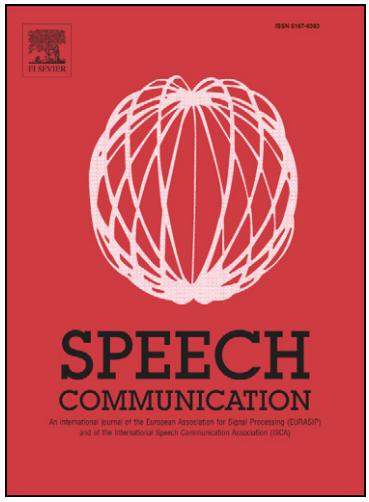
Cross entropic comparison of formants of british, australian and american english accents

Seyed Ghorshi, Saeed Vaseghi, Qin Yan



Cross entropic comparison of formants of british, australian and american english accents

Seyed Ghorshi, Saeed Vaseghi, Qin Yan

PII: S0167-6393(08)00043-5
DOI: 10.1016/j.specom.2008.03.013
Reference: SPECOM 1702

To appear in: *Speech Communication*

Received Date: 29 August 2006
Revised Date: 14 March 2008
Accepted Date: 25 March 2008

Please cite this article as: Ghorshi, S., Vaseghi, S., Yan, Q., Cross entropic comparison of formants of british, australian and american english accents, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.03.013

CROSS ENTROPIC COMPARISON OF FORMANTS OF BRITISH, AUSTRALIAN AND AMERICAN ENGLISH ACCENTS

Seyed Ghorshi *Saeed Vaseghi* *Qin Yan*

School of Engineering and Design, Brunel University, London

{Seyed.Ghorshi, Saeed.Vaseghi}@brunel.ac.uk

ABSTRACT

This paper highlights the differences in spectral features between British, Australian and American English accents and applies the cross-entropy information measure for comparative quantification of the impacts of the variations of accents, speaker groups and recordings on the probability models of spectral features of phonetic units of speech. Comparison of the cross entropies of formants and cepstrum features indicates that formants are a better indicator of accents. In particular it appears that the measurements of differences in formants across accents are less sensitive to different recordings or databases compared to cepstrum features. It is found that the cross entropies of the same phonemes across speaker groups with different accents (inter-accent distances) are significantly greater than the cross entropies of the same phonemes across speaker groups of the same accent (intra-accent distances). Comparative evaluations presented on cross-gender speech recognition shows that accent differences have an impact comparable to gender differences. The cross entropy measure is also used to construct cross-accent phonetic trees, which serve to show the structural similarities and differences of the phonetic systems across accents.

1. INTRODUCTION

This paper has two main objectives: (i) to reveal the differences across accents among the formants or cepstrum features of acoustic realisations of phonemic units of spoken English using analysis of large speech

databases and (ii) to investigate the use of cross entropy as an information measure for quantification of the objective differences among spectral features of English accents.

Since the effect of accent and speaker characteristics cannot be modelled in isolation or quantified individually, there is a need to develop a method for comparative evaluation of the effects of accent and speaker variables on speech parameters. In addition since the comparative modelling of several accents often involves the use of speech from different databases recorded by different equipments, there is also a need to explore the effect of different databases and recording equipment on the observed differences of acoustic features across accents. In this paper the focus is on the effect of accents, speakers and databases on formants and cepstrum features of phonetic speech units. The cross-entropy information measure is used as a metric for modelling the effects of accents on phonetic speech units. However, the same methodology can be applied more generally to measurements of the effects of accents on intonation, pitch and duration.

Accent is one of the most fascinating aspects of speech acoustics [1]. Accent variability within a language is a major source of limitation of the accuracy of un-adapted machine (and unaccustomed human) speech recognition systems [2-4]. The performance of large vocabulary continuous speech recognition systems fluctuates sharply with variations in accent, and deteriorates when the input speech contains a strong accent unseen in the database on which the speech models are trained [5]. Models of differences in accents can also be used for accents synthesis in text-to-speech synthesis [6], for accent morphing [7], accent identification [8, 9] and in the teaching/learning of the acoustics of pronunciation of an accent. Accent morphing, that is the changing of the original accent of a recorded voice to a different accent, has several applications including in text-to-speech synthesis (TTS) where the user can be given the option of selecting an accent for the TTS voice from a number of choices; in changing the accent of the voice of a character in a film or play or in computer toys/games and in multimedia systems.

The term *accent* may be defined as a distinctive pattern of pronunciation and intonation characteristics of a community of people who belong to a national, regional or social grouping. In Crystal's dictionary of linguistics [10], an accent refers to *pronunciation only* as "the cumulative auditory effect of those features of *pronunciation*, which identify where a person is from regionally and socially". Similarly, Wells [1] defines an English *accent* as a pattern of pronunciation used by a speaker for whom English is the native language or more generally, by the community or social grouping to which he or she belongs.

It is worthwhile clarifying the similarities and the differences between two closely linked linguistic terms, namely accent and dialect. The term dialect refers to the whole speech pattern, conventions of vocabulary, pronunciation, grammar, and the usage of speech by a community of people [1] while accent refers to a pattern of pronunciation and the abstract (phonological) representations which can be seen as underlying the actual (phonetic) articulation.

An accent may rely on the articulation of vowels, co-articulation, consonants, nasalisation, voice quality, melodic *clichés,* the particular forms of intonation, pitch range, pitch level at phrase boundaries, stress pattern, stress location differences or other shibboleths [1, 11].

Accents evolve over time, affected by a number of factors such as geographical variation, socio-economic classes, ethnicity, sex, age, cultural trends, the mass media and immigration. For example, the Australian accent is considered to have been influenced by waves of mass immigrations to Australia and in particular by London Cockney pronunciation [12], by Irish pronunciation and in relatively recent times by American pronunciation through migration and mass media. Similarly, the English Liverpool accent has been influenced by Irish immigration whereas the Northern Ireland accent has been influenced by Scottish immigration. Wells [1] provides an excellent introduction to the linguistic structures of the accents of English language within and beyond the British Isles.

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

- *Historical approach to accent development* compares the historical roots of accents and the evolutionary changes in sounds that accents have gone through as various accents merge or diverge. The historical approach compares the rules of pronunciation in accents and how the rules change and evolve over time.
- *Structural, synchronic approach,* first proposed by Trubetzkoy [13] models an accent in a system-oriented fashion in terms of the following systematic differences:
    - Differences in phonemic systems.
    - Differences in phonotactic (structural) distributions.
    - Differences in lexical distributions of words.
    - Differences in phonetic (acoustic) realisation.

In this work the differences between accents are modeled using a system-based approach as explained next.

1.1 Phonetics and Acoustics of Accents

Different accents of a language have distinct patterns of pronunciations and intonation. Generally the structural differences between accents can be divided into four main categories:

a) *Differences in the phonemic systems of accents*, i.e. in the number and/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 as described in Wells [1]. For example, in British Received Pronunciation (RP) the phoneme /ɒ/(*oh*) in *stop*, *dodg*e, *romp,* corresponds to /ɑ/(*aa*) in General American accent (GenAm) but the same phoneme /ɒ/(*oh*) in *cough*, *gone* and *Boston*, corresponds to /ɔ/(*ao*) in GenAm. Note that the symbols placed inside parentheses denote alphabet representation and the symbols placed outside parentheses denote the International Phonetic Alphabet (IPA) representation [14]. In British English RP the non-rhotic '*r*' is often changed into /ə/(*ax*) (schwa), giving rise to a new class of falling diphthongs: /ɛə/(*ea*) /iə/(*ia*) /uə/(*ua*), which means that for example the British diphthong /ɪə/(*ia*) most often corresponds to the GenAm two phoneme sequence /ɪr/ (e.g. the word *hear* is /hɪə/ in British RP and /hɪr/ in GenAm). Furthermore, the *er* sound of (stressed) *fur* or (unstressed) *butter* in British RP is realised in GenAm as a monophthongal *r*-coloured vowel, eg. *bird* is /bɝd/ in GenAm and /bɜːd/ in British RP and *dinner* is /dinɚ/ in GenAm and /dinə/ in British RP. Note that a distinction also exists among some American speakers in the following pairs: *purest* [ə] ~ *purist* [i] and *Lennon* ~ *Lenin* [1], *brewed* [ʉ] ~ *brood* [u] and *lute* ~ *loot* [15], *morning* [ɔ] ~ *mourning* [o] [6], *Mary* [e] ~ *merry* [ɛ] (in the Eastern United States) [16].

Australian English has three recognised varieties [12]. Broad Australian English is the most recognisable variety of Australian accent, it is characterised by a perceived drawl and by the prevalence of diphthongs. General Australian English is the accent used by the majority of Australians and it dominates the accents found in contemporary Australian-made films and television. This variety has noticeably shorter vowel sounds than Broad Australian English.

Cultivated Australian English has many similarities to British Received Pronunciation but it is now spoken by less than 10% of the population [17, 18]. Australian English has a number of vowels with significantly different realisations and phonetic quality such as /æi/ instead of /ai/ and /æɔ/ for /au/ [1, 12].

b) *Differences in the phonotactic distributions of accents.* Accents may differ in the environments (i.e. the contexts) in which phonemes occur. A prime example of this 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 realisation in a wide variety of contexts.

c) *Differences in the lexical realisations of words including phoneme substitution, deletion and insertion* [5, 14]. For example, *tomato* is pronounced as /tmeɪtəu/ in most American English accents but as /tmɑːtəu/ in British RP accent and the word *immediate* is pronounced as /ɪmiːdʒit/ in British RP and General American accents but it is pronounced as /əˈmiːdiːət/ in Australian accents [1].

d) *Differences in the phonetic realisations (acoustics) of accents.* The acoustic differences between accents of a language are due to the differences, during the realisation of sound, in the configurations, positioning, tension and movement of laryngeal and supra-laryngeal articulators. The differences due to articulation are manifested in the differences in the formant trajectories, pitch trajectory, intonation nucleus and duration parameters [1, 19]. The following four aspects of the acoustic correlates of accents are considered essential in the modeling of an accent: (1) vowel formants [20], (2) consonants [1] (3) intonation [19], (4) duration and speaking rate [21]. Note that the rank of acoustic correlates of accents depends on the accent, for examples for some accents intonation may be a more influential feature than in others. As far as the role of intonation is concerned, a number of studies demonstrated its importance in language variety or accent identification [19, 22, 23, 24, 25]. For instance, Australian English and Northern Ireland English may be partly characterised by a high-rising contour on utterances, which are grammatically statements. In many contexts, sentences end with a questioning tune, as if the speaker were asking for feedback from the listener [26].

