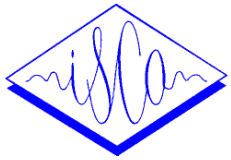

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

METHODS TO IMPROVE GAUSSIAN MIXTURE MODEL BASED LANGUAGE IDENTIFICATION SYSTEM

Eddie Wong and Sridha Sridharan

Speech Research Lab, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
{ee.wong, s.sridharan}@qut.edu.au

ABSTRACT

This paper investigates the use of Vocal Tract Length Normalisation (VTLN) and Output Score Fusion techniques to improve the performance of a Gaussian Mixture Model (GMM) based Language Identification (LID) system. The Universal Background Model (UBM) technique, which has been successfully employed in Speaker Verification, is incorporated into the GMM LID system to reduce the time requirement for both training and testing. The paper also presents a fast approach for selecting the normalisation factor for VTLN during the testing stage of LID which is based on the UBM technique. The output scores generated by the GMM system have been fused with a phonetic based LID system to improve the overall scores. Experimental results show that a reduction in the relative error rate by over 50% is possible for the 45-second test case in the NIST 1994 Evaluation data.

1. INTRODUCTION

Gaussian Mixture Model based Language Identification has the advantage of simplicity with no requirement of phonetic transcription of training data. However, recently focus has moved away from this method due to its poor performance compared to other LID techniques such as phonetic based system [1, 2]. We show in this paper that by incorporating Universal Background Model (UBM) [3], Vocal Tract Length Normalisation and Output Score Fusion techniques, the performance of GMM based LID can be improved significantly.

By using the UBM technique the time requirements for both training and testing with a set of GMMs can be significantly reduced. Due to the practical implementation considerations, in previous studies of GMM based LID systems the number of mixture components were restricted to low values (<100). When UBM techniques are applied to the GMM based LID system (GMM-UBM) [4], the number of mixtures of a GMM as well as the dimension of the feature vectors can be increased without the penalty of increased computations. The increase in the number of mixtures enables the characteristics of each language to be modelled more accurately and thus increase the discrimination between different languages.

The task of LID offers several challenges:

- The identification should be speaker independent
- The identification should be made with a test utterance from each speaker of less than one minute in duration.
- Most LID applications require a fast response time.

Speaker independence requires some kind of speaker normalization. Because of the limitation introduced by short test segments and the requirement for fast response by the system, the use of speaker adaptation technique such as MLLR adaptation is not feasible. Therefore, the use of Vocal Tract Length Normalisation (VTLN) is the most suitable approach for LID. In this paper, we also propose a novel method that can perform VTLN efficiently during testing by taking advantage of the UBM technique.

Previously we had studied the fusion of output LID score generated by different parameterisation methods [5]. The results showed that combining different sources of information could improve LID performance. In this paper, further investigation is performed on combining the GMM-UBM system with a phonetic base LID system which has been referred to in [1] as "Phoneme Recognition followed by Language Modelling performed in Parallel" (PRLM-P). This combined system incorporates two different information sources of the speech data. Firstly is the static acoustic features provided by the GMM-UBM system, and secondly is the phonetic information generated by the PRLM-P LID system.

The paper is organised as follows. An overview of the GMM-UBM system utilising the UBM technique and the details of the experiments are described in Section 2. The VTLN application in LID is presented at Section 3 followed by the experiment of output score fusion in Section 4. The conclusion is given in Section 5.

2. GMM-UBM LID SYSTEM

Figure 1 shows a block diagram of the GMM-UBM LID system. The characteristic of each target language is modelled by a GMM using the static acoustic information. The training with GMMs can be very time consuming especially for GMMs with a large number of mixture components as large amounts of speech data must be processed. To overcome this problem we have applied the technique of Universal Background Model to a GMM LID system [4] in order to alleviate this problem.

