Comparison of Linear Prediction Cepstrum Coefficients and Mel-Frequency Cepstrum Coefficients for Language Identification

Eddie Wong and Sridha Sridharan

Speech Research Lab, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology

ABSTRACT

The speech parameterisation methods: Linear Prediction Cepstrum Coefficients and Mel-Frequency Cepstrum Coefficients were compared with regard to language identification accuracy in a Gaussian Mixture Model based language identification system. Ten different languages were used to test against a set of ten second test files. The 12th order Linear Prediction Cepstrum Coefficients with delta and accelerate coefficients resulted the best accuracy of 60.0 percent. This has shown that information obtained from linear prediction analysis has increased the ability of discriminating different languages. It also shows that language identification performance may be increased by encompass temporal information by including delta and acceleration features. Besides the performance of our test system has proved the feasibility of modeling language by a single Gaussian Mixture Model instead of using complex system such as phonetic recogniser followed by language modelling or large vocabulary continuous speech recognition system.

1. INTRODUCTION

Previous research on Language Identification (LID) used Mel-Frequency Cepstrum Coefficients (MFCCs) extensively for parameterisation of speech and reasonably good results were obtained on varied classes of LID systems [2, 4, 12]. We propose the use the use of linear prediction cepstrum coefficients (LPCCs) which is a well known speech parameterisation technique and has been applied to speech recognition successfully.

The language identification system used in our experiment relys on Gaussian Mixture Models (GMMs) to model the characteristics of a language described by the features. This technique was investigated previously in [11] but poor results were obtained. This may be attributed to the features used. GMMs is an approach that widely applied in speaker recognition [8] and resulted in high accuracy. Adapted GMMs have shown a good robustness in speaker verification by providing coherence between target and non-target GMM distributions. The LID system incorporates the adapted Universal Background Model (UBM) technique used in speaker verification to greatly reduce the computation effort in both training and testing while providing robust performance [7].

This paper compares the accuracy of LPCC and MFCC in a language identification system and proves the feasibility of a GMM based LID system. The next section describes these two parameterisation techniques in greater details followed by depiction of the testing system in Section 3. The results from the two parameterisation methods will be presented and discussed in Section 4 and conclusions drawn in Section 5.

2. SPEECH PARAMETERISATION

2.1 Mel-Frequency Cepstrum Coefficients

The motivation for using Mel-Frequency Cepstrum Coefficients was due to the fact that the auditory response of the human ear resolves frequencies non-linearly. The mapping from linear frequency to mel frequency is defined as

$$\text{Mel}(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) \qquad (1)$$

Figure 1 shows this mapping with 20 triangular bandpass filters that are equally spaced along the mel-frequency scale with band-limiting between 300 and 3400 Hz.

The MFCC were computed using the Discrete Cosine Transform [10]

$$\text{MFCC}_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \qquad (2)$$

where N is the number of bandpass filters, m_j is the log bandpass filter output amplitudes.

MFCCs are one of the more popular parameterisation methods used by researchers in the speech technology field. It has the benefit that it is capable of capturing the phonetically important characteristics of speech. Also band-limiting can easily be employed to make it suitable for telephone applications. A small drawback is that MFCCs are more computationally expensive than LPCC due to

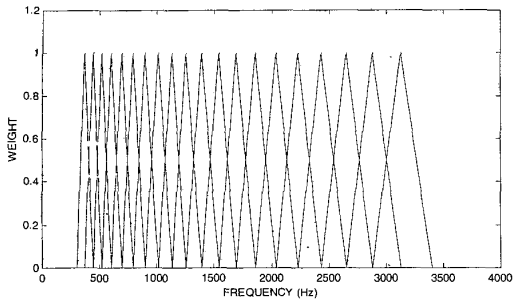

Figure 1. Filters for generating MFCCs with band-limiting between 300 to 3400Hz.

the Fast Fourier Transform (FFT) at the early stages to convert speech from the time to the frequency domain.

