HMM SPEAKER VERIFICATION WITH SPARSE TRAINING DATA ON TELEPHONE QUALITY SPEECH

M.E. Forsyth A.M. Sutherland J.A. Elliott and M.A. Jack
Centre for Speech Technology Research
UNIVERSITY OF EDINBURGH

ABSTRACT—Speaker verification experiments using discrete and semi-continuous HMMs with telephone quality isolated digits are reported. The models were trained with varying numbers of tokens, giving equal error rates of 14% and 16% respectively on single isolated digits, and 4% and 2% on a sequence of 7 isolated digits.

INTRODUCTION

The tasks of speech recognition and speaker verification have much in common. Currently the most widely used technique used in speech recognition is that of hidden Markov modelling (HMM). There are three main forms of HMMs, discrete (DHMM), continuous (CHMM) and semi-continuous (SCHMM) (also known as tied-mixture continuous). All three techniques can be applied to speaker verication.

Although verification and recognition are similar tasks, there is a basic difference. In speaker independent speech recognition the objective is to avoid discriminating between different speakers saying the same word. In verification the objective is to maximise this discrimination. Use of minimal training data per speaker is a perogative.

The first section describes the parameters of the HMM system used, including the performance measure used to evaluate the system. The second section contains the results obtained using a conventional DHMM for speaker verification using 5 and 10 tokens to train isolated word models. The third section shows the comparative performance of a SCHMM system using 10 training tokens.

THE SYSTEM

Database

A database of isolated digits recorded over the U.K. telephone network was used for all experiments. There are 12 digits, consisting of the digits 1 to 9 plus 'zero', 'nought' and 'oh'. The data was sampled at 8 kHz and was divided into blocks of 5 utterances of each digit. Three such blocks (A, B, C) were used in these experiments. Each block comes from a different recording session. The data were automatically endpoint-detected to remove excess silence, and to reduce the amount of memory required by the database.

An independent set of 20 speakers was used for training speaker independent codebooks. The first utterance of each word from the A block data of each of these speakers was used to train a discrete codebook of 256 codewords and a semi-continuous codebook of 50 probability density functions (pdfs). The codebook data consisted of a total of 35,000 cepstral vectors (20 ms frames with 15 ms overlap). A set of 12 cepstral coefficients were used as the parameter set for all experiments.

A second set of 11 speakers, 6 male and 5 female was used for training and an additonal 9 impostor speakers added for the verification tests. The 5 token models were trained with the A block data and the 10 token models were trained with the A and the B block data. In both cases the models were tested against the C block data, giving 5 true speaker tokens and 95 imposter tokens for each digit for each speaker. Results were collected over all speakers to obtain 55 true scores and 1,045 impostor tests for each digit.

The data were collected from paid volunteers calling from their own homes. Although the data have been collected over a period of 6 months, the channel for a given speaker will almost certainly have non-random characteristics, due mainly to the handset, but also from the network. It is possible that the system is helped by the channel characteristics when verifying a speaker. This source of bias would be very difficult to quantify without further experiments on another database. It will at worse affect any quantitative measure of the system performance compared to other systems. It should not affect the validity of using the system to compare various configurations of HMM models, since the all factors up to and including the extraction of the cepstral parameters have been keep constant.

A codebook of 50 codewords with diagonal covariance pdfs was used for the SCHMMs. Re-estimation of the codebook was not integrated into the Baum-Welch algorithm for the SCHMM models. A standard codebook, trained using modified k-means (J.G.Wilpon and Rabiner, 1985) was used for all speakers and all words.

Performance measures

In a verification system, two types of error are possible. The first, called false rejection (FR), occurs when a genuine speaker is rejected. The second, called false acceptance (FA) occurs when an impostor is accepted by the system. The two types of error will have varying importance depending on the application. False rejection errors cause inconvenience and frustration. False acceptance errors are a security risk.

The 'zero FA' rate is measure of the inconvenience of a system with perfect security. The 'zero FR' rate is a measure of how secure a very unobtrusive system would be.

The equal error rate (EER) is obtained by choosing a decision threshold such that FA = FR. In one sense this is an ideal threshold since it minimises the two error rates using knowledge of the correct data. In real applications a threshold will have to be determined without this knowledge, and so the equal error rate represents an upper bound to performance (it assumes a perfect choice of threshold).

