Transforming Voice Quality and Intonation

Ben Gillett

Thesis submitted for the degree of Masters of Science by Research

University of Edinburgh

30 January 2003

Acknowledgements

I would like to thank many people for their advice and support which has made this work possible. I am very grateful to my supervisor Simon King, and to other members of the department including Bob Ladd, Rob Clark, Joe Frankel, Korin Richmond, Jithedra Vepa, Shiga Yoshinori and James Horlock.

I am also grateful for the provided by Anthropics Technology Ltd.

Declaration

I declare that, apart from where properly indicated, the work contained in this thesis is entirely the product of my own work.

Abstract

Voice transformation is the process of transforming the characteristics of speech uttered by a source speaker, such that a listener would believe the speech was uttered by a target speaker. In this thesis two aspects of the transformation problem are addressed: voice quality and intonation.

The voice quality transformation component of our system has two main parts corresponding to the two components of the source-filter model of speech production. The first component transforms the spectral envelope as represented by a linear prediction model. The transformation is achieved using a Gaussian mixture model, which is trained on aligned speech from source and target speakers. The second part of the system predicts the spectral detail from the transformed linear prediction coefficients. A novel approach is proposed, which is based on a classifier and residual codebooks. The system has some similarities with earlier work by Kain, however the work reported here is not restricted to speech spoken in a monotone and with mimicked prosody. Also, on the basis of a number of performance metrics it outperforms existing systems.

We also present a new method for the transformation of F0 contours from one speaker to another based on a small linguistically motivated parameter set. The system performs a piecewise linear mapping using these parameters. A perceptual experiment clearly demonstrates that the presented system is at least as good as an existing technique for all speaker pairs, and that in many cases it is much better and almost as good as using the target F0 contour.

Contents

List of Figures

Chapter 1

Introduction

Voice transformation is the process of taking the speech of a source speaker and transforming the characteristics of the signal, such that a human listener would believe the speech was uttered by a target speaker.

1.1 Motivation

Throughout our lives we rely on our ability to identify speaker identity. For example, in a telephone conference or radio programme we can identify and distinguish between different speakers.

One of the main applications of voice conversion would be in the field of text-to-speech adaptation. Modern speech synthesisers generally work by joining together segments of speech to create a desired utterance. In order for such an approach to work, a large database of speech is required, and in addition many man hours must be spent labeling the data. A voice transformation system which could be trained on relatively small amounts of data would allow new voices to be created with much lower cost. In addition, such a system could be used in a situation where the speaker was not available and previous recordings had to be used, such as is the case where a patient had lost the power of speech through disease or injury.

Voice transformation also has other applications such as very low bandwidth speech encoding; the speech may be transmitted without speaker identity information, and this may be resynthesised at the decoding stage. Voice transformation may also prove useful

for multimedia entertainment, as a pre-processing step to speech recognition and also in the field of voice disguise. In addition, gaining a better understanding of the ways in which speakers differ is likely to be valuable more generally in both speech synthesis and recognition.

1.2 Summary of existing transformation approaches

As previously mentioned, voice conversion involves modifying the characteristics of source speech to be like that of the target speaker. There are a number of different parameters to be mapped including voice quality, fundamental frequency and timing characteristics.

There has been a considerable amount of research directed at the problem of transforming voice quality (Arslan 1999, Arslan & Talkin 1997, Stylianou, Cappe & Moulines 1995). The general approach has been to begin with a training phase in which material from source and target speakers is aligned and used to define a transformation which maps the acoustic space of the source speaker to that of the target. Residual Excited Linear Prediction (RELP) analysis has commonly been used to represent the spectral characteristics of the speech. A variety of approaches have been used to map the LPC parameters, including codebooks (Abe, Nakamura, Shikano & Kuwabara 1988), neural networks (Narendranath, Murthy, Rajendran & Yegnanrayana 1995) and most recently Gaussian mixture models (GMMs) (Stylianou et al. 1995). Codebooks had problems due to the discontinuities created when moving from one codebook to another over time. GMMs have been shown to be the most successful, since they avoid the discontinuities associated with the codebook approaches. Early approaches used the residual (spectral detail) of the source speaker unmodified. More recently filters (Arslan 1999) and codebooks (Kain 2001) have been used to transform the residuals.

Very little work has been directed at the problem of mapping the F0 contours of one speaker to another. All existing voice transformation systems simply normalise the mean and standard deviation of the fundamental frequency to be that of the target speaker.

1.3 Summary of proposed approach

In this work the two aspects of voice transformation (voice quality and intonation) will be addressed separately. It is easier to discuss and assess these two components individually, and it would be trivial to integrate the two systems together.

The voice quality transformation system has two main parts, which correspond to the two components of the source-filter model. The first component transforms the spectral envelope as represented by a linear prediction model. The transformation is achieved using a Gaussian mixture model, which is trained on aligned speech from source and target speakers.

The second part of the voice quality transformation system predicts the spectral detail from the transformed LPC parameters. A classifier is used to perform this task, in combination with separate magnitude and phase residual codebooks.

The system has some similarities with earlier work by Kain, however this system is extended to perform well with normal speech, rather than speech spoken in a monotone and with mimicked prosody. Specifically, the system represents residuals without the need for the harmonic sinusoidal model.

The F0 transformation system makes use of the parameterisation described by Patterson (2000). Frequency measurements were taken by Patterson at four selected target points in each sentence. These points are sentence-initial high (S), non-initial accent peaks (H), post-accent valleys (L), and sentence-final low (F). For each sentence there is one sentence-initial high, one sentence-final low and a varying number of peaks and valleys depending on the sentence. The mapping from source to target F0 is piecewise linear, where one segment runs through the points (F_{source}, F_{target}) and (L_{source}, L_{target}), another between (L_{source}, L_{target}) and (H_{source}, H_{target}), and a final segment through (H_{source}, H_{target}) and (S_{source}, S_{target}).

1.4 Outline

The remainder of the dissertation covers the following material:

- Chapter 2 gives an introduction to some of the properties of speech signals. It

also provides a review of the literature on speaker discrimination, and on existing voice transformation systems.

- Chapter 3 describes the component of the presented system that converts the spectral envelope of the speech signal.
- Chapter 4 explains how the system for transforming spectral detail functions.
- Chapter 5 proposes a new method for the transformation of F0 contours. It does so by applying a non-linear mapping function to the source contour, the parameters of which are derived from linguistically motivated features.
- Chapter 6 presents a perceptual experiment to assess the effectiveness of the F0 transformation system.
- Chapter 7 concludes the work and makes suggestions for future work.

Chapter 2

Background

2.1 Properties of the speech signal

2.1.1 Speech model

Human speech is produced by the vocal tract, which starts at the glottis (vocal folds) and ends at the lips. The lungs contract to force air through the trachea and pharynx and out through the nasal and oral cavities. In English there are four different types of sounds that can be created; aspiration noise, frication noise, plosion and voicing. Voicing is a quasi-periodic vibration of the vocal folds - for example the syllable in 'lay'. The frequency of the vibration is called the fundamental frequency or F0 and is perceived as pitch.

The sound wave produced at the glottis is modified by the vocal tract. One useful way of describing speech production is the source-filter model. In this view a source (excitation) waveform is modified by a filter. This model is able to represent most speech phenomena. A simple form of this model works as follows; during unvoiced speech the excitation may be modeled as noise, and during voicing as a series of impulses at the appropriate fundamental frequency. The filter simulates the effect of the vocal tract resonance, to create the resulting speech.

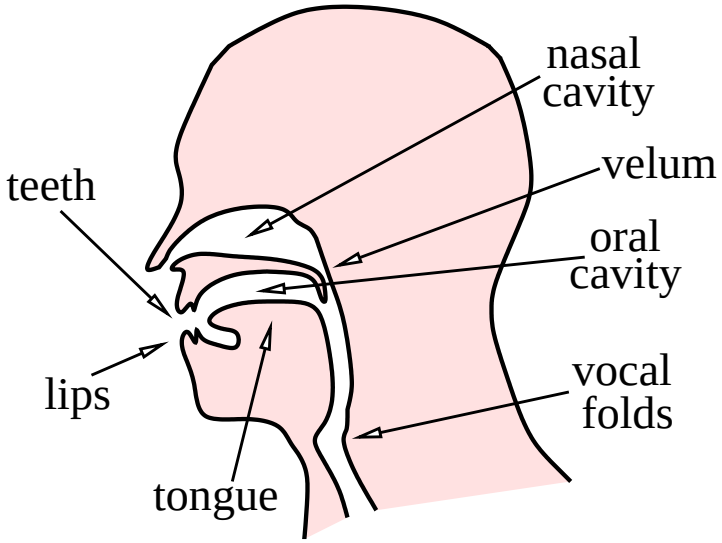

Figure 2.1: The human vocal tract - from (King 2002)

2.1.2 Speaker characteristics

There are a very large number of respects in which speech from different speakers differs. These can be broken down into three main types of cues to speaker identity:

- Segmental: Segmental characteristics describe the timbre of the voice. This encompasses information such as the location and bandwidth of the formants, as well as the frequency spectra of the glottal signal (glottal tilt). For example, glottal tilt dictates whether the speech would be described as breathy.

- Suprasegmental: These characteristics describe the prosodic features of the voice related to the style of speaking. This includes information about how the fundamental frequency (F0) varies during utterances, duration variation, and also how stress varies over the course of a sentence.

- Linguistic: Linguistic features include the choice of words, as well as pronunciation differences due to dialects. For example, if a speaker uses the word 'wee' rather than 'small', this suggests that they may be Scottish.

2.2 Speaker discrimination

Human recognition of speakers is by no means perfect, as was shown in an experiment by Ladefoged and Ladefoged (1980). They measured the ability of subjects to recognize a group of 53 voices, 29 of which were familiar to the speaker. The subjects were asked to name the speaker. 31% of the familiar voices were correctly identified from a single word, 66% from a single sentence and only 83% from 30s of speech!

Research by Necioglu et al (Necioglu, Burhan, Clements, Barnwell & Schmidt-Nielson 1998) indicates that the most important cues to speaker discrimination are as follows: median F0 and vocal tract features such as length for males; median F0, glottal tilt, and mean duration of unvoiced segments for females. Matsumoto et al (1973) investigated the ability of speakers to discriminate between Japanese vowels from different speakers. He found that average F0 accounts for 55% of the variance, F0 and spectral tilt together account for 71% and F0 and the lowest three formants accounts for 81% of the variation.

An investigation by van Lancker et al (1985) tested the ability of listeners to recognize voices when played normally as well as backwards. He found that for some speakers the listeners were able to recognize the speaker nearly as well, whereas for others they performed poorly. On the basis of this and other research, Van Lancker concluded that the critical cues for recognition are not the same for all speakers.

In a paper by Zetterholm (2000) it was shown that voice quality, pitch register, intonation and other prosodic aspects of the voice and speech style are important features to capture in order to succeed in imitating another voice. This was demonstrated through a series of perceptual tests.

Very little work has been done on the transformation of linguistic features. It is particularly hard problem since in order to do so, one must recognise the words spoken, identify those words which should be mapped, and then synthesise the appropriate word with the voice quality of the target speaker.

2.3 Existing voice transformation systems

In order to carry out voice transformation there are a number of different parameters to be mapped, including spectral dynamics, fundamental frequency and timing. These characteristics can broadly be decomposed into two parts; firstly voice quality, and secondly characteristics of the fundamental frequency and timing.

2.3.1 Voice Quality Conversion

There has been a considerable amount of research directed at the problem of voice quality transformation (Arslan 1999, Arslan & Talkin 1997, Stylianou et al. 1995). The general approach has been to begin with a training phase in which material from source and target speakers is aligned and used to define a transformation which maps the acoustic space of the source speaker to that of the target. There are two key questions to be addressed; how should the speech signal be represented and how should the mapping be achieved?

Parameter spaces for the representation of speech

The first approaches were based around linear predictive coding (LPC) (Makhoul 1975). The central idea behind LPC is that speech can be modeled by applying an appropriate filter to a pulse-like excitation signal. This approach had the disadvantage of creating voices which sound rather 'robotic' and unnatural. The technique was improved with a method in which the residual error was measured and used to produce an excitation signal. This technique is known as residual-excited linear prediction (RELP). This technique is much better at representing voice signals accurately, and has been used extensively in speech compression and synthesis. It has also been used in much of the work on voice transformation (Arslan 1999, Arslan & Talkin 1997, Stylianou et al. 1995).

Other work has made use of a technique called sinusoidal modeling (Bailly, Bernard & Coisnon 1998). The idea is to decompose the speech signal into a sum of sine waves. The discrete Fourier transform (DFT) is commonly used to perform this decomposition. The problem with this approach is that the dimensionality of the sinusoidal representation is high. This makes it very difficult to perform transformations on this data. In order to

overcome this problem McAulay and Quatieri (1995) proposed the use of all-pole models (like that of LPC) to represent the frequency spectrum at each point in time. Then as few as 10 to 20 linear prediction coefficients may efficiently represent the spectrum for each time segment. There are well known techniques for deriving these coefficients based on the frequency power spectrum, from the work on LPC.

This sinusoidal modeling technique models the harmonic part of the signal well, however it does not model the noise part effectively. One approach for dealing with this problem is the use of harmonic plus noise models (HNMs), which have achieved good results (Bailly 2001) in representing speech signals. This approach involves using the sinusoidal method for representing the harmonic part of the signal and stochastic models for the remaining part. Ahn and Holmes proposed a method for analysis using such methods (Ahn & Holmes 1997). Bailly (2001) developed appropriate synthesis methods. This technique is also used in the MBROLA synthesiser (Dutoit & Leich 1993).

