Incorporating Discriminating Observation Probabilities (DOP) into Semi-Continuous HMM for Speaker Verification.

M.E. Forsyth, P.C. Bagshaw, M.A. Jack

Abstract—
This paper describes the use of a semi-continuous hidden Markov models for speaker verification. The system uses a technique for discriminative hidden Markov modelling known as discriminating observation probabilities (DOP). Results are presented for text-dependent experiments on isolated digits from 25 genuine speakers and 84 casual imposter speakers, recorded over the public telephone network in the United Kingdom. Performance measures which are used to assess the DOP technique are equal error rate, zero false rejection rate, zero false acceptance rate and two measures of the distance between probability distributions for genuine and imposter speakers. The different performance measures are assessed with regard to their suitability for comparing speaker verification algorithms. This analysis further supports previous work which shows that the addition of DOP to an HMM system provides a significant advantage in speaker verification performance.

Keywords— **Semi-continuous HMM, speaker verification, discriminating observation probabilities (DOP), telephone speech, text-dependent, isolated digits, performance measures**

1. Introduction

The technique of incorporating discriminating observation probabilities (DOP) into an HMM has been reported as being beneficial in the speaker verification task [3]. This paper extends that work by employing a second data set and several performance measures to test the reliability of the initial results.

Section 2 describes the database used in these experiments and describes how a second set of data is created by rotating the database to improve the robustness of performance measures. Section 3 briefly describes the DOP technique, which is introduced in [3].

The various performance measures used in this paper are explained in Section 4, including a new distance measure specifically aimed at assessing the performance of verification systems. The parameter sets used in these experiments are cepstra, mel frequency cepstra (MFCC), and the corresponding difference parameters. Each parameter set is tested separately and in combination with the corresponding DOP scores.

2. Database

The procedure for parameter extraction and for training the HMM is the same as described in [3]. The database is also the same, except for the addition of 3 speakers to the training set. There are now 23 speakers (12 female and 11 male) with the set of 84 imposters remaining the same.

The training database is divided into 5 blocks each containing 5 tokens per word. These blocks are labelled a to e. The A data set referred to in these experiments involves training on the a block and testing against the b, c, d, e blocks. The B data set involves training on the b block and testing on the a, c, d, e blocks. The C data set involves combining the results obtained from the A and B data sets.

There are 20 genuine speaker utterances and 84 imposter speaker utterances in the test set for each digit. The data was end-point detected to remove excess silence and minimise storage requirements.

The data consists of twelve isolated digits (digits 'one' to 'nine' plus 'zero', 'nought' and 'oh'), recorded over the U.K. telephone network. The training data was recorded in a single session, with the test data being recorded over a period of six months.

3. Discriminating Observation Probabilities (DOP)

In order to address the lack of explicit discrimination between classes in conventional HMM, a technique using discriminating observation probabilities has been developed [3].

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 HMM model.

The DOP model is not a separate model but is treated similarly to the various codebooks in a multiple codebook HMM.

For these experiments the reference model is a general speaker independent model, trained with data from an independent group of 20 speakers. A reference model is trained for each digit.

DOP HMM has the following technical benefits:

- A DOP model can be derived from a conventional HMM with no extra training

Mark Forsyth, Paul Bagshaw and Mervyn Jack are all with the Centre for Speech Technology Research, 80 South Bridge, Edinburgh, EH1 1HN, SCOTLAND, UK. E-mail: forsyth@cstr.ed.ac.uk pcb@cstr.ed.ac.uk maj@cstr.ed.ac.uk

• 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 HMM, thus doubling the number of information sources available for the verification decision.
• DOP models require minimal extra processing.
• The results in Section 5 show that the combination of DOP scores and conventional HMM scores provides better speaker discriminating performance then either score alone.

4. Performance Measures

Speaker verification is concerned with the classification of unknown *bidders* into two classes, *genuine* speakers and *imposters*. There are two types of correct classification, the acceptance of genuine speakers, and the rejection of imposters. There are two corresponding types of errors, namely the rejection of genuine speakers, often called false rejection (FR), and the acceptance of imposters, often called false acceptance (FA).

