Combining Cepstral and Prosodic Features in Language Identification

Bo Yin[1], Eliathamby Ambikairajah[1], Fang Chen[2]
School of Electrical Engineering and Telecommunications UNSW[1], National ICT Australia Ltd.[2]
bo.yin@student.unsw.com.au, ambi@ee.unsw.edu.au, fang.chen@nicta.com.au

Abstract

A novel approach of combining cepstral features and prosodic features in language identification is presented in this paper. This combination approach shows a significant improvement on a GMM-UBM based language identification (LID) system which utilizes modern shifted delta cepstrum (SDC) and feature warping techniques. The proposed system achieves a high accuracy of 87.1% on a 10-language task, and outperforms the baseline system by 12%. The prosodic features are proven to be very effective in both tonal and non-tonal LID, as they deliver new language-discrimination information in addition to those from widely used cepstral features.

Additionally, the performance of MFCC and PLP features with different coefficient numbers in language identification tasks are researched and compared. Less number of coefficients is more likely to be sufficient or even better for language identification.

1. Introduction

This paper aims to propose a new feature set which will improve overall LID performance. Cepstral based features, such as Mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP), are the most popular features utilized in modern LID system[1]. On the other hand, prosodic features have drawn increasing attention in the speech processing area in recent years. Several attempts in language identification have been presented[2, 3]. However, these attempts are focusing on or only show positive results from tonal languages. In this paper, a novel approach of combining cepstral and prosodic features is presented which utilize the silence removing, feature warping and temporal information extraction techniques. This approach shows an impressive improvement both for identifying tonal and non-tonal languages.

Different cepstral features and different coefficient numbers will lead to different performances. However, there is still little research concentrated on analyzing how the number of coefficients will affect the performance of LID. The performances of different cepstral features with different coefficient numbers are compared in this paper.

In this paper, the Oregon Graduate Institute (OGI-92) multi-language telephone speech database is used for all experiments. A Gaussian Mixture Model – Universal Background Model (GMM-UBM) back-end based system is selected as the baseline system instead of other reported better back-end based systems, such as Parallel Phone Recognizer followed by Language Model (PPRLM) system[1], beside the effective features, a GMM-UBM based system is more efficient in computation.

2. Cepstral features

Cepstral based features, which typically represent the magnitude properties of speech spectrum, are widely used in speech processing. Choosing effective features is important to achieve a high performance. MFCC and one of its alternatives, PLP coefficients, are the two most popular cepstral features.

Previous analysis and experiments show the slight difference between MFCC and PLP feature extraction does not deliver much performance difference according to speech recognition task[4, 5]. Some research shows the system achieves a slightly better performance with MFCC than PLP[4], while others show that PLP gives slightly better performance[5]. Since the MFCC and PLP extraction process are similar, the information gathered from them will be similar too. That is the reason why combining these two features does not produce a valuable improvement.

Although the research of comparing the performance of using MFCC and PLP has been done in speech or speaker recognition tasks, little research has concentrated on the influence of the number of coefficients, especially in language identification task. In this paper, this influence is researched by conducting experiments with different cepstral features containing different number of coefficients. Different results are shown other than commonly used 12

0-7695-2521-0/06 $20.00 © 2006 IEEE

MFCC coefficients or 9 PLP coefficients, which are adapted from the experience of speech recognition.

3. Prosodic features

Prosody plays a key role in the perception of human speech. The information contained in prosodic features is partly different from the information contained in cepstral features. Therefore, more and more researchers from the speech recognition area are showing interests in prosodic features. Recently, some researchers also presented their approaches of utilizing prosodic information into language identification. These approaches include unsupervised learning pitch contour through GMM, following by fusion with other results from different features[3]; and unsupervised learning through HMM[2]. However, most results of this research showed the prosodic information is only beneficial for tonal language task.

