TIME SEQUENCE INFORMATION WITHIN A GAUSSIAN MIXTURE MODEL

R.P.Stapert and J.S.Mason
Speech and Image Research Group
Department of Electrical and Electronic Engineering
University of Wales Swansea
Department of Electrical and Electronic Engineering
SA2 8PP, UK
Robert.Stapert@aculab.com J.S.Mason@swan.ac.uk
http://galilee.swan.ac.uk *

ABSTRACT

This paper addresses the task of text independent speaker recognition and in particular looks at capturing time sequence information within the modelling process itself. A recent extension to the popular Gaussian mixture model (GMM) is the segmental mixture model (SMM), and its advantages are thought to be more pronounced as more and more training data becomes available. Here this idea is examined along with a hypothesis on model size, model complexity and their dependencies on the quantity of available training data. Experimental results on a 2000 speaker database show that an SMM does offer better recognition results than a GMM once a threshold in the amount of training data has been reached.

1 INTRODUCTION

Time sequence information (TSI) is the bastion of *speech* recognition where the task is to minimise speaker specific information in the models and maximise speech specific information. The temporal sequence in a signal is directly related to what is said, i.e. the speech content, and this is fundamental for speech recognition. In contrast in text independent *speaker* recognition it is sought to minimise speech specific characteristics and maximise characteristics associated with the speaker. To this end, vector quantisation and Gaussian mixture modelling do not include within the model any temporal constraints. The use of models that do use temporal information (such as hidden Markov models and dynamic programming techniques) has, until recently [1], shown no advantage over the GMM for text independent speaker recognition [2, 3]. However, TSI can also be harnessed from dynamic features, commonly used in both speech and speaker recognition and first proposed by Furui in 1981 [4] in the form of the regression features in the context of speaker recognition. It is the widespread use of such dynamic features in text independent speaker recognition that confirms the potential benefits of time sequence information.

In a recent paper [1] the current authors introduced the segmental mixture model (SMM). This uses dynamic time warping (DTW) to incorporate time sequence information. Yu *et al* [5, 6] has shown that DTW can be superior to vector quantisation VQ [7] and hidden Markov models HMM in a text dependent speaker recognition task and this provides motivation to combine DTW with the most popular text independent technique, namely the GMM.

The amount and quality of data are known to be influential factors in speaker recognition performance, and these factors are application dependent. At one end of the scale there might be just a few short utterances, at the other end of the scale, in terms of data, there is the situation where very large quantities of data are available, both for testing and training. Broadcast recordings present such a situation. It would be possible to collect large quantities of data from well-known broadcasters or entertainers, build models from this data and use these to search the archives for instances of these people. A recent paper by Doddington [8] addresses this latter scenario of large amounts of data, and shows that word frequencies are potentially useful in discriminating people. This idiolectic based approach in this case demonstrates the benefits of N-grams, thereby presenting an interesting contrast to the conventional atomic unit level spectral-based approaches which have dominated the field to date.

The VQ and GMM approaches can be thought of as operating on atomic levels in speech space, with potentially many thousands of components in the model. However, Doddington has shown very clearly that information at a completely different level, well beyond the atomic level, can be useful. He has shown that a speaker will display a degree of text-dependency and, as a consequence, recognition systems should incorporate a corresponding degree of text-dependence. This raises the question of how best to harness the TSI information. In other words, how might the approach change as more data becomes available? This paper is concerned

* Robert Stapert is now employed at ACULAB UK; this work was performed as part of his PhD while at the University of Wales Swansea.

with the approaches taken to speaker verification as the amount of speech data changes.

It is obvious from [8] that a practical system based solely on atomic unit level spectral features with no higher level TSI will be sub-optimal, certainly when large amounts of data are available. Here we take a small step in this direction by considering small segments of speech, just beyond the atomic level. The belief is that with ever increasing amounts of data, these sgements can become larger.

Dynamic time warping is traditionally employed in text dependent tasks. However, by applying it to short feature sequences the DTW constraints can be useful in a text independent mode especially as part of a GMM. The SMM is a step from the pure atomic level of frames spanning tens of milliseconds towards the much higher level of N-grams for which the speech might span a second or so. As presented here it is only a small but potentially useful step; in practice greater steps are likely to demand greater amounts of training data.

