WARPED MAGNITUDE AND PHASE-BASED FEATURES FOR LANGUAGE IDENTIFICATION

Felicity Allen[1], Eliathamby Ambikairajah[1] and Julien Epps[2,1]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052 Australia
[2]National ICT Australia, Eveleigh 1430, Australia
felicity.allen@optusnet.com.au, ambi@ee.unsw.edu.au, julien.epps@nicta.com.au

ABSTRACT

To date, systems for the identification of spoken languages have normally used magnitude-based parameterization methods such as the MFCC and PLP. This paper investigates the use of the recently proposed modified group delay function (MODGDF) coefficients in combination with traditional magnitude-based features in a Gaussian Mixture Model (GMM) based system. We also examine the application of feature warping to magnitude-based features and the MODGDF and find that it can offer a significant cumulative improvement. We find that the addition of a modified regression-based Shifted Delta Cepstrum (SDC) further improves system performance beyond that obtained by a more standard SDC configuration. The combination of PLP, feature warping and the proposed regression-based SDC achieved an accuracy of 88.4% in tests on 10 languages in the OGI TS Corpus, which compares very favourably with alternative language identification systems reported in the literature.

1. INTRODUCTION

Research into acoustic language identification has recently gained momentum after an acoustic based system out-performed more traditionally accepted phonetic systems in the NIST 2003 Evaluation task [1]. This revived efficacy of acoustic language identification systems stems largely from the extension of the traditional delta and acceleration cepstrum to the so-called Shifted Delta Cepstrum (SDC) and the use of more efficient Gaussian Mixture Model (GMM) adaptation algorithms for training and testing [2]. Systems employing these methods provide a fast and effective means of language identification. However, while acoustic language identification systems have improved markedly in recent years, the search continues for the most effective front-end processing methods for distinguishing between spoken languages.

Feature warping was proposed for use in language identification by the current authors and has been shown to provide significant improvements to language identification performance [3]. Recently used for speaker identification [4] and for robust speech recognition [5], feature warping maps the short-term distribution of each feature stream to a standardized distribution. It provides improved compatibility between training and test data, reduces channel mismatch and noise and provides greater compatibility with a GMM back-end.

Traditionally, language identification systems have only included features derived from the magnitude of the frequency spectrum, such as the MFCC and PLP. While some attention has been paid to the inclusion of information on the prosodic features in the signal, such as the pitch and intensity, very little attention has been paid to the inclusion of information relating to the phase of the frequency spectrum. It is well-known that quality reconstruction of speech signals is only possible from the magnitude spectrum if a reliable estimate of the phase is included. Recent research has also shown that human auditory perception may rely on both amplitude and frequency modulation for the comprehension of sounds [6]. We hypothesize that the phase carries important information about the speech signal that can be exploited to distinguish between languages.

Coefficients based on a modified calculation of the group delay of the phase spectrum have recently been proposed for language identification [7]. These coefficients have been shown to offer small improvements in language identification performance, with further small improvements obtained when they are combined with the MFCC using late fusion [7].

This paper makes a detailed comparison of the effects of feature warping and the SDC on system performance when applied to both magnitude (PLP or MFCC) and phase-based coefficients (MODGDF). We also propose a modified method for calculating the SDC using a regression-based calculation of the first order delta cepstrum.

2. FRONT-END TECHNIQUES

2.1 Feature Warping

Feature warping, also known in the image processing literature as histogram equalization and in the speech

processing literature as cumulative distribution mapping, has been previously shown by the current authors to produce significant improvements in language identification accuracy [3]. The technique maps each feature vector stream to a standardized distribution over a specified time interval. Speech recognition [5], speaker verification [4] and language identification [3] tasks using this technique have been found to exhibit superior performance to those using other methods including Cepstral Mean Subtraction (CMS) and mean and variance normalization.

Previous research by the current authors has also shown that feature warping is best applied to the MFCC or PLP only, providing no benefits when applied to the delta and acceleration or shifted delta cepstrum [3]. While feature warping can be applied using any standard probability distribution, the best results are obtained when a Gaussian distribution is used [3].

