A COMPARISON OF TWO UNSUPERVISED APPROACHES TO ACCENT IDENTIFICATION

Mike Lincoln[1], *Stephen Cox*[1] *and Simon Ringland* [2]

[1] School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.
[2] British Telecom Laboratories, Martlesham Heath, Ipswich IP5 7RE, U.K.

ABSTRACT

The ability to automatically identify a speaker's accent would be very useful for a speech recognition system as it would enable the system to use both a pronunciation dictionary and speech models specific to the accent, techniques which have been shown to improve accuracy. Here, we describe some experiments in unsupervised accent classification. Two techniques have been investigated to classify British- and American-accented speech: an acoustic approach, in which we analyse the pattern of usage of the distributions in the recogniser by a speaker to decide on his most probable accent, and a high-level approach in which we use a phonotactic model for classification of the accent. Results show that both techniques give excellent performance on this task which is maintained when testing is done on data from an independent dataset.

1. INTRODUCTION

Until recently, adaptation of speech recognisers to the voices of new speakers has been viewed purely as a signal-processing problem of compensating for differences in the acoustic signals used to train the recogniser and the signals from the new speaker. However, the variation in the speech signal caused by different accents is fairly systematic and can be dealt with more powerfully at a higher level than that of signal-processing. Different accents give rise to several differences in the realisation of a phrase, the most significant from the point of view of automatic speech recognition being that non-native accents may use a different subset of phonemes from those used by native speakers and/or a given text may be realised as a different sequence of phonemes in different accents. Since the pronunciation dictionaries used by speech recognisers are usually based upon pronunciations for a single accent group, differently-accented speech is likely to have a higher error-rate.

It is impractical to address this problem by adding more pronunciations to the dictionary, as the increase in the number of alternatives increases decoding time and generates additional confusions which may worsen performance [8]. However, some recent studies ([3], [6]) have shown that the use of an accent-specific pronunciation dictionary can improve the performance of a speech recogniser. Although it unlikely that accent-specific pronunciation dictionaries will ever be available for every identifiable accent of a language, such dictionaries already exist for the major accents of English and automatic techniques for constructing them are being researched [6].

In this paper, we compare two techniques to automatically classify accent. Our aim has been to examine the robustness of techniques which require as little as possible prior knowledge of and pre-processing of the speech and which could be fairly easily incorporated into current speaker-independent (SI) recogniser architectures. Accordingly, both of the techniques described here are unsupervised i.e. do not require a transcription of the speech uttered and operate within the framework of a single SI recogniser rather than using a separate recogniser for each accent (as in e.g. [7]).

We have examined a "low-level" technique, which works on the acoustic decoding level, and compared its performance with a technique which uses higher-level knowledge of the phonotactics of the accent. The low-level technique attempts to cluster speakers according to their accent. It bases the clustering on the way in which a speaker "uses" the distributions within the recogniser across the range of speech sounds. An attractive feature of this technique is that it could be added immediately to an SI recogniser with very little effort. We have compared this technique with an established technique for language identification (which has also been used for accent identification) in which a phonotactic model is constructed for each different accent and is used to classify the decoded phoneme string from the recogniser.

2. CLASSIFICATION TECHNIQUES

2.1. Mixture-component usage (MCU)

This technique is based on the premise that, if we assume that speech from all accents of interest has been used to train the recogniser, speech spoken with one of these accents will occupy a distinct set of regions in the pattern-space. This assumption will undoubtedly not hold for all speech from all speakers, and for some sounds, different accents may well occupy many of the same regions. However, if enough sounds are available, these effects should cancel to make classification possible using regions where the assumption is good. By estimating and recording at training-time the regions in the pattern space used by speakers with known accents, we can then classify the accent from a new speaker. It would be possible to identify the regions in which speech with a certain accent lies by clustering directly data from several speakers who use this accent, but such an approach would be difficult to apply within a recogniser. An alternative is to model the distributions of the speech sounds using a Gaussian mixture-density and to then identify which components of this mixture are most frequently used by speakers of a certain accent. If a small number of components were used to model the data within a state, this approach would be too coarse. For this reason, we use a semi-continuous HMM (SCHMM) [5] to model the speech. In an SCHMM, each state distribution is modelled as a weighted sum of a large set of Gaussian component densities (we used 256) which are shared between all the

states. This set of components covers the complete speech space i.e. all accents used in the training data.