3.1. Pitch and intensity

Generally, prosody means "the structure that organizes sound"[6]. Tone, loudness, and the rhythm (tempo) structures are the main components of prosody. To utilize them, suitable physical representations (features) have to be devised. Typically, these features include pitch, intensity, and the normalized duration of syllables. In this paper, as the limitation of feature combination (detailed in next section), the features have to be frame based. Pitch and intensity are selected to represent the prosodic information in this paper.

Silence, which does not present any useful information but only noise, is another issue when utilizing prosodic features. The zero values of pitch in silence segments will distort the GMM and produce inaccurate model[7]. In this paper, a pitch detection based silence removing approach is presented, which produce superior result than conventional methods like power detection or silence modeling.

3.2. Temporal information

Pitch and intensity are static features as they are calculated frame by frame. They only represent the exact value of the current frame in this situation. In order to reveal more pattern information, the temporal information has to be extracted.

In previous research, the shifted delta cepstrum (SDC) was reported to produce superior performance in cepstral feature based language identification systems[8]. The standard SDC is calculated as follows:

$$\Delta c(t+iP) = c(t+iP+D) - c(t+iP-D) \quad (1)$$

The final vector at time t is given by the concatenation of the $\Delta c(t+iP)$ for all $0 \le i < k$, where the D, P and k are the parameters of SDC, $c(t)$ refers to the original feature value at time t. More recently, a modified version of SDC (modified SDC) was presented and reported to have even higher performance[9], which is calculated as follows:

$$\Delta c(t+iP) = \frac{\sum_{d=-D}^{D} dc(t+iP+d)}{\sum_{d=-D}^{D} d^2} \qquad (2)$$

However, these temporal information extraction methods have not been reported for prosodic features. In this paper, to investigate the effects of introducing temporal prosodic information, delta/acceleration and modified SDC were conducted for prosodic feature only configuration. The result shows the better method for extracting temporal information from prosodic feature.

4. Feature combination

In previous research, fusion and feature concatenation are two commonly used methods for combining different features. Fusion technology, which fuses different sub-system's results together, is widely used in modern LID systems [10]. In a typical fusion based system, each feature set is independently used to create an independent model. These models are used in different sub-systems. From the view of statistics theory, features with a smaller number of components will lead to a less robust trained model. That means more samples are required for training those separate models to achieve comparable stability. Unfortunately, the prosodic features contain only two components in this case. Feature concatenation is a simple solution to avoid this problem.

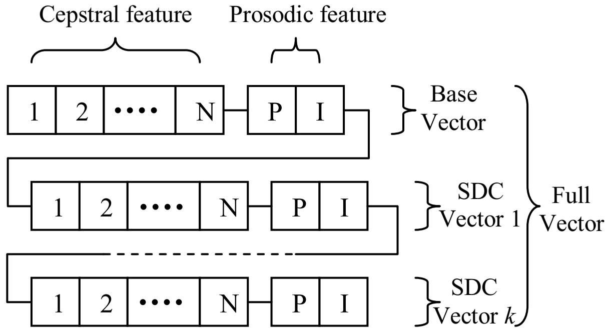

Fig. 1. The structure of combined feature vector

Fig. 1 shows the structure of the combined feature vector proposed in this paper. With this method, cepstral and prosodic features are directly concatenated to formulate one single base feature vector, which presents all information from each feature. The full vector is produced by concatenating this base feature vector and other delta vectors produced by SDC processing.

5. Experiments

Following experiments are conducted in this paper: comparing the performances between utilizing MFCC and PLP features with different coefficient numbers; comparing the performances between utilizing cepstral features only and combined features with prosody; and comparing the performances of different temporal prosodic information extraction methods.

