Modeling and synthesis of English regional accents with pitch and duration correlates

Qin Yan a,*, Saeed Vaseghi b

a College of Computer and Information, Hohai University, Nanjing 210098, China
b School of Engineering and Design, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

Received 16 March 2009; received in revised form 9 November 2009; accepted 15 February 2010
Available online 1 March 2010

Abstract

This paper provides an introduction to the acoustic–phonetic structure of English regional accents and presents a signal processing method for the modeling and transformation of the acoustic correlates of English accents for example from British English to American English. The focus of this paper is on the modeling of intonation and duration correlates of accents as the modeling of formants is described in previous papers (Yan et al., 2007; Vaseghi et al., 2009). The intonation correlates of accents are modeled with the statistics of a set of broad features of the pitch contour. The statistical models of phoneme durations and word speaking rates are obtained from automatic segmentation of word/phoneme boundaries of speech databases. A contribution of this paper is the use of accent synthesis for comparative evaluation of the causal effects of the acoustic correlates of accent. The differences between the acoustics–phonetic realizations of British Received Pronunciation (RP), Broad Australian (BAU) and General American (GenAm) English accents are modeled and used in an accent transformation and synthesis method for evaluation of the influence of formant, pitch and duration on conveying accents.

© 2010 Elsevier Ltd. All rights reserved.

Keywords: Accent; Formant; HMMs; Pitch; Intonation; Duration

1. Introduction

Modeling and synthesis of accents are some of the most challenging and fascinating aspects of speech research (Wells, 1982; Humphries et al., 1996). Accent models have applications in accent morphing, accent adaptation in speech recognition, accent clustering and accent identification. Accent morphing has applications in text to speech synthesis where the user can be given the option of selecting an accent from a number of choices; in the teaching/learning of accents; in changing the accent of a character in a film or play or in computer toys/games and in multimedia systems.

Accent is a distinctive pattern of pronunciation by a community of people who belong to a national, geographical or socio-economical grouping (Wells, 1982; Bezooijen van, 1995). The distinctive patterns arise from...
differences in vowels and consonants as well as the stress patterns, rhythm, and intonation. These features can vary in the way in which they are articulated (e.g. Formants, stress patterns), but they can also involve different underlying abstract elements (e.g. inventories of consonants and vowels used).

Accents evolve over time influenced mainly by such factors as mass immigrations, social and cultural trends, sex, education and the mass media (Yan et al., 2007). For example, the Australian accent has been influenced by the waves of immigrations to Australia, by London “Cockney”, Irish and American accents. Similarly, the Liverpool accent has been influenced by Irish immigration and the Northern Ireland accent has been influenced by Scottish immigration. African–American accents are thought to be rooted in African languages and southern US accents.

In general, there are two broad approaches to the classification of the differences between accents:

- **Historical approach** compares the roots of accents, the evolutionary changes in sounds as various accents merge or diverge, the rules of the pronunciation and how the rules have evolved over the time.
- **Structural, synchronic approach** (Trubetzkoy, 1931), models an accent in terms of the differences in: (a) phonemic systems, (b) phonotactic distributions, (c) lexical distributions, (d) prosodic properties/features (stress, rhythm, intonation and, in tone languages, lexical tone) and (e) phonetic realisation.

### 1.1. Acoustics and phonetics of accents

The differences between accents can be attributed to five types of properties (Wells, 1982).

(a) **Differences in the phonemic systems of accents**, i.e. in the number or identity of the phonemes. Most phonemes in American, British and Australian accents match but sometimes a phoneme in one accent can be matched by two phonemes in another accent and vice versa (Wells, 1982). For example, in British RP the phoneme /ɒ/ in *stop, dodge, romp*, corresponds to /ə/ in GenAm but the same phoneme /ɒ/ in *cough, gone* and *Boston*, corresponds to /ɜː/ in GenAm. In British RP the post-vocically approximant /r/ is not pronounced, but the preceding vowel is diphthongised, for example the British diphthong /ʊə/ corresponds to the GenAm two phoneme sequence /ɪə/ (e.g. the word “hear” is /hɪə/ in British RP and /hər/ in GenAm) (Wells, 1982; BEEP Dictionary; CMU Dictionary). Furthermore, the /s/ sound in British RP is realized in GenAm as a monophthongal r-colored vowel, e.g. “*bird*” is /ɜːd/ in GenAm and /bɜːd/ in British RP and “*dinner*” is /dɪnər/ in GenAm and /dɪnər/ in British RP (Wells, 1982; CMU Dictionary). Australian English also has a number of vowels with significantly different realizations such as /æɪ/ instead of /æ/ and /æz/ for /au/ compared with British RP.

