EVALUATION OF METHODS FOR PARAMETERIC FORMANT TRANSFORMATION IN VOICE CONVERSION

Emir Turajlic Dimitrios Rentzos Saeed Vaseghi Ching-Hsiang Ho*

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

ABSTRACT

This paper explores methods of estimation and mapping of parametric formant-based models for voice transformation. The main focus is the transformation of the parameters of a model of the vocal tract of a source speaker to a target speaker. The vocal tract parameters are represented with the linear prediction (LP) model coefficients and the associated formant frequencies, bandwidths, intensities and their temporal trajectories. Two methods are explored for vocal tract (formant) mapping. The first method is based on non-uniform frequency warping and the second is based on pole rotation. Both methods transform all parameters of the formants (frequency, bandwidth and intensity). In addition, the factors that affect the selection of the warping ratios for the mapping functions are presented. Experimental evaluation of voice morphing based on parametric models are presented.

1. INTRODUCTION

Voice conversion has applications in all voice output systems such as text to speech synthesis, voice editing, Karaoke, broadcasting and Internet voice applications.
An effective voice conversion system would need two essential components: (a) accurate models of the source and the target speakers' voice characteristics and (b) an effective signal processing method 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 (LPC) model-based methods [5] of mapping through 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 compares two different methods of parametric mapping of formant features. Section 4 describes experimental results and section 5 concludes the paper.

2. FORMANT TRAJECTORY ESTIMATION

To perform a formant-based spectrum mapping, an accurate formant model estimation is needed 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 by using a hidden Markov model (HMM) based formant estimation procedure [1,6], where an LP-analysis is performed on speech, and the poles of the LP model are converted into candidate formant features. The pole 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 2.
Once the formant models for phonemes are available they are used to estimate the formant trajectories as illustrated in figure 1 and described in this paper.

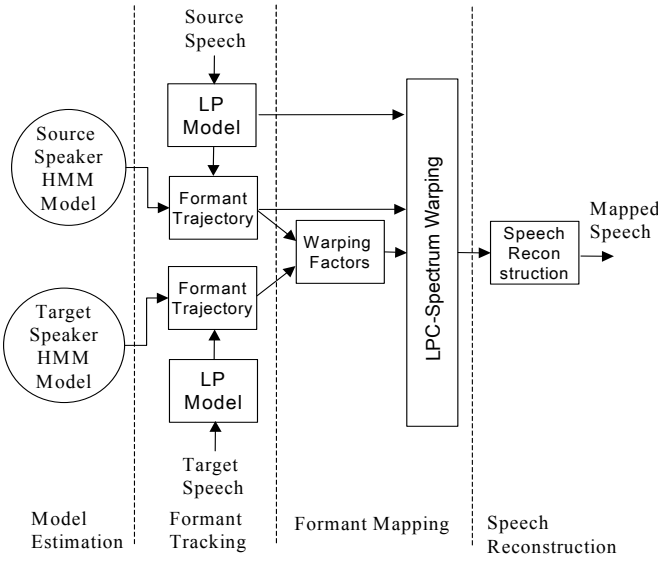

Figure 1: Spectrum Mapping Procedure

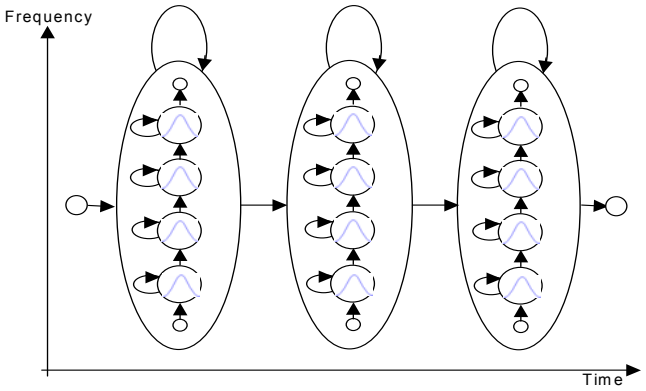

Figure 2: A 2-D HMM formant tracking model.

3. PARAMETRIC FORMANT MAPPING

Two methods for formant mapping are considered: (a) non-uniform spectral mapping and (b) LP frequency response mapping through rotation of the poles LP models. The proposed methods are all formant based in that the frequency axis is divided into N bands with each band centred on one formant. Both methods make use of the information derived from the formant estimation. In the spectrum warping method the formant estimation is used to divide the spectrum into subbands according to the formant positions as shown in figure(3), and in the pole rotation method, the poles are associated with formants. The inputs to the formant mapping block are the *LP-spectrum* and the *formant feature vector* of the current frame.

