LANGUAGE IDENTIFICATION USING MODIFIED MLKSFM FOR PRE-CLASSIFICATION WITH NOVEL FRONT-END FEATURES

[1,2]*Liang Wang* [1,2]*Eliathamby Ambikairajah* [2,1]*Eric H.C. Choi*

[1]School of Electrical Engineering and Telecommunications, the University of New South Wales
Sydney, NSW 2052, Australia
[2]ATP Research Laboratory, National ICT Australia
Sydney, NSW 1435, Australia
E-mail: l.wang@student.unsw.edu.au, ambi@ee.unsw.edu.au, eric.choi@nicta.com.au

ABSTRACT

This paper presents two novel contributions to automatic language identification. The first one is the use of the modified multi-layer Kohonen self-organizing feature map (MLKSFM) as a pre-classification for language identification (LID). Secondly, we discuss the novel application of empirical mode decomposition (EMD) to generate features for the LID pre-classification task. The use of instantaneous frequency (IF) and instantaneous amplitude (IA) of a speech signal as features for the pre-classifier is investigated. The experiment results on a 16-language speech database indicates that, the EMD by itself cannot perform well in the LID task, however it helps to improve the pre-classification rate when concatenated with other cepstral features. The overall LID performance is also increased when pre-classification is applied. We achieve LID rates of 85.2% and 62.3% for 45-sec and 10-sec test utterances, respectively.

Index Terms— Language identification, empirical mode decomposition, pre-classification, modified MLKSFM, modified group delay function

1. INTRODUCTION

The goal of automatic language identification (LID) is to identify the language spoken in a particular utterance. Approaches using phonotactic information, such as PPRLM (parallel phoneme recognition followed by language modeling) [1, 2] and PPR followed vector space modeling [3] are well studied. In PPRLM, the input speech utterances are firstly transcribed into phoneme strings by a set of phoneme recognizers. Then the N-gram language model is employed to estimate the probability of the occurrence of a particular phoneme sequence for the final scoring.

The performance of a language identification system depends on two important properties: the ability to extract sufficient information from the speech signal, and the ability to realize complex decision regions in the feature vector space [1, 2, 3]. These two properties refer to the front-end feature and the back-end model of the LID task.

Speech signals are typically non-stationary [4], and consequently Fourier-based signal analysis methods cannot precisely describe the variation of frequency with time. Empirical mode decomposition (EMD) [5, 6] is a new signal decomposition technique for analyzing data from non-linear and non-stationary process. This is usually followed by Hilbert transform (HT) to extract instantaneous amplitude and instantaneous frequency estimates from the decomposed components. Recently proposed practical applications of EMD are broadly spread over numerous scientific disciplines and investigations [6, 7, 8]. However, its efficacy for LID has never been studied. Thus in this paper the possibility of using EMD in LID is discussed.

Regarding the back-end model, previous research indicates that, by using tonal and non-tonal language pre-classification based on pitch information, the final language recognition rate will be increased [9]. So far the best tonal and non-tonal language classification rate is only 87.1% [9], and the lack of an efficient way to model the pitch changing pattern appears to be the primary limitation. Also some languages cannot be clearly classified as either tonal or non-tonal. For example, Japanese is normally classified as a pitch accent language, which is in some sense intermediate between stress-accent languages like English and tonal languages like Chinese. Thus the investigation of an appropriate pre-classification scheme is a challenging problem, especially when many more languages are involved.

Recently, a novel use of the multi-layer Kohonen self-organizing feature map (MLKSFM) [10, 11] is found to provide a promising result for the language identification task. MLKSFM is widely used in classification tasks where a small number of classes are to be classified [12, 13]. By using the MLKSFM, the language identification task is treated as a feature vector quantization problem [10]. In our experimentation with the MLKSFM [10], it is found that the system's performance will be decreased dramatically if a larger number of target languages are used for the evaluation.

Therefore in this paper we propose a novel language identification system by using the modified MLKSFM to perform the pre-classification. All languages are first pre-classified into several classes by the modified MLKSFM system, and then the final identification is performed by the corresponding PPRLM system for each class.