Mapping methods

There have been a number of approaches to the problem of determining the mapping of parameters from the source speech to the target speech. Arslan and Talkin proposed a system (Arslan 1999, Arslan & Talkin 1997), in which the speech of both speakers is marked up automatically into phonemes. Next, the line spectral frequencies for each frame of each utterance are calculated and labeled with the relevant phoneme. Following this, the centroid vector for each phoneme is calculated, and a one-to-one mapping from source to target codebooks is established. This process is also performed on the residual signal. The transformation may then be carried out by the use of codebook mapping. However, the quality suffered due to the fact that the converted signal is limited to a discrete set of phonemes.

Stylianou, Capped and Moulines suggested improvements to the method of Arslan and Talkin through the use of Gaussian mixture models of the speakers' spectral parameters (Stylianou et al. 1995). The source and target speech was first time-aligned using dynamic time warping. Mel-frequency ceptral coefficients (MFCC's) were computed for each frame of speech, and a vector was produced were each element consisted of the source MFCC's followed by the target MFCC's for the same frame. A Gaussian mixture

model was then fitted to this data, using the expectation-maximization algorithm (EM). This method using GMM's led to less unnatural discontinuities within the synthesized speech, than the method described above based on vector quantization.

Kain (2001) proposed a solution where he mapped the spectral envelope in the manner described by Stylianou et al. (1995), but then predicted the residual from the predicted spectral envelope. This resulted in fewer artifacts than existing systems, however this work was restricted to speech where the speakers were speaking in a monotone, and where the speakers where asked to mimic the segment and word durations of a template speaker.

2.3.2 Intonation Transformation

As was previously discussed, intonation plays an important role in speaker identity. The only approach which has so far been proposed to the problem of transforming the F0 contour of one speaker to another, simply consists of modifying the source F0 contour such that it has the mean and standard deviation of that of the target speaker (Arslan 1999). However, two contours may have the same mean and standard deviation, but differ greatly in how they are perceived, as was noted by Ladd and Terken (1995). Clearly, a more sophisticated approach would benefit voice transformation systems.

The remaining work discussed here was not performed with a view to the development of voice transformation systems, however it does throw light on the ways in which intonation may be described and measured. The Tones and Break Indices (ToBI) system proposed by Silverman et al (1992), offers a method for describing intonation contours in terms of a series of intonational events. These comprise tones which describe pitch accents and the nature of the contour at the end of a phrase, and break indices which describe the nature of pauses.

A study by Clark (1999) shows that the first tone group in a phrase has a higher mean F0 than the other tone groups. This indicates that phrase initial accents have a special status. In the Clark paper, a tone group describes any group which has a ToBI break index of 3 or more. This equates to a break which is larger than that typically between two words, and smaller than the break between sentences. Phrase final tones

also are lower than other categories. Medial tone groups appear to have very similar characteristics to one another.

Ladd and Terken (1995) conducted an investigation of pitch range variation within and across speakers. Schriberg et al. (1996) have since expanded on this work. The work relies on relatively invariant pitch accents in intonation contours, which broadly correspond to the tone accents within the ToBI accent system. Schriberg et al. (1996) investigate methods for mapping from a normal F0 contour to a 'raised' form, where the speaker is attempting to make themselves heard over a noisy communication channel. It was found that the raised mode can be accurately predicted using a linear function with speaker specific parameters. A two parameter model was used to predict the raised target (R) from the normal target (N) as follows:

$$R = aN + b \tag{2.1}$$

Patterson (2000) expanded on the work of Schriberg et al. by investigating a number of measures of pitch level and span, where pitch level is a measure of how high a voice is, and pitch span is a measure of how much the pitch varies between high and low. As part of this work Patterson proposed a method for measuring key features of a speakers intonation contours. Frequency measurements were taken at four selected target points in each sentence. These points were sentence-initial high (S), non-initial accent peaks (H), post-accent valleys (L), and sentence-final lows (F). For each sentence there is one sentence-initial high, one sentence-final low and a varying number of peaks and valleys depending on the sentence. Analysis was performed on approximately a minute of speech for each speaker. The values were collected into their respective categories and then averaged to get representative data for the speaker.

2.4 Corpora

In this thesis two corpora of data will be used. The following sections describe these collections of data.

2.4.1 The Boston University Radio Corpus

The Boston University Radio Corpus (Ostendorf, Price & Shattuck-Hufnagel 1995) was selected for use as both training and test data, since it provided a large amount of speech of a number of speakers (both male and female) uttering the same sentences. Furthermore, the speech is phonetically segmented and prosodically labeled. F0 tracks are also provided for each utterance, together with voiced/unvoiced labeling. The speech is sampled at 16kHz with 16bit resolution. Some of the waveforms in this corpus had inverted polarity, so these were corrected. It was important that all the waveforms had the same polarity, since it made extracting accurate pitchmarking easier.

2.4.2 The Patterson Corpus

The details of the collection of this data is described by Patterson (2000). It consists of eight passages, each of approximately a minute in length, each read by a total of 32 speakers. Each passage was read in a normal, natural style. No special instructions were given regarding the intonation to be used by the speakers. A number of statistics relating to the fundamental frequency of the speech for each speaker are also recorded. These statistics include mean, standard deviation, maximum frequency, etc.

2.5 Tools

F0 tracks were extracted from the speech signals using the pda program from Edinburgh Speech Tools (Taylor, Caley, Black & King 1999). Pitchmarks were determined using the pitchmark program which is also from Edinburgh Speech Tools.

All other processing was performed using specially written MATLAB code. The VOICEBOX speech processing toolbox for MATLAB was also used (Brookes 1998).

Chapter 3

Transforming the spectral envelope

As described earlier the problem of transforming voice quality may be decomposed into two parts corresponding to the two components of the source-filter model. The first component transforms the spectral envelope and will be described in this chapter. Howevert, we will first briefly describe how the amplitude contour may be transformed.

3.1 Transforming the Amplitude Contour

3.1.1 Computing transformation parameters

As previously discussed, one of the respects in which the speech of two speakers differs is the amplitude of the speech over the course of a sentence. In order to transform the amplitude envelope, we first compute the RMS amplitude of each frame of speech for all training speech for both source and target speakers. The mean and standard deviation of the amplitude of the voiced segments of speech was computed for the whole training set. These values were also computed for the unvoiced segments of speech. Sections of the waveform below the noise floor are excluded from these calculations.

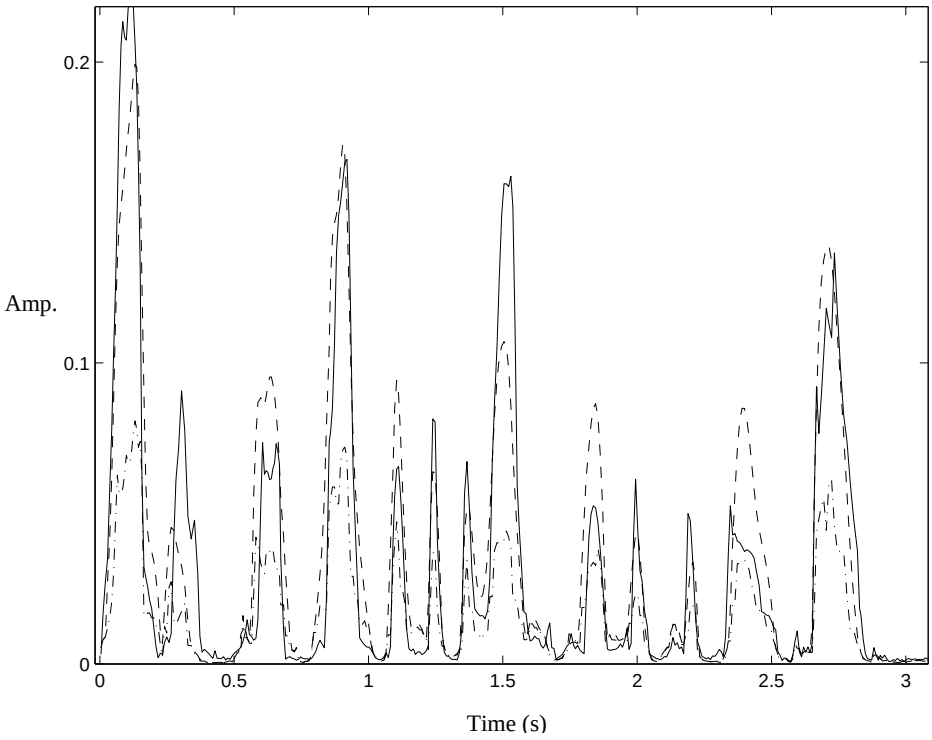

Figure 3.1: Graph showing an example source, target and predicted RMS amplitude contour. The solid line repesents the target contour, the dashed line the predicted target contour, and the dotted line the source contour)

3.1.2 Transformation

The unvoiced sections are normalized to have the mean and standard deviation of the unvoiced sections of the target speakers speech, and a similar process is carried out for the voiced sections. A Hanning window of length seven is then used to smooth the resulting amplitude envelope. The resulting predicted target amplitude envelope may then be applied to the source speech by scaling each frame of speech to have the predicted target amplitude of that frame.

3.2 Analysis

The periods of silence prior to, and following each passage of speech were first removed, since this would cause problems during later processing. This is because the dynamic

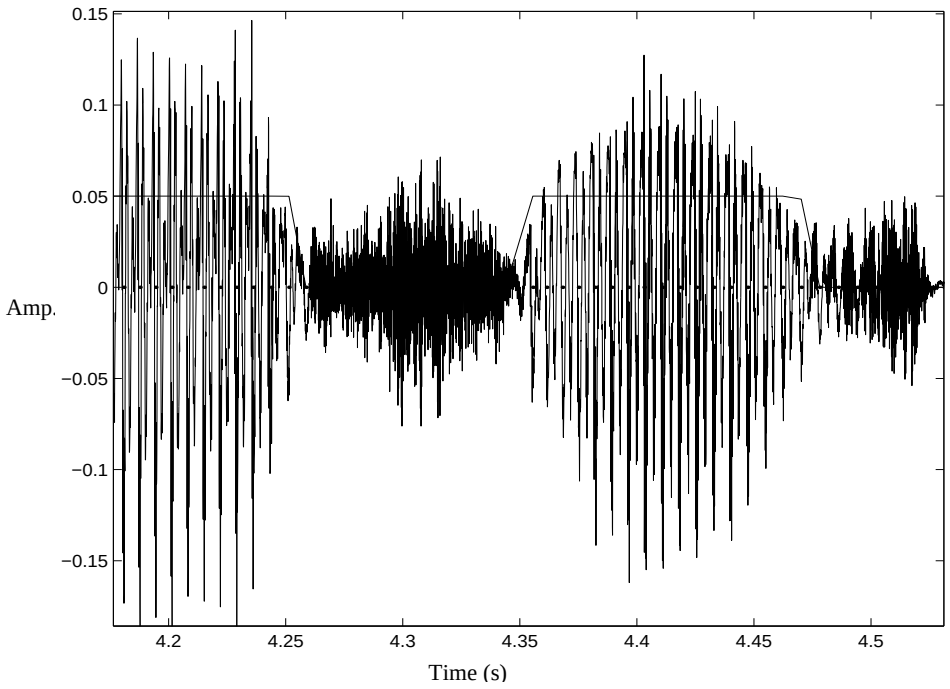

Figure 3.2: Example marked up speech segment.

time warping implementation we use may only scale each frame by at most a factor of two (see section 3.3.1). Therefore if the source piece of speech has a long silence at the beginning and the target does not, a good alignment cannot be found.

When pitchmarking was performed, the parameters to the program were carefully adjusted in order to avoid pitch-doubling and halving artifacts. The pitchmarks were also post processed to align them with waveform maxima, as it was found this gave more consistent pitchmarking. Where the speech was not voiced, pitchmarks were inserted at a constant frequency of 125Hz.

We carry out frame based analysis of the speech, since for short segments of speech the spectrum may be considered to be stationary. The speech was divided into short overlapping frames, where each frame was two pitch periods long and was centred around the current pitchmark. These frames were then windowed using a Hanning window as can be seen in figure 3.3. The Linear Prediction Coefficients (LPC) of the filter were computed using the autocorrelation method (Rabiner & Schafer 1978). The order of LPC analysis, O_{LPC} was one of the variables of the experiment. The LPC filter coefficients

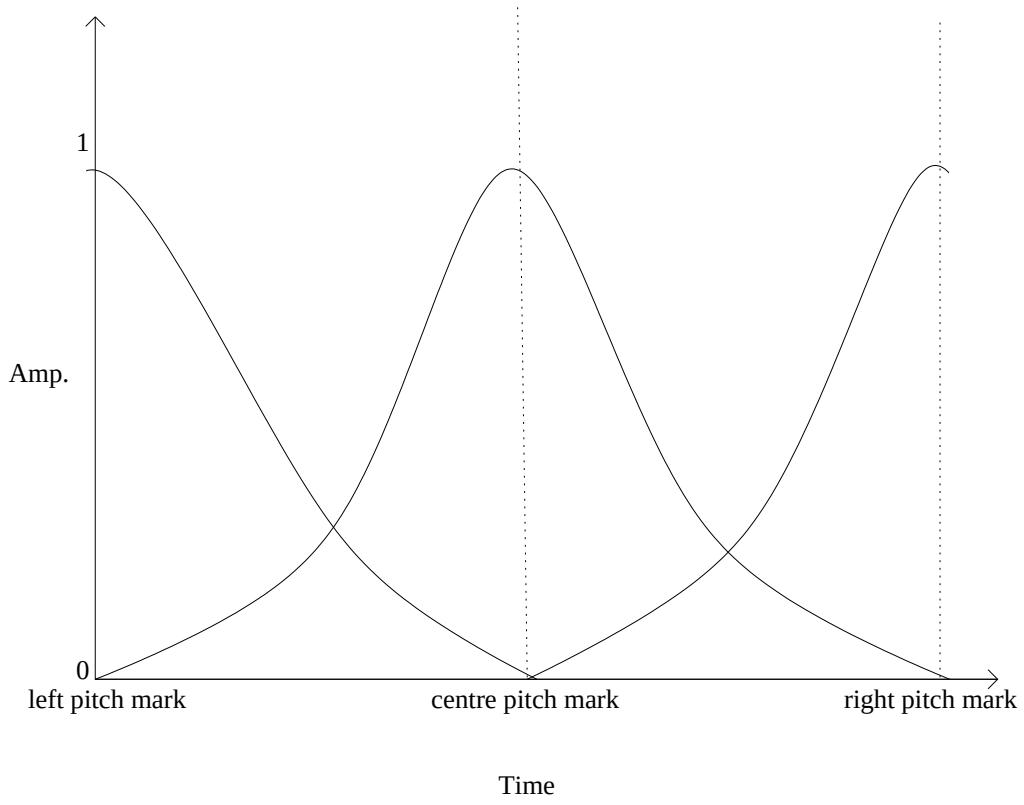

Figure 3.3: Diagram showing the windowing used

were converted into line spectral frequencies (LSFs) (Rabiner & Schafer 1978). Line spectral frequencies have better interpolation characteristics, which is important for this system since the target LSFs will be formed from a weighted sum of source LSFs.

