LANGUAGE IDENTIFICATION USING PITCH CONTOUR INFORMATION IN THE ERGODIC MARKOV MODEL

Chi-Yueh Lin, Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
d913920@oz.nthu.edu.tw, hcwang@ee.nthu.edu.tw

ABSTRACT

It had been shown that a segment of pitch contour represented by a set of Legendre polynomial coefficients was successful to the pair-wise language identification task. Feature vectors comprising these polynomial coefficients were formerly modeled by a Gaussian mixture model (GMM) for each language. However, the static model like GMM does not take advantage of the temporal information across several pitch contours. It is intuitive that the temporal information of prosodic features should be used for capturing the characteristics of a specific language. In this paper, a novel dynamic model in ergodic topology is proposed. The experiments show that the proposed method significantly improves the identification rate, even for stress-timed and syllable-timed languages.

1. INTRODUCTION

The automatic language identification (LID) is a process by which the language of a digitized speech utterance is recognized by a computer. Over the past decades, many approaches have been proposed to deal with the LID task [1][2][3]. They tried to capture the specific characteristics of each language. These characteristics roughly fall into three categories : the phonetic repertoire, the phonotactics, and the prosody. So far the most successful system is based on the phonotactics. However, the knowledge of phonotactics of a particular language can not be utilized without a linguistic expert. Moreover, manually labeling of speech data in the preparation is also a time-consuming task. The system based on phonetic repertoire utilizes the statistics of phone frequencies of occurrence. Many languages share a common subset of phones, but the frequency of occurrence of a common phone may differ among these languages. This idea was used in Muthusamy's [4] and Hazen's [5] LID systems. Prosody-based LID systems capture the duration, the pitch pattern, and the stress pattern in a language. LID systems based on prosody properties so far perform worse than those based on phonotactics or phonetic repertoire. The reason is the lack of efficient way to model these prosody characteristics. Therefore LID task based on prosody properties is still a challenging problem.

In this paper, we focus on the utilization of pitch information to LID task. Very few papers deal with the method using pitch information. Cummins [6] used Long Short-Term Memory (LSTM) model by applying differenced log-F0 and amplitude envelope information. He concluded that the better performance could be achieved by using F0 information only. His conclusion was also correspondent with Thymé-Gobbel's [7] work. Rouas [8] used fourth order statistics of pitch information combining with rhythmic parameters for LID task. A GMM method using Legendre polynomial coefficients of pitch contour has shown some success to LID task, especially for languages with special pitch patterns, like pitch-accent languages and tonal languages [9].

It is intuitive that the temporal information plays an important role while considering prosodic properties. Static modeling may not fully take advantages of such information as time progresses. Here we proposed a dynamic model in ergodic topology so that it can utilize more information as time proceeds. Experiment results show that the performance of language identification has been improved significantly by using this new model. Besides pitch-accent and tonal languages, stress-timed and syllable-timed languages are also benefited by the proposed model.

In the following sections, we first explain the pre-processing procedures such as the pitch contour extraction, the pitch contour segmentation, and the pitch contour representation. Then the proposed dynamic model will be introduced. The static model will also be revisited. Experiments are conducted to show the effectiveness of our proposed method. Finally a discussion is presented.

2. PRE-PROCESSING PROCEDURE

2.1. Pitch Contour Extraction

The pitch contour extraction is mainly with the help of Praat program [10]. The method we adopted is the one proposed by Boersma [11]. This method utilizes the autocorrelation function to detect vocalic segments and find pitch candidates. Then Viterbi algorithm is used to find the

most suitable contour path. Parameter settings used in this paper are the same as those were listed in [9]. The detail description of each parameter is described in Boersma's paper.

In the spontaneous speech, the vocalic portion of speech signal may across syllable or word boundaries. Some extracted pitch contours are somewhat too long. In order to segment those long pitch contours into shorter ones, we further utilize the information from the energy contour. Valley points of energy contour are candidates for contour segmentation. The additional constraint is that each segmented pitch contour should not be less than 50ms.

2.2. Pitch Contour Approximation

For each segmented pitch contour f_t, we approximate it by an *M*-th order Legendre polynomial in the sense of minimum mean square error.

$$\hat{f}_t = \sum_{i=0}^{M} a_{it} P_i \qquad (1)$$

, where *t* is the pitch contour index, *M* is the highest polynomial order, a_{it} is *i*-th order coefficient, and P_i is *i*-th order Legendre polynomial. In most cases, small value of *M* is sufficient. From our previous study, a_{1t} and a_{2t} are the most helpful coefficients to language identification task. Notice that P_0 stands for the height of pitch contour, P_1 stands for the slope of pitch contour, P_2 stands for the curvature of pitch contour, and P_3 stands for the S-curvature of pitch contour. With this representation, a feature vector $\vec{v}_t$ is formed including the length of pitch contour D_t and two coefficients, a_{1t} and a_{2t}.