5.1. Configuration and database

The baseline system was originally developed for GMM-UBM based LID research[8]. Since the purpose of all experiments is to find out the performance differences of utilizing different features in LID, this GMM-UBM based LID system is acceptable. The processing flows of training and testing are shown in Fig.2 and Fig.3. When training, after the features are extracted from speech data, they are expanded by modified SDC to enhance the temporal information. Then feature warping normalizes the distribution of feature data to Gaussian distribution, which will improve the GMM accuracy. Part of those processed feature data are used for training the Universal Background Model (UBM), while remaining parts are used for adapting the UBM to different GMMs corresponding to each language. When testing, the features of target speech are extracted and compared to each GMM. The language with the maximum log likelihood is determined as result. In this paper, the configuration of modified SDC is D=3, P=3, k=7. The GMM order is 256.

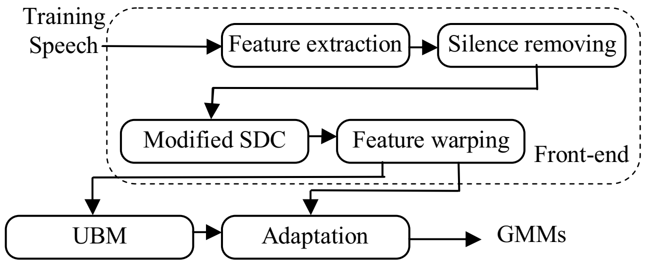

Fig. 2. Processing flow of training

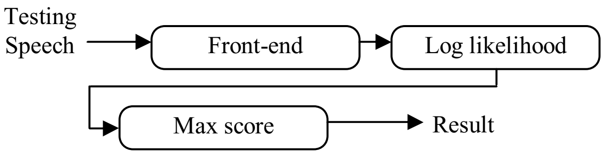

Fig. 3. Processing flow of testing

OGI-92 telephony speech database is utilized for all experiments in this paper. This database is a multi-language, multi-speaker database, composed of an average 122 calls (approx. 2 minutes each, different speakers for different calls) in each of 11 languages. Among them, 10 languages, English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnam, were used for experiments in this paper. For each language, 70 calls were used for training and adapting GMM, 20 calls were used for evaluation.

5.2. Results and discussion

Different coefficient numbers are used to conduct the experiments. The baseline is MFCC-12 (12 MFCC coefficients) and PLP-9 (9 PLP coefficients), which is widely used in LID. As shown in table 1, the change of coefficient number improves the final LID accuracy both for MFCC and PLP. The best coefficient number of MFCC is 7, which achieves a 2.4% improvement to the baseline. Similarly, 7 is the best coefficient number of PLP, which introduces 1.2% improvement. On average, MFCC and PLP show very similar performance in this LID task.

This result can be explained as follows: along with the increase of the coefficients number, more speech information is extracted, however, more noise is mixed as well. Therefore, there is a trade-off between these two influences. The result shows the best trade-off situation in this case. It also shows that fewer coefficients of cepstral feature produce sufficient or even better performance in this situation.

Feature	Corr.%	Feature	Corr.%
MFCC-5	73.4%	PLP-5	77.4%
MFCC-7	**77.8%**	**PLP-7**	**78.2%**
MFCC-9	76.6%	PLP-9	77.0%
MFCC-12	75.4%	PLP-11	77.4%
		PLP-13	72.6%

Table 1. Results from different cepstral features

Since MFCC and PLP features achieved similar accuracy, the prosodic feature combination experiments are conducted with both MFCC and PLP features. Table 2 shows the correction rate of both cepstral only feature set and combined feature set with different coefficient numbers. The highest performance is achieved by combining MFCC-7 with prosodic features, while the combination with MFCC feature shows better result on average.

