LANGUAGE IDENTIFICATION USING EFFICIENT GAUSSIAN MIXTURE MODEL ANALYSIS

E. Wong, J. Pelecanos, S. Myers and S. Sridharan
Speech Research Lab, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
ee.wong@qut.edu.au, j.pelecanos@qut.edu.au
sd.myers@qut.edu.au, s.sridharan@qut.edu.au

ABSTRACT: Automatic Language Identification (LID) is the automated process of identifying the language of a speech utterance. In this paper, we will describe a language identification system that utilises Mel-Frequency Cepstral Coefficients (MFCCs) and Gaussian mixture models (GMMs) to model the short-term characteristics of a language. We also compare this standard GMM language model to the models that are adapted from a universal, language-independent background model (UBM). Experiments show that model adaptation gave comparable performance. In addition, a computation speed-up approach was tested on the adapted language models. The accuracy of the system remained comparable while the computation time was reduced significantly.

1. INTRODUCTION

Automatic Language Identification (LID) is the process of identifying the language of a speech utterance using a computer. There are several important applications for Language Identification (Muthusamy et al., 1994). For example, telephone companies can handle foreign language calls with a LID system that routes each call to the operator that is fluent in the caller's language. This application can even extend to the handling of emergency call services. A LID system can also serve as a front-end for a multi-language translation system.

To accomplish the task of Language Identification, a variety of methods have been proposed (Muthusamy et al., 1994). These include Hidden Markov models (HMMs), expert systems, clustering algorithms, quadratic classifiers, and artificial neural networks. Our system uses the Gaussian mixture modelling (GMM) (i.e. a single state HMM) approach. This system operates in 2 phases: training and recognition. During the training phase, the system takes the speech utterances for a single language and converts them into feature vectors. A GMM is trained on the feature vectors for each language. During recognition, an unknown utterance is compared to each of the GMMs. The likelihood that the unknown utterance was spoken in the same language as the speech used to train each model is computed, and the most likely model is determined as the hypothesised language.

The GMM LID approach performs classifications using information from single observations while an LID system using HMMs has the ability to model sequential events of speech (Zissman, 1993). However, Zissman has reported that the performance of GMMs was comparable to that of HMMs and this is one of the reasons that we utilise GMMs. Note that with post-processing the performance of a phonetically based HMM system can be improved. The main reason that motivated us to utilise GMMs with Universal Background Modelling (UBM) was that this technique was successfully applied to speaker verification in a highly computationlly efficient manner (Reynolds, 1997). This paper begins with an overview of our basic GMM LID system. The approach for creating models by adapting the model from a universal, language-independent background model (UBM) is then described, followed by an approach that speeds up classifications during the recognition phase. Finally the results of the experiments are presented.

2. LANGUAGE IDENTIFICATION SYSTEM

2.1 Parameterisation

The feature vectors used for modelling languages comprised of 12 Mel-Frequency Cepstral Coefficients (MFCCs) (Rabiner et al., 1993) derived from 20 filterbanks. Each feature vector is extracted at 10 ms intervals using a 32ms window of bandlimited (300-3400 Hz) speech. Since the experiment involved telephone speech, cepstral mean subtraction was applied to the MFCCs to reduce the linear channel effects. The corresponding delta coefficients were computed over a

window length of 15 frames. Initially a shorter delta coefficient window length was trialled. Preliminary experiments indicated an improvement by extending this window length. A longer delta window length may be able to encapsulate more of the temporal information that is specific to language discrimination, particularly when the GMM does not use information across frames. Finally the delta coefficient of the frame energies (over the same window size) was appended to the features.

2.2 Gaussian Mixture Model (GMM) Classification

The GMM approach attempts to model the probability density function of a feature vector, $\vec{x}$, by the weighted combination of multi-variate Gaussian densities:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

with

$$b_i = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_i)} \tag{2}$$

where λ is the model described by

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \tag{3}$$

In equation 1, i is the mixture index ($1 \le i \le M$), p_i is the mixture weight such that $\sum_{i=1}^{M} p_i = 1$, and $b_i(\vec{x})$ is a multi-variate Gaussian distribution defined by the corresponding means $\vec{\mu}_i$ and diagonal covariance matrices, Σ_i.

The estimation of the GMM parameters is accomplished by an iterative process, termed the Expectation-Maximisation (EM) (Reynolds et al., 1995). For more rapid GMM convergence, the mixture means, weights and variances are seeded by statistics determined by a K-means (Schalkoff, 1989) vector quantisation estimate of the feature vectors (Pelecanos et al., 2000).

