of this case include vocal tract scaling or filtering. The following properties should hold for a good objective distance measure in this case:

  (i) $d_{ro} >> d_{ri} \rightarrow r_1 = d_{ro}/d_{ri} >> 1$

  (ii) $d_{ro} \sim d_{io} \rightarrow r_2 = d_{ro}/d_{io} \sim 1$

  (iii) $d_{io} >> d_{ri} \rightarrow r_3 = d_{io}/d_{ri} >> 1$

In summary, three performance measures are considered for performance comparison for each case: $r_1 = d_{ro}/d_{ri}$, $r_2 = d_{ro}/d_{io}$, and $r_3 = d_{io}/d_{ri}$. Depending on the signal pairs for which these ratios are computed and the signal processing algorithms employed, these ratios should be as close as possible to the specified values for the corresponding cases. For each case, we computed the respective objective distances and

applied a pair wise t-test to evaluate whether the given objective measure satisfies the requirement corresponding to that case.

As an example, one of the sample generation procedures for Case 2 is pitch scaling speaker X's utterance recording by a large factor (i.e. 0.6 or 1.8). When we compare the output of this procedure to another speaker's identical utterance recording, we expect to observe the reference to output distance to be close to the reference to input distance as pitch scaling does not change the vocal tract spectrum. In Table 5.2, the fourth row shows that all four objective distances satisfy this condition at a significance level of 95%.

Case # and condition	Input	Ref	Processing Algorithm	p-values and $r_1=d_{ro}/d_{ri}$ ratios for objective distances			
				SD	**SD_{fw}**	**LD**	**WCeps**
Case 1 $d_{ro}<<d_{ri}$?	X	Y	Transp1	**p=0.0000 r_1=0.9366**	**p=0.0000 r_1=0.9722**	**p=0.0000 r_1=0.5466**	**p=0.0000 r_1=0.5318**
Case 2 $d_{ro}\sim d_{ri}$?	X	Y	Transp2 or Transp3	p=0.7185 r_1=1.0027	**p=0.0005 r_1=1.0215**	**p=0.0000 r_1=0.7830**	**p=0.0178 r_1=0.4990**
Case 2 $d_{ro}\sim d_{ri}$?	X	Y	PScale1 or PScale2	**p=0.0000 r_1=1.2269**	**p=0.0000 r_1=1.2636**	**p=0.0000 r_1=1.0314**	**p=0.0319 r_1=0.6091**
Case 2 $d_{ro}\sim d_{ri}$?	X	Y	TScale1 or TScale2	**p=0.0000 r_1=1.0374**	**p=0.0000 r_1=1.0543**	**p=0.0027 r_1=1.0174**	p=0.4264 r_1=0.8433
Case 2 $d_{ro}\sim d_{ri}$?	X	Y	Noise1 or Noise2	**p=0.0000 r_1=1.1457**	**p=0.0000 r1=1.0439**	**p=0.0000 r_1=1.1432**	**p=0.0001 r_1=0.3679**
Case 3 $d_{ro}\sim d_{ri}$?	Y	Y	PScale1 or PScale2	**p=0.0000 r_1=1.3130**	**p=0.0000 r_1=1.3339**	**p=0.0000 r_1=1.0818**	p=0.9362 r_1=0.9891
Case 3 $d_{ro}\sim d_{ri}$?	Y	Y	TScale1 or TScale2	**p=0.0022 r_1=1.0152**	**p=0.0119 r_1=0.9898**	**p=0.0000 r_1=1.0535**	p=0.5001 r_1=1.0929
Case 3 $d_{ro}\sim d_{ri}$?	Y	Y	Noise1 or Noise2	**p=0.0000 r_1=1.1478**	**p=0.0000 r_1=1.0420**	**p=0.0000 r_1=1.2534**	p=0.3354 r_1=0.8681
Case 4 $d_{ro}>>d_{ri}$?	Y	Y	VScale1 or VScale2	**p=0.0000 r_1=1.1017**	**p=0.0000 r_1=1.0410**	**p=0.0000 r_1=1.5412**	p=0.3188 r_1=1.0656
Case 4 $d_{ro}>>d_{ri}$?	Y	Y	Filt1 or Filt2	**p=0.0000 r_1=1.4078**	**p=0.0000 r_1=1.6262**	**p=0.0000 r_1=2.5570**	**p=0.0068 r_1=1.7926**
Total Closest				**0**	**3**	**7**	**0**

Table 5.4. d_{ro} vs d_{ri} values for different objective measures and different speech processing algorithms

Table 5.4 shows that LSF distance using inverse harmonic weighting and frequency weighted spectral distance perform the best when d_{ro} and d_{ri} values are considered in different situations. The p-values which are less than 0.05 (i.e. satisfying the desired result at a confidence level of 95%) are marked in bold characters. The corresponding averaged ratios r_1 are also shown. Note that r is obtained by averaging d_{ro}/d_{ri} ratios over all samples. The r_1 values closest to the desired values for a given row are also marked for each case. Therefore, the objective distance measure which has the maximum number of closest r_1 values is a better distance that satisfies the properties discussed above.

Case # and condition	Input	Ref	Processing Algorithm	p-values and $r_2=d_{ro}/d_{io}$ ratios for objective distances			
				SD	**SD_{fw}**	**LD**	**WCeps**
Case 1 $d_{ro} << d_{io}$?	X	Y	Transp1	p=1.0000 r_2=1.3156	p=1.0000 r_2=1.3212	**p=0.0000 r_2=0.5812**	p=0.0836 r_2=0.8341
Case 2 $d_{ro} >> d_{io}$?	X	Y	Transp2 or Transp3	**p=0.0000 r_2=1.9029**	**p=0.0000 r_2=1.9411**	**p=0.0000 r_2=1.4831**	p=0.4540 r_2=1.0369
Case 2 $d_{ro} >> d_{io}$?	X	Y	Pscale1 or Pscale2	**p=0.0000 r_2=1.3868**	**p=0.0000 r_2=1.3455**	**p=0.0000 r_2=2.8084**	p=0.6871 r_2=0.8888
Case 2 $d_{ro} >> d_{io}$?	X	Y	Tscale1 or Tscale2	**p=0.0000 r_2=1.2491**	**p=0.0000 r_2=1.2883**	**p=0.0000 r_2=2.0782**	**p=0.0030 r_2=1.9243**
Case 2 $d_{ro} >> d_{io}$?	X	Y	Noise1 or Noise2	**p=0.0000 r_2=1.5022**	**p=0.0000 r_2=2.4557**	**p=0.0000 r_2=1.4770**	p=0.7701 r_2=0.8160
Case 3 $d_{ro} \sim d_{io}$?	Y	Y	Pscale1 or Pscale2	**p=0.0000 r_2=1.4895**	**p=0.0000 r_2=1.4500**	**p=0.0000 r_2=2.4762**	**p=0.0000 r_2=3.2430**
Case 3 $d_{ro} \sim d_{io}$?	Y	Y	Tscale1 or Tscale2	**p=0.0000 r_2=1.1627**	**p=0.0000 r_2=1.1341**	**p=0.0000 r_2=1.9324**	**p=0.0000 r_2=2.7134**
Case 3 $d_{ro} \sim d_{io}$?	Y	Y	Noise1 or Noise2	**p=0.0000 r_2=1.7778**	**p=0.0000 r_2=3.2166**	**p=0.0000 r_2=1.4289**	**p=0.0000 r_2=3.3189**
Case 4 $d_{ro} \sim d_{io}$?	Y	Y	Vscale1 or Vscale2	**p=0.0000 r_2=1.2985**	**p=0.0000 r_2=1.2922**	**p=0.0000 r_2=1.1489**	**p=0.0000 r_2=2.2441**
Case 4 $d_{ro} \sim d_{io}$?	Y	Y	Filt1 or Filt2	**P=0.0000 r_2=1.1219**	**p=0.0000 r_2=1.0709**	**p=0.0000 r_2=1.0708**	p=0.1376 r_2=1.5367
Total Closest				**0**	**4**	**6**	**0**

Table 5.5. d_{ro} vs d_{io} values for different objective measures and different speech processing algorithms

Similar convention is used in Tables 5.5 and 5.6. According to Table 5.5, LSF distance using inverse harmonic weighting performs the best when d_{ro} and d_{io} values are

considered. Table 5.6 shows that spectral distortion, frequency weighted spectral distortion, and weighted LP cepstral distance measures perform better when d_{io} and d_{ri} values are considered. Counting the total closest r_1, r_2, and r_3 values assigned to each distance measure, we obtain Table 5.7. Table 5.7 shows that LSF distance is assigned the maximum number of closest r_i values. Therefore, we decided to use LSF distance using inverse harmonic weighting for objective evaluations of vocal tract transformation performance.

Case # and condition	Input	Ref	Processing Algorithm	p-values and $r_3=d_{io}/d_{ri}$ ratios for objective distances			
				SD	**SD_{fw}**	**LD**	**WCeps**
Case 1 d_{io} ~ d_{ri}?	X	Y	Transp1	**p=0.0000 r_3=1.4046**	**p=0.0000 r_3=1.3590**	**p=0.0000 r_3=1.0634**	**p=0.0002 r_3=1.5685**
Case 2 d_{io} <<d_{ri}?	X	Y	Transp2 or Transp3	p=1.0000 r_3=1.8978	p=1.0000 r_3=1.9001	p=1.0000 r_3=1.8941	p=0.9894 r_3=2.0781
Case 2 d_{io} <<d_{ri}?	X	Y	Pscale1 or Pscale2	p=0.9999 r_3=1.1303	p=0.9830 r_3=1.0648	p=1.0000 r_3=2.7230	p=0.9336 r_3=1.4592
Case 2 d_{io} <<d_{ri}?	X	Y	Tscale1 or Tscale2	p=1.0000 r_3=1.2041	p=1.0000 r_3=1.2219	p=1.0000 r_3=2.0426	p=0.9998 r_3=2.2818
Case 2 d_{io} <<d_{ri}?	X	Y	Noise1 or Noise2	p=1.0000 r_3=1.3112	p=1.0000 r_3=2.3523	p=1.0000 r_3=1.2920	p=0.9996 r_3=2.2178
Case 3 d_{io} ~ d_{ri}?	Y	Y	Pscale1 or Pscale2	**p=0.0000 r_3=1.1344**	**p=0.0000 r_3=1.0870**	**p=0.0000 r_3=2.2890**	**p=0.0000 r_3=3.2788**
Case 3 d_{io} ~ d_{ri}?	Y	Y	Tscale1 or Tscale2	**p=0.0000 r_3=1.1453**	**p=0.0000 r_3=1.1458**	**p=0.0000 r_3=1.8342**	**p=0.0000 r_3=2.4827**
Case 3 d_{io} ~ d_{ri}?	Y	Y	Noise1 or Noise2	**p=0.0000 r_3=1.5489**	**p=0.0000 r_3=3.0871**	**p=0.0000 r_3=1.1400**	**p=0.0000 r_3=3.8230**
Case 4 d_{io} >> d_{ri}?	Y	Y	Vscale1 or Vscale2	**p=0.0000 r_3=1.1786**	**p=0.0000 r_3=1.2414**	p=1.0000 r_3=0.7455	**p=0.0000 r_3=2.1059**
Case 4 d_{io} >> d_{ri}?	Y	Y	Filt1 or Filt2	p=1.0000 r_3=0.7969	p=1.0000 r_3=0.6585	p=1.0000 r_3=0.4188	p=0.7070 r_3=0.8572
Total Closest				**1**	**2**	**2**	**0**

Table 5.6. d_{io} vs d_{ri} values for different objective measures and different speech processing algorithms

Distance Measure	**SD**	**SD_{fw}**	**LD**	**WCeps**
Total Closest	1	9	15	0

Table 5.7. Total closest r values that satisfy the corresponding requirements in Tables 5.4, 5.5, and 5.6 for each objective distance value

Table 5.8 shows the total number of triples used for each processing algorithm.

Case	Input Speaker	Ref. Speaker	Algorithm	Total Triples
1	X	Y	Transplantation from Y onto X's residual (Transp1)	100
2	X	Y	Transplantation from Y onto X's residual with mixing (0.4) or too much smoothing (20) (Transp2, Transp3)	30 (15+15)
2	X	Y	Pitch scaling of X (0.6, 1.8) (Pscale1, Pscale2)	30 (15+15)
2	X	Y	Time scaling of X (0.6, 1.8) (Tscale1, Tscale2)	30 (15+15)
2	X	Y	Noise addition to X (10 and 20dB SNR) (Noise1, Noise2)	30 (15+15)
3	Y	Y	Pitch scaling of X (0.6, 1.8) (Pscale1, Pscale2)	40 (20+20)
3	Y	Y	Time scaling of X (0.6, 1.8) (Tscale1, Tscale2)	40 (20+20)
3	Y	Y	Noise addition to X (10 and 20 dB SNR) (Noise1, Noise2)	40 (20+20)
4	Y	Y	Vocal tract scaling (0.6, 1.8) (Vscale1, Vscale2)	40 (20+20)
4	Y	Y	Filtering (LPF 2 KHz, BPF, 2-4 KHz) (Filt1, Filt2)	40 (20+20)

Table 5.8. Total number of triples for each processing algorithm

5.5. Evaluations

5.5.1. Objective Test: Comparison with the Baseline

In order to test the vocal tract transformation function estimation described in this chapter, an objective cross-lingual voice conversion test is performed. The target speaker was a compound bilingual female American English and Turkish speaker. The source speaker was a male Turkish speaker who can speak English with foreign accent. 100 utterances in English were used in training. For the tests, 10 utterances in Turkish were transformed using context-matching and the baseline algorithm. The average LSF distance between the transformation outputs and the target speaker recordings is

computed. For providing a baseline, the average LSF distance between the vocal tract transplantations and the target recordings was also computed.

The results are shown in Figure 5.2. We observe that the proposed method results in 0.5 LSF distance reduction as compared to the baseline method. We also note that the difference between the source and the target speaker is larger in this example because of gender difference.

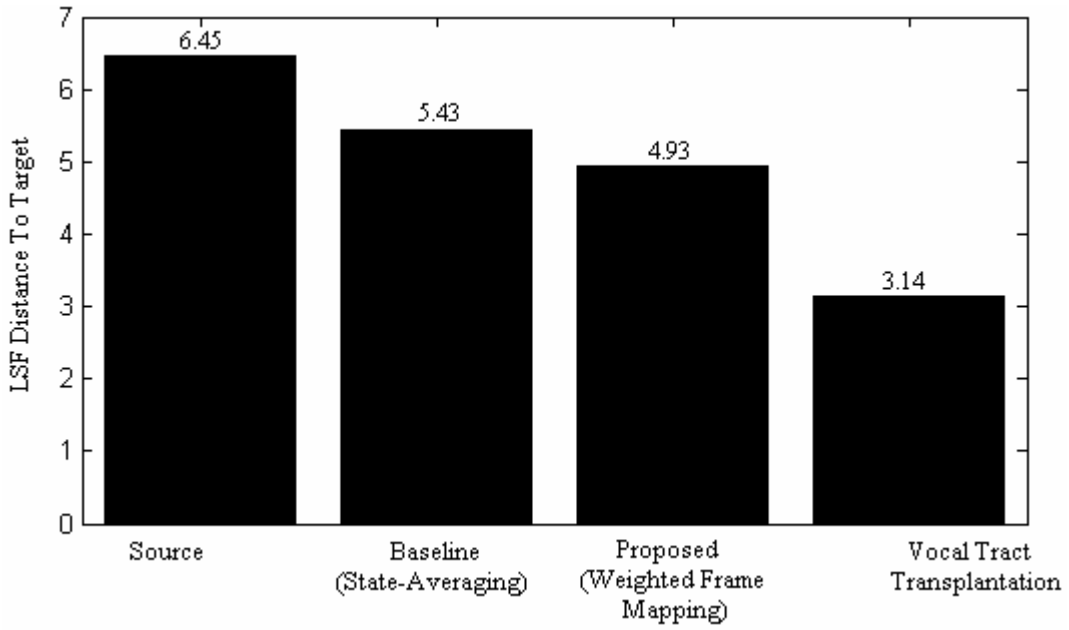

Figure 5.2. Result of objective test for the proposed vocal tract transformation function estimation method

Figure 5.3 shows samples of vocal tract spectra converted using the baseline method and the proposed method along with the corresponding target vocal tract spectra.

