CROSS-ENTROPIC COMPARISON OF THE EFFECTS OF ACCENT, SPEAKER AND DATABASE RECORDING ON SPECTRAL FEATURES OF ENGLISH ACCENTS

Seyed Ghorshi *Saeed Vaseghi* **Qin Yan*

School of Engineering and Design, Brunel University, London,
*School of Computer Information and Engineering, Hohai University, Nanjing, P.R.China,
{Seyed.Ghorshi, Saeed.Vaseghi}@brunel.ac.uk, *yanqin@ieee.org

ABSTRACT

This paper investigates the use of cross-entropy information measure for quantification and comparison of the impact of the variations of accents, speaker groups and recordings on the probability models of spectral features of phonetic units of speech. Cross-entropy measure can be used in applications such as accent identification, improved speech recognition, cross-accent phonetic-tree analysis and analysis of the influence of accents on different sets of speech parameters and models. For the purpose of this study the focus is on British English, Australian English and two different databases of American English accents (namely WSJ and TIMIT). Comparison of the cross entropies of formants and cepstrum features indicate that cepstrum features are less indicative of accents compared to formants. In particular it appears that the measurements of differences in formants across accents are less sensitive to different recording or databases. It is found that the cross entropies of the same phonemes across different accents (inter-accent distances) are significantly greater than the cross entropies of the same phonemes across different speaker groups of the same accent (intra-accent distances). 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.

Index Terms: *accent, formant, cepstrum, cross entropy, phonetic-tree clustering.*

1. INTRODUCTION

This paper considers the issues of modelling and quantification of the effects of accents on phonetic units of speech and evaluates the effects of variability of speaker and database recording on the measurements of accents. The modelling of differences in accents is an important issue for speech recognition and synthesis. Accent variability has a major impact on speech recognition [1-3]. Models of difference in accents can also be used for accents synthesis in text to speech synthesis [4] and in accent morphing [5].

Since accent and speaker characteristics variables 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 modelling of accent often involves the use of different databases recorded by different equipments, there is also a need to explore the effect of different databases 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.

The term *accent* may be defined as a distinctive pattern of pronunciation, including lexicon and intonation characteristics, of a community of people who belong to a national, regional or social grouping. In Crystal's dictionary of linguistics [6], 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". Accents evolve over time influenced mainly by large immigrations and social and cultural trends as well as the mass media. 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, Irish pronunciation and in relatively recently times by American pronunciation. Similarly, Liverpool accent has been influenced by Irish immigration whereas the Northern Ireland accent has been influenced by Scottish immigration. Geographical variation, socio-economic classes, ethnicity, sex, age, and cultural trends can affect accents. In [7] Wells provides an excellent introduction to the linguistic structures of the accents of English accents.

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

- *Historical approach* compares the historical roots of accents, the evolutionary changes in accents as various accents merge or diverge, the rules of pronunciation and how the rules evolve.
- *Structural, synchronic approach*, first proposed by Trubetzkoy [8] models an accent in a system-oriented fashion in terms of the systematic differences in:
- Differences in phonemic systems.
- Differences in phonotactic (structural) distributions.
- Differences in lexical distributions of words.
- Differences in phonetic (acoustic) realization.

In this work the differences between accents are modeled using a system-based approach as explained next. The remainder of this paper is organized in the following format. Section 2 briefly describes a method for modeling and estimation of formants and presents a comparative analysis of the formant spaces of British, American and Australian English accents. Section 3 introduces cross entropy and presents results of cross entropy analysis of cepstrum and formant features across accents. Section 4 uses cross entropy for phonetic tree clustering and introduces the concept of cross-accent phonetic- tree analysis and finally Section 5 concludes the paper.

