

# CROSS ENTROPY INFORMATION METRIC FOR QUANTIFICATION AND CLUSTER ANALYSIS OF ACCENTS

Seyed Ghorshi Saeed Vaseghi Qin Yan

School of Engineering and Design, Brunel University, London  
{Seyed.Ghorshi, Saeed.Vaseghi, Qin.Yan}@brunel.ac.uk

## ABSTRACT

This paper proposes a method for the measurement and quantification of the impact of accents on speech models. An accent metric is introduced based on the cross entropy (CE) of the probability models of speech from different accents. The CE metric has potentials for use in analysis, identification, quantification and ranking of the salient features of accents. The accent metric is used for phonetic-tree cluster analysis of phonemes, for cross-accent phonetic clustering and for quantification of the distances of phonetic sounds in different English accents. Experimental evaluation presented quantifies the effect of American, British and Australian accents on acoustic realisation of phonemes.

## 1. INTRODUCTION

Accent may be defined as a distinctive pattern of pronunciation, including phonemic systems, lexicon and intonation characteristics, of a community of people who belong to a national, regional or social grouping.

Accents evolve over time influenced by large immigrations and socio-economic and cultural trends. For example, the Australian accent is considered to originate from the London “Cockney” accent, the Liverpool accent has been influenced by the Irish immigrations and the Northern Ireland Ulster accent has been influenced by immigrations from Scotland [1].

Wells [1] provides an excellent introduction to the accents of English language within and beyond the British Isles. There are two general methods for the classification of the differences between accents:

- (1) *Historical approach*. Compares the historical roots of accents and the evolutionary changes that accents have gone through as various accents merge or diverge. The historical approach compares the rules of pronunciation and how the rules change and evolve.
- (2) *Structural, synchronic approach*, first proposed by Trubetzkoy in 1931, models an accent systematically in terms of the following differences in:
  - a) Phonemic systems, i.e. the number and identity of phonemes.
  - b) Lexical distributions of words.
  - c) Phonotactic (structural) distributions as in the pronunciation of ‘r’ in rhotic and non-rhotic accents.
  - d) Phonetic (acoustic) realisation.
  - e) Rhythms, intonations and stress patterns of accents.

In this work the differences of accents are modelled using a structural system-based approach.

Accent is a relatively under-explored aspect of automatic speech recognition (ASR), speaker identification and text-to-speech systems [2,3]. Experiments presented at the end of this paper show that the impact of accent on ASR is similar to or greater than even the impact of gender, in that the performance of a ASR trained say on British accent and tested on American accent can be as bad or worse than the performance of a ASR trained on one gender and tested on another of the same accent. Modelling and quantification of accents is also useful for speaker identification systems to exploit systematic pronunciation variation that might characterise a speaker’s accent and in text to speech synthesis systems to generate different accents from the voice of one stored speaker.

There is relatively little work on the quantification and measurement of accents with the exception of the ACCDIST recently introduced by Mark Huckvale at UCL [4], which follows up the work of Minematsu on the acoustic structures of speech [5].

This paper investigates a metric, based on cross entropy [6,7,8], for the quantification and clustering of accents. This metric is evaluated with general British (Br) accent also known as received pronunciation (RP), general American accent (Am) and Australian (Au) accents. The use of cross-entropy as an accent metric is interesting as entropy is the basic metric for quantification of information. Hence cross entropy could be interpreted as a measure of information (additional entropy) due to difference in accents.

## 2. CROSS ENTROPY ACCENT METRIC

A suitable choice for an accent metric should be able to take into account the systematic differences in the pronunciations across different accents and also remove the effect of the differences due to the speakers’ characteristics.

A measure of the differences in the pronunciation patterns of words in two accents may be defined by measuring the changes due to insertions, deletions or substitutions of phonemes in each word as well as the changes in the phonetic realisation of phonemes and the effect of accent in stress and intonation

characteristics of syllables and phrases. Even at the relatively simple level of the differences in the phonemic pronunciation and acoustic-phonetic realisations of words in different accents, an accent metric must be able to quantify the effects of a whole set of changes ranging from relatively subtle differences in acoustic realisation of a phoneme to more obvious changes due to substitution, deletion and insertion of phonemes.

In this section we propose the cross entropy as a measure of the differences between the acoustic units of speech spoken in different accents.