The methodology for performing the feature warping in this paper is the same as that described in [3] and [4]. A sliding rectangular window of N samples is applied, with the new warped value calculated for the cepstral feature in the centre of the window as shown in Figure 1, where R is the new index of the centre value when the window is sorted into ascending order and *norminv* is the inverse of the normal cumulative distribution function.

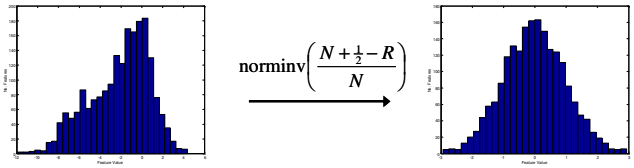

Fig. 1. Feature warping transformation

2.2 Modified Group Delay Coefficients

Traditionally, language and speaker recognition tasks use feature vectors containing cepstra derived from the magnitude of the Fourier transform such as the MFCC and the PLP. Recently Murthy *et al*. have proposed using coefficients based on the phase information via a modified calculation of the group delay function, their so-called MODGDF [7],[9]. They recently showed that the MODGDF could produce results comparable with the MFCC for language identification and that a small improvement could be obtained when the MODGDF were combined with the MFCC using late fusion [7]. However their results only considered the coefficients when used without further processing or with traditional delta and acceleration cepstra appended. No experiments have been published to date that consider the combination of the MODGDF with either the shifted delta cepstrum (SDC) or feature warping.

According to the methodology described in [7], the MODGDF coefficients are calculated as follows: A smoothed estimate of the group delay function is calculated from the speech signal $x[n]$ according to

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}}, \quad (1)$$

where $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x[n]$ and $nx[n]$ respectively and the R and I subscripts denote the real and imaginary parts respectively. $S(\omega)$ is a cepstrally smoothed version of $|X(\omega)|$ and $\gamma \in [0, 1]$ is a constant introduced to reduce the spiky nature of the formants. The final modified group delay coefficients are then calculated by taking a DCT of $\tau_m(\omega)$, which is given by

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)(|\tau(\omega)|^{\alpha}). \quad (2)$$

Here $\alpha \in [0, 1]$ is another constant introduced to reduce the spiky nature of the formants.

2.3 Shifted Delta Cepstrum

Traditionally, language and speaker recognition tasks use feature vectors containing cepstra and delta and acceleration cepstra. Recently, however, the Shifted Delta Cepstrum (SDC) has been found to exhibit superior performance to the delta and acceleration cepstra in a number of language identification studies [2],[3] due to its ability to incorporate additional temporal information, spanning multiple frames, into the feature vector.

The SDC is obtained by concatenating the first order delta cepstra computed across multiple frames of speech. Four parameters (N, D, P, and k) specify the standard computation of the SDCs as follows [8]: N cepstral coefficients are computed each frame, then for each of these cepstral streams, the final vector at time t is given by the concatenation of the $\Delta c(t+iP)$ for all $0 \le i < k$, where

$$\Delta c(t + iP) = c(t + iP + D) - c(t + iP - D). \quad (3)$$

For some of the experiments in this paper, a more robust regression-based SDC calculation is proposed in place of (3), based on a method originally reported by Furui [11] for calculating the first order delta MFCC values. This has not previously been used for SDC calculation but is expected to provide a smoother and more robust estimate of the local slope. The same four parameters specify the computation, but the final vector at time t is given by the concatenation of the $\Delta c(t+iP)$ for all $0 \le i < k$, where

$$\Delta c(t + iP) = \frac{\sum_{d=-D}^{D} d\, c(t + iP + d)}{\sum_{d=-D}^{D} d^2}. \quad (4)$$

3. PROPOSED LANGUAGE IDENTIFICATION SYSTEM

3.1 Feature Extraction

The complete proposed feature extraction configuration is shown in Figure 3. The magnitude-based coefficients (PLP or MFCC) and MODGDF coefficients were calculated from the input speech. In place of the late fusion method [7] for combining magnitude and phase coefficients, we concatenated them to form a single feature vector. Feature warping was then performed using a 3 second sliding window and a Gaussian distribution. The SDC were calculated from the un-warped features using the regression based method in (4). A parameter configuration of *N*-3-3-7 was used, where *N* is the total length of each feature. The SDC were then concatenated with the warped feature vectors to form the final features.

