COMPARIATIVE ANALYSIS AND SYSNTHESIS OF FORMANT TRAJECTORIES OF BRITISH AND BROAD AUSTRALIAN ACCENTS

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

Department of Electronic and Computer Engineering
Brunel University, UK UB8 3PH
*Fortune Institute of Technology, Kaohsiung, Taiwan
{Qin.Yan, Saeed.Vaseghi, Dimitrios.Rentzos, EmirTurajlic}@brunel.ac.uk *ch.ho@center.fjtc.edu.tw

Abstract

The differences between the formant trajectories of British and broad Australian English accents are analysed and used for accent conversion. An improved formant model based on linear prediction (LP) feature analysis and a 2-D hidden Markov model (HMM) of formants is employed for estimation of the formant trajectories of vowels and diphthongs. Comparative analysis of the formant values, the formant trajectories and the formant target points of British and broad Australian accents are presented. A method for ranking the contribution of formants to accent identity is proposed whereby formants are ranked according to the normalised distances between formants across accents. The first two formants are considered more sensitive to accents than other formants. Finally a set of experiments on accent conversion is presented to transform the broad Australian accent of a speaker to British Received Pronunciation (RP) accent by formant mapping and prosody modification. Perceptual evaluations of accent conversion results illustrate that besides prosodic correlates such as pitch and duration, formants also play an important role in conveying accents.

1. Introduction

Accents are differences in pronunciation by a community of people from a national or regional geographical area, or a social grouping. Accent is affected by differences in the phonetic transcriptions and the acoustic correlates of speech, including formants and their trajectories, pitch trajectory, pitch nucleus and duration parameters [1].

Accent is one of the main factors that impact automatic speech recognition (ASR) and text-to-speech synthesis (TTS). Performance of ASR systems is sensitive to accents and hence accent identification and modelling are essential for robust speech recognition [2]. Similarly accent models are applicable for accent morphing in text-to-speech synthesis systems.

The acoustics of accent are due to the differences in the configurations, positioning, tension and movement of laryngeal and supra-laryngeal articulator parameters. For example, in [3] Arslan and Hansen point out that generally non-native speakers of English do not produce the same tongue movement as native speakers, but produce accented sounds based on learned habits of tongue movements of their native language, which implies that their formants tend to move along the native language pronunciation. The difference in pitch and pitch trajectories in British and American English accents are analysed and presented in [4]. Recently, Harrington and Watson [5,6] explore the differences of formants between subclasses of Australian English: Broad Australian English, General Australian English and Cultivated Australian English and between New Zealand and Australian English.

The databases employed in this work for accent analysis are Australian National Database of Spoken Language (ANDOSL) for Australian English and Wall Street Journal Cambridge University (WSJCAM0) Database for British English. The subset of ANDSOL of broad Australian accent consists of 18 female and 18 male speakers with a total of 7200 utterances. The subset of WSJCAM0 of British English we use contains 40 female and 46 male speakers with 9476 utterances.

This paper presents a formant trajectory estimation method based on a two-dimensional time-frequency HMMs. A set of phoneme-dependent 2D-HMMs is trained to model the formants of a speaker's voice. The HMMs are then used to estimate the formant trajectory of each vowel. From the formant trajectories the target point of each vowel is estimated as the point where the formants reach steady status. Formant correlates of accents are ranked according to their individual importance to discrimination of accents. Finally a set of experiments in accent conversion is conducted via formant trajectory mapping, pitch modification and duration modification.

2. Formant Trajectory Estimation

Although automatic formant analysis of speech has received considerable attention and a variety of approaches have been developed, the calculation of accurate formant features from the speech signal remains a non-trivial problem. In this paper, a formant estimation method based on 2D-HMMs [7,8] is applied to estimate the formant trajectory.

