DISCRIMINATING SEMI-CONTINUOUS HMM FOR SPEAKER VERIFICATION.

M.E. Forsyth, M.A. Jack

Centre for Speech Technology Research
80 South Bridge, Edinburgh, EH1 1HN, SCOTLAND, UK

ABSTRACT

This paper describes the use of a multiple codebook SCHMM speaker verification system, which uses a novel technique for discriminative hidden Markov modelling known as discriminative observation probabilities (DOP). DOP can easily be added to a multiple codebook HMM system and require minimal additional computation and no additional training. The DOP technique can be applied to both speech and speaker recognition. Results are presented for text-dependent experiments on isolated digits from 27 true speakers and 84 casual imposters, recorded over the public telephone network in the United Kingdom.

DOP are shown to significantly improve speaker verification performance for several commonly used parameter sets.

1. INTRODUCTION

Semi-continuous hidden Markov Models (HMM) have previously been shown to be effective in the field of speech recognition [1], however this technique has only recently been applied to the field of speaker recognition [2, 3].

It has been shown that semi-continuous HMM (SCHMM) is superior to discrete HMM (DHMM) for speaker verification [3] and that state duration modelling (hidden semi-Markov models), and the use of multiple codebooks both provide significant benefits to a speaker recognition system [2].

This paper extends the work on multiple codebooks by testing a novel technique known as discriminating observation probabilities (DOP). The DOP technique is evaluated for cepstra, delta cepstra, mel-frequency cepstral coefficients (MFCC) and difference MFCC. DOP can be used in both speech and speaker recognition. Section 2 outlines the motivation and rationale for the DOP technique and section 3 describes the technique itself. Section 4 describes the database used in these experiments.

The multiple codebook SCHMM system and the novel technique used to isolate the speaker discriminating power of each codebook are described in section 5. The results are in section 6.

2. CONVENTIONAL MODELLING FOR SPEAKER VERIFICATION

The conventional way to apply HMM to the task of speaker verification is to make speaker-dependent models of a speaker. The verification procedure is then a matter of comparing the speech to be tested against the model. The Viterbi algorithm can be used to determine the probability of the speech having come from the model. If the probability is above a certain threshold the bid is accepted. The essence of this approach is *speech* modelling as opposed to *speaker* modelling. The probability of the speech coming from the model is, in a sense, a combined *speech* and *speaker* recognition probability. If the test data is noisy or distorted the false rejection rate will increase. This is because a noisy test utterance from a genuine speaker will not fit the speech model well, possibly leading to false rejection. Note that noise will not cause an imposter's speech to fit the speech model any better, and so will not increase the chance of false acceptance.

In some systems a normalisation technique has been used successfully to reduce the effect of speech modelling masking the speaker modelling . In particular it has been used to reduce the variation in speaker recognition scores caused by different telephone microphones [4]. This takes the form of an offset in the verification threshold which is proportional to the *speech modelling* quality of the test data. The size of the offset is determined by matching the test data with an independent set of reference models trained from speakers who are similar to the speaker whose identity is being verified.

Although normalisation has been shown to be a useful technique, it is simply a compensation for the fact that conventional HMM does not explicitly discriminate between speakers.

3. DISCRIMINATIVE MODELLING FOR SPEAKER VERIFICATION

In order to address the lack of explicit discrimination between classes in conventional HMM, a novel technique using discriminative observation probabilities (DOP) has been developed. The normalisation technique which is now commonly used in speaker verification is similar to, but significantly different from, a special case of DOP HMM. The procedure for generating a DOP HMM for a speaker (speaker A) is as follows.

- Train a conventional HMM for speaker A (model A)
- Train a conventional HMM as a reference model using appropriately chosen speech data (model R)
- Take the differences in the observation probabilities of model A and model R.
- Normalise the differences into probabilities in the range 0 to 1.
- Create a DOP model for speaker A by using these probabilities as the observation probabilities for the DOP model. The DOP model is not a separate model but is treated similarly to the various codebooks in a multiple codebook system

0-7803-1775-0/94 $3.00 © 1994 IEEE

For these experiments the reference model was a general speaker independent model. The effect of this is that the new observation probabilities reflect what is different about speaker A compared to the rest of the population. If an acoustic observation occurred frequently in speaker A's training data but not so frequently in the speaker independent training data then the appearance of that acoustic observation in the test data is a good indication that the speech came from speaker A, and therefore the discriminating observation probability (DOP) is high. Likewise, if a codeword occurs frequently in the speaker independent training set but not in the training data of speaker A , then the appearance of that codeword in the test data is an indication that the speaker is not speaker A and so the DOP will be low. If the frequency of a codeword is similar for speaker A and for the speaker independent set then that codeword will not be useful in distinguishing speaker A and the DOP will be neutral (around 0.5).

