SPEAKER VERIFICATION: BEYOND THE ABSOLUTE THRESHOLD

Xiaoyuan Zhu, Bruce Millar, Iain Macleod and Michael Wagner

TRUST[1] Project
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT - This paper examines the use of "discriminating probabilities" for text-dependent speaker verification. Conventional left-to-right continuous hidden Markov models are used to represent client speaker models, with the whole impostor speech database being considered during training of client models. In the speaker verification stage, the claimed speaker's discriminating probability is calculated based both on their own trained model and on the impostor distribution. The proposed method attempts to avoid using an absolute threshold. No additional training data for each client is required for this method and there is only a minor increase in computational overheads. One promising application of discriminating probabilities is in conjunction with cohort-normalised speaker verification, to help reduce the problem of false acceptances of dissimilar impostors associated with this latter approach.

INTRODUCTION

Although speaker identification and verification may both be based on similar pattern recognition techniques, the simple method of choosing the best fit (as used in the decision making stage while identifying one of a closed set of speakers) is not applicable for speaker verification. The problem of speaker verification can be formulated as one of testing a binary hypothesis, in which input speech is attributed either to the claimed speaker or to an (unspecified) impostor. The method of determining an appropriate threshold distance for this accept/reject decision is of central importance in any speaker verification system. Since the threshold has to be set on an *a priori* basis, it may not be optimum for given test data. The traditional method for assessment of speaker verification performance is measurement of the equal error rate (EER) on test data. Note, however, that the process of sorting scores to determine the threshold in measuring the EER is an *a posteriori* method. As such, this method has only limited use for calculating acceptance thresholds for a speaker verification system.

Rosenberg et al. (1992) proposed a cohort-normalised score to replace absolute acceptance thresholds. In their method, the likelihood score for the speaker whose identity is claimed was compared with the scores of a "cohort" of other speakers assigned to that speaker. A shortcoming of this method is that impostors who are quite dissimilar to the client are sometimes falsely accepted. As a result of earlier research in the TRUST Project, we have suggested a hybrid threshold approach for text-independent speaker verification (Chen, Millar & Wagner, 1994). The first stage of the hybrid threshold approach uses an absolute threshold to eliminate dissimilar impostors prior to a second stage of cohort-normalised verification. The hybrid approach reduces the incidence of false acceptances of dissimilar impostors, but the

threshold used in the first stage has to be large enough to allow for substantial variations (phonetic, intra-speaker and environmental) in the speech of the client.

In this paper, we define a "discriminating probability" (based on impostor distributions with the training data) for use at the decision making stage. The proposed method is theoretically superior to conventional speaker verification algorithms, since it makes use of *a priori* information about both the client and potential impostors in verifying the client. Compared with standard verification strategies, no additional training data for each client is required and there is little additional computation during the verification process.

CONTINUOUS HIDDEN MARKOV MODELS

Left-to-right HMMs have been shown to be very effective and practical for text-dependent speaker recognition (Rosenberg, Lee & Gokcen, 1991), since they are capable of modelling both static and dynamic speaker-dependent information within a sequence of phonemes. Our previous research (Zhu, Macleod & Millar) implemented and tested left-to-right HMMs for speaker identification. In this research, we have used continuous left-to-right HMMs to model clients' speech. Our 6-state HMMs had 4 mixtures assigned to each state, using the topology shown in Figure 1.

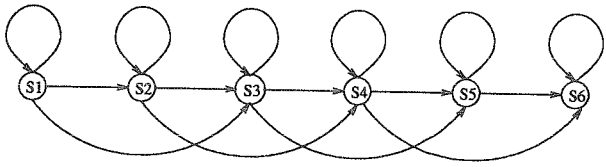

Figure 1: Topology of the 6-state left-to-right HMMs.

The state transition probabilities a_{ij}, which were estimated in the training procedure, were constrained as follows:

$$a_{ij} = 0, \qquad \text{for } j > i+2 \text{ or } j < i. \tag{1}$$

The probability density functions $b_j(x)$ in the HMMs have the form

$$b_j(x) = \sum_{k=1}^{M} c_{jk} \aleph(\mathbf{x}, \mu_{jk}, \mathbf{U}_{jk}), \quad 1 \le j \le N, \tag{2}$$

where $\aleph(\mathbf{x}, \mu_{jk}, \mathbf{U}_{jk})$ represents a multi-dimensional Gaussian distribution with mean vector μ_{jk} and covariance matrix $\mathbf{U}_{jk}$, and c_{jk} is the weight for the kth mixture component of state j. The HMMs were trained by the simplified segmental training procedure (Zhu, 1993).