In this paper we mainly focus on the comparison of the influence of vowels on accents. The motivation for this choice is due to the observation that vowels can be parameterised and represented in the formant spaces such as the well-known F1/F2 space. However the same methodology can be applied to other correlates of vowels or consonants.

1.2 Paper Outline

The remainder of this paper is organised in the following format. Section 2 describes the databases, features and models used for the training and evaluations in this paper. Section 3 describes a method for modeling and estimation of formants and their trajectories using phoneme-dependent hidden Markov models (HMMs) [20]. Section 4 presents a comparative analysis of the formant spaces of British, American and Australian English accents. Section 5 introduces cross entropy and presents comparative evaluation results of cross entropy analysis of cepstrum and formant features across accents. Section 6 uses cross entropy for phonetic tree clustering and introduces the concept of cross-accent phonetic-tree analysis. Section 7 compares the effects of gender and accent on speech recognition and finally Section 8 concludes the paper.

2. DATABASES, SPEECH FEATURES AND STATISTICAL MODELS

The databases used for accent analysis in this paper are Wall Street Journal (WSJ) and TIMIT databases for General American English accent, Wall Street Journal Cambridge University (WSJCAM0) database for British English accent and Australian national database of spoken language (ANDOSL) for Australian English accent. The use of two different databases for the American accent allows an assessment of the influence of different databases and recordings on observed differences in spectral features of accents.

The large-vocabulary continuous speech WSJ database was initiated during 1991 by the DARPA spoken language program. The WSJ0 and WSJ1 corpora consist primarily of read speech with texts drawn from a corpus of Wall Street Journal. The texts were selected to fall within either a 5,000 or a 20,000-word subset of the WSJ corpus. Subjects were recruited from the MIT community. An attempt was made to balance the speakers by sex, dialect, and age, particularly for the latter two groups, since the total number of speakers in these groups is relatively small. The subjects' age ranged from 17 to 52 years. Their regional affiliations were California, Utah, Colorado, Wyoming, Iowa, Missouri, Illinois, Mississippi, Alabama, Georgia, Michigan, Ohio, Pennsylvania, Maine, Vermont, New Hampshire, New York, Massachusetts, Rhode Island, Connecticut, New Jersey, Maryland, Virginia, and North Carolina. The subset of WSJ

database used here for modeling American English is from the Sennheiser recordings and contains 36 female and 38 male speakers with 9438 utterances.

The DARPA TIMIT continuous speech database was designed to provide acoustic phonetic speech data for the development and evaluation of automatic speech recognition systems. It consists of utterances of 630 speakers that represent the major dialects of American English. The corpus includes 438 males and 192 females. The talkers were each assigned one of eight regional labels to indicate their dialect as: New England, North, North Midland, South Midland, South, West, New York City, or Army Brat.

The WSJCAM0, developed at Cambridge University, is a native British English speech corpus for large vocabulary continuous speech recognition based on the texts of WSJ0. In addition to standard orthographic transcripts, WSJCAM0 also includes information on the time alignment between the sampled waveform and both words and phonetic segments. There are 90 utterances from each of 92 speakers that are designated as training set. An additional 48 speakers each read 40 sentences containing only words from a fixed 5,000-word vocabulary and another 40 sentences using a 64,000-word vocabulary, to be used as testing material. The age distribution of training and test speakers is from 18 to above 40 years old. The subset of WSJCAM0 of British English used here contains 40 female and 46 male speakers with 9476 utterances.

The Australian National Database of Spoken Language was prepared at Sydney University, the National Acoustic Laboratories, Macquarie University and Australian National University. The database of native speakers of Australian English used here consists of 36 speakers in each of the three categories of General, Broad and Cultivated Australian English. Each category is comprised of 6 speakers of each gender in each of three age ranges (18-30, 31-45 and 46+). Each speaker contributed in a single session, 200 phonetically rich sentences. The subset of ANDOSL we used is from broad and cultivated Australian accents and comprises 18 female and 18 male speakers with a total of 7200 utterances.

For speech segmentation and labeling, left-right HMMs of triphone units are employed and the Viterbi decoder is applied in the forced-alignment mode [5, 27] with phonemic transcriptions supplied. Each HMM has three states and in each state the probability distribution of speech features is modeled with a Gaussian mixture model with 20 Gaussian probability density function (pdf) components. The speech feature vectors used to train hidden Markov models consist of 39 features including 13 Mel-Frequency Cepstral Coefficient (MFCCs) and their 1^{st} derivative (velocity) and 2^{nd} derivative (acceleration) features. The multi-

pronunciation dictionaries used in this work include the BEEP dictionary [28] (British accent), the Macquarie dictionary [29] (Australian accent) and the CMU dictionary [30] (American accent).

3. FORMANT MODEL ESTIMATION

Speech formants carry information regarding phonemic labels [31], speaker identity and accent characteristics [32]. Although formant analysis has received considerable attention and a variety of automated approaches [33, 34] have been developed, the estimation of accurate formant features from the speech signal is a non-trivial problem that attracts continued research.

For formant trajectory estimation we employ a method proposed by Ho [20]. This method has been compared to a dataset of ground truth formant values obtained from manually derived and corrected formant trajectories. The results show that the formant trajectory estimation method produces highly accurate and reliable results [35, 36]. Further evaluations of this method and comparison of its performance with other formant estimation methods are presented in [37, 38]. The formant estimation method is briefly described in this section.

3.1 Formant Feature Extraction

In this work, formants are obtained from trajectories of the poles of linear prediction (LP) model of speech. The poles of the LP model of speech are associated with the resonant frequencies, i.e. the *formants,* of speech. The resonant frequency of each significant pole of an LP model of speech is a formant candidate. The pole angle relates to the resonant frequency. The pole radius relates to the concentration of local energy and the bandwidth of the spectral resonance. For each speech frame, a candidate formant feature vector is extracted from the poles of an LP model of speech. The formant feature vector $\boldsymbol{v}_k$ defined as

$$\boldsymbol{v}_k = [F_k, BW_k, M_k, \Delta F_k, \Delta BW_k, \Delta M_k] \quad (1)$$

where F_k, BW_k and M_k are the resonant frequency, bandwidth and magnitude of the k^{th} pole of the LP model respectively and ΔF_k, ΔBW_k, and ΔM_k are the slopes of the time trajectories of F_k, BW_k and M_k respectively. Depending on the speaker characteristics and the phonemes, typically voiced speech signals have five or six formants spanning a frequency range of 0-5 kHz.

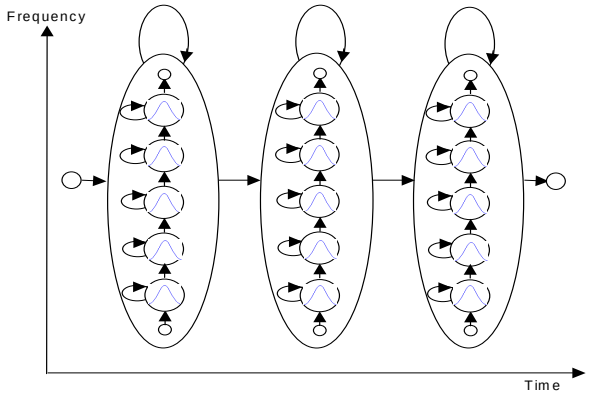


Figure 1: Illustration of a 2-D HMM model of the trajectories of formants of a phonetic segment of speech. Each state along time represents a subphonetic segment. Each state along the frequency represents a formant.

3.2 Hidden Markov Model of Formants

A hidden Markov model (HMM) [27] is an appropriate structure to model the probability distribution of formants. Figure (1) shows a phoneme-dependent formant model, based on a 2-D HMM with three left–right subphonetic states across time and five left–right formant states across frequency. 2-D HMMs are described in detail in [20] and applied in different speech applications [39-41]. Along the time dimension, each state of a 2-D HMM models the temporal variations of formants in a subphonetic state. Along the frequency dimension, the k^{th} state of the HMM models the distribution of the k^{th} formant. Note that along the frequency dimension of the 2-D HMMs there is no actual physical transition of formants from one formant-state to another formant-state, i.e. the formants exist concurrently at all states of the HMMs along the frequency dimension. However, the left-right HMM structure along the frequency dimension imposes a sequential constraint that allows constrained classification of the poles of LP models (sorted in the order of the increasing frequency) associated with the formant states of HMMs. This is necessary because in order to obtain a good LP fit, the order of the LP model is often set higher than twice the number of expected formants; usually an LP model order of 13 or more is used.

For most speakers, five states along the frequency dimension of formant HMMs are sufficient to represent the number of significant formants in speech, although some speakers may exhibit a sixth formant. The 2-D HMM of formants is subsequently used to extract formant trajectories from the poles of the LP model of speech segments.

Given a set of observations of the resonant frequencies $\boldsymbol{O}_n$, the maximum likelihood estimate of the associated formants is obtained, using Viterbi decoding, as

$$[\hat{F}_1, \hat{F}_2, ..., \hat{F}_N] = \arg\max_{F_1, F_2, ..., F_N} P(\boldsymbol{O}_n, [F_1, F_2, ..., F_N] \mid \Lambda_m) \quad (2)$$

where $\boldsymbol{O}_n$ is obtained from the poles of an LP analysis of a frame of a speech phoneme and sorted in terms of increasing frequency, Λ_m is HMM of the formants of phoneme m and N is the number of formants.