The technique uses a form of speaker clustering based on usage of mixture components. At training time, we record for each speaker, the index of the most likely mixture-component associated with each speech input frame. We also record the identity of the most likely state for each frame. When all the speech from the speaker has been processed, we find the identity of the most frequently occurring mixture component associated with each state of each speech model. Hence when training is complete, speaker S_i has associated with him/her a vector U_i of dimension 44 models $\times$ 3 states = 132. The components of $U_i, U_i(k), k = 1, \ldots, 132$, are the indices of the mixture components most often used by the speaker in each state of each model.

The speaker clustering then proceeds as follows:

1. Construct a matrix D of distances between each mixture component in the SCHMM, where $D(k,l)$ is the distance between components k and l;
2. Estimate the distance d_{ij} between each speaker pair S_i, S_j in the training-set: $d_{ij} = \sum_{k=1}^{k=132} D(U_i(k), U_j(k))$;
3. Cluster the speakers into N clusters, where N is the number of accents and record the centroids of the clusters. Associate an accent with each cluster.

We found that this speaker clustering procedure separated the accent groups reasonably well: one cluster contained 29 American speakers and 13 British speakers and the other no American and 16 British speakers. At testing time, the procedure of estimating a vector of the most-often used mixture components in each state is applied to the speech from the speaker. The resulting vector is then classified as belonging to one of the N accent clusters and hence the accent is classified.

2.2. Phonotactic model

Previous studies have shown that phonotactics (i.e. the syntax of phonemes in a language) can be utilised to aid identification of both language (e.g. [9]) and accent (e.g. [7]). These studies have generally estimated language- or accent-specific diphone probabilities using the phonetic output from the recogniser. The phoneme recogniser used in our experiments had a phone accuracy of about 45%, so that only about 20% of the diphones available for use in the models are correct. There may be difficulties in using the recogniser output to train diphone probabilities if the recogniser errors are inconsistent (i.e. if a certain input phoneme sequence is decoded differently on different occasions) or if there are certain error patterns which are "preferred" by the recogniser, regardless of the accent of the input speech. Therefore, we have experimented with estimating these probabilities directly from an accent-specific pronunciation directory (which we assume would be available in a real system) and then using only legal diphones (i.e. diphones observed in the pronunciation directory) for classification. In this case, many diphones output by the recogniser will not be legal diphones. Such diphones do not contribute to the classification of accent and are effectively ignored in our algorithm. Incorrect diphones which are legal contribute noise to the classification which should average out if enough diphones are used. We also experimented with using measures of confidence [4] to identify correctly decoded diphones but found that this gave only a very small improvement. A single set of phoneme-level acoustic models was used to provide output for both phonotactic models.

The phonotactic bigram models were constructed for both American and British English by using the phonetic pronunciations supplied in the BEEP pronunciation dictionary [1] for the British model and the pronunciations in the CMUDICT dictionary [2] for the American model. The probabilities of occurrence of diphone d_i in American accented speech ($\Pr(d_i|A)$) and in British accented speech ($\Pr(d_i|B)$) were estimated directly from the entries in these dictionaries by counting. The amount of information $I(d_i)$ for discrimination of the accent supplied by diphone d_i can be estimated as follows:

$$I(d_i) = \sum_{j=1}^{2} \Pr(A_j, d_i) \log_2 \frac{\Pr(A_j, d_i)}{\Pr(A_j)\Pr(d_i)} \quad \text{bits}, \qquad (1)$$

where $A_1 = A$ (American accent) and $A_2 = B$ (British accent). A high value for $I(d_i)$ implies that d_i supplies a high amount of information about the identity of the accent, but does not tell us which accent is more likely. Hence we define

$$J(d_i) = \text{sgn}(\Pr(d_i|B) - \Pr(d_i|A)) I(d_i). \qquad (2)$$

$J(d_i)$ is positive for any diphone that occurs more frequently in British accented speech than in American and negative if the situation is reversed. Any diphone not occurring in the pronunciation directory has $I(d_i) = 0$.