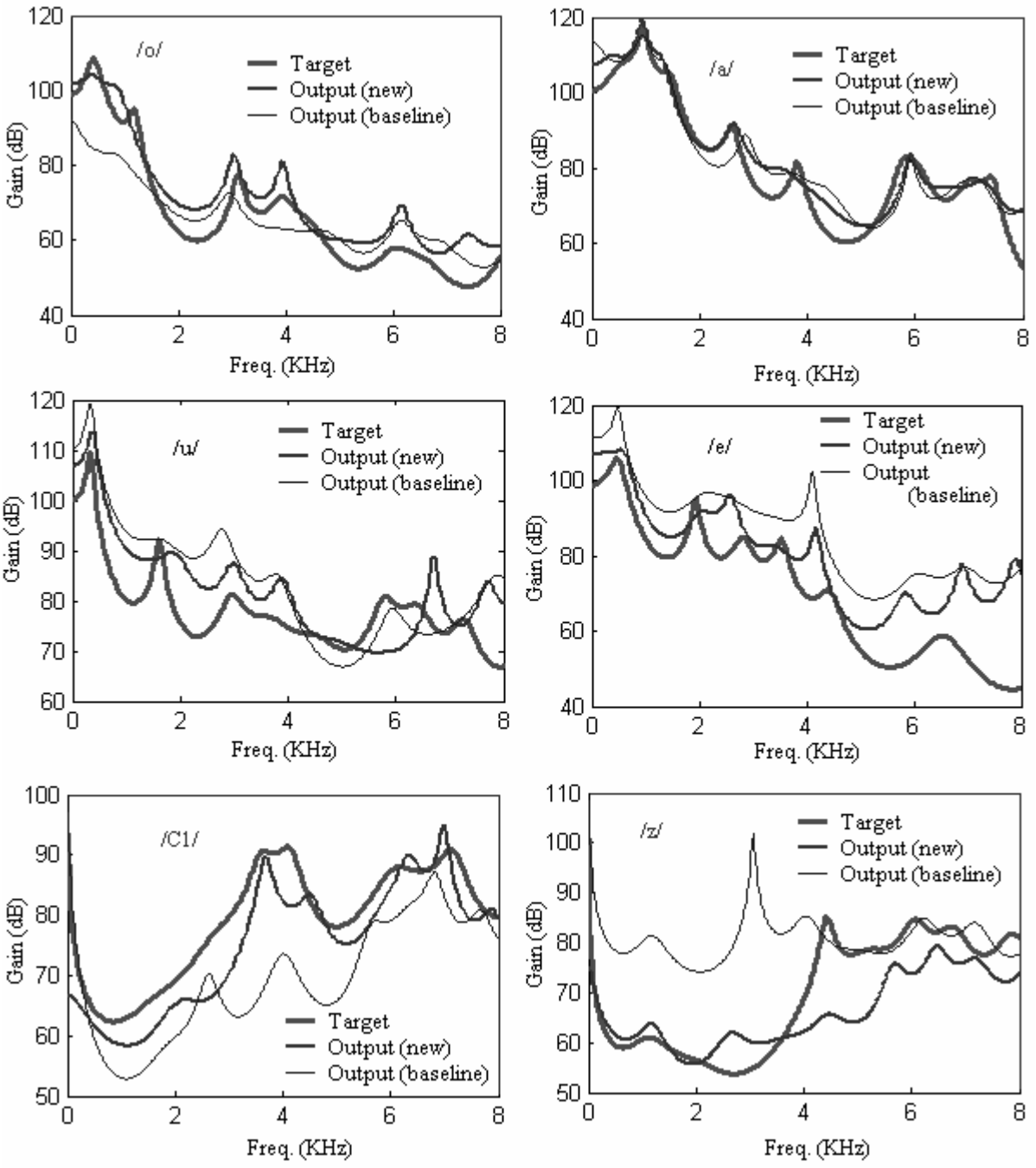

Figure 5.3. Examples of vocal tract spectra transformed using the new method, baseline method, and the corresponding target spectra

5.5.2. Subjective Test: Comparison of Parallel and Non-Parallel Training

The vocal tract transformation techniques described in this chapter enable the employment in non-parallel databases for voice conversion. Using non-parallel training, it is even possible to use source and target training databases that are in different languages. We have designed a subjective listening test to compare the performance of non-parallel with that of parallel training. Two types of non-parallel and a parallel training and transformation sessions were carried out:

- Parallel (P1): Training with 50 identical source and target utterances in English, transformation of four source utterances in Turkish using the baseline method
- Parallel (P2): Training with 50 identical source and target utterances in English, transformation of four source utterances in Turkish using the proposed method
- Non-parallel (NP1): Training with 50 non-identical source and target utterances in English, transformation of four source utterances in Turkish using the proposed method
- Non-parallel (NP2): Training with 50 source utterances in Turkish and 50 target utterances in English, transformation of four source utterances in Turkish using the proposed method

The target speaker was a male, native American English speaker and the source speaker was a male, bilingual Turkish and American English speaker. For all transformations, the mean of the pitch is transformed to match the target speaker's average pitch. As the source and the target speakers had close average pitch values (~112 Hz), the pitch scaling amount was fairly close to 1.0 in all cases. The subjects were presented with an output and a target recording and were asked to score the level of similarity to the target speaker's voice on a scale from 1 to 5. A score of "1" corresponds to minimum similarity to target and "5" corresponds to maximum similarity. As there were four methods and four outputs for each method, each of the

six subjects listened to 16 output and target pairs. Figure 5.4 shows the results of the similarity test.

The results indicate that the proposed method performs the best in terms of similarity to the target voice when parallel training is employed. The similarity to the target voice is reduced when non-parallel databases are used. This result is expected since the mapping between the source and the target speaker is likely to be better resulting in better estimation of the vocal tract transformation function in the case of parallel training. NP2 corresponds to the most difficult case because one speaker's phonemes in a specific language need to be matched to another speaker's phonemes in another language in the transformation stage. This results in a significant reduction in the similarity to the target speaker's voice. NP2 has the advantage of employing source target and training data in the same language. Therefore, the mapping of the source training and transformation data is an easier problem. However, it seems that once there are problems in training, the cross-lingual voice conversion algorithm is not able to cover from these errors.

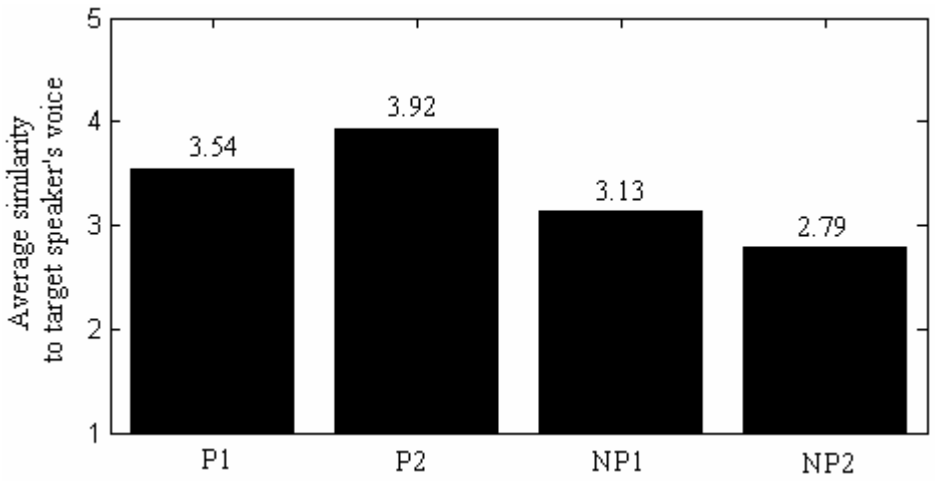

Figure 5.4. Results of the subjective similarity test

In a second subjective test, we have evaluated the MOS-based quality of the output signals. For each training method, six subjects were presented with four output recordings. Therefore, the subjects have listened to 16 outputs in total. Prior to test,

they were provided with the reference set of recordings in Table 5.9 to provide a reference in their judgments. The standard mean opinion scale is used for the judgments on quality as given in Table 5.10. The results of the MOS-based quality test are shown in Figure 5.5.

Coder or Recording format	**Bit rate (Kbps)**	**MOS**
PCM	64	4.4
ADPCM (G.726)	32	4.2
LD-CELP (G.728)	16	4.2
CSA-CELP (G.729)	8	4.2
CELP	4.8	4.0
LPC-10 (FS 1015)	2.4	2.3

Table 5.9. Reference set for the MOS test

MOS	**Meaning**
5	Very good quality. There is no noise, the conversation is clearly and distinctly understood.
4	Good quality. The noise does not disturb, the conversation is distinctly understood.
3	Normal quality. The noise disturbs a little, the conversation can be understood.
2	Low quality. The noise is disturbing but the conversation can be understood.
1	Very bad quality. The noise is very disturbing and the conversation can not be understood.

Table 5.10. Mean Opinion Score (MOS) scale on speech quality

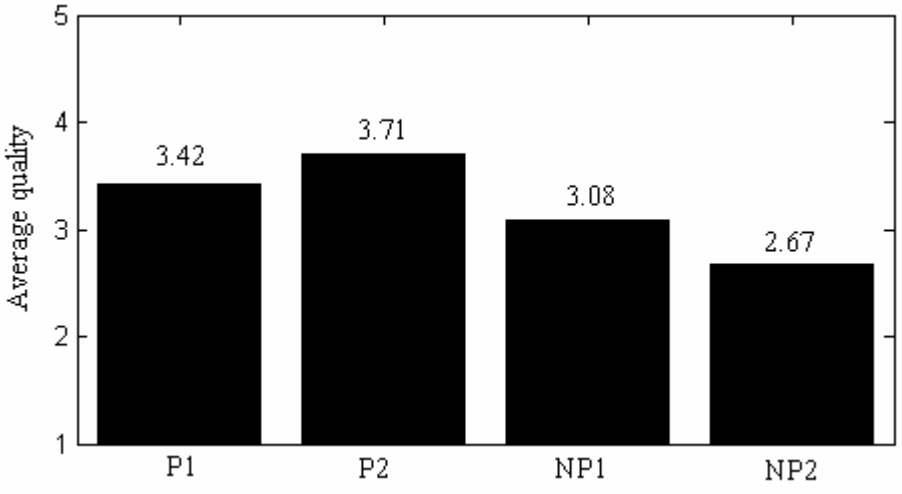

Figure 5.5. Results of the MOS-based quality test

In terms of quality, the proposed method using parallel training resulted in the best performance. It is followed by the baseline method. The proposed method enables non-parallel training which resulted in significant quality reduction. Therefore, there is still room for improvement in the automatic mapping of training databases as well as reliable estimation of the vocal tract transformation function in the case of non-parallel training.

6. STYLISTIC PROSODY TRANSFORMATION

6.1. Introduction

As prosody provides significant clues on a speaker's identity, a robust prosody modeling and modification module is an important component of a voice conversion system. Applications of prosody modeling, modification, and transformation are not only limited to voice conversion research. More detailed modeling of prosody resulted in significant improvements in TTS (van Santen, 1994), (Syrdal, et. al., 1998a), speaker identification (Sonmez, et. al., 1998), and speech recognition (Shriberg and Stolcke, 2004). Concatenative text-to-speech synthesis systems use prosody modification techniques to synthesize speech in a target prosodic setting (Syrdal, et. al., 1998b) as well as to perform smoothing during concatenation (Bozkurt, et. al., 2002). Prosody transformation provides a useful framework for investigating the effects of modifying prosody parameters in a controlled manner. Therefore, emotion research is another application field in which the relationship between perceived emotions and prosodic features are being investigated (Burkhardt, et. al., 2006).

Pitch, duration, and energy are the most prominent factors in prosody perception. In order to make the voice conversion output sufficiently close to the target voice, it is required to perform sufficient amount of prosody modification to match these characteristics. Among the three factors, pitch and duration are relatively more important in speaker identity perception (Ormanci, et. al. 2002). Modification of pitch and duration require application of techniques including phase vocoding (Flanagan and Golden, 1966), time domain or frequency domain pitch synchronous overlap-add (Moulines and Verhelst, 1995), and sinusoidal model based modification (Quatieri and McAulay, 1992). PSOLA based methods have become a popular choice in voice modification research since they enable modification of the pitch and the duration simultaneously in a robust manner. The frequency domain version of PSOLA is also appropriate for modifying other characteristics directly such as the vocal tract spectrum, or formant structure.

State-of-the-art voice conversion algorithms employ transformation of the long-term statistics of prosodic features such as the mean and the variance of f0, and the overall speaking rate. These modifications are able to make the average prosodic characteristics match those of the target speaker. However, they can not reliably model and transform prosody in more detail including local shapes and movements of the pitch contours, rhythm, and local speaking rate. Lack of a detailed prosody modeling and modification module can be a major drawback for cross-lingual applications. In cross-lingual voice conversion, the source and the target training databases are collected in one language and transformations are performed using a source database collected in another language. Significant differences in the source and the target prosody characteristics are likely to occur due to accent differences. Therefore, prosody transformation techniques to match average target prosodic characteristics might not result in sufficient similarity to the target speaker's style. This is particularly important for accent transformation in which the aim is to generate the target speaker's style in a different language. For example, consider a dubbing application in Turkish language where the aim is to make an American celebrity speak Turkish with American accent using cross-lingual voice conversion. In this particular example, having the target speak the transformation language with accent might make the voice conversion output more natural since listeners would not expect to hear perfect native-accented Turkish from that celebrity voice. In this case, conventional voice conversion algorithms can only be employed if a source speaker who is native in American English but can speak Turkish with accent could be found. It might be hard to find such source speakers especially when the transformation language is not a very common one. An alternative is to use a native Turkish speaker as the source who can speak American English as well and to modify his/her style to match the target prosody characteristics in the training language. In this chapter, we develop prosody transformation techniques that can be used for this purpose.

All prosody modification algorithms result in processing distortion especially when large amounts of modifications are performed. In order to perform detailed prosody modification without causing additional processing distortion, it is required to estimate the modification amounts carefully by avoiding large modification factors or

discontinuities at the output. In this chapter, we describe a new algorithm for stylistic prosody transformation. The algorithm consists of two stages: Stylistic pitch transformation and stylistic speaking rate transformation. Stylistic pitch transformation models and transforms the slopes of the pitch contour segments while trying to reduce the amount of pitch scale modification required. Stylistic speaking rate transformation modifies the speaking rate and rhythm by time invariant and time varying duration scaling as well as by expansion or contraction of pauses in the speech signal. We demonstrate the proposed algorithm in the evaluations for a cross-lingual voice conversion task. We also compare its performance with standard prosody transformation techniques. The results show that stylistic prosody transformation can generate closer output to the target voice with comparable quality to the standard methods.

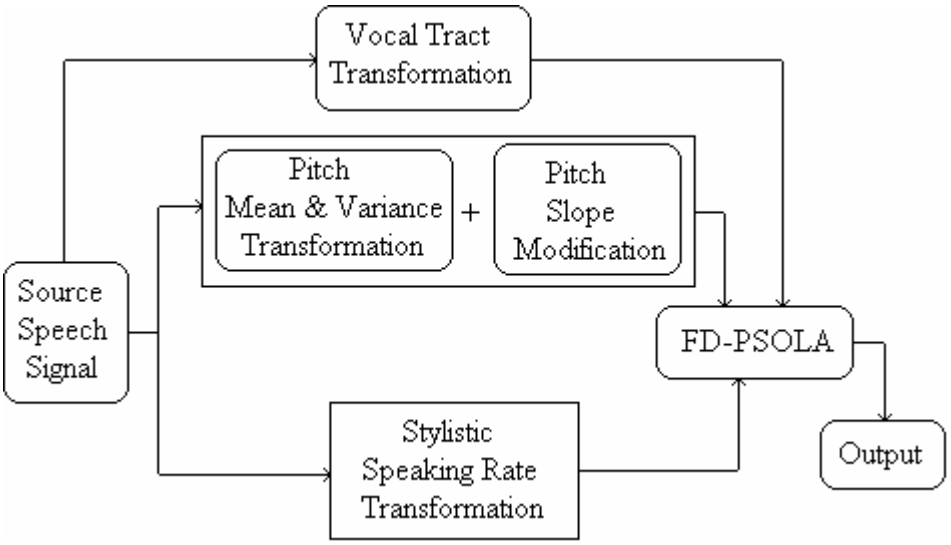

Figure 6.1. Flowchart of the proposed stylistic prosody transformation algorithm

Figure 6.1 shows the general flowchart of the proposed method. In Section 6.2, two algorithms for stylistic pitch transformation are described. The first algorithm fits least-squares lines to source and target pitch contours in the training database and estimates additional amounts of pitch modification during transformation to perform sentence-level pitch slope transformation. The second algorithm described in Section 6.2, is an extension of the first algorithm to fit least squares lines on voiced segments of

the source and the target pitch contours. It employs context information to estimate the average pitch slope modification amount to match the target pitch slope patterns in more detail. Estimating the slope modification factors on a voiced segment basis enables to reduce discontinuities in pitch manipulation. In Section 6.3, a new method for modifying the speaking rate is described. It involves duration modification with time invariant and time varying duration scaling as well as modification of the pauses between utterances to match the target speaking rate better.

6.2. Pitch Transformation

6.2.1. Conventional Pitch Transformation Methods

State-of-the-art voice conversion algorithms perform pitch transformation using a variance scaling and mean shifting approach to match the target f0 mean and variance (Arslan, 1999). In this approach, f0 values are assumed to be normal distributed. The distribution parameters (mean and variance) can be easily estimated from the training pitch contours. In the transformation stage, a time varying pitch scaling factor is determined from the instantaneous source f0 value $f_s(t)$, and source and target f0 distribution parameters using:

$$p(t) = \frac{f_s(t)\frac{\sigma_t}{\sigma_s} + \mu_t - \mu_s\frac{\sigma_t}{\sigma_s}}{f_s(t)} \tag{22}$$

where $f_s(t)$ is the instantaneous f0 value in the source transformation utterance, μ_s and μ_t are the mean of the source and target training f0 values, s_s and s_t are the variance of the source and target training f0 values, and $p(t)$ is the instantaneous amount of pitch scaling required for matching the target mean and variance. The limitation of this approach is that the local differences in the source and target pitch contour patterns can not be modeled and transformed. However, these differences may contain important information on a speaker's style.

