A Segmental Mixture Model for Speaker Recognition

Robert P. Stapert and John S. Mason

Department of Electronic and Electrical Engineering
University of Wales, Swansea
{robert.stapert@aculab.com, j.s.d.mason@swansea.ac.uk

Abstract

Standard Gaussian mixture modelling does not possess time sequence information (TSI) other than that which might be embedded in the acoustic features. Dynamic time warping relates directly to TSI, time-warping two sequences of features into alignment. Here, a hybrid system embedding DTW into a GMM is presented. Improved automatic speaker verification performance is demonstrated. Testing 1000 speakers in a fully text independent, world-model-adapted mode shows an equal error improvement over a standard GMM from 4.1% to 3.8%.

1. Introduction

All automatic speech and speaker recognition systems operate on a series of acoustic features extracted from measurements along the time course of the speech signal. The time sequence of these features is of critical importance in the case of *speech* recognition, since this task must distinguish between different utterances. This requirement is reflected in the structure of the recognition system, for example the long-standing dynamic time warping (DTW) [1] or the now universally accepted hidden Markov model (HMM) [2]. Both have in-built sequence processing, enabling them to capture the important time sequence information (TSI).

The situation for the complementary task of *speaker* recognition is very different. For it is unclear how much discrimination exists in the TSI when the goal is to recognise the person rather than the utterance. One indication comes from today's wide-scale use of Gaussian mixture modelling (GMM) [3]. The GMM is a popular state-of-the-art approach to automatic *speaker* recognition. It is a probabilistic representation of speech space which can be tailored to the speech of an individual speaker. Since there are no state transitions, the conventional GMM can possess no TSI other than that which might be embedded in the speech features, for instance the dynamic regression components proposed by Furui originally for speaker recognition [4].

Given the current popularity of GMM's, particularly in text-independent mode, the implication is that it has not proven possible to harness TSI using, for example, HMM-type structures. One suggestion for this is that the over-arching state transitions in a standard HMM are too course, especially for a text-independent mode. However, some useful TSI obviously does exist when the task is *speaker* recognition, as is evident from the wide-scale use of dynamic regression features. This paper addresses this apparent dichotomy with a new hybrid structure, namely a GMM-like structure with an embedded DTW, here termed a segmental mixture model (SMM)[5]. The motivation is to harness any additional TSI to aid text independent person recognition. This paper seeks to take a state-of-the-art GMM speaker recognition system and introduce into this model additional information in the form of TSI. The approach is to replace the kernel of the GMM similarity measure by a similarity measure which embraces TSI. This is done here using the classic DTW approach. A simple HMM could equally well be used, but here the DTW option was chosen on the grounds that it provides a high level of granularity and an insight into events.

2. Gaussian Mixture Models

Spectral based speech features such as mel frequency cepstral coefficients are assumed to have multivariate Gaussian densities. This leads to a similarity measure that can be interpreted in terms of likelihood or probablility density functions (pdf). In a GMM recognition test the conditional pdf of an input $p(\vec{x}|\lambda)$, is given by Equation 1, where $\vec{x}$ is the input and λ is the model; $b_i(\vec{x})$ is the density for component i of the model given the input vector $\vec{x}$, and w_i is the component weight; i ranges from $1 \ldots M$ where M is the number of components in the model. The pdf $p(\vec{x}|\lambda)$ is the weighted sum of M densities.

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \quad (1)$$

The density of a single component is given by Equation 2 where Σ_i is the covariance matrix and $\vec{\mu}_i$ is the mean vector. D is the feature order.

$$b_i(\vec{x}) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} exp\{-\frac{1}{2}(\vec{x} - \vec{\mu_i})' \Sigma_i^{-1} (\vec{x} - \vec{\mu_i})\} \quad (2)$$

The pdf of the input speech X given the model λ is given by Equation 3.

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x_t}|\lambda) \quad (3)$$

where T is the total number of inputs, $\vec{x}_t$, over time.

In a GMM a mixture of Gaussian components act together to model the overall probability density function. Full covariance matrices are usually deemed unnecessary on the assumption of statistical independence for the mixture components. This allows simplification by using only the diagonal of the covariance matrix.