1.1 Databases, Features and Models

The databases used for accent analysis in this paper are Australian National Database of Spoken Language (ANDOSL), American Wall Street Journal database (WSJ), American TIMIT and Cambridge University's British Wall Street Journal Database (WSJCAM0). TIMIT contains eight regional American accents; however, we use a mixture of these in order to compare the effect of using different databases of an accent on the measured cross entropy results. The subset of ANDSOL of broad Australian accent comprises 18 female and 18 male speakers with a total of 7200 utterances. The subset of WSJ database used for modeling American English contains 36 female and 38 male speakers with 9438 utterances. The subset of WSJCAM0 of British English used contains 40 female and 46 male speakers with 9476 utterances. For speech segmentation and labeling, left-right hidden Markov models (HMMs) of triphone units are employed and the Viterbi decoder is applied in the forced-alignment mode [9, 10] with phonemic transcriptions supplied. Each HMM has three states and each state is modeled with a Gaussian mixture model with 20 components. The speech feature vectors consist of 13 Mel-Frequency Cepstral Coefficient (MFCCs) and their 1st and 2nd derivatives. The dictionaries used in this work include the BEEP dictionary (British accent), the Macquarie dictionary (Australian accent) and the CMU dictionary (American accent).

2. COMPARISON OF FORMANTS ACROSS ACCENTS

The HMM-based formant estimation method used here is described in detail in [11, 12]. A set of phoneme-dependent HMMs are trained to model the probability of the formants of a group of speakers. In the formant estimation method employed 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 bandwidth of the spectral resonance. 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.

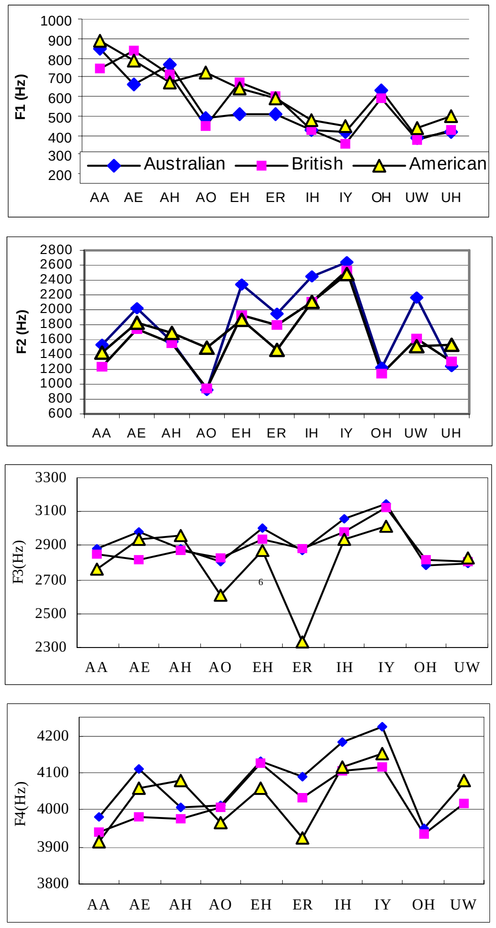

Figure 1: Comparison of the mean values of the formants of Australian, British and American.

Using the formant estimation method described in [11, 12], the average formants of the vowels of British, Australian and American accents are calculated. Figure (1) shows the average of the first, second, third and fourth formants of the monophonal vowels for female speakers for these three accents of English. Figure (2) shows a comparative illustration of the formants of British, Australian and American accents in F1/F2 space. Some significant differences in the formant spaces of these accents are evident from Figure (2). The results conform to previous findings regarding the effect of accent on F1/F2 space [13].

From Figures (1) and (2), it can be seen that, for most vowels, except for the *aa*(ɑ), *ah*(ʌ) , *iy*(ɪ) and *oh*(ɒ) the Australian vowels have a lower value of F1 than the British vowels. The Australian vowels also seem to have a larger F2 than British and Americans. The American vowels exhibit a higher value of F2 than British except for *er*(ɜː). On average, the 2nd formants of Australian vowels are 11% higher than those of British and 8% higher than those of American vowels.