(b) **Differences in the lexical realizations of words.** Accents may vary in the phonemes they use for the lexical representation of particular words or morphemes. For example some accents, such as those in parts of northern England, use /æ/ in “*bath*, “*staff*, “*grass*” and “*basket*”, while others use /ɑː/. As further examples, ‘*JOHN*’ is pronounced as /dʒɔːn/ in general American English accent but as /dʒən/ in British English accent and ‘*IMMEDIATE*’ is pronounced as /iːmɪdɪˈmeɪt/ in British RP but it is pronounced as /ɪˈmɪdɪmɛt/ in General and Broad Australian accents (Yan et al., 2007; Wells, 1982).

(c) **Differences in the phonotactic distribution.** The phonotactic distribution of a given phoneme is the set of phonetic contexts in which it may occur. A prime example of differences in phonotactics of accents is the division of English accents into rhotic and non-rhotic accents. In non-rhotic accents, such as RP British, the instances of the phoneme /r/ that occur before a consonant, or at the end of a word, are not pronounced whereas in rhotic accents, such as general American, /r/ can occur with an overt realization in a wide variety of contexts. For instance, the word “*farm*” is pronounced as /fɑːm/ in rhotic accents but as /fəm/ in non-rhotic accents. The rhotic accents include those typical of Scotland, Ireland, Canada and most of the United States. The non-rhotic accents include those typical of Australia, New Zealand, South Africa and most of England and Wales.

(d) **Differences in prosody of accents.** These are differences in the rhythms of intonations and the stress patterns associated with an accent. The acoustic parameters needed for the modeling of prosody of an accent are: (1) pitch contour, (2) phoneme durations and speaking rates and (3) intensity. In our exper-
iments we have not observed a consistent pattern of differences in intensity among our databases of different accents.

(e) Differences in the phonetic realizations of accents. The differences in phonetic realizations of the accents of a language are in part due to the differences, during the realization of sounds, in the configurations, the positioning, tension and movement of laryngeal and supra-laryngeal articulators (Yan et al., 2007; Wells, 1982). These differences can be modeled as differences in formant trajectories and spectral slopes of the acoustic of phonetics of accents. For instance the degree of vowel backness, diphthongization and brightness are usually associated with formants. These attributes are reliable acoustic–phonetic features for accent distinction (Clopper and Pisoni, 2004). For example, backness of /_e_/ and r-fulness is the only significant predictor variables for New England talkers (Clopper and Pisoni, 2004). Note that the possible correlations of glottal pulse shape with accent are not considered in this paper.

The remainder of this paper is organized as follows. Section 2 presents an overview of the accent model estimation method. Section 3 presents an investigation of the differences across the formant spaces of British RP, general American and Broad Australian English accents. Section 4 presents the intonation model and a comparative analysis of the correlates of pitch intonation. Section 5 presents an analysis of the pattern of phoneme durations and speaking rate variations across accents. Section 6 presents a set of accent conversion experiments for evaluation of the influence of each acoustic correlate of accents. Finally Section 7 concludes the paper.

2. Description of system outline for accent modelling

Fig. 1 illustrates an overview of the accent modeling method presented in this paper. The speech databases used for accent analysis are subsets of Australian National Database of Spoken Language (ANDOSL) for Broad Australian English, Wall Street Journal database (WSJ) for general American English and Wall Street Journal Database Cambridge University (WSJCAM0) for British English Received Pronunciation. The subset of ANDSOL contains 18 female and 18 male speakers with a total of 7200 utterances. The subset of WSJ database contains 36 female and 38 male speakers with 9438 utterances. The subset of WSJCAM0 contains 40 female and 46 male speakers with 9476 utterances. The style of speech in all databases, we have used, is declarative; this implies that there are no additional variations in intonation patterns due to varying speaking styles. Each utterance is a sentence and is read in a continuous manner. The multi-pronunciation dictionaries used in this work include BEEP dictionary (British accent) BEEP Dictionary, CMU dictionary (CMU Dictionary) (American accent) and Macquarie dictionary (Macquarie Dictionary) (Australian accent).

All the databases are down-sampled to 10 kHz. Note that the fundamental frequency information and the formant frequencies are within the 5 kHz bandwidth supported by a sampling rate of 10 kHz. The raw speech data is processed to extract cepstrum features, formant features, pitch trajectories and durations. First, a set of left–right HMMs for phonetic units of speech are trained on 39-dimensional feature vectors comprised of 13 Mel-Frequency Cepstral Coefficient (MFCCs) and their 1st derivative and 2nd derivative features (Young et al., 2002), Each HMM has three states and in each state the feature probability distribution is modeled with

![Diagram](image-url)
a Gaussian mixture model (GMM) with 20 components, this choice was made by monitoring the convergence
of likelihood of the HMMs. The HMMs are used, in forced-alignment recognition modes (i.e. with the transcriptions supplied), to extract the boundaries and durations of phonemes and words. A detailed investigation of the accuracy of HMM-based phoneme segmentation and boundary estimation is described in Sethy (Sethy and Narayanan, 2002). Finally, the probability distributions of formants, pitch intonation pattern and duration pattern are modeled by HMMs or GMMs (Vaseghi et al., 2009).

