Probability Models of Formant Parameters for Voice Conversion

Dimitrios Rentzos  Saeed Vaseghi  Qin Yan  Ching-Hsiang Ho*  Emir Turajlic

Department of Electronics and Computer Engineering Brunel University, Middlesex UB8 3PH, UK
*Fortune Institute of Technology, Kaohsiung, Taiwan, 842, R.O.C.
(Dimitrios.Rentzos, Saeed.vaseghi, Qin.Yan, Emir.turajlic)@brunel.ac.uk,  ch.ho@center.fjtc.edu.tw

Abstract
This paper explores the estimation and mapping of probability models of formant parameter vectors for voice conversion. The formant parameter vectors consist of the frequency, bandwidth and intensity of resonance at formants. Formant parameters are derived from the coefficients of a linear prediction (LP) model of speech. The formant distributions are modelled with phoneme-dependent two-dimensional hidden Markov models with state Gaussian mixture densities. The HMMs are subsequently used for re-estimation of the formant trajectories of speech. Two alternative methods are explored for voice morphing. The first is a non-uniform frequency warping method and the second is based on spectral mapping via rotation of the formant vectors of the source towards those of the target. Both methods transform all formant parameters (Frequency, Bandwidth and Intensity). In addition, the factors that affect the selection of the warping ratios for the mapping function are presented. Experimental evaluation of voice morphing examples is presented.

1. Introduction
Voice conversion has applications in all voice output systems such as text to speech synthesis, voice editing, broadcast systems, Internet, toys, robots, Karaoke and entertainment applications.

An effective voice conversion system requires two essential components: (a) accurate probability models of the distributions of sets of feature vectors that characterise the source and the target speakers’ voice and (b) an effective signal processing system for mapping the source speaker’s voice to the target speaker’s voice. There are two broad approaches to voice conversion: (a) non-parametric mapping of the spectral vectors of a source speaker to those of a target speaker using a source-to-target spectral codebook [3,4], and (b) parametric (mainly LP) model-based methods [5] of mapping via the modification of the source model parameters towards the estimates of the target model parameters. Parametric modelling allows a more flexible and selective modification of spectral parameters of the vocal tract and also allows modification of the glottal and prosodic parameters. In this paper we consider some of the practical issues in the parametric modelling and mapping of formants in the context of voice conversion. This paper is organised as follows: in section 2 formant estimation is described. Section 3 considers and compares two different methods of parametric mapping of formant features. Section 4 describes experimental results and section 5 concludes the paper.

2. Formant Probability Modelling and Trajectory Estimation
To achieve formant-based voice mapping, an accurate formant model estimation is required to deal with the problems of the variability of the number of formants across the phonemes and the merging and de-merging of neighbouring formants (such as F2 and F3) over time. The problems can be alleviated using a hidden Markov model (HMM) based formant estimation procedure [1,6]. The formant candidate vectors, with unknown labels, are derived from the poles of an LP model of speech. The pole (formant candidate) features are frequency, bandwidth, delta frequency, delta bandwidth and intensity. A 2-D HMM with N left-to-right states distributed across frequency, and M states distributed across time, is used to classify the formant observations as shown in figure (1). The distribution of formants within each HMM state may be modelled by a mixture Gaussian distribution as

\[ p(F_k) = \sum_{k=1}^{M} c_k N(F_k, \mu_{F_k}, \sigma_{F_k}) \]  

(1)

where \( N(F_k, \mu_{F_k}, \sigma_{F_k}) \) is a Gaussian model of Formant with mean \( \mu_{F_k} \) and variance \( \sigma_{F_k} \).

