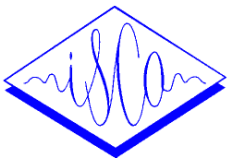


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

PARAMETER REDUCTION IN A TEXT-INDEPENDENT SPEAKER VERIFICATION SYSTEM

Roland Auckenthaler[1,2], *Michael Carey*[1], *John Mason*[2]

[1]Ensigma Technologies, Turing House, Station Road, Chepstow, NP16 5PB, UK.
[2]Department of Electrical & Electronic Engineering, University of Wales Swansea, SA2 8PP, UK

{Roland.Auckenthaler, Michael.Carey}@ensigma.com, J.S.D.Mason@swansea.ac.uk

ABSTRACT

Different methods for reducing parameters in a Gaussian mixture model (GMM) for text-independent speaker verification are investigated in this paper. The number of parameters is directly related to the memory requirement. Reducing the parameters is important in environments with limited memory resources or limited bandwidth for data transmission.

In contrast to standard approaches such as reducing the number of mixture components in the GMM or the dimension of the acoustic space, speaker specific parameters are selected from a global parameter set. Experiments reveal a small performance degradation when only a few parameters are chosen. Reducing the number of parameters to 25% of the original count gives a slightly better performance compared to a four times smaller global parameter set.

1. INTRODUCTION

Verifying the identity of a person becomes important with the emerging market of electronic commerce. In future, transactions will be conducted over mobile phones more often. The voice is the natural interface when using a phone and hence verification by voice would be an obvious choice.

There are two general strategies for implementing verification systems. In the first the verification system is distributed throughout the network. This implies that either speech or speaker characterising models have to be transferred across the network. The transmission of large amounts of data can take a long time given the low data bandwidth of today's mobile networks. Second, the verification system can be directly integrated in a mobile handset. This implies a better quality of speech input because data compression and transmission are avoided. However the implementation has to cope with limited resources of memory and processing due to the small sizes of the mobile devices.

In both cases the memory requirement is important. For high performance speaker verification systems, it is known that speaker models tend to use a large amount of memory. In [1] approaches to reduce the memory requirement for text-dependent speaker verification is discussed. In contrast, this work considers the requirements of a text-independent verification system using Gaussian mixture models [2].

A state-of-the-art GMM system uses speaker adaptation from a world or background model. Such a system obtains good performance for limited speaker training data. The world model components consist of mean and variance parameters and a component weight. Adapting the speaker model is performed on the means only. The variance and weight parameters remain identical to the world model. This is shown to achieve the best verification performance [2]. Further, the adaptation process also allows a fast scoring of the speaker model in testing, due to re-scoring the best mixture components only.

The memory requirement of GMMs is directly related to the number of parameters in the model. The goal is to reduce the requirement of a speaker model, by quantisation and/or parameter reduction, while minimising degradation in verification performance. Experiments on varying parameter quantisation have shown that with careful scaling 6 to 8 bits are sufficient to store each parameter. Further reduction in memory is possible by reducing the number of parameters.

The number of mean parameters in a speaker's GMM is given by the product of the feature vector dimension and the number of mixture components in the model. A brute force approach to reduce the number of parameters is to decrease one of these factors. This is achieved by lowering the dimension of the acoustic space or fewer components to describe the speech patterns. Both imply a degradation in performance. The approaches in this paper do not reduce the feature dimension nor the number of components but try to select important parameters from the set to define the speaker model.

The remainder of the paper is organised as follows: Section 2 describes the parameter selection methods and Section 3 the system outline and experiments. Section 4 concludes the experimental results.

2. PARAMETER SELECTION

The speaker model is a two dimensional array given by the features and the number of mixture components. From this mean parameter set only the important parameters are retained to create the speaker model. This model is also called a parameter sub-set of the global set.

Three different approaches are considered for the experiments to determine the retained parameters. The first picks important parameters from the array using a statistical significance test. The second approach selects a set of important mixture components. In the third, a fixed number of important parameters are selected within each component. This selection is performed individually for each component in the model.