The ear has better frequency resolution at lower frequencies (Ladefoged 1962). In order that the numerical distance between a pair of LSFs better reflect the perceptual distance between them, this non-linear frequency resolution must be accounted for. One scale that achieves this is the Bark scale. The Bark warping function b is as follows:

$$b(f) = 6.log(\frac{f}{1200} + \sqrt{(\frac{f}{1200})^2 + 1}) \tag{3.1}$$

The Bark-warping process was applied to the LSFs for each frame of speech. Residuals were computed by inverse filtering each frame of speech using the associated LPCs.

3.3 Training

The purpose of the training stage is to estimate the parameters of a transformation function that will map source features (LSF vectors) to target features with minimum error. In order to do so, the features of source and target must be time-aligned so that an appropriate mapping can be established. Approximately two minutes of speech was used for training. It was found that using less speech degraded the quality significantly (see section 3.6.3), and using more resulted in unacceptably high computation time.

3.3.1 Time-alignment

Time-alignment was carried out on each set of sentences for each source/target speaker pair. Firstly, Cepstral Coefficients (CCs) (Rabiner & Schafer 1978) for each Bark-warped set of LSFs were calculated, together with the log of the associated residual energy. CCs and log energy were chosen as parameters for alignment since it was found that they gave better results than when using LSFs or LPCs. The Dynamic Time-Warping (DTW) (Sakoe & Chiba 1978) algorithm was used to find the minimum error alignment of these two feature vectors. Slopes of 0.5, 1 and 2 were allowed for each point within the alignment. Features were either duplicated or deleted within the source feature vector in order to get matching source and target vectors.

3.3.2 Fitting the GMM

Pre-GMM estimation rejection of poorly matched data

There is a great deal of variability within and across speakers as to the way words such as 'the' and 'a' are spoken. In some cases they even leave out these words entirely, despite the fact that it was a read-text task. There are also sometimes significant differences due to differing dialects. For example the words 'lot off' may sometimes be pronounced as 'lotta'. When trying to compute a transformation function for mapping from one speaker to another, it is helpful to reject these extreme cases from the training set. Our approach is in contrast to all previous approaches which have not used a strategy of rejecting poorly matched data. Two strategies were employed to help isolate the frames where this occurred.

Those pairs of aligned frames of speech where one speaker's speech was voiced and the other speaker's was unvoiced were rejected from the training set. If they have different voicing classification, this suggests that they were poorly aligned. As described in section 3.1, the predicted amplitude envelope of the target is computed by modifying the source amplitude envelope. The frames of speech where the predicted amplitude is more than three times larger or smaller than the actual amplitude at that point are also rejected. When combined, these methods typically reject about 25% of the data, and were found to significantly improve quality in an informal listening test.

Estimation of the transformation function

The transformation function must map the features of the source speaker to the appropriate target speaker features. Gaussian mixture models are one possible approach to this problem. They have the useful property of being continuous, as opposed to a lookup table based approach such as that of Arslan and Talkin (1999). It has been shown that GMMs have as good as or superior performance at the task of voice transformation to other transformation approaches based on neural networks, vector quatization or linear regression (Baudoin & Stylianou 1996).

We use the joint density approach (Ghahramani & Jordan 1994) as applied to VT by Kain (2001). This approach involves fitting a GMM to the joint density $P(x, y)$ and then predicting y from x by finding $E[y|x]$ (the expected value of y given x). To do this we form a vector Z where each element is composed of the source features X and target features Y, where

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix}$$

The probability distribution of a GMM with Q_{LSF} components (Ghahramani & Jordan 1994) is given by:

$$p_{GMM}(x; \alpha; \mu; \Sigma) = \sum_{q=1}^{Q_{LSF}} \alpha_q N(x; \mu_q; \Sigma_q), \sum_{q=1}^{Q_{LSF}} \alpha_q = 1, \alpha_q \geq 0 \tag{3.2}$$

where α_q is the weight for component q, $N(x; \mu_q; \Sigma_q)$ is the n-dimensional normal distribution with mean μ_q and covariance Σ_q which can be computed by

$$N(x;\mu_q;\Sigma_q) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma_q|}} e^{(-\frac{1}{2}(x-\mu_q)^T\Sigma_q^{-1}(x-\mu_q))} \tag{3.3}$$

The probability of a datapoint x belonging to a particular class p may be computed using Bayes' rule, which is

$$P(c_p|x) = \frac{\alpha_p N(x;\mu_p;\Sigma_p)}{\sum_{q=1}^{Q_{LSF}} \alpha_q N(x;\mu_q;\Sigma_q)} \tag{3.4}$$

The Expectation Maximization (EM) (Ghahramani & Jordan 1994) algorithm is an iterative algorithm which may be used to find the most likely GMM parameters (α, μ, Σ) for a given set of data. To start the process we set α_q equal to $1/Q_{LSF}$ for all $q = 1...Q_{LSF}$, Σ_q equal to the identity matrix for all q, and set each μ_q by applying the K-means algorithm (MacQueen 1967). The EM algorithm was then run until either the likelihood $P_{GMM}(x;\alpha;\mu;\Sigma)$ was maximized, or 30 iterations were exceeded. After these 30 iterations, the maximum likelihood was only found to increase marginally. It is necessary to add a small quantity to the diagonal entries of the covariance matrices after each iteration, in order to stop them becoming too close to singular. The value 0.00001 was used, which was found by experimentation and is somewhat smaller than the value of 0.001 used by Kain (2001). The number of components of the GMM Q_{LSF} was one of the parameters which was varied in the experiment.

Post-GMM estimation rejection of poorly matched data

Once the GMM had been fitted to the training data, a second stage of rejecting poorly matched data took place. We rejected $R\%$ of the data which had lowest probability $P(c_p|x)$ under the GMM. These points may be regarded as remaining outliers and are due to poor alignment. A GMM was then re-estimated for the remaining data points. The optimum proportion for rejection (15%) was found through informal listening tests. The appropriate amount to reject is likely to depend on the extent to which the source and target speakers' accents and prosody differ.

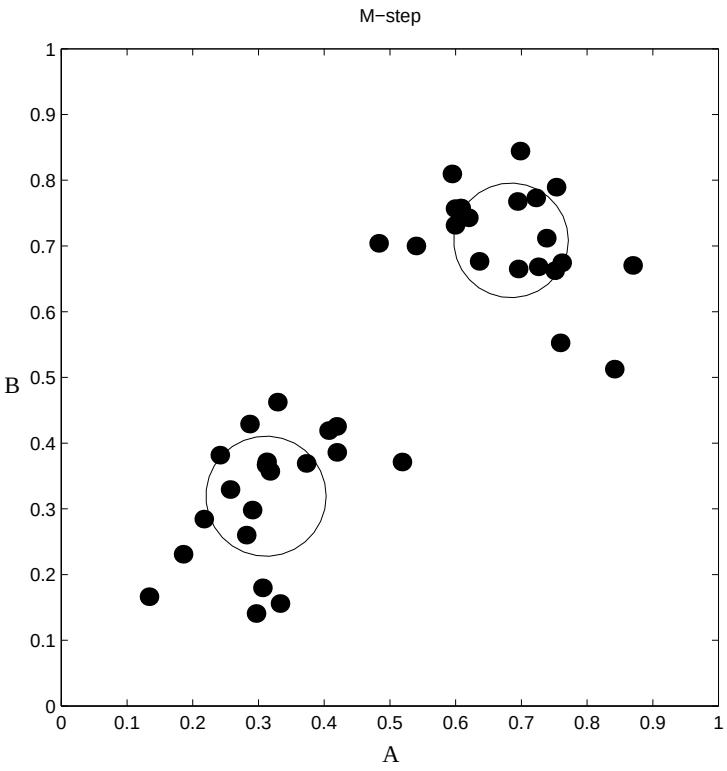

Figure 3.4: A simple two component GMM.

3.4 Transformation

In order to carry out transformation, the speech is first analyzed by computing Bark-warped LSFs for each frame. X and Y are the aligned source and target feature streams. For each frame of source LSFs, the most likely target LSFs are computed. The expected value of the target LSFs for a target frame, y, may be computed using the appropriate source frame LSFs x as follows:

$$E[y|x] = \int y.p(y|x)dy \tag{3.5}$$

$$E[y|x] = \sum_{q=1}^{Q_{LSF}} (\mu_q^Y + \Sigma_q^{YX}(\Sigma_q^{XX})^{-1}(x - \mu_q^X)).p(c_q|x) \tag{3.6}$$

where

$$\Sigma_q = \begin{bmatrix} \Sigma_q^{XX} \Sigma_q^{XY} \\ \Sigma_q^{YX} \Sigma_q^{YY} \end{bmatrix} \tag{3.7}$$

$$\mu_q = \begin{bmatrix} \mu_q^X \\ \mu_q^Y \end{bmatrix} \tag{3.8}$$

After the predicted LSFs have been computed, a smoothing function is applied to each of the LSF coefficients, in order to restrict the difference in value between neighbouring frames. The filter used is a 2nd order lowpass digital Butterworth filter with a cutoff frequency of F_{LP} of half the sampling rate. The sampling rate is the rate of pitchmarking. This low pass filtering of the LSFs is motivated by the fact that the components of the human speech system responsible for filtering the signal from the glottis are restricted in how rapidly they may change their response.

3.5 Synthesis

Once a vector of target LSFs has been predicted, the LSFs are then converted from Bark to Hertz and converted to LPCs. The associated target residuals are then found, and a Hanning window is applied prior to inverse filtering with the associated LPC parameters.

The resulting speech is then created by PSOLA (pitch synchronous overlap-add) of all the frames of windowed speech.

Since an overlap and add resynthesis method is used, it is possible to modify the pitch easily (Quatieri & McAulay 1992) by moving the pitchmarks closer together or further apart, and to modify duration by duplicating or deleting them.

3.6 Evaluation

3.6.1 Speech Data

Data from the Boston University Radio News corpus as described in 2.4.1 was used to both train and test the system. Four speakers, (two male and two female) were selected for the experiment. They are labelled as f1a, f2b, m1a, m2b within the corpus. The speakers are all native speakers of English and have North American accents. They are all professional news readers. f1a and f2b are female and m1a and m2b are male. Further details of the nature of the speakers can be found in documentation for the corpus (Ostendorf et al. 1995). Experiments were run on the following transformation combinations; m1 to m2, m2 to m1, f1 to f2 and f2 to f1. T_{train} seconds of data were used for training. The test set consists of one minute of speech. The training and testing sets do not intersect. All performance measures presented in this chapter were found on the test set.

3.6.2 Performance indices

The error between two aligned sets A and B of LSF vectors may be computed as follows:

$$E_{LSF}(A,B) = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{p}\sum_{i=1}^{p}(L_A^{m,i} - L_B^{m,i})^2} \tag{3.9}$$

where M is the number of frames, p is the LPC order and $L^{m,i}$ is the i_{th} LSF vector component in frame m.

This is however not a useful way to evaluate the performance of a transformation system since it doesn't take into account the 'difficulty' of the mapping, i.e. the difference between the source and target vectors. The difference between two speakers is called

Source Speaker	Target Speaker	Inter-Speaker Error	P_{LSF}
m1b	m3b	0.014	0.373
m3b	m1b	0.014	0.424
f1a	f2b	0.013	0.312
f2b	f1a	0.013	0.292

Table 3.1: Table showing the inter-speaker error and performance (P_{LSF}) for a variety of source and target speaker pairs. ($O_{LPC} = 20, Q_{LSF} = 12, F_{LP} = 0.3, T_{train} = 60s or 120s$)

the inter-speaker error $E_{LSF}(t(n), s(n))$, where $t(n)$ are the LSFs of the target speech. The LSFs of the predicted target speech are represented as $\hat{t}(n)$. The transformation error is the difference between the predicted and actual LSFs ($E_{LSF}(t(n), \hat{t}(n))$). Kain suggested an LSF performance index P_{LSF} for assessing the quality of transformation in a voice transformation system, as follows:

$$P_{LSF} = 1 - \frac{E_{LSF}(t(n), \hat{t}(n))}{E_{LSF}(t(n), s(n))} \tag{3.10}$$

A value of $P_{LSF} = 0$ indicates that the output of the system is no more similar to the target than the source is, whereas a value of $P_{LSF} = 1$ indicates that the output of the system is identical to the target. In general, a higher value for P_{LSF} suggests a better system.