3. Formant correlates of accent

The differences in formants across accents for Broad, General, and Cultivated Australian accents and between New Zealand and Australian accents were investigated by Harrington et al. (1997) and Watson et al. (1998). Weil et al. (2000) investigates the differences in the formants of the diphthongs between Southern US English accent and Standard American English accent.

3.1. Formant model estimation

The formant estimation method used in this work is described in detail in our previous paper (Yan et al., 2007) and outlined here. For each speech frame, a set of feature vectors for formant candidates \( \{F_k\} \) are extracted from the poles of a linear prediction (LP) model of speech as

\[
F_k = [F_k, B_k, M_k, \Delta F_k, \Delta B_k, \Delta M_k] \quad k = 1, \ldots, N
\]

where \( N \) is the number of formants, \( F_k, B_k \) and \( M_k \) are the frequency, bandwidth and magnitude of the \( k \)th pole of the LP model of speech respectively and \( \Delta F_k, \Delta B_k, \Delta M_k \) are the slopes of the time trajectories of \( F_k, B_k \) and \( M_k \) respectively.

HMMs (Young et al., 2002), widely employed in speech recognition, are used here for modelling the probability density functions (pdfs) of formants. Fig. 2 shows a phoneme-dependent formant model, based on a 2-D HMM with three left–right states across time and five states across frequency. The \( k \)th state of the formant HMM, along the frequency axis, models the distribution of the \( k \)th formant. Given a set of observations of resonance frequencies of speech \( O_n \), the maximum likelihood estimate of the formants is obtained as

\[
[F_1, F_2, \ldots, F_N] = \arg \max_{F_1, F_2, \ldots, F_N} P(O_n | F_1, F_2, \ldots, F_N | A_m)
\]

where \( O_n \) is obtained from the poles of an LP analysis of a segment of a speech phoneme and sorted in terms of increasing frequency, \( A_m \) is an HMM of the formants of phoneme \( m \) and \( N \) is the number of formants. Using a set of training data, the distribution of each formant vector in each HMM state is modeled by a multivariate GMM trained via the Expectation Maximization (EM) algorithm.

Fig. 2. An example of parallel left-to-right HMMs; along the vertical axis each state represents a different formant, along the time axis each state represents a different segment.
3.2. Comparison of the formant spaces of British, Australian and American accents

Using the formant estimation method outlined above, the average values of the first to fourth formants \((F_1–F_4)\) of the vowels and diphthongs of female speakers with British, Australian and American accents are calculated and compared in Figs. 3 and 4.

Some differences in the formant space of the accents are evident from Fig. 3. It can be seen that, except for the vowel /æ/, the Australian vowels have a lower value of \(F_1\) than those of American. The Australian vowels exhibit a higher value of \(F_2\) than American except for /ɜ/:. On average, the 2nd formants of Australian vowels are 11% and 8% higher than those of the British and the American accents respectively. The 3rd and 4th formants are consistently higher in the Australian accent compared to the British accent. With the exception of the vowel /ɜ:/, American vowels exhibit higher values of \(F_3\) and \(F_4\) than British speakers.

Fig. 4 illustrates the \(F_1\) vs. \(F_2\) formant spaces of British, American and Australian accents. In phonetics (Wells, 1982), the front or back movements of vowels are associated with the movements of \(F_2\) while high and low movements are associated with the movements of \(F_1\). Front movement of a vowel implies a higher value of \(F_2\) while back movement corresponds to a lower value of \(F_2\). Similarly, high movement of a vowel implies a lower value of \(F_1\) towards the so-called ‘close’ vowels, whereas low movement of a vowel corresponds to higher value of \(F_1\) towards ‘open’ vowels. It can be noted from Fig. 4 that in comparison to the British and American vowels, the Australian vowels exhibit the following distinctive characteristics:

- Raising of the vowels /æ/ (e.g. had) and /ɛ/ (e.g. head) (Fig. 4a).
- Fronting of the open vowel /ɑ:/ (e.g. hard) and the high vowels /ʊː/ (e.g. who’d) and /ʊ/ (e.g. hood) (Fig. 4b).
- Fronting and rising of the vowel /ɜ:/ (e.g. herd) (Fig. 4c).
- The vowels /i:/ (e.g. heed), /ɛ/ (e.g. head) and /æ/ (e.g. had) in Australian are closer to each other (Fig. 4a).

With the vowels /æ/ and /ɛ/ in Australian being raised towards /i:/, the result is that /i:/, /ɛ/ and /æ/ are closer in Australian formant space than in the other two accents. In Fig. 4d, it can be observed that these vowel movements form a trend such that the front short vowels in Australian are more in the upper region of the vowel space. In addition, the vowel /ɜ:/ in Australian is more closed (has a lower \(F_1\)) and more fronted (has a higher \(F_2\)) compared to British and American /ɜ:/ The noticeable fronting of /ɑ:/ in Australian makes /ɜ:/ the only long back vowel in Australian as shown in Fig. 4d. The American /a/ is slightly more open (has a

![Fig. 3. Average formant frequencies of British, Australian and American accents.](image-url)
higher \( F_1 \) compared to the British /\alpha/ and the American /\alpha/ is centralized compared to British and Australian accents as shown in Fig. 4e. One of the most striking differences in the formant space of the three accents is that of the American /\alpha/ which is a much lower (i.e. has a higher \( F_1 \)) and more fronted (i.e. has a higher \( F_2 \)) compared to British and Australian.