Using a set of formant training data, the distribution of each formant feature vector in each HMM state is modeled by a multivariate Gaussian mixture pdf trained via the Expectation Maximization (EM) algorithm [42]. The HMM states span the frequency axis such that the first state corresponds to the lowest frequency (first) formant and the last state corresponds to the highest frequency formant. For each state a Gaussian mixture model with four mixture components is used to model the distributions of the frequencies of the poles of LP models of speech segments.

There are a number of factors that may significantly affect the variance and the accuracy of LP-based formant estimation [43], these include:

- The LP model order,
- The influence of pitch on the observation and estimation of the first formant,
- Rapid formant variation that may occur in consonant-vowel transitions or diphthongs,
- Merging of neighbouring formants,
- Source vocal-tract interaction and the influence of the glottal pulse spectrum,
- Effects of lips radiation and internal loss on formant bandwidth and frequency.

To improve formant estimation five processing rules are applied as follows:

(a) A pre-emphasis filter is used to mitigate the effect of the spectral peak due to the pitch on the estimation of the first formant.

(b) Very short phonetic segments of duration less than a threshold of 4 frames equivalent to 40 ms, which may have excessive co-articulation of formants of neighbouring phonemes, are discarded.

(c) To further limit the effects of co-articulation, only formant candidates from speech frames within the central part (i.e. 50% of the speech segment around the centre) of phoneme segments are used.

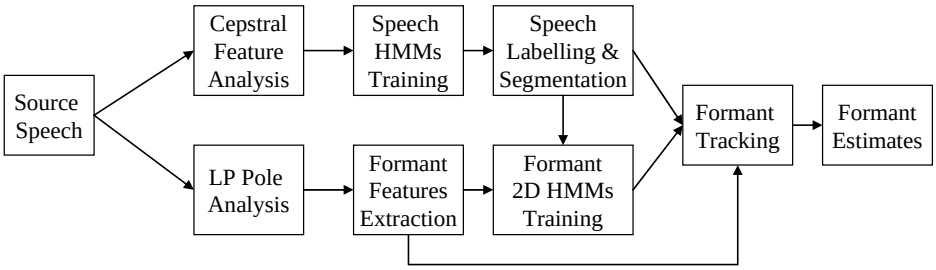


Figure 2: A block diagram illustration of the formant estimation method.

(d) Lower limits are placed on the bandwidths of the formant candidates (i.e. poles of the LP model) to avoid over-modeling of speech and the consequent adverse influence from inclusion of insignificant poles with large bandwidth.

(e) After the training of formant HMMs, in each state the mixture component with the largest variance is discarded. Large variance mixture components are associated with the values of pole frequencies that fall in between two successive formant frequencies. The probability weights of the remaining mixtures are then proportionally rescaled.

The formant estimation procedure is shown in Figure (2). Figure (3) shows comparisons of histograms and HMM pdfs of formants for the vowel /iː/(*iy*) from a female Australian speaker. Formant distributions obtained from the Gaussian mixture pdfs of 2-D HMMs are marked by the dash-dot line in Figure 3(d). The

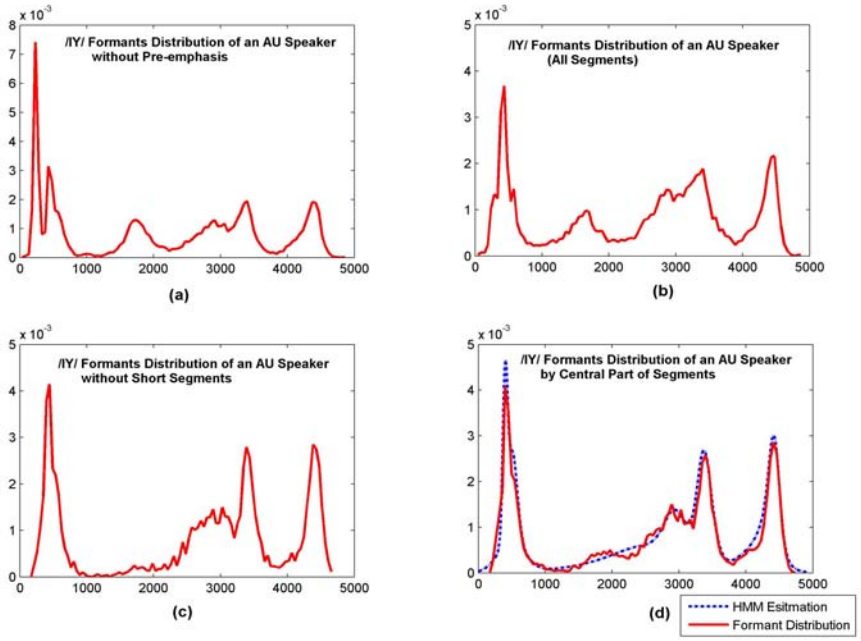


Figure 3: (a)-(c) Formant histograms showing the effect of improvement of the estimation method: (a) without pre-emphasis, (b) with pre-emphasis, (c) discarding short segments and using limits on bandwidth and LP order, (d) using the central part of segments, this plot illustrates a comparison of the histogram (solid line) and HMM of formants (dash dot line) of *iy* from a female Australian speaker.

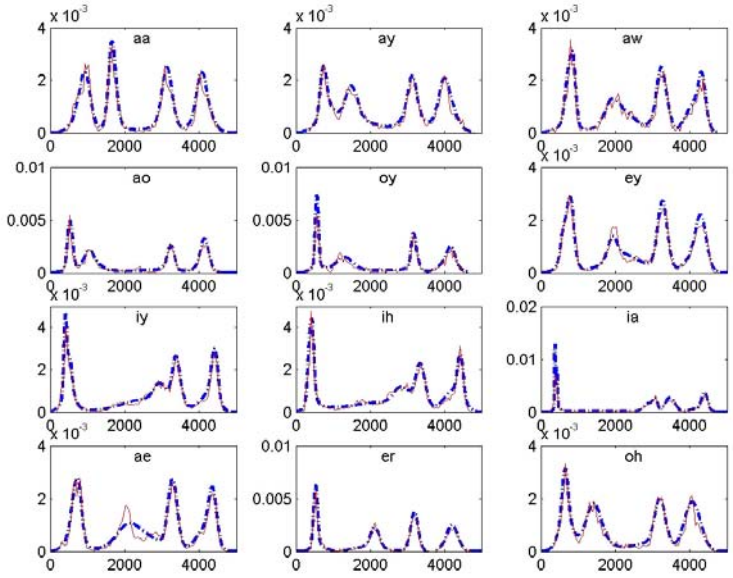


Figure 4: Superposition of Gaussian probability density functions and histograms of formants of vowels from an Australian Speaker. Note, Gaussian probabilities are extracted from HMMs of formants.

frequency at each peak of the distribution represents a formant. It can be noted from Figure 3(a) that in the absence of a pre-emphasis filter the spectral peak due to the vibrations of the glottal folds could be mistaken for the first formant (which is in fact the second peak), while in Figure 3(b) the effect of the spectral peak due to the vibrations of the glottal folds is eliminated through the use of a pre-emphasis filter. In Figure 3(a-c) the hump around 1700Hz is easily mistaken for the 2nd formant although the phoneme /iː/(*iy*) does not have a formant in that frequency range. After applying signal processing rules (b)-(e) described above, the hump disappears in Figure 3(d) and the second and third formants can be seen clearly.

Figure (4) shows histograms of the formant distributions for all vowels and diphthongs in broad Australian together with Gaussian mixture models obtained from 2-D HMMs. The peaks of estimated Gaussian pdfs and histograms, which occur at the formant frequencies, coincide. The close match between the histograms of formant candidates of a phoneme and the corresponding Gaussian models from HMM states indicates that HMMs are good models of the distributions of formants.

3.3 Formant Trajectory Estimation

Formant HMMs are used for the classification of formant candidates and for estimation of the trajectories of formants. The 2-D HMM formant classifier may associate two or more formant candidates $F_{i(t)}$, with the same formant label k. In these cases formant estimation is achieved through minimisation of a weighted mean square error objective function [20] as

$$\widehat{F}_k(t) = \min_{F_k(t)} \sum_{i=1}^{I_k(t)} w_{ki}(t) \left[\frac{\left(F_{i(t)} - F_k(t)\right)^2}{BW_i(t)^2} \right] \quad (3)$$

where t denotes the frame index, k is the formant index, $I_{k(t)}$ is the total number of the formant candidates classified as formant k.

In Equation (3) 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 $w_{ki}(t)=P(F_i|\lambda_k)$ where λ_k is the Gaussian mixture pdf model of the k^{th} state of a phoneme-dependent HMM of formants. The weighting of a pole frequency with the squared inverse of its bandwidth reflects the observation that narrow bandwidth poles are more likely to be associated with a speech formant whereas poles with a relatively larger bandwidth may be associated with more than one formant or may even be unrelated to any formant. Figure (5) is an example of a formant track estimated using 2-D HMMs. In order to refine the calculation of the mean values of formants, for each vowel and diphthong, a set of formant values can be obtained from the average of the mean value, or the mid value of the formant trajectories of all examples of the phoneme [12, 44].

The following method is used to obtain a set of mean formant trajectories for each phoneme. First, each speech example in the training database is processed to extract a set of N formant tracks [$F_1(t)$... $F_N(t)$]. Then, for each phoneme, the set of N formant tacks are dynamically time-aligned and interpolated or decimated such that they all have duration equal to the mean duration of the phoneme. The formants are then averaged to yield a set of N mean formant trajectories for each phoneme. Experiments indicate that the mean trajectories of context-independent phonemes are relatively flat and do not exhibit distinct features. To obtain the distinct curves of the fluctuations of the trajectories of the formants, context-dependent triphone units are used.

