ANALYSIS BY SYNTHESIS OF ACOUSTIC CORRELATES OF BRITISH, AUSTRALIAN AND AMERICAN ACCENTS

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

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

ABSTRACT

This paper presents analysis through synthesis of the acoustic correlates of British, Australian and American accents by transforming the correlates individually across the accents. The acoustic correlates of accents are grouped into three main categories: (a) the spectral features at formants, (b) the pitch intonation pattern and (c) duration. The modeling and transformation methods for each group of voice features are outlined. The spectral features at formants are modeled using two-dimensional (2D) phoneme-dependent HMMs. Subband frequency warping is used for spectrum transformation where the subbands are centred on estimates of the formant trajectories. The F0 contour is used for modeling the pitch and intonation patterns of speech. A method based on time domain pitch synchronous overlap and add algorithm(TD-PSOLA) is used for transformation of pitch intonation and duration pattern. Perceptual tests based on mean opinion score (MOS) are conducted to rank the main features of accents. Formants are regarded as the most important features of accents, followed by intonation pattern and duration.

1. INTRODUCTION

Accents are differences in pronunciation by a community of people from a national or regional geographical area, or a social grouping [1]. Accent is one of the main factors that impact the performance of automatic speech recognition (ASR) and text-to-speech synthesis (TTS). Accent identification and modeling can be used to improve the robustness of speech recognition [2]. Similarly accent models are applicable in accent morphing for text-to-speech synthesis systems.

Accent is affected by the differences in the phonetic transcriptions and the acoustic correlates of speech [1]. The acoustics of accent are due to the differences in the configurations, positioning, tension and movement of laryngeal and supra-laryngeal articulator parameters. The differences in articulation of phonemes across accents are manifested in the differences in the formants and their trajectories[3,4]; pitch intonation and duration parameters [1,5]. For example, in [6] 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 trajectories in British and American English accents are analysed and presented in [5].

In this paper, three main aspects of the acoustic correlates of accents, considered to be essential for modeling accent, are investigated. These are:

- Formant correlates of accents,
- Pitch intonation correlates of accents,
- Phoneme duration pattern and speaking rate correlates of accents.

The importance of the three acoustic correlates of accents is perceptually assessed after transforming them individually.

2. OVERVIEW OF ACOUSTIC FEATURES OF ACCENTS

Figure (1) illustrates a block diagram overview of the accent model estimation methods used in this paper. The variation of formants, pitch intonation pattern and duration pattern are individually modeled by statistical models.

The formants' trajectories of each phoneme are modeled by 2D-HMMs [7]. A 2D-HMM is a combination of a 1-D HMM along the time dimension and a 1-D HMM along the frequency dimension. The configuration parameters are listed in Table 1. Along the frequency axis, every state of a 2D-HMM models the distribution of one formant of the phoneme.

Pitch trajectories are used to model the broad intonation characteristics patterns of an accent. The pitch parameters used in modeling the broad intonation pattern are the slope of initial pitch rise, the slope of final pitch fall, pitch frequency range and pitch contour slope.

The statistics of phoneme duration pattern and speaking rate variations across accents are obtained from the segmentation and

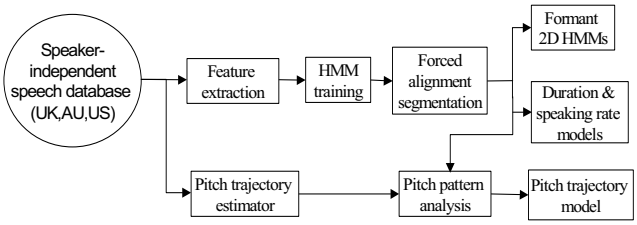

Figure 1 Overview of an acoustic model of accents.

Parameters	Values
Sampling frequency	10kHz
Frame size	25 ms
Frame rate	10 ms
No. of states of each cepstrum HMM	3
No. of Gaussian components per cepstrum HMM state	20
LP order	13
No. of states each formant HMM	4 or 5
No. of Gaussian components per state in formant HMMs	4

Table 1 Configuration of analysis of acoustic correlates of accents.

Database Name (accent)	No. Speakers (f/m)	No. Sentences
ANDOSL (Australian)	18/18	7200
WSJ (American)	36/38	9438
WSJCAM0(British)	40/46	9476

Table 2 Databases configurations.

labeling of speech databases using speaker-dependent HMMs and Viterbi segmentation in the 'forced-alignment' mood (Table 1) [8,9].

The databases used in this work for accent analysis are Australian National Database of Spoken Language (ANDOSL) for broad Australian English accent, Wall Street Journal database (WSJ) for American English accent and Wall Street Journal Database Cambridge University (WSJCAM0) for British English accent as listed in Table 2.