### 4. Intonation correlates of accents

Some of the distinctive characteristics of an accent are carried by the differences in intonations primarily conveyed by the variations of pitch trajectory and duration. Note that pitch is the auditory sensation of the fundamental frequency of a periodic audio signal; however, for the purpose of pitch extraction and modelling, the terms pitch and fundamental frequency are used interchangeably in this paper.

The most noticeable variation of pitch with accent, in regional British English accents, is the extensive use of final rising tones (where the pitch at the end of a declarative sentence is raised instead of falling as in RP) in many northern UK cities such as Birmingham, Liverpool, Glasgow, Newcastle and Belfast (Wells, 1982). American and Australian English are marked by a high rise tone beginning on the final accented syllable near the end of the statement and continuing to increase in frequency to the end of the intonational phrase (Wells, 1982). Cruttenden (1997) points out that the initial rise or the final fall/rise in pitch (i.e. the first or the last pitch accent) is an indicator of the differences in accents. For example, in Northern Ireland accent the pitch intonation curves tend to rise at the end of each sentence in some cases whereas in received pronunciation (RP) British English accent the end of a sentence is partly signaled by a fall in the pitch (Wells, 1982; Grabe and Brechtje, 2004).
4.1. Pitch trajectory estimation

The pitch trajectories are estimated from a sequence of candidate pitch values extracted from the peaks of the autocorrelation function, or equivalently the spectral envelope of speech frames. First, speech is segmented into overlapping windowed frames with a duration of 25 ms long and with an overlap of 15 ms between successive frames (Young et al., 2002). A speech frame is marked voiced if its autocorrelation value and the energy value remain higher than a threshold for three successive frames. The pitch contour for successive voiced segments is obtained from a Viterbi-type decoding of the trajectories of the candidate pitch values (Hirose et al., 1992; Fujisaki et al., 1987).

4.2. Pitch intonation models

Modelling the pitch intonation curve requires a symbolic representation of the elementary pitch events and methods of coding and decoding of the associated pitch contour signals. Popular models – such as Tones and Break Indices (ToBI) (Silverman et al., 1992), Rise–Fall-Connect model (RFC) (Taylor, 1994), Tilt (Taylor, 1998), Momel (Hirst et al., 2000) and Fujisaki’s superposition models (Fujisaki, 1988) – define a set of elementary symbols or parameters associated with the rise, fall, peaks and troughs of the pitch curve and associate these with significant phonological and phonetic events in the intonation curve. Similarly, in this paper we use a set of simple pitch intonation parameters to describe the broad characteristics of the pitch intonation curve as discussed next.

4.3. A pitch intonation model for accent

We have chosen the rise/fall/connection (RFC) model Taylor, 1994 for pitch intonation because of its simplicity in modelling the differences of the slopes of intonation curves at the positions and contexts where it correlates with accent. An RFC model represents a pitch contour using a sequence of rising and falling pitch segments with straight lines applied for the intervals without pitch values (unvoiced segments). The shape and the intensity of rise or fall in a pitch trajectory are modelled by the Legendre polynomial. The set of parameters that are considered essential for the modelling of broad intonation patterns include the pitch range, the rate of initial pitch rise in an intonation utterance, the rate of final pitch fall and the average pitch slope across an utterance Fig. 5. These intonation correlates of accent are described in more detail in the following paragraphs.

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

\[
\text{range}(F_0) = F_0 \pm 3 \times \text{std}(F_0)
\]  

(3)

![Fig. 5. A model of the broad pitch intonation characteristics of accent employed in accent modelling. Thick line is the pitch contour. The symbol ‘/’ represent boundaries of rise and fall.](image-url)
where std($F_0$) is the standard deviation of $F_0$ and $\overline{F}_0$ is the average pitch, the latter is not an indicator of accent as it is mainly determined by the physiological characteristics of speakers. The choice of $\pm 3 \times \text{std}(F_0)$ would cover about 99% for a Gaussian distributed variable. This is a reasonable choice even though the pitch distribution is not quite Gaussian.

*Pitch contour slope* ($F_{0\text{slope}}$) is the overall trend of change in the pitch frequency over an utterance. Pitch contour slope is indicated by pitch reference line in Fig. 5.

*Pitch peaks/troughs envelope lines*. These lines are obtained from a first order polynomial model of the lines that envelop the maxima and minima of the pitch curves Fig. 5.

*Pitch reference line* is the base frequency line that a speaker usually returns to after a high or a low pitch excursion. It is obtained as the mean line between the pitch peaks/troughs envelope lines.