On the other extreme, very detailed pitch transformation can be performed by an approach that replaces the whole source f0 contour with an estimated target contour. Examples of this approach include (Chappel and Hansen, 1998), and (Turk and Arslan, 2003). The estimation of the target f0 contour can be performed by matching the source input f0 contour with the training contours, finding an optimal match and transforming the pitch to match the corresponding target f0 contour. In (Turk and Arslan, 2003), we have developed a method to perform target f0 contour estimation in a weighted manner. However, this approach has a major drawback: When the estimated target f0 value is significantly different from the instantaneous source f0 value, large amounts of pitch modification is required to match the target f0 contour. This results in quality reduction as pitch modification algorithms can typically perform well for low to medium amounts of modification but fail to produce natural output for larger modification amounts. As an example, PSOLA based techniques which are employed in this study perform considerably well for pitch scaling factors in the range 0.7 to 1.5. The quality and naturalness may decrease for pitch scaling factors out of this range.

6.2.2. Stylistic Pitch Contour Modeling and Transformation

In order to avoid the shortcomings of the two pitch transformation approaches summarized in the previous section, we propose to model and transform the sentence and segment level slopes of the pitch contours for the source and the target speakers as an additional component to mean and variance transformation. Special care is taken to determine pitch modification factors to minimize the amount of pitch scaling required. Note that parallel source and target training data in the form of sentence utterance recordings are required for the proposed method.

The flowchart of the sentence-level pitch slope modeling and transformation algorithm is shown in Figure 6.2. The algorithm consists of two stages. In the training stage, source and target recordings of identical utterances are collected. Pitch contours are extracted with the RAPT algorithm (Talkin, 1995). The pitch contours are further smoothed with a median filter of 5 frames. The unvoiced regions are linearly interpolated starting from previous voiced frame's pitch value to the next voiced

segment's pitch value. Then, a line is fit to the pitch values using the least squares error criterion as shown in Figure 6.3.

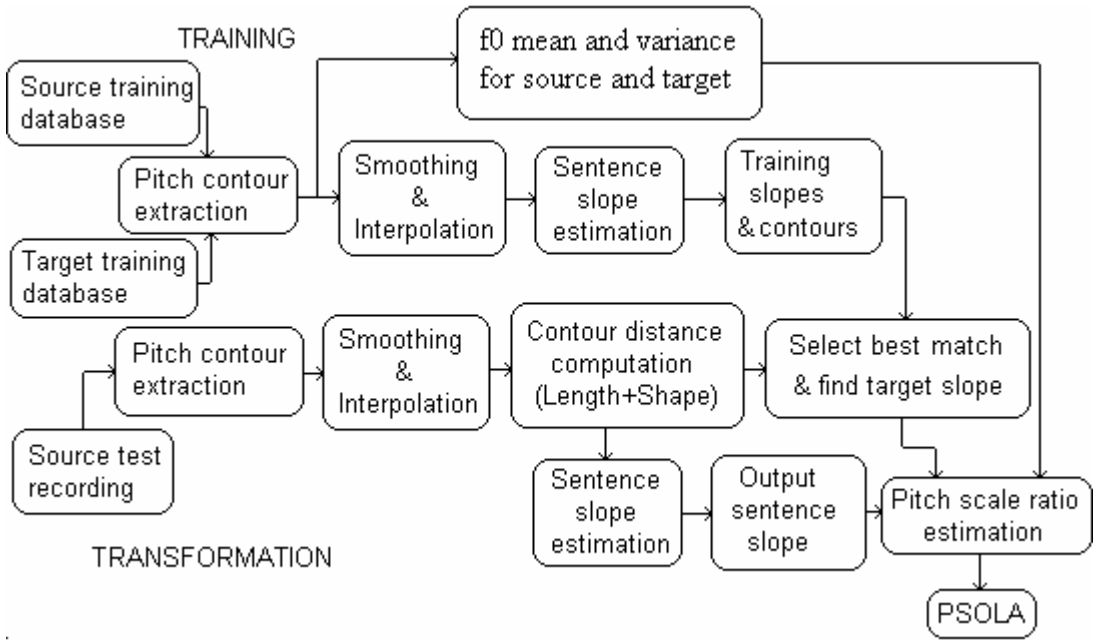

Figure 6.2. Sentence-level pitch slope modeling and transformation algorithm flowchart

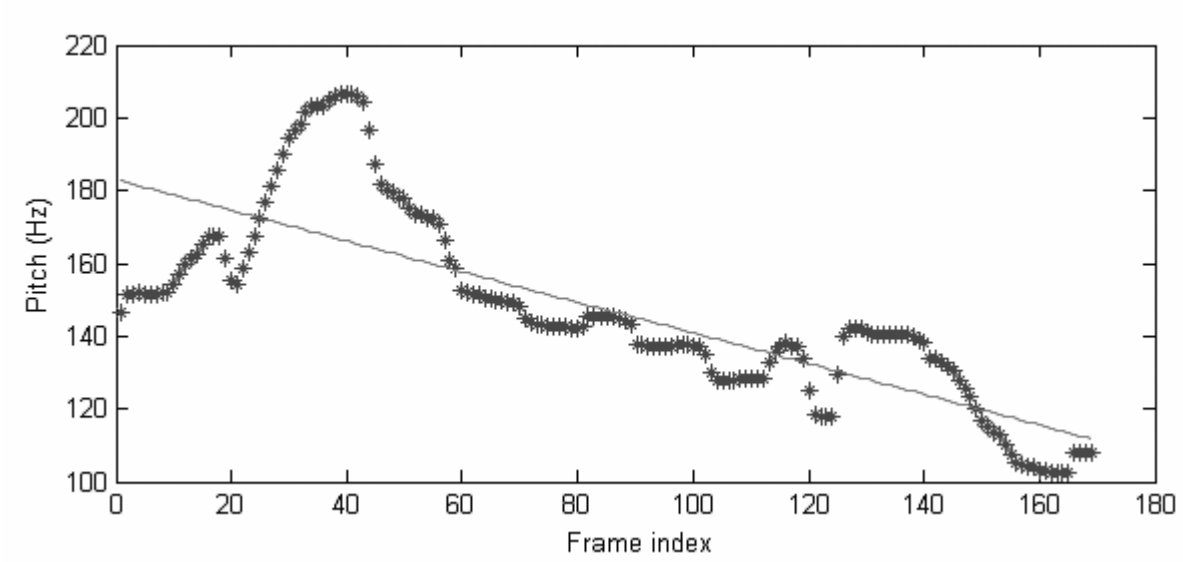

Figure 6.3. Least-squares line fit to the smoothed and interpolated pitch contour

Let K be the total number of parallel source and target utterance pairs. Let also k be the index of each parallel training utterance where $1 = k = K$. The average length of smoothed source training contours is determined using:

$$N = \frac{1}{K} \sum_{k=1}^{K} |p_s(k)| \tag{23}$$

where $p_s(k)$ is the smoothed version of the pitch contour of the k^{th} source training utterance. A least-squares line is fit to each smoothed source pitch contour and the corresponding line slopes, $m_s(k)$, are determined. The smoothed source pitch contours are linearly interpolated to the average source training pitch contour length N. $m_s(k)$, the slopes of the lines fitted to the smoothed source pitch contours; $|p_s(k)|$, the original length of each contour; and $x_s(k)$, the smoothed and linearly interpolated versions of the source training pitch contours, are reserved for the transformation stage. The target training pitch contours are also smoothed and a least-squares line is fit to each one of them. $m_t(k)$, the slopes of the least-squares lines fit to target training pitch contours are also reserved for the training stage.

In the transformation stage, identical pre-processing is performed on the source input pitch contour to obtain a smoothed and linearly interpolated pitch contour denoted by x. The distance to each smoothed source training contour is computed using:

$$d(k) = \alpha.d_{shape}(k) + (1 - \alpha).d_{length}(k) \tag{24}$$

where k is the source training contour index, and a is a weighting parameter between contour similarity and contour length. Setting a equal to 0.95 works well in practice. The normalized cross-correlation of the smoothed input pitch contour and the smoothed source training pitch contours are computed using:

$$r(k) = \frac{\sum_{i=1}^{N}(x(i) - \mu_x)(x_s(k,i) - \mu_{x_s(k)})}{\sqrt{\sum_{i=1}^{N}(x(i) - \mu_x)^2 . \sum_{i=1}^{N}(x_s(k,i) - \mu_{x_s(k)})^2}} \tag{25}$$

where *r(k)* is the normalized cross-correlation between the smoothed source input pitch contour *x* and k^{th} smoothed source training pitch contour $x_s(k)$. μ_x and $\mu_{xs(k)}$ are the mean of the f0 values in *x* and x_s respectively. The shape distance is computed using:

$$d_{shape}(k) = \begin{cases} 1.0 - r(k), & if\ r(k) \geq 0 \\ 1.0, & otherwise \end{cases} \tag{26}$$

Note that **f** $d_{shape}(k)$ turns out to be negative, it is set equal to the greatest possible shape distance value of 1.0. The length distance is computed using:

$$d_{length}(k) = \frac{||x| - |x_s(k)||}{max(|x|, |x_s(k)|)} \tag{27}$$

The final pitch contour distance metric is guaranteed to be a continuous value between 0.0 and 1.0. Higher distance values correspond to increased similarity between a given contour and a source training contour. The distance values are exponentially weighted and the weighted values are normalized to sum up to unity using:

$$w(k) = \frac{e^{-\beta d(k)}}{\sum_{k=1}^{K} e^{-\beta d(k)}} \tag{28}$$

ß=5.0 is used in practice. The normalized weights, *w(k)*, are then used to estimate the target sentence pitch slope by weighted averaging of the corresponding target sentence slopes as follows:

$$m = \sum_{k=1}^{K} w(k) m_t(k) \tag{29}$$

where m is the estimated sentence slope, $w(k)$ is the weight of the k^{th} target training contour, and $m_t(k)$ is the corresponding target training sentence slope.

Once the target sentence slope is estimated, a line equation is determined by computing a bias term which makes the middle point of the line and the line fit to the source input contour will intersect using:

$$l_t(i) = (i - \frac{I}{2})m + l_s(\frac{I}{2}) \tag{30}$$

where l_t is the estimated line, l_s is the least squares line fit to the source pitch contour, i is the speech frame index and I is the number of frames in the source input pitch contour. The difference between the two lines is assigned as the pitch scaling factor for that frame:

$$p(i) = l_t(i) - l_s(i) \tag{31}$$

Additional scaling and shifting is applied to the pitch scaling ratios $p(i)$ if mean and variance transformation to target will be applied simultaneously as follows:

$$p^o(i) = (p(i) - \mu_p)\frac{\sigma_t}{\sigma_p} + \mu_t \tag{32}$$

where μ_t and s_t are the mean and the standard deviation of target f0 values estimated from the target training pitch contours, and μ_p and s_p are the mean and the standard deviation of the original pitch scaling factors p. Figure 6.4 shows an example of joint sentence slope, mean and variance transformation.

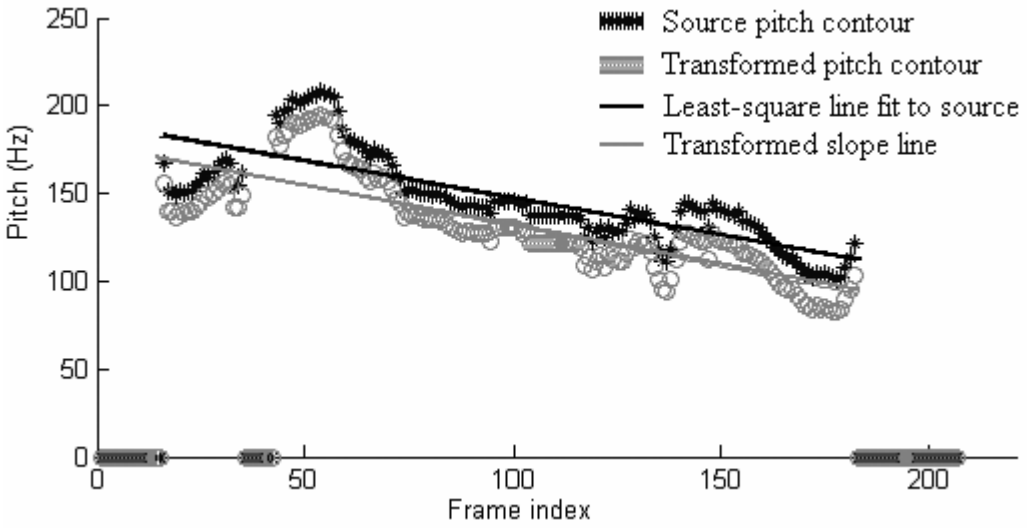

Figure 6.4. Source input pitch contour, least squares line fit to source input pitch contour, estimated target line with mean compensation (red line), and output pitch contour after scaling

Figure 6.5 shows the segment-level pitch slope modeling and transformation algorithm. The pre-processing steps for the segment-level pitch slope modeling and transformation algorithm are identical with the sentence-level algorithm. The main difference of the segment-level algorithm is the modeling of source and target pitch slopes in a local manner rather than at sentence-level. This results in more detailed transformation of the pitch contour movements.

In order to perform segment-level modeling and transformation of the pitch contour slopes, an additional segmentation step is required. Although it is possible to segment a given pitch contour according to pitch accent movements and syllable boundaries, we have used each voiced pitch contour segment as a single unit to model the segment slopes. Using this approach segment slopes can be transformed with less discontinuity as compared to a more detailed model which estimates more than one segment within a voiced pitch contour segment. In the latter case, significantly different pitch slope transformation amounts in neighboring units will cause discontinuities in

the time-varying pitch scale modification factor. Another advantage of using voiced segments as the basic units is the ease of automatic segmentation.

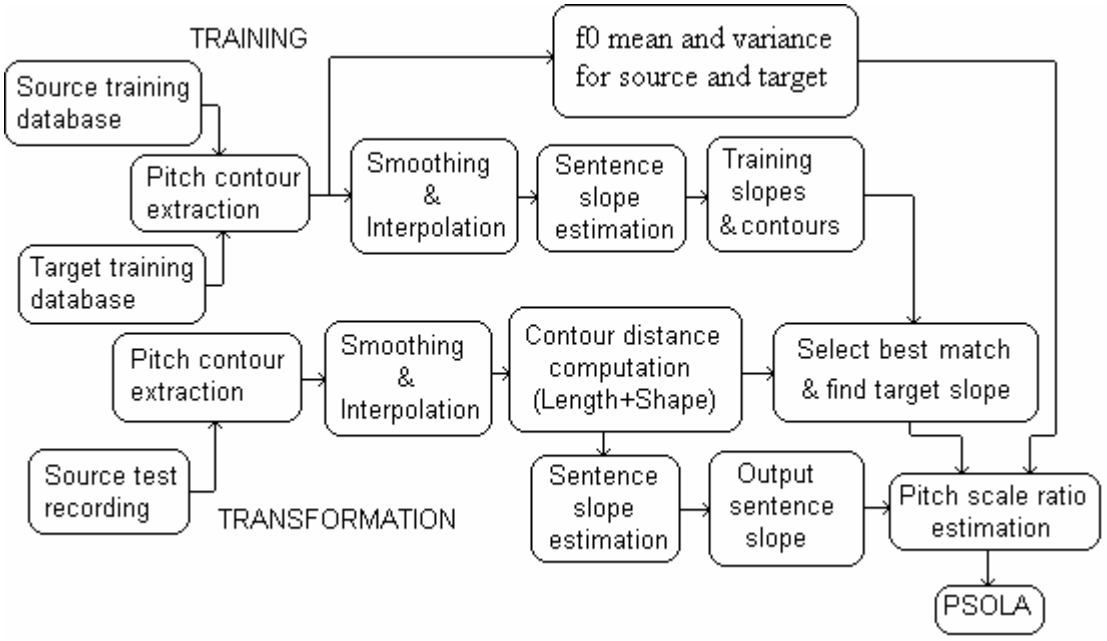

Figure 6.5. Segment-level pitch slope modeling and transformation algorithm flowchart

The segmentation algorithm searches for voiced segments that are separated by unvoiced regions in the median filtered pitch contour. A new segment is assigned only if the previous and next three speech frames are marked as unvoiced. Otherwise, the unvoiced values are linearly interpolated using the neighboring f0 values and the current voiced segment is extended. After all voiced segments are determined, a least squares line is fit to the each segment and the slope of the line is recorded. Depending on the slopes, each segment is classified into three groups:

- Decreasing (D): slope<-0.5
- Monotone (M): -0.5=slope<0.5
- Increasing (I): 0.5=slope

For all combinations of previous J source segment slope classes, current source segment slope class and the corresponding target segment slope class, the average

target slopes and the probability of observing that source-target slope sequence is computed. *J* is set to 0, 1, and 2 respectively. Note that for *J*=0, no previous source segment slope class context is used. For *J*=1, the statistics for the source slope sequences and the corresponding target sequence given in Table 6.1 are computed.