2.1 Global Sub-set - GS

Selecting a parameter sub-set is based upon a ranking of importance of each model parameter. Therefore a measure of importance is introduced. Given that only the mean is adapted from the background model, a student-t test can be used for measuring the significance of adaptation. The t-value is given by

$$t = \frac{\mu_S - \mu_W}{\sigma_W} \qquad (1)$$

where μ_S is the speaker's mean and μ_W the world or background mean with the corresponding variance σ_W. The statistical significance is obtained by applying the incomplete beta function described in [3]. The degrees of freedom for this test are related to the number of independent samples. The problem of calculating the significance is the exact evaluation of the degrees of freedom. The number of frames for the adaptation provides only an indication for the degrees of freedom due to the inter-frame correlation in the training speech.

A significance value is calculated for each model parameter. The obtained values are sorted and only the most significant model parameters are retained for the speaker model. The indices to the position in the two dimensions are stored for each individual parameter. The large indexing overhead is a disadvantage of this method.

Assuming a parameter quantisation of eight bits and sixteen bits for indexing, the overhead for storing the indices is 200%. Therefore selecting a sub-set of most significant parameters must at least offset the memory usage for indexing, i.e. one in three retained.

The next two approaches attempt to overcome the indexing overhead. They reduce the overhead by indexing in just one of the two dimensions of the parameter array.

2.2 Mixture Component Sub-set - MCS

The first approach of indexing in one dimension is to find a sub-set of important mixture components. All parameters of a chosen component are stored. The important components may include some less significant parameters. Here, the advantage over the GS approach is that only one index for the chosen component is stored rather than one for each parameter. This leads to a very small amount of indexing overhead.

The sub-set of components is selected according to their occurrence frequencies in the speaker training data. The occurrence frequencies are established by the highest scoring components of the world model during training.

Only the most frequent occurring components are kept for the speaker model. These components encompass the best trained speaker parameters due to the larger number of training samples. The occurrence frequencies are seen as an indication of importance and are not based on a statistical significance test.

2.3 Component Parameter Sub-set - CPS

The third approach selects a number of parameters within each mixture component. For this purpose the parameters of a component are divided into blocks of two or four parameters. For each block the single most significant parameter is chosen. The significance ranking of the parameters within a block is determined by the t value of student-t test, equation (1). This ranking is valid as there are similar degrees of freedom within a component. The t-value is directly related to the significance of the parameter. Therefore, the CPS selection overcomes the problem of estimating the degrees of freedom for statistical testing.

Given the division into blocks of two or four parameters, the index overhead is reduced to one or two bits for each chosen parameter respectively. In the case of selecting one out of four parameters and a quantisation of eight bits, the indexing overhead is only 25%. As mentioned earlier, the quantisation for a model parameter can be reduced to 6 bits without major performance degradation. This would lead to an eight bit representation for each model parameter, including two bits indexing overhead.

3. EXPERIMENTS

The NIST 1998 evaluation data are used for the experiments. The evaluation conditions are two sessions training, one minute of speech from each session, and ten second segments in testing. All results are plotted for trials when the same handset is used in training and testing.

The speaker verification system uses a filter-bank front-end to extract 39 features at a frame rate of 20ms. Features are extracted from a filter-bank with twelve mel-spaced band-pass filters and the absolute energy of the frame. The first and second order derivatives are calculated for the thirteen static coefficients across five and seven frames respectively. Channel normalisation is applied using cepstral mean subtraction.

The GMM system consists of a world model with 256 mixture components trained on eight hours of speech of telephone quality. The speaker models are adapted from the world model using mean only adaptation. The speaker scores are normalised by the world model score during testing. Speaker model parameters are selected according to the three approaches. Each parameter is quantised to eight bits.