3. TRANSFORMATION OF FORMANTS

Frequency warping through non-uniform adaptive sub-band spectral mapping [7] is deployed to transform the formants of a source speaker towards those of a target accent. In the spectrum warping method the estimates of the formant tracks are used to divide the signal spectrum into N sub-bands centred on the formant trajectories. The inputs to the spectrum mapping function are the LP-spectrum and the formant feature vector of the current source speech frame. The formant transformation is

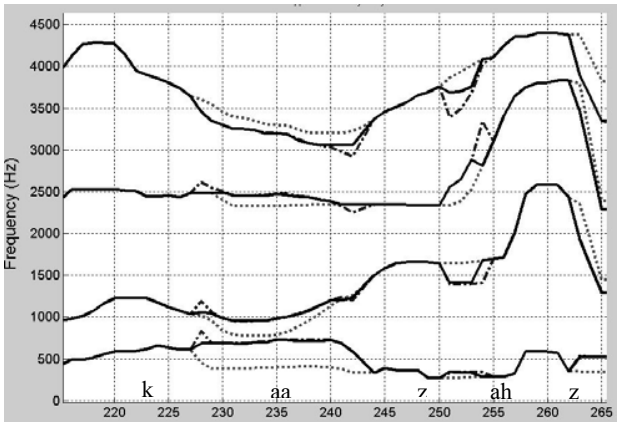

Figure 2: An illustration of: original formant trajectories (dot lines); mapped unsmoothed formant trajectories (thin dash-dot lines); mapped and smoothed formant trajectories (thick solid lines) of the spoken word "*causes*". X-axis is Frame No, Y-axis is frequency.

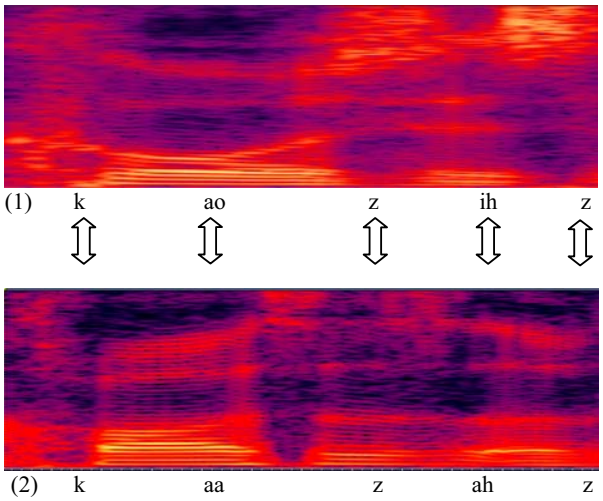

Figure 3: An example of different phonetic transcription across (1) British and (2) American accents for spoken word "*causes*".

achieved by multiplying the original source LP spectrum by a set of phoneme-dependent source-to-target warping ratios (Eq. 1)

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

The phoneme dependent warping ratios are obtained by

$$\alpha_{i(i+1)} = \frac{F_{i+1}^T - F_i^T}{F_{i+1}^S - F_i^S} \quad (2)$$

Where $F_{i+1}{}^T$ is the average i^{th} gender-dependent formant frequency of the target accent and F_i^S is the average i^{th} gender-dependent formant frequency of the source accent. The estimates of the average formant frequencies of each phoneme are obtained from HMMs of formants [7]. In frequency domain, the warping ratios at N formants are interpolated to produce a frequency warping function at K (typically 512) samples. In time domain, smoothing across phoneme boundaries is achieved using a polynomial interpolation (Figure (2)).

It is worthwhile pointing out that F_i^S and F_i^T are not necessarily from the same phoneme as the same word can have different phonemic transcriptions across the accents. This is particularly the case in accent conversion between British and American English because there are five fewer vowels (*ax ah ia ea ua*) in American as transcribed by CMU phonetic dictionary compared to British BEEP phonetic dictionary. Figure 3 illustrates an example of differences in phonetic transcription of the word "*causes*". In this case, F_i^S is the formant frequency of British vowel *ao*, F_i^T is the formant frequency of American vowel *aa*. Hence, for voice conversion from British to American, the formant frequencies of British *ao* are shifted towards the formant frequencies of American *aa*. Consequently phoneme identity can be changed through transformation of formants.

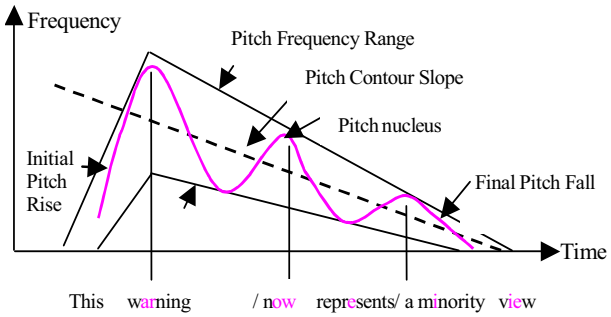

Figure 4: Broad Pitch Intonation Pattern in Accent Modelling Thick line is the pitch contour.