3.6.3 Results

As previously discussed, the experiments were carried on two pairs of speakers, with each speaker used once as source and once as target. In table 3.1 it is possible to see that the system has significantly different performance (P_{LSF}) depending on which voices are to be transformed. It can be seen that the performance for mapping speaker S to T is different to the performance when mapping T to S. Example WAV files of the output of our system may be found online (Gillett 2003).

A large number of different experiments were carried out in order to discover the effects of varying various parameters. The variables in the following experiments are as

	P_{LSF}	
Q_{LSF}	$T_{train} = 60s$	$T_{train} = 120s$
4	0.3484	0.3534
8	0.3506	0.3588
12	0.3512	0.3619
16	0.3508	0.3618
20	0.3490	0.3618

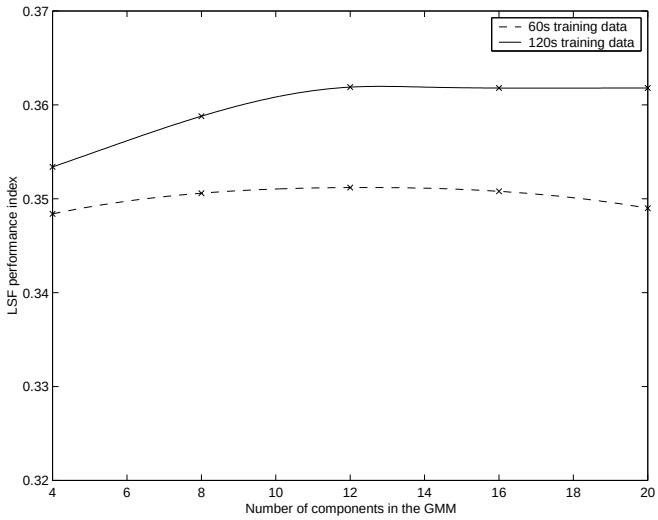

Figure 3.5: Graph showing the relationship between the number of components in the GMM (Q_{LSF}) and the mean performance of the resulting system (P_{LSF} averaged over all data). ($O_{LPC} = 20, F_{LP} = 0.3, T_{train} = 60s$.)

follows: order of LPC analysis O_{LPC}, the number of components in the GMM Q_{LSF}, the cutoff frequency of a low pass filter applied to the transformed LSFs F_{LP} and finally the amount of training data T_{train}. The results were averaged for all four combinations of source and target speakers.

Figure 3.5 shows the effect of changing the number of components in the GMM. A value of $Q_{LSF} = 12$ provided the best performance of the values we tested, regardless of the amount of training data used. The performance does not improve when there are more than 12 components of the GMM, and this is the case regardless of the amount of data trained on.

The relationship between the order of LPC analysis and the performance of the resulting system is shown in figure 3.6. It can be seen that a value of $O_{LPC} = 20$ provides the best performance. This is consistent with the order of LPC analysis used in similar tasks (Kain 2001).

The relationship between the cutoff of the low-pass smoothing filter and the performance of the resulting system is shown in figure 3.7. The optimum cutoff is a value of $F_{LP} = 0.3$. The performance when doing smoothing with an appropriate cutoff value is substantially higher than the performance with no smoothing. This indicates that the

O_{LPC}	P_{LSF}
8	0.3310
12	0.3402
16	0.3492
20	0.3512
24	0.3505

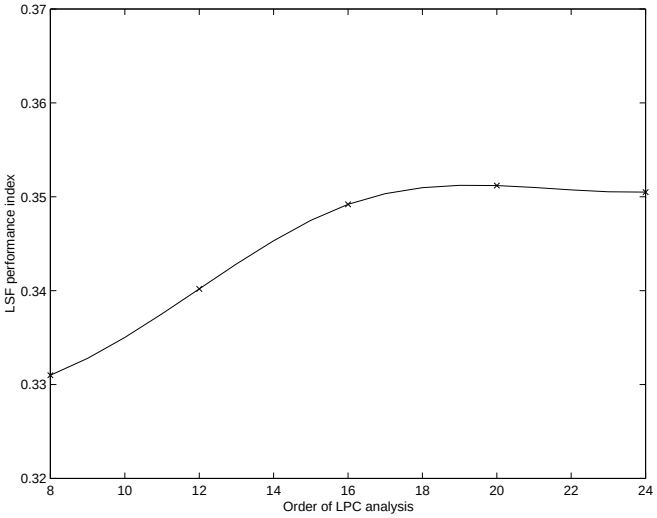

Figure 3.6: Graph showing the relationship between the order of LPC analysis (O_{LPC}) and the performance of the resulting system (P_{LSF}). ($Q_{LSF} = 12, F_{LP} = 0.3, T_{train} = 60s$)

smoothing plays a key role in obtaining good performance from the system. The movement of LSFs in natural speech is quite smooth. However, the transformation system works on a frame by frame basis resulting in noisy tranformed LSFs. Therefore, if there is too little filtering the transformed LSFs are still too noisy, and if there is too much filtering then information is lost.

Figure 3.8 shows how performance improves when a larger amount of training data was used. As the amount of training data is increased, the performance of the system improves. The largest amount of data used for training was 120 seconds which provided a value of $P_{LSF} = 0.3619$. After a value of $T_{train} = 30s$ is reached, the increase in performance when more training data is used is much smaller.

3.7 Conclusion

The results show that in order to gain the best performance, the following parameters should be used: $Q_{LSF} = 12, O_{LPC} = 20, F_{LP} = 0.3, T_{train} = 120s$. This leads to a performance of $P_{LSF} = 0.3619$. In existing research, the voice transformation system which has the highest performance is a system by Kain (2001), which has a performance

F_{LP}	P_{LSF}
0.1	0.3218
0.2	0.3501
0.3	0.3512
0.4	0.3496
0.5	0.3464
1.0 (No smoothing)	0.3333

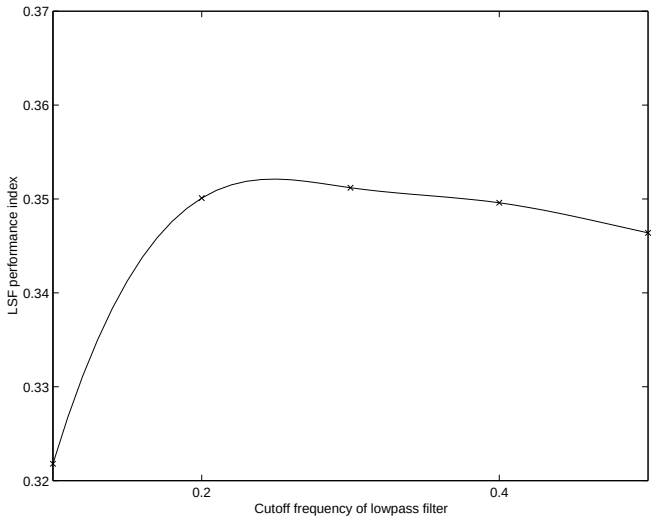

Figure 3.7: Graph showing the relationship between the cutoff (as a fraction of the Nyquist frequency) of a lowpass filter applied to the LSFs (F_{LP}) and the performance of the resulting system (P_{LSF}). ($O_{LPC} = 20, Q_{LSF} = 12\, T_{train} = 60s$)

T_{train}	P_{LSF}
7.5	0.2874
15	0.3255
30	0.3448
60	0.3512
120	0.3619

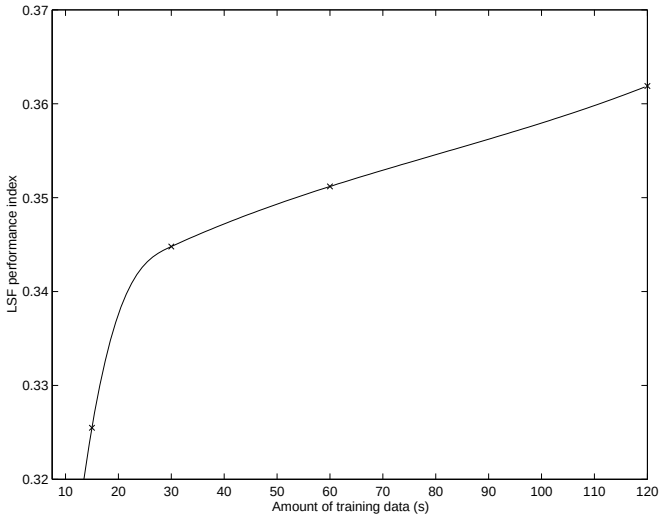

Figure 3.8: Graph showing the relationship between the amount of training data (T_{train}) and the performance of the resulting system (P_{LSF}). ($O_{LPC} = 20, Q_{LSF} = 12, F_{LP} = 0.3$)

of $P_{LSF} = 0.31$. Therefore, our system outperforms this system. Unfortunately we do not have access to the same test and training data as Kain used. Kain does not give the duration in seconds of the training data used, however he does state that 40 sentences were used for the training corpus. The text of these sentences was taken from the TIMIT (Garofolo, Lamel, Fisher, Fiscus, Pallett & Dahlgreen 1990) database, which has a typical sentence length of 4 seconds. Therefore approximately $40*4 = 160$ seconds were used for training. Therefore we conclude that it is likely that Kains system was trained on more than 120 seconds of speech. The training data used in our experiment is significantly different since it is prosodically varied. Kain asked the speakers to speak in a monotone, and to mimic the F0 contour, segment and word durations of a particular speaker to minimize intra-speaker error. It is easier to make the transformation if the timing and F0 are similar, since it is easier to find a good alignment, and also because the F0 of the speech does not need to be altered so much. Our system improves over Kain's system since it is able to deal with a more difficult problem: natural, prosodically varied speech. The improved performance index of our system over Kains could be due to the fact that our system rejects poorly aligned data from the training set, and also be due to the smoothing applied to the mapped LSFs.

Chapter 4

Transforming the spectral detail

In this chapter we address the problem of transforming the spectral detail, in order to get a higher quality transformation. The spectral detail is represented by the residual.

4.1 Motivation and overview

4.1.1 Motivation

A study by Kain and Macon (1998) shows that in an LPC VT system, (such as that described in chapter 3) when using source residuals for resynthesis the resulting speech was judged to be closer to the target speaker 52% of the time (i.e. close to chance). However, when the target residuals were used, the resulting speech was judged to be closer to the target speaker 100% of the time. This result shows that residuals play an important part in the characterisation of a speaker.

4.1.2 Approach

The system will predict the residuals from the transformed LSFs which were predicted in chapter 3. This is similar to the system put forward by Kain (2001), however our approach is different in a number of respects and it removes the requirement that all the speech is uttered in a monotone and with mimicked prosody. In addition, Kain's work made use of a sinusoidal model, whereas our work does not. It may seem strange to attempt to predict the characteristics of the source from the characteristics of the

spectral envelope, since the source-filter model is based on an assumption that the residual is independent of the spectral envelope. However, if only one speaker is considered we will show that the residual is sufficiently correlated with the spectral envelope that prediction is possible.

4.2 Analysis

The speech of a single speaker whose residuals are to be predicted was first analysed in the manner described in chapter 3. Although the pitchmarks were adjusted to lie at peaks in the waveform, after the inverse LPC filter was applied the centre of the frames no longer corresponded to the peaks in the residual. This was due to the phase shift introduced by the inverse LPC filter. Due to the nature of the processing which will be performed (see section 4.3), we wish the residuals to align with one another, so that their phases are most similar. Therefore, the residuals were further processed as follows: for all voiced frames, the peak amplitude within the middle third of the frame was found. If this peak was over an empirically determined peak threshold (0.0017), then the frame was moved such that this point lay at the centre of the residual.

Those residuals that did not have a peak above the threshold were marked as 'suspected unvoiced'. This was performed prior to the application of the Hanning window. All residuals which were suspected of being unvoiced, and whose neighbours were also suspects were set to be unvoiced. This process was performed to try to eliminate incorrectly marked-up voicing. For each residual, the magnitude and phase spectra were computed using the Fast Fourier Transform (FFT).

