Impostor Cohort Selection for Score Normalisation in Speaker Verification

R.A. Finan [a], R.I. Damper [b,c] and A.T. Sapeluk [a]

[a] *School of Engineering, University of Abertay Dundee, Scotland DD1 1HG.*
Emails: {r.a.finan|a.t.sapeluk}@tay.ac.uk

[b] *Department of Electronics and Computer Science, University of Southampton, England SO17 1BJ. Email:* rid@ecs.soton.ac.uk

[c] *Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, PO Box 9100, Portland, OR 97291-1000, USA.*
Email: damper@cse.ogi.edu

Abstract

Impostor cohort normalisation (ICN) may be used to improve speaker verification performance. The choice of impostors for the cohort is critical to the success of ICN. To select suitable impostors, they must be ranked according to their ability to impersonate the genuine speaker. In the technique described here, a genuine speaker's impostors are ranked according to how well their models score against the genuine-speaker model. This is faster than the traditional method of testing the model against the impostor training utterances, and is in good agreement with it. Having obtained rankings, impostors were selected in three different ways: of these, cohorts based on the closest impostors gave the best results, reducing the unnormalised equal error rate by more than half.

Key words: speaker verification, impostor cohort, score normalisation.

1 Introduction

The speaker recognition problem can be divided into speaker *verification* and speaker *identification* tasks. Either of these may in turn be *text-dependent* or *text-independent*. In verification, the aim is to determine if a given utterance was produced by the claimed speaker. This is done by testing the model of the claimed speaker with the utterance, comparing the score to a threshold, and deciding on the basis of this comparison whether or not to accept the claimant. In identification, the aim is to determine which speaker from among a known group produced the utterance. In this case, the test utterance is scored against all possible speaker models, and that with the best score determines the speaker identity. Of the two tasks, identification is generally accepted to be the harder, especially for large speaker populations.

In text-independent recognition, there are no limits on the vocabulary employed by speakers. This is in contrast to text-dependent recognition, where the presented utterance must be from a set of predetermined words or phrases. As text-dependent recognition only models the speaker for a limited set of phonemes in a fixed context, it generally achieves higher recognition rates than text-independent recognition, which must model a speaker for a variety of phonemes and contexts.

This letter is concerned with score normalisation for speaker verification, with primary interest in the text-dependent case. Score normalisation is a means of improving verification rates which relies on presenting the test utterance to more than just the claimed speaker model. Two of the commonest forms for speaker verification are normalisation by a *world model* and normalisation by a subset (the *cohort*) of the impostor scores. The former is often used in hidden Markov model (HMM) (Matsui and Furui, 1995) and predictive neural network (Hattori, 1994) systems, and the latter in systems such as vector quantisation (VQ) and dynamic time warping (DTW) (Li and Porter, 1988; Higgins *et al.*, 1991; Rosenberg *et al.*, 1992; Booth *et al.*, 1993; de Veth *et al.*, 1993). The focus here is on the second approach.

A vital, but currently unresolved, issue is how to select the most suitable impostors for the cohort (de Veth *et al.*, 1993). To facilitate this, we need to rank the impostors according to how well they impersonate the genuine speaker. This is most easily done *a posteriori*, when all the recordings have been collected, so that testing with the complete impostor population is possible. While this *a posteriori* approach is reasonable in a research setting, it may be impractical in a real system in which new users are being constantly enrolled, and such comprehensive testing would be too time-consuming. Impostor ranking may also be done *a priori* by testing the genuine-speaker model against the impostor training utterances, although this may take a long time depending on the number of impostors involved (Rosenberg *et al.*, 1992; de Veth *et al.*, 1993). We have developed a faster method of determining impostor rankings and used it to vary the selection of impostors for the normalisation cohort.

The remainder of this paper is structured as follows. Section 2 describes the speaker database used for the experiments, while Section 3 gives details of how the speech was analysed and the speakers modelled. Next, Section 4 covers the normalisation of verification scores using impostor cohorts. In Section 5, the means of generating the impostor rankings is explained. Section 6 concerns the effect on the verification error rate of varying the impostors in the normalisation cohort. Section 7 compares the results of using impostor cohort normalisation with those using unnormalised scores. Finally, Section 8 concludes with a brief discussion of cohort normalisation and its possible applications.