DOP HMM has the following technical benefits

- A DOP model can be derived from a conventional HMM with no extra training
- The DOP model can be easily implemented as another information stream in a multiple codebook system.
- DOP models can be generated for all parameter sets in a multiple codebook system, doubling the number of information sources available for the verification decision.
- The information from the DOP model is at least partially independent from the information from the conventional model
- DOP models require minimal extra preprocessing.

3.1. Generalised DOP models

In these experiments the DOP models have been used to discriminate between a single speaker and a general speaker independent set. By choosing an appropriate reference model a DOP model can be created to maximise discrimination between any two groups of one or more speakers. For example, an obvious extension to this work would be to follow the approach used with normalisation and use a group of speakers who are are similar to speaker A to make the reference model, thereby maximising the discrimination between speaker A and speakers who sound like speaker A (cohort speakers). Note that if this would not be the same as normalisation because the segmentation is based on the true speaker model and not on the cohort model. Also DOP allows more flexibility in the codebook weighting than is possible with normalisation.

If the requirement of a system was to discriminate between male and female speakers, a model of male speakers could be trained and a model of female speakers used as the reference model.

The application of DOP models is not limited to discrimination between speakers. In speech recognition DOP models could be used to increase the distinction between commonly confused speech units. For example DOP models could increase discrimination between two phones or between a phone and a group of similar phones.

4. DATABASE

The data consists of twelve isolated digits (digits 'one' to 'nine' plus 'zero', 'nought' and 'oh'), recorded over the telephone, over a period of six months. A group 20 speakers (9 males, 11 females) are modelled by the system and an independent set of 84 imposter speakers is used for testing. There are 20 true speaker utterances and 84 impostor utterances in the test set for each digit. The data are all end-point detected to remove excess silence and minimise storage requirements.

The database is similar to the one used in [2] but with more speakers in the training set, and more occurrences of noisy or distorted data.

The codebooks used are of size 32 and were trained from an independent set of 20 speakers. The frame size was 20ms with 15ms overlap. The delta (first order difference) cepstra data used a window of 5 frames (current frame plus 2 frames either side).

4.1. Training

As in [2, 3], an emphasis has been placed in this work on using a minimal amount of training data, in the belief that the amount of training data will be strongly constrained in most large scale telephone applications, such as telephone banking.

Another significant factor is that the training data was recorded in a single session, whilst the test data was recorded over a period of six months. This is the most difficult case, since there can be significant variation in both the speakers voice and the telephone channel over different recording sessions. Five training tokens were used for each word model, with 6 states per model. A Gaussian distribution was used for duration modelling. The top six codeword probabilities for each speech vector were used in the HMM verifier.

The multiple codebook models were trained using only the cepstral codebook. All parameter sets were re-estimated but only the cepstral codebook was used to calculate the observation probabilities which were used to optimise the model in the Baum-Welsh algorithm. In other words, the cepstral codebook was used for segmenting the data into states in the baum-welsh re-estimation. This could lead to a advantage for the cepstral parameter set over the other parameter sets. For example, the performance of MFCC against the cepstral parameters may be different if the MFCC parameters were used for segmentation.

5. ISOLATING EACH PARAMETER

The verification process involves a Viterbi search through the silence/word/silence HMM lattice to determine the path with the highest probability. This *Viterbi path* is then used to calculate a verification score. The Viterbi path can be given as a frame interval defined by a beginning frame $t_{b,s}$, an end frame $t_{e,s}$ and a duration T_s for each state s of N states.

The system uses four parameter sets in four codebooks for verification (cepstra, delta cepstra, MFCC, delta MFCC). For training and for finding the Viterbi path during verification only the cepstra codebook is used.

It is not proposed that all these parameter sets would be used in a verification system. Part of the aim of this research is to determine which parameter sets have the best speaker discriminating ability. It is likely that some combination of some of the parameters will prove to be optimum

The DOP for each of the parameter sets are treated within the HMM as if they came from an another parameter set, although the DOP obviously use the same codebook as the parameter they are derived from. The cepstra DOP, for example, will use the same codebook as the normal cepstra observation probabilities.

The verification score is calculated as shown in equation 1. The duration probability $P(T_s/s)$ has a weighting d.

A codebook m of the C codebooks has a weighting c_m. The set of observations for the frame interval $t_{b,s}$ to $t_{e,s}$ for codebook m is denoted $O_{m,s}$.

$$\prod_{s=1}^{N}\left[P(T/s)^{d}\prod_{m=1}^{C}P(O_{m,s}/s)^{c_m}\right] \qquad (1)$$

The probabilities from the front and back silence models are not included, as they contain no speaker discriminating information. For all experiments described here the duration weighting was kept fixed ($d = 0$).