4.3 Training

During training, a vector was created where each element comprised of the frames' Cepstral Coefficients. Those elements of the vector where the associated frame was not voiced were removed. The system only attempts to predict the residual for voiced frames of speech, since the residual in unvoiced frames contains very little information about the nature of the speaker, as there is no vocal fold activity. A GMM with Q_{rp} components

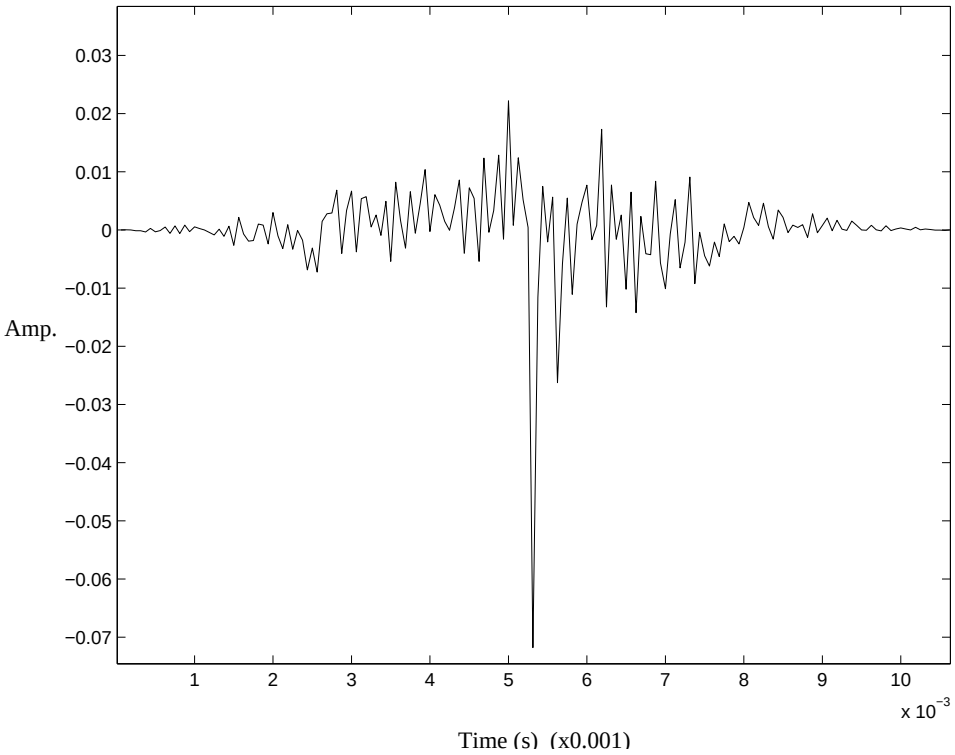

Figure 4.1: Typical windowed residual

was then fitted to this data. For each component of the GMM a codeword was calculated. This codeword has a magnitude spectrum, which was computed by summing the magnitude spectra of all the residuals, weighted according to the probability of each datapoint (frame of Cepstral Coefficients) belonging to that component. If $h_{q,i}$ is the posterior probability of C_{train} (the training data) for a class q and frame i, then the magnitude for codebook entry q is:

$$m_q = \sum_{i=1}^{N} M_i . \frac{h_{q,i}}{\sum_{j=1}^{N} h_{q,j}} \tag{4.1}$$

The codeword also contains a table of all the phases of the frames which have a 90% or greater probability of belonging to that component. The value of 90% was chosen in order to ensure there was a large enough number of entries in the table to provide reasonable spread of lengths of the associated phases, which will be important for reasons explained in 4.4.

4.4 Residual Prediction

Given the set of Cepstral coefficients associated with a voiced frame of speech, the residual may be predicted in the following way. The magnitude of the residual was computed by summing all the codeword magnitudes, weighted according to the probability of the datapoint belonging to the component that this codeword is associated with. This is:

$$\widehat{M_i} = \sum_{q=1}^{Q} m_q . h_{q,i} \tag{4.2}$$

This method for predicting magnitudes has the desirable property of the resulting magnitude spectrum changing smoothly provided the input parameters change smoothly. This avoids many of the artifacts associated with vector quatization methods (Arslan 1999).

Unfortunately, the same approach may not be taken with the phase, since phase may not be interpolated using a weighted sum analogous to equation 4.2 due to the way in which phase may 'wrap around' (i.e. a phase of 2π is equivalent 0). In addition, resampling a phase to be a different length also requires interpolation. Resampling would

be required to alter the residual to be the desired length for each frame. Although one might think that the phase could be unwrapped to overcome these difficulties, this is an error prone process (Gold & Morgan 2000, Kain 2001), and was therefore not used in this system. Instead, the following method was chosen: the phase was computed by finding the most likely component of the GMM and choosing the phase from the associated table that was closest in length to the desired frame length.

After a phase and magnitude vector had been obtained, the magnitude vector was resampled to be of the same length as the phase vector. An anti-aliasing FIR filter of length 10 was used during the resampling process. The inverse Fourier transform was then used to convert the magnitude and phase back into a time-domain signal.

4.5 Transformation

In order to perform the transformation, we require a set of Cepstral Coefficients for each frame of speech. These may either be predicted using the method of Chapter 3, or obtained directly from the target speech if the system is being used purely to do residual prediction. For each frame of speech, if the frame is voiced, then a residual is predicted on the basis of the Cepstral Coefficients of that frame. If it is unvoiced, then the source residual is used, though it is resampled to be of the correct length. It is acceptable to resample in this case, since the resampling process is being performed in the time domain rather than the complex frequency domain as was discussed earlier. Each frame of speech is resynthesised by filtering the residual using the appropriate LPC coefficients. Finally the speech is formed using the overlap and add method described in section 3.5.

4.6 Evaluation

4.6.1 Performance indices

In order to ascertain the relative effectiveness of the system depending on the parameter values used, it is necessary to have a method for measuring performance. The performance index used in the last chapter is not appropriate, since it measures only errors in

LSFs. The most common measure used in speech coding tasks is the signal-to-noise ratio (SNR). Therefore we have selected this for measuring the performance of our system. The signal to noise ratio is the ratio of the signal energy to the noise energy. Therefore

$$SNR(s(n), s_c(n)) = 10.log_{10} \frac{\sum |FFT(s(n))|^2}{\sum (|FFT(s_c(n))| - |FFT(s(n))|)^2} \tag{4.3}$$

gives the SNR on a dB scale, where $s(n)$ is the original speech, and $s_c(n)$ its coded form. The SNR of a whole utterance is computed by dividing the speech into a number of fixed length (20ms) frames, and then finding the average SNR of these frames, rather than simply finding the SNR of the whole utterance. A frame based approach better reflects the perceptual quality as errors in quiet and loud segments of the speech are computed separately. The error is computed on the magnitude spectrum, since this better reflects perceptual quality, as the human auditory system is not very sensitive to changes in phase. Higher SNR values indicate a better system.

4.6.2 Results

Figure 4.2 shows how the SNR of the system varies with the number of components (Q_{rp}) in the residual prediction GMM. It can be seen that the highest SNR values are obtained when $Q_{rp} = 64$. This is likely to be due to the fact that when fewer components are used, it is not possible for the system to fit to the data well enough, whereas when more components are used, the model is over-fitted to the data.

Figure 4.2 also shows how the SNR values change depending on whether LSF values are predicted or not. In the column 'residual prediction only', the values relate to an experiment where the target LSFs were used directly and only the residuals were predicted. In the column 'LSF and residual prediction' the LSFs were predicted (as described in the last chapter), and then these predicted LSFs were used for prediction of the associated residuals. It can be seen from figure 4.2 that the total error is higher when both the LSFs and residual are predicted. This is of course what one would expect.

The effect on the SNR of varying the amount of training data can be seen in table 4.3. As the amount of training data increases, the SNR also increases.

	SNR (dB)	
Q_{rp}	RP Only	LSF and RP
16	3.047	2.106
32	3.068	2.108
64	3.085	2.141
128	3.076	2.133
256	3.066	2.132

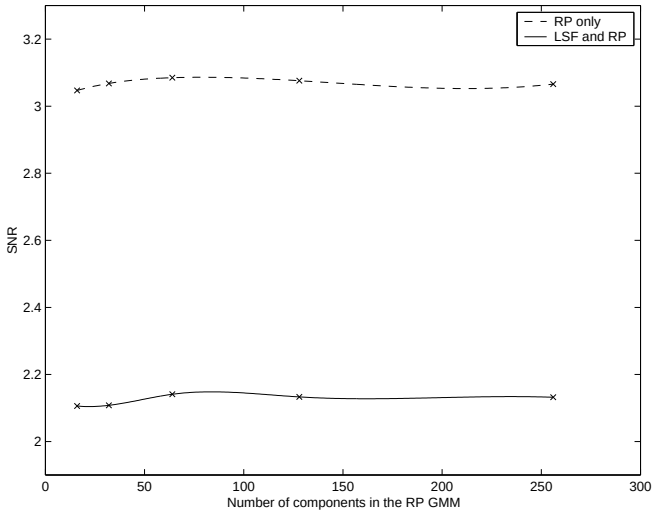

Figure 4.2: Graph showing the effect of changing the number of components in the residual prediction GMM, on the SNR in dB of the system. ($O_{LPC} = 20, Q_{LSF} = 12, T_{train} = 60s$.)

T_{train}	SNR (dB)
7.5	1.928
15	2.058
30	2.125
60	2.141
120	2.186

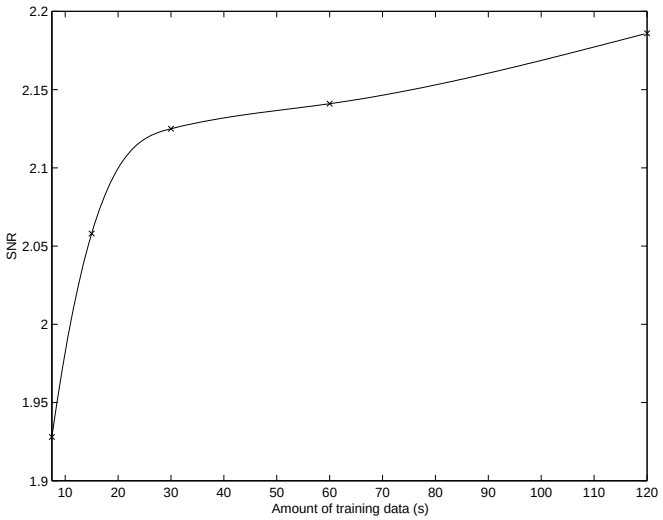

Figure 4.3: Graph showing the relationship between the amount of training data (T_{train}) and the SNR in dB of the resulting system. ($O_{LPC} = 20, Q_{LSF} = 12, Q_{rp} = 64$.)

4.7 Conclusion

We have presented a performance measure based on a signal to noise ratio for the magnitude spectrum of the speech. We have shown that a GMM with 64 components provides the highest SNR. The results also show that when residual prediction alone is performed a higher SNR is obtained (3.085 dB) than when full transformation is carried out (2.141 dB). These results are confirmed in informal listening tests, where it was found that when residual prediction alone is performed, the quality of the speech is extremely high, and it is quite hard to tell from the original speech. Example files may be found online (Gillett 2003). When LSF mapping and residual prediction are performed, the quality is also good and may easily be recognised as the target speaker. This perception of good transformation quality is reflected in figure 4.4 where it can be seen that the predicted waveform is very similar to the target waveform. However, there are buzzing and other artifacts, which are typically associated with RELP manipulation.

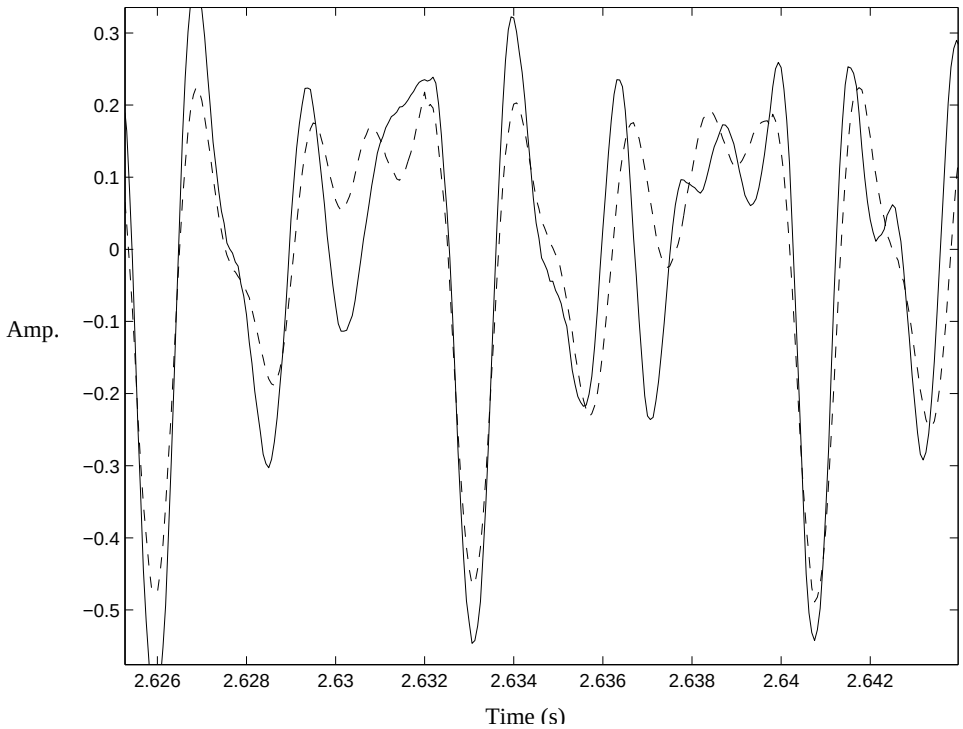

Figure 4.4: A predicted waveform overlaid over the target waveform. Both LSF prediction and residual prediction was performed. The solid line represents the target speech and the dashed line represents the predicted waveform.

Chapter 5

Transforming the F0 contour

5.1 Introduction

In this dissertation so far we have only addressed issues of voice quality. We now turn our attention to the problem of transforming F0 contours. In contrast to the other work on voice quality, there has been very little work in this area. The approach taken by all existing systems (Arslan & Talkin 1997, Arslan 1999, Stylianou et al. 1995, Toda, Ju, Saruwatari & Shikano 2000) is to simply normalise the F0 of the source speaker to be like that of the target. We will call this mapping function M_N, where

$$M_N(x) = ((x - \mu_{source})/\sigma_{source}) * \sigma_{target} + \mu_{target} \tag{5.1}$$

and μ_{source}, σ_{source} are the mean and standard deviation of the source speaker respectively, and μ_{target}, σ_{target} are the mean and standard deviation of the target speaker.

This mapping technique fails to capture many of the important features of F0 contours, which contain information about speaker identity. We present a method for the transformation of F0 contours from one speaker to another based on a small linguistically motivated parameter set. This was first presented in an earlier paper (Gillett 2002).