Source		**Target**		**Source**		**Target**
Previous segment	**Current segment**	**Current segment**		**Previous segment**	**Current segment**	**Current segment**
D	D	D		M	M	I
D	D	M		M	I	D
D	D	I		M	I	M
D	M	D		M	I	I
D	M	M		I	D	D
D	M	I		I	D	M
D	I	D		I	D	I
D	I	M		I	M	D
D	I	I		I	M	M
M	D	D		I	M	I
M	D	M		I	I	D
M	D	I		I	I	M
M	M	D		I	I	I
M	M	M				

Table 6.1. All combinations of source-target slope sequences for *J*=1

In order to determine the target slope given the source slope sequences, the target slope class with highest probability given the source slope sequence for *J*=2 is determined. If the number of sequences observed in the training data is less than a fixed threshold (we used 10 in practice), *J* is decreased by one and the highest probability target slope class is determined again. The process is repeated as required until *J*=0, i.e. no source context is employed. This approach is similar to probability estimation in language modeling using a back-off mechanism when reliable estimates can not be found in the training data.

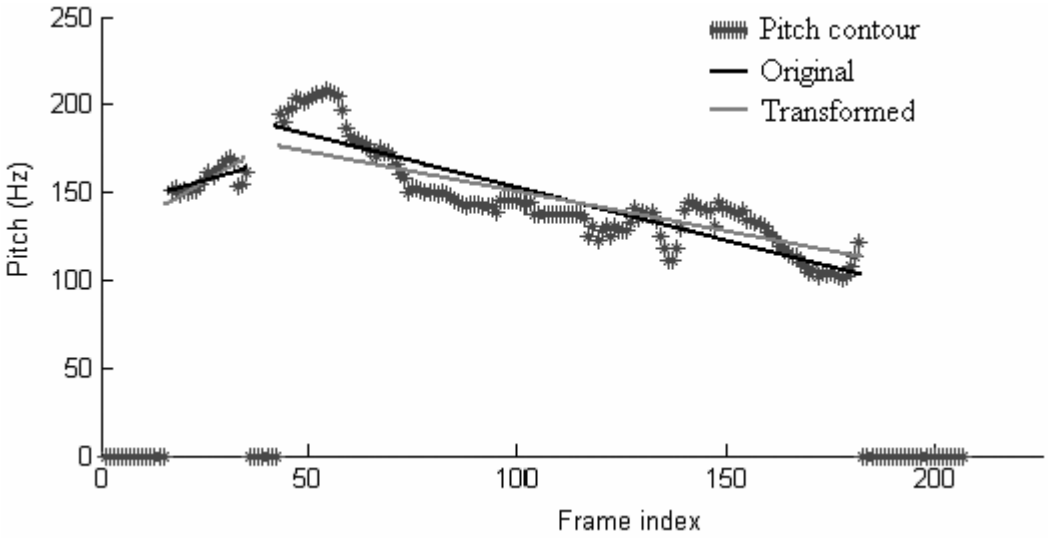

Figure 6.6. Original least-squares lines fit to the segments extracted from the smoothed and interpolated pitch contour and their transformed versions

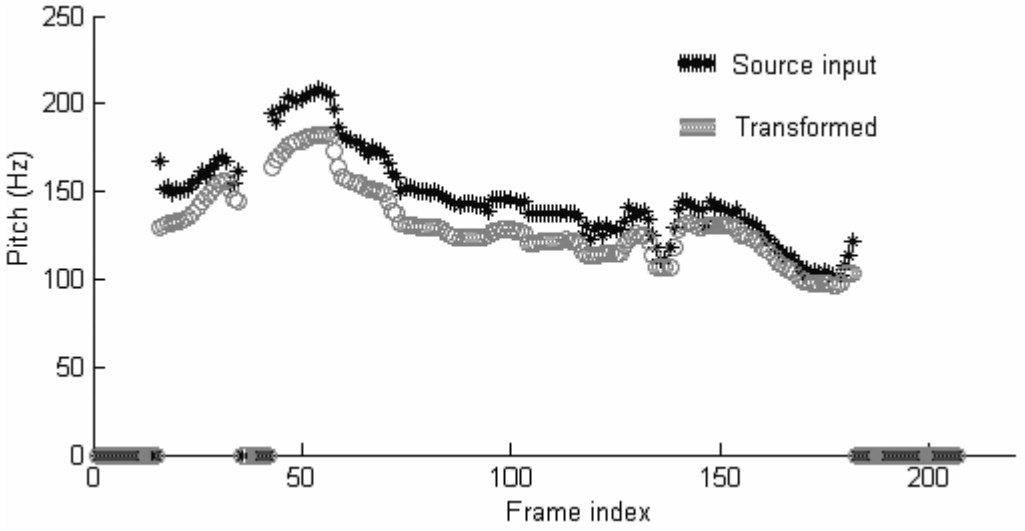

Figure 6.7. Source input pitch contour and the pitch contour after scaling and mean compensation using the segment based approach

Figures 6.6 and 6.7 show the segment slope lines fit to an utterance, their modified versions, and the output pitch contour respectively. In determining the pitch scale amount, we follow the method used in sentence-level pitch slope transformation.

6.3. Speaking Rate Transformation

Speaking rate is an important component of prosody that conveys clues on the identity as well as the emotional state of the speaker. It changes depending on the language background of the speaker, on psychological conditions and emotions, and on whether the speaker has speech and language impairment. As an example, in the sad, happy, and angry modes, people tend to use more variable speaking rate (Yildirim, et. al., 2004). It is well known that second-language speakers tend to have significantly different durational characteristics when compared to native speakers (Arslan and Hansen, 1997), (Tomokiyo, 2000). Hearing-impaired children have slower articulation skills as compared to normal children that results in slower speaking rate (Monsen, 1978), (Osberger and McGarr, 1982). Previous research has shown that speaking rate and other acoustic characteristics are not independent. As an example, Hirata and Tsukada analyzed the change in formant movements with vowel duration for Japanese vowels and showed that the formants of short mid-vowels /e/ and /o/ had significant change with speaking rate whereas the high vowel /i/ resisted to changes with speaking rate (Hirata and Tsukada, 2003). According to Zellner, slow rate speech has major qualitative effects on the speech waveform in French (Zellner, 1998).

Estimating the speaking rate and modifying it in a natural manner with signal processing techniques has potential applications in speech compression, speech recognition, text-to-speech synthesis (TTS), voice conversion, audio watermarking, helping handicapped people, and language education. Faltlhauser, Pfau, and Ruske developed a method for on-line speaking rate estimation with Gaussian Mixture Models and Artificial Neural Networks (Faltlhauser, et. al., 2000). Their phoneme rate estimates had a correlation coefficient of 0.66 with the actual phoneme rates. In a multilingual study, Pellegrino and his colleagues proposed a method based on unsupervised vowel detection for speaking rate estimation in multilingual spontaneous speech. The correlation coefficient between the outputs of the proposed method and the actual speaking rates was 0.84 on the average for 6 languages including English, German, Hindi, Japanese, Mandarin, and Spanish (Pellegrino, et. al., 2004). There has been extensive research targeted at increasing the speaking rate without loss of quality

and intelligibility as a means of reducing storage and bandwidth requirements for speech (Foulke and Sticht, 1969), (Beasley and Maki, 1976), (Duker, 1974). Arons provided a good summary on time-compressed speech with a broad list of relevant references (Arons, 1992). Several researchers focused on speaking rate compensation for speech recognition systems (Okuda, et. al., 2002), (Mirghafori, et. al., 1995). TTS engines employ duration modeling techniques based on sequential rules (Klatt, 1987), (van Santen, 1994), decision trees (Pitrelli and Zue, 1989), and neural networks (Campbell, 1992) to synthesize more natural sounding speech. Foote, Adcock and Girgensohn used time-scale modification in audio watermarking (Foote, et. al., 2003). Automatically slowing down speech without reducing intelligibility can be useful for manual transcription of spontaneous speech and helping speech and language disordered. As an example, Coyle and his colleagues used time-scale modification techniques for slowing down speech for the treatment of verbal apraxia (Coyle, et. al., 2004). In (Demol, et. al., 2004), the authors proposed a time-varying duration scaling algorithm for computer-aided language education applications.

A high-quality speaking rate transformation module is an essential part of a complete voice conversion system. Different speakers have varying speaking rates and timing characteristics due to their linguistic backgrounds, physiology of their vocal tracts, and their emotional states. In this section, we describe an algorithm that can be used to transform the speaking rate without introducing additional processing distortion.

6.3.1. Conventional Speaking Rate Transformation Methods

The overall speaking rate of the target speaker can be matched by applying a time invariant scaling factor once the average speaking rates of the source and the target speakers are estimated. Arslan proposed estimation of the time varying duration scaling factors from the training data (Arslan, 1999). In this approach, the codebook entries that are matched with the current source speech frame are analyzed in terms of durations. The ratio of the target and source state durations are used as an estimate of the local duration modification factor. The durations of the previous and next states are also

considered in order to provide a more robust duration estimate and to reduce the effects of alignment mismatches. This approach works considerably well for monolingual transformations. However, in cross-lingual voice conversion the speaking rate of the source speaker might be different in the training and transformation languages. Therefore, local modification factors estimated from training data in one language may not provide natural sounding output when applied to transform the durations in a recording in another language. Another disadvantage is the requirement for different amounts of duration modification in consecutive speech frames which may result in distortion.

6.3.2. Stylistic Speaking Rate Transformation

In order to minimize the possibility of additional distortion and to match the target speaking rate better, we propose a stylistic speaking rate transformation algorithm. The algorithm performs global transformation of the speaking rate as well as transformation of speech rhythm by analyzing and modifying long pauses to match the target characteristics. This approach results in significantly stable output quality with sufficient similarity to target voice. Figure 6.8 shows the flowchart of the stylistic speaking rate transformation algorithm. It consists of the following steps:

- The average sentence duration is transformed to match the overall target speaking rate.
- The patterns of long pauses in target speech are modeled and transformed.

In order to analyze speaking rate patterns, a separate paragraph is recorded from the target speaker in the training language. A paragraph transformed to the target speaker's voice is recorded from the source speaker in the transformation language. The target training paragraph and the source and target training utterances are used determining the stylistic speaking rate transformation parameters which include:

- A global duration modification factor (r)
- A global pause duration modification factor (d)

- A global pause rate modification factor (*f*)

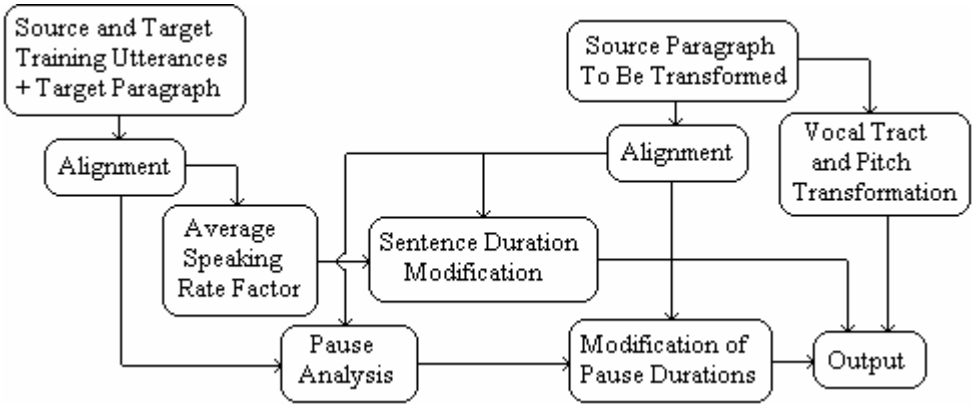

Figure 6.8. Stylistic speaking rate transformation

The global duration modification factor, *r*, is determined as the ratio of the average target training utterance duration to the average source training utterance duration using:

$$r = \frac{\sum_{i=1}^{N} D_t(i)}{\sum_{i=1}^{N} D_s(i)} \tag{33}$$

where N is the number of parallel training utterances, and $D_s(i)$ and $D_t(i)$ are the total duration of the i^{th} source and target training utterances respectively excluding silence in the beginning and at the end of the utterances. This factor is used in fixed amount duration scaling.

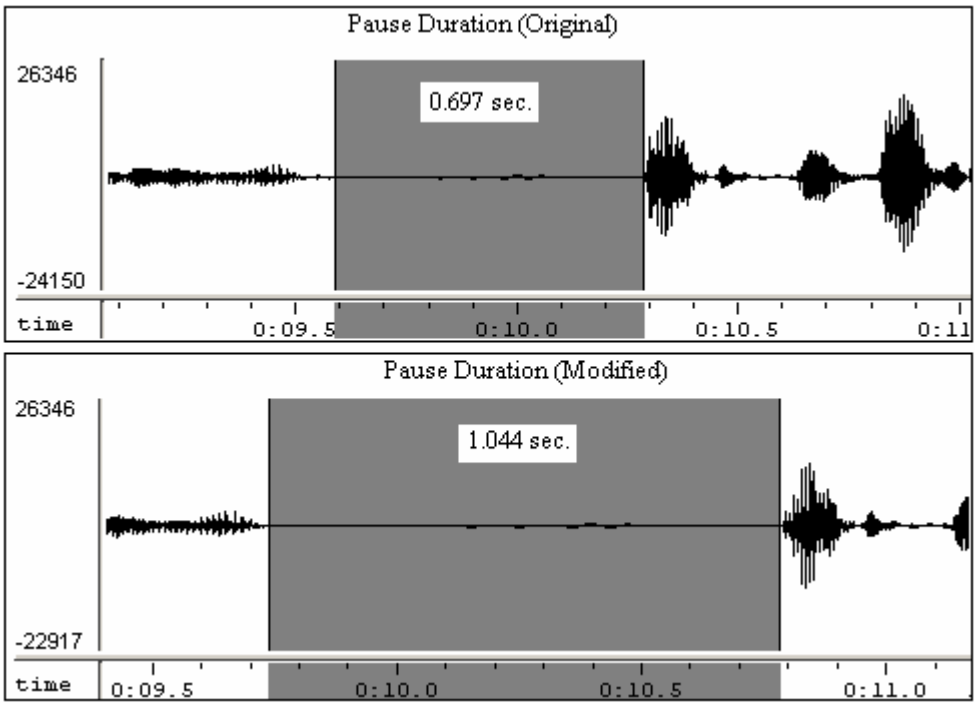

Figure 6.9. Long pause duration modification

The global pause duration modification factor, *d*, is determined using:

$$d = \frac{\frac{(M_s-1)}{(M_t-1)} \sum_{i=1}^{M_t-1} p_t(i, i+1)}{r \sum_{i=1}^{M_s-1} p_s(i, i+1)} \tag{34}$$

where M_s is the number of utterances in the paragraph to be transformed, M_t is the number of utterances in the target training paragraph, $p_s(i,i+1)$ and $p_t(i,i+1)$ are the pause duration between i^{th} and $(i+1)^{th}$ source and target utterances in the corresponding paragraphs respectively. This factor is used in fixed amount scaling of pauses between sentences during transformation. If $d>1.0$, the durations of the long pauses should be expanded and if $d<1.0$, they should be compressed. Figure 6.9 shows an example for long pause duration expansion.

The global pause rate modification factor, *f*, is determined using:

$$f = \frac{\frac{P_t}{D_t}}{r\frac{P_s}{D_s}} \tag{35}$$

where P_s is the total number of long pauses in the source paragraph recording to be transformed and D_s is its total duration in seconds. P_t is the total number of long pauses in the target training paragraph and D_t is its total duration in seconds. f is used in inserting or deleting pauses in the source utterances to be transformed. If f>1.0, pauses need to be inserted and if f<1.0 pauses need to be removed. For inserting pauses, the energy contour is extracted and low energy regions corresponding to silence are determined using the phonetic alignment information also. If an appropriate low energy region is found between two consecutive pauses, a new pause with average target long pause duration is inserted in the middle of the low energy region. If such a region cannot be found, no insertion is performed.

Note that r, the global duration modification should be considered when estimating d and f since global duration modification is first performed by PSOLA. Then, the modifications as required by d and f are applied. First, pause insertion/deletion is applied. Each inserted pause has the average target pause duration. Then, all pauses are extracted/compressed with the global pause duration modification factor d to match the target characteristics. For f>1.0, new pauses are inserted into appropriate locations. For f<1.0 part of the long pauses are deleted. In order to prevent distortion at label boundaries only the middle 80% of the pauses are deleted.

6.4. Evaluations

6.4.1. Correlation Analysis of Pitch Contour Slopes

We have analyzed the differences of sentence and segment-level slope with respect to the level of proficiency in the American English language of the source and the target speakers. For this purpose, we have used 106 phonetically balanced utterances in English collected from two male native American English speakers

(NAT1, NAT2) and one male non-native American English speakers (NON). We extracted the sentence and segment-level slopes as explained in the previous sub-sections. Two source-target pairs are formed as shown in Table 6.2. For each source-target pair, the normalized cross-correlation of the sentence-level and segment-level slopes extracted from the source and the target training data are computed. The results in Table 6.2 show that the correlation between the sentence slopes of the source and the target speaker is lower for the nonnative source speaker. Therefore, the difference between the slope patterns of the nonnative-native source-target pair is larger when compared to the native-native source-target pair. Performing slope transformation using the methods described above, it might be possible to make the nonnative source speaker closer to the target. In the following subjective listening test, we investigate the correctness of this hypothesis.

