Proceedings of the Institute of Acoustics

SPEAKER VERIFICATION IN TELEPHONY

A. M. Ariyaeeinia, P. Sivakumaran and B. Jefferies

DSP Applications Research Group,
Department of Electrical and Electronic Engineering,
University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB

1. INTRODUCTION

Automatic speaker recognition has been the subject of extensive research over the past two decades [1-4]. An outcome of this has been the development of systems that are highly reliable when used with relatively clean speech signals. Many practical applications of such systems, however, necessitate the transmission of speech over telephone channels [5,6]. As a consequence, the speech signals are subjected to degradation by various sources. Amongst these are the transmission channel filter, channel noise, and also environmental noise such as crowd babble or other forms of background sounds.

The main effect of the channel filter is the spectral distortion due to a non-flat frequency response in the pass-band. An additional problem is that channel filter characteristics change randomly from one call to another. This can lead to a significant mismatch between the training and test utterances from the same speaker. Such mismatches, and the consequent increase in the verification error, may also result due to the undesired noise events stated above or because of involuntary speaker generation variations. Since automatic speaker verification is being considered for such important applications as telephone banking and access to confidential information on databases, it is believed that an employable system must be able to exhibit an acceptable level of robustness under varying operational conditions.

This paper presents an investigation into methods for improving the performance of speaker verification under adverse conditions. A number of approaches considered for this purpose are compared, and details of the experimental work and results are presented.

2. LINEAR SPECTRAL DISTORTION

A significant problem in the automatic speaker verification operation in practical applications, as stated earlier, is due to the variability of the characteristics of the communication channel filter. It has been found that the short-term spectral features (e.g. cepstral, mel cepstral, perceptual linear predictive (PLP)-cepstral) which perform well when speech signals are recorded under clean and stationary conditions, are vulnerable to such variations [7-9]. To tackle this problem a number of techniques have been developed which aim to introduce robustness into spectral features [9-11]. Most of these are based on the post-processing of the

cepstral vectors with the aim of rejecting channel-related information and perhaps other causes of data mismatch. An effective method in this category is the cepstral mean normalisation [1,6,12]. Since the channel filter magnitude response is additive in the log spectra domain, its effects can be minimised by removing the global mean of the log spectra. The approach, which is based on the assumption that the channel is both linear and time invariant, involves the removal of the bias content of the cepstral feature vectors. This is carried out by computing the average cepstral feature vector across the given utterance, and then subtracting this from individual feature vectors before using them for the generation of a reference model or conducting a verification test [6]. The operation can be described mathematically by expressing the effects the channel filter on the cepstral vectors as:-

$$\tilde{\mathbf{c}}_i = \mathbf{c}_i + \mathbf{h}, \qquad 1 \le i \le M \tag{1}$$

where $\tilde{\mathbf{c}}_i$ are the observed cepstral vectors, $\mathbf{c}_i$ are the input speech cepstral vectors, $\mathbf{h}$ is the channel filter cepstral vector, and M is the number of feature vectors within the given utterance.

The compensated cepstral vectors are obtained by:-

$$\hat{\mathbf{c}}_i = \tilde{\mathbf{c}}_i - \tilde{\mathbf{c}}_m, \qquad 1 \le i \le M \tag{2}$$

where $\tilde{\mathbf{c}}_m$ is the global mean vector, i.e.

$$\tilde{\mathbf{c}}_m = \frac{1}{M}\sum_{i=1}^{M} \tilde{\mathbf{c}}_i \tag{3}$$

Inevitably, this method also removes the average speech spectrum which contains speaker specific information. The adverse effect of this on the verification performance might be particularly significant when the utterances are too short [13]. It has, on the other hand, been demonstrated that [14,15] the global mean of the speech feature vectors exhibit significant intra-speaker variability over time. This may unfavourably affect the verification accuracy. Furthermore, the mean speech spectrum is influenced by variations due to speech efforts and health [12]. It is therefore concluded that the use of the above method with clean speech results in the minimisation of the inter-session variability, and for telephone quality speech it also leads to the removal of the spectral shaping imposed by different communication channels.