Each parameter set in the multiple codebook system and its DOP counterpart was tested in isolation for verification performance. To do this the Viterbi path was found using the duration plus cepstra information which was used in training. The probability score that was calculated on the backtrace, was solely the contribution from the parameter being examined. The weightings for testing parameter i in isolation are shown in equation 5.

$$d = 0, \quad m \neq \imath \ c_m = 0, \quad m = i \ c_m = 1 \qquad (2)$$

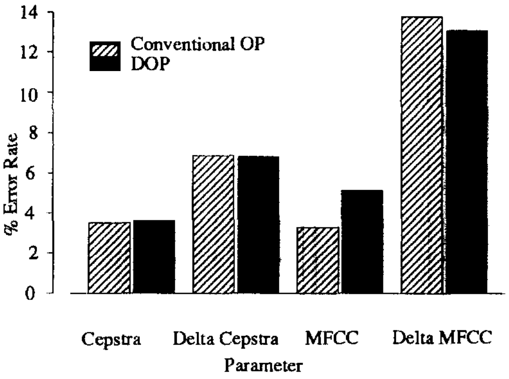

Figure 1. Comparison of each parameter set in isolation. Conventional HMM (striped) and the corresponding DOP on its own (black). 12 digit string EER for each parameter set

6. RESULTS

Figure 1 gives the EER for each parameter tested in isolation. The DOP all show significant speaker discriminating power, -comparable, in fact, to the conventional models. The test utterance consists of a concatenated sequence of the twelve isolated digits.

While the results in Figure 1 show that DOP have significant speaker discriminating power, the inclusion of DOP into a conventional HMM will only be useful if the speaker discriminating information of the DOP and the conventional observation probabilities are at least partially independent. In other words, if the conventional observation probabilities and the DOP make *different* errors then it may be possible to combine them to get a better result than is possible with either one alone.

Figure 2 show the difference in the EER between cepstra and DOP cepstra for each speaker. It can be seen from the mix of light and dark bars that the two information streams *do* complement each other. For speakers {2, 3, 5, 18, 19} DOP offers significantly fewer errors, while for speakers {4, 7, 14, 15, 20} straight cepstra produces fewer errors. This is encouraging , since not only does DOP provide speaker discriminating information but it provides *new* information.

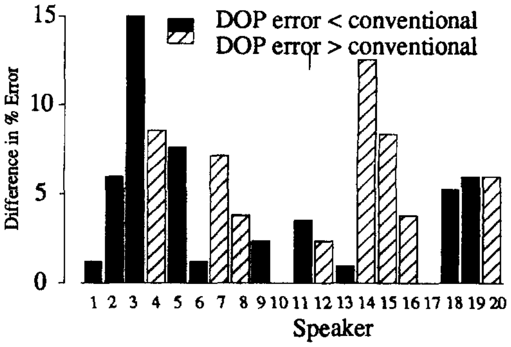

Figure 2. Independent speaker discriminating information. This plot shows the difference in conventional HMM and DOP HMM 12 digit sting EER for each speaker. Note that the two techniques have different strengths and weaknesses.

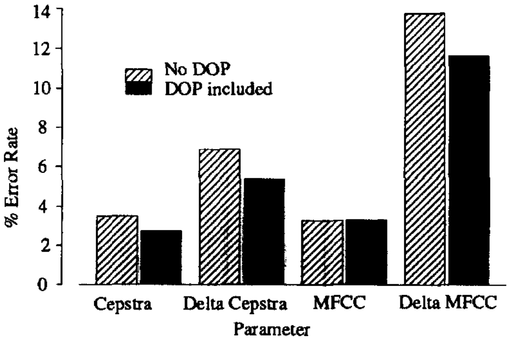

Figure 3. Comparison of conventional HMM (striped) and the conventional model with DOP included (black). 12 digit string EER for each parameter set

The next task is to combine the two information sources to produce a better EER. Initial attempts at including DOP into the conventional HMM using a weighted sum show that a clear advantage can be gained from the addition of DOP to the system for all the parameter sets. Figure 3 gives the comparative EER performance of the conventional model (striped) against the EER for the same parameter when the equivalent DOP are added (black).