During recognition, an unknown speech utterance, X, comprising of observations $\vec{x}_1, \vec{x}_2, ..., \vec{x}_T$, is classified by first calculating the average log likelihood that the language model produced the unknown speech utterance. This is given as

$$p(X \mid \lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda) \tag{4}$$

where λ is the model for the corresponding language. The maximum-likelihood classifier hypothesis, H can be calculated as

$$H = \arg\max_{l=1}^{L} p(X \mid \lambda_l) \tag{5}$$

where the language index l = 1, 2, ..., L for L languages. Figure 1 shows the block diagram of the two phases of the LID system.

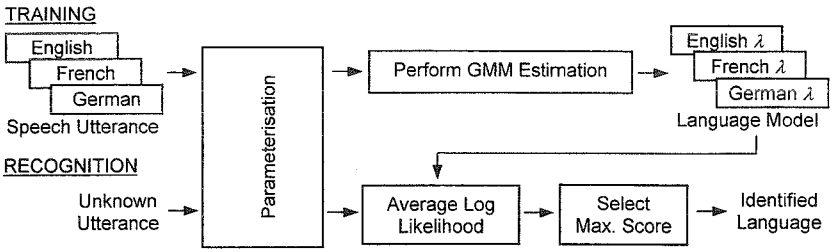

Figure 1. Block diagram of the LID system.

3. ADAPTATION USING UNIVERSAL BACKGROUND MODELS

We wish to investigate the method of Universal Background Modelling (UBM) as it applies to LID. This is a successful technique that has been applied previously to speaker verification (Reynolds, 1997). The application of the UBM approach to the LID problem has not been investigated in the past. The idea (in terms of speaker verification) is that instead of incorporating a group of background speakers for each target speaker, a universal, speaker-independent background model is used. After the creation of a UBM, the speaker model is formed by employing Bayesian adaptation (Gauvain et al., 1994) to train the speaker model given the prior information provided by the UBM. This approach is generally applied to situations where there is insufficient training data. There are other speed advantages associated with this approach that will be discussed later. A necessity of the UBM is that there must be sufficient speech in order to cover the general acoustic features of all speakers (speaker verification) or all languages (LID) and not be overly turned to any particular class.

This Universal Background Model technique can be applied to LID. Firstly, a universal, language-independent background model is created using a portion of the training data from all languages. Then, by using Bayesian adaptation, all language models are trained by adapting from the UBM. With all language models obtained, the recognition is performed the same as the standard LID system. An advantage of employing UBMs in the LID system is that the quantity of training data required can be reduced. This is important when a new language is added to the system, as collecting enough data suitable for training a language can be difficult. We will also investigate the possibility of substituting a standard GMM from a single language for the UBM and examining the performance variation. This enables us to avoid retraining the UBM when further languages are added to the system. Another benefit of the UBM includes a drastically reduced model training time due to models being adapted and not fully retrained.

4. COMPUTATION SPEED-UP APPROACH

To accurately model the acoustic characteristics of a language, a relatively large number of mixtures is recommended to create the GMM. During the training phase, the computation involved is acceptable due to the fact that no real time processing is required. However, the time requirement for performing recognition may not be suitable for a real world application. Therefore, computational speed-ups of the LID system are mandatory and this is one of the main benefits of employing UBMs. The idea of this computation speed-up is based on the fact that each language model is adapted from the UBM. Thus the language model and the UBM are sharing a partial correspondence of their mixtures and this idea was successfully employed for speaker verification optimisations (McLaughlin et al., 1999).

The actual implementation of this idea is very simple. For each test feature vector, all the UBM mixtures are scored to determine the top 5 highest scoring mixtures. Using the property that each language model is adapted from the UBM, the calculation of the language model likelihood only requires the testing of the 5 mixtures that correspond to the top 5 mixtures from the UBM (McLaughlin et al., 1999). By employing this approach to the LID system, the computation speed was rapidly improved. This can be shown as follows.

Given that both the GMM and UBM have N mixtures, we choose to test the top C mixtures for L languages. The number of mixture tested, R, is:

$$R = N + C \times L \tag{6}$$

Alternatively, for the standard GMM system with all mixtures tested, the number of mixture tests will be

$$R = N \times L \tag{7}$$

In our case, we tested 10 languages using a 512 mixture GMM and determined the top 5 mixtures from the adapted models. This gives only R = (512 + 5 × 10) = 562 mixture tests compared to R = (512 × 10) = 5120 mixture tests for standard GMM system. This gives us a 900% computation improvement.

One of the pitfalls of this method is the possible degradation of accuracy. However McLaughlin (1999) has indicated that the sacrifice in accuracy by using this computation speed-up is small. This has motivated us to apply this approach to the LID system.

5. EXPERIMENT AND RESULTS

The Language Identification experiment was trialled using the 10 language version of the Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus (Muthsamy et al., 1992). This corpus contains 90 speech messages in each of the following languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The 1994 National Institute of Standards and Technology (NIST, 2000) specification was used as a guideline for extracting the training and testing data to perform the experiment. Worthy of note is the partitioning of (approx.) 10 second speech segments for testing purposes. There were 560 test segments with 4 experiments trialled: the standard GMM system, the GMM-UBM system, the GMM-UBM system with the top 5 mixtures being tested and the use of a standard English GMM in place of the UBM. We also analysed the variation in performance due to the number of significant mixtures selected for UBM testing.