The 2D HMM-based formant classifier [7] may associate two or more formant candidates (i.e. LP model pole frequencies) $F_{i(t)}$, with the same formant b, in these cases formant estimation is achieved through minimization of a weighted mean square error objective function

$$\widehat{F}_b(t) = \min_{F_b(t)} \sum_{i=1}^{I_b(t)} w_{bi}(t) \left[\frac{(F_{i(t)} - F_b(t))^2}{BW_i(t)^2} \right] \qquad (1)$$

where t denotes the frame index, b is the formant index, $I_{b(t)}$ is the total number of the formant candidates classified as formant b. The squared error function is weighted by a perceptual weight $1/(BW_i)^2$ where BW_i is the formant bandwidth, and a probabilistic weight defined as $w_{bi}(t)=$

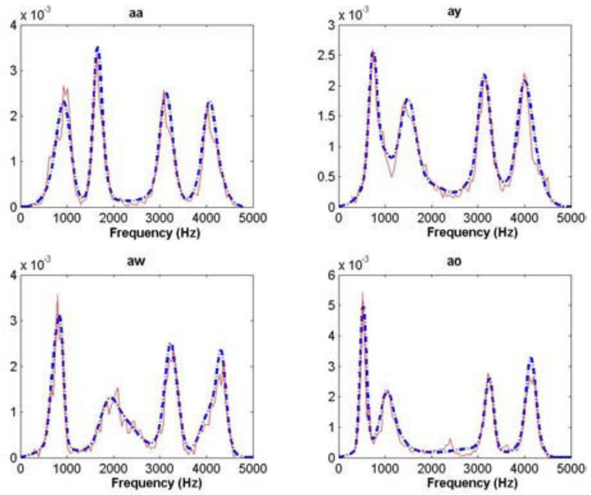

Figure 1*:* Comparison of histograms and HMMs of Formants for broad Australian English
Bold dashed line: Gaussian curves modelled by HMMs
Thin solid line: Histograms
X axis: Frequency (Hz); Y axis: probability.

$P(F_i|\lambda_b)$ where λ_b is the Gaussian mixtures models of b^{th} formant state of a phoneme-dependent HMM of formants.

The success of this method can be seen in Figure 1 showing the very close match between the histograms of formant candidates of a phoneme and the corresponding Gaussian models of HMM states. It can also be seen that the peaks of estimated Gaussian curves and histograms, which occur at formant frequencies, coincide.

3. Formant Target Estimation

In order to obtain an average formant trajectory model for each phoneme, the estimated formant trajectories of the different examples of each phoneme are linearly time-warped to ensure the same duration. Then all the time-normalized trajectories for each phoneme are averaged.

During the realisation of each phoneme, the acoustic configuration of vocal tract traverses a time trajectory aiming to reach a *target* point in the formant space. Using the method described in [5,6], the acoustic *target* point of a vowel is marked as a single time point (a "static" spectral slice [9]) between the acoustics onset and offset of the vowel. For high front vowels, the target is marked at the point where F2 reaches a peak; for back vowels, the target is marked at the point where F2 reaches a trough; for open vowels, the target is marked at the point where F1 reaches a maximum. When there is no acoustic evidence of any kind for a target point, the target is marked at the vowel's temporal mid-point. In rising diphthongs, two targets are marked using the same method as for monophthongs.

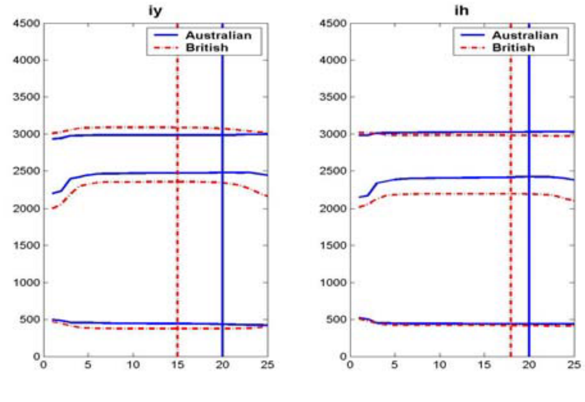

Figure 2: Average Formant Trajectories (female) of British, Australian and American after alignment
X-axis is the normalized time. Y-axis is frequency (Hz).