To illustrate the accuracy with which the distributions of the formants are modelled by Gaussian mixture probability

Figure 1: A 2-D HMM formant tracking model.
models of HMM states, the plots of histograms of formants were superimposed on the plots of the distributions of the mixture Gaussians of HMM states as shown in figure (2). The raw experimental data used are the frequencies of the poles of the LP model for all training data segments of each phoneme of a speaker. In figure (2) the peaks of the histogram and mixture Gaussian distribution curves correspond to the formants. Figure (2) illustrates, that as expected, the HMMs of formants closely fit the histograms of the distribution of formants.

Given the phoneme-dependent HMM formant classifier and the associated formant bandwidths, formant contour estimation for a speech waveform is achieved through minimisation of a weighted mean square error objective function as

$$\hat{f}_k(t) = \min \sum \frac{(f_k(t) - \hat{f}_k(t))^2}{BW^2} w_k(t)$$  \hspace{1cm} (2)

where $L_k(t)$ is the total number of poles in the $k$th speech frame classified as formant $k$ and $w_k(t)$ is a probabilistic weight derived from the model.

3. Parametric Formant Mapping

In this section two alternative methods for voice conversion via formant mapping are considered: (a) non-uniform spectral mapping and (b) LP frequency response mapping through rotation of the poles of LP models.

3.1 Spectrum Mapping Through Frequency Warping

The equation for the transformation of the spectrum of a source speaker to a target speaker is expressed as

$$Y(f, t) = \gamma(f, t)X[W(f, t)T]$$  \hspace{1cm} (3)

where $X$, $Y$, $f$, and $t$ denote the source spectrum, the transformed spectrum, and the time and frequency variables respectively. The method for warping the frequency axis is based on a formant-dependent function for rescaling the frequency axis in order to achieve the desired change in the frequency and bandwidth of the formants. The composite frequency warping function $W(f,t)$ includes the mapping functions for both the formant frequency and bandwidth. The spectral amplitude shaping function $\gamma(f, t)$ is used to map the energy intensities between the source and target formant frequencies.

In order to derive the parametric mapping functions $W(f,t)$ and $\gamma(f, t)$ the source to target formant parameter vector conversion ratios are calculated; namely the formant frequency ratio vector $\alpha(i,t)$, the formant bandwidth ratio vector $\beta(i,t)$ and formant intensity ratio vector $\gamma(i, t)$. The ratio $\alpha(i,t)$ is calculated from the ratio of frequency differences between two successive formants of the target speaker, $F_{i+1,t}^T$, to that of the source speaker, $F_{i+1,t}$, as,

$$\alpha(i,t) = \frac{F_{i+1,t}^T}{F_{i+1,t}} = \frac{F_{i+1,t}^T - F_{i,t}^T}{F_{i+1,t}^S - F_{i,t}^S} \hspace{1cm} i = 1, \cdots, N$$  \hspace{1cm} (4)

where $F_{i,t}^T$ and $F_{i,t}^S$ are the formant frequency values for the target and source respectively, $i$ denotes the formant index and $N$ is the total number of formants. The conversion ratios for bandwidth of resonance at formants, $\beta(i,t)$, are obtained as

$$\beta(i,t) = \frac{BW_{i,t}^T}{BW_{i,t}^S} \hspace{1cm} i = 1, \cdots, N$$  \hspace{1cm} (5)

where $BW_{i,t}^S$ and $BW_{i,t}^T$ denote the source and target bandwidths for formant $i$. The conversion ratios for intensity of resonance at formants are derived in a similar way to the frequency warping ratios in equation 5. To apply the formant parameter conversion vectors to an LP spectrum with $K$ spectral samples, the discrete warping ratios at the formants are interpolated to obtain $K$ spectral values. The interpolation of frequency and intensity functions can be performed using one of the widely used polynomial interpolation methods such as the Hermite or

![Figure 2: Formant histograms and models for male speaker models of HMM states, the plots of histograms of formants were superimposed on the plots of the distributions of the mixture Gaussians of HMM states as shown in figure (2).](image)

![Figure 3: Spectrum Mapping Procedure](image)
where \( \varphi(p_i) \), \( \varphi(p'_i) \) are the angles of the original and mapped poles, and \( \alpha(i,t) \), \( \beta(i,t) \) are the frequency and bandwidth mapping ratios (equations 3,4).