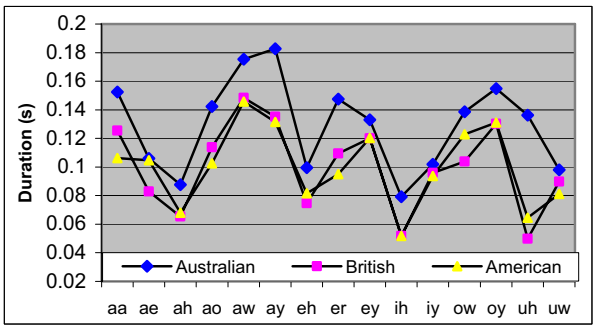

Figure 6: Comparison of average duration of Australian, British and American vowels.

4. TRANSFORMATION OF PITCH INTONATION CORRELATES OF ACCENTS

Intonation is one of the most important characteristics of accents. Popular intonation models include Tone and Break Index (ToBI)[10], Tilt[11], Momel[12] and Fujisaki[13]. In this paper, a model of broad pitch intonation pattern is employed to describe the pitch intonation difference across the accents [9] (Figure 4). A set of features describing the broad patterns of variation of pitch contour are pitch frequency range, pitch contour slope (F_{0slope}), rate of initial pitch rise (∂F_{0I}), rate of final pitch fall (∂F_{0F}). The pitch trajectory features are used for modeling intonations as illustrated in Figure (4):

Pitch frequency range represents the range of F_0 bounded from the lowest to the highest frequency. However, in this work we define pitch range as

$$\text{range}(F_0) = \overline{F_0} \pm 3 \times \text{std}(F_0) \qquad (3)$$

where $std(F_0)$ is the standard deviation of F_0. Standard deviation is used instead of the maximum and minimum values of the pitch frequency over an utterance to avoid errors in the estimation of pitch range due to outliers.

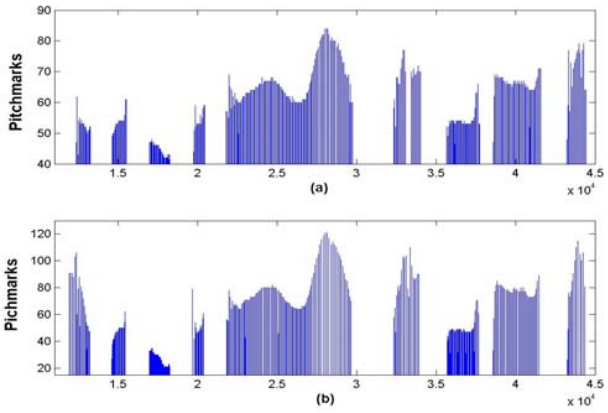

Figure 5: Original pitch marks (a) from a broad Australian accented sentence and transformed pitch marks (b) towards to British accent. Pitch frequency range increases 50%; F_{0slope} increases 5%; ∂F_{0I} increases 50%; ∂F_{0F} decreases 5%.

Pitch contour slope (F_{0slope}) is the overall slope of pitch trajectory over an utterance.

Rate of initial pitch rise (∂F_{0I}) at the beginning of a pitch contour. This is described by the rate and direction of the F_0 change at the initial pitch contour segment at the beginning of the utterance.

Rate of final pitch fall (∂F_{0F}) is described by the rate and direction of the F_0 change at the end of the utterance.

A pitch transformation method, based on TD-PSOLA[14], is developed that allows independent modification of the pitch intonation parameters. The modification of intonation features across British, Australian and American accents is based on the analysis and estimates of the features for these accents [9]. Figure 5 gives an example of broad pitch intonation modification.

5. TRANSFORMATION OF DURATION CORRELATES OF ACCENTS

Accents are partly conveyed by the differences in vowel duration patterns and speaking rate [9]. Variation of the average phoneme durations across accents is shown in Figure 6. In our experiments, the transformation of duration pattern of accents is accomplished by adjusting the duration of each phoneme to the mean duration of the corresponding phoneme in the target accent.

The modification of phoneme duration pattern is achieved using TD-PSOLA [14] and estimates of the ratio of the average duration of the phoneme in the source accent to that of the corresponding phoneme in the target accent.