Training F0 contour generation models for speech synthesis requires a large corpus of speech (Black & Hunt 1996, Kochanski & Shih 2000). If it were possible to adapt the F0 contour of one speaker to sound more like that of another speaker, using a small, easily obtainable parameter set, this would be extremely valuable for speech synthesis.

5.2 Parameterisation

We use the parameterisation described by Patterson (2000), which was based on work by Ladd and Terken (1995). Patterson took F0 measurements at four selected target points in each sentence. These points were sentence-initial high (S), non-initial accent peaks (H), post-accent valleys (L), and sentence-final low (F). For each sentence there is one sentence-initial high, one sentence-final low and a varying number of peaks and valleys depending on the sentence. Patterson carried out analysis on approximately a minute of speech for each speaker. The values were collected into their respective categories and then averaged to get representative data for the speaker. Figure 5.1 shows diagramatically where the four points lie. The mean and standard deviation of the frequency of the voiced segments of speech for each speaker were also computed. In this work we make use of these values of S,H,L,F, mean and standard deviation collected by Patterson. All the following work using these parameters to carry out mappings is the work of the author and was not proposed by Patterson. We are simply using his data set and measurements of the parameters.

5.3 Mapping

The mapping from source to target F0 is then defined by a piecewise linear mapping, where one segment runs through the points (F_{source}, F_{target}) and (L_{source}, L_{target}), another between (L_{source}, L_{target}) and (H_{source}, H_{target}), and a final segment through (H_{source}, H_{target}) and (S_{source}, S_{target}). An example mapping is shown in figure 5.2, where one can see how a value x may be transformed to a value $M_{PL}(x)$. The mapping function M_{PL} is:

$$M_{PL}(x) = \begin{cases} F_{target} + \frac{(x - F_{source})(L_{target} - F_{target})}{(L_{source} - F_{source})} & \text{if } x < L_{source} \\ L_{target} + \frac{(x - L_{source})(H_{target} - L_{target})}{(H_{source} - L_{source})} & \text{if } L_{source} \leq x \leq H_{source} \\ H_{target} + \frac{(x - H_{source})(S_{target} - H_{target})}{(S_{source} - H_{source})} & \text{if } x > H_{source} \end{cases} \quad (5.2)$$

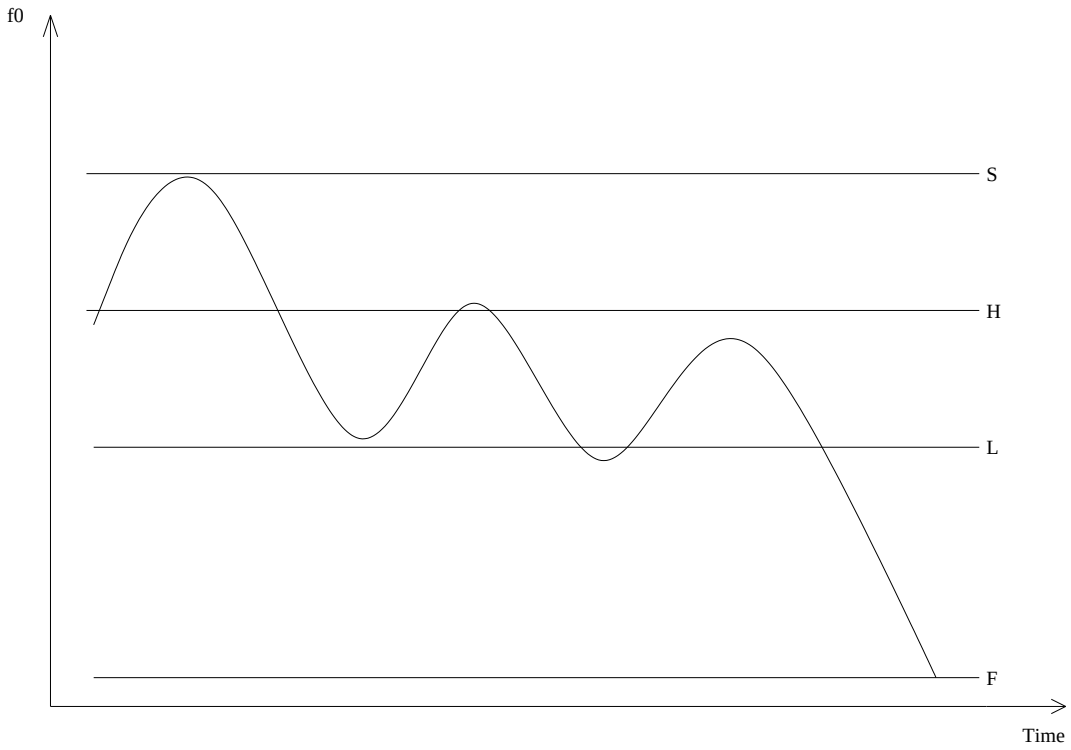

Figure 5.1: Measurement locations on an idealised speaker contour.

5.4 Transformation

Pitchmarks and F0 tracks are first found for the speech to be transformed, using the same tools as described earlier in section 2.5. The four parameters (S,H,L,F) were then obtained for both source and target speaker (from Patterson's thesis). These eight parameters were then used to define the mapping M_{PL}. Then for each voiced element of the F0 track, the F0 value was converted using M_{PL}. Finally, pitchmarks were generated from the transformed F0 track, and the speech was resynthesised using pitch synchronous overlap and add (PSOLA) (Gold & Morgan 2000).

5.5 Discussion

It is possible for the parameters of the mapping function to assume values such that that mapping function M_{PL} is in practice identical to the mapping M_N. There will also be cases where the mappings differ a great deal.

As can be seen in figure 5.3, the example target F0 contour is more closely matched by the method presented in this thesis, than by a method based on transforming the mean and standard deviation. In particular it can be seen that the F0 contour generated by the presented method more closely follows the sentence initial high of the target speaker. This may also be perceived from the associated waveforms (Gillett 2003). In addition, if one examines the distribution of the F0 values, the track transformed using the presented method has a distribution which more closely matches the distribution of the target speaker. This can be seen in figure 5.4.

These are simply examples of the result of the mapping function. In the next chapter a perceptual experiment will be presented to measure the effectiveness of the presented method.

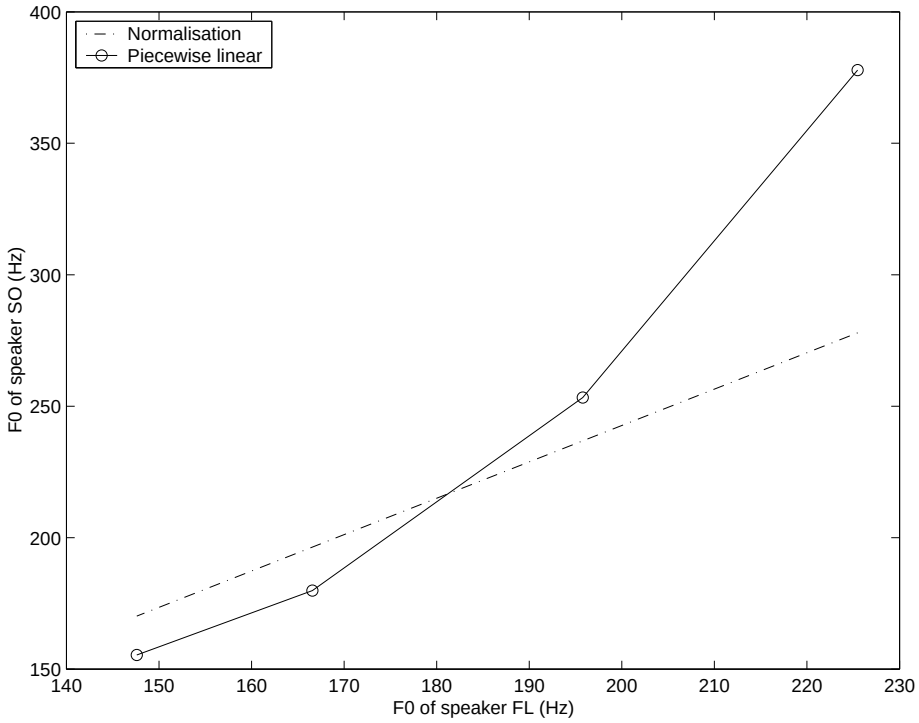

Figure 5.2: Female-female F0 map

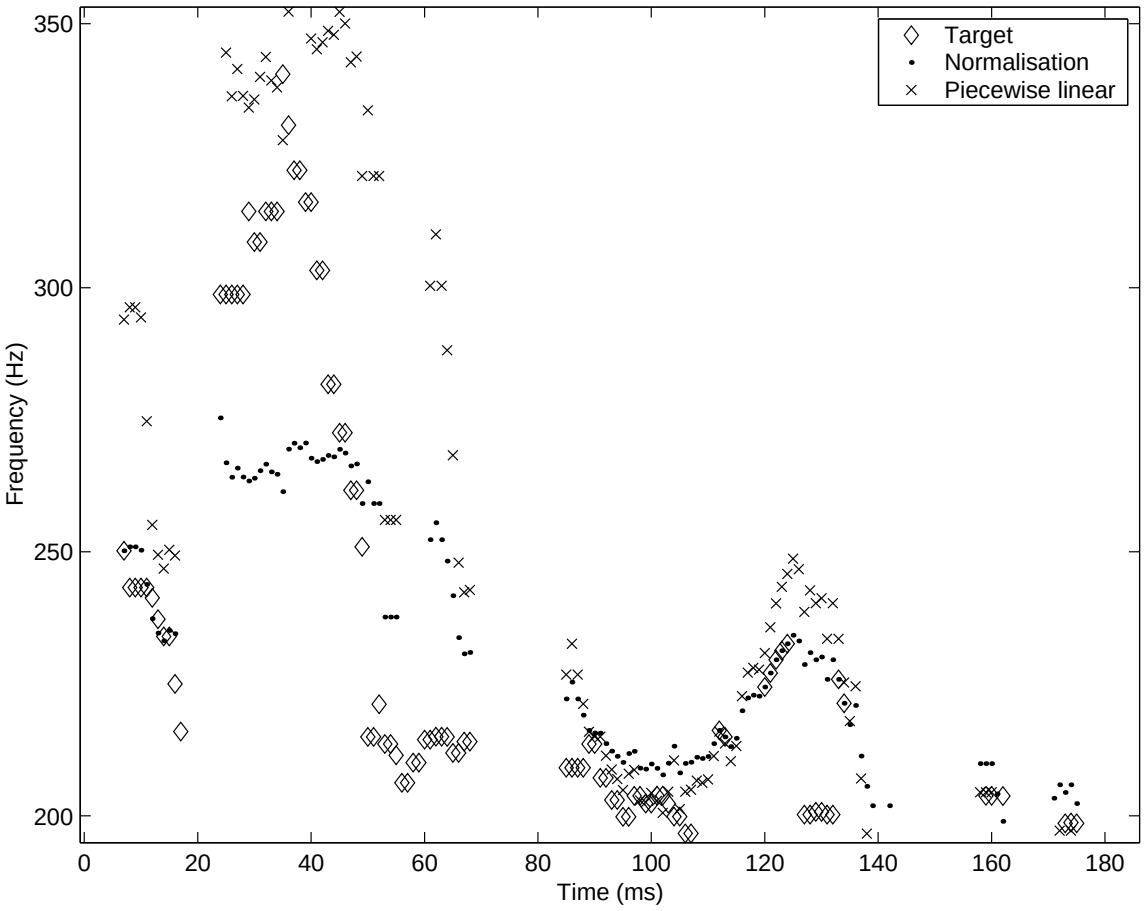

Figure 5.3: Target and mapped f0 tracks

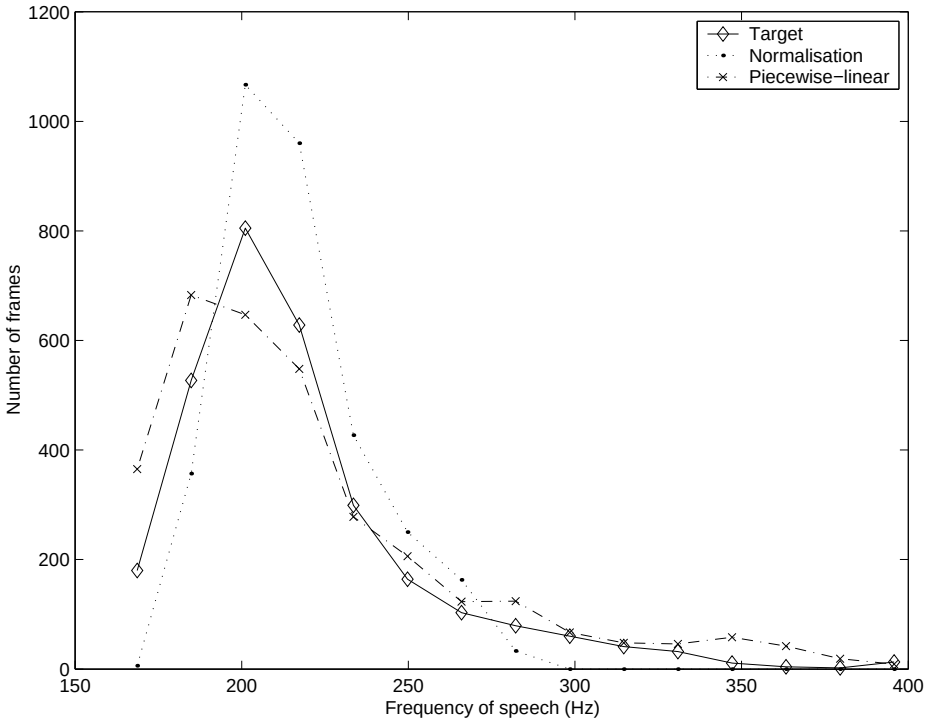

Figure 5.4: Histogram of frequencies for one minute of speech from Patterson corpus

Chapter 6

Evaluating the F0 transformation system

6.1 Introduction

The previous chapter demonstrated a new approach to the problem of transforming F0 contours from one speaker to another. We wish to ascertain if the proposed method is perceived as producing contours that are more similar to the F0 contours of the target speaker, than the existing technique. In order to this, we conducted a perceptual experiment.

