TEXT-INDEPENDENT SPEAKER IDENTIFICATION AND VERIFICATION USING THE TIMIT DATABASE

Nuala C. Ward[1], *Dominik R. Dersch*[2]

Department of Electrical Engineering, University of Sydney, NSW 2006, Australia
[1]Nuala.Ward@alcatel.com.au, [2]dersch@speech.usyd.edu.au

ABSTRACT

This paper presents a neural network inspired approach to speaker recognition using speaker models constructed from full data sets. A similarity measure between data sets is used for text-independent speaker identification and verification. In order to reduce the computational effort in calculating the similarity measure, a fuzzy Vector Quantisation procedure is applied. This method has previously been successfully applied to a database of 108 Australian English speakers [1].

The purpose of this paper is to apply this method to a larger benchmark database of 630 speakers (TIMIT Database). Using the full 630-speaker database, an accuracy of 98.2% (one test sentence) and 99.7% (two test sentences) was achieved for text-independent speaker identification. On a 462-speaker subset of the database a 98.5% successful acceptance and 96.9% successful rejection rate for text-independent speaker verification was achieved.

1. INTRODUCTION

There has been great interest in the area of Automatic Speech Processing over recent years, and many different methods have been used for speaker recognition, for example Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Vector Quantisation and Artificial Neural Networks (ANNs) [2- 4].

There are many applications for speaker recognition, including telephone banking and security control for access to restricted systems. An application which has recently been proposed allows a verified speaker to gain access to a protected WWW page [5].

This paper investigates models created from complete data sets constructed from 12 mel-frequency cepstral coefficients (mfcc) of the speech signal. A fuzzy Vector Quantisation (VQ) approach is used to reduce the computational complexity of calculating the similarity between a speaker model and an utterance.

This method has been demonstrated to achieve an accuracy of 100% for text independent speaker identification, a 100% successful acceptance and 99.81% successful rejection rate for text-independent speaker verification on a set of 108 speakers from the ANDOSL database [1].

In order to allow a comparison with results in the literature [6, 7] and to investigate the performance of this approach on a larger database, this method was applied to the TIMIT Database [8] - a large benchmark database widely used for speaker recognition systems.

2. EXPERIMENTAL SET-UP

2.1. The TIMIT Database

The TIMIT Database contains 10 sentences from each of 630 speakers (438 males and 192 females). The text corpus consists of three different types of sentences: dialect sentences (*sa* sentences), phonetically compact sentences (*sx* sentences), and phonetically diverse sentences (*si* sentences). The dialect sentences are spoken by all speakers, whereas the other sentences are different for each speaker. The arrangement of the speech data for training and testing was chosen to provide a comparison with the work done by Reynolds [6].

The database is divided into two sections, a “test” section and a “train” section, containing 168 and 462 speakers respectively. These sub-sections were used to divide the database into different sets of speakers to investigate the effect of the number of speakers on accuracy.

2.2. Pre-Processing of Speech Data

The speech data (utterances) were sampled at a rate of 20kHz, parameterised by 12 mel-frequency cepstral coefficients (mfcc) to produce a series of 12-dimensional data vectors. The sampling was performed using a Hamming window of 16-msec duration and 5-msec step-size. The mfcc spectrum was pre-emphasised by a filter coefficient of 0.97. Silence detection was performed by cutting frames which are less than a threshold of 0.1 of the normalised log energy, and removes any noise from the speech data.

3. METHOD

3.1. Speaker Model Construction

Figure 1 provides an overview of the procedure used to construct the speaker models:

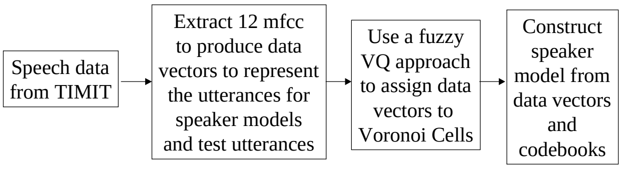
Speech data from TIMIT
Extract 12 mfcc to produce data vectors to represent the utterances for speaker models and test utterances
Use a fuzzy VQ approach to assign data vectors to Voronoi Cells
Construct speaker model from data vectors and codebooks

Figure 1: Constructing models from the speech data

"*The best model for a data set is the data set*" [1]

Using this concept, models were constructed for each speaker within the 12-dimensional mfcc space from the full data set, with no reduction in the input data prior to constructing the speaker models. This ensures that the model contains enough data vectors to capture the characteristic features of the speaker within the local distribution of data vectors for the speaker [1]. Many methods reduce the complexity of the data set prior to classification, but we argue that reducing the complexity of the data set incurs a loss of information which is relevant for a classification task.