2 Speaker Database

The text-dependent British Telecom Millar database was used for the major part of this work. In its entirety, this consists of 60 native English speakers saying the digits *one* to *nine*, *zero*, *nought* and *oh* 25 times each in random order. Recordings were made in 5 sessions spaced over 3 months, in order to capture the variation in speakers' voices over time which is one of the most important aspects of speaker recognition (Furui, 1974; 1997). The speech was recorded digitally in a quiet environment using a high-quality microphone, and a sampling rate of 20 kHz with 16-bit resolution. In this work, a version of the database was employed in which the speech was bandpass filtered to telephone quality and then down-sampled to 8 kHz.

A subset of 31 speakers saying the words *one* to *nine* and *zero* was used for the experiments. They were all male and from the same age group. (The database divides the speakers into 4 different age groups.) Data from the first two recording sessions were used for training, and data from the last three sessions were used for testing. This gave a total of 310 training utterances and 465 test utterances per word.

3 Speaker Models and Scoring

The feature set used for recognition consisted of cepstral coefficients derived from linear prediction analysis (Picone, 1993). Cepstral coefficients are well recognised as one of the best speech parametrisations for both speaker and speech recognition (Furui, 1997). Using an analysis frame of 20 ms, Hamming windowed and overlapping by 50%, 12th order linear predictor coefficients were obtained. These were then used to create cepstral coefficients via the recursion described by Atal (1974).

Speaker models were then produced by vector quantisation. VQ has been used extensively for both text-dependent and text-independent speaker recognition (Rosenberg and Soong, 1987; Booth *et al.*, 1993; Yu *et al.*, 1995). It is a data-reduction technique (Picone, 1993) in which similar vectors are grouped together and represented by their centroid. The grouping is repeated iteratively until the distance (see below) between each vector and its group centroid has been minimised. These centroids make up the codebook which models the

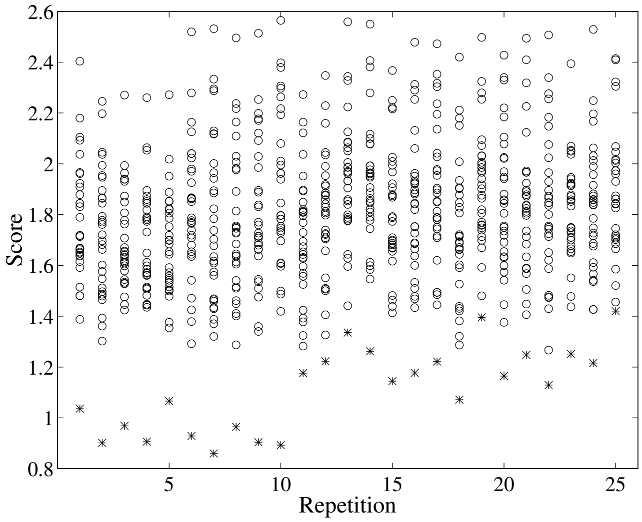

Fig. 1. Typical distributions of genuine and impostor speaker scores for 25 repetitions of a word ('o' denotes impostor scores and '$*$' denotes genuine-speaker scores).

data, and their number will determine the accuracy of the modelling. The standard LBG algorithm (Linde *et al.*, 1980) was applied to calculate the centres.

Distances between vectors j and k (codebook centres or test utterance frames) were calculated using the 'city-block' measure as:

$$d_{jk} = \sum_{i=1}^{12} |c_{ij} - c_{ik}|$$

where c_{ij} is the ith cepstral coefficient of vector j, similarly c_{ik} for vector k. (As usual, c_0 was discarded.) When testing an utterance against a VQ speaker model, the minimum absolute distance between each utterance vector and the codebook for each analysis frame was summed over the total number of frames of the utterance. This sum was then averaged over the number of frames to give the final score.

4 Impostor Cohort Normalisation Method

Impostor cohort normalisation (ICN) was originally introduced by Li and Porter (1988). It has subsequently been used in several speaker-recognition systems (Higgins *et al.*, 1991; Rosenberg *et al.*, 1992; Booth *et al.*, 1993; de Veth *et al.*, 1993). However, *a priori* selection of impostors for the normalisation cohort has always been problematic.