*Rate of initial pitch rise* ($\partial F_{0I}$) at the beginning of a pitch contour. It is described by the slope of the $F_0$ change at the initial pitch segment, i.e. the first pitch accent, at the beginning of the utterance. This slope is measured from the beginning of the segment to the peak point of the first pitch accent.

*Rate of final pitch fall or rise* ($\partial F_{0F}$) at the terminal pitch segment of pitch contour. The final rise or fall in pitch depends on the discourse, context or pragmatics, e.g. a rise terminal may convey an intention to continue speaking or may signal a question. The rate of final pitch fall is described by the slope of the $F_0$ change at the end of the utterance. It is measured from the last peak in pitch contour to the end of the pitch utterance. It is worth noting that factors such as attitude and emotion can also affect the intonation contour although in the experiments here only declarative reading speeches are employed.

### 4.4. Comparison of pitch intonation of British, Australian and American accents

Table 1 presents the estimates of the average pitch frequency and the pitch range for male and female speakers across British, Australian and American accents. The number of speakers used is given in Section 1. All spoken sentences in the database were declarative. The average pitch frequency is a speaker-dependent parameter that varies significantly across speakers. As expected, no statistically significant differences is observed in the average value of $F_0$ across the three accents in Table 1. For example, the difference between the average pitch value of Australian and American is around 1%, which is a small fraction of their standard deviations as shown in Table 1.

Table 1 also shows that the pitch frequency range, as measured by standard deviation, is affected by accent and gender. Female speakers have a larger pitch range than male speakers. Both British male and female speakers have the largest pitch range among the three accents. On average the pitch frequency range of British speakers is 41% and 43% larger than those of Australian and American respectively. Australian and American speakers have a similar pitch range. It appears that the pitch frequency range is closely related to the intonation patterns employed by different accents. British speakers tend to employ extensive usage of low-rise tone in non-final intonation groups (tone unit) to form an oratorical and formal speaking style whereas Americans prefer the high-rise tones, in what is considered to be a more causal speaking style. There also appears to be an increasing tendency of using high-rise tones among Australians (Cruttenden, 1997).

Table 2 presents the variation of pitch utterance slope. Female speakers consistently have steeper pitch contour slope than male speakers regardless of their accents. It can be noted that the American accent has a slightly steeper utterance slope than the British and the Australian accents, with a difference of 6% and 1% respectively.

Table 6 displays the initial slopes of pitch rise of the British, the American and the Australian accents. It is noticeable that female speakers consistently have steeper pitch slopes in the initial rising part of pitch contour.

<table>
<thead>
<tr>
<th>Accent</th>
<th>Female (avg.pitch ± 3 × std)</th>
<th>Male (avg.pitch ± 3 × std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>British</td>
<td>184 ± 3 × 37.30</td>
<td>121 ± 3 × 29.93</td>
</tr>
<tr>
<td>American</td>
<td>194 ± 3 × 21.21</td>
<td>119 ± 3 × 17.98</td>
</tr>
<tr>
<td>Australian</td>
<td>195 ± 3 × 21.78</td>
<td>121 ± 3 × 17.17</td>
</tr>
</tbody>
</table>
than male speakers. British speakers have the steepest rate of initial pitch rise among the three accents; about 31% and 21% steeper than Australians and Americans respectively. This coincides with the observation that they have the largest frequency range of the three accents. Australian speakers have the lowest final fall rate in pitch compared to British and American speakers (Table 2). The results illustrates that British speakers tend to have a steeper pitch rise and pitch fall rates than American speakers. Furthermore, American speakers tend to have a lower pitch in final parts of sentences compared to British speakers (Yan and Vaseghi, 2002).

The method described above is a relatively simple model for investigation of the effect of accents on pitch trajectory. Much more extensive investigations on the suitability of different intonation models for the modelling and synthesis of pitch intonation of accents are needed.

5. Duration and speaking rate correlates of accent

Recent studies of English accents suggest that phoneme duration patterns and speaking rates are two important features of an accent. For example Watt and Ingham suggest that a noticeable feature of Scottish English accent is the lengthening of the vowel duration (Yan and Vaseghi, 2002). Cox presents an extensive investigation of durations of Australian vowels (Watt and Ingham, 2000). Angkitirakul (Cox, 2006) found that the durations of some phonemes were sensitive to accents. In this paper, the statistics of phoneme durations and the speaking rates of British, American and Australian English speakers are estimated using phoneme boundaries obtained from a segmentation process based on HMM-based Viterbi decoder from the HTK software (Young et al., 2002). These duration statistics are used in subsequent sections as part of the input parameters for accent synthesis.

Note that in the study of the dependency of the variations of durations and speaking rates on accents, as in the study of other speech features, a question arises as to what extent the observed variations can be attributed to the variations of speakers or phonetic environments. The results presented here average-out the effect of phonetic environments since we use phonetically-balanced databases with many examples of each phone. Note that in a phonetically-balanced database the frequency of occurrence of the phonemes is approximately equal to their actual value. Similarly, the results are gender-dependent and averaged on a number of different speakers. Furthermore, the style of speech across all accents is declarative and read (as opposed to conversational) speech.