6.2 Measuring the difference between techniques for given speaker pairs

In this experiment we will be investigating the relative effectiveness of the new method we presented in the last chapter (M_{PL}), against the existing technique based on the normalisation of mean and standard deviation (M_N). As previously discussed, the extent to which the results of the two methods differ is dependent on the particular parameters of the two speakers involved. For example, if the four points of the piecewise linear mapping ((F_{source}, F_{target}), (L_{source}, L_{target}), (L_{source}, L_{target}), (H_{source}, H_{target})), lie on the line defined by the mean and standard deviations (μ_{source}, σ_{source}, μ_{target},

σ_{target}), then the result of applying these two mappings will be identical. Figure 6.1 shows one mapping where the difference is large, and another where the difference is small.

The extent of any preference for one or other technique is likely to be proportional to the degree by which the two techniques differ for the speaker pair being tested. Therefore we have devised a method for determining how different the mappings are, for a particular speaker pair.

The difference between the two techniques for a given speaker pair is computed by taking the sum of the squares of the differences between the mapped frequencies generated by each of the two methods, at points corresponding to $S_{source}, H_{source}, L_{source}, F_{source}$. All frequencies are measured on an equivalent rectangular bandwidth (ERB) scale (Gold & Morgan 2000). This difference can be represented as:

$$\begin{aligned} D'(A,B) &= (M_N(S_{source}) - S_{target})^2 + (M_N(H_{source}) - H_{target})^2 \\ &\quad + (M_N(L_{source}) - L_{target})^2 + (M_N(F_{source}) - F_{target})^2 \end{aligned}$$

This difference measure is not symmetric. In other words, $D'(A,B) \neq D'(B,A)$. This is as expected, since the mapping function defined in the last chapter is non-linear. However, it is likely that there will be a high correlation between the two values. It is useful to have an overall distance measure between two speakers. We define this to be:

$$D(A,B) = \frac{D'(A,B) + D'(B,A)}{2} \tag{6.1}$$

6.3 Stimuli

The speech used in this experiment was recorded previously by Patterson (2000) as described in 2.4. Two sentences were selected from this corpus, chosen for their relatively short duration.

1) 'Madonna has been lined up as a key backer along with Ossie Kilkenny, the accountant to the stars.'

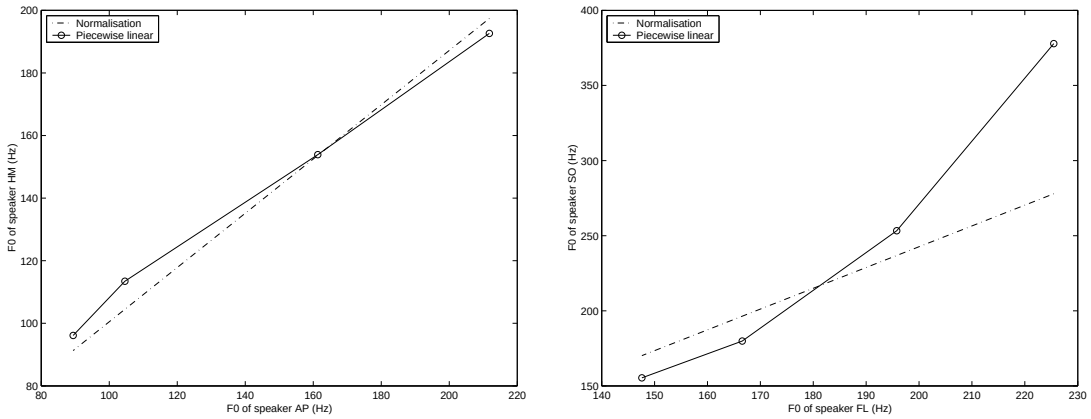

Figure 6.1: Graphs showing an example of a frequency mapping where the piecewise linear mapping is very similar to the normalisation mapping (left), and a frequency mapping where they are very different (right).

2) 'Kilkenny, whose clients include the rock band U2, will be employed as a consultant.'

Seven male and seven female speakers were selected, all of whom are native speakers of English, and have an accent commonly spoken by people from the Home Counties. Their ages range from 19 to 65.

For each same sex speaker pair (S, T), and for each of the two sentences, we created three stimuli. Firstly, the sentence uttered by source speaker S with its F0 modified to have the mean and standard deviation of the target speaker T. Secondly, the source speech with its F0 modified using the new method presented in this thesis. Finally, the source speech with the actual F0 contour of the target applied to it. This final sentence is the ideal output of an F0 transformation system. For all three stimuli types, although the F0 was modified, the voice quality was not. Therefore these stimuli have the voice quality of the source speaker with the intonation of the target speaker.

The experiment was of an XABX type, where X was the sentence spoken by the target speaker. A and B were the same sentence spoken by the source speaker modified to have an F0 contour like the target speaker, by one of the three methods described earlier. The decision to make the experiment XABX rather than ABX was based on the fact the utterances are relatively long, and in pilot experiments it was found that

	sm	gf2	rc	me	rl	vr	jb
sm	-	6+	1-	1	1-	3	4
gf2	-	-	7	8+	3	10+	8
rc	-	-	-	0	1	3	5+
me	-	-	-	-	1	2	4+
rl	-	-	-	-	-	2-	1-
vr	-	-	-	-	-	-	2
jb	-	-	-	-	-	-	-

Table 6.1: Table showing a measure of the difference (D) between the two F0 mapping techniques for English male speaker pairs. (Members of S_{same} are marked with '-' and members of $S_{\text{different}}$ are marked with '+').

playing the target twice helped the subjects decide which of A or B was better.

In order to ensure that any result we obtain is due to improvements in our method over the existing technique, we created two sets of stimuli. One of these groups consisted of those where we expect to get a clear preference for one or other method. These were the speaker pairs where the distance between the two methods, as defined in equation 6.2, was large. We call this group $S_{\text{different}}$. The other group was of speaker pairs where the two methods do not differ greatly. We call this group S_{same}. For each sex we selected five pairs where $D(S,T)$ is large, and five where it is small. We did not necessarily select the four largest or smallest values, since we were also trying to ensure that each speaker is chosen about the same number of times. In table 6.1 the distances between each male speaker pair is recorded, and table 6.2 contains similar information for the female speakers. These also indicate which speaker pairs were selected for the two groups.

In order to control for ordering effects, any given pair of stimuli was always presented both ways round to a given subject. Since there are three methods (A,B,C), there are 6 possible combinations as follows: XABX, XBAX, XCAX, XACX, XBCX, XCBX.

Since the concentration span of our subjects is limited, and we are most interested in the result concerning the destinction between the existing method (A), and our newly

	fl	so	nc	jk	jv	rs	mt
fl	-	22+	15	8	1	7-	21+
so	-	-	4-	10	22+	6	2
nc	-	-	-	2-	14+	1	1
jk	-	-	-	-	7	1-	6
jv	-	-	-	-	-	7	21+
rs	-	-	-	-	-	-	4-
mt	-	-	-	-	-	-	-

Table 6.2: Table showing a measure of the difference (D) between the two F0 mapping techniques for English male speaker pairs. (Members of S_{same} are marked with '-' and members of $S_{\text{different}}$ are marked with '+').

presented method (B), each subject is either presented with (XABX, XBAX, XCAX, XACX) or (XABX, XBAX, XCBX, XBCX). This is carried out such that there are always paired groups, so that all those subjects with odd subject numbers are presented with the first set, and all those with even numbers with the second set.

Since there are two groups of stimuli (one for each sex), and for each of these there are five pairs of speakers which must be presented both ways round, and there are two sentences, and four method combinations as described in the last paragraph, there are $2 * 10 * 2 * 2 * 4 = 320$ trials to be run. However, since each trial takes approximatly 30s, and we wish to restrict the experiment to not more than 35 minutes so that concentration is not impaired, we can have at most 70 trials per subject. Therefore, for each subject we select four speaker pairs, where half are from S_{same} and half from $S_{\text{different}}$, to give a total of 64 stimuli for each subject. The exact correspondence between subject numbers and which speaker pairs they listen to can be seen in tables 6.3 and 6.4.

6.4 Subjects

Twenty-five subjects were selected of whom approximately half were native and half non-native speakers of English. Similarly, approximately half were male and half were female. The task of discriminating between two similar F0 contours is difficult. Therefore

Source Speaker	Target Speaker	Sex	Group	Subject Numbers
gf2	vr	M	$S_{\text{different}}$	1,11,21
vr	gf2	M	$S_{\text{different}}$	2,12,22
me	jb	M	$S_{\text{different}}$	3,13,23
jb	me	M	$S_{\text{different}}$	4,14,24
sm	gf2	M	$S_{\text{different}}$	5,15,25
gf2	sm	M	$S_{\text{different}}$	6,16
rc	jb	M	$S_{\text{different}}$	7,17
jb	rc	M	$S_{\text{different}}$	8,18
gf2	me	M	$S_{\text{different}}$	9,19
me	gf2	M	$S_{\text{different}}$	10,20
rl	jb	M	S_{same}	1,11,21
jb	rl	M	S_{same}	2,12,22
sm	rc	M	S_{same}	3,13,23
rc	sm	M	S_{same}	4,14,24
rl	vr	M	S_{same}	5,15,25
vr	rl	M	S_{same}	6,16
rc	me	M	S_{same}	7,17
me	rc	M	S_{same}	8,18
sm	rl	M	S_{same}	9,19
rl	sm	M	S_{same}	10,20

Table 6.3: Table showing which male speaker pairs each subject listened to.

Source Speaker	Target Speaker	Sex	Group	Subject Numbers
fl	so	F	$S_{\text{different}}$	1,11,21
so	fl	F	$S_{\text{different}}$	2,12,22
jv	mt	F	$S_{\text{different}}$	3,13,23
mt	jv	F	$S_{\text{different}}$	4,14,24
fl	mt	F	$S_{\text{different}}$	5,15,25
mt	fl	F	$S_{\text{different}}$	6,16
nc	jv	F	$S_{\text{different}}$	7,17
jv	nc	F	$S_{\text{different}}$	8,18
so	jv	F	$S_{\text{different}}$	9,19
jv	so	F	$S_{\text{different}}$	10,20
jk	rs	F	S_{same}	1,11,21
rs	jk	F	S_{same}	2,12,22
so	nc	F	S_{same}	3,13,23
nc	so	F	S_{same}	4,14,24
nc	jk	F	S_{same}	5,15,25
jk	nc	F	S_{same}	6,16
rs	mt	F	S_{same}	7,17
mt	rs	F	S_{same}	8,18
fl	rs	F	S_{same}	9,19
rs	fl	F	S_{same}	10,20

Table 6.4: Table showing which female speaker pairs each subject listened to.

a relatively large number of subjects where selected, with a view to removing those who were not good at the task from the analysis.

6.5 Experiment

The E-Prime experiment design system was used for this experiment (Psy 2002). Prior to the experiment being run on the computer, the sex of the subjects together with their first language, and if non-native, an estimate of their ability at English was noted. Any academic background relating to linguistics was also recorded. The subjects were then placed in a quiet booth with headphones, computer monitor and an input box for recording responses. The subjects were given on-screen instructions regarding the procedure for inputing data. The following instructions were then given to the subjects:

> You will be presented with four pieces of speech. First you will be presented with a piece of target speech. Then two attempts of a speaker at imitating the F0 of the target speaker. Finally you will be presented with the target speech once more.
>
> (Target) target speech
> (1) first attempt by the imitator
> (2) second attempt by the imitator
> (Target) target speech
>
> You must decide whether attempt 1 or 2 has a more similar pitch pattern to the target. You shouldn't make your decision based on any aspects of the voice apart from pitch.
>
> If you think the first attempt sounds most like the target then press 1. If you think the second attempt sounds most like the target then press 2.
>
> In some cases it will be very hard to distinguish a difference between attempts 1 and 2. If so, just choose one or other.
>
> If you have any questions, please ask the experimenter now, otherwise press either button to continue.

The subjects were given three practice trials, followed by 70 actual trials. The experiment took approximately 35 minutes for each subject to complete. The order in which the stimuli were presented was randomized for each subject.

6.6 Results

For each subject, the number of times they selected the ideal contour in preference to a contour formed by either of the mapping techniques was counted. In order for the results to be meaningful, the subject must be capable of telling that the 'correct' F0 contour is better at representing the target speaker than a contour formed by either of the mapping techniques. Since the task is difficult, a relatively low level (60%) of preference for the ideal contour was selected as a criteria for rejecting subjects from further analysis. Just over half (13 of 25) of the subjects were able to tell that the ideal contour was better than mapped contours. It may seem suprising that such a high proportion of the subjects were not able to distinguish effectively, however making such judgements is difficult for naive listeners.

For the remaining set of subjects, the number of times the subject preferred contours mapped with M_{PL} over M_N was counted for each of the two data sets, $S_{\text{different}}$ and S_{same}. Table 6.5 shows the results of these calculations. No correlation between the nature of the subject (i.e. whether they were native speakers, their sex), and their preferences was found. Also, the sex of the speaker does not appear to make any difference to the preferences expressed.