2.2 Linear Prediction Cepstrum Coefficients

Linear Prediction Cepstrum Coefficients are Linear Prediction Coefficients (LPC) represented in the cepstrum domain. The idea of LPC is based on the speech production model which the characteristic of the vocal tract can be modeled by an all-pole filter. LPC is simply the coefficients of this all-pole filter and is equivalent to the smoothed envelope of the log spectrum of the speech. LPC can be calculated either by the autocorrelation or covariance methods directly from the windowed portion of speech and the LPCC [3] were acquired from the LPC as

$$\mathrm{LPCC}_i = \mathrm{LPC}_i + \sum_{k=1}^{i-1} \frac{k-i}{i} \mathrm{LPCC}_{i-k} \mathrm{LPC}_k \quad (3)$$

LPCC have been widely used for a few decades and it has been proven that it is more robust and reliable than LPC. However LPCC has also inheriting the disadvantages from LPC. One of the main disadvantages is that LPC approximates speech linearly at all frequencies. This is inconsistent with the perception of human hearing. Also LPC includes the details of the high frequency portion of a speech where containing mostly noise. This inclusion of noise information may affect the system performance.

2.3 Delta Coefficients

It has been proved that system performance may be enhanced by adding time derivatives to the static parameters. The first order derivatives are referred to as delta features and can be calculated as [10]

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (4)$$

where d_t is the delta coefficient at time t, computed in terms of the corresponding static coefficients $c_{t-\theta}$ to $c_{t+\theta}$ and Θ is the size of delta window.

3. TEST SYSTEM

The testing system [9] utilise Gaussian Mixture Model (GMM) to model the characteristic of each target language. 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:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (5)$$

with

$$b_i = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_i)} \quad (6)$$

where λ is the model described by

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad (7)$$

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.

A block diagram of the system is shown at Figure 2. In order to reduce the training and testing time, a novel technique [7] that has successfully applied to speaker verification is adapted in this system.

3.1 Universal Background Model (UBM)

In terms of speaker verification, an UBM is a representation of all the speakers. In the LID case, an UBM is representing the characteristic of all different languages and thus is trained using data from all target languages. Note that these data will be reused at later stage, however dedicated data are prepared for UBM in speaker verification. With this universal language model, the models of each language are created by employ the Bayesian adaptation [1] from this UBM instead of performing a full Expectation-Maximisation (EM) estimation. Therefore an enormous amount of time are saved from the training of each language model.

3.2 Significant Mixture Testing

In the standard GMM approach, all mixtures of a model are examined in order to calculate the likelihood score. However, by use of the properties that only a few of the mixtures of a GMM contribute significantly to the likelihood value [7] and that the model of each language will be sharing a certain correspondence with the UBM since each model is adapted from it. The testing procedure requires scoring all the mixtures of the UBM to determine the top 5 highest scoring mixtures. The likelihood score of the corresponding language GMM is obtained by testing the 5 mixtures of the model that corresponded to the top 5 scoring mixtures of the UBM. Thus the

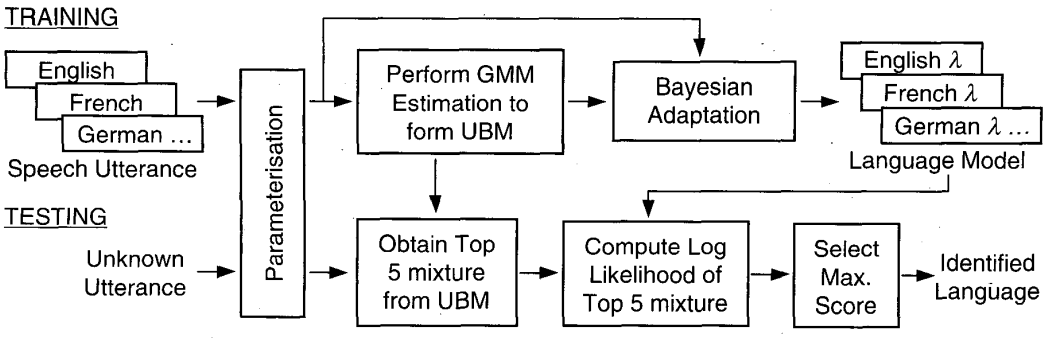

Figure 2. Block diagram of the LID system.

computation required for testing is reduced significantly.