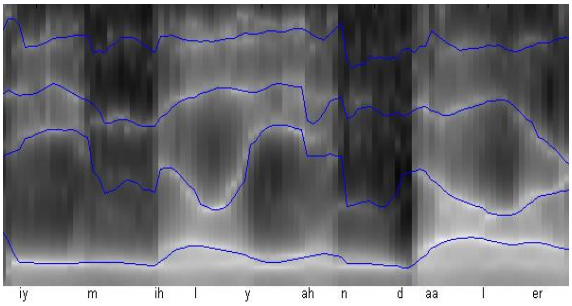


Figure 5: An example of formant tracks estimated using 2-D HMMS superimposed on the LPC spectrum from an American male speaker.

4. COMPARISON OF FORMANTS OF BRITISH, AUSTRALIAN AND AMERICAN ACCENTS

This section presents a comparative investigation of the differences between the formant spaces of British, Australian and American accents. The four most significant formants (namely F1, F2, F3 and F4) are shown. Distinctive characteristics of vowel trajectories in the formant spaces of British, Australian and American accents are studied.

As described in Wells [1], in phonetics, the front or back articulation of vowels are associated with high and low values of F2 while high (close) and low (open) articulation of vowels are associated with the low and high values of F1 respectively [1]. Note that back and front attributes refer to the horizontal tongue position during the articulation of a vowel relative to the back of the mouth. In front vowels, such as /i/(iy), the tongue is positioned forward in the mouth, whereas in back vowels, such as /u/(uw), the tongue is positioned towards the back of the mouth. The height of a vowel refers to the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw. In high vowels, such as /i/(iy) and /u/(uw), the tongue is positioned high in the mouth, whereas in low vowels, such as /a/(aa), the tongue is positioned low in the mouth.

4.1 Formant Spaces of British, Australian and American Accents

Using the modified formant estimation method described in section 3, the average formants of the vowels and diphthongs of British, Australian and American accents are calculated. Figure (6) shows the average of the first, second, third and fourth formants of the monophonic vowels for male and female speakers for these three accents of English. Figure (7) shows a comparative illustration of the formants of British, Australian and American accents in the F1/F2 space. Some significant differences in the formant spaces of these accents are evident from Figure (7). The results conform to previous findings regarding the effect of accent on the F1/F2 space [12, 45].

It can be seen that, except for the vowels /a/(*aa*), /ʌ/(*ah*) and /ɒ/(*oh*), the Australian vowels have a lower value of F1 than the British vowels. The American vowels exhibit a higher value of F2 than British except for /ɜ:/(*er*). On average, the 2nd formants of Australian vowels are 11% higher than those of British and 8% higher than those of American vowels. The 3rd and 4th formants are consistently higher in the

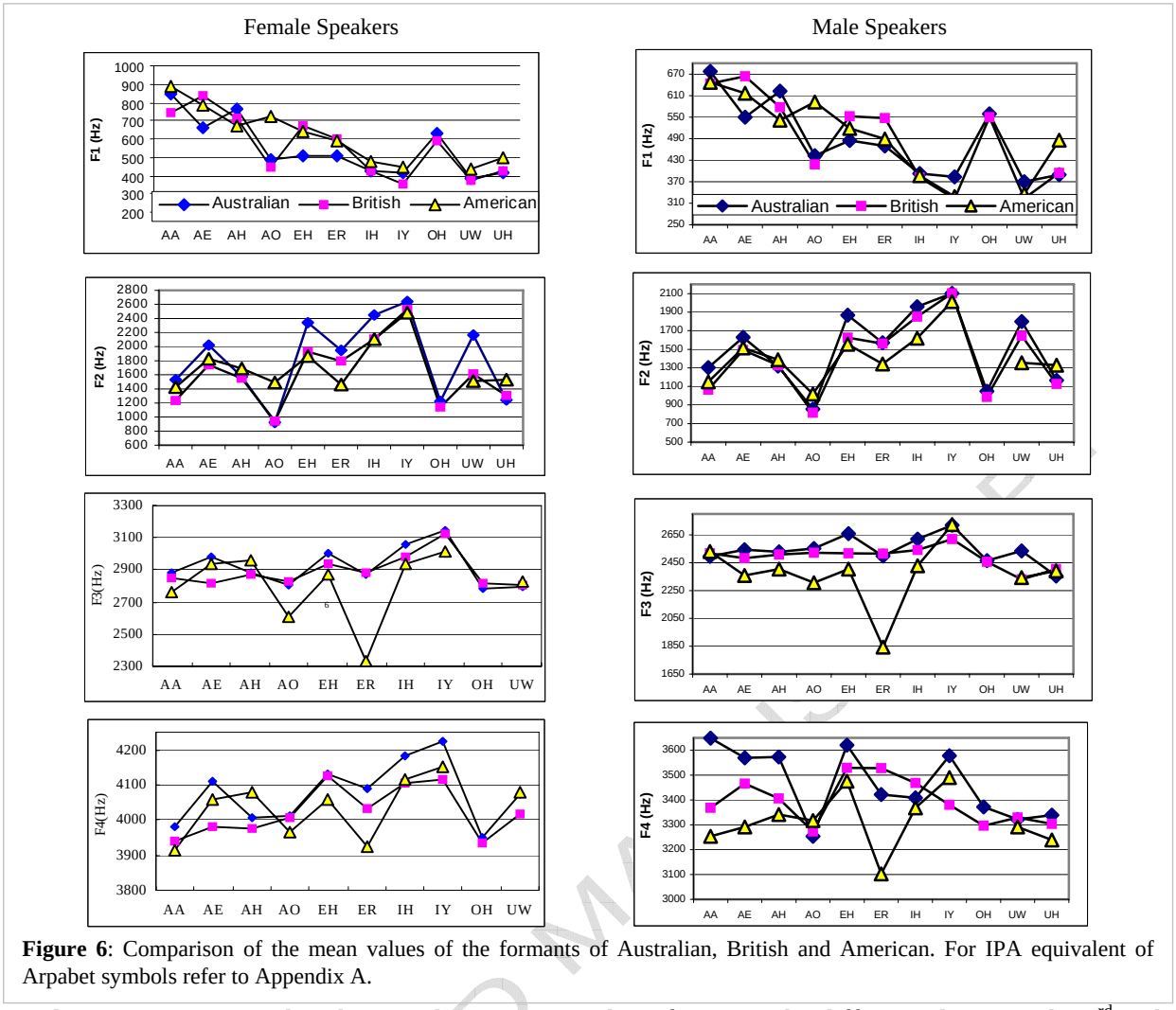


Figure 6: Comparison of the mean values of the formants of Australian, British and American. For IPA equivalent of Arpabet symbols refer to Appendix A.

Australian accent compared to the British accent. A striking feature is the difference between the 3rd and 4th formants of the American vowel /ɜ:/(*er*) compared to those of the British and Australian accents. Generally there are apparent differences in the values of F3 and F4 across accents as can be seen in Figure (6). The results show that American males have a lower F3 and F4 compared to British and Australian accents. The lower frequencies of F3 and F4 in American vowels compared to those in British and Australian English accents are consistent with the rhoticity of American English [45].

An analysis of the formants of vowels, in Figure (6), shows that the most dynamic of the formants is the 2nd formant with a frequency variation of up to 2 kHz. For the Australian female accent, the average vowel frequency of the 2nd formant varies from about 900 Hz for the vowel /ɔ:/(*ao*) to 2600 Hz for /i:/(*iy*). The range of variations of formants is converted to the Bark frequency scale to determine how many auditory critical bands the variations of each formant covers [46]. The second formant F2 covers 8 Barks while F1, F3 and F4 span about 5, 2 and 2 Barks respectively. The results indicate that the 2nd formant is the most

significant resonant frequency contributing to accents. Male speakers display a similar pattern. This result also supports the argument in [39] that the 2nd formant is essential for the correct classification of accents. Note that the second formant also occupies a frequency range with a relatively high-energy concentration and high auditory sensitivity. This result indicates that the 2nd formant may be the most significant formant for conveying accents. The 1st formant, with a frequency range of up to 1 kHz, is regarded as the second most important formant for accent classification.

Figure (7) illustrates the (F1 vs F2) formant spaces of British, American and Australian accents. It can be noted that in comparison to the British and American vowels, the Australian vowels exhibit the following distinctive characteristics in the formant space:

- The raising of the vowels /æ/(*ae*) and /ɛ/(*eh*) in Australian,
- The fronting of the open vowel /ɑː/(*aa*) and the high vowels /uː/(*uw*) and /ʊ/(*uh*) in Australian,
- The fronting and raising of the vowel /ɜː/(*er*) in Australian,
- The vowels /iː/(*iy*), /ɛ/(*eh*) and /æ/(*ae*) in Australian are closer together.

With the vowels /æ/(*ae*) and /ɛ/(*eh*) in Australian articulated higher towards /iː/(*iy*), the result is that /iː/(*iy*), /ɛ/(*eh*) and /æ/(*ae*) are articulated much closer in the Australian formant space than in those of the other two accents.