The mean and standard deviation of each category was then computed. In order to establish the statistical significance of these results, we used Student's t-test for equal variances. A one tailed analysis was performed, since we are trying to determine the probability of a particular method being better than the other, rather than looking for a preference either way. The results of this analysis are contained in table 6.6. Students t-test provides a value of α, where α indicates the probability of the result being purely due to chance. A value of $\alpha \leq 0.01$ is generally accepted as being a statistically significant result. It is therefore clear from table 6.6 that the preference for M_{PL} over M_N for $S_{\text{different}}$ is highly significant. Similarly the preference for the target contour over the

	Preference for M_{PL} (%)		
Subject	on $S_{\text{different}}$	on S_{same}	for target
1	72	56	72
2	78	56	86
3	61	56	81
4	56	44	69
5	83	56	89
6	56	44	63
7	67	50	72
8	61	56	64
9	56	75	72
10	61	56	67
11	83	50	78
12	67	50	78
13	72	50	61

Table 6.5: Table showing the individual subject preferences for different mapping methods.

	Mean (%)	Std. Dev.	α	t
Preference for M_{PL} over M_N for $S_{\text{different}}$	67	10	< 0.0000001	-8.711448
Preference for M_{PL} over M_N for S_{same}	54	8	~ 0.02	-2.488684
Preference for target over mapped contours	73	9	< 0.00000000001	-13.805951

Table 6.6: Table showing the subject preferences for different mapping methods.

mapped contours is also significant. However, the significance of the preference for M_{PL} over M_N for S_{same} is not very high, as is to be expected, since on the data set S_{same}, the two methods (M_{PL} and M_N), are almost identical (see section 6.2).

6.7 Conclusion

It was found that 73% of the time subjects expressed a preference for the ideal contour over a mapped contour. The remaining 26% of the time the subject chose the mapped contour, this is likely to be due to the fact that the contours were so similar that the subject was not able to distinguish between them. A clear preference for our method is shown in the experiment, with subjects selecting the speech modified with the presented mapping (M_{PL}) in preference to M_N for the dataset where the two methods are most different 67% of the time. This result compares very favourably with the preference for the ideal contour of 73%, suggesting that using M_{PL} is almost good as using the actual contour. In the cases where the mapping techniques differ least, there was a preference for M_{PL}, although it is on the border of not being statistically significant.

It has been clearly shown that the presented method based on a piecewise-linear mapping is at least as good as the only existing technique for F0 contour mapping for *all* speaker pairs, and that in *many cases* it is much better and almost as good as using the target F0 contour.

Chapter 7

Conclusion

7.1 Summary

In this thesis we have tackled two of the major tasks necessary to produce an effective voice transformation system. The voice quality transformation component of our system has two main parts corresponding to the two components of the source-filter model. The first component transforms the spectral envelope as represented by a linear prediction model. The transformation was achieved using a Gaussian mixture model, which was trained on aligned speech from source and target speakers. Using Kain's LSF conversion performance measure (Kain 2001), our system achieves a value of $P_{LSF} = 0.36$, whereas Kain's system which is representative of the best existing sytems achieves a value of only $P_{LSF} = 0.31$ on a similar amount of training data.

The second component of the voice quality conversion system predicts the spectral detail from the transformed LSFs. In the training phase a Gaussian mixture model is used to cluster the space of all voiced LSFs. Residual phases and magnitudes are stored in codeword tables for each component. When performing prediction, the Gaussian mixture model and codeword tables are used to predict a residual for each frame of LSFs. We also made measurements of a spectral magnitude domain signal to noise ratio measure. The results show that when residual prediction alone is performed a higher SNR is obtained (3.085) than when full transformation is carried out (2.141).

We presented a new method for the transformation of F0 contours from one speaker

to another based on a small linguistically motivated parameter set. Mean sentence initial highs, sentence medial highs, sentence medial lows, and sentence final lows were found for the source and target speakers. These eight parameters then define a three segment piecewise-linear mapping (M_{PL}).

A perceptual experiment was conducted, to ascertain how well the mapping performs relative to the standard approach based on normalisation of mean and standard deviation (M_N). A clear preference for our method is shown in the experiment, with subjects selecting the speech modified with the mapping M_{PL} in preference to M_N for the dataset where the two methods are most different 67% of the time. This result compares very favourably with the preference for the ideal contour of 73%.

7.2 Conclusion

The thesis advances the state of the art in a number of key respects. Methods for the rejection of poorly matched data have been described which enhance the quality of voice quality transformation. This is particularly useful for transforming natural speech where there is differing pronunciation, dialects and disfluencies. Our system also uses a lowpass filter to smooth the LSF tragectories, which significantly improves performance. Due to these advances, our LSF transformation system outperforms existing techniques.

We have also presented a new method for transformation of spectral detail from one speaker to another, which produces high quality results. The system presented is capable of transforming utterances given a small amount of speech from two speakers, where the speech is naturally spoken and prosodically varied. The transformed speech can be easily recognized, however there are significant signal processing artifacts introduced.

The second area in which this thesis offers a major contribution, is in F0 transformation. The perceptual experiment clearly demonstrates that our system is at least as good as the only existing technique for F0 contour mapping for *all* speaker pairs, and that in *many cases* it performs much better and is almost as good as using the target F0 contour.

7.3 Future Work

The voice quality transformation system does produce output with noticeable signal-processing artifacts. Further work must be done to reduce these artifacts. Improvements in the residual prediction module are likely to yield the most noticeable improvements.

The work on F0 transformation makes use of a number of parameters that were extracted by hand. However, for this approach to be useful, methods must be developed which extract these parameters automatically. The problem of finding these parameters is likely to be much easier than finding the pitch accents in a sentence. In order to find the sentence intial high, one may simply find the highest F0 in the first one second of speech, and similarly one may find the sentence final low by finding the minimum of the last second of speech for the sentence. To find sentence medial highs and lows, an approach based on finding maxima and minima in a smoothed F0 contour may well produce good results.

Appendix A

Perceptual Experiment Materials

A.1 Screen 1

Welcome to the experiment.

Input in this experiment will be given using the white box with five buttons which is on the desk in front of you. Button 1 is the left-most button, and button 2 is the button adjacent to it.

A.2 Screen 2

You will be presented with four pieces of speech. First you will be presented with a piece of target speech. Then two attempts of a speaker at imitating the pitch of the target speaker. Finally you will be presented with the target speech once more.

(1) first attempt by the imitator
(2) second attempt by the imitator
(Target) target speech

A.3 Screen 3

You must decide whether attempt 1 or 2 has a more similar pitch pattern to the target. You shouldn't make your decision based on any aspects of the voice apart from pitch. If you think the first attempt sounds most like the target then press 1. If you think the second attempt sounds most like the target then press 2. In some cases it will be very hard to distinguish a difference between attempts 1 and 2. If so, just choose one or other. If you have any questions, please ask the experimenter now, otherwise press either button to continue.

A.4 Screen 4

Press button '1' if you think the first attempt was closest to the target.
Press button '2' if you think the second attempt was closest to the target.

References

Abe, M., Nakamura, S., Shikano, K. & Kuwabara, H. (1988), Voice conversion through vector quatization, *in* 'Proceedings of the International Conference on Acoustics, Speech and Signal Processing', IEEE, pp. 655–658.

Ahn, R. & Holmes, W. H. (1997), An accurate pitch detection method for speech using harmonic-plus-noise decomposition, *in* 'Proceedings of the International Congress of Speech Processing', pp. 55–59.

Arslan, L. (1999), 'Speaker transformation algorithm using segmental codebook', *Speech Communication Journal* .

Arslan, L. M. & Talkin, D. (1997), Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum, *in* 'Proc. Eurospeech '97', pp. 1347–1350.

Bailly, G. (2001), *Improvements in Speech Synthesis*, J. Wiley and Sons Ltd, chapter 1, pp. 22–38.

Bailly, G., Bernard, E. & Coisnon, P. (1998), 'Sinusoidal modelling'.

Baudoin, G. & Stylianou, Y. (1996), On the transformation of the speech spectrum for voice conversion, *in* 'Proceedings of the International Conference on Spoken Language Processing', pp. 1405–1408.

Black, A. & Hunt, A. (1996), Generating f0 contours from tobi labels using linear regression, *in* 'Proceedings of the International Conference on Speech and Language Processing'.

Brookes, M. (1998), *VOICEBOX : Speech Processing Toolbox for MATLAB*, Department of Electrical and Electronic Engineering, Imperial College. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

Clark, R. (1999), Using prosodic structure to improve pitch range variation in text to speech synthesis, *in* 'Proceedings of the International Congress of Phonetic Sciences'.

Dutoit, T. & Leich, H. (1993), 'Mbr-psola : Text-to-speech synthesis based on an mbe re-synthesis of the segments', *Speech Communication* **13**.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J., Pallett, D. S. & Dahlgreen, N. L. (1990), Darpa timit acoustic-phonetic continuous speech corpus, Technical report, National Institute of Standards and Technology.

Ghahramani, Z. & Jordan, M. I. (1994), *Advances in Neural Information Processing Systems*, Morgan Kaufmann, San Mateo, California, chapter Supervised learning from incomplete data via an EM approach, pp. 120–127.

Gillett, B. (2002), Transforming pitch contours. Presented to Cambridge Intonation Discussion Group.

Gillett, B. (2003), 'Audio examples to accompany 'transforming voice quality and intonation", http://www.cstr.ed.ac.uk/ beng.

Gold, B. & Morgan, N. (2000), *Speech and Audio Signal Processing*, Processing and Perception of Speech and Music, John Wiley and Sons, Inc.

Kain, A. (2001), High Resolution Voice Transformation, PhD thesis, OGI School of Science and Engineering.

Kain, A. & Macon, M. (1998), Spectral voice conversion for text-to-speech synthesis., *in* 'Proceedings of ICASSP '98'.

King, S. (2002), Lecture notes.

Kochanski, G. & Shih, C. (2000), Prosody modeling with soft templates, *in* 'Proceedings of the International Conference on Speech and Language Processing'.

Ladd, B. & Terken, J. (1995), Modelling intras- and inter-speaker pitch range variation, *in* 'Proceedings of the International Conference of Phonetic Sciences', Vol. 2, pp. 386–389.

Ladefoged, P. (1962), *Elements of Acoustic Phonetics*, University of Chicago Press, chapter 5.

Ladefoged, P. & Ladefoged, J. (1980), The ability of listeners to identify voices., *in* 'UCLA Working Papers in Phonetics', Vol. 49, UCLA.

Lancker, D. V., Kreiman, J. & Emmorey, K. (1985), 'Familiar voice recognition: patters and parameters. part 1: Recognition of backwards voices.', *Journal of Phonetics* **13**.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the 5th Symposium of Mathematics and Probability'.

Makhoul, J. (1975), Linear prediction: A tutorial review, *in* 'Proceedings of the IEEE', Vol. 63, pp. 561–580.

Matsumoto, H., Hiki, S., Sone, T. & Nimura, T. (1973), 'Multidimensional representation of personal quality of vowels and its acoustical correlates', *IEEE Transactions on Audio and Electroacoustics* **21**.

McAulay, R. & Quatieri, T. (1995), *Speech Coding And Synthesis*, chapter Sinusoidal Coding.

Narendranath, M., Murthy, H. A., Rajendran, S. & Yegnanrayana, B. (1995), 'Transformation of formants for voice conversion using artificial neural networks', *Speech Communication* **16**(2), 207–216.

Necioglu, Burhan, F., Clements, M., Barnwell, T. & Schmidt-Nielson, A. (1998), Perceptual relevance of objectively measured descriptors for speaker characterization, *in* 'Proceedings of ICASSP'.

Ostendorf, M., Price, P. J. & Shattuck-Hufnagel, S. (1995), The boston university radio news corpus, Technical report, Boston University, SRI International, MIT.

Patterson, D. (2000), A Linguistic Approach to Pitch Range Modelling, PhD thesis, University of Edinburgh.

Psy (2002), *E-Prime Users Guide*. http://www.pstnet.com/e-prime/default.htm.

Quatieri, T. & McAulay, R. (1992), 'Shape invariant time-scale and pitch modifications of speech', *IEEE Transactions on Signal Processing* **40**.

Rabiner, L. R. & Schafer, R. W. (1978), *Digital Processing of Speech Signals*, Signal Processing Series, Prentice-Hall.

Sakoe, H. & Chiba, S. (1978), 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Transactions Acoustics, Speech, and Signal Processing* **ASSP-26**.

Schriberg, E., Ladd, D. R., Terken, J. & Stolcke, A. (1996), Modeling pitch range variation within and across speakers: predicting f0 targets when "speaking up", *in* 'Proceedings of the International Conference on Spoken Language Processing'.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992), Tobi: A standard for labeling english prosody, *in* 'Proceedings of the International Conference on Spoken Language Processing', pp. 867–870.

Stylianou, Y., Cappe, O. & Moulines, E. (1995), Statistical methods for voice quality transformation, *in* 'Proc. EUROSPEECH, 1995'.

Taylor, P., Caley, R., Black, A. W. & King, S. (1999), *Edinburgh Speech Tools Library : System Documentation*, Centre for Speech Technology, University of Edinburgh.

Toda, T., Ju, J., Saruwatari, H. & Shikano, K. (2000), Straight-based voice conversion algorithm based on gaussian mixture models, *in* 'Proceedings of the International Conference on Speech and Language Processing'.

Zetterholm, E. (2000), The significance of phonetics in voice imitations, *in* 'Proceedings of SST', pp. 342–347.