DISCRIMINATING PROBABILITY

To overcome the lack of explicit discrimination between classes in conventional HMMs, Forsyth, Bagshaw and Jack (1994) introduced discriminating observation probabilities (DOPs) in their HMM-based speaker verification algorithm. The basic idea of their approach is that by incorporating DOPs within an HMM, the resulting DOP HMM depends not only on the client's speech but also on that of impostors, which should thus increase the HMM's ability to make client/impostor distinctions.

In the present research, we propose the use of discriminating probabilities as another alternative (in addition to cohort-normalised scores) to the use of pre-set absolute thresholds in speaker verification systems. Claimed speakers are verified based on their (individually-trained) HMMs and on information relating to impostor likelihood distributions (for each client and utterance). Our proposed discriminating probability approach, which differs from Forsyth, Bagshaw and Jack's in that our impostor distributions are represented separately from the HMMs, can be summarised as follows:

1. Each speaker's HMMs are trained using all their training data for each utterance — only the client's characteristics are represented in each model.
2. Each trained model is subsequently tested against all the same utterances from other speakers, these other speakers being regarded as representative potential impostors. An overall impostor distribution for each client HMM is then created from the normalised distribution of individual impostor likelihood scores. Absolute values of logarithmic likelihood scores u are used as the distance measure in the impostor distribution $i(u)$ calculation.
3. In the test stage, the distance S of the input utterance from the client's model is computed and used to partition the client's impostor distribution $i(u)$. The discriminating probability P is calculated via the following formula:

$$P = \int_0^S i(u)du. \tag{3}$$

The discriminating probability score P indicates the position of the input speech in the space between the centre of a client model and its impostor distribution. Because this "discrimination score" is based on client- and utterance-specific impostor distributions, the proposed method should compensate to some extent for intra-speaker variations, as they interact with the characteristics of other speakers. The intention here is to allow us to sidestep some of the problems arising from unmodelled variations in the true-speaker characteristics (Lund et al., 1994).

EXPERIMENT

The speech data used in our experiments were from the initial TRUST speech data corpus (Millar et al., 1994), which contains corpora A and B. Corpus A consists of 30 short utterances — mainly one or two word commands (on average about one second long). Speech data from 45 speakers were acquired in 3 sessions, with 5 repetitions of each utterance in each session. Corpus B consists of 10 utterances which were recorded only in the second and third sessions, with 2 repetitions in each session. The intervals between the sessions ranged from one to several weeks.

The speech data were digitised to 16 bits/sample at a rate of 20000 samples/s. The digitised signal was then pre-emphasised with a transfer function $1 - 0.95z^{-1}$, and windowed using a 512-point (25.6 ms) Hamming window with a 10 ms frame shift.

The speech features used in the experiments were 15 mel-frequency cepstral coefficients (MFCCs). The MFCC analysis was as described by Davis and Mermelstein (1980). We used 40 110-mel bands spaced 55 mel apart covering a range of 55 to 2310 mels.

For our text-dependent experiment, we used 27 utterances (omitting the 3 longest utterances)

from Corpus A as the test set. The speech of 37 native English speakers (20 male and 17 female) was used to evaluate the algorithm. Each HMM was trained by the 10 repetitions of each utterance in the first and second sessions. The 5 repetitions of each utterance in the third session formed the test data. There were thus 999 HMMs (37 speakers $\times$ 27 utterances) and the same number of impostor distributions. Each impostor distribution was created by 360 likelihood scores (36 impostors $\times$ 10 repetitions). In the test procedure, each HMM was tested against 5 client utterances and 180 impostor utterances.

The impostor distributions vary across both speakers and utterances. The aggregated impostor distribution (speakers/utterances) is shown in Figure 2.

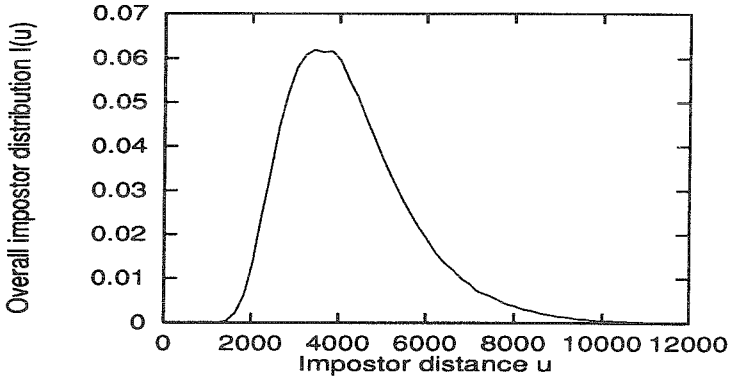

Figure 2: Overall impostor distribution.

When we used the EER to evaluate the discriminating probability approach, we achieved an average EER of 15.72% for testing all 37 clients. For comparison with a conventional HMM algorithm, we used the same data sets to do speaker verification without using impostor distributions. The average EERs of the proposed approach and the conventional HMM are shown in Table 1. Table 1 also gives the standard deviations of EERs across 37 speakers and 27 utterances of each approach.