The distribution of the diphones in the dictionaries is highly skewed, some diphones occurring thousands of times and some a handful. Hence the estimates of the probabilities of occurrence for diphones which occur very infrequently are subject to large uncertainty. Associated with each diphone probability estimate $\Pr(d_i)$ is a variance $V_i = \Pr(d_i)(1 - \Pr(d_i))/N_i$ where N_i is the number of times diphone d_i occurred in the dictionaries. In order to alleviate the problem of poor estimates of $\Pr(d_i)$ caused by infrequently occurring diphones (which could have spuriously high information associated with them), we approximated the variance of $I(d_i)$ by V_i and normalised $I(d_i)$ by dividing by $\sqrt{V_i}$.

To classify the accent, the input speech is decoded using the phone recogniser and adjacent pairs of phones are concatenated into diphones. We use a sequential technique in which classification is achieved when at time T a score J_T is outside one of two thresholds. J_T is derived as follows: we propose a null hypothesis $\mathcal{H}_0$ that the speaker is "mid-Atlantic" i.e. that the frequency of his/her diphone usage is taken in equal proportions from American and British accented speech. Define $I_k = I(d_{f(k)})$ where $f(k)$ gives the index of the k'th diphone in the sequence of diphones output by the recogniser. Under $\mathcal{H}_0$, the expected value of $J_T = \sum_{k=1}^{T} I_k$ is zero and the variance $Var(J_T)$ of $J_T = \sigma_I^2 / T$ where σ_I^2 is the variance of the set of values of $I(d_i)$. Hence if at time T, the value J_T is outside $\pm 2 * SD(J_T)$ (where $SD(J_T) = \sqrt{Var(J_T)}$), then with 95% confidence, the accent is British if J_T is positive and American if J_T is negative.

Figure 1 shows the value of J_T for a typical American-accented sentence. The two 95% confidence thresholds (which follow a $1/\sqrt{T}$) curve) are shown as dotted lines. It can be seen that the lower threshold is exceeded after about 30 diphones have been processed and the accent is then classified as American.