VQ is an approach used for information compression, and is widely used in speech and image compression to reduce the complexity of the data set [9]. The fuzzy VQ procedure used in this approach is obtained from a minimal free energy criterion [10, 11] and the underlying update rule belongs to a class of co-operative competitive learning rules [12]. This particular VQ procedure has been chosen because of its excellent convergence properties, and the representation properties of the data set by the codebook are well suited to this problem.

VQ is used to map the speaker model $\mathbf{X}\subset\Re^{12}$ with $\mathbf{X} \equiv \{\mathbf{x}_i\in\Re^{12}|i=1,..,N_x\}$, where N_x is the number of data vectors in the speaker model $\mathbf{X}$, onto a finite set of N (N_x>>N) codebook vectors $\mathbf{w}_r\in\mathbf{W}\{\mathbf{w}_r\in\Re^{12}|r=1,..,N\}$. Each data vector $\mathbf{x}\in\mathbf{X}$ is assigned to the codebook vector $\mathbf{w}_r'\in\mathbf{W}$ by the condition:

$$\left\|\mathbf{x} - \mathbf{x}'_r\right\| = \min_r \left\|\mathbf{x} - \mathbf{w}_r\right\| \qquad (1)$$

Simulated annealing is used to find the optimal positions for the codebook vectors in the feature space. A *Voronoi tessellation* is used to allocate equal numbers of data vectors to each Voronoi Cell. The combination of the codebook vectors and the data distribution uniquely describe the speaker's subspace, and form a model for the speaker (Figure 2).

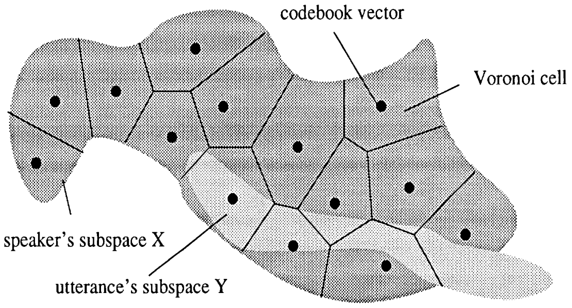

Figure 2: The model for a speaker.

3.2. Feature Extraction

A speaker's utterance **Y** follows a trajectory through the speaker's subspace **X** (Figure 2), where the utterance subspace is $\mathbf{Y} \equiv \{y_k\in\Re^{12}|k = 1,\dots,N_y\}$. Feature extraction is performed by calculating a similarity measure between the data set **X** and the test utterance **Y** called the mean next neighbour distance d_{nn} with

$$d_{nn}(\mathrm{X},\ \mathrm{Y}) = \frac{1}{\mathrm{N}_y}\sum_{N_y}\left\|\mathbf{x}'(\mathbf{y}) - \mathbf{y}\right\| \qquad (2)$$

where $\mathbf{x}'(\mathbf{y})$ is the next neighbour in the speaker's subspace **X** to the data vector $\boldsymbol{y}$ in the test utterance. An average is taken over the N_y data vectors in **Y**. Here a high similarity measure corresponds to a small mean next neighbour distance, d_{nn}.

The search for the next neighbour using equation (2) is computationally expensive. How do we make the next neighbour search more efficient without losing information which may be relevant for classification? We perform a hierarchical search to reduce the computational complexity of the feature extraction procedure.

A large number of data vectors far away from $\boldsymbol{y}$ may be excluded from the next neighbour search, allowing for a reduced search. For each data vector $\boldsymbol{y}$, the next codebook vector $\mathbf{w}_r'$ is obtained using equation (1), then the Voronoi Cell of $\mathbf{w}_r'$ is searched for the next neighbour $\tilde{\mathbf{x}}'(\mathbf{y})$ in the speaker's model. This corresponds to a slightly different value for the mean next neighbour distance:

$$\tilde{d}_{nn}(\mathrm{X},\ \mathrm{Y}) = \frac{1}{\mathrm{N}_y}\sum_{N_y}\left\|\tilde{\mathbf{x}}'(\mathbf{y}) - \mathbf{y}\right\| \qquad (3)$$

which allows for a computation saving of approximately N [1]. This distance measure implements a Manhattan metric rather than an Euclidean distance measure to further reduce the number of numerical operations. The values for d_{nn} and $\tilde{d}_{nn}$ may differ slightly, for example when $\mathbf{x}'(\mathbf{y})$ is not in the subset of $\mathbf{w}_r'$, when $\boldsymbol{y}$ is on the border of a Voronoi cell, $\tilde{d}_{nn}$ will be an over-estimation of the mean next neighbour distance.