The normalized formant trajectories of some vowels are shown in Figure 2. Vertical lines are superimposed on these trajectories to mark the average time at which the vowel targets occur. It is noticeable that Australian vowels have considerably delayed target points compared with those of British vowels. This is particularly evident in /iy/. However, for /ih/ the target is a good deal closer to the vowel's temporal mid-point. These results conform to existing acoustic studies of Australian English that show Broad Australian possesses a rather delayed vowel target point compared with New Zealand English, Cultivated Australian and General Australian English [5,6].

4. Ranking of Formant Correlates of Accents

In [8] it is pointed out that second formant usually has the widest frequency range compared to other formants. In order to assess the importance of each formant on conveying an accent *A* compared to a reference accent *B* the formants are ranked according to some distance measure**.** A simple formulae for ranking the formant of accents *A* with reference to the formants of accent *B* is proposed as

$$\underset{i}{Rank}\left(\sum_{v=1}^{V}\left[\frac{F_{vi}^{A}-F_{vi}^{B}}{0.5(F_{vi}^{A}+F_{vi}^{B})}\right]^{2}\right) \tag{2}$$

Where $Rank(\cdot)$ can be a sorting function that sorts the formants in increasing or decreasing importance, F_{vi}^{A} is the average gender-dependent i^{th} formant of the vowel *v* from accent *A*, *V* is the number of vowels. Average formants of the vowels of each pair of accents are used in Equation (2) to obtain a ranking of formants influence in conveying accents. This formula is applied in other accent pairs as well (such as British and American, American and broad Australian). The experimental results rank the 2nd and the 1st formants as the most important two formants for accents, which are also consistent with perceptual evaluation in [8].

5. Accent Conversion

Analysis of formant spaces [8] and trajectories of different accents indicates that formants and their trajectories are important acoustic correlates of accents in addition to prosodic

Accent Pair	Formant Ranking Order			
	1	2	3	4
British & Australian	2st	1nd	3th	4rd

Table 1: Ranking of Formant Correlates of Accents Importance is ranked from 1 (high) to 4 (low).

and phonetic correlates which are traditionally regarded as correlates of accents. This section presents a set of experiments in accent conversion to change the accent of a broad Australian speaker into RP British accent. Each acoustic correlate of accents is independently modified. Results are perceptually evaluated by XAB [7] experiments.

5.1 Formant Transformation Methods

Since formant frequencies are among the most important features for voice synthesis, various frequency-warping methods have been developed. LP-spectrum warping and DFT-spectrum warping are two commonly used methods for changing the formants and the spectral shape of speech. In the following experiments, an LP-spectrum based phoneme-dependent frequency warping approach is used for synthesis of accent through mapping the formants.

In this method, three sets of parameters need to be estimated and transformed (Figure 3):

(a) Frequency difference between successive formants, $[F_{01}, F_{12}, F_{23}, F_{34}, F_{45}]$

(b) Bandwidth associated with the resonance at each formant, $[BW_1, BW_2, BW_3, BW_4]$

(c) Spectral intensity differences between formants, $[I_{12}, I_{23}, I_{34}]$

For simplicity, we assume that accents are not affected by bandwidth and intensity of resonance at formants [3]. Formants of each vowel are shifted according to the target to source formant ratios. The frequency warping function in the sub-band between the i^{th} and the $i^{th}+1$ formants can be described as:

$$\overline{f}_{i(i+1)} = \alpha_{i(i+1)} f_{i(i+1)} \qquad (3)$$

where $i(i+1)$ is the sub-band index along the frequency and $i = 0,1,2,\ldots$ (Figure 3), $f_{i(i+1)}$ is the frequency of source speech and $\overline{f}_{i(i+1)}$ is the mapped frequency. The frequency warping ratio $\alpha_{i(i+1)}$ is calculated by

$$\alpha_{i(i+1)} = \frac{f_{i+1}^{T} - f_{i}^{T}}{f_{i+1}^{S} - f_{i}^{S}} \qquad (4)$$

where $f_{i+1}{}^{T}$ is the average i^{th} gender-dependent formant frequency of target accent. The average formant frequencies (Table 2) of both British and Broad Australian accents are obtained from HMMs of formants [8]. They are used to obtain the formant warping ratios. A smooth spectrum warping function is then interpolated through these warping ratios.