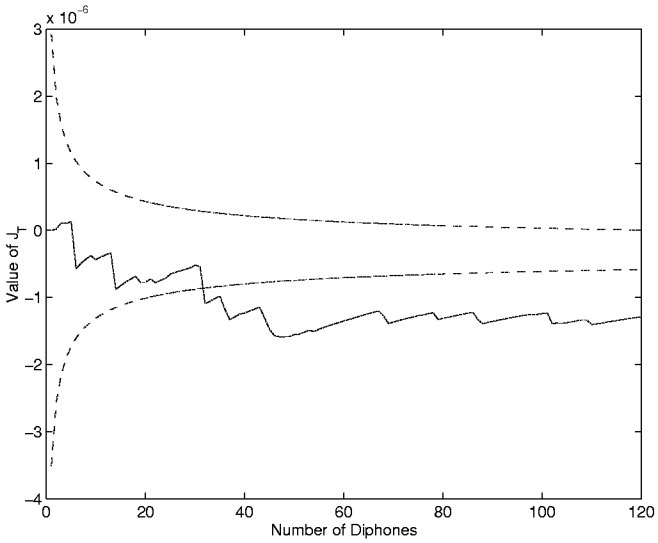

Figure 1: Value of J_T for an American-accented sentence

3. EXPERIMENTAL DETAILS AND RESULTS

3.1. The data and the models

We used the WSJ database to provide American-accented speech and the WSJCAM0 database to provide the British-accented speech. Speech from 98 speakers from WSJ and 98 speakers from WSJCAM0 was used for training, a total of 8596 sentences. This speech was processed to give a 12-component MFCC vector every 100 ms to which velocity, acceleration coefficients and a log-energy coefficient were added. Cepstral mean normalisation was applied to each sentence processed. The WSJCAM0 utterances were provided with an (automatically generated) phonetic segmentation of each utterance and the WSJ data was segmented automatically using pronunciations in the CMUDICT dictionary to force alignment. The speech-data was pooled and used with the appropriate segmentations to train a set of 44 monophone models and a silence model. Each model consisted of three emitting states with no skips allowed between states. For the MCU experiments, the models shared a common set of 256 mixture components in an SCHMM structure (section 2.1). The speaker-clustering (section 2.1) was done using a subset of 29 speakers from each of the American and British databases. For the phonotactic experiments, "conventional" HMMs using two-component mixture distributions for each state were used. In both cases, a mixture component had a separate diagonal covariance-matrix associated with it. For testing, speech from a set of 40 speakers from WSJ and 19 speakers from WSJCAM0 was used.

3.2. Results on original databases

For both methods, classification accuracy was tested after $1, 2, \ldots, 8$ sentences were provided by each of the test-set speakers. Results for the MCU technique are shown in figure 2. 10 of the 59 speakers are mis-classified after 3 sentences are available but this falls to 4 speakers after 4 sentences are available and 2 speakers after 6. Classification using the phonotactic technique

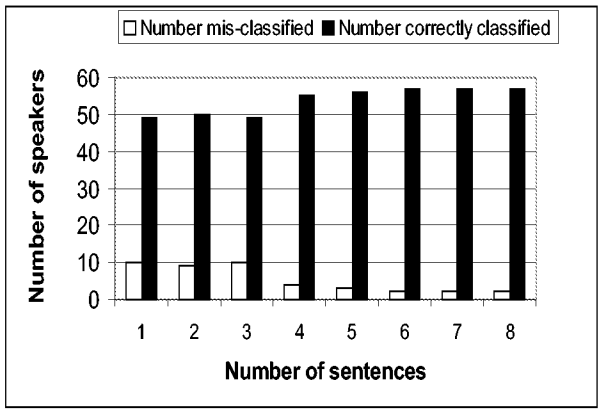

Figure 2: Results using mixture component usage

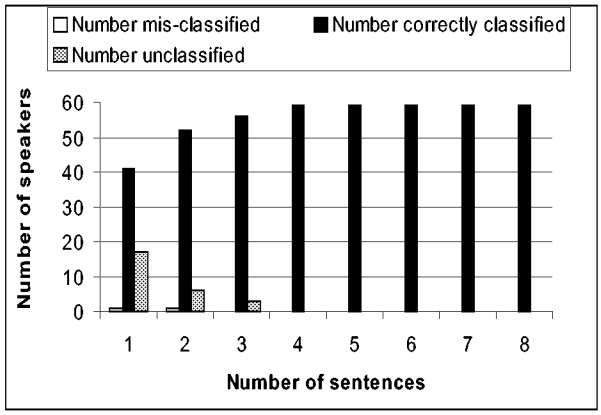

Figure 3: Results using phonotactic models

is done by noting the duration for which the score J_T lies outside each of the two 95% confidence thresholds. The accent is classified as the accent whose threshold was exceeded more often. In practice, very few speakers produced scores which lay outside both thresholds and the most commonly-observed behaviour was for J_T to exceed one of the thresholds and then remain outside it (as shown in figure 1). However, if the score remains within the thresholds after all the diphones have been seen, the result is "unclassified". The results in figure 3 show that when there is only a small amount of data available, the technique is liable to produce the result "unclassified". However, after 4 sentences are available, there are no unclassified or misclassified speakers.

3.3. Results on an independent database

The American and British accented speech was derived from two separate databases recorded under different conditions. Cepstral mean normalisation was used on the data in an attempt to alleviate any overall spectrum differences between the two datasets, but we were concerned that the "accent recognition" demonstrated here might be no more than identification of two sets of data which differed in their acoustic characteristics and which were represented in both the training and the test data. We therefore ran an experiment to verify the techniques on a

