Speaker Identification Based on Log Area Ratio and Gaussian Mixture Models in Narrow-Band Speech: Speech Understanding...

Speaker Identification Based on Log Area Ratio and Gaussian Mixture Models in Narrow-Band Speech

David Chow and Waleed H. Abdulla

Electrical and Electronic Engineering Department, The University of Auckland, Auckland, New Zealand
Email: ccho071@ec.auckland.ac.nz, w.adbulla@auckland.ac.nz,
http://www.ele.auckland.ac.nz/~wabd002

Abstract. Log area ratio coefficients (LAR) derived from linear prediction coefficients (LPC) is a well known feature extraction technique used in speech applications. This paper presents a novel way to use the LAR feature in a speaker identification system. Here, instead of using the mel frequency cepstral coefficients (MFCC), the LAR feature is used in a Gaussian mixture model (GMM) based speaker identification system. An F-ratio feature analysis was conducted on both the LAR and MFCC feature vectors which showed the lower order LAR coefficients are superior to MFCC counterpart. The text-independent, closed-set speaker identification rate, as tested on the down-sampled version of TIMIT database, was improved from 96.73%, using the MFCC feature, to 98.81%, using the LAR features.

1 Introduction

Feature extraction is the key to the front-end process in speaker identification systems. The performance of a speaker identification system is highly dependent on the quality of the selected speech features. Most of the current proposed speaker identification systems use mel frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) as feature vectors [1]. Currently, researches are focusing on improving these two cepstral features. Orman had developed a new filter-bank to replace the mel frequency filter banks used in MFCC calculation [2]. On the other hand, there are many new features proposed to be used along with MFCC and LPCC [3, 4, 5]. Although MFCC and LPCC were proved to be two very good features in speech recognition, they are not necessarily being as good in speaker identification. In 1976, Sambur proposed to use orthogonal linear prediction coefficients as features in speaker identification [6]. In his work, he pointed out that for a speech feature to be effective, it should reflect the unique properties of the speaker's vocal apparatus and contains little or no information about the linguistic content of the speech [6]. As a result, he had tried to use linear prediction coefficients (LPC), parcor and log area ratio coefficients (LAR) as the speech features and then using orthogonal technique to reduce the linguistic content in those features. According to his work, log area ratio feature and parcor feature gave better results than LPC feature [6]. In this paper, LAR feature are chosen instead of parcor feature because it has a linear spectral sensitivity and is more robust to quantization noise [7]. In 1995, Reynolds demonstrated a Gaussian mixture model (GMM) based classifier work well in text-independent speaker identification even with speech feature that contains rich linguistic information like MFCC [3, 8]. With the above results, the authors believe that using LAR based features as feature vectors in the GMM-based speaker identification system will yield a very good identification result.

In this paper, LAR feature is investigated thoroughly by using the F-ratio analysis. A series of experiments about the performance of LAR feature on a speaker identification system had been conducted. This paper is organized as follows; section 2 gives a description of the LAR feature. Section 3 explains the GMM-based speaker identification system used in this paper. Section 4 compares the performance of LAR feature with the MFCC feature. Section 5 derives the conclusions of this work.

2 Log Area Ratio Coefficients

The log area ratio (LAR) coefficients are derived from the linear prediction (LPC) coefficients. Linear prediction coefficients are a highly effective representation of the speech signal. In this analysis, each speech sample is represented by a weighted sum of *p* past speech samples plus an appropriate excitation. The corresponding formula for the LPC model is:

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + Gu_n \qquad \textbf{(1)}$$

where *p* is the order of the LPC filter, s_n is nth speech sample and a_k is the kth coefficients of the LPC vector. The LPC are found by Durbin algorithm which minimizes the mean square prediction error of the model [7, 9].

The LPC model characterizes the vocal tract of a person. It can be transformed into other coefficients called Log area ratio coefficients (LAR). In LAR analysis, the vocal tract of a person is modelled as a non-uniform acoustic tube formed by cascading *p* uniform cylindrical tubes with different cross-section areas having equal lengths [9]. The glottis connected to the first tube is assumed to have zero area while the lips connected to the last tube is assumed to have infinite area. Figure 2.1 illustrates the acoustic tubes speech production model.