2 SEGMENTAL MIXTURE MODEL

In a Gaussian Mixture Model (GMM) each component consists of a mean, a covariance matrix and a weight. The density for component i of the model given the input vector $\vec{x}$ is given by Equation 1 where Σ_i is the covariance matrix and $\vec{\mu_i}$ is the mean vector. D is the dimension of the vector. A simplified form, popular in practical speaker recognition, has each component consisting of a mean, the diagonal of the covariance matrix and a weight.

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

In the segmental mixture model, proposed in [9, 1], each mixture component of the standard GMM becomes a short sequence of single components called a segment. The segments are compared using dynamic time warping (DTW) [10], which has proved successful in both speaker and speech recognition. It is regarded as a template pattern matching approach where two sequences are optimally aligned and matched according to prescribed similarity scores.

For the SMM, the GMM similarity measure is modified to apply to segments rather than single vectors. The probability of input segment $\Box x$ given a model is shown in Equation 2 as the sum of M weighted segment densities, w_i is the segment weight and ranges from 1 to M where M is the total number of model components.

A segment density $b_i(\Box x)$ is equal to the simplified Equation 3, where d_w is the DTW warp difference between an input segment $\Box x$ and a model segment, and $\prod_k^K |\Sigma_i|^{-\frac{1}{2}}$ is the product of the diagonal covariance matrices taken along the warp path. K is the size of the segment measured in vectors.

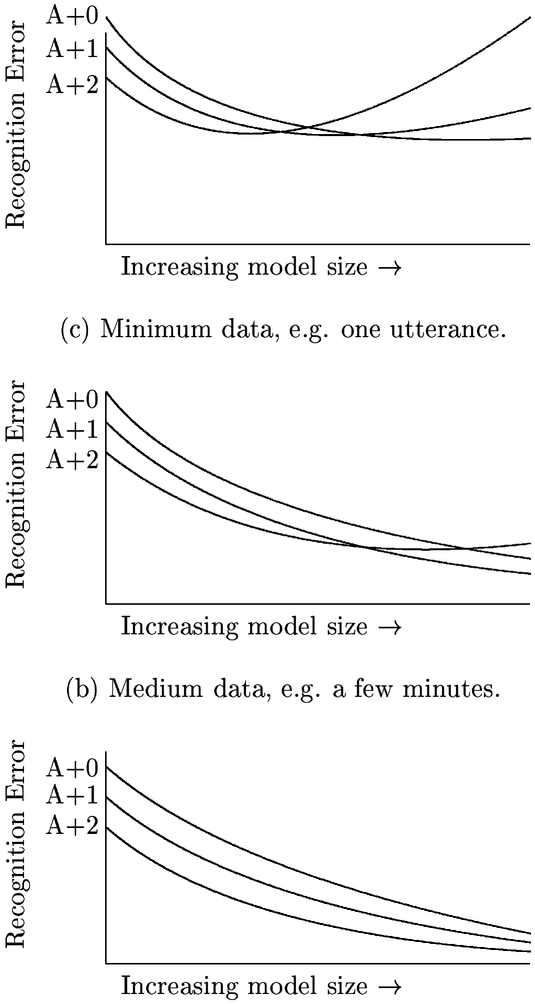
The DTW warp difference is given in Equation 4, where W is the normalised sum along the warp path, $\vec{x}$ and $m\vec{u}_{ik}$ are individual vectors of the test and model segments.

$$p(\Box x|\lambda) = \sum_{i=1}^{M} w_i b_i(\Box x) \tag{2}$$

$$b_i(\Box x) = \ln\{\prod_k^K |\Sigma_i|^{-\frac{1}{2}} exp[-\frac{1}{2}d_w]\} \tag{3}$$

$$d_w = W_k^K((\vec{x_k} - \vec{\mu_{ik}})'\Sigma_i^{-1}(\vec{x_k} - \vec{\mu_{ik}})) \tag{4}$$

3 MODEL COMPLEXITY



(c) Minimum data, e.g. one utterance.