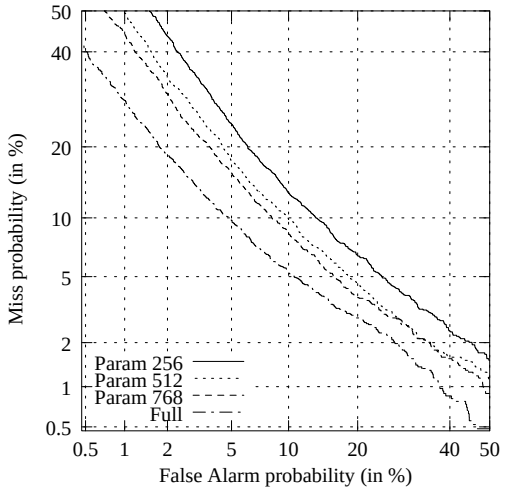
In the case of GS selection, 256, 512 or 768 parameters are selected from the global set of 9984 parameters given by 256 mixture components with 39 coefficients. For the MCS approach, sets of 20, 40 or 60 components are selected to form the speaker model. The CPS selection is based on 1 out of 2 (1/2), one out of four (1/4) or two out of four (2/4) parameter block selections.

A legitimate comparison across different methods is permitted when 768 parameters for GS, 60 components for MCS or one out of four for CPS are used to reduce the number of model parameters. In these cases the required memory for the speaker models is around 2500 bytes.

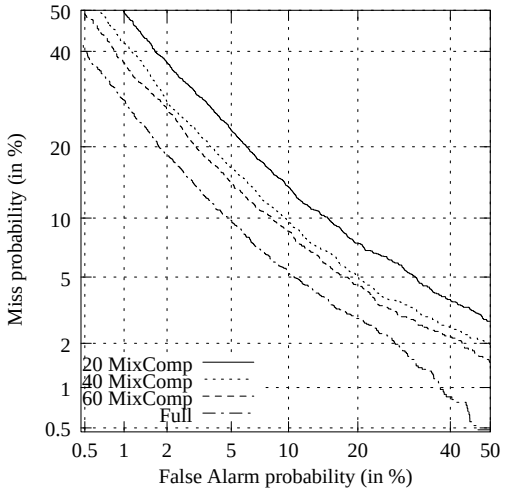
Figure 1 shows the verification performances for the different parameter selection strategies. In Figure 1a the results for the GS approach reveal an equal error rate, EER, of around 12% when 256 parameters are selected for the speaker model. Increasing to 512 parameters reduces the EER to 10%. The verification performance improves only slightly when another 256 parameters are added to the speaker model. Using all parameters for the speaker model obtains an EER of 7.1%.

A similar scenario can be seen for the MCS approach in Figure 1b. Using more components improves the verification performance. Compared to Figure 1a, the MCS approach reveals a slightly better performance for low false alarm rates. For areas with low miss probabilities the GS approach is superior.

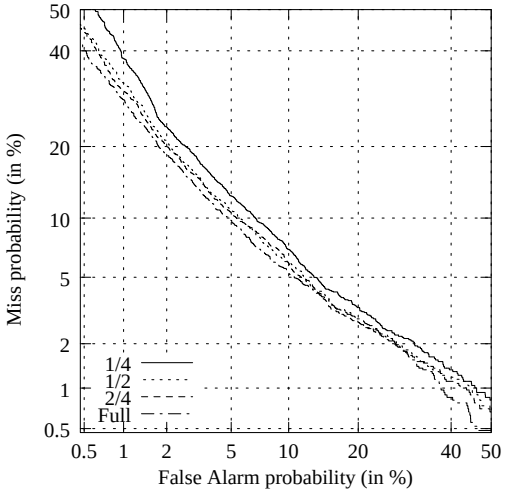
The results for the CPS strategy can be seen in Figure 1c. Again a similar picture is seen for the one out of four and one out of two selection settings. The two out of four setting leads to the same performance as the one out of two setting. It shows that using smaller block sizes does not lead to a loss of important parameters when all parameters in a block are very important. The main reason for this might be the de-correlation of vector parameters due to the cepstral representation [4].



a) GS Selection



b) MCS Selection



c) CPS Selection

Figure 1: Verification Performance for Different Parameter Selection Strategies

The performance degradation due to parameter selection is quite high for all three approaches but given the number of selected parameters the degradation is surprisingly low. In the case of GS selection, the EER increases from 7.1% to 12%. On the other side the memory requirement for the speaker model shrinks from 9984 bytes to about 768 bytes.