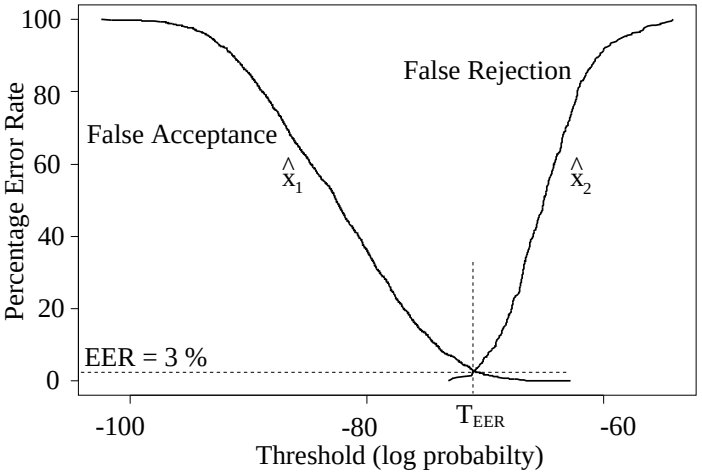

Fig. 1. Typical plot of FR rate and FA rate against choice of decision threshold. The EER, ZFR, and ZFA can be determined from this plot.

Figure 1 is a typical plot of FA rate and FR rate against the choice of decision threshold. Notice that there is a trade-off between FR and FA. Error rates for any given threshold can be determined from this plot. It is also possible for the trained eye to make some assessment of the robustness of the system to an imperfect choice of threshold. However, an objective measure of the separation of the genuine and imposter probabilities is still required to compare various algorithms and systems reliably.
There are several performance measures available for comparing speaker verification systems which measure different *aspects* of performance. The ZFR rate is the FA rate when no genuine speakers are rejected and the ZFA rate is the FR rate when no imposters are accepted. These measures are critically dependent on the worst genuine speaker score and the best imposter score, respectively. The ZFR and ZFA measures cannot be used as the sole basis for selecting one algorithm over another, since slight changes in the data could easily reverse the rankings of the algorithms, as can be seen in Section 5.2.

4.1 Equal Error Rate (EER)

The most common performance measure referred to in the literature is the equal error rate. This involves applying an *a posteriori* threshold T_{EER} which makes the percentage of FA and FR errors equal. It is important to make a distinction between whether T_{EER} is speaker-specific or speaker-independent [3], [4]. T_{EER} is speaker independent in these experiments.
The use of an EER implies a perfect choice of threshold, which is not possible in a real application since the threshold would have to be determined *a priori*. Therefore the EER provides an upper bound on performance and does not indicate how robust the system is to variations in data. Although EER is an important performance measure, it is also useful to have a measure of how well a system separates the probability distributions for the genuine speakers and the imposters. Such a measure would give an indication of the robustness of the system to an imperfect choice of threshold.

4.2 Mahalanobis Distance (MD)

A parametric measure of the distance between two statistical populations is the Mahalanobis distance [6], which assumes that the two populations have normal (Gauss-Laplacian) distributions. Consider that the two populations of log probabilities from imposter ($i = 1$) and genuine ($i = 2$) speakers are respectively represented by the sets,

$$x_i = \{x_{i,k} | k = 1, 2, \ldots, N_i\} \qquad i = 1, 2 \tag{1}$$

These populations are normal distributions with a Lilliefors' probability [5] of approximately one, although it is noted that their greatest deviations from normal distributions are above the 90^{th}-percentile for the imposter scores ($i = 1$) and below the 10^{th}-percentile for the genuine speaker scores ($i = 2$).
The Mahalanobis distance of two univariate normal distributions is given by,

$$D^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{\sigma_{12}} \tag{2}$$

where,

$$\bar{x}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} x_{i,k} \qquad i = 1, 2$$

$$\sigma_{12} = \frac{1}{N_1 + N_2 - 2} \sum_{i=1}^{2} \sum_{k=1}^{N_i} (x_{i,k} - \bar{x}_i)^2$$

The MD gives a measure of the separation between genuine speaker scores and imposter scores. Unfortunately, as is shown in section 5.1, this is not an ideal measure for the purpose of quantifying speaker discriminating performance. This is because the primary goal of a new algorithm is to reduce errors and most imposters are never mistaken for genuine speakers and most genuine speakers are not usually falsely rejected. Thus, the scores which most need to be improved are those near the equal error threshold.

The Mahalanobis distance assigns equal importance to all scores. A distance measure which targets the most important scores is required.

4.3 Targeted Distance Measure (TDM)

A figure of merit called the *targeted distance measure* is used in this paper. TDM targets the most important scores, namely the highest third of the imposter scores and the lowest third of the genuine speaker scores. It is calculated by the addition of two distance measures — TDM_{imp} for the imposter scores and TDM_{gen} for the genuine speaker scores.

$$TDM = TDM_{imp} + TDM_{gen} \tag{3}$$

where,

$$TDM_{imp} = 100. \left[\frac{1}{|\bar{x}_1 - \bar{x}_2|} . \frac{3}{N_1} \sum_{k=\lceil 2N_1/3 \rceil}^{N_1} (T_{EER} - \hat{x}_{1,k}) \right]$$

$$TDM_{gen} = 100. \left[\frac{1}{|\bar{x}_1 - \bar{x}_2|} . \frac{3}{N_2} \sum_{k=1}^{\lfloor N_2/3 \rfloor} (\hat{x}_{2,k} - T_{EER}) \right]$$

$$\hat{x}_{i,k} = k^{th} \text{member of } x_i \text{ sorted in ascending order}$$

This calculation takes an average *signed* distance from T_{EER} and normalises it with respect to the distance between the means of the two distributions. Note the reversal of sign between the calculation of TDM_{imp} and that of TDM_{gen}, so that a higher number corresponds to better performance in both cases.