Overall equal error rates are obtained using a common threshold for all digits and all speakers. Average equal error rates are also quoted. This is the average of the equal error rates from each of the digits. There is an implicit assumption in taking this average that there is a different threshold for each digit, but that the thresholds are common across all speakers. It would be possible to have different thresholds for each speaker, which would result in different performance levels for each speaker. Having thresholds which are both speaker and digit specific would probably reduce the average equal error rate. In a real application however, the threshold would have to be determined from the training data. If the training data is sparse, it would be very difficult to reliably estimate useful speaker-specific thresholds. The average equal error rate measure used in this paper is therefore a realistic one.

In addition to the error rates for each digit, results are given for strings of isolated digits, such as could be obtained from someone reading out a credit card number, with pauses between each digit. Increasing the length of the digit sequence reduces the error rate, up to a limit caused by correlation of the data samples. The digit string probability is obtained by a summation of the probabilities of each of the digits. The seven digit string consists of the seven digits with the smallest equal error rates.

Training

Speaker independent speech recognition allows detailed modelling to be done with large amounts of training data. Speaker verification models are inherently speaker dependent. This means that all users of the system must supply training data. In some small scale applications such as door entry systems and computer security, it may be feasible to collect large amounts of data for each speaker. On mass market telephone based applications, such as telephone banking the amount of training data that can be gathered is severly constrained by the level of inconvenience to the user. The target market, consisting of people who require a rapid, convenient service, may be unwilling to record a set of data 20 times.

The performance of a speaker verification system varies greatly with the amount of training data used

(Rosenburg et al., 1990). While excellent performance can be obtained from systems with large amounts of training data, this is not a good measure of whether HMM based verification systems are of any practical use. The training data available for speaker verification applications will be strongly constrained.

Adaptive training can be used with HMMs to make use of new data gathered as the system is being used (Rosenburg et al., 1990). Here however, the decision on whether new data are valid and should be used for adaptive training, must be made by the verification system itself. This is a potentially unstable situation.

Training data is therefore a crucial factor in determining whether HMMs will be useful in speaker verification. The aim of this research is to get the best performance possible given a small amount of training data.

DHMM VERIFICATION

Two series of experiments were carried out with a DHMM system using 5 and 10 tokens of training data respectively.

Standard HMMs use a fixed transition probability to model temporal features. Hidden semi-Markov models (HSMM) can be used for more detailed modeling of the durational aspects of the speech data. Although it is generally accepted that duration modelling improves the performance of an HMM speech recognition system (Russel and A.E.Cook, 1987; Levinson, 1986), full duration modelling is rarely used because of the large increase in computational requirement.

Speaker verfication does not involve searching through a lattice or grammar, and so extra computation can be accomodated without sacrificing real-time performance. It seems reasonable that durational information may assist in discrimating between speakers, and so time duration modelling using Guassian duration probability density functions was used in these experiments.

One of the many variable parameters in an HMM system is the number of states used to model a speech unit. The number of states should correspond roughly to the number of distinct events in the speech unit. The ideal number of states is also affected by the amount of training data used. Sparse data may not be sufficiently rich in information to train a large number of states. Models with 3 states and with 6 states were constructed for each digit. With 10 tokens the average EER for 3 and for 6 states is very similar which suggests that the number of states is not critical.

Using the correct number of states is possibly more important with only 5 training tokens. In speaker recognition systems, interpolation is used to smooth the output probabilities when training data is sparse. This involves a blending of the output probabilities from a speaker independent model with those derived from the training data. In speaker verification, where discriminating between speakers is the objective, this technique is unlikely to be helpful, and has not been used here.

Equal error rates for individual digits ranged from 12% to 28% for the 5 token models and 8% to 17% for the 10 token models. Figure 1 shows the average EER for the 3 state and 6 state models with 5 and 10 training tokens.

SCHMM VERIFICATION

It has been shown that SCHMMs offer improved performance in speaker recognition (Huang and M.A.Jack, 1988a; Huang and M.A.Jack, 1988b; Huang et al., 1990a). This is due to the more accurate representation of speech offered by a codebook of pdfs, which avoids the quantisation errors inherent in the DHMM approach. Despite the success in speech recognition, SCHMMs have not been tried in speaker verification.