3. MODEL DESCRIPTION

3.1. Gaussian Mixture Model

Gaussian mixture model (GMM) is one of most well-known static modeling methods and has been successfully applied to various engineering problems. In the language identification task, a GMM $\Lambda^{GMM,\ell}$ is created for each language ℓ. Under GMM assumption, the likelihood of a feature vector feature $\vec{v}_t$ is represented by a weighted sum of multi-variant Gaussian density:

$$p(\vec{v}_t \mid \Lambda^{GMM,\ell}) = \sum_{n=1}^{N} w_n^{\ell} \mathrm{N}(\vec{v}_t \mid \mu_n^{\ell}, \Sigma_n^{\ell}) \qquad (2)$$

During the recognition, an unknown speech utterance is represented by a sequence of feature vectors. Then the log-likelihood L_{ℓ}^{GMM} produced by the model $\Lambda^{GMM,\ell}$ is calculated as follows,

$$L_{\ell}^{GMM} = \sum_{t=1}^{T} \log p(\vec{v}_t \mid \Lambda^{GMM,\ell}) \qquad (3)$$

3.2. Dynamic Model in Ergodic Topology

The temporal information of prosodic features is important in capturing the characteristics of a specific language. It is obvious that a static model like GMM can't describe temporal information across several pitch contours. Here a novel dynamic model was proposed to compensate for the weakness of static model.

In brief, the proposed dynamic model $\Lambda^{DM,\ell}$ for each language ℓ is composed of a set of states and a set of transition probabilities. Each state is modeled by a GMM, and transition probabilities are modeled by (i) bigram, (ii) trigram, or (iii) mixture of bigrams. This topology is the same as ergodic Markov model in speech recognition.

In the training phase, we first define a rule $R_D(\cdot)$ to cluster feature vectors $\vec{v}_t$ into six groups according to their duration components D_t. These six groups correspond to six states. The rule is described as follows,

$$S_t = R_D(\vec{v}_t) = \begin{cases} S^{(1)} & \text{if } D_t \in [50ms, 100ms) \\ S^{(2)} & \text{if } D_t \in [100ms, 150ms) \\ S^{(3)} & \text{if } D_t \in [150ms, 200ms) \\ S^{(4)} & \text{if } D_t \in [200ms, 250ms) \\ S^{(5)} & \text{if } D_t \in [250ms, 300ms) \\ S^{(6)} & \text{if } D_t > 300ms \end{cases} \qquad (4)$$

, where S_t is the state. After clustering, each state $S^{(k)}$ is modeled by a GMM $\lambda_{S^{(k)}}^{\ell}$. It should be noted that in this training step, feature vector $\vec{v}_t$ is modified to $\vec{u}_t$ which consists of only two components, a_{1t} and a_{2t}.

Transition probabilities can be modeled by the conventional method like bigram probabilities $p(S_t \mid S_{t-1})$ or trigram probabilities $p(S_t \mid S_{t-1}, S_{t-2})$. In this paper, we also adopt another modeling technique, i.e., the mixture transition distribution model [12]. The main purpose of this method is to approximate the high-order Markov model with mixtures of low-order Markov models. For example, the transition probability matrix of a conventional *N*-th order Markov model is specified by $O(m^{N+1})$ elements, where m is the number of states. With mixture transition distribution model, $O(m^{N+1})$ elements can be reduced to only $O(Nm^2)$ elements. Here we approximate trigram probabilities with mixture of two bigrams as follows.

$$p(S_t \mid S_{t-1}, S_{t-2}) \approx \sum_{n=1}^{2} \beta_n p(S_t \mid S_{t-n}) \qquad (5)$$

, where β_n are mixture weights and $\sum \beta_n = 1$, $0 < \beta_n < 1$. During the recognition, feature vectors $\{\vec{u}_t\}$ are used. The log-likelihood L_{ℓ}^{DM} produced by the model $\Lambda^{DM,\ell}$ is calculated as

$$L_\ell^{DM} = \sum_{t=1}^{T} \log p(\vec{u}_t \mid \Lambda^{DM,\ell}) = \sum_{t=1}^{T} \left[\alpha \log p(\vec{u}_t \mid \lambda_{S_t}^{\ell}) + (1-\alpha) \log TP^{\ell} \right] \quad (6)$$