Source	**Target**	**Source ID**	**Target ID**	**Correlation (Sentence)**	**Correlation (Segment)**
Non-native	Native	NON	NAT2	0.4578	0.4243
Native	Native	NAT1	NAT2	0.7286	0.6167

Table 6.2. Correlation analyses for sentence and segment-level slopes for nonnative-native and native-native source-target speaker pairs

6.4.2. Subjective Test 1: Stylistic Pitch Transformation

In order to examine the effect of the proposed stylistic pitch transformation methods on voice conversion performance, we have designed a forced-choice AB preference test. For this purpose, a male native American English speaker who speaks Turkish with American English accent was employed as the target. The source speaker was a male, compound bilingual Turkish and American English speaker. Five subjects were first presented with ten target recordings in Turkish and were told to focus on the speaker's accent in speaking Turkish. They were then presented with pairs of voice conversion outputs using two different pitch transformation methods. They were asked to select the item in the pair that is closest to the target speaker's accent. There were three different cases concerning pitch transformation:

- P0: No pitch transformation
- P1: Pitch mean and variance transformation
- P2: Pitch mean, variance, sentence slope, and segment slope transformation

Vocal tract transformation was identical for all three cases using the frame weighting based method described in Chapter 4. All six combinations of pairs using the three methods are generated for four utterances in Turkish. Therefore, the subjects have scored 24 pairs in total. Table 6.3 shows the preference percentages of the subjects among all pairs:

Pair	**Result**
P0 vs P1	P1 was preferred over P0 by 75.0%
P0 vs P2	P2 was preferred over P0 by 85.0%
P1 vs P2	P2 was preferred over P1 by 67.5%

Table 6.3. Preference percentages among all pairs

In the second part of the test, the subjects were asked to rate the MOS quality of the three different transformation outputs. For each case, they were presented with four recordings. Therefore, the subjects have listened to twelve outputs in total. Prior to test, they were provided with the reference set of recordings in Table 5.7 to provide a baseline in their judgments. The standard mean opinion scale is used for the judgments on quality as given in Table 5.10. The results of the subjective quality test are shown in Table 6.4. The MOS quality of using no pitch transformation is the highest as expected since there is no additional pitch processing distortion. The MOS-based quality scores for the two pitch transformation methods are fairly close. Therefore, the proposed slope modification strategy does not result in a large quality reduction. Considering the contribution it makes for making the output closer to the target speaker's accent, the slight reduction in quality is negligible.

Method	MOS
P0	3.9
P1	3.5
P2	3.4

Table 6.4. Results of MOS-quality test for stylistic pitch transformation

6.4.3. Subjective Test 2: Stylistic Speaking Rate Transformation

For the evaluation of the proposed stylistic speaking rate transformation algorithm, the same source and target speaker pair used in the subjective tests of Section 6.4.2 was employed. A separate paragraph in English was recorded from the target. The source speaker recorded another paragraph in Turkish and this paragraph was transformed using the proposed stylistic speaking rate transformation algorithm. As the baseline method, we use the time varying duration transformation approach as described in Section 6.3.1 and in (Arslan, 1999). The training set for vocal tract transformation consisted of 50 utterances in English. The vocal tract transformation method described in Chapter 4 was used with identical parameters in both cases. For pitch transformation, only the mean of the target speaker is matched in order to minimize additional distortion from pitch processing. The global duration modification factor was estimated using 50 training utterances for vocal tract transformation. The global pause duration modification factor and the global pause rate modification factor were estimated using the target paragraph recording in English and the source paragraph recording in Turkish.

In the first part of the subjective test, ten subjects were presented with the original target recording for the paragraph in English and the transformed paragraph in Turkish using two duration transformation methods: Time-varying duration transformation and stylistic speaking rate transformation. The subjects were asked to decide which method sounded closer to the target speaker's style in terms of speaking rate and rhythm. They were also allowed to respond that they did not observe any differences. Table 6.5 shows the subject responses. Seven subjects selected the proposed method over the baseline

method and two subjects preferred the baseline method over the proposed method. One subject reported that he did not hear significant difference between the two methods.

	Baseline	**Proposed**	**No Difference**
Total Preferred	2	7	1

Table 6.5. Subject preferences between the baseline and the proposed speaking rate transformation algorithms

In the second part of the test, ten utterances from each paragraph were presented to five subjects for MOS scale based quality assessment. The reference set along with the corresponding MOS values given in Table 5.7 was presented to the subjects prior to the experiment. They assessed the quality of the outputs using the instructions given in Table 5.10. Table 6.6 shows the results. The sentences extracted from the paragraph using the stylistic speaking rate transformation algorithm were assigned slightly higher MOS values. However, the difference is not very significant and we can conclude that the proposed method and the baseline method result in comparable quality.

	Baseline	**Proposed**
MOS	3.6	3.7

Table 6.6. MOS values for baseline and proposed duration transformation algorithms

7. EVALUATIONS

7.1. Database

A cross-lingual database is designed for evaluations in this study. The database consists of recordings of native American English speakers who can read Turkish texts, native Turkish speakers who speak American English with foreign accent, and compound bilingual speakers who can speak both language without foreign accent.

As the text material, 240 utterances in English were written down first. The utterances contained sentences that are commonly used in daily-life and easy to read, i.e. “I had a cheese sandwich for breakfast”. Special care was taken to cover each phoneme in the SAMPA phoneme set for American English at least three times. An English language teacher checked and corrected all the utterances in terms of semantics and grammar. TIMIT phonetic transcriptions are obtained by using the TIMIT (Garofolo, et. al., 1990) and CMU (CMU, 1996) pronunciation dictionaries. The transcriptions are converted to SAMPA by using the conversion table given in Appendix B.

As the test set, 40 utterances in Turkish were written down that cover all phonemes of the Turkish SAMPA phoneme set. The SAMPA transcriptions were obtained by using tables in the Appendix B. As Turkish is a phonetic language, no pronunciation dictionary was required for the conversion. The list of English training, English transformation and Turkish transformation sets are given in Appendix A.

A set of twenty utterances in English and five utterances in Turkish were selected for trial recording sessions. The trial sessions were performed at a private language school. Five female and five male English teachers were selected as the native American English target voices. The teachers were also able to speak Turkish at different levels of proficiency. The trial session was performed in order to do initial voice conversion tests and to evaluate the proficiency of the speakers in Turkish

language. Among these ten speakers, three male and three female speakers who can speak Turkish better were selected as the final target set.

As the source speaker set, two groups of speakers were recorded: Native Turkish speakers who can speak American English with foreign accent (three female, three male), and compound bilingual American English and Turkish speakers (two female, one male). Two paragraphs were recorded from all source speakers: A paragraph in English describing New York City, and another paragraph in Turkish describing Istanbul. The target speakers also recorded the paragraph in English. These paragraphs were used for prosody transformation tests as described in Chapter 6. The final cross-lingual voice conversion database consisted of the source and target speakers given in Table 7.1.

Speaker Type	**Proficiency in Training Language**	**Proficiency in Transformation Language**	**# Female**	**# Male**	**Total**
Source	L2 Advanced	Native	3	3	6
Source	Native	Native	2	1	3
Target	Native	L2 Beginner	3	3	6

Table 7.1. Speakers in the cross-lingual voice conversion database

The final recordings were collected in an acoustically isolated recording room at a sampling rate of 44100 Hz and were stored as 16-bit, mono, PCM files. A set of high quality recording equipments were used:

- M-Audio Fast Track USB sound card
- Rode NT2-A multi-directional condenser microphone with a windscreen that prevents pops
- TubePre microphone pre-amplifier
- Phillips HP95 headphones

- A PC with an LCD monitor, mouse, and keyboard inside the recording room and the hardcase outside the room in order to prevent fan noise from the computer
- High quality XLR microphone cables

7.2. Subjective Test 1: Effect of Source Speaker Proficiency in Training and Transformation Languages on Performance

In our previous work, we have focused on the problem of source speaker (donor) selection from a set of available speakers that will result in the best quality output for a specific target speaker's voice (Turk and Arslan, 2005). For this purpose, we have collected a well-controlled monolingual voice conversion database consisting of 20 native Turkish speakers (10 male, 10 female). 180 conversions that cover all male-to-male and female-to-female voice conversion combinations were performed using a codebook mapping based method. A listening test was carried out in order to determine the subjective scores for similarity of the output to the target speaker's voice and the output quality. The results indicated that selecting the appropriate source speaker is likely to improve monolingual voice conversion performance.

In the case of cross-lingual voice conversion, we expected to observe performance variation depending on the proficiency of the source and the target speakers in the training and transformation languages. In order to examine the effect of source speaker proficiency on training and test languages on voice conversion performance, a subjective listening test is designed. The cross-lingual voice conversion database described in Section 7.1 was employed. The training language was American English and the transformation language was Turkish. Three types of source speaker voices were transformed to native American English target speakers' voices using the vocal tract transformation method described in Chapter 4:

- S1: Native American English, L2 Turkish speakers
- S2: Compound bilingual Turkish and American English speakers
- S3: Native Turkish, L2 American English speakers (3 female, 3 male)

There were three female and three male target speakers. Only male-to-male and female-to-female transformations were considered in order to reduce distortion due to excessive amounts of pitch scaling for prosody transformation. The source-target combinations are shown in the following table.

Source Type	# Sources	# Targets	# Source-Target pairs
S1	2F, 2M	3F, 3M	2Fx3F + 2Mx3M = 12
S2	2F, 1M	3F, 3M	2Fx3F + 1Mx3M = 9
S3	3F, 3M	3F, 3M	3Fx3F + 3Mx3M = 18
		TOTAL	39

Table 7.2. Source-target combinations (F: Female, M: Male)

Training was performed using 50 identical utterances in American English for all source and target pairs. Then, one utterance in Turkish was transformed using the vocal tract transformation method described in Chapter 4. For prosody transformation, only mean pitch is adjusted to match the target mean.

The subjects were presented with triples of sound recordings which contained two voice conversion outputs with different source speaker types and a target recording in American English. They were asked to select the voice conversion output that sounds closer to the target speaker's voice. For each of the three male and three female target speakers, S1-S2, S1-S3, and S2-S3 combinations are prepared as shown in the following table.

Gender	S1-S2	S1-S3	S2-S3	TOTAL
Male-to-male	9	9	9	27
Female-to-female	9	9	9	27
TOTAL	18	18	18	54

Table 7.3. Subjective listening test material

Five subjects participated in the listening test. The test results are shown in Table 7.4. We observe that:

- The subjects did not identify significant difference between S1 and S2 since the preference rate the S1-S2 group is not much greater than 50%, i.e. percentage of choosing one source speaker type over the other by chance. However, there is a slight tendency to prefer native American English source speakers over bilingual source speakers.
- Native American English source speakers resulted in significantly better performance as compared to native Turkish speakers. This result is expected and it confirms our previous informal observations. In the case when a native Turkish source speaker is employed in an American English to Turkish voice conversion application, the source and the target training databases are likely to contain more variation in terms of accent. These differences cause a reduction in similarity to the target speaker's voice.

S1 preferred over S2	**S1 preferred over S3**	**S2 preferred over S3**
56.7%	81.1%	68.9%

Table 7.4. Preference rates between different source speaker types

Source Speaker Type	**MOS**
S1: Native American English	3.97
S2: Compound Bilingual	3.70
S3: Native Turkish	3.55

Table 7.5. MOS test results for the effect of source speaker proficiency in the training and test languages

In the second part of the test, MOS-based quality of transformation utterances was evaluated. For this purpose, eight transformation outputs for each group were presented to the subjects. The MOS quality testing procedure described in Section 6.4.2 was performed. Table 7.5 shows the test results. Although there is no significant difference between S1 and S2 type of source speakers considering the similarity to target speaker,

the quality is significantly better when a native American English source speaker is used.

7.3. Subjective Test 2: Comparison of the Proposed and Baseline Algorithms

In order to compare the proposed methods with the baseline algorithm, a subjective listening test is designed. For this purpose, all methods described in this study are applied. The target was a native male American English speaker and the source was a male bilingual Turkish and American English speaker. The training set consisted of 106 sentence utterance recordings in English. The transformation set was 20 sentence utterance recordings in Turkish. In an ABX test, the outputs of the proposed cross-lingual algorithm are presented to 14 subjects along with the corresponding baseline results and an original target recording. The baseline algorithm employed Sentence-HMM based alignment, state-averaging to estimate the vocal tract transformation function, and transformation of the mean and variance of pitch. The proposed method employed: Phonetic-HMM based alignment, frame weighting based vocal tract transformation function estimation, transformation of the mean and variance of pitch, and stylistic pitch transformation. Speaking rate transformation was not applied since the transformed recordings were single sentence utterances. The subjects were asked to judge which output is more similar to the target voice. A total of 20 pairs of outputs using the baseline and the proposed methods are presented. The proposed method was preferred 251 times out of the 280 cases resulting in a preference rate of 89.6%. All test pairs were presented in random order. The first item in each pair was also shuffled randomly.

In the second part of the test, the subjects were asked to rate the MOS-based quality of the baseline and the proposed methods, presented in random order. The testing procedure is identical with the one described in Section 6.4.2. The subjects have scored the 20 outputs used in the preference test. The MOS quality for the baseline method was 3.59 and 3.74 for the proposed method.

8. CONCLUSIONS

Development of high quality cross-lingual voice conversion techniques will provide a deeper understanding of human language and its perception. It will be possible to improve naturalness in man-machine interaction by providing the means to represent voices in a compact and easily adaptable manner. Inspired by these facts, a cross-lingual voice conversion framework is developed in this study. The proposed framework has a number of advantages over the state-of-the-art counterparts:

- Phonetic-HMM based alignment that can handle multi-lingual data and perform alignment and segmentation in cross-lingual databases in a robust manner is developed
- Context-matching based training can be employed to reduce one-to-many mapping problems as well as to employ non-parallel training databases in cross-lingual voice conversion
- Weighted speech frame mapping based vocal tract transformation function estimation enables detailed transformation of the vocal tract characteristics while keeping continuity and smoothness at desired levels
- Stylistic prosody modeling and transformation is integrated which may help to make the voice conversion output sound closer to the target voice without increasing additional processing distortion significantly in cross-lingual voice conversion applications
- A high-quality cross-lingual database is collected from bilingual speakers with different levels of proficiency in training and transformation languages and employed in subjective and objective performance evaluations

In Chapter 4, a robust automatic alignment and segmentation module that can handle multi-lingual data is developed. In order to compare the alignment performance in an objective framework, an alignment mismatch score is proposed. The alignment mismatch score is a measure of the mismatch between the mappings among source and

target speech frame indices using two different alignment methods. This measure is employed to show that Phonetic-HMM based alignment results in significantly better performance as compared to the Sentence-HMM based counterpart. We have also analyzed the performance of the Phonetic-HMM aligner by combining speech databases in English and Turkish. We have shown that adjusting the number of states for each phoneme according to the number of observations, performance of the Phonetic-HMM based alignment can be significantly improved. Another objective test was performed to show that Phonetic-HMM based alignment results in significantly lower LSF distance to the target voice both in text-independent and text-dependent modes.

The proposed Phonetic-HMM based segmentation and alignment framework opens up a number of interesting topics for future research. First of all, it will be interesting to investigate the extension of the phonetic models to more than two languages and perform evaluations using multi-lingual databases. Another future research topic is the employment of language independent phoneme models as used in multi-lingual TTS engines. This may help to improve alignment performance further in cross-lingual voice conversion. Additionally, integration of cross-lingual voice conversion techniques with multi-lingual text-to-speech synthesis will become an easier task once phonetic information is extracted in a unified framework.

In Chapter 5, a number of methods were proposed for detailed estimation of the vocal tract transformation function. The development of the robust and speaker-independent alignment method in Chapter 4 enabled the employment of information in the training data at the level of individual speech frames. In the transformation stage, a weighted average of the closest source and target speech frame parameters were employed in the estimation of a detailed time-varying vocal tract transformation filter. The proposed method reduced the problems of excessive smoothing. Combined with context-based matching to restrict the search range in transformation, one-to-many mapping problems were reduced. Restricting the search range resulted in lower memory requirements during transformation since all training data need not be loaded into the memory to perform the search. The employment of context information in

transformation also enabled using non-parallel training databases in cross-lingual voice conversion. The final topic of Chapter 5 was the assessment of vocal tract transformation performance using objective measures. The performances of well-known objective measures in speech processing research including spectral distortion, frequency weighted spectral distortion, LSF distance and weighted LP cepstral distance were compared in a statistical testing framework. Results indicated that LSF distance performs significantly better in a number of vocal tract modification and transformation scenarios. Therefore, it was employed in evaluations regarding vocal tract transformation performance throughout the study. We have compared the performances of the baseline and proposed methods in objective and subjective tests. The proposed method performed significantly better both in terms similarity to the target speaker's voice and quality as compared to the baseline method when parallel training databases are available. There is a significant performance reduction in the case of non-parallel databases.