2. INSTANTANEOUS FREQUENCY

The concept of instantaneous amplitude and instantaneous frequency is a powerful idea for the characterization of the non-linear and non-stationary time series data. It can help capture better the time varying characteristics of speech signal that would be otherwise not possible by using other short-time Fourier-based

methods. Recently Huang et al. [5] proposed a technique for extracting instantaneous frequencies from non-stationary signals consisting of two parts: (i) the EMD, which generates a collection of intrinsic mode functions (IMF) from the original data using a sifting process, and (ii) the Hilbert transform which provides information about instantaneous amplitude (IA), instantaneous phase and instantaneous frequency (IF) [5, 6].

2.1. Empirical Mode Decomposition

The EMD is necessary pre-processing of the data before the Hilbert transform can be applied. Using EMD any multi-component signal can be decomposed into a set of IMFs, which can be defined as hidden oscillation modes that are embedded in the data series. They are non-stationary and can be amplitude and/or frequency modulated. According to Huang [5, 6], the IMF is defined as a function satisfying the following conditions: (1) in the whole data set, the number of extrema and the number of zero-crossings must be either equal or differ at most by one; and (2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

The first condition ensures that the local maxima of the data are always positive and the local minima are always negative. The second condition guarantees the physically meaningful of the instantaneous frequency such that the IF will not have the unwanted fluctuations induced by asymmetric wave forms [5, 6]. For an arbitrary signal $X(t)$, EMD is performed and the signal expressed as a sum of IMFs and a residual trend as follows [5, 6]:

$$X(t) = \sum_{j=1}^{n} c_j(t) + r_n(t) \tag{1}$$

Where $c_j(t)$ is the j-th IMF of the signal and $r_n(t)$ is the residual trend. The completeness and orthogonality of IMFs are shown by Huang [5]. It should be noted that, as the order of the mode increases, the time scale increases while the mean frequency of the mode decreases.

In the case of the speech processing, it was observed that most of the speech information is contained in the first five IMFs, and the difference between the original signal and the signal reconstructed by the first five IMFs is not very significant. Hence only the first five IMFs are used in this research.

2.2. Instantaneous Frequency and Amplitude

With the IMF components in hand, one should have no difficulty in applying the Hilbert transform to each IMF component. Applying HT to every IMF component $c_j(t)$, we have a new data series $y_j(t)$ in the transform domain:

$$y_j(t) = \frac{1}{\pi} P \int \frac{c_j(\tau)}{t-\tau} d\tau \tag{2}$$

where P indicates the Cauchy principle value. Then an analytic signal $z_j(t)$ can be obtained as:

$$z_j(t) = c_j(t) + i y_j(t) = a_j(t) e^{i\theta_j(t)} \tag{3}$$

where $a_j(t)$ is the instantaneous amplitude (IA) and $\theta_j(t)$ is the instantaneous phase.

$$a_j(t) = \sqrt{c_j^{\,2}(t) + y_j^{\,2}(t)} \tag{4}$$

$$\theta_j(t) = \arctan \frac{y_j(t)}{c_j(t)} \tag{5}$$

Also the instantaneous frequency (IF) is defined as:

$$\omega_j(t) = \frac{d\theta_j(t)}{dt} \tag{6}$$

As mentioned previously, only the first five IMFs is used in the experiment, so five instantaneous amplitudes and five instantaneous frequencies can be obtained for each frame of the speech signal.

To compactly represent the instantaneous amplitude and instantaneous frequency, the discrete cosine transform (DCT) is applied. The first ten DCT coefficients are used in this experiment, along with the delta and acceleration parameters for the DCT of IA and IF, defined in a manner similar to that of the delta and acceleration parameters of MFCC.

In order to concatenate instantaneous amplitude and instantaneous frequency features to the other cepstral features such as MFCC and the modified group delay function (MODGDF) [11], identical time windows are used to extract IA and IF.