In unnormalised speaker verification, a test utterance is presented to the claimed-speaker model only. The decision to accept or reject the claimed identity is then made by comparing the resulting score with a threshold. To a certain extent, however, verification using ICN is similar to *identification* in that the test utterance is presented not only to the genuine-speaker model but also to a subset of impostor models. The test scores against the impostor models are then used to normalise the genuine-speaker model score. Only after normalisation is a threshold used to accept or reject the claimed identity. Normalising all the scores means that it becomes possible to set a single threshold for all speakers. In practice, however, the results are rarely as good as setting speaker-specific thresholds.

ICN is based on the observed fact that a genuine-speaker's test utterances score lower against their own model than the impostor models, although the absolute values of the score may vary considerably depending on the test utterance in question (Li and Porter, 1988). A typical example of this is shown in Figure 1, which depicts the results of scoring a single speaker's 25 repetitions of one word against that speaker's own model and against all the impostor models for that word. Scores against impostor models are represented by 'o' while those against the genuine-speaker model are represented by '$*$'. The first 10 utterances were those used to create the (genuine) speaker model and, hence, they are expected to give generally low genuine-speaker scores (as seen in the figure). The genuine-speaker scores for the remaining, unseen utterances are rather higher. However, the genuine-speaker score for a particular unseen utterance is invariably lower than its *corresponding* scores against the impostor models, even though the test utterances came from later recording sessions than the training utterances. This fact – that the genuine-speaker scores remain low relative to the impostor scores when both are tested against the same utterance – may therefore be used to normalise the genuine-speaker score.

To normalise a score, an utterance is presented to the genuine-speaker model and a subset of impostor models (rather than all possible models as in Figure 1). The genuine-speaker score is then normalised by the mean and standard deviation of the scores of the impostor cohort as follows:

$$S_{ICN} = \frac{S_{gen} - \mu_{coh}}{\sigma_{coh}}$$

where S_{ICN} is the normalised score, S_{gen} is the original genuine-speaker model score, and μ_{coh} and σ_{coh} are the mean and the standard deviation of the impostor cohort scores respectively.

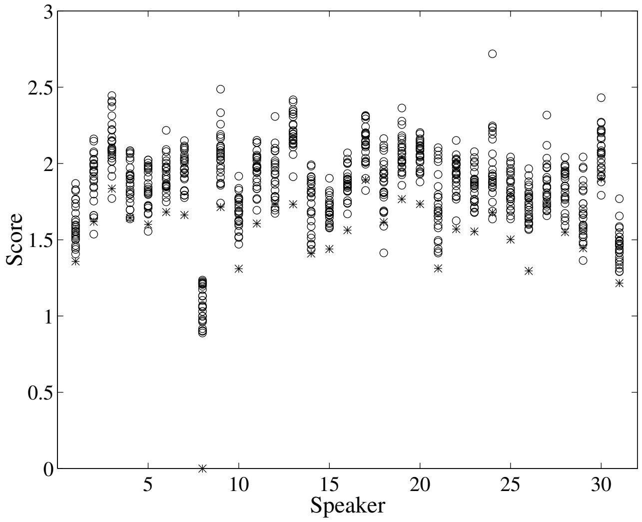

Fig. 2. *Model vs. model* scores and *test vs. model* scores for Speaker 8.

5 *Model vs. Model* Scoring

The most obvious way of determining impostor rankings for a speaker is to test all the impostor training utterances against that speaker's model. As mentioned before, this can be time-consuming if there is a large number of training utterances or impostors. It would be quicker to perform only one test per impostor: comparing speaker models with each other allows us to do this. In effect, each impostor speaker model is treated as a generalised test utterance for that impostor. Testing with the model generates a *single* score for each impostor which may be used to determine the impostor's ranking against the genuine speaker.