5. Results

5.1 Comparing Performance Measures

Figure 2 is a comparison of 5 different parameter sets using seven different performance measures and three different data sets.

The ordinate measures performance, with the top representing the score of the best algorithm and the bottom representing the score of the worst algorithm. This means that the lowest error rates and the greatest distances are at the top. The ordinate is linear and has no absolute scale.

The seven performance measures, EER, ZFR, ZFA, TDM_{gen}, TDM_{imp}, TDM, and MD all have three vertical columns, one for each of the data sets. Each column has been normalised so that the *relative* performance of the five algorithms can be directly compared over all the performance measures, and all the data sets.

This figure is a comparison of performance measures as well as a comparison of algorithms. TDM shows a clear ranking of the algorithms. Not only is the ranking the same over all three data sets, but the relative differences in performance of the algorithms are the same over the three sets. This is an indication of a reliable performance measure, because it means that the relative merits of one algorithm over another can be assessed without undue sensitivity to the data set being used.

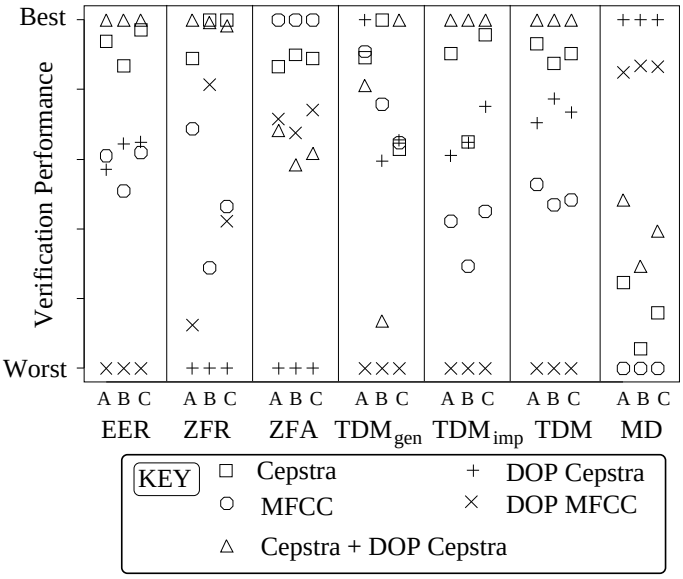

Fig. 2. Comparison of 5 different parameter sets using seven performance measures. Three different data sets are used, *A B* and *C*. The top of any vertical column represents the best algorithm for the given data set and performance measure.

Contrast this with the ZFR rate and the TDM_{gen} results. The relative positions of the algorithms change considerably between data set *A* and data set *B*, even though they are derived from the same database. These measures must therefore be used with caution.

The ZFA rate appears to be more reliable, although it should suffer from the same sensitivity as the ZFR rate, because it is a similar type of measure. It is interesting to note that the rankings from ZFA are different from the ranking of the other measures. This does not mean that it is a poor performance measure. It is a good measure of a different *aspect* of performance. The ZFA rate is a measure of system performance when security is the key requirement, taking priority over convenience and ease of use.

The Mahalanobis distance maintains the ranking for the different data sets and but does not appear to be measuring the same thing as the EER and the TDM. The MD favours the DOP algorithm in all cases. This means that the DOP scores are better separated overall than the conventional scores, but this has not lead to a corresponding reduction in real or potential misclassifications. This supports the need for the TDM.

Finally, the TDM_{imp} was more stable than the TDM_{gen}, which can probably be explained by the fact that TDM_{imp} is derived from 644 scores while TDM_{gen} is calculated from only 154 scores.