### 2.1 Cross Entropy of Accents

Cross entropy is a measure of the difference between two probability distributions [6]. There are a number of different definitions of cross entropy. The definition used here is also known as Kullback-Leiber 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 of accents defined as:

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

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 [6, 7].

The cross entropies between two different left-right  $N$ -state HMMs of speech with  $M$ -dimensional (cepstral) features is the sum of cross -entropies of their respective states obtained as

$$CE(P_1, P_2) = \sum_{s=1}^N \sum_{i=1}^M \int_{-\infty}^{\infty} P_1(x_i | s) \log_2 \frac{P_1(x_i | s)}{P_2(x_i | s)} dx_i \quad (2)$$

where  $p(x_i/s)$  is the probability distribution of the  $i^{\text{th}}$  mixture of speech in state  $s$ . Cross entropy is asymmetric  $CE(P_1, P_2) \neq CE(P_2, P_1)$ . A symmetric cross entropy measure can be defined as

$$CE_{sym}(P_1, P_2) = (CE(P_1, P_2) + CE(P_2, P_1)) / 2 \quad (3)$$

In the following the cross entropy distance refers to the symmetric measure and the subscript *sym* will be dropped. The total distance between two accents can be defined as

$$AccDist = \sum_{i=1}^{N_u} P_i CE(P_1(i), P_2(i)) \quad (4)$$

where  $N_u$  is the number of speech units and  $P_i$  the probability of the  $i^{\text{th}}$  speech unit.

The cross-entropy distance can be used for a wide range of purposes including:

- (a) To find the differences between two accents or the voices of two speakers.
- (b) To cluster phonemes, speakers or accents.
- (c) To rank voice or accent features.

### 2.2 The Effects of Speakers on Cross Entropy

Speech models would inevitably include the characteristics of the individual speaker or 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 speakers is due to the difference in their accents and how much of it is due to the differences of the voice characteristics of the individual 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 difference between the cross entropies of inter-accent models (e.g. when one model is trained on a group of British speakers and the other on American speakers) and intra-accent models obtained from models trained on speaker groups of the same accent. The adjusted accent distance is

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

The relative differences of the intra-accent and the inter-accent cross entropies are shown in section 4.

## 3. MINIMUM CROSS ENTROPY (MCE) FOR 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 trees for different accents of English.

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

$$CE(C_i, C_j) = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} CE(P_{i,k}, P_{j,l}) \quad (6)$$

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

$$[C_i, C_j] = \underset{i=1:N}{\operatorname{arg\,min}} \underset{\substack{j=1:N \\ j \neq i}}{\operatorname{arg\,min}} CE(C_i, C_j) \quad (7)$$

The results of the application of MCE clustering for construction of phonetic-trees of American, Australian and British English are shown in Figures 1,2 and 3.

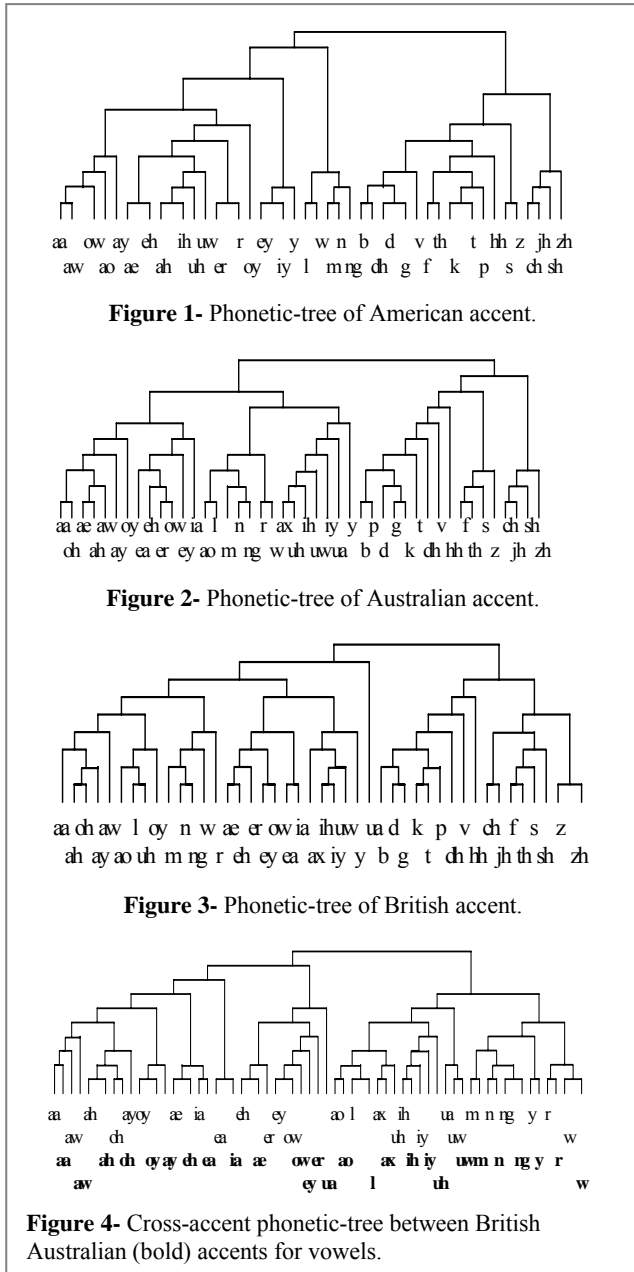
The speech databases used for accent analysis are Australian National Database of Spoken Language (ANDOSL) for Broad Australian English, Wall Street Journal database (WSJ) for general American English and Wall Street Journal Database Cambridge University (WSJCAM0) for Received Pronunciation British English. The subset of ANDSOL of (broad, general and cultivated) Australian accent consists of 18 female and 18 male speakers with a total of 7200 utterances in each category. 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. The style of speech in all databases