Table 1 shows all the experimental results given as the percentage of utterances correctly identified. The performance of the standard GMM system is 56.6% and this compares favourably to 50.0% of accuracy obtained by a similar system presented by Zissman (1996). The results of the UBM-GMM system and the equivalent system employing the speed up approach have shown that these systems can obtain a comparable performance to the standard GMM system. We also found that using a few of the more significant mixtures for testing resulted in little to no measurable degradation in performance. The use of this mixture selection process results in significant savings in testing time. In addition, by replacing the UBM with a standard language model, (ie. An English model) the system could be made to accommodate further language models without retraining the UBM. In this configuration, the use of the English language model for the UBM reduced the performance slightly. This is a viable approach of building a LID system under the situation of lack of training data or when flexibility of adding new language models is required.

	% correct
Standard GMM	56.6
GMM-UBM	53.2
GMM-UBM with Top 5 mixture test	53.4
Replace UBM by standard English GMM	51.6
Replace UBM by standard English GMM with Top 5 mixture test	51.9

Table 1. Experimental results of all LID system configurations.

Another experiment examined the varying performance of the adapted LID system when the number of significant UBM mixtures tested was altered. The results are shown in Figure 2 and it indicates that the system obtained almost constant accuracy with a varying number of significant mixtures tested. Note that the slight increase of 0.04% in accuracy for the smaller number of mixtures is not a significant improvement.

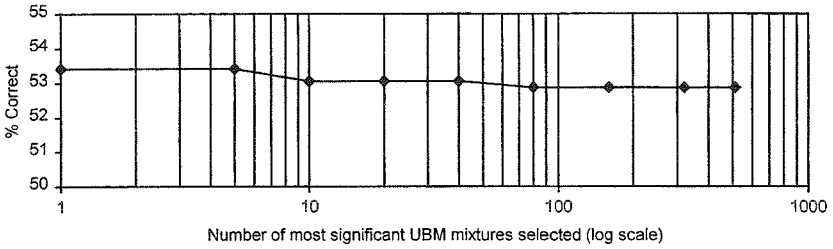

Figure 2. LID performance variation with respect to the number of significant mixtures selected.

6. CONCLUSIONS

This paper discussed the use of GMM-UBMs for language modelling as a speed enhanced alternative to the standard GMM system. This approach has the additional benefit that less training data is required for a similar performance. Also, the model training time is decreased due to the use of the adaptation procedure. In testing, by employing the computation speed-up method of the GMM-UBM system, the efficiency of testing was improved dramatically while accuracy remained comparable. The possibility of replacing the UBM by an English or language specific model would allow the GMM-UBM system to be efficiently extended when additional languages are to be included.

7. ACKNOWLEDGEMENTS

This work is sponsored with a research contract from the Australian Defence Science and Technology Organisation (DSTO).

8. REFERENCES

Gauvain J. & Lee C. (1994) Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains, Transactions on Speech and Audio Processing, vol 2, pp. 291-298.

McLaughlin J, Reynolds D.A. & Gleason T. (1999) A Study of Computation Speed-Ups of the GMM-UBM Speaker Recognition System, EuroSpeech, vol 3, pp. 1215-1218.

Muthusamy Y.K., Cole R.A. & Oshika B.T. (1992) The OGI multi-language telephone speech corpus, In ICSLP Proceedings, vol 2, pp.895-898.

Muthusamy Y. K., Barnard E. & Cole R. A.(1994) Reviewing automatic language identification, IEEE Signal Processing Magazine, vol 11, No 4, pp. 33-41.

NIST (2000) Spoken Natural Language Processing Group, Web page: http://www.nist.gov/speech/.

Pelecanos J., Myers S., Sridharan S. & Chandran V. (2000) Vector Quantization based Gaussian Modelling for Speaker Verification, International Conference on Pattern Recognition, paper number 1219.

Rabiner L. & Juang B.H. (1993) Fundamentals of Speech Recognition, Prentice-Hall, New Jersey.

Reynolds, D. A. & Rose R. C. (1995) Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83.

Reynolds, D. A. (1997) Comparison of Background Normalization Methods for Text-Independent Speaker Verification, EuroSpeech, vol. 2, pp. 963-966.

Schalkoff R. (1989) Pattern Recognition, New York: John Wiley & Sons.

Zissman M. A. (1993) Automatic language identification using Gaussian mixture and hidden Markov models, International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 399-402.

Zissman M. A. (1996) Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, IEEE Transactions on Speech and Audio Processing, vol. 4, no. 1, pp. 31-44.