Proceedings of the Institute of Acoustics

A LARGE SPEAKER IDENTIFICATION AND VERIFICATION EXPERIMENT

E. Abadjieva Rohwer(1), J. A. Lear(2), R. J. Rohwer(3)

(1) LINKON Corporation, 140 Sherman St., Fairfield, Connecticut, USA
(2) British Telecom Research Laboratories, Martlesham Heath, Ipswich, UK
(3) Dept. of Computer Science and Applied Mathematics , Aston University, Birmingham, UK

1. INTRODUCTION

Research in Speaker Recognition (SR) has reached the stage where it can be implemented successfully in many different practical voice systems applications such as voice dialling, automatic access to database information and dialogue systems. For these purposes it is required that the SR system performs well when a large number of users is involved.

SR can be based on several different techniques. Good performance results are reported using a hidden Markov model (HHM) technique[1,2], variants of vector quantisation (VQ) [3], autocorrelation techniques, orthogonal linear prediction (OLP) [4,5] and a neural net approach [6]. In most cases the results reported are for a database of between 6 [6] and 100 [3] speakers. For real-life applications it is important to evaluate a given approach for a larger number of speakers. This is the main purpose of the research presented here, which demonstrated error rates well below 1% on a speaker verification task involving 630 speakers using an OLP technique.

A popular database with a suitable structure available for academic research and comparison is the TIMIT database. It contains speech samples from 8 American dialect regions, recorded from 630 speakers, each of whom provided 10 sentences drawn from 3 sentence types. These features enable the comparative study of the influence of various factors on the performance of the SR system. This research was conducted in two directions:

1. Evaluating the influence of the dialect regions used for training and testing.
2. Evaluating the influence of the type of the text used for training and testing.

2.BACKGROUND

The SR task has two different aspects: Speaker Verification (SV) and Speaker Identification (SI). The main task in Speaker Verification is to accept or reject the claimed identity of a tested speaker. The voice templates of the verified speaker are compared only once to the voice reference of the claimed identity. If the match is close enough, below a given threshold, the speaker is deemed the true one; otherwise the speaker is deemed an impostor. The setting of the acceptance/rejection threshold influences the performance of the verification system significantly. The setting is obtained as part of the training stage.

A LARGE SPEAKER IDENTIFICATION AND VERIFICATION EXPERIMENT

For Speaker Identification the speech sample from an unknown speaker is compared with every one of the stored references from a set of known users. The closest match identifies the speaker. The larger the set of possible users the more difficult the identification task is.

The Orthogonal Linear Prediction technique uses an eigenvector analysis and decomposition to separate speakers by emphasising the differences between their vocal tract parameters. The method uses the eigenvalues and eigenvectors of the covariance matrix of measurements made on a set of utterances from an enrolled speaker to transform the measurements made on an utterance from the unknown speaker. The method can be used in both text-dependent and text independent mode and is language independent.

3. THE SYSTEM

The system reported here is a baseline system, intended to demonstrate the suitability of the OLP approach for speaker identification and verification purposes and to study the importance of various factors on its performance. Substantial optimisation is envisaged. The system is made up of three sub-systems, a signal processing module, a training module and a recogniser module.

3.1 Signal Processing

The signal processing module calculated the 12^{th} order cepstral coefficients via the following method:

1. The speech (sampled at 16khz) was segmented with a frame size of 256 samples and an overlap of 50%.
2. Each frame was pre-emphasised (by 1.067).
3. A Hamming window was applied to each frame.
4. Linear Predictive Coefficients (LPC) were calculated, from each frame, using the autocorrelation and Levinson-Durbin approach.
5. The LPCs were transformed into cepstral coefficients (LPCC) using:

$$c_k = a_k + \sum_{i=1}^{k-1} \left(\frac{i}{k}\right) c_i a_{k-i}$$

where the a_k are LPCs and the c_k are the required LPCCs, for $k = 1...12$.

3.2 Training

For each speaker to be enrolled, the covariance matrix is computed for the combined sequence of LPCC vectors for all the input utterances. (To save re-computation in new experiments, covariance matrices are computed for each utterance and then combined in a sum weighted by utterance length.) Each speaker's covariance matrix is diagonalised, producing eigenvalues λ_{ir} and eigenvectors b_{ir}, for speaker r, with $i = 1...12$. The speakers cepstral vectors are then

projected onto this eigenbasis, and averaged over time. These average values along with the eigenvectors and eigenvalues form the reference data for each user.