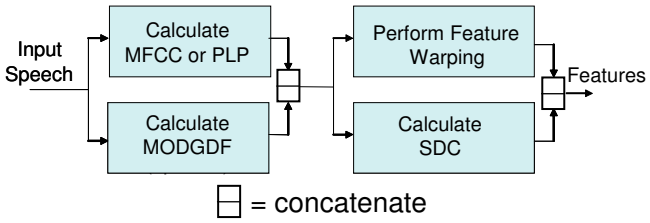

Fig. 3. Proposed feature extraction procedure

3.2 Classification

The back-end used for all experiments in this study was a GMM-based acoustic language identification (LID) system identical to that used in [3]. A single GMM was trained on data from all languages using the expectation maximization algorithm. Then for each of the 10 languages, a separate GMM was adapted using data from that language only. Testing was performed using the fast scoring technique described in [10], considering only the five most significant components. The chosen language was selected based on the adapted language GMM that produced the highest average log likelihood score. All tests were conducted on 20-second speech samples.

4. EXPERIMENTS

4.1 Corpus

All the experiments described in this paper were conducted on the 1994 OGI Multi-language Telephone Speech Corpus using the data sets defined for the National Institute of Science and Technology (NIST) 1994 Evaluation task. All used closed-choice tests on 10 languages. The training and development sets were both used to train the LID system and the evaluation set was used for evaluation testing.

4.2 Results

Investigations were conducted to compare the performance of the MFCC, PLP and MODGDF individually, combined with delta and acceleration coefficients, feature warped, concatenated with the standard 7-1-3-7 configuration of the SDC (calculated using equation 3) and concatenated with the 7-3-3-7 regression based SDC (calculated using equation 4). All these experiments used GMMs with 256 mixtures.

The results are given in Table 1, and demonstrate that feature warping improves system performance in all cases, and substantially so in many cases. However it is shown to be less beneficial when applied to the MODGDF, providing an average relative improvement in accuracy of only 3.3%, compared to an average relative improvement of 54.7% when applied to the PLP and 29.6% when applied to the MFCC.

The results also show that while the MODGDF achieve comparable performance with the MFCC and PLP in experiments where either just the 7 coefficients were used or where 12 coefficients and the delta and acceleration cepstra were used, they performed relatively poorly when combined with the SDC and feature warping. The poorer performance when combined with the SDC is likely to be because the SDC configuration that performs best for the MFCC and PLP is not the optimal configuration for the MODGDF.

A significant improvement is found to be obtained in most cases by using the 7-3-3-7 proposed regression based method for calculating the SDC in equation (4) rather than the commonly used 7-1-3-7 simple subtraction method of equation (3) [2],[3].

Table 1. Accuracy comparison of MFCC, PLP and MODGDF features alone, with feature warping and with the SDC. The effect of the proposed regression-based SDC on warped MFCCs is shown in bold.

	Without Feature Warping (% Correct)			With Feature Warping (% Correct)		
	MFCC	PLP	MODGDF	MFCC	PLP	MODGDF
7 Coefficients only	23.0	27.8	24.9	-	-	-
12 Coefficients with Delta and Acceleration Cepstra	41.9	31.7	37.9	60.8	64.4	38.3
7-1-3-7 SDC (calculated using eq. 3)	55.9	56.0	-	**73.4**	76.4	-
7-3-3-7 SDC (calculated using eq. 4)	71.9	61.4	50.7	**80.9**	76.4	53.5

An average relative improvement of 12.1% is obtained by using this former method over the latter.

4.3 Combining Magnitude and Phase features

In these experiments, the MODGDF coefficients were concatenated with the magnitude cepstra (MFCC and PLP) to form a single feature vector. Feature warping was performed in all tests and the 7(14)-3-3-7 regression-based SDC (calculated according to equation (4)) was appended. 256 order GMMs were used in all tests. The results yielded accuracies of 69.2% for MFCC+MODGDF and 63.4% for PLP+MODGDF. Given that the concatenated features contain additional information, one reason for the poorer performance in these experiments may be because 256 mixtures are not sufficient to accurately model the increased dimensionality of the input feature vectors.