(b) Medium data, e.g. a few minutes.

(a) Maximum data, hours to lifetime.

Figure 1: Hypothetical performance of three models of increasing complexity, A (Atomic) + 0 to 2, for three amounts of data, minimum (a), medium (b) and maximum (c).

It is proven that information is available both at the atomic level and at the N-gram level. To make use of this higher-level information much more training data must be available than that conventionally considered in speaker verification studies. And more data means potential for models with greater complexity.

Figure 1 depicts the performance of three hypothetical modelling approaches of increasing complexity, moving from the atomic (A) upwards (labelled A+0, A+1 and A+2). For this thought experiment the models are trained on a given (constant) amount of speaker specific data and they do not utilise general global speech information. In Figure 1(a) the amount of training data is small, for instance a single utterance, in Figure 1(b) more data is available, a number of minutes worth, and in Figure 1(c) it is very large, hours or even a lifetime's worth. The figure shows the hypothetical effect of increasing the model size on the recognition error for the three amounts of speaker-specific training data.

In Figure 1(a) The profiles show that the most complex model (A+2) curves upwards first, followed by the second most complex model, A+1. The amount of training data is small and is not sufficient to estimate complex models.

Figure 1(b) shows a similar scenario to Figure 1(a) except that more data is available. The most complex model begins to curve upwards at a certain point, due to under-training. This point represents the model size at which the amount of training data is no longer sufficient to accurately train all the components. However, the other models continue to improve as they are less complex and, importantly, not under-trained.

In Figure 1(c) the profiles remain monotonic with the error rates dropping as the model sizes increase. Note, in this case the profiles do not cross each other. The most complex model consistently has the lowest error rates and the least complex, the greatest error rates. In this case the amount of training data is large, sufficient to utilise the complexity of the most complex model (labelled A+2) without encountering under-training.

The three figures serve to illustrate that increasing the amount of training data will not only allow larger models, but also more complex ones, the benefit of which is seen in the error rates. Note however, that in the limits of large model size, quantity of data and minimum complexity then such profiles must always cross, or at least converge. For the optimum classifier under these conditions is the nearest-neighbour [11] and thus the least complex of all models is the best when the training data is infinite!

Considering the hypothesis illustrated in Figure 1 the current authors recently addressed the question of the nature of a model with increasing data, albeit over a constrained data range [12]. It is shown that increasing the complexity of a model leads to a reduction in error given a sufficient quantity of training data. This theme is examined further below.

4 EXPERIMENTS

The data comes from 2000 speakers recorded over the public switched telephone network [13]. 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. A total of about 8 hours of data is used for the world model. Testing is text independent using one digit utterance per speaker per test giving 1000 tests in total. The features are standard MFCC-10 static and 10 first order regression. Tests are run on GMM and SMM for comparison purposes.

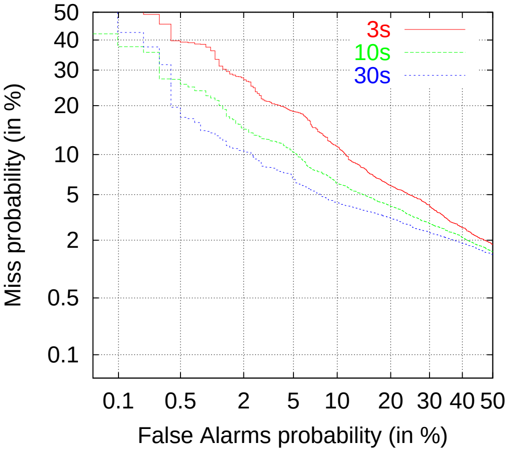


Figure 2: GMM DET curves for (approximately) 3, 10 and 30 seconds speaker training data.