5.1. Phoneme duration pattern

Figs. 6 and 7 show the average phoneme durations of female and male speakers from British, Australian and American accents. The duration of vowels and consonants are shown on separate graphs. The average phoneme duration of female speakers are generally longer than those of male speakers. From the phoneme duration analysis results, Australian accent has consistently longer average vowel duration and consistently

<table>
<thead>
<tr>
<th>Accent</th>
<th>Avg(F)</th>
<th>Std(F)</th>
<th>Avg(M)</th>
<th>Std(M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch contour slope (Octs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>British</td>
<td>0.1597</td>
<td>0.06</td>
<td>0.1318</td>
<td>0.08</td>
</tr>
<tr>
<td>Australian</td>
<td>0.1675</td>
<td>0.08</td>
<td>0.1347</td>
<td>0.1</td>
</tr>
<tr>
<td>American</td>
<td>0.1698</td>
<td>0.08</td>
<td>0.1391</td>
<td>0.07</td>
</tr>
<tr>
<td>Pitch rise rate (Octs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>British</td>
<td>0.66</td>
<td>0.4</td>
<td>0.55</td>
<td>0.4</td>
</tr>
<tr>
<td>Australian</td>
<td>0.45</td>
<td>0.3</td>
<td>0.43</td>
<td>0.3</td>
</tr>
<tr>
<td>American</td>
<td>0.52</td>
<td>0.4</td>
<td>0.44</td>
<td>0.3</td>
</tr>
<tr>
<td>Pitch fall rate (Octs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>British</td>
<td>1.59</td>
<td>0.9</td>
<td>1.56</td>
<td>0.7</td>
</tr>
<tr>
<td>Australian</td>
<td>1.55</td>
<td>0.6</td>
<td>1.54</td>
<td>0.8</td>
</tr>
<tr>
<td>American</td>
<td>1.61</td>
<td>0.4</td>
<td>1.57</td>
<td>0.6</td>
</tr>
</tbody>
</table>
shorter consonant duration compared to American and British accents. The duration of vowels and diphthongs of female Australians is on average 21% and 14% longer than those of British and American accents respectively. In particular, diphthongs such as /ʌ:/ and /ɔ:/ in Australian are over 30% longer than those in British, and some 28% longer than those in American. For female speakers, Australian consonants are 13% and 16% shorter than those of American and British respectively.

The duration results for British and American accents do not show significant differences in vowel duration except at the start and end of the sentences (Fujisaki, 1988). British speakers have shorter vowels durations at the start and the end of sentences compared to American speakers. In contrast to the vowel duration pattern, the average durations of consonants across the three accents show a different picture. Male speakers across the three accents exhibit similar patterns of dependency of duration on accent. The differences in the duration pattern of phonemes are reflected in the differences in the speaking rate as discussed next.

### 5.2. Speaking rate variation

Speaking rate is measured in terms of the average number of phonemes per second. Table 3 and Fig. 8 shows the average speaking rates and the distributions of British, Australian and American accents. A speaking rate sample value is calculated for each sentence. Note from Fig. 8 that the distribution of the British speaking rate is significantly shifted relative to those of American and Australian. Although the peaks of the distributions of the Australian and American speaking rates seem to coincide, they have different shapes and hence different means. The average speaking rates of Australian and American are respectively 23% and 15% lower than that of British English respectively. The fact that Australians have the lowest speaking rate also conforms to the observations regarding the elongation of durations of the Australian vowels (Watt and Ingham, 2000). We should note at this point that the experimental results on the variations of durations

### Table 3

<table>
<thead>
<tr>
<th>Accent</th>
<th>Speaking rate phoneme (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>British</td>
<td>12.1</td>
</tr>
<tr>
<td>American</td>
<td>11.6</td>
</tr>
<tr>
<td>Australian</td>
<td>10.8</td>
</tr>
</tbody>
</table>
with accents are derived from speaker-independent databases with the speech produced in a declarative ‘news reading’ style from a prescribed text. Furthermore, there is a good level of consistency in the results concerning both genders across the accents.

6. Accent modification and evaluation

In this section the perceptual influence of modification of the parameters of formants, pitch and duration on conveying accents are assessed through accent transformation of British, Australian and American spoken sentences. The modification of accent is accomplished by a voice morphing software developed by the authors.

6.1. Transformation of acoustic correlates of accents

Two sets of accent transformation experiments were performed: accent transformation from Australian to British accent and from British to American accent. A non-uniform spectrum warping (Yan et al., 2007) is employed to move the formant tracks of the vowels of the source accent speakers (e.g. Australian) towards the expected values of the formants of vowels of the target accent (e.g. British). In total four formants are mapped from the source accent to the target accent using the method described in our previous paper (Yan et al., 2007). We found that the formant mapping is also able to perform phoneme substitution across accents (e.g. transform a vowel such as /æ/ in accent A to a different vowel such as /ɔ/ in accent B).