The minimisation of the transmission channel effects may also be achieved through the use of delta cepstral parameters [15,16]. The original motivation for the use of delta cepstral vectors was to capture the transitional spectral features of speech. These vectors are obtained as the weighted combination of the differences between K pairs of the observed cepstral vectors which are $2k$ frames apart, where $k = 1,2,...,K$. i.e.

$$\Delta \mathbf{c}_i = \zeta_K \sum_{k=1}^{K} k(\tilde{\mathbf{c}}_{i+k} - \tilde{\mathbf{c}}_{i-k}), \qquad 1 \le i \le M \tag{4}$$

where $\Delta \mathbf{c}_i$ is the i^{th} delta cepstral vector and ζ_K is a coefficient whose value depends only on K.

If the channel filter is assumed to be linear and time-invariant, then based on equation (1) it becomes evident that delta cepstral vectors are channel invariant, i.e.

$$\Delta \mathbf{c}_i = \zeta_K \sum_{k=1}^{K} k(\mathbf{c}_{i+k} - \mathbf{c}_{i-k}), \qquad 1 \le i \le M \tag{5}$$

It has, however, been demonstrated that in most cases combining delta cepstral and mean-normalised cepstral features does not lead to a significant improvement over using mean normalised cepstral features alone [12].

3. SPEECH VARIABILITY DUE TO ANOMALOUS EVENTS

Another very important factor to consider in automatic speaker verification is that of undesired variations in speech characteristics due to anomalous events. These anomalies can range from channel and environmental noise to uncharacteristic speech sounds from the speakers. The resultant mismatch between the corresponding test and training patterns is known to reduce the verification accuracy significantly [17-19].

Due to the absence of accurate information about the existence, level, and nature of speech degradation in practical applications, it has been proposed to introduce robustness into the verification operation by normalising the verification scores appropriately [18-21]. The approach is based on the concept that if anomalous events in the test utterance cause a speaker's score against his (her) own model to degrade, then the scores obtained for the same speaker against certain other models in the set are also affected in the same way. As a result, the ratio of the score for the target model to a statistic of scores for other considered models remains relatively unchanged. The use of this ratio instead of the absolute score for the target speaker should therefore be expected to improve the verification performance considerably.

An effective score normalisation method for the above purpose is that based on comparing the given test utterance against the model of the proposed speaker, as well as, against the models of a cohort of other speakers assigned to that particular speaker [19]. The assigned speakers are those whose models are most similar to that of the proposed speaker, and are selected a priori. In this technique, the ratio of the score for the proposed speaker to the mean cohort score can be compared against a pre-set threshold for the verification purpose.

A somewhat different approach in this category [21] consists of normalising the score for the proposed speaker by the average of the top N scores for all the registered speakers in the set. As a consequence, a dynamic normalisation factor is provided by allowing the selection of the competitive speakers in each test to depend on their relative scores on that particular occasion.

Although, the above two methods have been examined in independent studies, their relative effectiveness has not previously been investigated. It is therefore proposed to compare the performance of the two approaches under identical experimental conditions. Details of this comparative study is presented in section 5. In the remainder of this paper these normalisation methods are referred to as cohort and unconstrained cohort respectively.

4. SPEECH DATA AND FEATURE EXTRACTION

The speech data adopted for this study was a subset of the BT Brent database [6]. The subset consisted of 47 repetitions of digit utterances one to nine and zero spoken by 11 male and 9 female speakers. It was collected from telephone calls made by speakers from various locations. For each speaker, the first 3 utterance repetitions, which were recorded in a single call, formed the training set. The remaining 44 repetitions (one recorded per week) were reserved for testing.

For the purpose of the experimental study, utterances were pre-emphasised using a first-order digital filter. Each utterance was then segmented into 25 ms frames at intervals of 12.5 ms using a Hamming window, and subjected to a 12^{th}-order linear prediction analysis. The resultant linear predictive coding (LPC) coefficients for each speech frame were appropriately analysed using a 10^{th}-order fast Fourier transform, a filter bank, and a discrete cosine transform to extract a 12^{th}-order mel-frequency cepstral feature vector [22,23].

In order to reduce the effects of linear spectral shaping imposed by the communication channel, cepstral feature vectors were subjected to the mean normalisation technique described in section 2.