An example of this is shown in Figure 2. The *model vs. model* scores are represented by '$*$' and the *test vs. model* scores by 'o'. It is important to remember that the test utterances come from the final three recording sessions while the models were created using data from the first two sessions. Despite this inter-session variability, which can be so detrimental to speaker verification, the *model vs. model* scores would appear to give a very good indication of the final impostor rankings – being generally positioned at or below the low end of the *test vs. model* distributions. (As Speaker 8 is actually the genuine speaker, the *model vs. model* score is zero in this case.)

To determine how good the *model vs. model* rankings are, we require some *reference* rankings for comparison. Unfortunately, there is no definitive way of ranking impostors. Even using final, *a posteriori* test scores, the rankings could be arrived at on the basis of the best impostor score, the mean impostor score, the number of impostor scores below the threshold, etc. Clearly, rankings based on different criteria are unlikely to agree completely with each other, so that small differences should be expected when comparing any two rankings. In these experiments, all reference rankings were determined on the basis of the mean test utterance score per speaker.

The average difference between the *model vs. model* and reference rankings was calculated as follows. First, the absolute difference between an impostor's position, r_m, in the *model vs. model* ranking and that person's position, r_r, in the reference rankings is calculated as $|r_m - r_r|$. This is done for each impostor in the rankings and the average calculated. Once this difference has been evaluated for each speaker, it is then averaged across all speakers. The smaller this value, the closer is the agreement between the ranking method and the *model vs. model* scoring strategy under investigation. Using simple combinatorics, it is easily shown that – since there are only 30 possible ranking positions – the upper bound on the possible average difference is 15 places. When 100 completely random rankings (uniformly distributed) were tested against each other, the average distance was 9.67 places.

The average ranking differences for the 10 test words are shown in Figure 3. The *model vs. model* rankings are compared to three reference rankings, generated by taking the mean score of all possible utterances (both test and training), of the training utterances alone and of the test utterances alone. The figure shows that the *model vs. model* rankings are closest to the training utterance rankings. This is only to be expected, as they were used to create the models. Also, as expected, differences based on the test utterances alone were furthest from the *model vs. model* scores, while the combination of both test and training utterances gave differences somewhere in between. We emphasise again that none of these reference rankings is in any way definitive, as the mean of the test scores may not be an exact measure of an impostor's likelihood of impersonating a speaker. However, the average difference in ranking of about 3.5 per word is significantly lower than the value of 9.67 generated by random rankings. So, although by no means perfect, the rankings were judged to be sufficiently good to make it worthwhile creating and evaluating impostor cohorts.

The rankings also showed, in the VQ case at least, that just because Speaker A was a good impersonator of Speaker B did not mean that Speaker B would be a good impersonator of Speaker A. That is, if Speaker A was ranked at position i against Speaker B's model, and Speaker B was ranked at position j in Speaker A's rankings, then $i \neq j$ in general. This kind of asymmetric relation is common in pattern recognition, but is nonetheless worth remarking on. Figure 4 shows the difference $|i - j|$ for a particular impostor position averaged across speaker pairs. The value is quite different from zero – indeed, it may be anything from 5 to 9 places. The implication for ICN is that although Speaker A may be similar to Speaker B, this does not mean that an impostor cohort for Speaker A would be suitable for Speaker B.

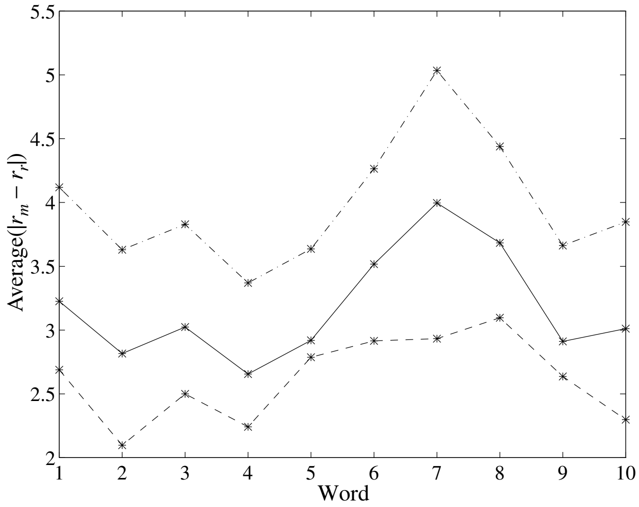

Fig. 3. Average difference in *model vs. model* and reference rankings for each word. There are 3 different reference rankings: '—' denotes rankings based on all possible utterances, '– –' denotes those based on the training utterances alone and '– · –' denotes rankings based on the test utterances alone. See text for specification of $|r_m - r_r|$.