From Figure (1), the 3rd and 4th formants are consistently higher in the Australian accent compared to the British accent. A striking feature is the difference between the 3rd and 4th formants of the American vowel *er*(ɜː) 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 (1). The results show that American females 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 [14].

An analysis of the formants of vowels, in Figure (1), 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 for the vowel *ao*(ɔː) to 2600 Hz for *iy*(iː). The range of variations of formants is converted to the Bark frequency scale to determine how many auditory critical bands the variations

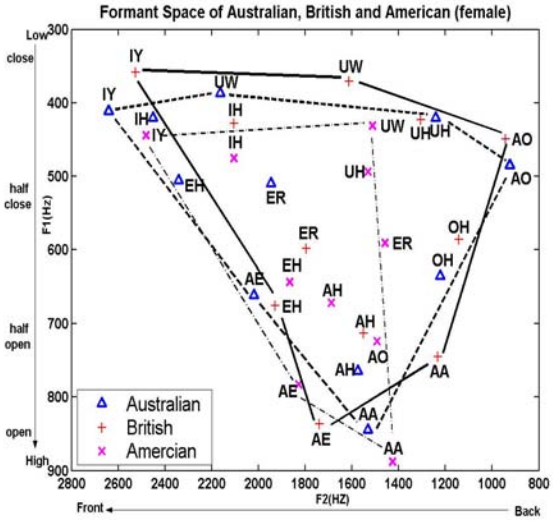

Figure 2: Comparison the formant spaces of British, Australian and American English.

of each formant covers [15]. 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 [16] that the 2nd formant is essential for the correct classification of accents. The 1st formant, with a frequency range of up to 1 kHz, is regarded as the second most important formant for accent classification.

3. 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 (formants and cepstrum) of phonetic units of speech spoken in different accents. The effect of speaker groups and databases on the calculation of cross entropy measures is also explored.

3.1 Cross Entropy of Accent Models

Cross entropy is a measure of the difference between two probability distributions [17]. The cross entropy 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:

$$\begin{aligned} CE(P_1,P_2) &= \int_{-\infty}^{\infty} P_1(x)\log_2 \frac{P_1(x)}{P_2(x)}dx \\ &= \int_{-\infty}^{\infty} P_1(x)\log_2 P_1(x)\,dx - \int_{-\infty}^{\infty} P_1(x)\log_2 P_2(x)\,dx \end{aligned} \quad (1)$$

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 [17].

Cross entropy is asymmetric $CE(P_1,P_2) \neq CE(P_2,P_1)$. A symmetric cross entropy measure can be defined as

$$CE_{sym}(P_1,P_2) = \left(CE(P_1,P_2) + CE(P_2,P_1)\right)/2 \quad (2)$$

In the following the cross entropy distance refers to the symmetric measure and the subscript *sym* will be dropped. The cross entropy between two left-right *N*-states HMMs with *M*-dimensional features, and state Gaussian mixture pdfs, may be obtained as the sum of the cross-entropies of their respective states:

$$CE(P_1,P_2) = \sum_{s=1}^{N} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P_1(\boldsymbol{x}\,|\,s)\log_2 \frac{P_1(\boldsymbol{x}\,|\,s)}{P_2(\boldsymbol{x}\,|\,s)}dx_1 \cdots dx_M \quad (3)$$

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

$$P(\boldsymbol{x}\,|\,s) = \sum_{i=1}^{K} P_i N(\boldsymbol{x},\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \quad (4)$$

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. In Equation (3) the corresponding states of the two models are compared with each other, this is reasonable for short duration units such as phonemes. Given the