The broad features of the pitch intonation pattern are modified by changing the pitch frequency range, the pitch contour slope, the rate of initial pitch rise and the rate of final pitch fall using the time domain pitch synchronous overlap and add (TD-PSOLA) (Moulines and Charpentier, 1990) method. The modification of the duration patterns of phonemes from those of the source accent to those of the target accent is accomplished by appropriate insertion or deletion of pitch periods using TD-PSOLA. The required amounts of change in duration patterns are obtained from the ratios of the average durations of the corresponding phonemes in the source accent to those in the target accent. In order to minimize signal processing artifacts, the TD-PSOLA stages for the modifications of pitch intonation and duration patterns are combined into a single signal processing stage.

6.2. Perceptual tests

An experiment was conducted to use the mean opinion score (MOS) (Kim and Chang, 2000) to quantify the influence of each acoustic correlate of accents. Seven listeners with British accent are used as listeners. The test
Each listener is given seven sets of sentences from different speakers via headphones. The listeners were allowed to listen to speech samples more than once and no fixed time has been set for their responses. In each set, there are five sentences (A B C D E), one of which is from a British speaker, one from an Australian speaker and the remaining three sentences are speech that were transformed from Australian to British by modification of either the formants or the pitch or the duration of the source accent towards the expected values in the target accent. All sentences in each set have the same text and the order of the presentation of the sentences in each set, to different listeners, is changed randomly. The randomization of the presentation order of the sentences in the test sets is intended to reduce possible prediction of accents from the order of the accents and sentences in the test set. In addition, listeners are given two known accent sentences: a source Australian accent speech S and a target British accent speech T. Both sentences have different text from those of test sentences. The listeners are asked to give similarity scores between the given sets of speech (sentences A B C D E) and the target accent speech (sentence T) and source accent speech (sentence S). The similarity score is in the range between 10 (identical) and 0 (completely different). The listeners are also asked to rate the naturalness of each test sentence. The score of naturalness is in the range from 1 (poor) to 5 (excellent). The two types of ratings are conducted in the same tests. Before the test starts, each listener is given a practice session that involves instructions on how to use the system and a number of samples sentences from the source and target accents. Subsequently the listeners are provided with seven sets of sentences for the formal rating tests. In total 49 ratings are collected. Similar tests are conducted on British and American accent pairs.

The evaluation results are shown in Table 4. It can be seen that for British and American accent pairs, the modification of the formants of speech is more influential in affecting change of accent than the modification of duration and intonation patterns. However, in Australian to British accent conversion, Australian speech, after modification of duration or pitch intonation, is considered to have slightly more British accent compared to that obtained after modification of formants. In both cases, changing duration pattern alone exhibits little impact on accent.

In the next experiment, accent conversion is accomplished by transforming the three correlates of accent together. ABX (Stylianou et al., 1998) and MOS tests are designed to further assess the influence of the correlates of accents. An ABX test involves a source speaker A, a target speaker B and a transformed version of the source speaker’s voice X. The aim is to test how perceptually close the transformed voice X is to the target voice B. In the experiments described here it is the accent of the transformed voice that is compared to the accent of the target speaker’s voice. In the ABX test, ten listeners are involved and all of them are British native speakers or fluent British speakers.

The source, target and transformed speech sentences are presented to listeners through headphones. The listeners were allowed to listen to speech samples more than once and no fixed time has been set for their responses.

Table 4

<table>
<thead>
<tr>
<th>Accents</th>
<th>British accent</th>
<th>American accent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>British accent speech → American accent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original British speech</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>After modification of formants</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>After pitch intonation modification</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>After duration modification</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Original American speech</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accents</th>
<th>Australian accent</th>
<th>British accent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Australian accent → British accent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Australian Speech</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>After modification of formants</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>After pitch intonation modification</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>After duration modification</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Original British Speech</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

**Scoring scheme:** 9–10: Identical; 7–8: very similar; 5–6: fairly similar; 3–4: a little similar; 1–2: fairly different; 0: very different.
responses. The listeners are then asked to indicate if the transformed voice has a similar accent to the target accent or to the source speaker’s accent. Each listener is given a practice session that involves instructions on how to use the system and a number of sample sentences from the source and target accents. Subsequently, the listeners are provided with seven sets of $ABX$ sentences for the formal rating tests. Note here that the speech samples $A$ and $B$ are taken from two sets of seven different speakers with source ($A$) and target ($B$) accents. The speech sample $X$ is the transformed source speech. The source, the transformed and the target speech have the same text contents. In cases where listeners decide that the accent of some transformed sentences $X$ is not similar to either the source $A$ or the target $B$ then those results are excluded. The order of presentation of the source accent sentence $A$ and the target accent sentence $B$ in each $ABX$ set are randomized. The order of the $ABX$ sentences presented to different listeners is also random. The randomization of the presentation order of the sentences in the test sets is intended to reduce possible prediction of accents from the order of presentation of the accents and the sentences in the test sets. In total 67 test ratings are collected from the listeners.