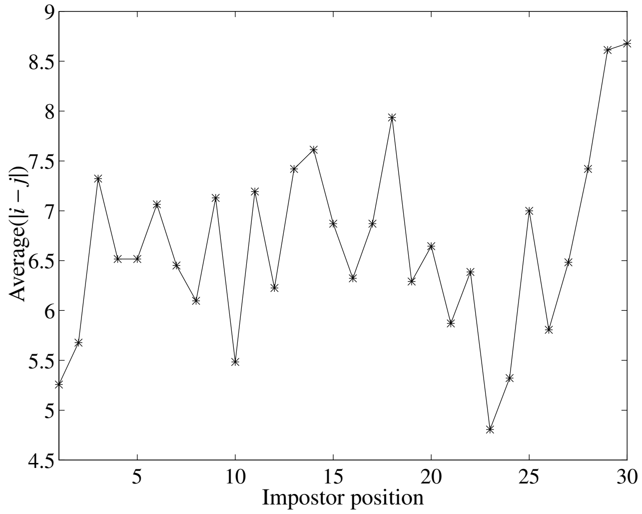

Fig. 4. The average ranking asymmetry when comparing two speakers; i.e. comparing Speaker A as an impostor of Speaker B with Speaker B as an impostor of Speaker A. See text for specification of $|i - j|$.

We have also found that impostors vary in their ability to impersonate the genuine speaker depending on the word in question. This is shown by the fact that, for a particular speaker, the ranking of the impostors varies from word to word. To measure this, the absolute difference between rankings according to a particular word, r_w, and every other test word, r_t, was calculated as $|r_w - r_t|$. This was then averaged across the remaining 9 test words. The results are illustrated in Figure 5, where the rankings are seen to vary from 7 to 8.6 places – indicating that impostor position depends somewhat upon

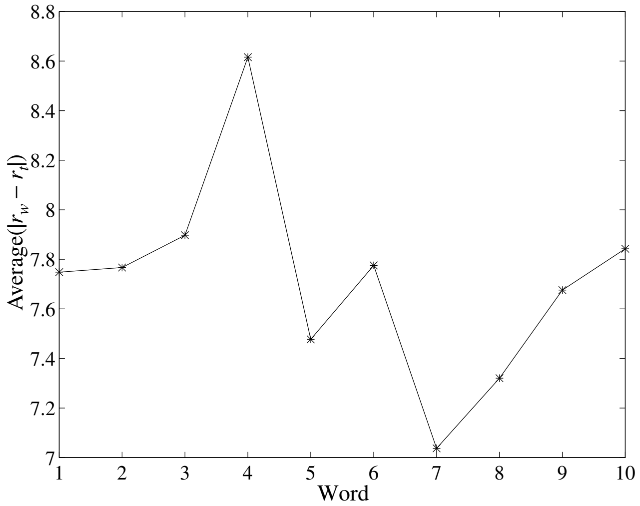

Fig. 5. Average difference between ranking position on the basis of a specific word, r_w, and ranking position on the basis of every other test word, r_t.

specific test words. Thus, testing with more than one word should be beneficial. The likelihood of accepting an impostor will be reduced in this case, as any impostor is unlikely to be consistent across all test words in impersonating the genuine speaker.

Preliminary experiments have indicated that *model vs. model* impostor rankings are also useful in the text-independent case (Finan *et al.*, 1997). Using all 38 speakers of the first dialect region from the TIMIT database, training was done using the dialect (SA) and phonetically compact (SX) sentences and testing with the phonetically diverse (SI) sentences. The average $|r_m - r_r|$ value, calculated using only test utterances, was around 4 places. (Random rankings had an average difference of around 12 places.) This is very similar to the average ranking difference of just over 4 for the test utterances in the text-dependent case. Because the text-independent data were recorded in a single session, one might expected the average ranking difference would be smaller than for the text-dependent data, as the latter are subject to inter-session variability. However, the text-independent models are much more general than the text-dependent models, where only a limited number of phonemes and contexts are being considered. These two competing factors appear to balance one another.