Vector quantisation makes a hard, thresholded, decision by labelling each frame as being either 'a' or 'b' or 'x'. Pdf codebooks soften this decision by decribing a frame as 'partly a' 'partly b' and 'partly x'. The binary weighting of codewords in DHMM becomes continuously variable weighting in SCHMM. The softening of this decision makes the models more tolerant of minor variations between the test data and the training data, which is very important in speaker independent speech recognition. The results obtained to date indicate that the softer decision is a disadvantage in speaker verification.

The average EER for the SCHMM with 6 states and 10 training tokens is compared with the results for the DHMM in Figure 1. A comparison of the ERR curves between SCHMM and DHMM for 10 training tokens is shown in Figure 2. It can be seen from the slope of the EER curves that the spread of scores from the SCHMM system is significantly less than in the DHMM system. This is probably due to the 'soft' decision. The lack of spread makes the SCHMM system very sensitive to the choice of threshold. However, when several scores are added together, as with the 7 digit string, the spread is increased, so the sensitiviy is reduced and the EER becomes better for SCHMM than for DHMM.

A key factor which may be reducing the performance of the SCHMM is the codebook. The power of the SCHMM is partly due to the ability to integrate codebook re-estimation into the training process. This was not done for this experiment, the modified k-means algorithm being used instead. Also it has been suggested (Huang et al., 1990b) that the use of diagonal covariance pdfs with a cepstrum parameter set is inappropriate due to the high correlation of the components of the cepstral vector.

CONCLUSION

Although the EER for the SCHMM system is a little higher than DHMM for single digits, the ERR for the 7 digit string is slightly better. Larger amounts of true speaker data are needed before any statisically significant conclusions can be drawn from this.

There is clearly room for further speech recognition techniques to be applied to speaker verification. The results obtained from the SCHMM indicate that the full power of this technique is not being exploited. A new parameter set, with multiple codebooks (perhaps cepstrum, delta cepstrum and pitch) together with full covariance pdfs and integration of codebook re-estimation into the training process should help to rectify this.

REFERENCES

Huang, X. D., Lee, K., and Hon, H. (1990a). Large-vocabulary speaker-independent continuous speech recognition with semi-continuous hidden-Markov models. In *Eurospeech*, pages 163–166.

Huang, X. D., Lee, K., and Hon, H. (1990b). On semi-continuous hidden-Markov modeling. In *Proc. IEEE International Conference on Acoustics, Speech, Signal Proccessing*, pages 689–692.

Huang, X. D. and M.A.Jack (1988a). Performance comparison between semi-continuous and discrete hidden Markov models of speech. *Electronics Letters*, 24(3):149–150.

Huang, X. D. and M.A.Jack (1988b). Semi-continuous hidden-Markov models in isolated word recognition. In *IEEE 9th International Conference on Pattern Recognition*, pages 406–408.

J.G.Wilpon and Rabiner, L. (1985). A modified k-means clustering algorithm for use in isolated word recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 33(3):587–594.

Levinson, S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. In *Computer Speech and Language*, pages 29–45.

Rosenburg, A., Lee, C.-H., and Soong, F. (1990). Sub-word unit talker verification using hidden Markov models. In *Proc. IEEE International Conference on Acoustics, Speech, Signal Proccessing*, pages 269–272.

Russel, M. and A.E.Cook (1987). Experimental evaluation of duration modelling techniques for automatic speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, Signal Proccessing*, pages 2376–2379.