An important factor that affects the performance of a GMM is the model size (the number of components). Determining the number of components is important because a model that is too small or too large is likely to be sub-optimal for a given training set. In practice the size is dependent on the quantity of training data and determining the optimum size is largely empirical. By using a world model [6], the problem of an upper limit on the model size is alleviated to an extent. The optimum size is now

a function of the training data for both the world model and the client speaker. Here, a world model is employed and client models adapted from the world model [3]. Performance against model size is examined.

The initial seed for the world model can be created by any suitable algorithm, and it is common practice to use a vector quantisation (VQ) algorithm. Here VQ is adapted to work on speech segments, similarly to the matrix quantisation of [7] except that here DTW is incorporated and variances and weights are estimated.

3. Segmental Mixture Models

In the proposed segmental mixture model (SMM) each mixture component, λ, of the standard GMM becomes a short sequence of single components called a segment. The segments are compared using DTW. It is regarded as a template pattern matching approach where two sequences are optimally aligned and matched according to prescribed similarity scores. Here its potential benefit lies in the granularity of time-warping and matching of two given segments.

For the SMM, the similarity measure comes from the DTW process. Here this uses the same form of spectral measure as that given in Equation 2 leading to an equivalent pdf interpretation for the SMM output, now applied to a segment rather than single speech feature input. Thus the pdf for an input segment $\square x$ given a model is the sum of M weighted segment scores and is shown in Equation 4 where w_i is the segment weight and i ranges from 1 to M where M is the total number of model components.

A segment score $b_i(\square x)$ is given in Equation 5, where d_w is the DTW measure between an input segment $\square x$ and a model segment, and $\prod_k^K |\Sigma_i|^{-\frac{1}{2}}$ is the product of the diagonal covariance matrices taken along the DTW path. K is the size of the segment measured in vectors. The DTW measure is given in Equation 6, where W is the normalising term along the warp path of $\vec{x}_k$ and $\vec{\mu}_{ik}$.

$$p(\square x|\lambda) = \sum_{i=1}^{M} w_i b_i(\square x) \tag{4}$$

$$b_i(\square x) = \ln\{\prod_k^K |\Sigma_i|^{-\frac{1}{2}} exp[-\frac{1}{2} d_w]\} \tag{5}$$

$$d_w = W_k^K((\vec{x_k} - \vec{\mu_{ik}})' \Sigma_i^{-1} (\vec{x_k} - \vec{\mu_{ik}})) \tag{6}$$

4. Experiments

The data comes from 2000 speakers recorded over the public switched telephone network [8]. One thousand of the 2000 speakers are used to create a world model and the other 1000 speakers are used for speaker model training and testing. Training is done on approximately 30 seconds of phonetically rich sentences per speaker with a total of about 8 hours for the world model. Text independent testing uses one digit per speaker per test. giving 1000 tests in total. Features are standard MFCC-14 static and 14 first order regression.

The effect of the time sequence information introduced by DTW is examined through several segment sizes, 1 to 5, where segment size 1 is equivalent to 32 milliseconds of speech and segment size 5 equals 96 milliseconds of speech, given a 50% overlap of the segments. Segment size 1 (which is where the SMM reduces to a standard GMM) is tested with and without

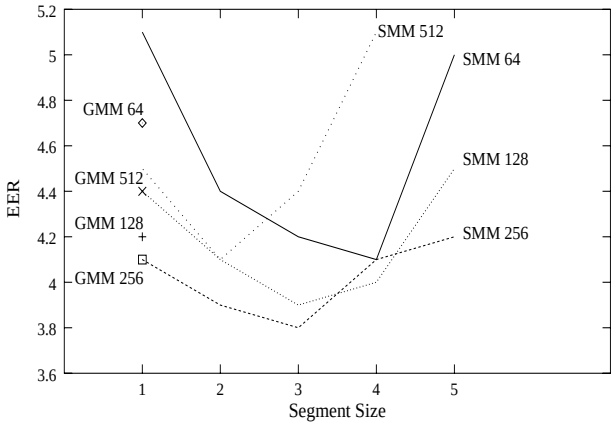

Figure 1: Speaker Verification % EER against SMM segment size for different model sizes. Also included are GMM results with re-estimation training.