### 3.3 Selection of Mapping Parameter Resolution

The voice conversion ratios for mapping the parameters of a source speaker towards the parameters of a target speaker can be made frame-dependent, phoneme-dependent or just one constant set of average mapping values may be used for a pair of source and target speakers.

**Average conversion vector:** One set of average parameter vector conversion ratios are obtained over all vowels and used for mapping the source speech to target voice.

**Phoneme-based conversion vector:** A set of Phoneme-dependent transformation ratios are obtained and applied to each cluster of triphones. This method required the phonemic labelling/segmentation of source speech prior to transformation.

**Frame-based conversion vector:** The source and target speech frames are time-aligned and a set of transformation ratios is obtained for each frame.

### 4. Experiments

Experiments were performed to train formant models of speakers and to determine the effectiveness of the mapping methods in converting the source speech to target speech. Three speaker-dependent databases, collected from three American English speakers, two males and one female were used as a case study. The databases consist of ten minutes of speech per speaker with a sampling rate of 10 kHz. An LP order of 13 is used and the LP coefficients are estimated every 10 ms using a 25 ms signal window.

The first set of experiments obtained phoneme-dependent formant models for the source and target speakers, the models are used to obtain the best mapping function between the source and target speakers. Figure 5 shows a comparison of the formant models between two male speakers (M1/M2) and a male and a female (M1/F). It is interesting to note that for the male-to-male case the differences in frequencies are not large. This is expected since the two male speakers are of the same sex, accent and have a similar age. Thus for males or females of a similar age and physique the most important formant characteristics that distinguish two voices are the bandwidths and intensities of formants which can differ significantly as can also be shown in table 1. In the female-to-male example the formant frequency difference is important (about 12%). In the male-to-male voice conversion often the bandwidth and in particular the

### Table 1: Frequency (F), bandwidth (BW) and intensity (I) differences for formant 2 of phoneme /æ/

<table>
<thead>
<tr>
<th></th>
<th>M1/M2</th>
<th>M1/F</th>
</tr>
</thead>
<tbody>
<tr>
<td>F (%)</td>
<td>4.84</td>
<td>12.12</td>
</tr>
<tr>
<td>BW (%)</td>
<td>49</td>
<td>24</td>
</tr>
<tr>
<td>I (dB)</td>
<td>-11.28</td>
<td>-10.75</td>
</tr>
</tbody>
</table>
spectral intensity differences are most significant correlates of voice characteristics.

To evaluate the effectiveness of the phoneme-dependent mapping ratios, the average formant trajectories for phonemes of the source and target speakers were extracted (figure 6). Speech was mapped using both the spectrum warping and pole rotating methods. Two mapping examples can be seen in figure 7. Significant voice modifications can be achieved as in the first plot where the source has one less formant than the target which is “reproduced”. Pole rotating is a more flexible method in which theoretically any change in spectral shape can be performed. On the other hand its ability to produce large modifications in voice spectrum also makes it a less robust method than frequency warping. Experiments using all three conversion ratios resolutions (constant parameters, phoneme-based and frame based) were performed and it was established that overall a smoothed phoneme-based method produced preferred results.

5. Conclusion

The problems of modelling and mapping of the formants of a source speaker to a target speaker were explored. It is shown that he distributions of the formants are accurately modelled with 2D-HMMs of formants. A main focus of the paper was presentation and evaluation of different methods for formant mapping in voice conversion. A spectrum warping and a pole rotation method were evaluated and it was established that both methods are effective. Different mapping resolutions were evaluated, and the phoneme based mapping was found to produce the best results.

Acknowledgement We acknowledge the support of UK’s Engineering and physical sciences research council Grant No GR/M98036/01.

6. References