is read (as opposed to conversational) speech.

For speech segmentation and labeling, left-right hidden Markov models (HMM) of monophone units are employed. Each HMM has three states and in each state the feature probability distribution is modeled with a Gaussian mixture model with 20 components. The speech feature vectors used to train HMMs consist of 39 features including 13 Mel-Frequency Cepstral Coefficients their 1<sup>st</sup> derivative (velocity) and 2<sup>nd</sup> derivative (acceleration) features. The pronunciation dictionaries used in this work include BEEP dictionary (British accent), CMU dictionary (American accent) and Macquarie dictionary (Australian accent).

The phonetic-tree of the American accent, Figure 1, 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 Australian and British accents, Figures 2 and 3, 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 next section.

Figure 4 shows a cross-accent phonetic-tree. This tree shows how the vowels in British accent cluster with the vowels in Australian accent.



#### 4. CROSS ENTROPY QUANTIFICATION OF ACCENTS OF ENGLISH

In this section we describe experimental results in application of cross entropy for quantification of accents. The plots in Figure 5 illustrate the results of measurements of inter-accent and intra-accent cross entropies. Eighteen speakers were used to obtain each set of models for each group in each accent. The results clearly show that in all cases the inter-accent model differences are significantly greater than the intra-accent model differences. Furthermore, the results show that in all cases the difference between Australian and British are less than the distance between American and British (or Australian). The results further show that the largest differences of Australian accents are in diphthongs.

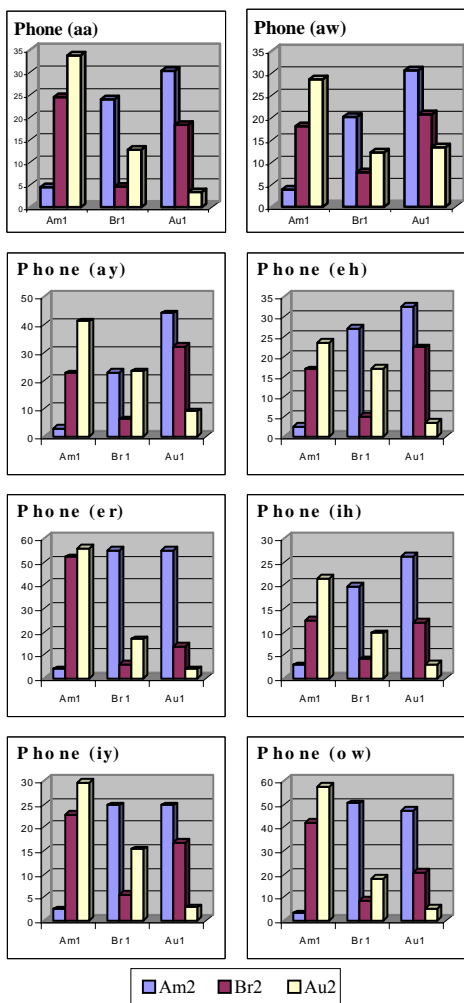
Table 1 shows the cross entropy distances between different accents. It is evident that of the three accent pairs Australian and British are closest. Furthermore American is closer to British than to Australian. Table 2 shows cultivated Australian (CulAu) is closer to British than broad Australians (BrAu) and general Australian (GenAu) as expected.