3. LID SYSTEM WITH MLKSFM BASED PRE-CLASSIFICATION

Fig. 1 shows the proposed LID system. All the languages are first pre-classified into 3 classes by the MLKSFM. Then the final identification is performed using different PPRLM for each class.

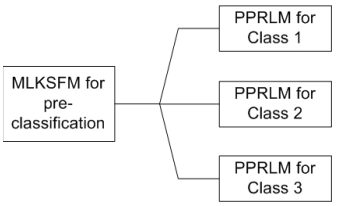

Fig. 1. The structure of LID system with MLKSFM for pre-classification

3.1. Pre-classification with Modified MLKSFM

A good introduction of MLKSFM can be found in [10], where a 3-layer KSFM was used as the classification model, but without any explicit pre-classification, for the language identification.

For the pre-classification here, each neural unit in the competitive layer is to be labeled firstly. Then according to similarities observed from the top competitive layer, all the languages are manually grouped into several classes. Therefore each neural unit in the competitive layer is sensitive to only a certain class of language. It should be noted that, a modified weighted voting function is used in this work, as opposed to the

simple voting function used in [10]. A description of the modification is as follows.

During evaluation, for each unknown testing utterance $\boldsymbol{L}$, the set of feature vectors are first calculated as $\boldsymbol{L = \{ l_1, l_2, ..., l_t, ..., l_T \}}$. For each feature vector $\boldsymbol{l_t}$ the best match is found from each of the neural units in the competitive layer with a corresponding weight vector $\underline{\boldsymbol{w}}_{\boldsymbol{p}}$. The label $\boldsymbol{i}$ in $\underline{\boldsymbol{w}}_{\boldsymbol{p}}$ is added to the corresponding feature vector $\boldsymbol{l_t}$, where $\boldsymbol{i \in \Lambda, \Lambda = \{ 1, 2, \ldots, M \}}$ and $\boldsymbol{M}$ is the number of classes. The pre-classification is performed as:

$$\Phi(L) = \arg\max N(i \mid L), \quad \mathbf{i} \in \mathbf{\Lambda} \tag{7}$$

where

$$N(i \mid L) = \sum_{t=1}^{T} \| l_t - \underline{w}_p \| * \tau(l_t \in i) \tag{8}$$

is the weighted votes for each classes **i, i** $\in \mathbf{\Lambda}$, and

$$\tau(l_t \in i) = \begin{cases} 1, \text{if the unit for language } i \text{ is the best matched} \\ 0, \text{otherwise} \end{cases} \tag{9}$$

3.2. Language Identification with Modified MLKSFM Pre-classification

Once the input speech is pre-classified, it is fed into the corresponding PPRLM [1, 2] for the final identification. The PPRLM systems use only the 13 MFCCs plus its delta and 13 MODGDF as the feature vectors, which results in a 39-dimension feature vector. In the phone recognizers of the PPRLM, each phone symbol is modeled by a 3-state HMM and each state distribution is modeled by 6 Gaussians; Witten-Bell discounting method is used in the language model of the PPRLM.

4. EXPERIMENT RESULTS AND DISCUSSION

4.1. Corpora Description

The data sources for this experiment are the multi-language CALLFRIEND corpus, the OGI-TS corpus and the OGI 22-language corpus [14]. NIST LRE databases are not used because we want to evaluate our system on data sets that contains more languages. Table 1 lists the languages used in the experiments. For evaluation, 45-sec and 10-sec utterances are used. All the evaluation utterances are unseen in training.

Table 1. The languages being used for this LID task

Arabic	*Cantonese*	*English*	*Farsi*
French	*German*	*Hindi*	*Japanese*
Korean	*Malay*	*Mandarin*	*Russian*
Spanish	*Swedish*	*Tamil*	*Vietnamese*

4.2. Experiment Results and Discussion

For the pre-classification using the modified MLKSFM, the map size for the first, second and third competitive layer are chosen as 75x45, 22x15 and 8x6 respectively. As mentioned earlier, a label is first added to each neural unit which is actually accomplished by performing the language identification. The top competitive layer after the labeling is shown in Fig. 2.