6 Selection of the Impostor Cohort

As it is often impractical to use all possible impostors in the normalisation (de Veth *et al.*, 1993), some means is needed for selecting a cohort of impostors which gives the best improvement in speaker verification rate. Three different cohort selection schemes, based on *model vs. model* scoring, were studied. For the first, the closest impostors were chosen for each speaker. For the second, impostors were evenly distributed throughout the rankings. Finally, impostors were selected randomly. In each case, the size of the impostor cohort was varied from 5 to 30 speakers (the whole impostor population).

Results for the three types of cohort are shown in Figure 6. Here, d' (adapted from the sensitivity measure of signal detection theory (Green and Swets, 1966)) is a measure of separation between the genuine-speaker and impostor score distributions. It is calculated as:

$$d' = \frac{\mu_{imp} - \mu_{gen}}{\sqrt{\sigma_{imp}\sigma_{gen}}}$$

where μ_{imp} and μ_{gen} are the means and σ_{imp} and σ_{gen} are the standard deviations of the impostor and genuine-speaker distributions respectively.

The larger d', the greater the separation. According to Hannah *et al.* (1993a), a d' of around 4 represents good separation between genuine-speaker and impostor score distributions. (In fact, for their small set of 10 speakers, a d' of about 4 corresponded to total separation and, hence, to zero verification error rates.) It is clear that the closest impostors give the best results, both in terms of consistency and actual d'. The d' rises steadily as the size of the impostor cohort is increased, staying generally higher than for either of the other selection methods. The randomly-selected cohort often gives a better d' than does the evenly-distributed one, but its d' fails to increase smoothly, and is generally well below that for the closest-impostors cohort. Hence, the latter is used in the remaining work.

7 Results using the Closest-Impostors Cohort

Speaker verification error rates are often measured in terms of the equal error rate (EER). This occurs when the percentages of false rejections (the genuine speaker being rejected as an impostor) and false acceptances (an impostor being accepted as the genuine speaker) are equal. To find the EER, a threshold is set and then varied until the two error rates are equal. However, quantisation due to the small number of genuine-speaker test utterances often means that an exact *equal* error rate is not possible. The average EER is the mean of the speaker-specific EERs.

Table 1 compares the d' and average equal error rates for the unnormalised and normalised results. An impostor cohort of 15 speakers was used for these

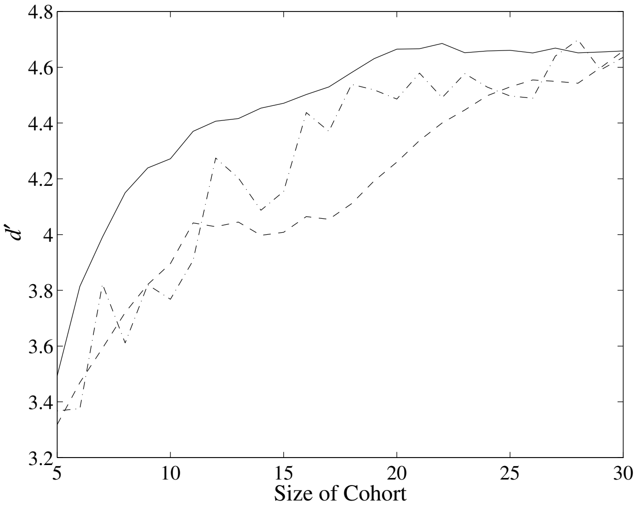

Fig. 6. Average d' separation for the cohorts: '—' denotes closest impostors, '– –' denotes evenly distributed rankings and '– · –' denotes randomly-selected impostors).

	d'	F.R.(%)	F.A.(%)	Total(%)
Unnormalised	3.84	3.23	4.03	4.00
Normalised	4.47	1.16	1.84	1.82

Table 1
Average d' and average equal error rates using an individual threshold for each speaker (F.R. and F.A. denote false rejections and false acceptances respectively).

tests and the average EER calculated across all speakers and all words. The benefit of the normalisation is clearly evident, as the normalised scores have an average EER of 1.82%, which is less than half that of the unnormalised scores (4.00%). This is also reflected in the average d' which shows a marked increase from 3.84 to 4.47.