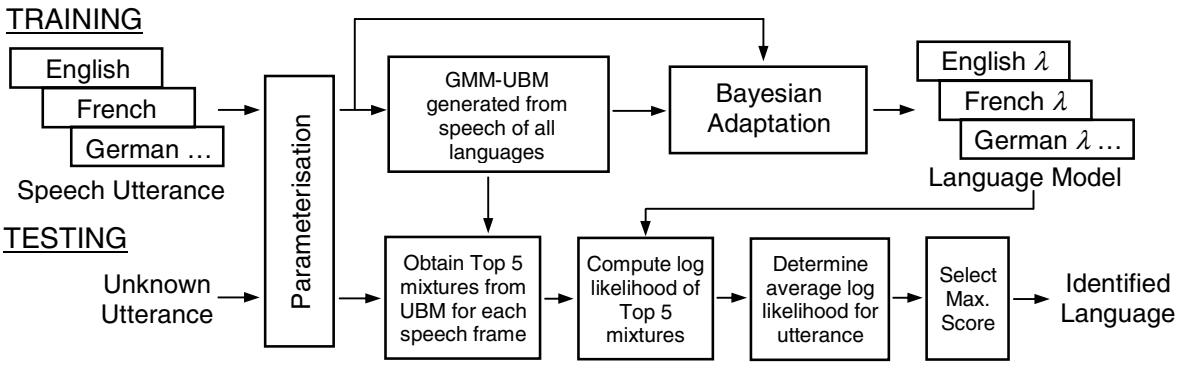

Figure 1: Block diagram of the GMM-UBM LID system.

2.1. The Universal Background Model Technique

A Universal Background Model (UBM) in the LID case is a GMM representing the characteristics of all different languages. Instead of performing the Maximum-Likelihood training, the models of each language can be created by employing Bayesian adaptation from the UBM using the language-specific training speech. Therefore, significant time will be saved in the training of each language model. This adaptation also allowed the time requirement for LID testing to be reduced significantly where will be described later. In addition, with this form of training, any test observations not seen by the models would typically not discriminate on the bias of any particular LID models. The upper part of Figure 1 shows this training operation.

From previous experiments conducted for speaker verification, Reynolds [3] has found that only a few of the mixtures of a GMM contribute significantly to the likelihood score for a speech feature vector. In addition, the mixture components of the adapted model of each language share a certain correspondence with the UBM (since each model is adapted from it). Therefore, the average log-likelihood score of the language models can be calculated by scoring only the more significant mixtures (the top 5 mixtures [3] for this system). According to the correspondence of mixtures between the UBM and the model of a language, these significant mixtures can be obtained by selecting the mixtures from the UBM that have the highest score. By employing this mixture testing strategy, the computation required for testing will also be reduced significantly. This approach is shown at the lower part of Figure 1.

2.2. The LID Experiment

The experiment was performed using the Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus [6] which included the following 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The 1994 National Institute of Standards and Technology (NIST) LID evaluation specification was used as a guideline for extracting the training and testing data to perform the experiment. Both the training and extended data from the corpus (678 male and 268 female speakers) were used for creating the UBM and adapting language models. The 45 second (187 test segments) and 10 second (625 test segments) utterances were compiled for testing. PLP cepstral coefficients (5th order) were used in this experiment. Each feature vector was extracted at 10ms intervals using a 32ms window. Delta energy, PLP delta and PLP acceleration coefficients were appended to the feature vectors with PLP mean and variance normalisation applied. All models consisted of GMMs using 512 mixtures. The LID result of this baseline system is shown at the first row (PLP Baseline) of Table 1.

3. VTLN FOR LANGUAGE IDENTIFICATION

VTLN has been widely used in speech recognition, especially for large vocabulary speaker independent continuous speech recognition [7, 8]. The basic idea behind VTLN is that each person has a different vocal tract length and this length has an inverse relationship to the formant frequencies [9]. Therefore, speaker normalisation may be performed by re-scaling the frequency axis (compress or stretch the spectrum) according to the length of the vocal tract or a corresponding normalisation factor. The task of VTLN, therefore, is the process of obtaining the normalisation factor, α, for each speaker. The selection process is repeated until there are no more significant changes in α for all speakers. This approach does not select an α that is directly related to the physical vocal tract length of each speaker. However, it does guarantee that the normalised version of the speech will provide equal or better likelihood scoring against the given model(s).

3.1. Integration of VTLN

We have adapted the piecewise linear transformation method by Wegmann [9] to normalise the speech data. The α selection process for each speaker from the training data is to choose the α value that maximises the likelihood score against a GMM, λ_{GMM}, as

$$\hat{\alpha} = \arg\max_{\alpha} p(\mathbf{X}_{\alpha} \mid \lambda_{GMM}) \qquad (2)$$

where there are 13 different normalisation factors ($0.88 \le \alpha \le 1.12$, and stepping by 0.02) selected for each speaker. The entire iterative selection process is as follows:

i - iteration	45s	10s
PLP Baseline	23.0	34.4
PLP_VTLN (i = 1)	16.6	29.4
PLP_VTLN (i = 2)	16.0	28.3
PLP_VTLN (i = 3)	14.4	27.7
PLP_VTLN (i = 4)	13.4	27.5
PLP_VTLN (i = 5)	13.4	27.0
PLP_VTLN α-model	14.4	27.0

Table 1. Test results (% error) for applying VTLN to the baseline LID system compared with test results of PLP_VTLN α-model are after 5 iterations.