5. EXPERIMENTAL WORK AND RESULTS

This section provides the details and results of the experiments conducted to examine the effectiveness of the considered normalisation methods for speaker verification in telephony. The study was carried out using a text-dependent Hidden Markov Model (HMM) speaker verification system. Speakers were modelled by a set of four-state left to right HMM's representing individual digit utterances. The observation probability for each state was a

continuous density function described by a mixture of two Gaussian densities. The covariance matrix of the probability distribution was assumed to be diagonal, and the initial model parameters were estimated using a modified K-means algorithm [24].

The experimental work consisted of speaker verification trials using individual digit utterances. The main aims of these tests were to compare the performance of the cohort and unconstrained cohort normalisation techniques, and also to examine the dependence of their effectiveness on the size of the set of competitive speakers.

For the purpose of cohort normalisation the selection and assignment of competitive speakers was carried out in the following manner. The verification system was trained for individual speakers, and then, for each digit utterance, the training repetitions from each speaker were compared against the corresponding models of all other speakers in the set. The speakers who were most competitive* with each registered individual were selected using the pair-wise comparison method [19]. The process was repeated ten times to cover all different ten digit utterances. As a result, each registered speaker was assigned a set of text-dependent speaker cohorts. For unconstrained cohort normalisation, on the other hand, competitive speakers were determined during the actual test trials.

The baseline verification scores were expressed as log likelihoods, i.e.

$$l_i = \log L_i \tag{6}$$

where L_i denotes the likelihood for the i^{th} speaker. The normalised scores were obtained by:-

$$\tilde{l}_i = l_i - \frac{1}{N}\sum_{j=1}^{N} \log L'_j \tag{7}$$

where N is the size of the adopted set of competitive speakers and L'_j are the likelihoods for these.

The first set of verification experiments was conducted by incrementing N from a minimum of 1 to a maximum of 19. This was due to the fact that there were only 20 speakers in the set. Figure 1 illustrates the results of this study in terms of the average equal error rate (EER) as a function of N, in each considered case. The EER obtained using unnormalised verification scores is also given in this figure as the baseline. These results clearly confirm that the verification accuracy can be significantly improved by normalising the test scores appropriately. It is observed that the EER for the cohort method decreases almost exponentially by increasing N from 1 to 15. As the cohort size is increased beyond 15, the verification error starts to increase gradually. Figure 1 also shows that only for values of N of up to 2 this error rate is larger than that obtained using unnormalised verification scores.

A very important part of the experimental results is the considerable difference between the performance of the two normalisation techniques in favour of the unconstrained cohort approach, particularly for small values of *N*. For very large values of *N*, the two methods are almost identical and lead to very similar results. The superior performance of the unconstrained cohort method is believed to be due to the way the competitive speakers are selected in this method. Although it can be argued that in the case of true speakers the two methods are almost equally effective. When the test utterance is spoken by an impostor the unconstrained cohort method tends to exhibits a better performance. This is because in this method, unlike in the cohort approach, the competitive talkers associated with registered speakers are not pre-selected. As a result, when a speaker from within the set (as has been the case in the above experiments) claims the identity of another speaker, he (she) is very likely to score higher against his (her) own model, and also perhaps against some other models in the set, than against the target model. The small EERs obtained with this method for very small values of *N*, and in particular for a value of *N* of 1, confirms the above argument.

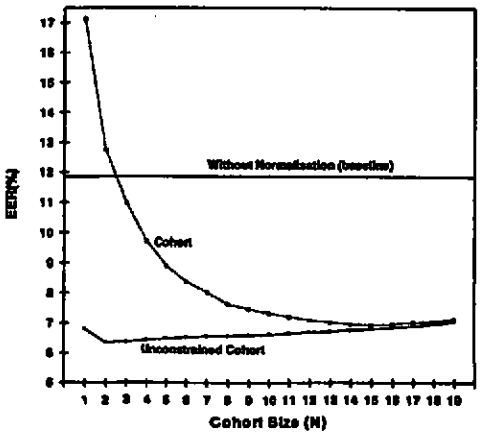

Figure 1. Verification performance of cohort and unconstrained cohort normalisation methods based on drawing impostors from within the set of registered speakers.