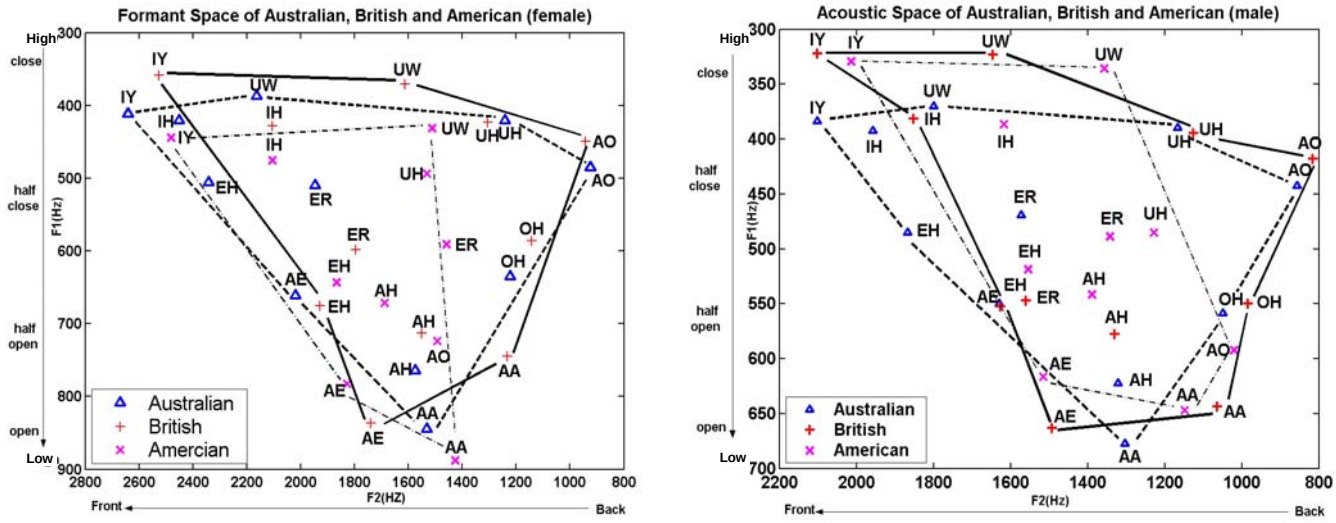


Figure 7: Comparative illustration of the formant spaces of British, Australian and American English. For IPA equivalent of Arpabet symbols refer to Appendix A.

Wells [1] suggests that in Australian /ɪː/(*iy*), /ɛ/(*eh*) are pharyngealised and /æ/(*ae*) is nasalised. From the formant space of Figure (7), it can be seen that these vowel movements form a trend such that the front short vowels in Australian are more squashed into the upper part of the vowel space. In addition, the vowel /ɜː/(*er*) in Australian is relatively more closed (has a lower F1, 510 Hz for female speakers in Figure 7) and more fronted (has a higher F2, 1950 Hz for female speakers in Figure 7) compared to British and American /ɜː/(*er*) which have F1 of 598 Hz and 590 Hz and F2 of 1800 Hz and 1460 Hz respectively. The noticeable fronting of /aː/(*aa*) in Australian makes /ɔː/(*ao*) the only long back vowel in Australian.

The formant spaces of Figure (7) also reveal that American /ɑ/(*aa*) is slightly more open (it has a higher F1, 888 Hz for female speakers in Figure 7) compared to the British /ɑː/(*aa*) which has F1 of 745 Hz, and American /ʌ/(*ah*) is centralised compared to British and Australian accents. The most striking feature in the formant spaces of the three accents is that of the American /ɔ/(*ao*) which is a much lower (it has a higher F1, 724 Hz for female speakers in Figure 7) and more fronted vowel (it has a higher F2, 1493 Hz for female speakers in Figure 7) compared to British and Australian vowels due to the tendency of the vowels /ɔ/(*ao*) and /ʊ/(*oh*) to merge in American English.

A further distinct feature of some dialects of American English is the Northern cities vowel shift, so called as it is taking place mostly in an area beginning some 50 miles west of Albany and extending west through Syracuse, Rochester, Buffalo, Cleveland, Detroit, Chicago, Madison, and north to Green Bay [47]. In this shift, the vowels in the words *cat*, *cot*, *caught*, *cut*, and *ket* have shifted from IPA: /[æ],[ɑ],[ɔ],[ʌ],[ɛ]/ toward [ɪə],[a],[ɑ],[ɔ],[ə], and, in addition, the vowel in *kit* (IPA [ɪ]) becomes more mid-centralised [47]. Consideration of the differences in the formant spaces of vowels and diphthongs indicate that formants play a central role in conveying different English accents.

5. CROSS ENTROPY OF FORMANTS AND CEPSTRUM FEATURES ACROSS ACCENTS

In this section the cross entropy information metric is employed to measure the differences between the acoustic features (here formants and cepstrum) of phonetic units of speech spoken in different accents.

Specifically, this section addresses the measurement of the effect of accent on the probability distributions of spectral features of phonetic units of sounds and compares the differences across speaker groups of the same accent with the differences across speaker groups of different accents. The effect of different databases on the calculation of cross entropy measures is also explored.

5.1 Cross Entropy of Accent Models

Cross entropy is a measure of the difference between two probability distributions [48]. There are a number of different definitions of cross entropy. The definition used here is also known as Kullback-Leibler distance. Given the probability models $P_1(x)$ and $P_2(x)$ of a phoneme, or some other sound unit, in two different accents a measure of their differences is the cross entropy defined as:

$$CE(P_1, P_2) = \int_{-\infty}^{\infty} P_1(x) \log_2 \frac{P_1(x)}{P_2(x)} dx = \int_{-\infty}^{\infty} P_1(x) \log_2 P_1(x)\, dx - \int_{-\infty}^{\infty} P_1(x) \log_2 P_2(x)\, dx \tag{4}$$

Note that the integral of $P(x)$ log $P(x)$ is also known as *the differential entropy*.

The cross entropy is a non-negative function. It has a value of zero for two identical distributions and it increases with the increasing dissimilarity between two distributions [48, 49]. Cross entropy is asymmetric $CE(P_1,P_2) \neq CE(P_2,P_1)$. A symmetric cross entropy measure can be defined as

$$CE_{sym}(P_1, P_2) = \big(CE(P_1, P_2) + CE(P_2, P_1)\big)/2 \tag{5}$$

In the following the cross entropy distance refers to the symmetric measure and the subscript *sym* will be dropped.

The cross entropy between two different left-right N-state HMMs of phonetic speech units with M-dimensional (cepstral or formant) features, and Gaussian mixture pdfs in each HMM state, may be obtained as the sum of the cross-entropies of their respective states as

$$CE(P_1, P_2) = \sum_{s=1}^{N} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P_1(\boldsymbol{x} \mid s) \log_2 \frac{P_1(\boldsymbol{x} \mid s)}{P_2(\boldsymbol{x} \mid s)} dx_1 \cdots dx_M \tag{6}$$

where the Gaussian mixture pdf of the feature vector $\boldsymbol{x}$ in each state $\boldsymbol{s}$ of an HMM is obtained as

$$P(\boldsymbol{x} \mid s) = \sum_{i=1}^{K} P_i N(\boldsymbol{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{7}$$

where P_i is the prior probability of i^{th} mixture of state s, K is the number of Gaussian pdfs in each mixture and $N(\boldsymbol{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is an M-variate Gaussian density. Note that in Equation (6) the corresponding states of the

two models are compared with each other, this is a reasonable procedure for comparing short duration units such as phonemes. The cross-entropy distance can be used for a wide range of purposes including: (a) quantification of the differences between two accents or the voices of two speakers, (b) clustering of phonemes, speakers or accents, (c) ranking of voice or accent features.

5.2 The Effects of Speakers Characteristics on Cross Entropy

Speech models would inevitably include the characteristics of the individual speaker or the averaged characteristics of the group of speakers in the database on which the models are trained. For accent measurement a question arises: how much of the cross entropy between the voice models of two speaker groups is due to the difference in their accents and how much of it is due to the differences of the voice characteristics of the speakers?

In this paper we assume that the cross entropy due to the differences in speaker characteristics and the cross entropy due to accent characteristics are additive. We define an accent distance as the differences between the cross entropies of inter-accent models (e.g. when one set of models are trained on a group of British speakers and the other on a group of American speakers) and intra-accent models obtained from models trained on different speaker groups of the same accent. The adjusted accent distance between two speech unit models may be expressed as

$$AccDist(P_1, P_2) = InterAccDist(P_1, P_2) - IntraAccDist(P_1, P_2) \tag{8}$$

where P_1 and P_2 are two models of the same phonetic units in two different accents. Inter-accent distance is the distance between models trained on two speaker groups across accents whereas intera-accent distance is the distance between models trained on different speaker groups of the same accents. The total distance, due to all variables, between *Nu* phonetic models trained on speech databases from two different accents, A_1 and A_2, can be defined as

$$Dist\left(A_1, A_2\right) = \sum_{i=1}^{N_U} P_i \, AccDist\left(P_1(i), P_2(i)\right) \tag{9}$$

where N_u is the number of speech units and P_i the probability of the i^{th} speech unit. In the following the inter-accent and intra-accent cross entropies of accent of English are quantified.

5.3 Cross Entropy Quantification of Formants and Cepstrum Across English Accents

In this section we describe experimental results on application of cross entropy information metric for quantification of the influence of accents, speakers and database on formants and cepstrum features. The cross entropy distance is obtained from HMMs trained on different speaker groups using the American WSJ and TIMIT, the British WSJCAM and the Australian ANDOSL databases.

Formants are expected to be a better predictor of differences of English accents than cepstrum. This is because: (1) dialects of English are known to differ mostly in vowel production and to maintain largely similar consonant patterns, (2) vowels are largely represented acoustically by formants while consonants involve transient and noise-related acoustic patterns and (3) the formants largely reflect the shape of the vocal tract for a particular sound, independent of phonation type, pitch range, etc. Furthermore, cepstral measures include additional influences relevant to differences across individuals or groups of individuals rather than across accents. In this section the cross-entropy method is used to validate the expectation that formants are better indicators as of accents than cepstrum.

The plots in Figures (8) and (9) illustrate the results of measurements of inter-accent and intra-accent cross entropies, across various speaker groups, for formant features Figure (8) and cepstrum features Figure (9). The motivation for using groups of speakers is to achieve a reasonable degree of averaging out of the effect of individual speaker characteristics. One can then compare the differences in speech features between speaker groups within and across accents. Eighteen different speakers of the same gender were used to obtain each set of models for each speaker group in each accent. The choice of the number of speakers was constrained by the available databases.