3.3. Set-Up of Identification and Verification Experiments

Eight sentences from each speaker (2 *sa*, 3 *si* and 3 *sx* sentences) were used to create a model for the speaker. The speaker models contain the complete set of mfcc vectors extracted from the sentences and 10 codebook vectors. The test sentences for identification and verification were always different from the sentences in the speaker's model to ensure text-independency. Both identification and verification experiments were carried out on sets of all male (M), all female (F), and mixed (M+F) speaker sets of varying size.

The two remaining *sx* sentences from each speaker were used in identification testing. In verification, a threshold value was set so that the identity claim could be accepted or rejected. In order to set this threshold, a "training" stage was performed, using a set of 10 background speakers randomly selected from the speaker set. These 10 background speakers were different for each speaker, and the remaining speakers in the set were used in the test set as imposters.

The first *sx* sentence from the true and background speakers were used to set the threshold where the number of false acceptances (acceptance of an imposter) equals the number of false rejections (rejection of a true speaker). The second *sx* sentence from the true and imposter speakers were used to test the threshold. An identity claim was accepted if the d_{nn} value falls below the threshold and was rejected otherwise.

4. RESULTS

4.1. Speaker Identification

Table 1 shows the results for speaker identification using the two *sx* sentences from each speaker individually.

Speaker Set	# speakers	# incorrect	Correct %
M	112	0	100.00
F	56	1	99.11
M+F	168	1	99.70
M	326	12	98.16
F	136	5	98.16
M+F	462	17	98.16
M	438	16	98.17
F	192	7	98.18
M+F	630	23	98.17

Table 1: Identification results for one sentence

Four of the incorrect speaker identifications in the male 326-speaker set were caused by the same very short utterance which contained an average of only 153 data vectors (corresponding to 0.75 seconds of speech). It was found that the shorter sentences had a higher incorrect identification rate.

Two sentences were concatenated for the second set of identification tests to explore the effect of a longer utterance length on identification accuracy (Table 2).

Speaker Set	# speakers	# incorrect	Correct %
M	112	0	100.00
F	56	0	100.00
M+F	168	0	100.00
M	326	1	99.69
F	136	1	99.26
M+F	462	2	99.57
M	438	1	99.77
F	192	1	99.48
M+F	630	2	99.68

Table 2: Identification results for two sentences

As the number of speakers was increased from 462 to 630 speakers, the identification accuracy increased for both the one sentence and two sentence experiments. Dersch [1] and Reynolds [6] found that females achieve a lower speaker identification accuracy than males. However our results show that for the single sentence tests the female accuracy was actually higher, but using the longer utterance length, the accuracy was distinctly lower for females.

The rank of the true speaker is a performance measure used to indicate how accurately the true speaker was correctly identified. If the speaker is correctly identified they are assigned a rank of one, if not, the speaker is assigned a rank according to the number of speakers who had a smaller d_{nn} value than them. Table 3 shows that the rank increases with an increase in the number of speakers. There is more scope for incorrect identification in a larger group of speakers. The rank is much smaller for the longer utterance length, which is to be expected, since a longer utterance provides greater opportunity to capture the characteristic features of a speaker.

speaker set	# speakers	1 sentence	2 sentences
M	112	1.000	1.000
F	56	1.009	1.000
M+F	168	1.003	1.000
M	326	1.044	1.003
F	136	1.037	1.007
M+F	462	1.043	1.004
M	438	1.048	1.006
F	192	1.034	1.007
M+F	630	1.044	1.006

Table 3: Mean rank of true speaker for one and two sentences

Another performance measurement to evaluate the quality of results measures the relative distance between the smallest and the second smallest mean next neighbour distances. Using one sentence for the 168 speaker subset, we found a value of 12.8%, and for the 462 speaker subset, 10.1%. This value is used as a confidence measure, and a larger percentage indicates that the correct identification is more reliable. It is interesting to note that for the incorrect utterance identifications, we found a very small value of 0.64% indicating that the incorrect identification was caused by a speaker who was very similar to the true speaker.

4.2. Speaker Verification

Table 4 shows the results for speaker verification, and includes the number of true and imposter speakers within each group.

Speaker Set	#true	#imp	Correct Accept %	Correct Reject %
M	112	11312	98.2	98.0
F	56	2520	91.1	98.5
M+F	168	26376	99.4	97.7
M	326	102690	98.2	97.2
F	136	17000	94.9	95.4
M+F	462	208362	98.5	96.9

Table 4: Verification results using random background speaker selection

Both in the male and mixed speaker groups, the correct acceptance rate is higher than the correct rejection rate. The correct acceptance and rejection rates decreased with an increase in speaker set size, with the exception of the female set. Generally we found that the female set results were much lower than the other sets, and it is unclear why this is so, but it agrees with other results in the literature [6, 13].