5.2 Adding DOP Scores to Conventional HMM scores

Several experiments were conducted using various combinations of normal cepstra and DOP cepstra. A simple weighted sum of the probabilities was employed, using the same method described in detail in [3]. Figure 3 shows the performance of the best of these combinations against

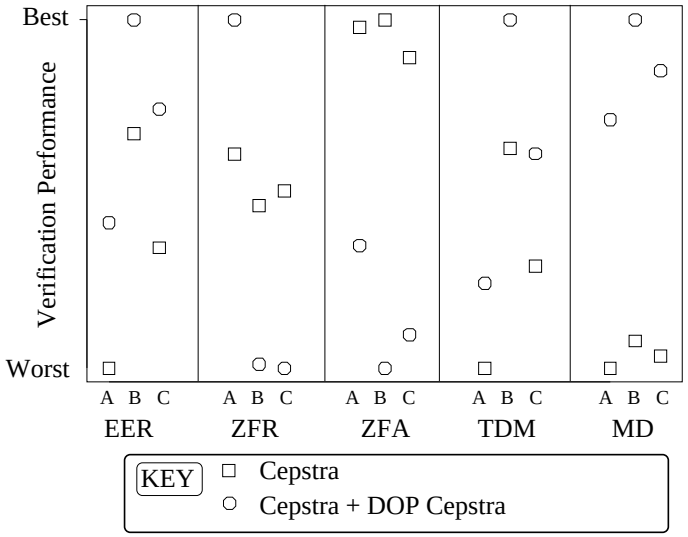

Fig. 3. Comparison of cepstra alone versus a weighted sum of cepstra plus DOP cepstra. Three different data sets are used, A B and C (which is the combination of the A and B sets). The top of any vertical column represents the best performance for the given performance measure.

cepstra alone. This figure differs from Figure 2 in that the results are normalised for each performance measure, instead of for each data set withing each performance measure. This allows some indication of the significance in the differences in the algorithms relative to the difference caused by using different data sets.

It can be seen that the EER, TDM and MD results were better for data set B than for data set A. As would be expected from a reliable performance measure, the results for data set C lie about half way between those for A and for B. The EER, TDM and MD performance measures all clearly illustrate the advantage of adding DOP to the system.

The results from ZFR and ZFA require some comment, since they illustrate the points made in Section 4. Cepstra without DOP gave clearly the best ZFA rate for all data sets, and on balance it was also superior for ZFR rate. It is interesting to note, however, that the *best* ZFR rate was obtained by DOP+CEP on data set A which other measures found to be the *hardest* of the data sets. This supports the proposition in Section 4 that these measures need to be used with caution.

The absolute values of the performance measures for the two algorithms can be seen in Table 5.2, along with the results for the other parameters tested. *No DOP* denotes only the conventional scores for that parameter were used, while *+DOP* denotes a combination of conventional and DOP scores. Note that since the TDM is a distance, the higher the number, the better the performance, while the reverse is true for EER.

The addition of DOP improves both performance measures for all the parameters tested. Comparison of these results with other studies in the literature [1], [7], [8] is not really possible because of the lack of a common database. Also note that state duration probabilities from the HMM have not been used in these experiments so that each algorithm can be examined in isolation. It has previously been shown that the inclusion of state duration probabilities significantly improves the EER [2].

TABLE I

The effectiveness of including DOP for several different parameters. All values are for the C data set.

Parameter	EER		TDM	
	No DOP	+DOP	No DOP	+DOP
Cepstra	2.95	2.49	3.53	3.69
Δ Cepstra	6.74	5.47	2.45	2.95
MFCC	3.88	3.86	3.42	3.49
Δ MFCC	12.71	11.19	0.36	0.97

6. Conclusions

A targeted distance measure has been developed which is a reliable complement to the conventional EER. It is easily calculated using the EER threshold. The TDM is a more useful measure for speaker verification than a total distance between the genuine and imposter probability distributions, such as the Mahalanobis distance.

The results of earlier work [3] on DOP HMM have been confirmed by experiments on a second data set. The incorporation of DOP scores lead to improvements in the EER and the TDM for a variety of parameters. Further investigation is required to find an optimal combination of multiple speaker discriminating information streams.

References

[1] J. de Veth, G. Gallopyn, and H. Bourlard. Speaker verification over telephone channels based on concatenated phonemic hidden Markov models. In *Eurospeech*, volume 3, pages 2279–2282, September 1993.

[2] M.E. Forsyth and M.A. Jack. Duration modelling and multiple codebooks in semi-continuous HMMs for speaker verification. In *Proc. European Conference on Speech Communication and Technology*, pages 319–322, September 1993.

[3] M.E. Forsyth and M.A. Jack. Discriminating semi-continuous HMM for speaker verification. In *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, April 1994. (in press).

[4] M.E. Forsyth, A.M. Sutherland, J.A. Elliott, and M.A. Jack. HMM speaker verification with sparse training data on telephone quality speech. In *Speech Communication*. (in press).

[5] H.W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of American Statistical Association*, 64:399–402, 1967.

[6] P.C. Mahalanobis. On the generalized distance in statistics. *Proc. National Institute of Sciences of India*, 2(1):49–55, 1936.

[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*, pages 381–384, 1991.

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