Accents Metric	Am-Au	Am-Br	Br-Au
Cross Entropy	32	25.15	19.84

**Table 1-** Cross entropy between American, British and Australian accents.

Accents Metric	Br-BrAu	Br-GenAu	Br-CulAu
Cross Entropy	20.8	20.7	18.6

**Table 2-** Cross entropy between British and different Australian (Broad, General and Cultivated) accents.



**Figure 5-** Plots of inter-accent and intra-accent cross entropies of a number of phonemes of American, British and Australian accents. Note each colour-keyed column shows the cross entropy of a group of one speech accent from the another indicated on the horizontal axis.

## 5. A COMPARISON OF THE IMPACT OF ACCENT AND GENDER ON ASR

The results of evaluation of the impact of same/different gender or same/different accent on the error rate of automatic speech recognition are shown in Table 3. These results reveal that an accent mismatch can be more detrimental to the accuracy of automatic speech recognition than a gender mismatch. Table 4 shows the results of accent identification for sentences of an average duration of 5 seconds. The very low accent classification error shows the importance of the effect of accent on speech signals.

## 6. CONCLUSION

In this paper we applied the cross-entropy measure for quantification of accents and for classifications of speech models. The cross entropy is a general information measure that can be

Model \ Input	Br Fem	Br Male	Am Fem	Am Male	Au Fem	Au Male
BrF	<b>30.1</b>	43.3	53.7	60.2	43.6	51.6
BrM	45.7	<b>33.1</b>	62.5	53.4	52.0	44.4
AmF	51.3	62.0	<b>33.6</b>	40.3	53.1	66.8
AmM	61.0	51.3	45.4	<b>34.8</b>	58.6	53.1
AuF	43.3	52.2	53.0	56.4	<b>31.0</b>	44.7
AuM	56.0	48.8	63.8	54.2	48.8	<b>36.0</b>

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

Model \ Input	Br Fem	Br Male	Am Fem	Am Male	Au Fem	Au Male
BrF	<b>0.0</b>	0.0	0.0	0.0	0.0	0.0
BrM	0.0	<b>10.0</b>	0.0	5.0	5.0	0.0
AmF	8.0	0.0	<b>9.0</b>	1.0	0.0	0.0
AmM	0.0	1.0	0.0	<b>1.0</b>	0.0	0.0
AuF	0.0	0.0	0.0	0.0	<b>0.0</b>	0.0
AuM	0.0	0.0	0.0	0.0	1.0	<b>1.0</b>

**Table 4-** % Identification error Between British, American and Australian.

used for a wide range of speech applications from quantification of differences between accents and voices to maximum cross entropy discriminative speech modelling. 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 models due to accents. Further work is being carried out on the use of cross entropy to measure the accents of individuals.

## REFERENCES

- [1] J. C. Wells, "Accents of English," Volume: 1,2 Cambridge University Press, 1982.
- [2] P. Angkititkul and J. H. L. Hansen, "Use of Trajectory Model for Automatic Accent Classification," *Proc EuroSpeech*.pp.1353-1356, Geneva, Switzerland, Sep. 2003.
- [3] J. J. Humphries and P. C. Woodland, "The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training," *Proc. IEEE Inter. Conf. Acoustic, Speech, Signal Processing*, vol. 1, pp.317-320, Seattle, USA, May. 1998.
- [4] M. Huckvale, "ACCDIST: a Metric for Comparing Speakers' Accent," *Proc. On spoken Language Processing, ICSLP 2004*.
- [5] N. Minematsu, "Mathematical Evidence of The Acoustic Universal Structure In Speech," *Proc ICASSP 2005*.pp.889-892.
- [6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [7] J. E. Shore and R. W. Johnson, "Properties of cross-entropy minimization," *IEEE Trans. Inform. Theory*, vol. IT-27, pp.472-482, July. 1981.
- [8] E.T. Jaynes, "On the rationale of maximum entropy methods," *Proc. IEEE*, vol. 70, pp. 939-952, Sep. 1982.