In practical applications, the impostors are more likely to be from the outside the set of registered speaker than the inside. In such cases, the unconstrained cohort method would again be expected to produce a lower false acceptance error than the cohort method. This is because, as stated before, the method allows the speakers in the set whose models are most close to the impostor's utterance to compete with the target model. It is thought that if the speaker set is

adequately large, there is always a high probability that an impostor targeting a particular speaker, will score higher against one or more models in the set other than the target one. In the cohort method, on the other hand, the competitive speakers are those whose models are closest to the target speaker model, and not necessarily to the impostor's utterance. As a consequence, the impostor's score against these competitive speakers may not necessarily be higher than that against the target speaker.

In order to investigate this case further, experiments were conducted by dividing the speaker set into two equally populated subsets. In the first set of experiments the verification system was trained for speakers in subset 1, and the impostors were drawn from subset 2. A second set of experiments was then carried out by interchanging the roles of speakers in the two subsets. Results of these experiments are given in Figure 2.

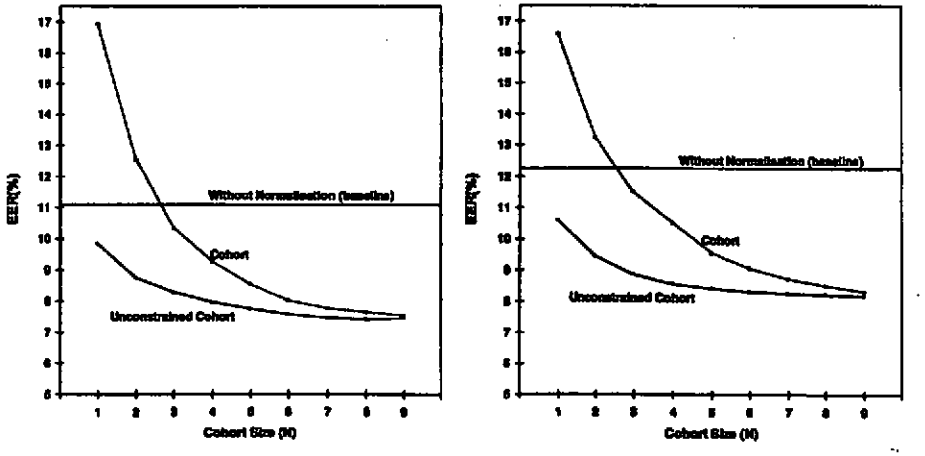

Figure 2. Equal error rates in speaker verification experiments using impostors from outside the group of registered speakers.

(a) *Subset 1*: registered speakers, *Subset 2:* impostors.
(b) *Subset 1*: impostors, *Subset 2*: registered speakers.

It is observed that in both cases the EERs obtained for normalised verification scores decrease almost exponentially with an increase in *N*. The results also show that the EER for the unconstrained cohort method is consistently lower than that for the cohort method. This, as suggested earlier, is due to the lower false acceptance error in the former method. It is,

however, seen that since in the present cases the impostors are drawn from the outside the set of registered speakers, the EERs obtained with the unconstrained cohort method for a cohort size of 1 are not as small as before.

6. CONCLUSIONS

An investigation into the performance of automatic speaker verification in telephony was presented. The study included a coverage of the effects, on the verification performance, of the telephone channel and also methods for reducing these. For the purpose of minimising the effects of mismatch between the corresponding test and training utterances, two score normalisation methods, i.e. cohort and unconstrained cohort, were considered and compared.

Verification trials were conducted using impostors from the within as well as without the set of registered speakers. The results of the investigation showed that the verification performance could be significantly improved by normalising the verification scores appropriately. The results also showed the performance of the unconstrained cohort normalisation method to be consistently superior to that of the cohort method. This is believed to be due to the fact that the cohort method attempts to improve the verification accuracy only by reducing the false rejection error. The unconstrained cohort method, on the other hand, reduces both false rejection and false acceptance errors. As a result a lower equal error rate can be achieved with this method.

7. ACKNOWLEDGEMENT

The authors wish to express their thanks to Mr Mark Pawlewski and Mr Simon Downey of BT Laboratories for their support, stimulating discussions, and the provision of the Brent database.