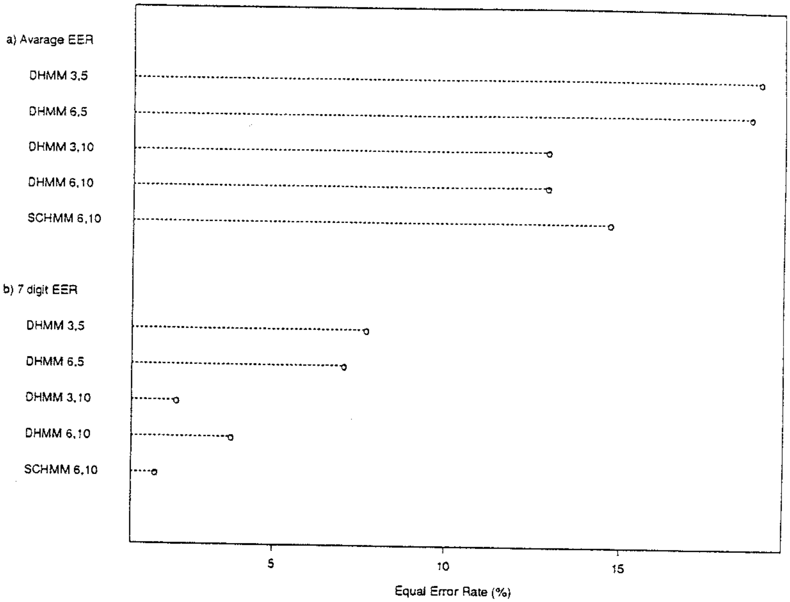
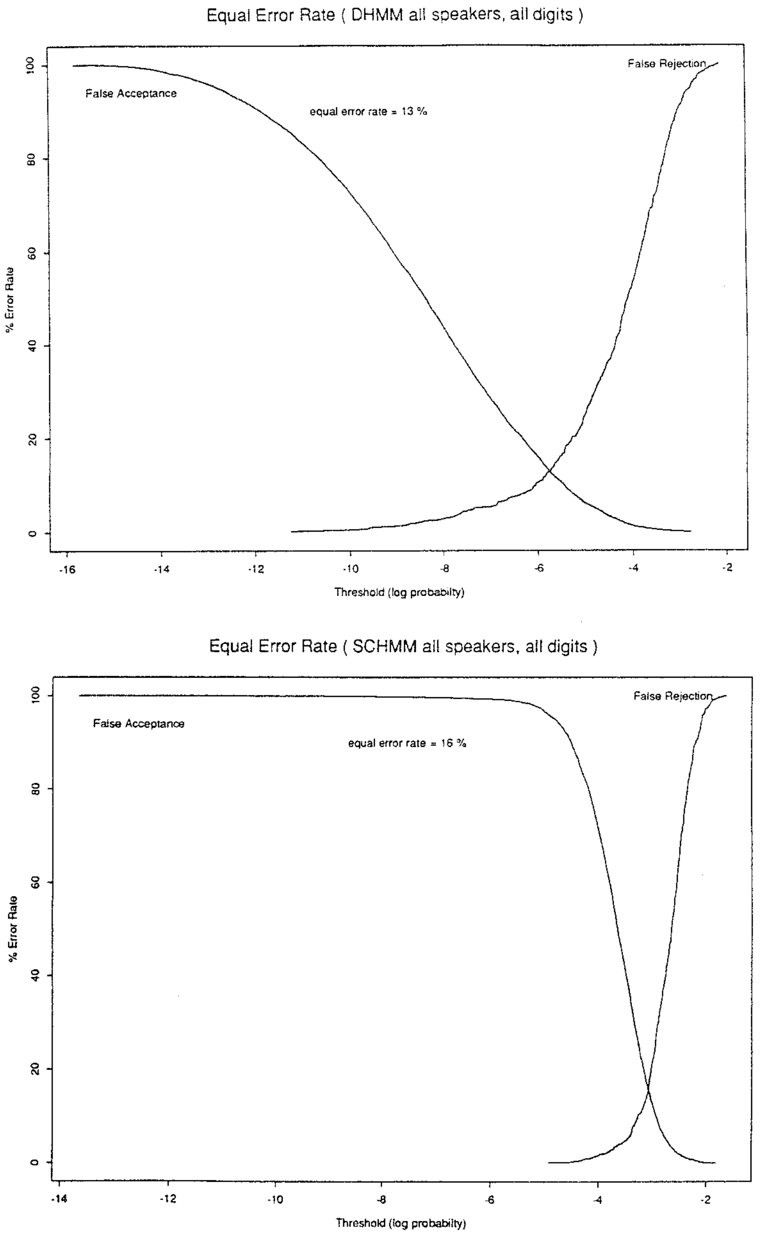


Figure 1. Average equal error rates for:

DHMM 3,5	:	DHMM	3 states	5 training tokens
DHMM 6,5	:	DHMM	6 states	5 training tokens
DHMM 3,10	:	DHMM	3 states	10 training tokens
DHMM 6,10	:	DHMM	6 states	10 training tokens
SCHMM 6,10	:	SCHMM	6 states	10 training tokens


Equal Error Rate (DHMM all speakers, all digits)

Equal Error Rate (SCHMM all speakers, all digits)

Figure 2. Overall equal error rate curves calculated over all speakers and all digits.
a) DHMM 6 states, 10 tokens
b) SCHMM 6 states, 10 tokens