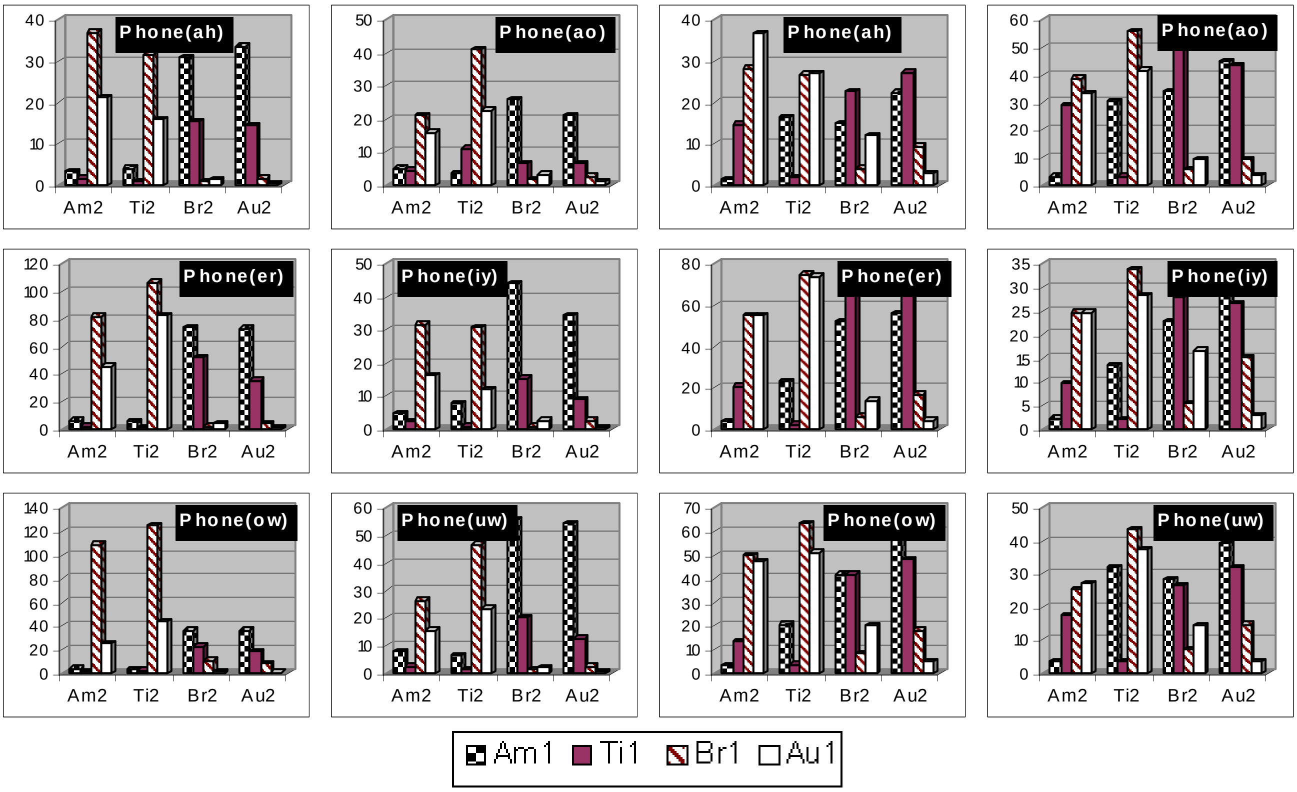

Figure3: Cross-entropies of vowels modelled with formant features. **Figure4**: Cross-entropies of vowels modelled with cepstrum features.

Plots of inter-accent and intra-accent cross-entropies of the probability models of a number of phonemes of American, British and Australian accents. Note each colour-keyed column shows the cross entropy of a HMM probability model trained on a group of speakers of one accent from another model trained on a different group of speakers of either the same accent (intra-accent) or different accent (inter-accent) as indicated on the horizontal axis.

pdfs $P_1(x)$ and $P_2(x)$, the cross entropy (in Equations 1 and 3) can be calculated through numerical integration in a range of L times the variance, with L chosen large enough ($L > 10$) so that it includes all of the non-zero pdf values.