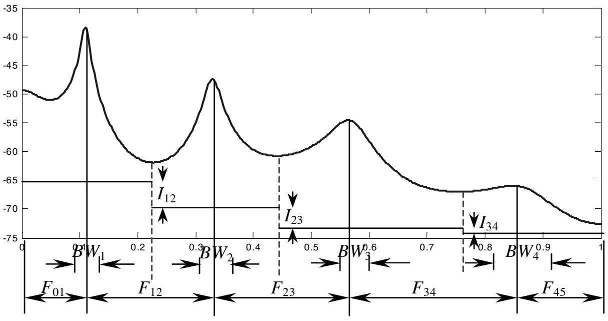

Figure 3: Illustration of LP spectrum based method for frequency warping and spectral shaping.

Prosodic parameters such as pitch and duration are modified using VoiceMorph, a software developed in our lab. For each voiced group, pitch trajectory and duration are adjusted according to the parameters of the corresponding voiced/unvoiced group of the target speech. However, the average pitch of the source speaker remains unchanged to maintain the speaker identity.

Vowel	Australian English (Female) Formants			
	F1	F2	F3	F4
AA	845	1532	3027	4058
AE	661	2020	3070	4212
AH	764	1575	2979	4014
AO	486	924	3039	3901
EH	506	2341	3105	4184
ER	510	1947	2909	4046
IH	421	2451	3099	4202
IY	411	2640	3126	4242
OH	636	1221	3010	3927
UW	387	2164	2844	3934
UH	420	1243	2927	4073
AY	697	1427	2944	3946
AW	697	1901	2993	4103
EY	636	2005	2982	4147
OW	611	1768	2872	4001
OY	512	1193	2863	3938
EA	498	2340	3040	4190
IA	392	2557	3159	4196
UA	653	1471	2813	3751

Vowel	British English (Female) Formants			
	F1	F2	F3	F4
AA	745	1234	2952	3816
AE	836	1740	2922	3999
AH	713	1551	2935	3916
AO	450	942	2971	3799
EH	675	1929	2961	4212
ER	598	1795	2936	4079
IH	428	2105	2972	4202
IY	358	2526	3102	4128
OH	586	1142	2926	3725
UW	371	1613	2783	3764
UH	423	1306	2840	3796
AY	712	1613	2929	3889
AW	765	1634	2897	3979
EY	509	2226	3037	4134
OW	528	1738	2830	3962
OY	506	1381	2950	3903
EA	654	1961	2936	4152
IA	453	2192	2930	4070
UA	460	1793	2962	4023

Table 2: Average Formant Frequencies of Broad Australian and British English (female)

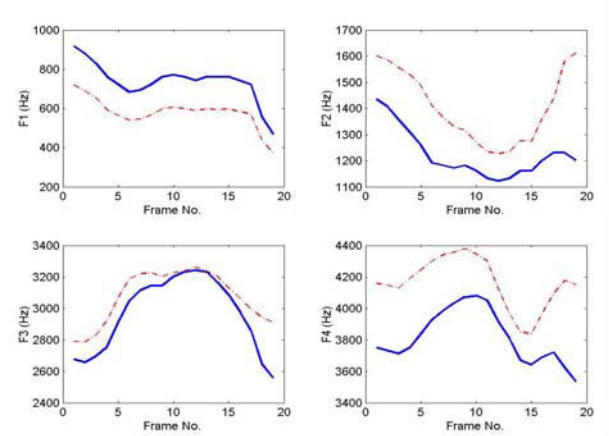

Figure 4: Illustration of warped (solid line) and original (dash dot line) formant trajectories of /aa/
X axis: Frequency (Hz); Y axis: Frame (time).

5.2 Experiments and Results

The source and target speech are down-sampled to 10kHz from the original sampling rates of 20kHz (ANDOSL) and 16kHz (WJSCAM0). The aim is to avoid unnecessary high frequencies and formants above 5kHz. The speech are then segmented and labeled. HMMs of formants are then trained on LP poles of segmented and phonemically labeled speech.