, where α is a balance factor between the observation log-likelihood score and the transition log-likelihood score, and TP^{ℓ} is the transition probability in one of following three forms:

$$\begin{aligned} & p(S_t \mid S_{t-1}) && \text{Bigram} \\ & p(S_t \mid S_{t-1}, S_{t-2}) && \text{Trigram} \\ & \sum_{n=1}^{2} \beta_n p(S_t \mid S_{t-n}) && \text{Mixture of Bigrams} \end{aligned} \quad (7)$$

Finally, a maximum-likelihood classifier hypothesizes $\hat{\ell}$ as the language of the unknown utterance, where

$$\hat{\ell} = \arg\max_{1 \le \ell \le 2} L_\ell \quad (8)$$

4. EXPERIMENTS

The pair-wise LID experiment was conducted using the Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus [13] which including the following 10 languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. For each language, 50 speakers in TrainSet were used to train both static model and dynamic model. 20 speakers in EvalSet were used to evaluate the system performance. Only 45-sec and 10-sec utterances in EvalSet are chosen for the evaluation. The identification rate is calculated as the number of correctly identified utterances out of all evaluation utterances.

Experimental results are listed from Table 1 to Table 3. In Table 1, each identification rate is averaged from 45 pair-wise LID tasks. Rates in brackets are relative performance improvement with respect to the performance of static model, GMM. Relative performance is defined as ($\text{Rate}_{DM} - \text{Rate}_{GMM}$) / Rate_{GMM} . With dynamic modeling, we achieve significant improvement on 45-sec utterances and somewhat minor improvement on 10-sec utterances. It is also worth noting that transition probabilities modeled by mixture of bigrams achieves the highest identification rate, since this modeling technique benefits from more information across longer history like trigram but avoids insufficient training data problem that the trigram may encounter.

In Table 2, we look into identification rate of each language. Each rate in column 1 and column 2 is averaged from 9 pair-wise LID tasks. It reveals that all 10 languages benefit from the proposed dynamic model, even for stress-timed languages, like English and German, and syllable-timed languages, like French and Spanish. For both 45-sec and 10-sec utterances, German has the highest improvement. On the other hand, Japanese and Mandarin are not benefit much from dynamic model. Japanese even has little degradation on 10-sec utterances.

Results of all 45 pair-wise LID tasks on 10-sec and 45-sec utterances are given in Table 3. All rates listed in the table are derived from dynamic model with mixture of bigrams. Rouas's work on 45-sec utterances is also shown in square brackets. Compare to our results, our proposed dynamic model performs better for almost all pair-wise identification tasks. Only 4 out of 45 pairs perform worse.

Table 1. Comparison of performance of difference models

	45-sec	10-sec
GMM	68.91%	65.45%
Dynamic Model – Bigram	80.23% (16.43%)	69.83% (6.69%)
Dynamic Model – Trigram	79.62% (15.54%)	68.84% (5.18%)
Dynamic Model – MixBigram	**81.35%** (18.05%)	**70.02%** (6.98%)

Table 2. Identification rate of each language

		GMM	Dynamic Model - MixBigram	Relative Improvement
EN-other	45s	67.03%	81.84%	22.09%
	10s	59.31%	65.99%	11.26%
FA- other	45s	74.48%	85.05%	14.18%
	10s	68.05%	70.98%	4.32%
FR- other	45s	61.00%	71.51%	17.23%
	10s	60.00%	66.91%	11.52%
GE- other	45s	63.77%	84.65%	**32.75%**
	10s	61.32%	70.02%	**14.19%**
JA- other	45s	79.10%	86.08%	8.82%
	10s	81.60%	79.48%	-3.83%
KO- other	45s	67.67%	82.71%	22.22%
	10s	64.47%	69.57%	7.91%
MA- other	45s	76.54%	83.41%	8.97%
	10s	71.69%	73.09%	1.94%
SP- other	45s	61.26%	73.31%	19.67%
	10s	58.32%	64.23%	10.13%
TA- other	45s	63.05%	76.75%	21.73%
	10s	61.53%	67.71%	10.05%
VI- other	45s	74.82%	88.21%	17.90%
	10s	68.20%	73.24%	7.39%

5. DISCUSSION

Examine Table 2 and Table 3 in more detail, we can observe that the performances of syllable-timed languages like French and Spanish are not as good as others. Though their relative improvements look good in some extent, the proposed dynamic model may not be good enough for this type of languages. On the other hand, the performance of stress-timed languages, like English and German, boosts a lot. The duration of vocalic portion varies often as time proceeds. This is one of

Table 3. Confusion matrix of pair-wise LID task on 10 languages.