3.2 The Effects of Speakers Characteristics on Cross Entropy

Speech models include the characteristics of the individual speaker or group of speakers 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 accents and how much of it is due to the differences of the voice characteristics of the speakers and databases?

In this paper we assume that the cross entropies due to the differences in speaker characteristics, accents and recordings 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 models may be expressed as

$$AccDist(P_1,P_2) = InterAccDist(P_1,P_2) - IntraAccDist(P_1,P_2) \quad (5)$$

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

$$Dist\left(A_1, A_2\right) = \sum_{i=1}^{N_U} P_i \; AccDist\left(P_1(i), P_2(i)\right) \quad (6)$$

where P_i is the probability of the i^{th} speech unit.

3.3 Cross Entropy of Spectral Features of English Accents

This section describes experimental results on application of cross entropy 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.

The plots in Figure (3) and (4) illustrate the results of measurements of inter-accent and intra-accent cross entropies, across various speaker groups, for formant features Figure (3) and cepstrum features Figure (4). The motivation for using groups of speakers is to achieve a reasonable degree of averaging out of the effect of individual speaker characteristics. 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 number of speakers was constrained by the available databases.

A consistent feature of the results, as evident in Figures (3) and (4), is that in all these cases the inter-accent differences are significantly greater than the intra-accent differences. 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 in comparison to American English.

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. 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 should indicate that American WSJ and TIMIT are close to each other than to British WSJCAM0 or Australian ANODSL. It can be seen 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 conditions 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 quite conditions.

4. CROSS ACCENT PHONETIC TREE CLUSTERING

In this section the minimum cross entropy (MCE) information criterion is used, in a bottom-up hierarchical clustering process, to construct phonetic trees for different accents of English. These trees show the structural similarities and the differences of phonetic units from different accents. To illustrate the 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

$$CE(C_i, C_j) = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} CE(P_{i,k}, P_{j,l}) \quad (7)$$

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

$$[C_i, C_j] = \underset{i=1:N}{\arg\min} \; \underset{\substack{j=1:N \\ j\neq i}}{\arg\min} \, CE(C_i, C_j) \quad (8)$$

The results of the application of MCE clustering for construction of phonetic-trees of American, British and Australian English are shown in Figures (5), (6) and (7). The phonetic-tree of the American accent, Figure (5), confirms the reputation of the American English as being a ‘phonetic accent’ (i.e. an accent in which phonemes are clearly pronounced).

The clustering of American phonemes more or less corresponds to how one would expect the phonemes to cluster. The phonetic trees of British and Australian accents, Figures (6) and (7) are more similar to each other than to American phonetic tree. Figure (8) 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.

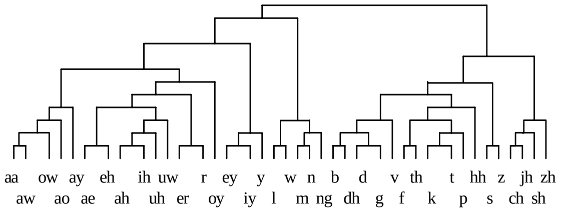
aa ow ay eh ih uw r ey y w n b d v th t hh z jh zh
aw ao ae ah uh er oy iy l m ng dh g f k p s ch sh

Figure 5: Phonetic-tree of American accent.

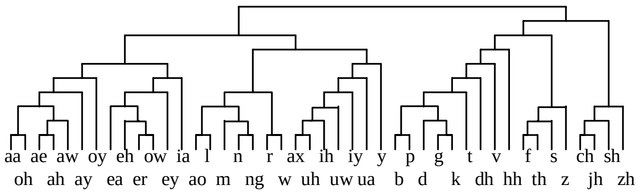
aa ae aw oy eh ow ia l n r ax ih iy y p g t v f s ch sh
oh ah ay ea er ey ao m ng w uh uw ua b d k dh hh th z jh zh

Figure 6: Phonetic-tree of British accent.