Since vocal tract characteristics are one of the most important features that characterize a given speaker's voice, numerous methods have been proposed for vocal tract transformation as discussed in Chapter 1. Our preliminary results and comparisons with the outputs of other voice conversion methods show that weighted codebook mapping as well as the proposed method outperforms state-of-the-art algorithms in terms of vocal tract transformation performance. However, a formal evaluation in the context of cross-lingual voice conversion has not yet been performed which will be a subject for future research. A combination of different approaches may improve robustness to speaker and language variations and enhance voice conversion performance.

As the baseline algorithm does not support non-parallel training at all, the proposed method provides a useful framework for future research in cross-lingual voice conversion using non-parallel databases. Improvements in mapping phonemes of different languages to each other and performing language independent phonetic training as in the case of multi-lingual text-to-speech synthesis are likely to improve cross-lingual voice conversion performance using non-parallel training.

Chapter 6 focused on the problem of prosody transformation in a cross-lingual context. A pitch contour modification method was developed which performs stylistic pitch transformation. The algorithm involved modeling and transformation of sentence and voiced segment based slopes of pitch contours in addition to the standard mean and variance transformation method. A speaking rate transformation algorithm is developed in order to compensate for overall speaking rate differences as well as rhythm and local differences. The long pause modification strategy resulted in closer output to the target speaker's voice in terms of accent. We have evaluated the performance of the proposed stylistic prosody transformation techniques in the context of accent transformation and have shown that these techniques help to improve similarity to target speaker's voice without adding extra processing distortion.

Future improvements in cross-lingual prosody transformation could be possible by employing large, multi-speaker databases which are also rich in prosodic content. Significant progress has already been achieved in text-to-speech synthesis and speaker identification research by employing data driven methods for prosody modeling. Integration of information from these large databases into the cross-lingual voice conversion framework is very likely to improve prosody transformation performance. As an example, it would be interesting to examine the differences in the stylistic prosody patterns of two languages using databases rich in prosodic content, classify speakers according to their styles, and develop robust prosody modification methods between different prosody-style classes. Once the prosody-style classes are determined, the source and the target speaker can be assigned to one of these prosody-style classes. Then, prosody transformation can be performed by using all information available in the databases rich in prosodic content.

The proposed stylistic pitch transformation method requires parallel training databases. In order to extend the proposed method to non-parallel databases, a standard prosody labeling strategy like the Tones and Break Indices (ToBI) might be employed (Silverman, et. al., 1992). Using a prosody labeling framework, the mapping between the source and target prosodic events can be determined and used in prosody transformation.

Another important topic for future research in cross-lingual prosody transformation is the development of new and robust signal processing techniques for prosody modification. State-of-the-art methods perform signal processing either in a fully parametric manner or in a non-parametric manner. Parametric methods like sinusoidal modeling and modification have the advantage of producing more stable output by providing direct control on the model parameters. Usually large amounts of modifications can be performed with ease and less quality reduction as compared to the non-parametric case. The disadvantage is the reduction in naturalness due to model based speech signal generation. On the contrary, non-parametric methods like the FD-PSOLA result in very natural output for small to medium amounts of modifications. The major disadvantage of non-parametric methods is the absence of control over the output signal in a parametric manner. For example, phase discontinuities and mismatches may result in severe distortions in FD-PSOLA based prosody modifications. It is fairly easy to control and correct such problems in a parametric framework by simply adjusting a set of parameters. A combination of parametric and non-parametric approaches may help to improve the quality and robustness of signal processing techniques for prosody transformation. Similar hybrid models have already been applied for text-to-speech synthesis with success (Min and Ching, 1998). The employment of hybrid techniques that enable a larger range of natural sounding modifications can be used for more detailed stylistic prosody modification.

Chapter 7 provided a description of the cross-lingual voice conversion database that is employed in the evaluations. A subjective test is performed to determine the dependence of cross-lingual voice conversion performance on source speaker proficiency in the training and transformation languages. We have shown that native American English source speakers and compound bilingual source speakers perform equivalently well regarding similarity to the target speaker's voice. The quality was better in the case of native speakers. However both bilingual and native American English source speakers resulted in better quality output as compared to native Turkish source speakers. The performance of the proposed cross-lingual voice conversion algorithm was also compared with the baseline algorithm in a subjective test. We have

shown that the proposed algorithm outperforms the baseline algorithm both in terms of similarity to target voice and quality in cross-lingual voice conversion.

For subjective evaluations, we have focused on a restricted case where the training language was American English and the transformation language was Turkish. Testing more than two languages was fairly difficult since native subjects in all transformation languages would be required. A future collaboration among voice conversion researchers from different countries may facilitate testing all combinations of training and transformation language combinations and comparison of results. For this purpose, a cross-lingual voice conversion with parallel and non-parallel training databases will be required. The database collected in this study may be translated to other languages easily and can be used as part of the multi-lingual database.

State-of-the-art voice conversion systems can only make use of information from two speakers. Employing large multi-speaker databases with data-driven knowledge extraction techniques lead to significant improvements in different areas of speech processing technology. Integration of these data-driven techniques may help to improve both monolingual and cross-lingual voice conversion performance significantly. For this purpose, a given speaker can be mapped on a large speaker database and can be represented as a combination of other speakers for which large amounts of data is available. Our preliminary results in monolingual voice conversion have shown that using a multi-speaker mapping and weighting strategy, the vocal tract spectrum of a given speaker can be successfully modeled. However, further research is required for the assessment of performance in monolingual voice conversion as well as for problems in automatic mapping and weighting in cross-lingual voice conversion.

REFERENCES

Abe, M., S. Nakamura, K. Shikano, and H. Kuwabara, 1988, "Voice Conversion Through Vector Quantization", *Proc. of the IEEE ICASSP*, pp. 565-568.

Abe, M., and S. Sagayama, 1990, "Statistical Study On Voice Individuality Conversion Across Different Languages", *Proc. of the ICSLP 1990*, pp. 157-160.

Acero, A., 1993, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Dordrecht.

Arons, B., 1992, "Techniques, Perception, and Applications of Time-Compressed Speech", *Proc. of the 1992 Conference, American Voice I/O Society*, pp. 169-177.

Arslan, L. M. and D. Talkin, 1997, "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum", *Proceedings of the EUROSPEECH 1997*, Rhodes, Greece, Vol. 3, pp. 1347-1350.

Arslan, L. M. and J. H. L. Hansen, 1997, "A Study of Temporal Features and Frequency Characteristics in American English Foreign Accent", *Journal of the Acoustical Society of America*, vol 102(1), pp. 28-40.

Arslan, L. M., 1999, "Speaker Transformation Algorithm using Segmental Codebooks", *Speech Communication*, 28, pp. 211-226.

Beasley, D. S., and J. E. Maki, 1976, "Time- and Frequency-Altered Speech", in N. J. Lass, (editor), *Contemporary Issues in Experimental Phonetics*, vol. 12, pp. 419-458. Academic Press.

Biology Online: http://www.biology-online.org, 2007.

Black, A. and K. Lenzo, 2004, "Multilingual Text-to-Speech Synthesis", *Proc of the IEEE ICASSP 2004*, vol. 3, pp. 761-764.

Black, A. and P. Taylor, 1997, "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", *Proc. of the Eurospeech* 1997, pp. 601-604.

Bozkurt, B., T. Dutoit, R. Prudon, C. D'Alessandro, and V. Pagel, 2002, "Improving quality of MBROLA synthesis for non-uniform units synthesis", *Proceedings of the IEEE Workshop on Speech Synthesis 2002*, pp. 7- 10.

Burkhardt, F., N. Audibert, L. Malatesta, O. Turk, L. M. Arslan, and V. Auberge, 2006, "Emotional Prosody - Does Culture Make A Difference?", *Proc. of the Speech Prosody 2006*, Dresden, Germany.

Campbell, W. N., 1992, "Syllable-Based Segmental Duration", in G. Bailly, C. Benoit, and T. Sawallis, (eds.), *Talking Machines: Theories, Models, and Designs*, pp 211–224. Elsevier.

Chappel, D.T. and J. H. L. Hansen, 1998, "Speaker-Specific Pitch Contour Modeling and Modification", *Proc. of the IEEE ICASSP 1998*, Seattle, Washington, vol. 2, pp. 885-888.

Childers, D. G., and C.-K. Lee, 1991, "Vocal Quality Factors: Analysis, Synthesis, and Perception", *Journal of the Acoustical Society of America* 90, pp. 2394-2410.

Childers, D. G., 1995, "Glottal Source Modeling for Voice Conversion", *Speech Communication*, vol. 16 (2), pp. 127-138.

Coyle, E., O. Donnellan, E. Jung, M. Meinardi, D. Campbell, C. MacDonaill, and P. K. Leung, 2004, "Time-Scale Modification as a Speech Therapy Tool for Children with Verbal Apraxia", *Proc. of the 5th Intl. Conf. Disability, Virtual Reality & Assoc. Tech.*, Oxford, UK.

Crosmer, J. R., 1985, *Very Low Bit Rate Speech Coding Using the Line Spectrum Pair Transformation of the LPC Coefficients*, Ph.D. Dissertation, Elec. Eng., Georgia Inst. Technology.

d'Alessandro, C. and B. Doval, B., 1998, "Experiments in Voice Quality Modification of Natural Speech Signals: The Spectral Approach", Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis 1998, pp. 277-282.

Davis, S. and P. Mermelstein, 1980, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357-366.

Demol, M., K. Struyve, W. Verhelst, H. Paulussen, P. Desmet, and P. Verhoeve, 2004, "Efficient Non-Uniform Time-Scaling of Speech with WSOLA for CALL Applications", *Proc. of the InSTIL/ICALL 2004 – NLP and Speech Technologies in Advanced Language Learning Systems*, Italy.

Drioli, C., 1999, "Radial Basis Function Networks for Conversion of Sound Spectra", *Proc. of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, NTNU, Trondheim.

Duker, S., 1974, *Time-Compressed Speech: an Anthology and Bibliography in Three Volumes*, Scarecrow, Metuchen, N.J.

Dutoit, T., 1997, "High-Quality Text-to-Speech Synthesis: an Overview", *Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis*, vol. 17, no 1, pp. 25-37.

Duxans, H. and A. Bonafonte, 2003, "Estimation of GMM in Voice Conversion Including Unaligned Data", *Proc. of the EUROSPEECH 2003*, Geneva, Switzerland.

Duxans, H, Bonafonte, A., Kain, and J. Santen, 2004, "Including Dynamic Information in Voice Conversion Systems", *Proc of the XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural SEPLN 2004*, Barcelona, Spain.

Erdogan, H., O. Buyuk, and K. Oflazer, 2005, "Incorporating Language Constraints in Sub-word Based Speech Recognition", *IEEE Automatic Speech Recognition and Understanding Workshop*, Cancun Mexico.

Faltlhauser, R., T. Pfau, and G. Ruske, 2000, "On-Line Speaking Rate Estimation Using Gaussian Mixture Models", *Proc. of the IEEE ICASSP 2000*, vol. 3, pp. 1355-1358.

Fant, G., J. Liljencrants and Q. Lin, 1985, "A Four-Parameter Model of the Glottal Flow", *Speech Transmission Laboratory Quarterly Progress and Status Report*s, No. 4, Royal Institute of Technology, Stockholm, Sweden, pp. 1-13.

Flanagan, J. L. and R. M. Golden, 1966, "Phase Vocoder", *Bell Systems Technical Journal*, Vol. 45, pp. 1493-1509.

Foote, J., J. Adcock, and A. Girgensohn, 2003, "Time Base Modulation: A New Approach to Watermarking Audio", *Proc. of the IEEE International Conference on Multimedia and Expo 2003*, vol. 1, pp. 221-224.

Foulke, W. and T. G. Sticht, 1969, "Review of Research on the Intelligibility and Comprehension of Accelerated Speech", *Psychological Bulletin*, vol. 72, pp. 50-62.

Furui, S., 1986, "Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques", *Speech Communication*, vol. 5 (2), pp. 183-197.

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, 1990. *DARPA-TIMIT acoustic-phonetic continuous speech corpus [CDROM]*.

Hieronymus, J. L., 1993, "ASCII Phonetic Symbols for the World's Languages: Worldbet", *Journal of the International Phonetic Association*.

Hines, W. W. and D. C. Montgomery, 1990, *Probability and Statistics in Engineering and Management Science*, John Wiley and Sons, Inc., NY.

Hirata, Y. and K. Tsukada, 2003, "The Effects of Speaking Rates and Vowel Length on Formant Movements in Japanese", in A. Agwuele, W. Warren, and S. H. Park (eds.), *Proc. of the 2003 Texas Linguistics Society Conference: Coarticulation in Speech Production and Perception*. Somerville, MA: Cascadilla Proceedings Project, pp. 73-85.

Holmes, W., J. Holmes, and M. Judd, 1990, "Extension of the Bandwidth of the JSRU Parallel-Formant Synthesizer for High Quality Synthesis of Male and Female Speech", *Proc. of the IEEE ICASSP 90*, vol. 1, pp. 313-316.

Hunt, A. and A. W. Black, 1996, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", *Proc. of the ICASSP 1996*, pp. 373-376.

Itakura, F., 1975a, "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. ASSP-23, No. 1, pp. 67-72.

Itakura, F., 1975b, "Line spectrum representation of linear predictor coefficients of speech signals", *Journal of Acoust. Soc. of America*, vol. 57, S35 (A).

Itoh, K. and S. Saito, 1982, "Effects of Acoustical Feature Parameters of Speech on Perceptual Identification of Speaker", *IECE Transactions*, J65-A, pp. 101-108.

Kain, A., and M. Macon, 1998, "Personalizing A Speech Synthesizer by Voice Adaptation", *Proc. of the Third ESCA/COCOSDA International Speech Synthesis Worksho*p, pp. 225-230.

Kain, A., and Y. Stylianou, 2000, "Stochastic Modeling Of Spectral Adjustment for High Quality Pitch Modification", *Proc. of the IEEE ICASSP 200*0, vol. 2, pp. 949-952, Istanbul, Turkey.

Kain, A. B., 2001, *High Resolution Voice Transformatio*n, Ph.D. Dissertation, OGI School of Science and Engineering at Oregon Health and Science University.

Klatt, D., 1987, "Review of text-to-speech conversion for English", *Journal of Acoust. Soc. of America*, vol. 82, pp. 737-793.

Knohl, L. and A. Rinscheid, 1993, "Speaker Normalization with Self-Organizing Feature Maps", *Proc. IJNN-93-Nagoya Int. Joint Conf. on Neural Networks*, 243-246.

Kreyszig, E., 1970, *Introductory Mathematical Statistics*, John Wiley and Sons, Inc.

Kuwabara, H. and Y. Sagisaka, 1995, "Acoustic Characteristics of Speaker Individuality: Control and Conversion", *Speech Communication*, vol. 16, pp. 165-173.

Laroche, J., Y. Stylianou, and E. Moulines, 1993., "HNS: Speech Modification Based on A Harmonic + Noise Model", *Proc. of the IEEE ICASSP-93*, Minneapolis.

Latorre, J., K. Iwano, and S. Furui, 2005, "Polyglot Synthesis Using a Mixture of Monolingual Corpora", *Proc. of the IEEE ICASSP 2005*, vol.1, pp.1-4.

Makhoul, J., 1975, "Linear Prediction: A Tutorial Review", *Proc. of the IEEE 1975*, vol. 63, 561-580.

Mashimo, M, T. Toda, K. Shikano, and N. Campbell, 2001, "Evaluation of Cross-Language Voice Conversion Based On GMM and STRAIGHT", *Proc. of the Eurospeech 2001*, pp. 361-364.

Matsumoto, H., S. Hiki, T. Sone, and T. Nimura, 1973, "Multidimensional Representation of Personal Quality of Vowels and Its Acoustical Correlates", *IEEE Trans. AU*, AU-21, pp. 428-436.

McAulay, R. J., and T. F. Quatieri, 1995, "Sinusoidal Coding", in Kleijn and Paliwal (eds.), *Speech Coding And Synthesis*, pp. 121-173, Elsevier Science B.V., Netherlands.

Meshabi, L., V. Barreaud, and O. Boeffard, "Comparing GMM-based Speech Transformation Systems", Proc. of the Interspeech 2007.

Min, C. and P. C. Ching, 1998, "A Hybrid Approach to Synthesize High Quality Cantonese Speech", Proc. of the IEEE ICASSP 1998, vol. 1, pp. 277 – 280.

Mirghafori, N., E. Fosler, and N. Morgan, 1995, "Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes", *Proc. of the EUROSPEECH 1995*, pp. 491-494.

Mizuno, H. and M. Abe, 1995, "Voice Conversion Algorithm Based on Piecewise Linear Conversion Rules of Formant Frequency and Spectrum Tilt", *Speech Communication*, vol. 16, pp. 153-164.

Monsen, R. B., 1978, "Toward Measuring How Well Hearing Impaired Children Speak", *Journal of Speech and Hearing Research*, vol. 21, 1978, pp. 197-219.

Moulines, E. and F. Charpentier, 1990, "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones", *Speech Communication*, vol. 9, pp. 453-467.

Moulines, E. and W. Verhelst, 1995, "Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech" in Kleijn and Paliwal (eds.), *Speech Coding And Synthesi*s, pp. 519-555, Elsevier Science B.V., Netherlands.

Moulines, E. and Y. Sagisaka, (Eds.), 1995, "Voice Conversion: State of the Art and Perspectives (Special Issue of Speech Communication)", Elsevier Science B.V., Netherlands, 16(2).