8. REFERENCES

[1] B. S. Atal, "Automatic Recognition of Speakers from Their Voices", *Proc. IEEE*, vol. 64, No. 4, pp. 460-475, April 1976.

[2] A. E. Rosenberg, "Automatic Speaker Verification : A Review", *Proc. IEEE*, vol. 64, No. 4, pp. 475-487, April 1976.

[3] G. R. Doddington, "Speaker Recognition-Identifying People by Their Voices", *Proc. IEEE*, vol. 73, pp. 1651-1664, 1985.

[4] J. M. Naik, "Speaker Verification: A Tutorial", *IEEE Commun. Mag.*, pp. 42-48, Jan. 1990.

[5] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker Verification Over Long Distance Telephone Lines", *Proc. ICASSP*, pp. 524-527, 1989.

[6] M. Pawlewski, B. P. Milner, S. A. Hovell, D. G. Ollason, S. P. A. Ringland, K. J. Power, S. N. Downey and J. Bridges, "Advances in Telephony Based Speech Recognition", *BT Tech. J.*, pp. 127-150, Jan. 1996.

[7] J. Hernando, C. Nadeu, C. Villagrasa and E. Monte, "Speaker Identification in Noisy Conditions Using Linear prediction of the One-Sided Autocorrelation Sequence", *ICSLP 94*, pp. 1847-1850, 1994.

[8] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A Comparative Study of Robust Linear Predictive Analysis Method with Applications to Speaker Identification", *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 117-125, Mar. 1995.

[9] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, "RASTA-PLP Speech Analysis Technique", *Proc. ICASSP*, vol. I, pp. 121-124, 1992.

[10] Y. H. Kao, J. S. Baras and P. K. Rajasekaran, "Robustness Study of Free-Text Speaker Identification and Verification", *Proc. ICASSP*, vol. II, pp. 379-382, 1993.

[11] K. T. Assaleh and R. J. Mammone, "Robust Cepstral Feature for Speaker Identification", *Proc. ICASSP*, vol. I, pp. 129-132, 1994.

[12] D. A. Reynolds, "Robust Text-Independent Speaker Identification using Gaussian Mixture Models", *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 72-83, Jan. 1995.

[13] S. Furui and M. M. Mohan, "Advances in Speech Signal Processing", *Marcel Dekker, Inc., New York*, Chapter 22, pp. 701-738, 1992.

[14] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on ASSP*, vol. 29, pp. 254-272, April 1981.

[15] C. Bernasconi, "On Instantaneous and Transitional Spectral Information for Text-Dependent Speaker Verification", *Speech Commun.*, vol. 9, pp. 129-139, April 1990.

[16] F. K. Soong, and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *IEEE Trans. on ASSP*, vol. 36, pp. 871-879, June 1988.

[17] H. Gish, M. Schmidt and A. Mielke, "A Robust, Segmental Method for Text Independent Speaker Identification", *Proc. ICASSP*, vol. I, pp. 145-148, 1994.

[18] K-P. Li and J. E. Porter, "Normalisations and Selection of Speech Segments for Speaker Recognition Scoring", *Proc. ICASSP*, pp. 595-598, 1988.

[19] A. E. Rosenberg, J. Delong, C. H. Lee, B. H. Huang, and F. K. Soong, "The Use of Cohort Normalised Scores for Speaker Verification", *Proc. ICSLP*, pp. 599-602, 1992.

[20] M. J. Carey and E. S. Parris, "Speaker Verification Using Connected Words", *Proc. IOA*, vol. 14, pp. 95-100, 1992.

[21] T. Matsui and S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition", *Proc. ICASSP*, pp. 391-394, 1993.

[22] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. on ASSP*, vol. 28, pp. 357-366, Aug. 1980.

[23] J. R. Deller, J. G. Proakis and H. L. Hansen, "Discrete-Time Processing of Speech Signals", *Macmillan Inc. New York*, 1993.

[24] L. R. Rabiner, J. G. Wilpon and B-H. Juang, "A Segmental K-Means Training Procedure for Connected Word Recognition", *AT&T Tech. J.*, vol. 65, pp. 21-31, May/June 1986.