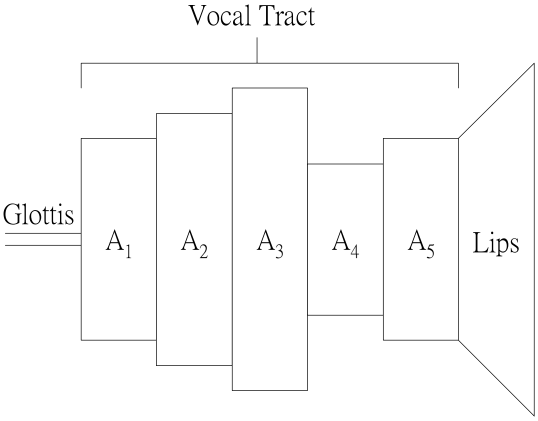

Fig. 2.1. The acoustic tubes speech production model.

In this model, the length of each cylindrical tube is closely related to the time different between two speech samples in LPC model which is equal to the sampling period. Therefore, in calculating the LAR coefficients, the length of the vocal tract is not necessary.

The LAR coefficients are formed by the log area ratio between the cross-section areas of every two connected tubes. The number of cylindrical tubes in the model is equal to the number of LAR coefficients plus one. The relationship between the LAR coefficients and the LPC is:

$$LAR_i = \log\left(\frac{A_i}{A_{i+1}}\right) = \log\left(\frac{1+\alpha_i}{1-\alpha_i}\right), A_{p+1} = 1 \qquad (2)$$

where α_i is the i$^{\text{th}}$ parcor coefficients which can be found by:

$$\alpha = a_i^{(i)}, 1 \leq i \leq p \qquad (3)$$

where $a_i^{(i)}$ is the *i*th LPC calculated by the i$^{\text{th}}$ order LPC model [9].

3 Computation Complexity Analysis

According to Karpov, the time to compute MFCC feature is about 1.2 times slower than computing LPCC feature [10]. The computation of LAR feature is very similar to LPCC feature. Both the LAR and LPCC algorithms required to compute the autocorrelation matrix and both require the Durbin algorithm to solve the system of equations formed by the autocorrelation matrix. However, the last step to compute LAR feature is different to LPCC feature. The computation complexity of the last step of LAR feature is 4p operations while the LPCC counterpart is p(p+1) operations [10] where p is the order of the analysis. In conclusion, the computation complexity of LAR feature is slightly less than LPCC and thus less than MFCC.

4 Gaussian Mixture model based speaker identification system

In this speaker identification system, each speaker enrolled in the system is represented by a Gaussian mixture model (GMM). The idea of GMM is to use a series of Gaussian functions to represent the probability density of the feature vectors produced by a speaker. The mathematical representation is:

$$P(\vec{x} \mid G_s) = \sum_{i=1}^{M} w_i G_i(\vec{x} \mid \overline{\mu}_i, \Sigma_i) \qquad \mathbf{(4)}$$

where M is the number of mixtures, $\vec{x}$ is the feature vector, w_i is the weight of the i-th mixture in the GMM, $\overline{\mu}_i$ is the mean of the i-th mixture in the GMM and $\sum_i$ is the covariance matrix of the i-th mixture in the GMM [3, 8]. The Model parameters $(w_i, \overline{\mu}_i, \Sigma_i)$ characterize a speaker's voice in the form of a probabilistic density function. They are determined by the Expectation maximization (EM) algorithm [11].

In the identification phase, the log-likelihood scores of the incoming sequence of feature vectors coming from each speaker model are calculated by:

$$L(X, G_s) = \sum_{t=1}^{F} P(\vec{x}_t \mid G_s) \qquad \mathbf{(5)}$$

where $X = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_F)$ is the sequence of feature vectors and F is the total number of feature vectors [3, 8]. The speaker whose speaker model generates the highest score is identified as the producer of the incoming speech signal. This decision method is called maximum likelihood (ML).

5 Experimental Method and Results

The speech data used in our speaker identification experiments consist of 112 males and 56 females selected from the testing set of the TIMIT database. TIMIT is a clean speech database recorded using a high quality microphone sampled at 16 kHz. In this paper, the speech signal used was down sampled from 16 kHz to 8 kHz in order to test the identification accuracy under narrow-band (0–4000Hz) speech. Each speaker produces 10 sentences, the first 8 sentences were used in training and the last 2 sentences were used in testing. The average length of each sentence is 3 seconds. In other word, there was about 24 seconds of speech for training and 3 seconds for testing.