10/45 sec	FA	FR	GE	JA	KO	MA	SP	TA	VI
EN	64.3/**89.7** [76.3]	64.1/**70.3** [51.5]	64.4/**76.9** [59.5]	77.2/**78.9** [67.6]	68.4/**88.9** [79.4]	73.2/**83.8** [75.0]	63.7/**77.8** [67.7]	56.4/**78.8** [77.4]	61.9/**91.4** [67.7]
FA	--	72.0/**86.8** [68.6]	75.5/**95.0** [71.8]	81.8/**94.9** [66.7]	69.3/**75.7** [75.0]	73.4/**84.2** [76.3]	62.0/**70.2** [66.7]	67.5/**88.2** [69.7]	73.1/**80.6** [66.7]
FR	--	--	57.6/**65.8** [55.9]	81.6/**81.1** [55.9]	65.1/**65.8** [54.8]	75.2/**75.0** [60.6]	62.4/57.1 [64.3]	53.8/59.3 [60.1]	70.3/**82.3** [58.1]
GE	--	--	--	79.4/**82.1** [65.8]	70.1/**89.2** [71.4]	73.5/**92.1** [62.2]	65.4/**81.1** [59.4]	72.1/**85.3** [69.7]	72.0/**94.4** [65.7]
JA	--	--	--	--	78.6/**88.9** [65.7]	72.2/**83.8** [54.1]	75.5/**86.1** [62.5]	77.8/**81.8** [59.4]	82.2/**97.1** [68.6]
KO	--	--	--	--	--	73.2/**91.4** [73.5]	57.5/**76.5** [75.9]	71.1/**74.2** [62.1]	72.8/**93.9** [56.2]
MA	--	--	--	--	--	--	65.6/77.1 [80.6]	75.2/**75.0** [74.2]	76.3/**88.2** [50.0]
SP	--	--	--	--	--	--	--	55.4/58.1 [65.4]	70.5/**75.8** [62.1]
TA	--	--	--	--	--	--	--	--	80.0/**90.0** [71.4]

the characteristics of stress-timed languages. It makes stress-timed languages look like "Morse-code" languages. Our dynamic model captures this special property because the transition probabilities are explicitly based on the duration of pitch contours. At last, for those pitch-accent and tonal languages, the higher performance is mainly due to the novel representation of pitch pattern by Legendre polynomials. The consideration of duration changes as time proceeds for this kind of language contributes only a little.

5. ACKNOWLEDGEMENT

This research was sponsored by the National Science Council, Taiwan, under contract number NSC-93-2213-E-007-019.

6. REFERENCES

[1] Y.K. Muthusamy, E. Barnard, and R.A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Mag.*, Vol. 11, no. 4, pp.33-41, 1994.

[2] M.A. 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.

[3] M.A. Zissman, and K.M. Berkling, "Automatic language identification," *Speech Communication*, Vol. 35, pp. 115- 124, 2001.

[4] Y.K. Muthusamy, "A segmental approach to automatic language identification," *PhD. dissertation of Oregon Graduate Institute of Science and Technology*, 1993.

[5] T.J. Hazen, and V.W. Zue, "Segment-based automatic language identification," *Journal of Acoustical Society of America*, Vol. 101, No. 4, pp. 2323-2331, 1997.

[6] F. Cummins, F. Gers, and J. Schmidhuber, "Language identification from prosody without explicit features," in *Proc. EUROSPEECH'99*, Budapest, Hungary, 1999, pp.371-374.

[7] A.E. Thymé-Gobbel, and S.E. Hutchins, "On using prosodic cues in automatic language identification," in *Proc. ICSLP'96*, Philadelphia, USA, Vol. 3, 1996, pp. 1768-1772.

[8] J. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Modeling prosody for language identification on read and spontaneous speech," in *Proc. ICASSP'2003*, Hong Kong, China, Vol. I, 2003, pp. 40-43.

[9] C.Y. Lin, H.C. Wang, "Language identification using pitch information," in *Proc. ICASSP 2005*, Philadelphia, USA, Vol. 1, 2005, pp.601-604.

[10] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org.

[11] P. Boersma, "Accurate short-term analysis of the fundalmental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, University of Amsterdam, pp. 97-110, 1993.

[12] Raftery, A. "A model for high-order Markov chains," *Journal of the Royal Statistical Society B*, *47*, 528–539, 1985.

[13] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika, "The OGI Multilanguage telephone speech corpus," in *Proc. ICSLP'92*, Banff, Alberta, Canada, Vol. 2, 1992, pp.895-898.