Given that both the GMM and UBM have N mixtures and the top C mixtures are tested for L languages. The number of mixtures tested, R, is

$$R = N + C \times L \tag{8}$$

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

$$R = N \times L \tag{9}$$

In our case (10 languages, 512 mixture GMMs with top 5 mixtures tests), only 562 mixture tests are required compared to 5120 mixture tests for the standard GMM system. A 900% computational improvement is obtained. For more languages, the speed benefits become apparent.

4. EXPERIMENTS AND RESULTS

The Language Identification experiment was trialled using the 10 language version of the Oregon Graduate institute Telephone Speech (OGI-TS) Corpus [5] which included the following languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The 1994 National Institute of Standards and Technology [6] (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 were used for creating the UBM and adapted language models. There were 560 test segments with a duration of 10 seconds.

Each feature vector is extracted at 10ms intervals using a 32ms window and Cepstral mean subtraction was applied for reducing linear channel effects. These settings were used in all experiments reported in this paper. Experiments were trailled in an attempt to determine the characteristics and performance of MFCCs and LPCCs for language identification. The optimal configuration for both parameterisation methods was sought. Different features were added to the standard feature vectors in the experiment to comparing the effect on the accuracy.

The experiment used 12th order MFCCs with 20 filterbanks and 12th order LPCCs using 14 LPCs. The results are shown in Table 1 and it indicates that both parameterisation methods were affected consistently when additional energy parameters were concatenated to the feature vectors. The addition of delta, acceleration and delta energy coefficients to the feature vector improved the accuracy of the system. Note that adding the static log energy coefficient reduced the performance. This is not a surprising result as one cause is the different recording levels over telephone line. Another explanation is that static short-term features do not encapsulate the language specific information in contrast to transient features. As noted before, both standard parameterisation methods improved accuracy by adding features. It is clear that the use of additional features with LPCCs rather than with MFCCs was advantageous. An example can be found by observing the "Delta and Acceleration coefficient with window size of 15 frame" test where MFCCs gave a 29.6% improvement from the "no feature" test compared to 37.3% for LPCCs. Also worthy a note is that LPCCs outperformed MFCCs in all tests.

5. CONCLUSION

This paper has compared two methods to parameterise speech for a language identification system. Results show that the Linear Prediction Cepstrum Coefficients (LPCC) outperform Mel-Frequency Cepstrum Coefficients (MFCC) in all tests. The 12th order LPCC with delta and accelerate coefficients has resulted the best accuracy of 60.0 percent. The 12th order MFCC with 20 filterbanks, delta and delta energy coefficients obtained an accuracy of 55.6 percent. The superior result obtained by LPCC has shown that LPCC is capable of capturing extra information from speech that is increasing the ability to discriminate different languages. The experimental results also show that the UBM with the top 5 mixture test approach of the adapted GMM LID system is feasible and reasonable results can be obtained using this technique.

FEATURE ADDED	% Correct	
	MFCC	LPCC
No feature added	40.5	43.7
Delta coefficient with window size of 15 frame	51.1	57.4
Delta and Acceleration coefficient with window size of 15 frame	52.5	60.0
Delta coefficient with window size of 15 frame + log energy	48.9	53.1
Delta coefficient with window size of 15 frame + delta log energy	54.1	59.1
Delta coefficient with window size of 9 frame + delta log energy	55.6	59.5
Delta coefficient with window size of 23 frame + delta log energy	50.2	55.2

Table 1. Results comparing different feature vector combination, 12th order MFCC with 20 filterbanks and 12th order LPCCs from 14 LPCs are used in this experiment.

6. ACKNOWLEDGEMENTS

This work is sponsored by a research contract from the Australian Defence Science and Technology Organisation (DSTO). Jason Pelecanos has participated in discussions of both the LID system and related experiments and his contribution is greatly appreciated.

7. REFERENCES

[1] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291-298, 1994.

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

[3] J. D. Markel and A. H. Gray, *Linear prediction of speech*, New York: Springer-Verlag, 1976.

[4] S. Mendoza, L. Gillick, Y. Ito, S. Lowe and M. Newman, "Automatic language identification using large vocabulary continuous speech recognition," *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 785-788, 1996.

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

[6] NIST, Spoken natural Language Processing Group 2000, http://www.nist.gov/speech/

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

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

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

[10] S. Young, *The HTK Book: for HTK Version 2.1*, Cambridge, England: Cambridge University Press, 1997.

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

[12] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 31-44, 1996.