1. Train λ_{GMM} with data from all languages and select α for each speaker using Equation (2).
2. Re-train the λ_{GMM} with newly selected normalised data.
3. Select α using Equation (2) for each speaker against the new λ_{GMM} from step 2.
4. Repeat step 2 and 3 until there are no further significant changes in α.

The UBM and models for each language of the LID system are then trained using the normalised data.

VTLN during testing selects the normalisation factor that maximises the likelihood of the normalised λ_{GMM} given the input test speech and then tests it against the normalised set of model.

The LID result after applying VTLN is shown in Table 1 (PLP_VTLN, i = 5). A relative reduction of 42% in error rate was achieved for the 45 seconds test case and 22% for the 10 seconds test. The reason for the smaller improvement for the 10 seconds test case might be due to the lack of data for the α selection process to properly select a correct α value for the input test utterance. This result indicates that speaker normalisation is mandatory when the number of speakers presented in the database is high, and the reduction of speaker variation of the speech data can improve the performance of the system.

3.2. Rapid Normalisation Factor Selection Approach

An alternative approach of α selection during testing has been proposed by Lee [10]. Instead of re-training a new GMM with the final normalised data, the data that are selected with the same α value are grouped together. For each α group, the *unnormalised* version of data is used to train an α-model, λ_α, to represent the characteristic of that normalisation factor. The α selection process during testing thus becomes

$$\hat{\alpha} = \arg\max_{\alpha} p(\mathbf{X} \mid \lambda_\alpha) \tag{3}$$

and the overall selection process is performed with *unnormalised* data only. However, this method requires that enough data is available to train each α-model.

By applying the UBM technique to the aforementioned approach, the time required to perform α selection during testing can be further reduced. Rather than carrying out a full Maximum-Likelihood training to each α-model, UBM is first created and the α-models are then obtained by adaptation. Thus, when certain α-models lack training data, this approach is still feasible. With the UBM and its corresponding adapted models, the top-5 mixture test technique as described in Section 2.1 can be applied during the selection process. With this speed improvement, faster then real-time performance for normalisation factor selection during testing can easily be achieved. The LID result utilising this approach is shown in Table 1 (PLP_VTLN α-model). Although the α selection approach is different between the training and testing stage, this mismatch does not significantly affect the final LID performance.

4. OUTPUT SCORE FUSION

We have previously [5] investigated the usefulness of fusing output scores generated by the GMM-UBM systems utilising different parameterisation methods. The result showed that different parameterisation methods do provide certain information that helps to increase the discriminative ability of the LID system despite the fact that the scores are generated by the same classifier structure and the analysis is restricted to using short-term spectral based features. Thus, a further improvement can be expected when the output scores are generated by a different classifier. We conducted an experiment in which a phonetic information based LID system namely "Phoneme Recognition followed by Language Modelling performed in Parallel" (PRLM-P) [1] is used for fusion with the GMM-UBM system.

As the name implies, the PRLM-P LID system utilises a phoneme recogniser as front-end to decode the speech data into phoneme sequences, and an n-gram language model is used to model these sequences for each language. The parallel part of the name means that a set of phoneme recogniser front-ends that are trained with different languages is running in parallel in order to capture the diverse phonetic sequence information specific to different languages. The PRLM-P system for this experiment has six phoneme recogniser front-ends in the languages of English, German, Hindi, Japanese, Mandarin and Spanish. Each feature vector consisted of 12 Mel-Frequency Cepstral Coefficient (MFCC) with energy, plus their delta and acceleration coefficients (39 dimension). A phoneme is modelled with a Hidden Markov Model (HMM) with 8 Gaussian mixtures and the characteristics of a language are modelled using a bi-gram language model. The fusion method Linear Score Weighting (LSW) is used in this experiment which is defined as

$$Z = \sum_{i=1}^{N} \omega_i Y_i \tag{4}$$

where Z is the final output score, Y_i is the output score generated by classifier *i*, *N* is the number of classifiers and ω_i is the score weighting for classifier output *i* such that