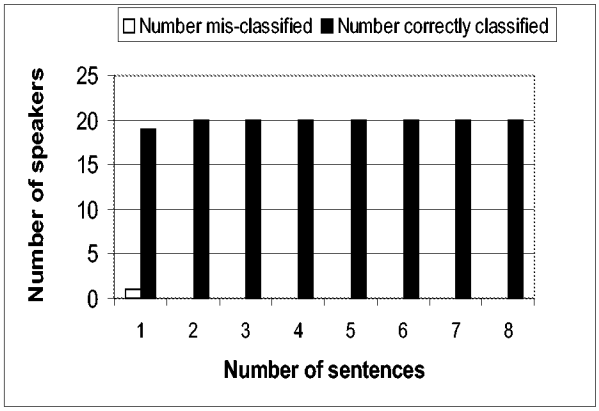

Figure 4: Results using mixture component usage on TIMIT data

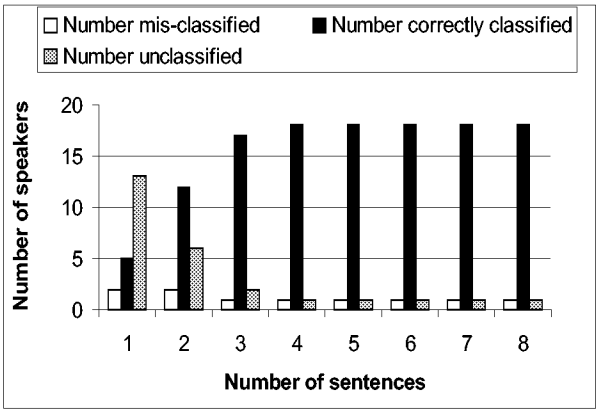

Figure 5: Results using phonotactic models on TIMIT data

third independent set of data. Sentences from twenty speakers from the American-accented TIMIT database (dialect region one) were tested using the same method as described in section 3.2. Results are shown in figures 4 and 5. For both techniques, the same pattern of fewer unclassified and misclassified speakers as more data becomes available is shown and the final classification performance is comparable to that achieved on non-independent data. These results encourage us to believe that both techniques are robust, at least for data recorded under laboratory conditions. At time of writing, we are validating the techniques on an independent British-accent database.

4. DISCUSSION

In this paper, we have investigated two approaches (low- and high-level) to automatically identifying accent and reported results on the problem of discriminating American- and British-accented speech. Both approaches used simple techniques which did not require training multiple recognisers for each accent and which could be easily integrated into a real recogniser. Both were effective and achieved high classification performance. Moreover, when a completely independent dataset was used, performance was maintained. We are encouraged by these results and now intend to compare the techniques on a more difficult problem in which there are several accents. We aim to improve the MCU technique by associating with each state, for each speaker, a *distribution* rather than a single mixture component and by refining the classification technique to use a k nearest-neighbour approach. The phonotactic technique will be enhanced by improving the diphone probability estimates and extending the information measure to multiple accent classes.

ACKNOWLEDGMENT

This work was funded by a grant from British Telecom Laboratories.

5. REFERENCES

[1] The British English Example Pronunciation (BEEP) dictionary is available from ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/.

[2] Available from ftp://ftp.cs.cmu.edu/project/fgdata/dict/cmudict.0.4.

[3] V.L. Beattie et al. An integrated multi-dialect speech recognition system with optional speaker adaptation. In *Proc. 4th European Conference on Speech Communication and Technology*, pages 1123–1126, September 1995.

[4] S.J. Cox and R.C. Rose. Confidence measures for the SWITCHBOARD database. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 511–515, 1996.

[5] X.D. Huang, I. Ariki, and M.A. Jack. *Hidden Markov models for speech recognition.* Edinburgh University Press, 1990.

[6] J.J. Humphries and P.C. Woodland. The use of accent-specific pronunciation dictionaries in acoustic model training. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 317–320, May 1998.

[7] K. Kumpf and R.W. King. Automatic accent classification of foreign accented australian english speech. In *Proc. of Fourth International Conference on Spoken Language Processing. ICSLP'96*, pages 1740–1743, 1996.

[8] K.F. Lee. *Automatic speech recognition—the development of the SPHINX system.* Kluwer Academic Publishers, 1989.

[9] M.A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1):31–44, January 1996.