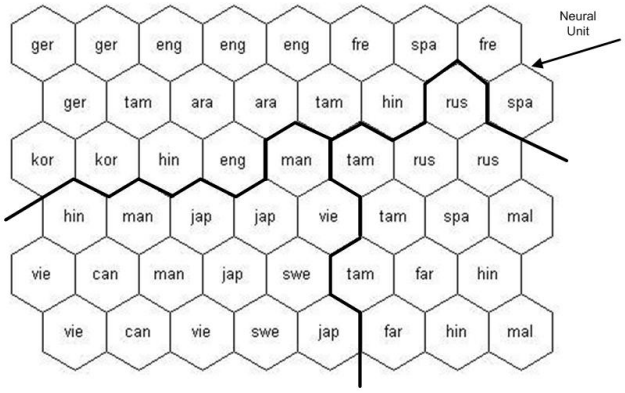

Fig. 2. The top competitive layer after the labeling

Based on the topology of the labeling and the number of target languages, all 16 languages are grouped into 3 classes manually (shown in Table 2).

Table 2. Three classes for the pre-classification

Class 1	Ara	Eng	Fre	Ger	Kor	Spa
Class 2	Can	Jap	Man	Swe	Vie	
Class 3	Far	Hin	Mal	Rus	Tam	

As discussed in Section 2 and also based on previous research, feature vectors for each frame (50ms width and 25ms overlap) in the pre-classification consists of: 13 MFCCs, 13 MODGDF, 10 IA and 10 IF, plus their corresponding delta and acceleration features, where applicable. Due to the limitation of the hardware, it is not feasible to concatenate MFCC, MODGDF, IF and IA with all the delta and acceleration features, but still some promising results are obtained. The overall pre-classification rates with different types of input feature vectors are listed in Table 3:

Table 3. Results for the pre-classification ('D' and 'A' refers to the delta and acceleration features, respectively)

	45-sec	*10-sec*
MFCC+D+A	89.1%	65.2%
MODGDF+D+A	84.2%	62.4%
(IF+D+A)+(IA+D+A)	66.7%	50.3%
(MFCC+D)+MODGDF	92.1%	67.6%
(MFCC+D)+IF+IA	89.4%	66.4%
(MFCC+D)+MODGDF+IF+IA	**93.0%**	**75.2%**
Pre-classification with un-weighted voting, (MFCC+D)+MODGDF+IF+IA	89.7%	71.2%

As no previous result for the 3 classes pre-classification is available, the result of 87.1% for the tonal and non-tonal language classification for 30-sec utterances in [9] can be used as a reference. Our results indicate that, the MODGDF+D+A by itself is able to provide a comparable performance compared with MFCC+D+A features. When combining the MFCC+D and

MODGDF features, we obtained the classification rates of 92.1% and 67.6% for 45-sec and 10-sec utterances, respectively. The IF+D+A and IA+D+A features by themselves only provide the classification rates of 66.7% and 50.3% for 45-sec and 10-sec utterances, however when only the IF and IA are combined with the MFCC+D and MODGDF features, the classification rates of 93% and 75.2% for 45-sec and 10-sec utterances can be obtained.

To prove the effectiveness of the modified MLKSFM, the pre-classification by using the standard un-weighted voting MLKSFM [10] with the MFCC plus its delta, and MODGDF together with IF and IA is also performed. The classification rates are 89.7% and 71.2% for 45-sec and 10-sec utterances, respectively. This indicates the 3.7% and 5.6% relative improvements by using the modified MLKSFM with the same feature vectors.