5. SUMMARY AND CONCLUSIONS

The values obtained for the mean next neighbour distance, $\tilde{d}_{nn}$ using equation (3), were found by searching only one Voronoi Cell, therefore achieving the highest possible computational saving with this method. This may result in an over-estimation in the calculations for the mean next neighbour distance. Dersch [13] found that an increased search to surrounding Voronoi Cells achieves better results, but this increases the complexity of the search. So a trade-off exists between the accuracy of the method and the computational complexity of the next neighbour search.

The time required to verify an identity claim is the amount of time required to calculate the similarity measure for the test utterance on the claimed speaker's model, and can be carried out in real time. Identification on the other hand, requires that the test utterance be compared with every speaker model in the database. It is a linear function of the number of speaker models and can be very time consuming for large databases. The mean amount of space required to store each speaker model is 376.9 kB, and the time required to train a speaker model is much faster than any other approach.

It was important to follow the set-up used in [6] to allow for a direct comparison of the results. Our method has been shown to give 98.2% accuracy for one sentence speaker identification on the full 630 speaker database. An increase to 99.7% accuracy was achieved by concatenating two sentences from each speaker. These results are comparable to the results achieved by Reynolds: 99.5% for speaker identification for the full 630 speaker database.

Background speaker selection methods used for verification in this study differ from those of Reynolds [6]. Our *eer* result is 2.3%, compared with 0.24% achieved by Reynolds. The results achieved by Fakotakis et al. [7] were a 98.09% identification accuracy and an *eer* of 1.72% for verification. Although the same arrangement of the database was not used, it is still useful to compare results with those achieved using the TIMIT Database.

As a result it has been shown that this method achieves very good performance on a large benchmark database. In both speaker identification and verification, the results confirm previous results obtained on a smaller database of 108 Australian speakers [1]. It would be interesting to evaluate how this new approach performs on noisy speech, eg telephone speech or speech with background office noise.

The current method does not disadvantage the utterance if the path is not aligned with the speaker's model. A new feature is being investigated which exploits the dynamic properties of the utterance, incorporating the time sequence of the data vectors. Investigation of this feature will be the subject of further work [14].

6. REFERENCES

1. Dersch, D.R., "The Acoustic Fingerprint: A Method for Speaker Identification, Speaker Verification and Accent Identification", *Proc. SST-96*, Canberra, 1996, pp. 307-312.

2. Soong, F.K., Rosenberg, A.F., Rabiner, L.R., Juang, B.H., "A Vector Quantisation approach to speaker recognition", *Proc ICASSP-85*, pp 387-390, 1995.

3. Waibel, A., Lee, K-F., *Readings in Speech Recognition*, Morgan Kaufmann Publishers, California, 1990.

4. Chollet, G. "Long index of References on Automatic Speaker Verification", http://www-sig.enst.fr/~chollet/ForMehdi/SpRecV1.l_ind.html,

5. Sokolov, M., "Speaker Verification on the World Wide Web", *Proc. Eurospeech-9*, Vol. 2, September 1997, pp. 847-850.

6. Reynolds, D.A., "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech Communication* 17, 1995, pp. 91-108.

7. Fakotakis, N., Georgila, K., Tsopanoglou, A., "A Continuous HMM Text-Independent Speaker Recognition System based on Vowel Spotting", *Proc. Eurospeech-9*, Vol. 5, September 1997, pp. 2347-2350.

8. National Institute of Standards and Technology (NIST), "TIMIT acoustic-phonetic speech corpus", *NIST speech disk*, 1-1.1, documentation, 1990.

9. Gray, R.M., "Vector Quantisation", *IEEE ASSP Magazine*, Vol. 1, April 1984, pp. 4-29.

10. Dersch, D.R. and Tavan, P., "Control of annealing in minimal free energy vector quantisation", *Proc. ICNN-94*, pp. 698-703.

11. Rose, K., Gurewitz, E., Fox, G.C. "Statistical mechanics and phase transitions in clustering", *Physical Review Letters*, Vol. 65, No. 8, 20 August 1990.

12. Dersch, D.R. "Eigenschaften Neuronaler Vektor-quantisierer und Ihre Anwendung in der Sprachverarbeitung", 1996, Verlag Harri Deutsch, Thun, ISBN 3-8171-1492-3.

13. Dersch, D.R. and King, R.W., "Speaker models designed from complete data sets: A new approach to text independent speaker verification", *Proc Eurospeech-9*, Vol. 5, pp.2283-2286, September, 1997.

14. Ward, N.C. and Dersch, D.R. in preparation.