3.3 Recognition

The test utterance is subjected to a 12th order LPCC calculation as described above. These coefficients are then orthogonalised and averaged as when training, but using the eigenvectors from the reference speaker's template. A distance between the test and reference speakers is then defined by:

$$D = \sum_{i=1}^{12}\left(\frac{\overline{\phi_{ir}} - \overline{\phi_{it}}}{\sqrt{\lambda_{ir}}}\right)^2 + \frac{1}{2}\sum_{i=1}^{12}\left(\frac{v_{it} - \lambda_{ir}}{\lambda_{ir}}\right)^2$$

where $\overline{\phi_{ir}}$ is the average value of the ith orthogonal cepstral coefficient for the reference speaker,

$\overline{\phi_{it}}$ is the average value of the ith orthogonal cepstral coefficient for the test speaker,

λ_{ir} is the reference eigenvalue for the ith orthogonal parameter for the reference speaker,

v_{it} is the variance of the ith orthogonal parameter of the test speaker.

The first term is simply the Mahalanobis distance, comparing the average coefficients of the test speaker with those of the reference speaker. It can be viewed as a quantity proportional to the difference of mean, in units of the measurement uncertainty of $\overline{\phi_{ir}}$ [4]. The second term compares the variances. As noted in [5], it approximates a more principled comparison, and leaves scope for debate about the weight it should be given relative to the first term.

This method makes use of the statistical characteristics of a speaker's speech without relying on the detailed structure of any particular utterance. This makes it especially well suited to text-independent verification. Provided that enough speech material is used to provide good measurements of each speaker's parameters, the precise content of the material is unimportant.

4. DESCRIPTION OF THE CORPUS TEXT MATERIAL

The structure of the TIMIT database presents certain advantages and limitations to speaker recognition experiments.. The text material consists of:

1. The Dialect (SA) Sentences - total number: 2. These are intended to expose the dialectical variants of the speakers and were read by all 630 speakers.
2. The Compact (SX) Sentences - total number: 450. The phonetically compact sentences were designed to provide good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Each speaker read 5 of these sentences and each text was spoken by 7 different speakers.
3. The Diverse (SI) Sentences - total number: 1890. The phonetically-diverse sentences were selected from existing text sources so as to add diversity in sentence types and phonetic

contexts. The selection criteria maximised the variety of allophonic contexts found in the texts. Each speaker read 3 of these sentences, with each sentence being read only by a single speaker.

The following table summarises the speech material in TIMIT:

Sentence type	Sentences	Speakers	Total	Sentences/Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3

The database evidently presents ample opportunities for text-independent tests involving different types of training and testing material. Text-dependent tests are not possible because no sentence is ever repeated by any single speaker. Text-independent tests using training and testing on just one type of material are possible, but somewhat limited by the need to split the small number of sentences of each type spoken by each speaker into training and testing sets.

Speech in the TIMIT database is sampled at 16kHz. The database also contains information on word boundaries which was used to endpoint the utterances.

5. EXPERIMENTAL PROCEDURE

The phonetic types of the sentences were used as a basis for separating training and test material. Because the database is most amenable to this type of experiment, it was decided to use a single type of sentence for training and a different type for testing. This excludes training and testing on the same type of sentences.

The Speaker Verification experiments were carried out as follows:

1. Training templates were produced for the 630 speakers using either 2 'Shibboleth'(SA), 3 Compact(SX) or 3 Diverse(SI) sentences. This is a similar amount of speech material to the 5 or 6 sentences used in previous studies [4,5], perhaps somewhat less.
2. Each of the speakers was compared against themselves (true speaker), and the other 629 speakers using sentences from one of the other classifications (SA, SI or SX). Thus the total number of trials was 793800 (=630 speakers × 630 speakers × 2 categories). Of these, there were
   - true speaker trials: 1260 = 1 × 630 × 2,
   - impostor trials: 792540 = 629 × 630 × 2.
3. As in previous studies [4,5], only one sentence was used per verification trial.
4. An ideal individual acceptance/rejection threshold for each speaker was calculated based on the Equal Error Rate (IndEER) point, at which the probability of a false rejection matches

that of a false acceptance. The threshold dependence of these probabilities was estimated from the test data because of the small number of utterances available per speaker. This expedient will give slightly better results than a perfectly fair method.

5. A speaker-independent threshold was calculated as an average of the individual thresholds (AvgEER). This will produce poorer results for which the number of false acceptances will generally be different from the number of false rejections. It gives a rough idea of how much the system suffers if very little attention is given to the threshold setting.