3.1 Spectrum Mapping Through Frequency Warping

The equation for spectrum mapping is expressed as

$$Y(f,t) = \gamma(t,f)X[A(f,t)f] \qquad (1)$$

where X, Y, t, and f denote the source spectrum, the transformed spectrum, the time and frequency axis respectively. The composite *frequency warping* function $A(t,f)$ includes mapping for both the formant frequency and bandwidth and $\gamma(t,f)$ is the *intensity shaping* function.

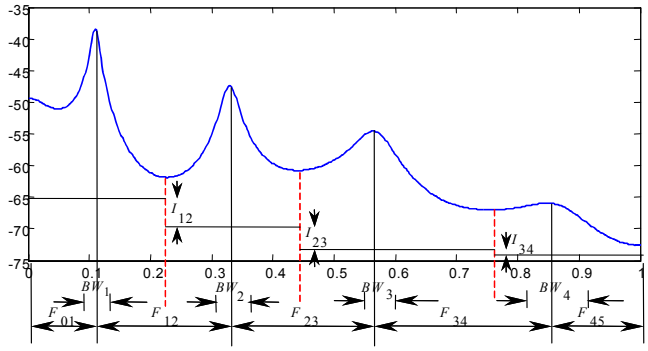

Figure 3: Formant-based spectrum warping

Warping frequency axis is based on a formant-dependent function for rescaling the frequency axis to achieve the desired change in the frequency and bandwidth of the formants. One set of factors controls the shift of the formant frequencies and another the modification of the formant bandwidths. To calculate the frequency warping function $\alpha(f,t)$ for the frequency axis at time t, the frequency differences between two successive formants of source and target speakers are used, as

$$a(f,t) = \frac{F^T_{f,f+1,t}}{F^S_{f,f+1,t}} = \frac{F^T_{f+1,t} - F^T_{f,t}}{F^S_{f+1,t} - F^S_{f,t}} \qquad (3)$$

where F^T and F^S are the formant values for the target and source respectively.

Bandwidth warping function $\beta(f,t)$ is obtained from interpolation of the target to source formant bandwidth ratios

$$\beta(f,t) = Interpolation\left[BW^T_{k,t} \,/\, BW^S_{k,t}\right] \quad k = 1,..5 \qquad (4)$$

The frequency and bandwidth warping functions are combined to form

$$A(f,t) = \alpha(f,t) * \beta(f,t) \qquad (5)$$

A method for transformation of the bandwidth parameters is illustrated in figure 3.

Spectral Amplitude Shaping function $\gamma(f,t)$ is used to map the energy intensities between the source and target formant frequencies. The values in between formants are linearly interpolated and applied to re-shape source spectrum. The intensity coefficients are derived in a similar way to the frequency warping coefficients in equation 3.

3.2 Spectrum Mapping Through Pole Rotation

The main problems in spectral mapping via changing the position of the poles of the LP transfer function are: (a) labelling poles with formants, (b) the unpredictability of the magnitude and bandwidth of the transformed formants due to the interaction of neighbouring poles and (c) the formant magnitude cannot be directly controlled only by moving the pole radius (bandwidth) and frequency. The first problem, the pole-formant correspondence is solved using the formant tracking function. To achieve frequency warping and spectrum shaping, the poles' angular frequencies are *rotated* to shift formant frequencies and the poles' radii is *modified* to increase or decrease the formant bandwidth.

Pole rotation - For each frame t, the rotation of the i^{th} pole $F_i(t)$ associated with the formant b is defined as:

$$\hat{F}_i(t) = \alpha(t)\,F_b(t) + \alpha'(t)\left(F_i(t) - F_b(t)\right) \qquad (6)$$

where F_b is the formant frequency of the b^{th} formant, $F_i(t)$ and $\hat{F}_i(t)$ are the original and shifted pole frequencies

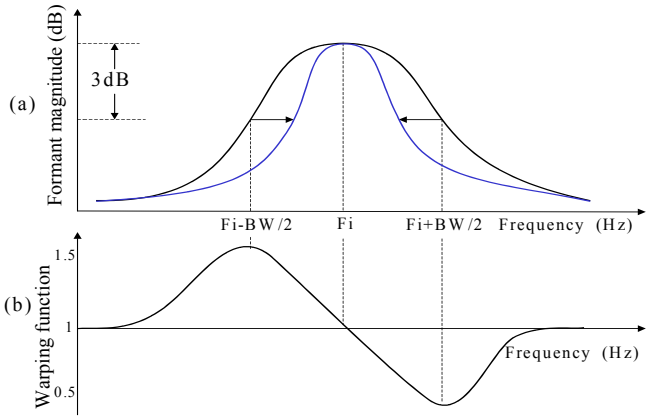

Figure 4 (a) Bandwidth warping process and (b) warping function for reducing the bandwidth of formant F_i by half.

respectively, α is the frequency warping factor and α' is the modified warping factor to maintain the distribution of poles between two formants.