The speech signal was extracted by energy based algorithm modified from the Rabiner's one [12]. No pre-emphasis filter was applied to the signal. The analysis of the speech signal was conducted over the speech frames of 20ms duration with 10ms overlapping. The windowing function used was Hamming window. The length of the window is chosen so that there are enough speech samples in each frame to estimate the speech spectrum and make it insensitive to window placement with respect to pitch periods. The classification engine used in this experiment was a 32 mixtures GMM initialized by vector quantization (VQ) [13].

5.1. F-ratio analysis

F-ratio is a figure of merit to evaluate the effectiveness of each feature coefficient. The formula of the F-ratio is:

$$F - ratio = \frac{\text{speaker variance among classes}}{\text{speaker variance within classes}} \qquad \textbf{(6)}$$

Figure 5.1 shows the F-ratio of the MFCC feature and LAR feature. It can be clearly seen that the lower order coefficients of LAR feature has higher F-ratio score than the MFCC counterpart. For the application of text-independent speaker identification, the F-ratio scores provide a good indication on the quality of the features but it is not perfect. It is because the three assumptions of F-Ratio are not fully satisfied. The three assumptions are:

The feature vectors within each class must have Gaussian distribution.
The features should be statistically uncorrelated.
The variances within each class must be equal [14].

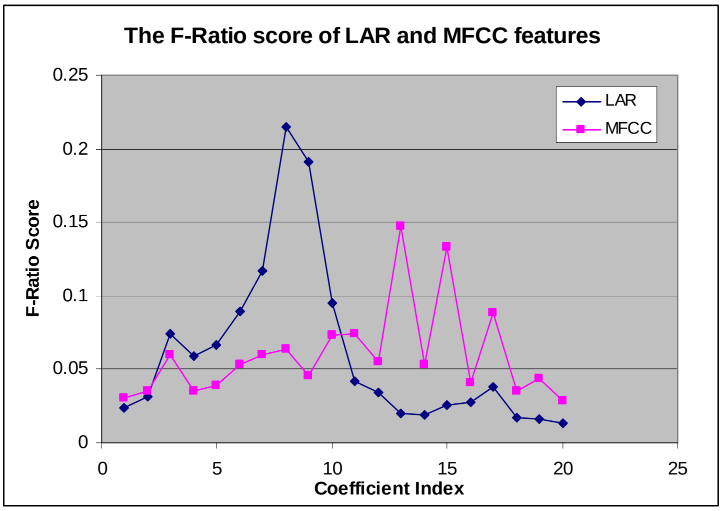

Fig. 5.1. The F-ratio score of the LAR and MFCC features.

5.2. Identification results

The identification tests were conducted by 168 speakers according to the experimental setup described at the beginning of section 5. In each test, each speaker conducted 2 trials on the system.

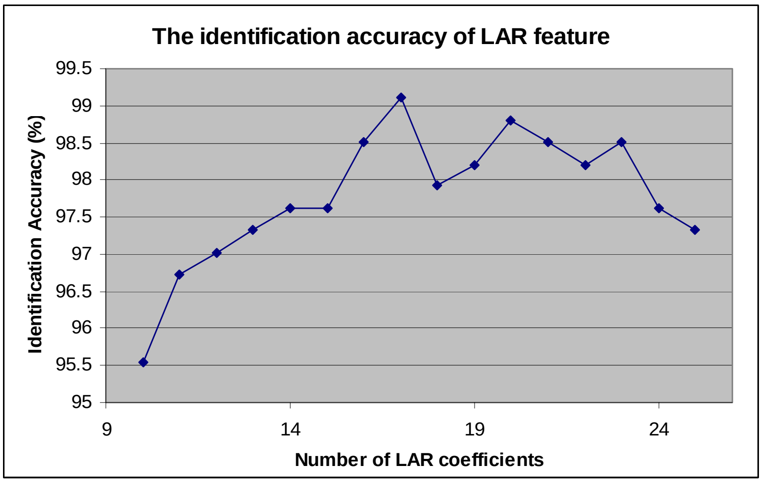

Fig. 5.2. The identification rate of the LAR feature.