Table 4. Results for the final language identification for 16 languages

	45-sec	*10-sec*
PPRLM LID system	78.7%	55.8%
MLKSFM LID system	82.3%	56.7%
PPRLM with MLKSFM for pre-classification	**85.2%**	**62.3%**

Table 4 shows the final language identification rates. The standard PPRLM system and the MLKSFM system (without pre-classification) [10] are used as the baseline systems. For the baseline systems, the 13 MFCCs plus their delta features and 13 MODGDF are used as the feature vectors. For the novel LID system with the PPRLM and the modified MLKSFM for pre-classification, the feature vectors for the pre-classification system are the 13 MFCCs plus the delta features, 13 MODGDF, 10 IF and 10 IA; while the feature vectors for the PPRLM system are the same features used in baseline systems. The results indicate that, for both the 45-sec and 10-sec utterances, the PPRLM with modified MLKSFM for pre-classification outperforms the baseline systems. Compared with both the baseline systems, the novel MLKSFM pre-classification based LID system provides at least 3.5% and 9.9% relative improvements for 45-sec and 10-sec utterances respectively.

5. CONCLUSION

Empirical mode decomposition is a new and powerful method for analyzing non-linear and non-stationary time series. While the best ways to use it appropriately in LID task and in general speech processing are still unclear, we have demonstrated its use along with that of the modified MLKSFM pre-classification system for language identification. Though instantaneous frequencies and amplitudes by themselves cannot perform well in the LID task, when represented compactly using discrete cosine transform and concatenated with cepstral features, they help improve the LID rate.

Also pre-classification appears to improve the LID rate when handling larger amount of target languages, and work on identifying the optimal number of classes for the pre-classification and the criterion for setting up the classes is still in progress.

6. REFERENCES

[1] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 31-44, 1996.

[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*, pp. 1345-1348, 2003.

[3] H. Li, B. Ma and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Speech and Audio Processing*, vol. 15, no. 1, pp. 271-284, 2007.

[4] X. D. Huang, A. Alex and H. W. Hon, "Spoken language processing: a guide to theory, algorithm and system development," Prentice Hall, 2001.

[5] N. E. Huang, Z. Shen, S. R. Long, M. L. Wu, H. H. Shin, Q. Zheng, N. C. Yen, C. C. Tung and H. H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proc. Roy. Soc. London* A, vol. 454, pp. 903-995, 1998.

[6] N. E. Huang, M. L. Wu, W. Qu and S. R. Long, "Applications of Hilbert-Huang transform to non-stationary financial time series analysis," in *Appl. Stochastic Models Bus. Ind.*, vol. 19, pp. 245-268, 2003.

[7] S. Q. Tan, J. W. Huang, Z. H. Yang and Y. Q. Shi, "Steganalysis of JPEG2000 lazy-mode steganography using the Hilbert-Huang transform, " in *Proc. IEEE ICIP*, pp. 101-104, 2006.

[8] W. Wang, X. Y. Li and R. B. Zhang, "Speech detection based on Hilbert-Huang transform," in *Proc. IEEE IMSCCS*, vol. 1, pp. 290-293, 2006.

[9] L. Wang, E. Ambikairajah and E. H.C. Choi, "Automatic language recognition with tonal and non-tonal language pre-classification," in *Proc. Eusipco*, pp. 2375-2379, 2007.

[10] L. Wang, E. Ambikairajah and E. H.C. Choi, "Multi-layer Kohonen self-organizing feature map for language identification," in *Proc. Interspeech*, pp. 174-177, 2007.

[11] L. Wang, E. Ambikairajah and E. H.C. Choi, "A comparisonal study of the multi-layer Kohonen self-organizing feature maps for spoken language identification," in *Proc. IEEE ASRU*, accepted and to appear.

[12] J. Koh, M. Suk and S. M. Bhandarkar, "A multi-layer Kohonen's self-organizing feature map for range image segmentation" in *Proc. IEEE ICNN*, vol. 3, pp. 1270-1275, 1993.

[13] A. Tomczyk, P. S. Szczepaniak and B. Lis, "Generalized multi-layer Kohonen network and its application to texture recognition," in *Proc. ICAISC*, LNCS, vol. 3070, pp. 760-767, 2004.

[14] Linguistic Data Consortium. http://www.ldc.upeen.edu