The manipulation of the formant position and shape can be achieved by moving the poles in the LP spectrum of speech. The frequency of the formant depends on the angle of the pole, and its bandwidth on the pole radius. The equations describing the frequency and bandwidth of a pole *p* are given by

$$F_p = angle(p) * \frac{Fs}{2\pi} \tag{7}$$

$$BW_p = \log(abs(p)) * \frac{Fs}{\pi} \tag{8}$$

where *Fs* is the sampling frequency.

Cascade LP Implementation The cascade implementation of LP models allows independent changes in the frequency and bandwidth of each second order section. The main problem is that the intensity and hence the spectral shape of each section cannot be controlled independently without effecting the overall response at all frequencies. Also interaction of changes in the radii (the bandwidths) of closely spaced poles can severely affect the spectral shape. If two poles lie close together, they may both contribute significantly to the resulting overall intensity. To alleviate this problem a band-pass filter was applied centred at the each formant frequency and with a bandwidth set to twice the formant bandwidth. In this way only the desired formant band is effected. An alternative is to adjust the intensity of the entire spectrum after the poles have been moved using the spectral shaping procedure described in the previous section. Experiments showed the second method to be simpler and more accurate.

Parallel LP Implementation Through a partial fraction expansion of the linear predictive coefficients, the model can be expressed in a parallel form and each formant parameter, including the intensity, can be independently adjusted. The main problem with the parallel configuration is that a substantial change in the frequency and or bandwidth of each parallel second order section can have a substantial additive change in the overall spectral function of the model. From our experiments for voice conversion the more accurate of the two configurations for spectral mapping of a source to a target has proved to be the cascade pole rotation method and this was used for the experiments.

3.3 Selection of Mapping Resolution

The conversion factors for mapping of the source to target parameters can be frame-dependent, phoneme dependent or just one set of average mapping values for a pair of source and target speakers.

Overall means: One set of average formant parameters over all vowels are obtained and used for mapping the source speech to target voice.

Phoneme based: Phoneme-dependent transformation ratios are obtained for each cluster of triphones.

Frame based: The source and target frames are aligned and a set of transformation ratios is obtained for each frame.

4. EXPERIMENTS

Experiments were performed 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 to test the methods and to evaluate the issues related to voice conversion. The databases of the test speakers, consist of ten minutes of speech per speaker with a sampling rate of 10 kHz down-sampled from an original sampling rate of 16 kHz. An LP order of 13 is used and the LP coefficients are estimated every 10 ms using a 25 ms signal window. A male to female (M1/F) and a male to male (M1/M2) voice conversion was performed.

The first set of experiments obtained accurate formant models for the source and target speakers in order to establish the best mapping function between them. A phoneme-based formant analysis for each speaker was performed and some of the results are illustrated in figure 5. It shows the target to source ratios for a number of vowels for the male-to-male (M1/M2) and the male-to-female (M1/F) conversion. It is interesting to note that for the male-to-male case the differences in frequencies are not great. This is expected since the 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, which can differ significantly. In the female-to-male case the formant frequency difference is most important and for the case shown in figure (5) it is about 15%. To evaluate the effectiveness of the phoneme-dependent mapping ratios, the average formant trajectories for the phonemes of the source and target speakers were extracted (figure 6).

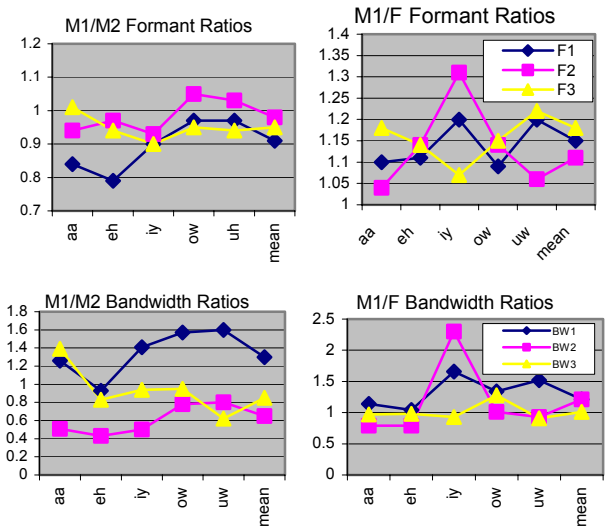

Figure 5: Frequencies and bandwidths ratios for formants 1,2 and 3 for the two pairs of speakers. The last column shows the mean across all vowels

Speech signals were mapped using both the spectrum warping and the pole rotation methods. Although for both the mapping is accurate in most cases, when the target differs significantly the pole rotating method gave better results because the poles can be rotated as much as it is desired. In order to establish the best way for deriving the ratios experiments using all three mapping resolutions were performed. The formant dependent resolution was preferred Two mapping examples can be seen in figure 7. In the first plot the spectra differ significantly, i.e the target has one less formant, but moving a pole close to the unit pole can reproduce it.