A consistent feature of all these evaluation results, as evident in Figures (8) and (9), is that in all these cases the inter-accent differences between HMMs of different speaker groups across different accents are significantly greater than the intra-accent differences between HMMs of different speaker groups of the same accents.

Furthermore, the results show that the cross entropy differences between Australian and British accents are less than the differences between American and British (or Australian), indicating that Australian and British accents are closer to each other in comparison to American English.

Accents / Metric	Am_{wsj}-Timit	Am_{wsj} -Au	Am_{wsj} -Br	Br-Au
Cross Entropy	4.9	25.8	36.8	2.9

Table 1: Cross entropy between American, British and Australian accents (*Formant* Features).

A comparison of the cross entropies of formant features versus cepstrum features within and across the databases shows that the formant features are more indicative of accents than the cepstrum, that is difference in formants across accents is more pronounced.

A particularly interesting comparison is that of two different American databases namely WSJ and TMIT versus each other and other databases. A good accent indicator (i.e. one that is robust to the differences due to speakers and recordings in different database) should indicate that American WSJ and TIMIT are close to each other than to British WSJCAM0 or Australian ANDOSL.

It can be seen that the formant features consistently show a much closer distance between the HMMs trained on American TIMIT and the HMMs trained on American WSJ compared to the distances of these models from HMMs trained on databases of British WSJCAM0 or Australian ANDOSL accents.

This shows that difference across speaker groups from different accents is not due to the recording condition of the databases since in all these databases care have been taken to ensure that the recording process does not distort the signal and all the databases used have been recorded in quiet conditions.

The following may explain why formants seem to be better indicators of accents than cepstrum. The cepstrum features contain all the variables that affect speech in particular speaker information and recording environment. On the other hand formants features are extracted from peak energy contours which, by the very nature of the process of formant estimation, are less affected by the recording environment.

Table (1) shows the cross entropy distances between different accents. It is evident that among the four accent pairs Australian and British are closest. Furthermore American is closer to Australian than to British.

Table (2) shows the ranking of vowels for different accent pairs in the descending order of the cross-entropy distance across the accents. The ranks of vowels were obtained from the sum of the cross-entropies of the first four formants.

Accent Pair	*Cross Entropy Distance Ranking of Most Distinct Phonemes*															
American & Australian	***ER***	***UW***	***OW***	***AH***	***OY***	***R***	***EH***	***IY***	***AA***	***EY***	***AY***	***AW***	***AE***	***UH***	***AO***	***IH***
American & British	***ER***	***OW***	***UW***	***EY***	***IY***	***AY***	***AH***	***OY***	***R***	***EH***	***UH***	***AA***	***AE***	***AW***	***AO***	***IH***
Australian & British	***UH***	***OW***	***EH***	***ER***	***R***	***AO***	***OY***	***IY***	***EY***	***UW***	***AA***	***AW***	***AY***	***AH***	***IH***	***AE***

Table 2: Illustration of the cross entropy ranking of the most distinct phonemes across pairs of accents (*Formant* Features).

6. CROSS ACCENT PHONETIC TREE CLUSTERING

Clustering is the grouping together of similar items. In this section the minimum cross entropy (MCE) information criterion is used, in a bottom-up hierarchical clustering process, to construct phonetic cluster trees for different accents of English. These trees show the structural similarities and the differences of phonetic units from different accents [50].

To illustrate the bottom-up hierarchical clustering process, assume that we start with M clusters C_1, …, C_M. Each cluster may initially contain only one item. For the phoneme clustering process considered here, each cluster initially contains the HMM probability model of one phoneme.

At the first step of the clustering process, starting with M clusters, the two most similar clusters are merged into a single cluster to form a reduced set of M-1 clusters. This process is iterated until all clusters are merged.

A measure of the similarity (or dissimilarity) of two clusters is the average CE of their merged combination. Assuming that the cluster C_i has N_i elements with probability models $P_{i,k}$, and cluster C_j has N_j elements with probability models $P_{j,l}$, the average cross entropy of the two clusters is given by

$$CE(C_i, C_j) = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} CE(P_{i,k}, P_{j,l}) \tag{10}$$

The MCE rule for selecting the two most similar clusters, among N clusters, for merger at each stage are

$$[C_i, C_j] = \arg\min_{i=1:N} \arg\min_{\substack{j=1:N \\ j \neq i}} CE(C_i, C_j) \tag{11}$$

The results of the application of MCE clustering for construction of phonetic-trees of American, Australian and British English are shown in Figures (10), (11) and (12).

The clustering of American phonemes more or less corresponds to how one would expect the phonemes to cluster. The phonetic trees of Australian and British accents, Figures (11) and (12) are more similar to each other than to American phonetic tree. This observation is also supported by the calculation of the cross entropy of these accents, presented in the previous section in Table (1).

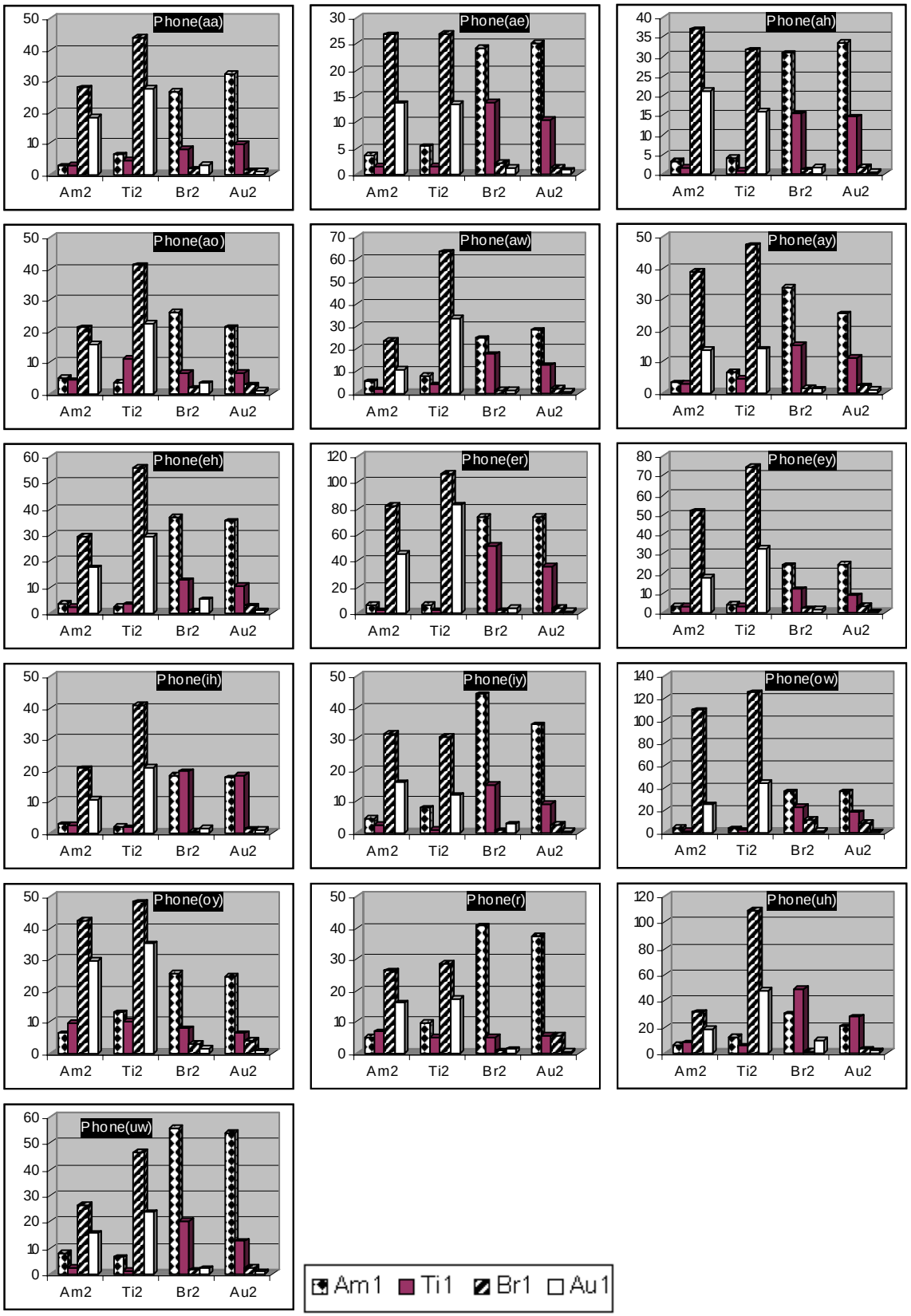


Figure 8: Plots of Female Speakers (Formants) for inter-accent and intra-accent cross entropies of a number of phonemes of American, British and Australian accents. Note that each coloured column shows the cross entropy of a group of one speech accent from another indicated on the horizontal axis. For IPA equivalents of Arpabet symbols refer to Appendix A.