Table 5.1 compares the identification accuracies obtained by three different experiments with similar setup and using similar speaker identification system. As can be seen in the table, the result obtained in this paper is virtually the same as the results obtained by Reynolds. This proved the correct implementation of MFCC and GMM based speaker identification used in this paper.

Table 5.1. The identification rate of MFCC based speaker identification system under wide-band speech.

	Identification rate
This paper (168 speakers from TIMIT)	99.4%
Reynolds's result in [8] (630 speakers from TIMIT)	99.5%
Reynolds's result [15] (168 speakers from TIMIT)	99.1%

Figure 5.2 shows that the identification rate of using 20 LAR coefficients produces the best result where the identification rate of 98.81% was achieved. Table 5.2 compares the identification rate using MFCC and LAR features. It also shows the identification rate obtained by Reynolds using similar setup. From the table, the identification rate of MFCC is just 96.73% as opposed to 98.81% obtained by LAR. A 2.08% improvement is achieved. The identification results obtained in this paper is higher than that obtained by Reynolds. One reason that explains the slightly worst result obtained by Reynold is the MFCC feature used by him only covered the telephone pass-band where the MFCC feature used in this paper covered the whole 0 – 4000Hz bandwidth.

Table 5.2. The identification rate of LAR and MFCC features

	Identification rate
20 LAR coefficients	98.81%
20 MFCC coefficients	96.73%
Reynolds's result [15]	95.2%

6 Conclusions

This paper presents a novel way of utilising the LAR feature in a GMM-based speaker identification system. The new speaker identification system using 20 LAR coefficients achieved an identification rate of 98.81% as opposed to 96.73% obtained by the MFCC-based speaker identification system.

The F-ratio analysis showed that the LAR feature is more efficient to capture the speaker's related information than the MFCC feature.

The computation of LAR feature has less computation complexity than the MFCC counterpart. Also, LAR feature is robust to quantization. These advantages make LAR feature extraction method easy to be implemented in an embedded system.

7 Acknowledgement

This work is supported by The University of Auckland Vice Chancellor Development Fund.

References

1. Premakanthan P. and Mikhad W. B. (2001) Speaker Verification/Recognition and the Importance of Selective Feature Extraction: Review. *MWSCAS.* **Vol 1,** 57-61.
2. Orman O. D. (2000) Frequency Analysis of Speaker Identification Performance. Master thesis, Boğaziçi University.
3. Sanderson S. (2002) Automatic Person Verification Using Speech and Face Information. PhD thesis. Griffith University.
4. Petry A. and Barone D. A. C. (2001) Fractal Dimension Applied to Speaker Identification. *ICASSP (Salt Lake City).* May 7-11. 405-408.
5. Liu C. H., Chen O. T. C. (2002) A Text-Independent Speaker Identification System Using PARCOR and AR Model. *MWSCAS.* **Vol 3,** 332-335.
6. Marvin R. S. (1976) Speaker Recognition Using Orthogonal Linear Prediction. *IEEE Transactions on Acoustic, Speech and Signal Processing.* **Vol 24,** 283-289.
7. Makhoul J. (1975) Linear Prediction: A Tutorial Review. *Proceedings of the IEEE.* **Vol 63,** 561-579.
8. Reynolds D. A. (1995) Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication.* **Vol 17,** 91 – 108.
9. Campell J.P. and Jr. (1997) Speaker recognition: a tutorial. *Proceeding of the IEEE.* **Vol 85,** 1437-1462.
10. Karpov E. (2003) Real-Time Speaker Identification. Master thesis, University of Joensuu.
11. Bilmes J. A. (1998) A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report, University of Berkeley.
12. Rabiner L. and Sambur B. (1975) An Algorithm for Determining the Endpoints of Isolated Utterances. *The Bell System Technical Journal.* **54**, pp 297 – 315.
13. Linde Y., Buzo A., Gray, R. (1980) An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications.* **Vol. 28(1)**, 84-95.
14. Paliwal K. K. (1992) Dimensionality Reduction of the Enhanced Feature Set for the HMM-Based Speech Recognizer. Digital Signal Processing. **Vol. 2.** 157-173.
15. Reynolds D. A., Zissman M. A., Quatieri T. F., O'Leary G. C., Carlson B. A. (1995) The Effects of Telephone Transmission Degradations on Speaker Recognition Performance. *ICASSP (Detroit).* May 9-12. 329-331.