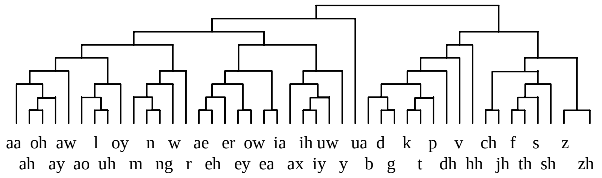
aa oh aw l oy n w ae er ow ia ih uw ua d k p v ch f s z
ah ay ao uh m ng r eh ey ea ax iy y b g t dh hh jh th sh zh

Figure 7: Phonetic-tree of Australian accent.

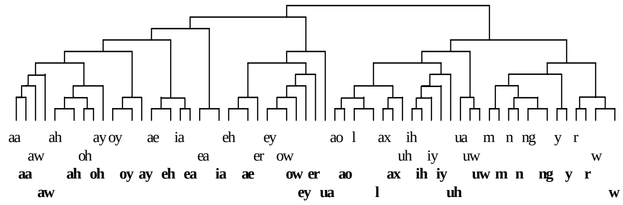
aa ah ay oy ae ia eh ey ao l ax ih ua m n ng y r
aw oh ea er ow uh iy uw w
aa ah oh oy ay eh ea ia ae ow er ao ax ih iy uw m n ng y r
aw ey ua l uh w

Figure 8: Cross-accent phonetic-tree between British and Australian (bold).

5. CONCLUSION

In this paper the cross-entropy information measure is applied for the quantification and comparison of differences of accents and for construction of cross accent phonetic-trees. Through the use of cross entropy measure of comparison of the probability distributions it is established that formants are a stronger indicator of accents than cepstrum features. It is also established that the major difference observed between the cross entropies of different intra-accent and inter-accent groups is also evident across speaker groups taken from different databases of the same accent. The cross entropy was used to construct phonetic trees of American, British and Australian 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] Hansen J. H. L, Yapanel U., Huang R., Ikeno A., (2004), “Dialect analysis and modelling for automatic classification”, *Interspeech*, Jeju, pp. 1569–1572.

[2] Köhler J., (1996), “Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds”, *ICSLP*, Philadelphia, pp. 2195–2198.

[3] Ten Bosch, L., (2000), “ASR, dialects and acoustic/phonological distances”, *ICSLP*, Beijing (pp. 1009–1012).

[4] Miller, C.A., (1998), “Pronunciation modeling in speech synthesis*”, PhD thesis, University of Pennsylvania.*

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

[6] Crystal D., (2003), *“A dictionary of linguistics and phonetics”*, Blackwell: Malden. Fletcher.

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

[8] Trubetzkoy, N.S., (1931), Phonologie et géographie linguistique. *Travaux du Cercle Linguistique de Prague* 4.228-234.

[9] Humphries J., (1997), “Accent Modeling and Adaptation in Automatic Speech recognition”, *PhD Thesis, Cambridge University Engineering Department.*

[10] Young S., Evermann G., Kershaw D., Moore G., Odell J. Ollason D., Dan P., Valtchev, Woodland P., (2002), “The *HTK Book”. V3.2*.

[11] Ho Ching-Hsiang, (2001), “Speaker Modelling for Voice Conversion”, *PhD thesis, School of Engineering and Design, Brunel University.*

[12] Acero A., (1999), “Formant Analysis and Synthesis Using Hidden Markov Models”, *Proc. of the Eurospeech Conference*, Budapest.

[13] 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.

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

[15] Zwicker E., Flottorp G., Stevens S.S., (1957), “Critical bandwidth in Loudness Summation” *J. Acoustic. Soc. Am.* 29 pp.548-557.

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

[17] Jaynes E.T., (1982), “On the rationale of maximum entropy methods,” *Proc. IEEE,* vol. 70, pp. 939-952, Sep.