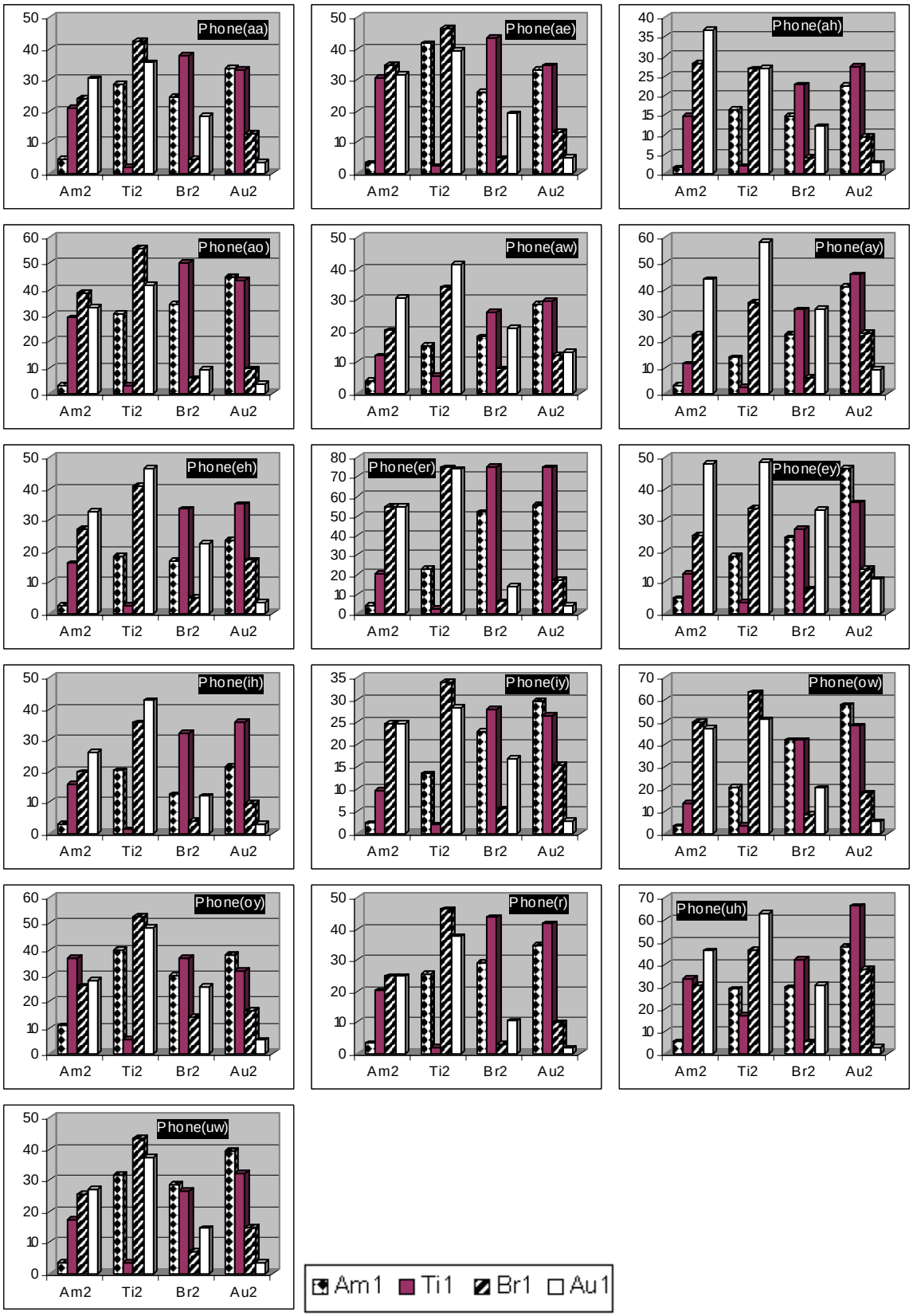


Figure 9: Plots of Female Speakers (MFCCs) for inter-accent and intra-accent cross entropies of a number of phonemes of American, British and Australian accents. Note that each different coloured column shows the cross entropy of a group of one speech accent from another indicated on the horizontal axis. For IPA equivalents of Arpabet symbols refer to Appendix A.

Figure (13) shows a cross-accent phonetic-tree between British and Australian accents. This tree shows how the vowels in British accent cluster with the vowels in Australian accent.

7. A COMPARISON OF THE IMPACT OF ACCENT AND GENDER ON ASR

The importance of accent variability is illustrated by comparing the effect of accent variations versus gender variation on the performance of HMM-based speech recognition systems. Here HMMs were trained on

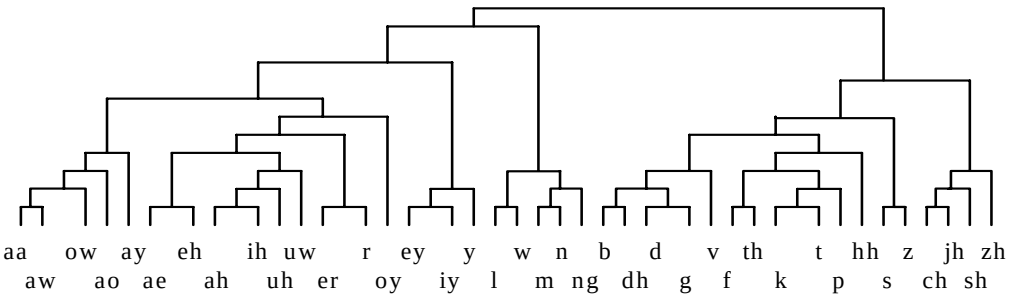


Figure 10: Phonetic-tree clustering of American accent.

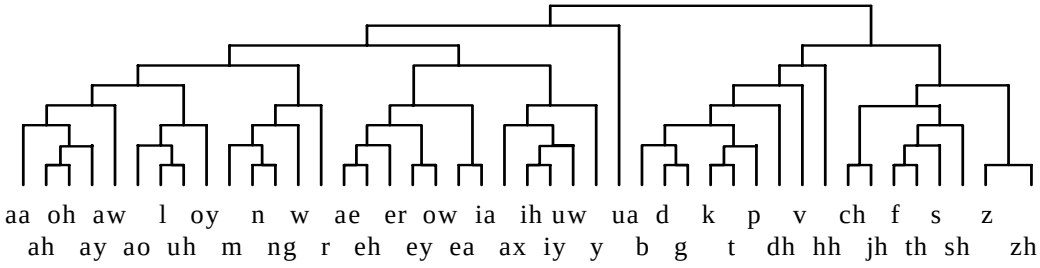


Figure 11: Phonetic-tree clustering of Australian accent.

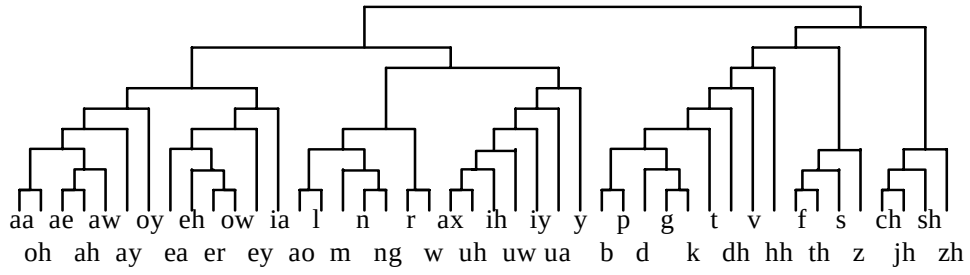


Figure 12: Phonetic-tree clustering of British accent.

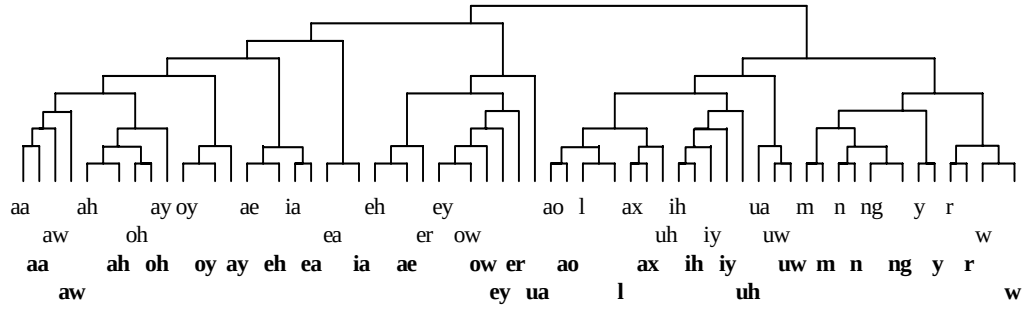


Figure 13: Cross–Accent phonetic-tree clustering of British and Australian (bold) accents for vowels. For IPA equivalent of Arpabet symbols refer to Appendix A.

phonetic units of speech using 39-dimensional features comprising of cepstrum, delta and delta delta features.

The results of evaluation of the impact of same/different gender or same/different accents on the error rate of automatic speech recognition are shown in Table (3) for sentences of an average duration of 5 seconds. These results reveal that an accent mismatch can have a similar detrimental effect on the accuracy of automatic speech recognition as gender mismatch. The table shows that in some cases differences in accents can be more detrimental to speech recognition than gender mismatch. This result highlights the importance of research in the modelling and quantification of accents.

8. CONCLUSION

In this paper the cross-entropy information measure is applied for quantification of the differences between accents and for construction of cross accent phonetic-trees. The formants features were extracted from 2-D HMMs. The plots of formant spaces for different accents of English show that formant features are significantly affected by accent. Furthermore the effect of accent on formants is non-uniform and depends on the phonemes and accents. For example, there is a relatively substantial difference between the formants of the American pronunciation of the phoneme '*r*' and its Australian and British pronunciations signifying the rhotic nature of American accent.

Through the use of cross entropy, as a measure of the differences of the models of probability distributions of speech across accents, it is shown that formants are a stronger indicator of accents than cepstrum features. This is particularly the case for vowels which are well characterised by formants. However, it is noted that an advantage of cepstrum features is their versatility as they provide a more appropriate parameterisation of the consonants than formants.

MODEL / INPUT	BrF	BrM	AmF	AmM	AuF	AuM
BrF	**30.1**	43.3	53.7	60.2	42.3	53.4
BrM	45.7	**33.1**	62.5	53.4	51.4	43.4
AmF	51.3	62.0	**33.6**	40.3	52.8	66.9
AmM	61.0	51.3	45.4	**34.8**	57.9	51.9
AuF	41.8	52.2	51.6	56.2	**29.0**	47.2
AuM	54.9	45.4	62.3	51.1	46.6	**31.9**

Table 3: The effect of accent and gender on the (%) error rate of automatic recognition accuracy of phonetic units of speech.