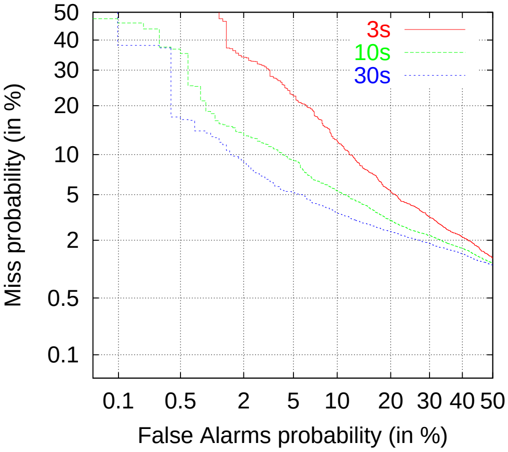


Figure 3: As for Figure 2 but for SMM.

The effect of the time sequence information introduced by DTW is examined through three levels of training: 3 seconds, 10 seconds and 30 seconds of phonetically rich sentences per speaker. The goal is to demonstrate that with sufficient traing data the SMM with its (albeit small) level of TSI can out perform the standard GMM. A model size of 256 components is used throughout based on preliminary results for this configuration [12].

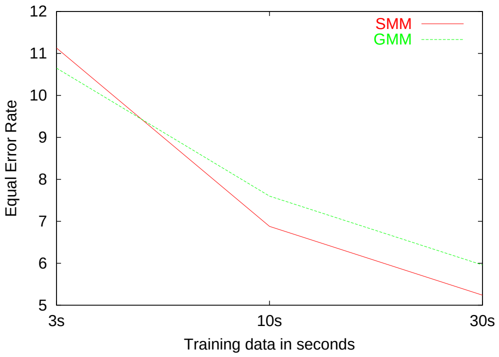


Figure 4: Speaker verification % EER for GMM and SMM against amount of speaker training data in seconds.

Figure 2 shows GMM verification results. A detection error tradeoff (DET) curve is given for each of the three levels of training. It is clear from the figure that error rates drop with increasing training data. The difference between 3 seconds and 10 seconds training is larger than the difference between 10 seconds and 30 seconds.

Figure 3 is similar to Figure 2. Here the verification results are for SMM. The error rates drop with increasing training data, but the rate of improvement is faster than for GMM. For 3 seconds of training data the GMM has a lower error rate than the SMM but at 10 seconds the SMM error rates are lower. This is more clearly seen in Figure 4 which plots the equal error rates for GMM and SMM as taken from the DET curves.

5 CONCLUSIONS

The SMM offers a variable step away from the atomic level. The greater the step the more complex the model becomes. The additional complexity, which is in the form of time sequence information, takes advantage of large amounts of speaker specific data when they are available. This is ultimately shown in improved speaker recognition results.

REFERENCES

[1] R. Stapert and J. S. Mason. A Segmental Mixture Model for Speaker Recognition. In *Proc. Eurospeech*, volume 4, pages 2509–2512, 2001.

[2] 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.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41, 2000.

[4] S. Furui. Comparison of speaker recognition methods using static features and dynamic features. *IEEE Trans. on ASSP*, 29:342–350, 1981.

[5] K. Yu, J. Mason, and J. Oglesby. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *Proc IEE vision, image and signal processing*, 142:313–318, 1995.

[6] K. Yu. Text Dependency and Adaptation in Training Speaker Recognition Systems. *Ph.D. Thesis, University College Swansea*, 1997.

[7] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang. A vector quantization approach to speaker recognition. In *Proc. ICASSP*, volume 1, pages 387 – 390, March 1985.

[8] G. Doddington. Speaker recognition based on idiolectal differences between speakers. In *Proc. Eurospeech*, volume 4, pages 2521–2524, 2001.

[9] R. Stapert. A segmental mixture model: maximising data usage with time sequence information. *PhD Thesis, University of Wales Swansea*, March 2001.

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

[11] A. L. Higgins, L. G. Bahler, and J. E. Porter. Voice Identification Using Nearest-Neighbor Distance Measure. *IEEE*, 2:375, 1995.

[12] R. Stapert and J.S. Mason. Speaker recognition and the acoustic speech space. In *Odyssey Speaker Recognition Workshop, Crete*, pages 195 – 199, June 2001.

[13] 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.