	Average EER	Std across speakers	Std across utterances
Discriminating algorithm	15.72	15.43	3.37
Conventional HMM	15.74	15.49	3.35

Table 1: Average equal error rates and standard deviations across speakers and utterances for the discriminating algorithm and the conventional HMM.

This level of performance is not significantly better than the results of the conventional HMM. Forsyth and Jack (1994) investigated their DOP HMM algorithm with various speech parameters, and found that the discriminating algorithm gave minor improvement in speaker verification performance using linear cepstral coefficients, delta linear cepstrals and delta mel-frequency cepstrals.

DISCUSSION

Problems which arise from use of predetermined absolute thresholds in deciding whether to accept or reject an input speaker's claimed identity are by now well understood. Unmodelled changes in the input utterances resulting from environmental, phonetic and intra-speaker variations lead to significant overlap between the distributions of similar speakers — a fixed threshold for goodness of fit with the client's model must therefore lead to an undesirably large EER.

Normalisation of measures of "goodness of fit" with respect to a cohort of similar speakers is one method for moving beyond fixed absolute thresholds, based on the observation that the above variations will tend to reduce the match with cohort models as well as with the client's model. The approach explored herein also attempted to move beyond an absolute threshold, but on a different basis.

The reasons for the failure of both Forsyth, Bagshaw and Jack's experiments and our experiments to obtain the theoretically expected benefits warrant further investigation. We have however shown that the proposed approach has a good prospect for moving beyond absolute threshold in speaker verification.

Thus, it appears that the most appropriate use of our calculated discrimination score is not as a primary criterion for speaker verification, but rather as one component of an overall method in combination with other techniques. An obvious application here is to use the discrimination score in place of the fixed absolute threshold that we have suggested for use in the first stage of a hybrid cohort-normalised speaker verification system (Chen, Millar & Wagner, 1994). The fact that the discrimination score is calculated on a very different basis from cohort normalised distances should help minimise problems with false acceptance of dissimilar impostors encountered with use of this technique.

ACKNOWLEDGEMENT

This research has been carried out on behalf of the Harry Triguboff AM Research Syndicate.

NOTES

(1) Technology for Robust User-conscious Secure Transactions.

REFERENCES

Chen, F., Millar, B. & Wagner, M. (1994) *Hybrid threshold approach in text-independent speaker verification*, Proc. Int. Conf. Spoken Language Processing, Yokohama, Japan, 1855–1858.

Davis, S. & Mermelstein, P. (1980) *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans. Acoust., Speech and Signal Processing, Vol.28, No.4, 357–366.

Forsyth, M.E., Bagshaw, P.C. & Jack, M.A. (1994) *Incorporating discriminating observation probabilities (DOP) into semi-continuous HMM for speaker verification*, Proc. Workshop on Automatic Speaker Recognition, Identification & Verification, Martigny, Switzerland, 19–22.

Forsyth, M.E. & Jack, M.A. (1994) *Discriminating semi-continuous HMM for speaker verification*, Proc. Int. Conf. Acoust., Speech and Signal Processing, Adelaide, Australia, I-313–316.

Lund, M.A., Lee, C.T., Lee, C.C. & Bossemeyer, B.W. (1994) *A distributed decision approach to speaker verification*, Proc. Int. Conf. Acoust., Speech and Signal Processing, Adelaide, Australia, I-141–144.

Millar, B., Chen, F., Macleod, I., Ran, S., Tang, H., Wagner, M. & Zhu, X. (1994) *Overview of speaker verification studies towards Technology for Robust User-conscious Secure Transactions*, Proc. Int. Conf. on Speech and Technology, Perth, Australia.

Rosenberg, A.E., DeLong, J., Lee, C.H., Juang, B.H. & Soong, F.K. (1992) *The use of cohort normalized scores for speaker verification*, Proc. Int. Conf. Spoken Language Processing, Banff, Canada, I-599–602.

Rosenberg, A.E., Lee, C.H. & Gokcen, S. (1991) *Connected word talker verification using whole word hidden Markov models*, Proc. Int. Conf. Acoust., Speech and Signal Processing, Toronto, Canada, 381–384.

Zhu, X. (1993) *A combined neural network and hidden Markov model approach for speaker recognition*, Proc. IEEE Region 10 Int. Conf. on Computers, Communications, Control and Power Engineering, Beijing, China, 1074–1077.

Zhu, X., Macleod, I. & Millar, B. (1994) *The effect of training data on the performance of an HMM-based speaker recognition system*, Proc. Int. Conf. on Systems, Control, Information — Methodologies & Applications, Wuhan, China, to appear.