Speaker Identification experiments took place simultaneously with the speaker verification experiments. The Speaker Identification experiments were carried out as follows:

1. Training templates were produced for the 630 speakers using either 2 'Shibboleth'(SA), 3 Compact(SX) or 3 Diverse(SI) sentences.
2. A test utterance was compared against all of the 630 templates.
3. The test utterance is classified as coming from the speaker which, when compared against their reference template, has the lowest score.

The error rate (IdenER) is then given by: $100\% - \frac{\text{correct classifications}}{\text{total number of tests}} \times 100\%$

6. RESULTS

6.1 Training on 'Shibboleth' Sentences (SA)

The 2 SA sentences were used for producing the reference templates. Then the 5 SX or the 3 SI sentences were used for testing. The following table shows the results from these experiments:

Test On...	IdenER	AvgEER	IndEER
SI	38.73%	5.85%	1.76%
SX	36.41%	3.63%	2.33%

6.2 Training on Phonetically Diverse Sentences (SI)

The 3 SI type sentences were used for producing the reference templates. Then the 2 SA or the 5 SX sentences were used for testing. The following table shows the results from these experiments:

Test On...	IdenER	AvgEER	IndEER
SA	30.48%	3.13%	0.53%
SX	33.94%	5.69%	2.61%

6.3 Training on Phonetically Compact Sentences (SX)

The 3 SX type sentences were used for producing the reference templates. Then the 2 SA or the 3 SI sentences were used for testing. The following table shows the results from these experiments:

Test On...	IdenER	AvgEER	IndEER
SA	31.35%	3.78%	0.56%
SI	37.62%	5.84%	2.05%

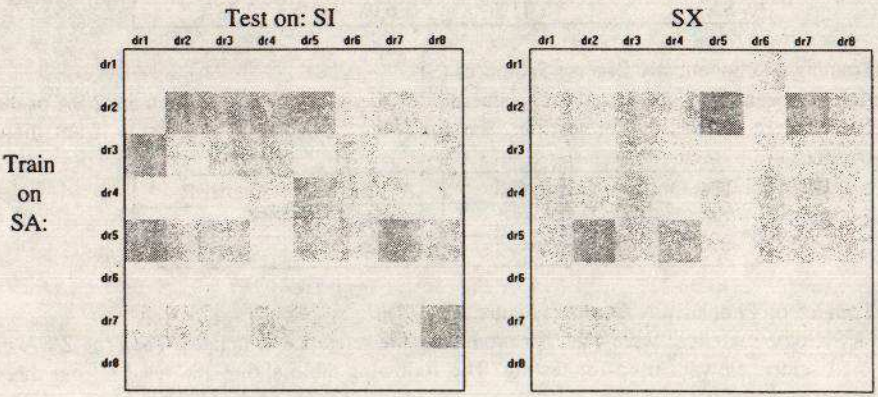
6.4 Discussion

First of all, the results demonstrate that the OLP system holds up well, even when a large number of speakers are enrolled, achieving verification error rates close to 0.5%. Also it is clear that the phonetic composition of the speech material matters greatly, with the Shibboleth sentences (SA), designed to expose dialectic variations, performing much better than the others as test sentences. Interestingly, they do not emerge as the best training sentences. That distinction goes marginally to the phonetically diverse SI sentences, with the SI/SA combination producing the best overall results. Training with SA sentences gives comparable results to the other types when testing on the same type of material. These results might have been somewhat better if there had been more than 2 SA sentences available for training. Three sentences were used for training with SI and SX sentences.