8 Conclusions

We have shown that the *model vs. model* scores may be used to create impostor rankings. These rankings may in turn be used to select impostors for impostor cohort normalisation. The results indicated that the best impostors for a cohort were those closest to the genuine speaker. It may also be possible to use this approach to determine the most likely impostors in HMM and Gaussian mixture model systems as the centre vectors of the systems could be used as codebooks and compared with each other.

The average *model vs. model* score can also be used as a means of determining if a speaker model is higher or lower scoring than the average speaker model. This is most often the case when speaker models have been weighted to accentuate speaker differences (Hannah *et al.*, 1993b). The average *model vs. model* score may then be used to generate an offset to bring model scores back into line with the average scores and enable better identification (Finan *et al.*, 1997).

Thus, the described technique is simple and effective, reducing equal error rates for text-dependent verification based on VQ by more than a half relative to the unnormalised case. It also has potential for application to other speaker-recognition tasks: such as text-independent verification, and techniques based on statistical modelling or feature-weighting schemes.

References

Atal, B.S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America* **55**, 1304–1312.

Booth, I., Barlow, M. and Watson, B. (1993). Enhancements to DTW and VQ decision algorithms for speaker recognition. *Speech Communication* **13**, 427–433.

de Veth, J., Gallopyn, G. and Boulard, H. (1993). Limited parameter hidden Markov models for connected digit speaker verification over telephone channels. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'93*, Minneapolis, MN, vol. 2, pp. 247-250.

Finan, R.A., Sapeluk, A.T. and Damper, R.I. (1997). VQ score normalisation for text-dependent and text-independent speaker recognition. *Proceedings of 1st International Conference on Audio- and Video-Based Biometric Person Authentication*, Crans-Montana, Switzerland, pp. 211–218.

Furui, S. (1974). An analysis of long-term variation of feature parameters of speech and its application to talker recognition. *Electronic Communications* **57-A**, 34–42.

Furui, S. (1997). Recent advances in speaker recognition. *Proceedings of 1st International Conference on Audio- and Video-Based Biometric Person Authentication*, Crans-Montana, Switzerland, pp. 237–252.

Green, D.M. and Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Hannah, M.I., Sapeluk, A.T. and Damper, R.I. (1993a). The effect of utterance content and length on speaker-verifier performance. *Proceedings of Eurospeech '93*, Berlin, Germany, vol. 3, pp. 2299–2302.

Hannah, M.I., Sapeluk, A.T., Damper, R.I. and Roger, I.M. (1993b). Using genetic algorithms to improve speaker-verifier performance. *Proceedings of IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, Chelmsford, UK, pp 24/1–24/9.

Hattori, H. (1994). Text-independent speaker verification using neural networks. *Proceedings of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, pp. 103–106.

Higgins, A., Bahler, L. and Porter, J. (1991). Speaker verification using randomized prompting. *Digital Signal Processing* **1**, 89–106.

Li, K-P. and Porter, J.E. (1988). Normalizations and selection of speech segments for speaker recognition scoring. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '88*, New York, NY, vol. 1, pp. 595–598.

Linde, J., Buzo, A. and Gray, R.M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications* **28**, 84–95.

Matsui, T. and Furui, S. (1995). Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Communication* **17**, 109–116.

Picone, J. (1990). Signal modeling techniques in speech recognition. *Proceedings of the IEEE* **81**, 1215–1247.

Rosenberg, A.E., DeLong, J., Lee, C-H., Juang, B-H. and Soong, F.K. (1992). The use of cohort normalized scores for speaker verification. *Proceedings of International Conference on Spoken Language Processing, ICSLP '92*, Banff, Canada, pp. 599–602.

Rosenberg, A.E. and Soong, F.K. (1987). Evaluation of a vector quantization talker recognition system in text dependent and text independent modes. *Computer Speech and Language* **22**, 143–157.

Yu, K., Mason, J. and Oglesby, J. (1995). Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *IEE Proceedings: Vision, Image and Signal Processing* **142**, 313–318.