% Error	45s	10s
Development Test		
GMM-UBM_VTLN	17.5	25.4
PRLM-P	16.5	25.5
Fused System	11.3	17.0
Score Weight	ω_{GMM}=0.94	ω_{GMM}=0.88
Evaluation Test		
GMM-UBM_VTLN	13.4	27.0
PRLM-P	16.6	24.8
Fused System	10.2	18.4

Table 2. Test results (% error) of fusing output scores between GMM-UBM (VTLN applied) and PRLM-P system.

$$\sum_{i=1}^{N} \omega_i = 1 \tag{5}$$

LID results before and after fusion are shown in Table 2. The GMM-UBM system has VTLN applied. Note that the development test data (194 45 second segments and 635 10 second segments) are used to obtain the first optimal score weighting by exhaustively searching through all the score weighting combinations. The search is done by stepping through the ω (of Equation 4) from 0.0 to 1.0 with each step incremented by 0.01. In addition, the *posteriori* probabilities (assumed each language is equal probable) are calculated to in place the raw likelihood scores generated by both systems before fusion. As showed by the LID results, further improvement in accuracy has been achieved. The average relative error rate reduction is 31% for 45 seconds test and 29% for the 10 seconds test.

5. CONCLUSIONS

This paper investigated several improvements to a Gaussian Mixture Model base Language Identification system. Experiments are conducted on the 1994 NIST Language Identification evaluation, which is based on the OGI-TS Corpus. The time requirement for training and testing of such a system was significantly reduced by employing the Universal Background Model technique. Due to this reduction, a higher number of mixtures GMM (512 mixtures in this experiment) is allowed to better model the characteristic of a language. Vocal Tract Length Normalisation has been applied to the system to reduce the speaker variation, and a relative 42% reduction in error rate for the 45 seconds test case and 22% for 10 seconds test has been achieved. Fusing the output score generated by the GMM system to a phonetic based LID system (PRLM-P) has made further improvement to the system. An average relative error rate reduction of 31% for 45 seconds test and 29% for the 10 seconds test has been achieved.

Overall LID error rate has been reduced from 23.0% to 10.2% with 45 second test segments and 34.4% to 18.4% for the 10 second test segments. One of the attractive features of this GMM-UBM system is its low computational complexity. It is suitable for operation in real time and no transcriptions of training data are required. This makes the implementation and adaptation to new languages relatively straightforward.

6. ACKNOWLEDGEMENTS

This work is sponsored by a research contract from the Australian Defence Science and Technology Organisation (DSTO). The author would like to thank Jason Pelecanos for valuable advice provided.

7. REFERENCES

[1] Zissman, M. A. and Singer, E., "Automatic Language Identification of Telephone Speech Messages using Phoneme Recognition and N-Gram Modelling," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 305-308, 1994.

[2] Hazen, T. J. and Zue, V. W., "Automatic language identification using a segment-based approach," *Eurospeech*, vol. 2, pp. 1303-1306, 1993.

[3] Reynolds, D. A., "Comparison of background normalization methods for text-independent speaker verification," *Eurospeech*, vol. 2, pp. 963-966, 1997.

[4] Wong, E., Pelecanos, J., Myers, S., and Sridharan, S., "Language identification using efficient Gaussian Mixture Model analysis," *Australian International Conference on Speech Science & Technology*, pp. 78-83, 2000.

[5] Wong, E. and Sridharan, S., "Fusion of Output Scores on Language Identification System," *Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[6] Muthusamy, Y. K., Cole, R. A., and Oshika, B. T., "The OGI multi-language telephone speech corpus," *International Conference on Spoken Language Processing*, vol. 2, pp. 895-898, 1992.

[7] Bacchiani, M., "Automatic Transcription of Voicemail at AT&T," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 25-28, 2001.

[8] Hain, T., Woodland, P. C., Evermann, G., and Povey, D., "New Features in the CU-HTK System for Transcription of Conversational Telephone Speech," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 57-60, 2001.

[9] Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B., "Speaker Normalization on Conversational Telephone Speech," *International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, USA, vol. 1, pp. 339-341, 1996.

[10] Lee, L. and Rose, R., "Speaker Normalization Using Efficient Frequency Warping Procedures," *International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, USA, vol. 1, pp. 353-356, 1996.