Similar $ABX$ tests are conducted for accent conversion experiments between British and American accents. In this work the selection of test sentences for accent conversion between British to American English excludes the rhotic ‘r’ sound as much as possible. Where there are ‘lost’ ‘r’ as in “here” /hir/ in British English, it will be converted to /ir/ as in “hear” /hir/ in American English. Rhoticity will be considered in a future work. The results in Table 5 shows that after accent transformation some 71% of the listeners perceived that the British source speakers have acquired American accent. In experiments on conversion of Australian accent to British accent, some 78% of listeners’ answers indicate that, after accent transformation, the source speech have a similar accent to the target British accent. Note that since the original source accent sentence, and its converted accent version, are given in the $ABX$ test, one may expect a bias towards selection of the source, for this reason the accent conversion results are even more significant.

To further assess the influence of formants on accents, another MOS test was designed. Ten listening listeners are involved in this test and seven of them are native British speakers and the others are fluent British speakers. Seven sets of source, target and transformed speech samples are presented to listeners through headphones. The listeners were allowed to listen to speech samples more than once and no fixed time has been set for their responses. Each test set includes three sentences ($A B C$), two of which are original Australian and British accent speech and the third is the converted speech from Australian accent to British accent. The sentences in each set are presented in a random order to listeners to reduce possible prediction of accents from the order of the accents and sentences in the test set. In addition, two different known-accent speech sentences ($S$ and $T$) are given to the listening listeners. Sentence $S$ is the original Australian speech and Sentence $T$ is the original British speech. The listeners are then asked to give a score in terms of the similarity between the test speech sentences ($A B C$) and the known-accent speech sentences ($S$ and $T$). Again, the scores range from 0 for “very different” to 10 for “identical”. Listeners are also asked to give a score of naturalness for each sentence. The scores range from 1 (poor) to 5 (excellent). The two types of ratings are conducted in the same tests. Listeners are given a practice session to familiarize them with the rating scale. In total 70 ratings are collected. Similar MOS tests are conducted for accent conversion experiments between British and American accents. The results are shown in Table 6. It is clear that the transformed speech is perceived to carry a great deal of the accent of target speech although some source accent can still be heard.

In addition, listeners believe that accent conversion from Australian to British is more successful than that from British to American accent. One possible reason is that both Australian and British accents are non-rhotic accents where the phoneme ‘r’ occurring before a consonant or at the end of a word is not well pronounced while American is rhotic accent where the phoneme ‘r’ is well pronounced. The transformation between Australian and British accents will be relatively easier than that from a non-rhotic accent to a rhotic accent. The transformation of rhotic to non-rhotic sounds and vice versa is a subject of further research.

Table 5
Evaluation of accent conversion by $ABX$ Test.

<table>
<thead>
<tr>
<th>Accent conversion</th>
<th>$ABX$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian accent speech → British accent</td>
<td>78</td>
</tr>
<tr>
<td>British accented speech → American accent</td>
<td>71</td>
</tr>
</tbody>
</table>
Table 6
Evaluation of accent conversion by MOS test.

<table>
<thead>
<tr>
<th>Speech</th>
<th>Australian (source)</th>
<th>British (target)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Australian accent → British accent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original Australian speech</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Converted speech</td>
<td>3.3</td>
<td>7.6</td>
</tr>
<tr>
<td>Original British speech</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td><strong>Speech</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>British accent → American accent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original British speech</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Converted speech</td>
<td>3.1</td>
<td>6.8</td>
</tr>
<tr>
<td>Original American speech</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

Scoring scheme: 9–10: Identical; 7–8: very similar; 5–6: fairly similar; 3–4: a little similar; 1–2: fairly different; 0: very different.

7. Conclusion

A model of the broad pitch intonation pattern is used for capturing the differences in intonation across accents. The pitch intonation model includes: pitch frequency range, the slopes of initial rise and final fall in pitch and pitch utterance slope. Statistical evaluations indicate that the average pitch is not dependent on accent. British speakers have the largest pitch frequency range. Furthermore, Australian speakers tend to have the lowest pitch rise and fall rates among three accents. Australian vowels and diphthongs have longer duration than those of British and American accents while Australian consonants have shorter duration than those of British and American accents. In a set of accent morphing experiments, three correlates of accents, namely formants, pitch intonation and duration patterns, are transformed from the source accent to the target accent to assess their individual influence on accents. The results show that these acoustic correlates are effective in accent morphing. The transformation of the formant correlates changes the source accent most markedly. The applications of the accent analysis include accent recognition, accent compensation in automatic speech recognition and accent identification; there are subjects of further investigations.

Acknowledgement

We thank the UK’s EPSRC for funding the project GR/M98036. We would also like to thank the support of Hohai University Natural Science fund (2008428811), doctoral fund of Education Ministry of China (200802941007), Hohai University talent start Fund (20080604) and the special fund for the basic research in Chinese central managed universities.

References