re-estimation for direct comparison with a standard GMM using no DTW. Increasing the segment size implies longer patterns of speech sounds included in the DTW alignment. In addition to the five segment sizes, results for model sizes 64, 128, 256 and 512 are given, to show the effect of the model-size / amount-of-training-data ratio on the SMM. Equal error verification results are given in Figure 1 for the range of model sizes. Two scores are given for each model size at segment size one, those labelled GMM include re-estimation and show improvements over no re-estimation for all model sizes except 256. Both the GMM and SMM error rates drop with increasing model size from 64 to 256. At 512 the results degrade due to insufficient training data; see also Table 1. Furthermore, the SMM scores improve with increasing segment size up to an optimum after which they degrade; the amount of training data is again likely to be a factor here. When comparing the two systems, it is evident that that SMM offers improved speaker discrimination. This is shown in Table 1 where the SMM with segment size three is compared with a GMM (with re-estimation). The overall best score is obtained using an SMM with segment size three and model size 256. Further comparison between the SMM and GMM is

	64	128	256	512
GMM	4.7	4.2	4.1	4.4
SMM	4.2	3.9	3.8	4.4

Table 1: GMM and SMM % equal error rates for model sizes 64 to 512.

given in the form of detection error trade-off (DET) curves in Figures 3 to 5. The figures show that the error rates of the SMM (lower profile) are consistently lower than those of the GMM (upper profile) for model sizes 64, 128, 256 and 512.

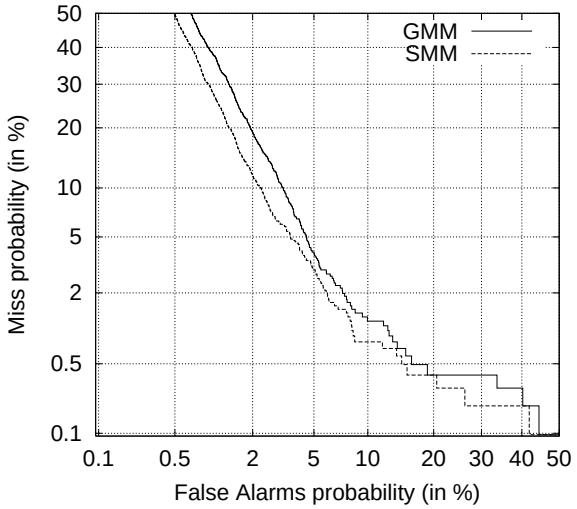

Figure 2: Speaker Verification DET curve, model size 64. Showing SMM performance (segment size 3, the bottom profile) and standard GMM with re-estimation (the top profile).