Narendranath, M., H. M. Murthy, S. Rajendran, and B. Yegnanarayana, 1995, "Transformation of Formants for Voice Conversion Using Artificial Neural Networks", *Speech Communication*, vol.16, pp. 207-216.

Necioglu, B. F., M.A. Clements, , T. P. Barnwell III, and A. Schmidt-Nielsen, 1998, "Perceptual Relevance of Objectively Measured Descriptors for Speaker Characterization", *Proc. of the ICASSP 1998*, vol. 2, pp. 869-872.

Okuda, K., T. Kawahara, and S. Nakamura, S., 2002, "Speaking Rate Compensation Based On Likelihood Criterion in Acoustic Model Training and Decoding", *Proc. of the ICSLP 2002*, vol. 4, pp. 2589-2592.

Oppenheim, A.V., and R.W. Schafer, *Digital Signal Processing*, Englewood Cliffs, NJ, Prentice-Hall, 1975.

Ormanci, E., U. H. Nikbay, O. Turk, and L. M. Arslan, 2002, "Subjective Assessment of Frequency Bands for Perception of Speaker Identity", *Proc. of the ICSLP 200*2, vol. 4, pp.2581-2584, Denver, Colorado, USA.

Osberger, M. J., and Mcgarr, N. S., 1982, "Speech Production Characteristics of the Hearing-Impaired", *Speech and Language*, vol. 8, pp. 222-283.

Pellegrino, F., J. Farinas, and J.-L. Rouas, 2004, "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech", *Proc. of the Speech Prosody 2004*, pp. 517-520.

Pitrelli, J., F., and V. W. Zue, 1989, "A Hierarchical Model for Phoneme Duration in American English", *Proc. of the EUROSPEECH 1989*, pp. 324–327.

Quatieri, T. F. and R. J. McAulay, 1992, "Shape Invariant Timescale and Pitch Modification of Speech", *IEEE Transactions On Signal Processing*, vol. 40, no. 3, pp. 497-510.

Rabiner, L. R. and R. W. Schafer, 1978, *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs. New Jersey.

Rabiner, L. R., 1989, "A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. of the IEEE*, vol 77 (2), pp. 257-286.

Shriberg, E., and A. Stolcke, 2004, "Prosody Modeling for Automatic Speech Recognition and Understanding", in M. Johnson (Editor) *Mathematical Foundations of Speech and Language Processing*.

Silverman, K., M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg, 1992, "ToBI: A Standard Scheme for Labelling Prosody", *Proc. of the ICSLP 1992*, pp. 867-879.

Sonmez, M. K., E. Shriberg, L. Heck, and M. Weintraub, 1998, "Modeling Dynamic Prosodic Variation for Speaker Verification," *Proc. of the ICSLP 1998* , vol. 7, pp. 3189-3192.

Stylianou, Y., O. Cappe, and E. Moulines, 1998, "Continuous Probabilistic Transform for Voice Conversion", *IEEE Trans. on Speech and Audio Proc.*, vol. 6, no. 2, pp. 131-142.

Suendermann, D., and H. Ney, 2003, "VTLN-Based Voice Conversion", *Proc. of the 3rd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2003),* Darmstadt, Germany.

Suendermann, D., A. Bonafonte, A., H. Ney, and H. Höge, 2004, "Voice Conversion Using Exclusively Unaligned Training Data", *Proc. of the XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural SEPLN 2004*, Barcelona, Spain.

Syrdal, A., G. Moehler, K. Dusterhoff, A. Conkie, and A. Black, 1998a, "Three Methods of Intonation Modeling", *Proc. of the 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 305–310.

Syrdal, A., Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, 1998b, "TD-PSOLA versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis", *Proc. of the IEEE ICASSP 1998*, vol. 1, pp. 273-276.

Talkin, D., 1995, "A Robust Algorithm for Pitch Tracking (RAPT)", in Kleijn and Paliwal (eds.), *Speech Coding And Synthesis*, pp. 121-173, Elsevier Science B.V., Netherlands.

The CMU Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2007.

Tomokiyo, L. M., 2000, "Linguistic Properties of Non-native Speech", *Proc. of the ICASSP 2000*, vol. 3, pp. 1335-1338.

Turajlic, E., D. Rentzos, S. Vaseghi, and H. Ching-Hsiang, 2003, "Evaluation of Methods for Parametric Formant Transformation in Voice Conversion", Proc.of the IEEE ICASSP 2003, vol. 1, pp. 724-727.

Turk, O. and L. M. Arslan, 2002, "Subband Based Voice Conversion", *Proc. of the ICSLP 200*2, vol. 1, pp.289-292, Denver, Colorado, USA.

Turk, O. and L. M. Arslan, 2003, "Voice Conversion Methods for Vocal Tract and Pitch Contour Modification", *Proc. of the Eurospeech 2003*, pp. 2845-2848.

Turk, O., 2003, *New Methods for Voice Conversion*, M.S. Thesis, Bogazici University.

Turk, O. and L M. Arslan, 2003, "Konusmaci Donusturme Icin Uc Yeni Yontem ", *Proc. of SIU 2003*, pp. 398-401, Istanbul, Turkey.

Turk, O. and L. M. Arslan, 2005, "Donor Selection for Voice Conversion", *Proc. of the EUSIPCO 2005*, Antalya, Turkey.

Turk, O. and L. M. Arslan, 2006, "Robust Processing Techniques for Voice Conversion", *Computer Speech and Language 20 (2006)*, pp. 441-467.

van Santen, J. P. H., 1994, "Assignment of Segmental Duration in Text-to-Speech Synthesis", *Computer Speech and Languag*e, vol 8, pp. 95–128.

Wells, J. C., 1997, "SAMPA computer readable phonetic alphabet" In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter, part IV, section B.

Wikipedia, The Free Encyclopedia, http://en.wikipedia.org, 2007.

Woodland, P. C., J. J. Odell, V. Valtchev, and S. J. Young, 1994, "Large Vocabulary Continuous Speech Recognition Using HTK", *Proc. of the IEEE ICASSP 1994*.

Yildirim, S., M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, 2004, "An Acoustic Study of Emotions Expressed in Speech", *Proc. of the ICSLP 2004*, pp. 2193-2196.

Zellner, B., 1998, "Fast and Slow Speech Rate: a Characterisation for French", *Proc. of the ICSLP 1998*, pp. 3159-3163.

Zhang, W., L. Q. Shen, and D. Tang, , 2001, "Voice Conversion Based On Acoustic Feature Transformation", *Proc. of the 6th National Conference on Man-Machine Speech Communications*.

APPENDIX A: TEXT MATERIAL FOR CROSS-LINGUAL VOICE CONVERSION DATABASE

English Training

- A clean plate please!
- Do you need a new toothbrush?
- I walk to school everyday.
- I suppose you will need to decrease the microphone gain level for less noise.
- When do you serve breakfast?
- What would you like to drink?
- She will be able to travel to fifteen countries with a single ticket.
- Don't forget to buy ham on your way home!
- I don't usually drive at nights.
- Can you please tell the guests to go to the meeting room?
- He should send his CV to big companies.
- What's your shoe size?
- Three bottles of fresh orange juice please!
- You should not drink too much cold water.
- You should try to see things from his viewpoint.
- Can you please send me the receipt via fax?
- Do you approve the changes?
- You don't want to stay at a second-class pension.
- Can you please mail me the invoice?
- When does the next train leave?
- You should change clothes before going out.
- Three employees were poisoned due to gas leakage.
- These brown mushrooms are extremely poisonous.
- He will join our department next week.
- All equations should be in boldface characters.

- Can you switch to the news channel?
- Please adjust your tables to an upright position.
- She will need to cancel the flight reservation.
- You should swim half an hour each day to strengthen your arm muscles.
- This street is known for being crowded every time of the day.
- Her ex-boyfriend moved to Washington last year.
- Can you help me open the window?
- Is it expected to rain today?
- It is smart to choose the most reliable travel agency for international journeys.
- We prefer traveling by train for its comfort and cheap tickets.
- She had two glasses of diet coke.
- We hope that you'll enjoy staying at our hotel.
- Tell him to bring a cup of coffee and cream!
- You will be responsible for updating databases and checking the old entries.
- We are going bowling tonight.
- Maintaining your composure is important for job interviews.
- The computer monitors are produced using the most recent technology.
- My favorite meals are meatballs and French fries.
- Could you give me instructions to find your office by car?
- I want to register for a new phone number.
- He should work hard for success.
- He told me that you have started playing tennis.
- He likes fishing and long walks.
- Inappropriate storage conditions may result in food spoilage.
- Can you please buy a newspaper form the market?
- She passed all her fifth grade exams.
- Where can I find the cleaning stuff?
- Proceed to the right at the second turn.
- The hotel has its own beach and bar.
- Would you like more ice for your drink?
- Wealthy people often prefer self contained houses.

- The most beautiful present was a cashmere sweater.
- Is it forbidden to play football in the school yard?
- The teacher can speak six languages perfectly.
- A group of migrating birds is approaching the lake.
- The total amount of sales turned out to be greater than expected.
- She is acting too paranoid about safety.
- Please ask her if she needed anything else.
- Did you manage to catch your appointment?
- When will the project be completed?
- We'll buy new paintings for the living room .
- Her yellow skirt faded after she washed it.
- Speech therapy helps mentally disordered children.
- Jet charters have started between Paris and Prague.
- Only a thief might have broken the rear window.
- Push the green button to get off!
- How long will the whole journey take?
- The optical illusions presentation attracted all students.
- We ran out of time for tests and enhancements.
- Did she enjoy the two weeks holiday?
- The doors will be fixed on Monday.
- Sit down and fasten your seat belt as soon as you get on the aircraft!
- His sister dislikes meat dishes.
- How long have you been teaching decision making classes?
- My cousin is brilliant in physics and chemistry.
- Can you please serve a bottle of red wine with two glasses?
- I don't believe that this old movie will be in theaters again!
- They have just finished painting the room.
- Can you please forward a copy of this fax to our partner?
- The company has increased its profits by twenty percent as compared to the previous year.
- It's hard to find a suitable carpet for the living room.

- I had a cheese sandwich for breakfast.
- How many employees does your company have?
- She wanted to have nothing but a long sleep.
- The explosion destroyed most of the statues.
- The whole album was recorded in just two weeks.
- Never exceed fifty miles per hour while driving!
- Mental confusion might be a symptom of a serious problem in the brain.
- Where can I find the cheapest computer hardware?
- Have you ever had a dessert that contained green apples?
- The joint conference on speech and music technology was quite interesting.
- Monthly subscription will cost you thirty five dollars.
- An identification card and a recent photograph are required for application.
- How much cement will they need?
- They own a large factory in which two thousand eight hundred and thirty workers are employed.
- An increase of fifteen percent of salaries is expected.
- The English alphabet consists of twenty six symbols.
- How many attorneys did you employ?
- Hold your breath for thirty seconds!
- It was just a casual conversation.
- Cubic yard is a measure of volume.
- The color of the curtains does not match that of the rug.
- They liked the sofa at the corner of the room.
- There is not sufficient gas pressure on the fourth floor.
- They've decided on the kitchen ceramics.
- There were two coils of wire near the door.
- Can I examine the figures once more?
- They ate out at the restaurant opposite their house.
- The letter he saw on the desk was an invitation.
- How hot is boiling water?
- Sweden is in the north of Europe.

- What is the inverse operation for division?
- He was appointed as the president of the council.
- I forwarded the telephone message to the secretary.
- There are no problems with the internet connection.
- The trajectories of the planets around the sun are perfect ellipsoids.
- My father collects toy cars.
- The recordings were not intelligible because of noise.
- The new fertilizer will accelerate the growth of corn.
- He is the author of twenty three books.
- Can you turn off the air conditioning?
- The hotels on the southwest are fully booked during August.
- The box was covered with a golden foil.
- Turn off the television before you go out!
- The soil is pretty moist in the backyard.
- Please pay your debt by the end of the month.
- I would like to get the recipe of this sauce.
- He was appointed as the manager of the new public relations company.
- Air pressure measurements are correlated with daily temperature.
- He is the manager of a financial supervision company.
- Jogging is a favorite leisure activity in our town.
- For refunds you should apply to the office on the right.
- Tomorrow will be the longest day of the year.
- Can I pay the check from my personal account?
- Can I return the shirt I purchased last week?
- Can I pay the ticket fare by credit card?
- Place the napkins on the right of the spoons!
- I had a long bike ride for five hours.
- The red t-shirt looks better.
- There are twelve species that can survive in the desert.
- The teacher wrote the famous equation on the board.
- I always had an interest in overseas voyages.

- A piece of cloth lies on the floor.
- Where is the missing items desk?
- Did you get the letter from him?
- They left before the show was over.
- The voice of the actor changed as he got older.
- The annual advertisement budget was over half a million dollars.
- The house had two toilets.
- The visit of the prime minister is postponed.
- The product will be available in the market this weekend.
- The project will be cancelled in the event of financial difficulties.
- The compass will show the direction we are going to.
- It's hard to believe that they started as an amateur garage band.
- What is the maximum speed limit in kilometers?
- Can you throw this heavy ball over the fence?
- Did you hear any rumors about the earthquake?
- I strongly recommend the soup of the day.
- She works as an anesthesiologist.
- I was late because of the snow.
- A five degree increase in temperature is expected.
- The accountant will send the annual financial reports.
- It takes half an hour for her to reach the office from the house.
- She was promoted as a software engineer.
- Drink your milk before having the chocolate!
- We need your signature for the approval of the modifications.
- She was a very thin and tall girl.
- Mary is playing with her wooden toy.
- For how long have you been a doctor?
- Can I have a glass of water please?
- You can taste the most delicious pasta in the city at this cafe.
- I'm very happy to hear that your test results are not pathological.
- The cookies that his aunt made did not last for more than ten minutes.

- Is it possible to avoid moisture at the basement floor?
- The company went bankrupt in the next crisis.
- Vision correction surgeries have become very popular in recent years.
- He should not leave without an umbrella.
- The farm workers went on a new strike.
- How much water should one drink every day for a healthy diet?
- This mixed salad didn't have enough tomatoes and lettuce.
- He had a sore throat and slept all Thursday.
- Who will do the washing?
- Did you get tired of flying for ten hours?
- Don't open the oven!
- She should watch her diet.
- The meeting will be held in the conference room.
- It usually rains during this period of the year.
- Some inclusions affect the quality of diamonds.
- A powerful storm has started at the ocean coast.
- Can I try on these pants?
- I bought a new game for my sister as a birthday present.
- You can find the application forms at the help desk.
- You should be at the post office near the bank at half past three.
- What are the most valuable companies in the stock market?
- Did I miss the eight fifteen ship?
- You should first cut the egg plant into slices and then boil it in hot water for twenty minutes.
- The basic goal of massage therapy is to help the body heal itself.
- The projected light died suddenly.
- The taxi fares will be re-adjusted.
- You should get off at the next stop.
- Can you show the direction it followed?
- Could you please prepare the invoice according to the company information?
- Planning serves for better performance.

- Did everyone attend the board meeting?
- The oily substance slipped out of his hand.
- Avoid long hours of exposure to direct sunlight!
- In which month should I pay the real estate taxes?

English Test

- How many hours do sports classes take in a week?
- The headphone cable is broken.
- You should take a bus when you leave the subway.
- Could you please turn the volume of your mobile down?
- Which vegetables can be cooked the most easily?
- I like to read and listen to music in my spare time.
- Would you prefer soybeans instead of meat?
- Don't buy fruits that are not fresh!
- Using double precision helps to avoid overflow problems in computing.
- He joined the voice experts group several weeks ago.
- You need private lessons to improve your painting skills.
- You should switch off your mobile phone in public transportation vehicles.
- Which department did you graduate from?
- The new travel card provides discounts for older people and children.
- The ratio of the area and the circumference of a circle is proportional to its radius.
- My son is very interested in mathematics.
- The director will be appointed to a new position.
- There was a significant drop in the unemployment rate.
- When was your mother's birthday?
- This it the most crowded room in the building.
- Can I ask a question regarding the last section?
- There is a limit of twenty kilograms for luggage.
- Does the salad lack lemon and salt?

- Wait for half an hour before taking the sugar and flour mixture out of the refrigerator!
- Can you check the second drawer on the left?
- Both the cat and the dog are sleeping.
- The balcony had a nice view of the sea and the forest.
- Does this bus go to the town center?

Turkish Test

- Ucuz ve konforlu bir yolculuk için treni tercih ediyoruz.
- Deprem söylentilerini duydun mu?
- Isleminize onay veriyor musunuz?
- Bu pantolonu deneyebilir miyim?
- Bu sosun tarifini verir misin?
- Içkinize buz ister misiniz?
- Polis komsusunun ifadesine basvurdu.
- Borsada en degerli bes hisse senedini biliyor musun?
- Bu otobüs sehir merkezine gider mi?
- Yeni jetlerdeki otomatik pilot sistemi yenilendi.
- Son bölümle ilgili bir soru sorabilir miyim?
- Yengem taze erik göndermis
- Siparisler nedeniyle hafta içi izinler iptal edildi.
- Is seyahati nedeniyle mi evde degildin.
- Sirketinize otomobil ile nasil gelecegimi tarif eder misiniz?
- Yeni düzenlemeler nedeniyle yolun bir seridi trafige kapali tutuluyor.
- Yaz tatilini büyük havuzu olan bir otelde geçirecek.
- Tahtayi süngerle silebilirsin.
- Telefonda aldigim mesaji sekretere ilettim.
- Proje kaç ay sonra bitecek?
- ,Firinin kapagini açma!
- Bulasiklari kim yikayacak?