Although equal error rates (EER) are the most common performance measure used in the literature, feedback from potential speaker verification users, such as banks, indicates that a negligible false rejection rate is crucial to the acceptability of a verification system [5] and so the zero false rejection (ZFR) error rate is perhaps a more useful measure of a systems performance. The ZFR rate is the false acceptance rate when the threshold is set such that there are no false rejection errors. The drawback of the ZFR rate is that it

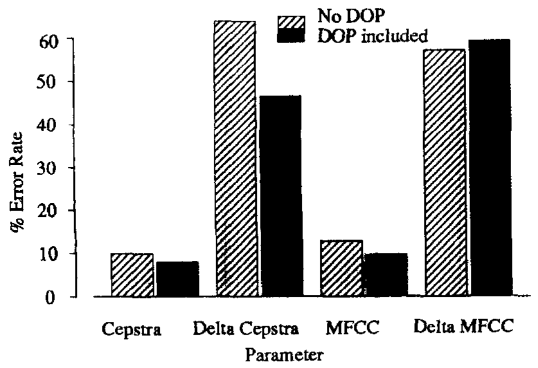

Figure 4. Comparison of conventional HMM (striped) and the conventional model with DOP included (black). 12 digit string ZFR for each parameter set

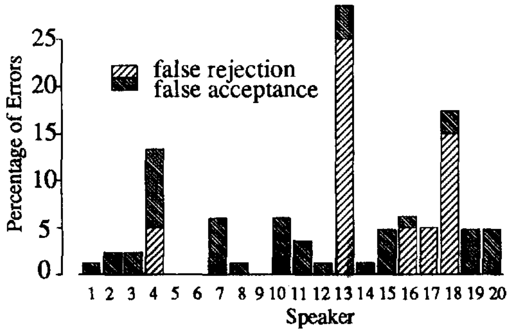

Figure 5. Breakdown of FR and FA errors by speaker, showing how the lack of speaker specific threshold increases the EER.

is sensitive to outliers in the database and so is not reliable when comparing systems using different databases. However if the database is the same it can be a useful measure for comparison.

Figure 4 illustrates the comparative ZFR rate performance of the conventional model (striped) against the ZFR rate for the same parameter when the equivalent DOP are added (black). There is clearly a general increase in performance as measured by ZFR when DOP are added. The weightings used for the results in Figure 3 were obtained from some simple trial and error experimentation, and are not optimal in any sense. They are , however , good enough to show that DOP is a useful addition to an HMM system. Optimal weightings could be obtained by many methods including discriminant analysis or by using a simple neural network. These approaches will be investigated in future work.

Some studies in the literature use speaker-specific thresholds to calculate EER results. Refer to citeForsyth93a for some discussion on why such EER are unlikely to be a realistic performance measure. In this work the EER thresholds are digit-specific but speaker independent. Figure 5 has a breakdown of false rejection (FR) and (FA) errors by speaker. The light bar represents FR errors and the dark bar represents FA errors. The potential advantage of using speaker-specific thresholds is clear. Thirteen out of the seventeen speakers with errors have fewer FR errors than FA errors. The other four speakers have far more FA errors than FR errors. This means that for each of the speakers with errors, the speaker independent threshold is either too low or too high. The difficulty in using speaker-specific thresholds arises from the limited amount of training data available. If a reliable threshold could be estimated for each speaker solely from closed test data a large improvement in performance could be gained.

7. CONCLUSIONS

DOP is a novel technique used to increase the discriminating power of HMM, which has been successfully used in a semi-continuous HMM speaker verification system to produce significant improvements in error rate. Although direct comparisons with other systems are not possible, due to the lack of a common database, the addition of DOP models shows a significant improvement over conventional HMM which are similar to those used in other systems [6, 7]. The technique is applicable to all applications of discrete, semi-continuous, or tied-mixture continuous HMM including speech recognition.

REFERENCES

[1] X. Huang, H. Hon, and K.-F. Lee, "Large-vocabulary speaker-independent continuous speech recognition with semi-continuous hidden Markov models," in *Proc. European Conference on Speech Communication and Technology*, vol. 1, pp. 163–166, Sept. 1989.

[2] M. Forsyth and M. Jack, "Duration modelling and multiple codebooks in semi-continuous HMMs for speaker verification," in *Proc. European Conference on Speech Communication and Technology*, 1993.

[3] M. Forsyth, A. Sutherland, J. Elliott, and M. Jack, "HMM speaker verification with sparse training data on telephone quality speech," in *Speech Communication*, Dec 1993. *In press.*

[4] A. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. Soong, "The use of cohort normalised scores for speaker verification," in *International Conference on Speech and Language Processing*, 1992.

[5] "DTI biometrics workshop," report, U.K.Department of Trade and Industry, 24 June 1992.

[6] A. Rosenberg, C.-H. Lee, and F. Soong, "Sub-word unit talker verification using hidden Markov models," in *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 269–272, 1990.

[7] A. E. Rosenberg, C.-H. Lee, and S. Gokcen, "Connected word talker verification using whole word hidden Markov models," in *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 381–384, 1991.