Using the estimates of the warping ratios, LP-spectrum mapping is employed, to move the formants of vowels in utterances from broad Australian speakers towards the average formants of vowels of British RP accent. In total four formants of the source speaker are mapped towards the target accent. The mapped and original formant trajectories of some vowels are displayed in Figure 4. Formants trajectories are moved up or down according to the ratios of the formants from target accent (British) and source accent (broad Australian). It can be observed from Figure 4 that the overall shape of each formant trajectory of the vowels remains unaffected after formants are moved up and down while the start and end of formant trajectories are slightly changed. The reason is that the formants trajectories of each pair of adjacent phonemes need to remain continuous at the segment boundaries to keep the formant trajectories smooth. It is also possible to transform a vowel to another vowel in a different accent by adjusting the formants towards those of the target vowel. After moving the formants of broad Australian accent towards those of British RP accent, a certain amount of residue broad Australian accent still can be heard. This is due to the remaining correlates of accent such as those contained in pitch trajectory and prosody. Hence prosodic parameters are also modified.

XAB method is used to perceptually evaluate the success of accent conversion techniques used here. The source, target and transformed speech are presented to listeners. The listeners are then asked if the transformed voice has a similar accent to the target or the source speaker. In XAB test, the speech samples *A* and *B* are from two set of different speakers with British and broad Australian accents. The speech sample *X* is the transformed source speech. The source, transformed and target speech have same content. Each listener has to identify 7 XAB test sets. In our experiments 78% of listeners' answers indicated that, after accent transformation, the broad Australian source speech have a similar accent to British target accent. The experiments demonstrate that formant correlates are among the important factors of accents.

6. Conclusion

This paper presented a comparative analysis of formant trajectories of British and Broad Australian English accents. 2D HMM formant classifiers are used for estimation of formant trajectories. From the formant trajectories the target point of each vowel in the formant space is estimated as the point where the formant values reach a "static" status. It is found that broad Australian possesses a delayed vowel target compared with British. The formants are ranked according to their contribution to accent and the first and second formants are determined as the most crucial formants to accent as to their individual contribution to discrimination of accent pairs.

A set of experiments in accent conversion is performed. Formants frequencies of vowels are shifted according to the estimated formant frequency warping factors. After modifications of the formant trajectories, the pitch trajectory and the timing (duration) of the source broad Australian speech, listener subjects agree that the converted speeches have similar accent to British.

For future work more efforts are required on modeling prosody structure of accents and on extension of the experiments to other accent pairs of such as British and American.

7. Acknowledgements

We would like to thank EPSRC for funding project no GR/M98036.

8. References

[1] Wells J.C., *Accents of English*, Cambridge University Press, (1982).

[2] Humphries J., "Accent Modelling and Adaptation in Automatic Speech recognition", PhD Thesis, Cambridge University Engineering Department (1997)

[3] Arslan L. M., Hansen H., "A Study of Temporal Features and Frequency Characteristics in American English Foreign Accent", *Journal of Acoustic Society of America*, vol. 102(1), p. 28-40, (1997)

[4] Yan Q., Vaseghi S., "A Comparative Analysis of UK and US English Accents In Recognition and Synthesis", *ICASSP,* Orlando (2002)

[5] Harrington J., Cox F., Evans Z., "An Acoustic Phonetic Study of Broad, General, and Cultivated Australian English Vowel*s*", *Australian Journal of Linguistics* 17: 155-184 (1997)

[6] Watson C., Harrington J., Evans Z., "An Acoustic Comparison between New Zealand and Australian English Vowels", *Australian Journal of Linguistics* (1996)

[7] Ho Ching-Hsiang, "Speaker Modelling for Voice Conversion", PHD thesis, Department of Electronic and Computer Engineering, Brunel University (2001)

[8] Yan Q., Vaseghi S. "Analysis, Modelling and Synthesis of British, American and Australian Accents", *ICASSP* (2003)

[9] Harrington. J., Cassidy, S. "Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English", Language and Speech 37 p357-373 (1994)