- Halilari yikadiktan sonra odayi havalandiralim.
- Ikinci sapaktan saga dönün.
- Konuklari toplanti odasina alir misiniz?
- Kitabin üçüncü bölümünün ilk alti sayfasini okuyun.
- Kizartma yemek zararli mi?
- Kirmizi gömlek daha güzel görünüyor.
- Oturma odasina birkaç tablo satin alacagiz.
- Bugün yagmur yagacagi söyleniyor.
- Bes dakika içinde havaalanina ulasmis olacagiz.
- Kira kontratina uygun davranmadigi için dava açtim.
- Yazilim uzmani olarak basladigi görevinde hizla yükseldi.
- Eski müdür baska bir okula atanacak.
- Maaslara yüzde on bes zam yapilacak.
- Tenis oynamaya basladigini duyduk.
- Çölde yasamini sürdürebilen on farkli hayvan türü var.
- Deniz kiyisinda küçük bir tatil köyünde kalmislar.
- Halamla amcami ziyarete gittik.
- Genç yasta emekli olacak.

English Paragraph

New York City is the most populous city in the United States and the most densely populated major city in North America. Located in the state of New York, the city has a population of over eight point one million within an area of three hundred and twenty one square miles.

The city is a center for international finance, fashion, entertainment, and culture, and is widely considered to be one of the world's major global cities with an extraordinary collection of museums, galleries, performance venues, media outlets, international corporations and financial markets. It is also home to the headquarters of the United Nations.

The New York metropolitan area has a population of about eighteen point seven million, which makes it one of the largest urban areas in the world. The city proper consists of five boroughs which would be among the nation's largest cities if considered independently. Popularly known as the Big Apple the city attracts large numbers of immigrants. Over a third of its population is foreign born. In addition, people from all over the United States come for its culture, energy, cosmopolitanism, and economic opportunity. The city is also distinguished for having the lowest crime rate among the twenty five largest American cities.

Turkish Paragraph

Türkiye'nin en kalabalik sehri olan Istanbul, dünyada iki kitada yer alan tek sehirdir. Sehri iki yakaya ayiran Bogaziçi, Avrupa ile Asya kitalarinin sinirini belirler. Artik bir baskent olmasa da Türkiye'nin endüstri, ticaret ve kültür merkezidir.

Tarihi Istanbul sehri üç tarafini Marmara Denizi, Bogaziçi ve Haliç'in sardigi bir yarim ada üzerinde yer almaktadir. Sehir stratejik bir bölgede bulunmasi nedeniyle hep önemli bir ticaret merkezi olmustur. Üç dünya imparatorluguna baskent olan Istanbul'da 1600 yili asan bir süre boyunca 120'den fazla imparator ve sultan hüküm sürmüstür.

Ikinci Dünya Savasi'ni takip eden yillarda baslayan ve 1950'den sonra hizlanan plansiz gelisme sehrin eski dokusuna zarar vermistir. Disaridan yapilan göçler ile nüfusu hizla artan Istanbul kisa sürede tarihi surlarin disina tasmis, sur içi alanlar atölye, fabrika ve is yerlerinin istilasina ugramistir. 1980'li yillarda baslayan kurtarma hamleleri ile Istanbul yeniden yapilanma sürecine girmistir. Roma sehir surlarinin restorasyonuna baslanmis, daha önceki yillara göre temizlik ve bakim konusunda Avrupa standartlari yakalanmistir

APPENDIX B: COMPARISON OF SAMPA PHONEME SETS FOR AMERICAN ENGLISH AND TURKISH

CONSONANTS				
SAMPA Symbol	**English Word**	**SAMPA Transcription**	**Turkish Word**	**SAMPA Transcription**
p	pin	p I n	ip (thread)	i p
b	bin	b I n	balik (fish)	b a 5 1 k
t	tin	t I n	ütü (iron)	y t y
d	din	d I n	dede (grandfather)	d e d e
k	kin	k I n	akil (brain)	a k 1 5
g	give	g I v	karga (crow)	k a r g a
tS	chin	tS I n	seçim(choice)	s e tS i m
dZ	gin	dZ I n	cam (glass)	dZ a m
f	fin	f I n	fare (mouse)	f a r e
v	vim	v I m	ver (give)	v e r
s	sin	s I n	ses (sound)	s e s
z	zing	z I N	azik (food)	a z 1 k
S	shin	S I n	asi (graft)	a S 1
Z	measure	m E Z @`	müjde (good news)	m y Z d e
h	hit	h I t	hasta (ill)	h a s t a
m	mock	m A k	dam (roof)	d a m
n	knock	n A k	ani (memory)	a n 1
N	thing	T I N	süngü (bayonet)	s y N g j y
r	wrong	r O N	raf(shelf)	r a f
l	long	l O N	lale (tulip)	l a l e
w	wasp	w A s p	tavuk (chicken)	t a w u k
j	yacht	j A t	yat (yacht)	j a t

Table B.1. Common consonants in American English and Turkish SAMPA phoneme sets.

VOWELS				
SAMPA Symbol	**English Word**	**SAMPA Transcription**	**Turkish Word**	**SAMPA Transcription**
i	ease	i z	kil (clay)	c i l
e	raise	r e z	keçi (goat)	c e tS i
u	lose	l u z	kul (slave)	k u 5
o	nose	n o z	kol (arm)	k o 5

Table B.2. Common vowels in American English and Turkish SAMPA phoneme sets.

SILENCES AND PAUSES	
SAMPA Symbol	**Meaning**
+	Epenthetic silence
#	Pause
##	Begin/End

Table B.3. Common silence and pause symbols in the American English and the Turkish SAMPA sets.

CONSONANTS		
SAMPA Symbol	**Word**	**SAMPA Transcription**
T	thin	T i n
D	this	D i s

Table B.4. Distinct American English consonants that do not exist in the Turkish SAMPA set.

VOWELS		
SAMPA Symbol	**Word**	**SAMPA Transcription**
I	pit	p I t
E	pet	p E t
{	pat	p { t
A	pot	p A t
V	cut	k V t
U	put	p U t
O	cause	k O z
aI	rise	r aI z
OI	noise	n OI z
aU	rouse	r aU z
3`	furs	f 3` z
@	allow	@ l a U
@`	corner	k o r n @`

Table B.5. Distinct American English vowels that do not exist in the Turkish SAMPA set.

CONSONANTS		
SAMPA Symbol	**Word**	**SAMPA Transcription**
c	kedi (cat)	c e d i
gj	genç (youth)	gj e n tS
G	sagir (deaf)	s a G 1 r
5	hala (aunt)	h a 5 a

Table B.6. Distinct Turkish consonants that do not exist in the American English SAMPA set

VOWELS		
SAMPA Symbol	**Word**	**SAMPA Transcription**
y	kül (ash)	c y l
2	göl (lake)	gj 2 l
1	kil (hair)	k 1 5
a	kal (stay)	k a 5

Table B.7. Distinct Turkish vowels that do not exist in the American English SAMPA set

CONSONANTS				
TIMIT Symbol	**SAMPA Symbol**	**English Word**	**TIMIT Transcription**	**SAMPA Transcription**
b	b	bee	b iy	b i
ch	tS	choke	ch ow k	tS o k
d	d	day	d ey	D e
dh	D	then	dh eh n	D e n
dx	d	muddy	m ah dx iy	m V d i
f	f	fin	f ih n	f I n
g	g	guy	g ay	g aI
hh	h	hay	hh ey	h e
hv	h	ahead	ax hh eh d	@ h e d
jh	dZ	joke	jh ow k	dZ o k
k	k	key	k iy	k i
l	l	lay	l ey	l e
m	m	mom	m aa m	m A m
n	n	noon	n uw n	n u n
ng	N	sing	s ih ng	s I N
nx	n	winner	w ih nx axr	w I n r
p	p	pea	p iy	p i
r	r	ray	r ey	r e
s	s	sea	s iy	s i
sh	S	she	sh iy	S i
t	t	tea	t iy	t i
th	T	thin	th ih n	T I n
v	v	van	v ae n	v { n
w	w	way	w ey	w e
y	j	yacht	y aa t	j A t
z	z	zone	z ow n	z o n
zh	Z	azure	ae zh er	{ Z 3`
el	l	bottle	b aa t el	b A t l
em	m	bottom	b aa t em	b A t m
en	n	button	b ah q en	b V ? n
eng	N	washington	w aa sh eng t ax n	w A S N t @ n
q	t (or ?)	bat	b ae q	b { t

Table B.8. Mapping between SAMPA and TIMIT phoneme sets for American English. The table is a shortened version of the list given in (Hieronymus, 1993)

VOWELS				
TIMIT Symbol	**SAMPA Symbol**	**English Word**	**TIMIT Transcription**	**SAMPA Transcription**
aa	A	bott	b aa t	b A t
ae	{	bat	b ae t	b { t
ah	V	but	b ah t	b V t
ao	O	bought	b ao t	b O t
aw	aU	bout	b aw t	b aU t
ax	@	about	ax b aw t	@ b aU t
axr	@`	butter	b ah dx axr	b V d r
ax-h	@	suspect	s ax-h s p eh k t	s @` s p e k t
ay	aI	bite	b ay t	b aI t
eh	E	bet	b eh t	b E t
ey	e	bait	b ey t	b e t
er	3`	bird	b er d	b 3` d
ih	I	bit	b ih t	b I t
ix	I	debit	d eh b ix t	d E b I t
iy	i	beet	b iy t	b i t
ow	o	boat	b ow t	b o t
oy	OI	boy	b oy	b OI
uh	U	book	b uh k	b U k
uw	u	boot	b uw t	b u t
ux	u	toot	t ux t	t u t

Table B.9. Mapping between SAMPA and TIMIT phonemes for American English. The table is a shortened version of the list given in (Hieronymus, 1993)

Silences and pauses		
TIMIT Symbol	**SAMPA Symbol**	**Meaning**
epi	+	Epenthetic silence
pau	#	Pause
h#	##	Begin/End

Table B.10. Mapping between SAMPA and TIMIT phonemes for American English. The table is a shortened version of the list given in (Hieronymus, 1993)

CLOSURE INSTANTS OF STOPS		
SAMPA Symbol	**TIMIT Symbol**	**Meaning**
-	bcl	Closure instant before/after b
-	dcl	Closure instant before/after d
-	gcl	Closure instant before/after g
-	kcl	Closure instant before/after k
-	pcl	Closure instant before/after p
-	tcl	Closure instant before/after t

Table B.11. Mapping between SAMPA and TIMIT phonemes for American English. The table is a shortened version of the list given in (Hieronymus, 1993)

APPENDIX C: THE PAIR WISE t-TEST

The procedure for performing a pair wise t-test is given in (Hines, et. al. 1990):

"…Let (X_{11}, X_{21}), (X_{12}, X_{22}), …, (X_{1n}, X_{2n}) be a set of n paired observations, where we assume that $X_1 \sim N(\mu_1, s_1^2)$ and $X_2 \sim N(\mu_2, s_2^2)$. Define the differences between each pair of observations as $D_j = X_{1j}-X_{2j}$, j=1, 2, …, n.

The D_j's are normally distributed with mean

$$\mu_D = E(X_1-X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2 \quad \text{(C1)}$$

so testing hypotheses about equality of μ_1 and μ_2 can be accomplished by performing a one-sample t-test on μ_D. Specifically, testing H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 ? \mu_2$ is equivalent to testing

$$H_0: \mu_D = 0 \quad \text{(C2)}$$

$$H_1: \mu_D ? 0 \quad \text{(C3)}$$

The appropriate test statistic for Equations C2 and C3 is:

$$t_0 = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} \quad \text{(C4)}$$

where

$$\bar{D} = \frac{\sum_{j=1}^{n} D_j}{n} \quad \text{(C5)}$$

and

$$S_D^2 = \frac{\sum_{j=1}^{n} D_j^2 - \frac{(\sum_{j=1}^{n} D_j)^2}{n}}{n-1} \quad \text{(C6)}$$

are the sample mean and variance of the differences. We would reject H_0: $\mu_D = 0$ (implying that $\mu_1 ? \mu_2$) if $t_0 > t_{a/2, n-1}$ or if $t_0 < -t_{a/2, n-1}$."

Note that a is the significance level and n-1 is the degrees of freedom for the test. a = 0.05 or a = 0.01 corresponding to levels of 95% and 99% respectively are commonly used in the tests performed in this study. The $t_{a/2,\, n-1}$ values can be found in tables of t-distributions.

APPENDIX D: LIST OF PUBLICATIONS

Journals

Türk, O., and Arslan, L. M., 2006, “Robust Processing Techniques for Voice Conversion”, Computer, Speech and Language (20), 2006, pp. 441-467.

Türk, O., and Arslan, L., M., 2007, “Donor Selection for Voice Conversion”, Journal of the Acoustical Society of America, *under review*.

Patent Applications

Türk, O., and L. M. Arslan. “Speech Conversion System and Method”, Voxonic Inc., U.S. Patent Application Filed: 10 Nov 2005, No: 11/271,325.

Türk, O., and L. M. Arslan. “Donor Selection for Voice Conversion”, Voxonic Inc., Provisional U.S. Patent Application, 14 Mar 2005.

International Conference Proceedings

Türk, O. Schröder, M., Bozkurt, B., and Arslan, L. M., “Voice Quality Interpolation for Emotional Text-To-Speech Synthesis”, INTERSPEECH 2005, Lisbon, Portugal.

Türk, O., and Arslan, L., M., “Donor Selection for Voice Conversion”, EUSIPCO 2005, Antalya, Turkey.

Türk, O., and Arslan, L. M., 2004, "Pronunciation Scoring for the Hearing-Impaired", 9th International Conference Speech and Computer - SPECOM 2004, St. Petersburg, Russia.

Arisoy, E., Arslan, L., Demiralp, M., N., Ekenel, H., K., Kelepir, M., Meral, H., M., Özsoy, A., S., Sayli, Ö., Türk, O., Yolcu, B., C., "Acoustic Analysis of Turkish Sounds", 12th International Conference on Turkish Linguistics, 11-13 August 2003, Izmir, Turkey.

Türk, O., and Arslan, L. M., 2003, "New Methods for Vocal Tract and Pitch Contour Transformation ", *EUROSPEECH 2003*, Geneva, Switzerland.

Türk, O., and Arslan, L. M., 2003, "Subjective Evaluations for Perception of Speaker Identity Through Acoustic Feature Transplantations", *EUROSPEECH 2003*, Geneva, Switzerland.

Türk, O., and Arslan, L. M., 2002, "Subband Based Voice Conversion", *Proceedings of the ICSLP 2002*, Vol. 1, pp.289-292, September 2002, Denver, Colorado, USA.

Ormanci, E., Nikbay, U. H., Türk, O., and Arslan, L. M., 2002, "Subjective Assessment of Frequency Bands for Perception of Speaker Identity", *Proceedings of the ICSLP 2002*, Vol. 4, pp.2581-2584, September 2002, Denver, Colorado, USA.

Türk, O., Sayli, O., Dutagaci, H., and Arslan, L. M., 2002, "A Sound Source Classification System Based On Subband Processing", *Proc. of the 3rd IEEE Benelux Signal Processing Symposium (SPS-2002),* Leuven, Belgium.

National Conference Proceedings

Türk, O., and Arslan, L. M., 2004, "Dayanikli Konusmaci Dönüstürme Yöntemleri", SIU 2004, Kusadasi, Turkey.

Türk, O., and Arslan, L. M., 2004, "Konusma Terapisine Yönelik Konusma Tanima Yöntemleri", SIU 2004, Kusadasi, Turkey.

Türk, O., and Arslan, L. M., 2003, “Konusmaci Dönüstürme Için Üç Yeni Yöntem ”, SIU 2003 Bildirileri Kitapçigi, pp. 398-401, Istanbul, Turkey.

Türk, O., and Arslan, L. M., 2004, “Konusma Terapisine Yönelik Otomatik Konusma Tanima Yöntemleri”, II. Ulusal Dil ve Konusma Bozukluklari Kongresi 2004, Eskisehir, Turkey.

Türk, O., Sayli, O., Ozsoy, A., S., and Arslan, L., M., “Türkçe'de Ünlülerin Formant Frekans Incelemesi”, 18. Ulusal Dilbilim Kurultayi 2004, Ankara, Turkey.

Türk, O., and Arslan, L. M., 2003, “Konusmaci Kimligi Algilanmasinda Akustik Özniteliklerin Karsilastirmali Analizi”, *SIU 2003 Bildirileri Kitapçigi*, pp. 394-397, Istanbul, Turkey.

Türk, O., Sayli, O., Dutagaci, H., and Arslan, L. M., 2002, “Alt-bant Islemeye Dayali Bir Ses Siniflandirma Sistemi”, *Proc. of the SIU 2002*, Denizli, Turkey.