4.4 Final Configuration

In order to arrive at an optimum configuration for language identification based upon the preceding results, the number of mixtures in the GMMs was increased from 256 to 1024 mixture components for three system configurations. Given the poor performance of the MODGDF features in the preceding tests, they were not included here. The results are shown in Table 2. Increasing the number of mixtures resulted in improved performance in all cases and an average relative improvement in performance of 6.9%. The best performing system achieved a final accuracy of 88.4% using PLPs with feature warping and the proposed regression-based SDC with 8-3-3-7 configuration.

Table 2. Language identification accuracies (% correct) for optimized configurations using higher order GMMs

	256 Mixtures	1024 Mixtures
12 MFCCs with Delta and Acceleration cepstra	60.9	64.1
PLP with 7-1-3-7 standard (eq. (3)) SDC and feature warping	76.4	83.6
PLP with 8-3-3-7 regression (eq. (4)) SDC and feature warping	**83.4**	**88.4**

5. CONCLUSION

The results in this paper have shown that feature warping offers improvements to both magnitude and phase-based front-end feature vectors. The proposed regression-based method for calculating the SDC achieves an average relative improvement in accuracy of 12.1% over the standard SDC calculation for the OGI TS corpus. The MODGDF coefficients were found to produce comparable results to the MFCC and PLP when used individually, but performed relatively poorly when concatenated with MFCC or PLP coefficients. Our experiments confirm that magnitude-based features offer the best accuracy for language identification. The best performing system used the PLP with feature warping and an 8-3-3-7 regression-based SDC to achieve an accuracy of 88.4% in closed choice tests between 10 languages on the OGI TS Corpus.

Future work will conduct a more thorough investigation of the optimal SDC configuration and feature warping distribution to use with the MODGDF.

6. REFERENCES

[1] P. Matejka, J. Cernocky and M. Sigmund, "Introduction to Automatic Language Identification," in *Proc. Radioelektronika Conf., 2004.*

[2] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell and D. A. Reynolds, "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition", in *Proc. EUROSPEECH*, 2003, pp. 1345-1348.

[3] F. Allen, E. Ambikairajah and J. Epps, "Language Identification Using Warping and the Shifted Delta Cepstrum", in *Proc. IEEE Int. Workshop on Multimedia Signal Processing* (Shanghai, China), 2005.

[4] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification", in *Proc. A Speaker Odyssey, The Speaker Recognition Workshop*, 2001, pp 243-248.

[5] A. Torre, J. Segura, C. Benitez, A. M. Peinado, and A. J. Rubio, "Non-Linear Transformations of the Feature Space for Robust Speech Recognition", in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. 401-404.

[6] K. Nie, Stickney, G., and Zeng, F.-G., "Encoding frequency modulation to improve cochlear implant performance in noise", *IEEE Trans. Biomedical Engineering*, vol. 52, no. 1, January 2005, pp. 64-73.

[7] R. M. Hedge, and H. A. Murthy, "Automatic Language Identification and discrimination using the modified group delay feature," in *Proc. Int. Conf. on Intelligent Sensing and Information Processing*, Chennai, 2005, pp. 395-399.

[8] P. A. Torres-Carassquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, Jr., "Approaches to language identification using Gaussian Mixture Models and shifted delta cepstral features" in *Proc. ICSLP*, 2002, pp. 89-92.

[9] H. A. Murthy and V. Gadde, "The Modified Group Delay Function and its Application to Phoneme Recognition" in *Proc. IEEE ICASSP*, 2003, pp. 68-71.

[10] E. Wong and S. Sridharan, "Methods to improve gaussian mixture model based language identification system", in *Proc. ICSLP*, 2002, pp. 93-96.

[11] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", *IEEE Trans. Acoust., Speech and Sig. Proc.*, vol. 34 no.1, 1986, pp. 52-59.