6. RANKING OF CORRELATES OF ACCENTS

In order to perceptually assess the importance of the acoustic correlates of accents, mean opinion score tests (MOS)[15,16] are performed. In total 7 British speaker subjects are used. Each subject is given 5 sets of sentences from different Australian speakers. In each set, there are five sentences (*A B C D E*). Two of them are original British accented speech and Australian accented speech. The remaining three are transformed speeches from Australian accented speech by modification of formant, pitch and duration correlates of the source accent towards the target accent respectively. All the sentences in one set have the same text content and are listed in random order. In addition, subjects are given two known accent sentences: Sentence *S* and

sentence *T*. Sentence *S* is the source Australian accented speech and sentence *T* is the target British accented speech. Both sentences have different text from test sentences. Subjects are then asked to give scores of the similarity between given set of speech (Sentence *A B C D E*) and the target accented speech (Sentence *T*) and source accented speech (Sentence S). The score is in the range between 10 (identical) to 0 (completely different). Similar tests are conducted on British and American accent pair.

The MOS results are shown in Table 3. It can be seen that for British and American accent pair, modification of formants of speech is more influential in affecting accent change than modification of duration and broad pitch intonation pattern. However, for Australian and British accent pair, Australian speech after modification of broad pitch intonation are considered to have slightly more British accent characteristics compared to that obtained after modification of formants. In both cases, duration pattern exhibits little impact on accent.

7. CONCLUSTION

This paper presented an analysis the acoustic correlates of accents: formants, pitch intonation and duration. Each correlate is transformed individually for perceptual assessment. Formants correlates of accents are transformed by non-uniform subband frequency warping based on 2D HMMs. A TD-PSOLA based method is used for transformation of pitch intonation patterns and duration pattern. The MOS test indicates that formants are the most important correlate accents of the three.

Future works include extending the experiments to other accent pairs (i.e. American and Australian accent pair) with more native American and Australian subject.

8. ACKNOWLEDGEMENTS

We wish to thank the UK's EPSRC for funding project no GR/M98036.

Accents	British Accents	American Accents
Original British speech	8	2
British speech after modification of formants towards American	4	6
British speech after pitch intonation modification towards American	6	4
British speech after duration modification towards American	7	3
Original American speech	1	9

Table 3(a) MOS scores of transformation of acoustic correlates of British accented speech to American accent

Accents	Australian Accents	British Accents
Original Australian Speech	7	3
Australian speech after formants modification towards British	5	5
Australian speech after pitch intonation modification towards British	3	7
Australian speech after duration modification towards British	6	3
Original British Speech	1	9

Table 3(b) MOS scores of transformation of acoustic correlates of broad Australian accented speech to British accent

9. REFERENCE

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

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

[5]Yan Q. , Vaseghi S., "A Comparative Analysis of UK and US English Accents in Recognition and Synthesis", *ICASSP*, Florida pp 413-416 (2002)

[6] 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), pp. 28-40, (1997)

[7] Yan Q., Vaseghi S., "Analysis, Modelling and Synthesis of formants of British, American and Australian Accents" *ICASSP*, Hong Kong, pp. 712-715 (2003)

[8] Sethy A., Narayanan S., "Refined Speech Segmentation for Concatenative Speech Synthesis" *ICSLP* (2002)

[9]Yan Q., Vaseghi S., Rentzos D., Ho C., Turajlic E., "Analysis of Acoustic Correlates of British, Australian and American Accents" IEEE Automatic Recognition and Understanding Workshop (2003)

[10] Silverman, K., Beckman, M., Pitrelli,J., Ostendorf, M., Wightman,C., Price,P., Pierrehumbert,J., Hirschberg,J. "ToBI; A standard for labeling English prosody", *ICSLP*. (1992).

[11] Talyor, P.A., "The Tilt intonation model" *ICSLP* Sydeny Australian. (1998).

[12] Hirst D., Espesser R., "Automatic Modelling of Fundamental Frequency Using A Quadratic Spline Function" Travaux de l'Institute de Phonetique d'Aix vo. 15 pp.75-85, (1993).

[13] Fujisaki H., "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour." In Fujimura, O. (Ed.) Vocal Fold Physiology: Voice Production, Mechanisms and Functions. Raven, New York, NY, pp.135-149 (1988).

[14]Moulines, E., Charpentier, F. "Pitch-Synchronous Waveform Processing techniques for Test-to-Speech Synthesis Using Diphones", Speech Communication, 9 pp.453-467 (1990)

[15] Stylianou, Y. Cappe, O. Moulines, E. "Continuous Probabilistic Transform for Voice Conversion" *IEEE Transaction on Speech and Audio Processing*, Vol6, No. 2 pp 131-142 (1998)

[16] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., "Voice Conversion Through Vector Quantization", *ICASS*P pp 565-568 (1998)