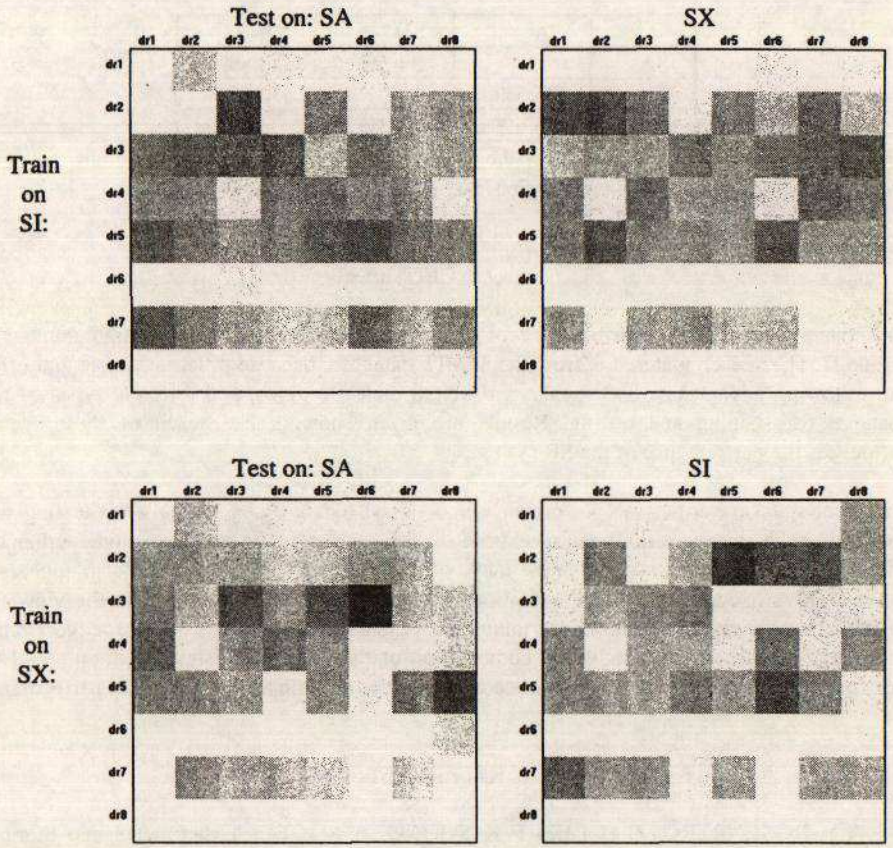
It seems natural to suggest that the SA sentences perform best for testing because their phonetic diversity provides a statistically good sample. Perhaps this matters less for training because of the larger amount of material used.

6.5 The Influence of the different Dialectic Regions

The TIMIT database splits the speakers into subsections based on the Dialectic Region of the US that they come from. The results from these experiments were further analysed to investigate whether confusions were mainly between speakers from their own dialectic region or from others. The following diagrams show the confusions between dialectic regions for the six experiments. Correct identifications have been removed.



A LARGE SPEAKER IDENTIFICATION AND VERIFICATION EXPERIMENT



A tendency to produce confusions within dialectic regions would show up as a darkened diagonal in these plots. It would appear that the identification errors are distributed roughly evenly throughout the dialectic regions. The appearance of dark horizontal stripes gives the slight impression that speakers from some regions, such as dr5 and dr7, are more easily imitated than others.

Another way to approach the issue of whether the method is sensitive to dialectic variations is to ask whether recognition confusions occur predominantly within dialectic regions. If a correct identification is acknowledged when the system simply gets the dialectic region correct, then the error rates are as follows:

	SA	SI	SX
SA	X	34.5%	33.5%
SI	26.5%	X	30.6%
SX	26.6%	33.3%	X

These are only marginally lower than the identification error rates for individuals, so again there is little evidence that dialectic regions have much importance.

7. CONCLUSION

This paper evaluates the performance of a simple SR algorithm using a large number of speakers. The speech material is from the TIMIT database, because it was available and offers the following advantages: 630 speakers, different dialect regions, and different types of text material for training and testing. Results are given showing how each of these factors influences the performance of the SR system.

Excellent speaker verification performance is obtained (nearly 0.5% error) when testing with the Shibboleth sentences designed to expose dialectic variations in pronunciation. Other test material gives verification scores in the range of 1.7% to 2.61%. The method also appears to be insensitive to regional dialectic variations. This desirable property supports the view that this technique measures the relatively immutable physical characteristics of the speakers rather than their manner of speaking. The combination of the chosen statistical technique and the selected front end feature extraction procedure provides reliable text-independent performance.

8. REFERENCES

[1] A E ROSENBERG, C H LEE, F K SOONG, A McGEE, 'Experiments in Automatic Talker Verification Using Sub-word Unit Hidden Markov Models', ICASSP 1990;

[2] J de VETH, G GALLOPYN, H BOURLARD, 'Limited Parameter Hidden Markov Models for Connected Digits Speaker Verification Over Telephone Channels', ICASSP,(1993).

[3] F K SOONG, A E ROSENBERG, 'A Vector Quantisation Approach to Speaker Recognition', ICASSP, (1985);

[4] M R SAMBUR, 'Speaker Recognition Using Orthogonal Prediction', IEEE Trans. Acoustics, Speech and Signal Processing, vol.ASSP-24, No.4 (1976);

[5] R E BOGNER, 'On Talker Verification Via Orthogonal Parameters', IEEE Trans. Acoustic, Speech and Signal Processing, vol.ASSP-29, No.1 (1981);

[6] M J CAREY, E S PARRIS, J S BRIDLE, 'A Speaker Verification Using Alpha - Nets', ICASSP,(1991);