Cepstral feature	Corr.% cepstral only	Corr.% cepstral + prosodic
MFCC-5	73.4%	86.3%
MFCC-6	75.0%	86.4%
MFCC-7	77.8%	**87.1%**

MFCC-9	76.6%	84.7%
MFCC-12	75.4%	80.2%
PLP-5	77.4%	81.9%
PLP-7	78.2%	85.5%
PLP-9	77.0%	80.6%

Table 2. Combining prosodic and cepstral features

It is shown that, although the correction rate drops quickly when the coefficients number of MFCC decrease from 7, the combined correction rate drops slowly. It can be explained that prosodic features contribute more when overall coefficient number become smaller, because the weight of prosodic features become larger, e.g. 2/7 when combining with MFCC-5 and 2/9 when combining with MFCC-9.

This result reveals that combining prosodic features with cepstral features achieves a significant improvement of overall performance. The accuracy of 10 languages LID task increased from 77.8% to 87.1% for the MFCC-7 case. Though this absolute result may be not suitable for directly comparison with those based on different databases, the effects should be similar. Therefore, this performance is still comparable to other modern LID systems, and even better than those 256-orders GMM based systems. Prosodic features, by delivering new information of speech, are proven to be very effective add-on features for the widely used cepstral features in LID task.

This result shows: the prosody patterns are significantly different among all different languages. The prosodic features play an important role in discriminating different languages of both tonal and non-tonal, though it may not deliver much valuable information for content recognition.

To research how the temporal information affects the performance of prosodic feature, experiments utilizing different delta methods on a prosodic features only system were conducted. From table 3, modified SDC gives a remarkably superior result than standard delta-acceleration. This result proves the analysis in section 3.2, and supports applying modified SDC to prosodic features in section 4.

Delta method	**Correction%**
None	13.7%
Delta and acceleration	46.4%
Modified SDC 3-3-7	71.4%

Table 3. Different delta configuration for prosodic features

6. Conclusion

There are two main conclusions that can be drawn from the experiments and analysis in this paper. First, prosodic feature can significantly improve the LID performance by combining with the widely used cepstral features, because it contains new important language-discrimination information in addition to those from standard cepstral features. By combining pitch, intensity and MFCC features properly, a modern GMM-UBM based language identification system improves remarkably: a 12.0% relative correction rate improvement is achieved while the overall accuracy increasing to 87.1%. On the other hand, for both MFCC and PLP features, less number of coefficients is more likely to be able to achieve similar or even better performance in LID. In the case of OGI-92 database, the MFCC and PLP coefficients numbers which achieved the highest performance are both 7, which is remarkably smaller than commonly used numbers of 12 and 9.

To avoid the limitation of the database, the Call-Friend database will be used in future research. Different feature combination methods and the relationship among a larger number of different features will also be concentrated on in future.

7. References

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

[2] N. S. Yasunari Obuchi, "Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization," presented at ICASSP, 2005.

[3] H.-C. W. Chi-Yueh Lin, "Language Identification Using Pitch Contour Information," presented at ICASSP, 2005.

[4] B. Milner, "A comparison of front-end configurations for robust speech recognition," presented at Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP), 2002.

[5] J. P. Openshaw, Z. P. Sun, and J. S. Mason, "A comparison of composite features under degraded speech in speaker recognition," presented at Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP), 1993.

[6] A. Cutler, D. Dahan, and W. v. Donselaar, "Prosody in the Comprehension of Spoken Language: A Literature Review," *Language and Speech*, vol. 40, pp. 141-201, 1997.

[7] J. Mariethoz and S. Bengio, "An Alternative to Silence Removal for Text-Independent Speaker Verification," IDIAP, Research Report IDIAP-RR 03-51, December 19, 2003 2003.

[8] F. Allen, E. Ambikairajah, and J. Epps, "Language Identification Using Warping and the Shifted Delta Cepstrum," presented at International Workshop on Multimedia Signal Processing (IEEE MMSP'05), Shanghai, 2005.

[9] F. Allen, "Automatic Language Identification," Thesis, University of New South Wales, Sydney, Australia, 2005.

[10] J. Gutierrez, J. L. Rouas, and R. Andre-Obrecht, "Fusing language identification systems using performance confidence indexes," presented at Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP), 2004.