Figure 2 compares the three different approaches to a full GMM model with 64 components. All four approaches use about the same amount of memory to store the speaker model.

It can be seen that the GS approach reveals a similar behaviour to the smaller global model size with 64 mixture components. The figure also shows the clear cross-over between the MCS and the GS approach. From this it seems that significant parameters are important for the impostor rejections while the use of a whole feature vector for scoring improves the target rejection.

The CPS approach obtains the best performance of the three approaches. It achieves a compromise between the two other approaches whilst retaining the performance advantages of both. The CPS approach also outperforms the smaller global model, particularly at low false alarm rates.

The comparison between the CPS and a smaller global model is repeated for a larger model size. A system with 1024 mixture components and CPS is compared to a smaller global model with 256 components. The results show similar DET performances for both cases. This leads to the conclusion that the parameter selection is superior only for small global model sizes.

4. CONCLUSIONS

This paper discusses several approaches of parameter reductions in a text-independent speaker verification system. This is important when a speaker verification system is implemented in an environment with limited memory resources. Examples for these are mobile telephones or data networks with low bandwidth.

Contrary to standard approaches which reduce the feature space or the number of mixture components in a speaker model, a sub-set of important parameters is selected from a global set. Three different approaches are investigated.

The results reveal that all selection methods lead to similar performance degradations. Only a small number of parameters contribute to the speaker discrimination. When a sub-set of 256 parameters is chosen from 9984 parameters the EER increases from 7.1% to only 12%. This leads to a memory reduction from 9984 bytes to 768 bytes.

The CPS selection obtains a slightly better performance

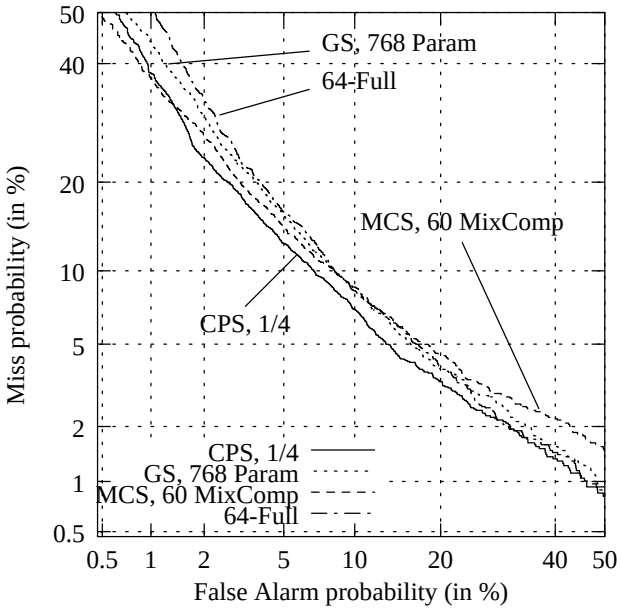


Figure 2: Comparison Between Different Selection Methods and a Smaller Global Model

compared to a smaller global model size. This is true for a global model size of 256 mixture components.

Future work on the parameter selection might include different normalisation techniques. It is of interest how specific parameters can be selected using cohort speakers. It may lead to new, selective, normalisation techniques due to the usage of knowledge from the important parameters.

5. REFERENCES

[1] Schalkwyk and N. Jain and E. Barnard, *Speaker Verification with Low Storage Requirements*, in Proceedings ICASSP '96, vol. 2, pages 693-696, Atlanta, GA, 1996

[2] D. Reynolds and T. Quatieri, *Speaker Verification Using Adapted Gaussian Mixture Models*, in Digital Signal Processing A Review Journal, vol. 10, no. 1-3, pages19-41, Academic Press, 2000

[3] W. Press, S. Teukolsky, W. Vetterling and B. Flannery, *Numerical Recipes in C*, Second Edition, Cambridge University Press, 1992

[4] S. Furui, *Cepstral Analysis Technique for Automatic Speaker Verification*, in IEEE Trans. Acoustics, Speech and Signal Processing, vol. 29, no. 2, pages 254-272, 1981