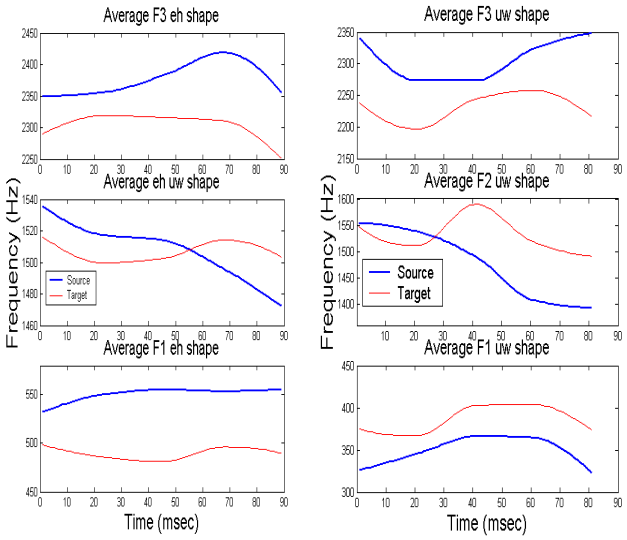

Figure 6: Average formants trajectories vowels /eh/ and /uw/ for source (male1) and target (male2) speakers.

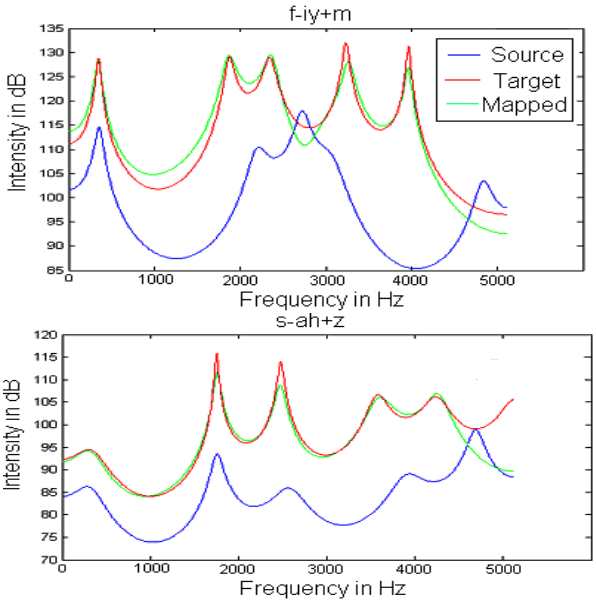

Figure 7: Source, target and mapped spectra of phonemes /iy/ and /ah/. Warping ratios for /iy/ :α=[0.97; 0.85; 0.87; 1.06; 0.82], β=[0.7823; 0.7; 0.9; 0.4; 0.4], γ=[5; 9.5; 2.5; 9; 15.5]

5. CONCLUSION

The problems of modelling and mapping of the formants of a source speaker to a target speaker were explored. The main focus was the evaluation of different methods for formant mapping in voice conversion. A spectrum warping and a pole rotation method were evaluated and the latter proved to be more effective due to its greater flexibility on the amount the formants can be warped. Different mapping resolutions were evaluated, and the phoneme based mapping was found to produce the best results.

REFERENCES

[1] Acero A. (1999), "Formant Analysis and Synthesis using Hidden Markov Models", *Proc. of the Eurospeech Conference*, Budapest .

[2] Allen J., Hunnicutt S., Klatt D. (1998) "From text to speech: the MITalk system", MIT Press, MA.

[3] Stylianou Y., Cappe O., Moulines E. (1998), "Continuous Probabilistic Transform for Voice Conversion", IEEE, SAP, Vol.6, No.2, pp.131-142.

[4] Kain A., Macon M., "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction". Proceedings of ICASSP, May 2001.

[5] Abe M. (1991), "A segment-based approach to voice conversion," in *Proc. ICASSP*, pp 765-768.

[6] Ho C.H., Rentzos D., Vaseghi S. (2002), "Formant Model estimation and transformation for Voice Morphing" in *Proc. ICSLP*, pp. 2149-2152.