Experiments on comparative evaluation of cross entropies of speech signals from two different databases of American accents (namely WSJ and TIMIT) and Australian and British databases show that in most cases there is a closer cross entropy between the two regional American speech databases in comparison to their cross entropy distances from British and Australian. Hence, assuming that in production of speech databases the recording conditions are similar and high quality microphones are used, the cross entropies differences across different speech databases mostly exhibit the differences in speaker characteristics and accents. However, in relation to the use of TIMI and WSJ databases it should be also noted that there exist major regional differences across American English, which is reflected.

The cross entropy has been used for phonetic-tree clustering within and across accents. The consistency of phonetic-trees for different groups of the same accent shows that cross-entropy is a good measure for hierarchical clustering. The cross entropy of inter-accent groups compared to that of the intra-accent groups clearly shows the level of dissimilarity of phonetic models due to effect of accents. Further work is being carried out on the use of cross entropy to measure the accents of individuals.

REFERENCES

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

[2] Köhler J., 1996. Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds. In: ICSLP, Philadelphia, pp. 2195–2198.

[3] Hansen J. H. L., Yapanel U., Huang R., Ikeno A., 2004. Dialect analysis and modelling for automatic classification. In: Interspeech, Jeju, pp. 1569–1572.

[4] Ten Bosch L., 2000. ASR, dialects and acoustic/phonological distances. In: ICSLP, Beijing, pp.1009–1012.

[5] Humphries J., 1997. Accent Modeling and Adaptation in Automatic Speech recognition. PhD Thesis, Cambridge University Engineering Department.

[6] Miller C. A., 1998. Pronunciation modeling in speech synthesis. PhD thesis, University of Pennsylvania.

[7] Yan Q., Vaseghi S., 2002. A Comparative Analysis of UK and US English Accents in Recognition and Synthesis. In: IEEE Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 413-417.

[8] Woehrling C., Boula de Mareüil P., 2006. Identification of regional accents in French: perception and categorization. In: Interspeech, Pittsburgh, pp. 1511–1514.

[9] Van Bezooijen R., Gooskens C., 1999. Identification of Language Varieties. Contribution of Different Linguistic Levels. Journal of Language and Social Psychology, 18:1, pp. 31–48.

[10] Crystal D., 2003. A dictionary of linguistics and phonetics. Blackwell: Malden.

[11] Boula de Mareüil P., Vieru-Dimulescu B., 2006. The Contribution of Prosody to the Perception of Foreign Accent. Phonetica, 63, pp. 247–267.

[12] Harrington J., Cox F., Evans Z., 1997. An Acoustic Phonetic Study of Broad, General, and Cultivated Australian English Vowels. Australian J. of Linguistics 17, pp. 155-184.

[13] Trubetzkoy N.S., 1931. Phonologie et géographie linguistique. Travaux du Cercle Linguistique de Prague 4. 228-234.

[14] IPA, available at International Phonetic Association website: http://www2.arts.gla.ac.uk/IPA/ipa.html .

[15] Labov W., 1994. Principles of Linguistic Change. Vol. 1: Internal features, Blackwell: Oxford & Cambridge.

[16] Nagy N., Roberts J., (forthcoming). New England: phonology. In Schneider E., Burridge K., Kortmann B., Mesthrie R., Upton C., eds. A Handbook of Varieties of English. Volume 2: Varieties of English of the Americas and the Caribbean. Berlin, NY: Mouton de Gruyter.

[17] Mitchell A. G., Delbridge A., 1965. The speech of Australian adolescents. Melbourne: Angus and Robertson.

[18] Przewozny A., 2004. Variation in Australian English. La Tribune internationale des langues vivantes, 36, novembre, pp. 74-86

[19] Cruttenden A., 1997. Intonation. Cambridge University Press.

[20] Ho Ching-Hsian., 2001. Speaker Modelling for Voice Conversion, PhD thesis, School of Engineering and Design, Brunel University.

[21] Yan Q., Vaseghi S., Rentzos D., Ho C. H., Turajlic E., 2003. Analysis of Acoustic Correlates of British, Australian and American Accents. In: ASRU Proc, pp. 345-350.

[22] Nolan F., Grabe E., 1997. Can ToBI transcribe intonational variation in British English?. In: ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications, Athens, pp. 259-262.

[23] Grabe E., Post B., Nolan F., 2001. Modelling intonational variation in English: the IViE system. In: Prosody 2000 Workshop, Kraków, pp. 51–57.

[24] Fletcher J., Grabe E., Warren P., 2004. Intonational variation in four dialects of English: the high rising tune. In S. -A. Jun (Ed.), Prosodic typology and transcription — a unified approach, Oxford University Press: Oxford.

[25] Ikeno A., Hansen J. H. L., 2006. The Role of Prosody in the Perception of US Native English Accents. In: Interspeech, Pittsburgh, pp. 437–440.

[26] Ladd D. R., 1996. Intonational phonology. Cambridge University Press: Cambridge.

[27] Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Dan P., Valtchev, Woodland P., 2002. The HTK Book. Version 3.2. Cambridge University Engineering Department.

[28] BEEP Dictionary available: http://mi.eng.cam.ac.uk/~ajr/wsjcam0/node8.html.

[29] Macquarie Dictionary, available http://handbook.mq.edu.au/p1/pt1b_046.htm.

[30] CMU Dictionary available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[31] Deller J.R., Jr., Proakis, J.G., Hansen, J.H.H., 1993. Discrete-Time Processing of Speech Signals. New York: Macmillan Publishing Company.

[32] Arslan, L.M., Hansen H., 1997. A Study of Temporal Features and Frequency Characteristics in American English Foreign Accent. J. Acoustic. Soc. Am, Vol. 102(1), pp. 28-40.

[33] Rabiner L., Schafer R., 1978. Digital Processing of Speech Signals. Prentice-Hall.

[34] Snell R., Milinazzo F., 1993. Formant Location from LPC analysis Data. IEEE Trans. on Speech and Audio Proc. Vol. 1, No. 2, pp. 129-34.

[35] Yan Q., Vaseghi S., Zavarehei E., Milner B., Darch J., White P., Andrianakis I., 2007. Formant-Tracking Linear Prediction Model Using HMMs and Kalman Filters for Noisy Speech Processing. In Press, Computer Speech and Language.

[36] Yan Q., 2005. Analysis, Modelling and Synthesis of British, Australian and American English Accents. PhD thesis, Brunel University.

[37] Darch J., Milner B., 2007. A Comparison of Estimated and MAP Predicted Formants and Fundamental Frequencies with a Speech Reconstruction Application. In Interspeech, pp. 542-545, Antwerp, Belgium.

[38] Darch J., Milner B., Vaseghi S., 2006. MAP prediction of formant frequencies and voicing class from MFCC vectors in noise. Speech Communication, vol. 48, no. 11, pp. 1556–1572.

[39] Weber K., Bengio S., Bourlard H., 2001. HMM2-Extraction of Formant Structures and Their Use for Robust ASR. In: Proc. Eurospeech, Aalborg, Denmark, pp. 607-610.

[40] Vergin R., Farhat A., O'Shaughnessy D., 1996. Robust Gender-Dependent Acoustic-Phonetic Modeling in Continuous Speech Recognition Based on a New Automatic Male/Female Classification. In: Proc. ISCLP, pp. 1081-1084.

[41] Kim C., Sung W., 2001. Vowel Pronunciation Accuracy Checking System based on Phoneme Segmentation and Formants Extraction. In: Proc. Int. Conf. Speech processing, pp. 447-452. Daejeon, Korea.

[42] Dempster A., Laird N., Rubin D., 1977. Maximum Likelihood from Incomplete Data via the {EM} Algorithm. J. Roy Stat. Soc., 39(B). pp. 1-38.

[43] Childers D.G., Wu K., 1991. Gender Recognition from Speech Part II: Fine Analysis. J. Acoustic. Soc. Am. Vol 90, pp. 1841-1856.

[44] Watson C., Harrington J., Evans Z., 1996. An Acoustic Comparison between New Zealand and Australian English Vowels. Australian J. of Linguistics.

[45] Boyce S. E., Espy-Wilson C. Y., 1997. Coarticulatory Stability in American English /r/. J. Acoustic. Soc. Am., 101 (6), pp.3741-3753.

[46] Zwicker E., Flottorp G., Stevens S.S., 1957. Critical bandwidth in Loudness Summation. J. Acoustic. Soc. Am. 29 pp. 548-557.

[47] Labov W., Sharon A., Charles B., 2006. The Atlas of North American English. Berlin: Mouton-de Gruyter.

[48] Shore J. E., Johnson R. W., 1981. Properties of cross-entropy minimisation. IEEE Trans. Inform. Theory, vol. IT-27, pp. 472-482, July.

[49] Jaynes E. T., 1982. On the rationale of maximum entropy methods. Proc. IEEE, vol. 70, pp. 939-952, Sep.

[50] Huckvale M., 2004. ACCDIST: a Metric for Comparing Speakers' Accent. Proc. On spoken Language Processing, ICSLP.

APPENDIX

A: IPA Phonetics Symbols

IPA	Arpabet	IPA	Arpabet
ɪ	ih	d	d
iː	iy	ð	dh
ɛ	eh	f	f
æ	ae	g	g
aː	aa	h	hh
ʌ	ah	ʤ	jh
ɒ	oh	k	k
ɔː	ao	l	l
u	uh	m	m
uː	uw	N	n
əː	er	Ŋ	ng
ə	ax	P	p
ei	ey	R	r
ai	ay	S	s
au	aw	ʃ	sh
əu	ow	T	t
ɔi	oy	Θ	th
iə	ia	V	v
eə	ea	W	w
uə	ua	J	y
b	b	Z	z
tʃ	ch	ʒ	zh