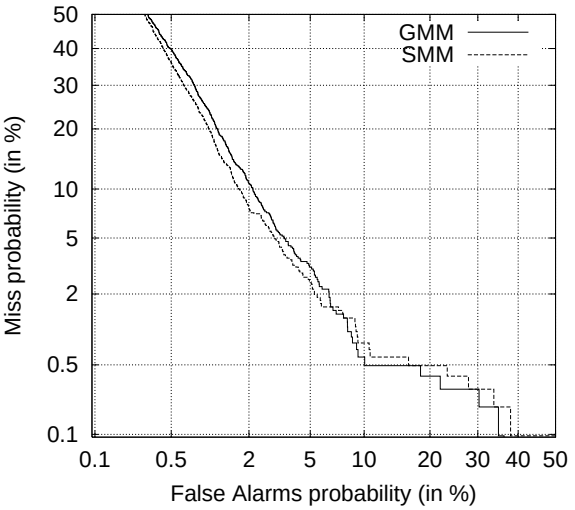

Figure 4: As for Figure 3 except model size 256

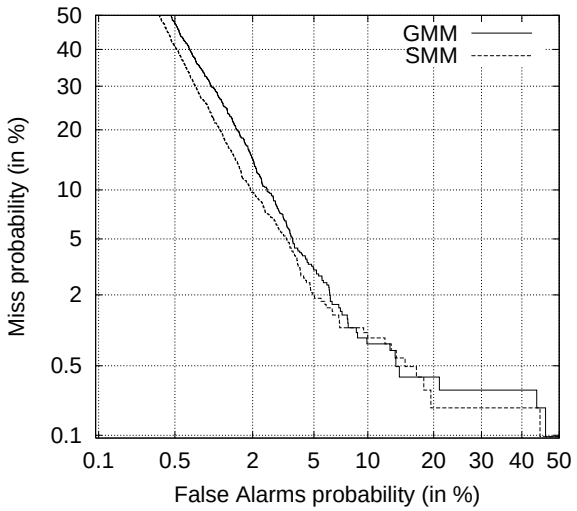

Figure 3: As for Figure 3 except model size 128

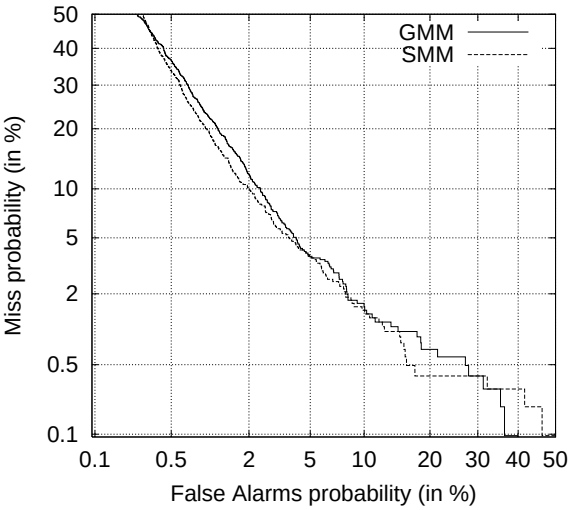

Figure 5: As for Figure 3 except model size 512.

5. Conclusion and Discussion

The framework of a state-of-the-art GMM has been extended to include time sequence information (TSI) in the kernel of the similarity measure. The resultant model is termed a segmental mixture model (SMM) [5]. TSI has been harnessed using DTW although it is likely that HMM sub-structures could also be used.

It is shown that the DTW extracted TSI supplements that inherent in the first-order dynamics of the cepstra. The sensitivity of model size to quantity of training data is well known. Here, using approximately 8 hours of speech to train a text independent world background model and about 30 seconds to adapt each speaker model, a model size of 256 mixture components is shown to be the optimum for the prevailing conditions.

The optimum sequence length for the SMM is found to be in the order of 100 milliseconds, though this also is shown to be sensitive to model size / quantity of training data. This duration corresponds well with that used in first-order derivative features. It should be emphesised that such features are used throughout the experiments reported here, and thus the TSI gained from the segmental approach is in addition to that coming from the first-order derivative features.

With a model size of 256 and segmental sequence length of about 100ms the SMM gives an EER of 3.8% compared with 4.1% for a standard GMM, when using single digit test tokens from 1000 speakers.

6. References

[1] H. Sakoe and S. Chiba. A Dynamic Programming Approach to Continuous Speech Recognition. *Seventh ICA*, page 20 C13, 1971.

[2] J. K. Baker. The Dragon system - an overview. *IEEE Trans. on ASSP*, 23:24–29, 1975.

[3] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72 – 83, 1995.

[4] S. Furui. Speaker-Independent Isolated Word Recognition using Dynamic Features of speech spectrum. *IEEE Trans. on ASSP*, 34:52–59, 1986.

[5] R. P. Stapert. A Segmental Mixture Model, maximising data use with time sequence information. *Ph.D. Thesis, University of Wales Swansea*, 2000.

[6] M. J. Carey, E. S. Parris, and J. S. Bridle. A speaker verification system using alpha nets. In *Proc. ICASSP*, volume 1, pages 397–400, 1991.

[7] D. K. Burton. Text-Dependent Speaker Recognition using Vector Quantization Source Coding. *IEEE Trans. on ASSP*, pages 133–143, February 1987.

[8] R. J. Jones, J. S. D. Mason, R. O. Jones, L. Helliker, and M. Pawlewski. SpeechDat Cymru: A large-scale Welsh telephony database. In *Proc. LREC Workshop: Language Resources for European Minority Languages*, 1998.