LANGUAGE IDENTIFICATION USING PITCH CONTOUR INFORMATION

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

In this paper, an approach to automatic language identification (LID) using pitch contour information is proposed. A segment of pitch contour is approximated by a set of Legendre polynomials so that coefficients of polynomials form a feature vector to represent this pitch contour. The Gaussian Mixture Model (GMM) method based on feature vectors extracted from pitch contours is suggested for the LID. Our experiments show that only two or three coefficients are necessary to obtain reasonably good identification rates. We also find that the length of segmented pitch contour is another useful feature for LID so that it is included to further improve the performance. Pair-wise language identification experiments on OGI-TS corpus show that our proposed approach is a very promising method. Besides, we find that tonal languages and pitch accent languages achieve better performance in our system.

1. INTRODUCTION

Automatic language identification (LID) is the 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 LID task [1][2][3]. They tried to capture the specific characteristics of each language. These characteristics roughly fall into three categories [4]: phonetic repertoire, phonotactics, and 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, manual labeling of speech data in the preparation is also a time-consuming task. System based on phonetic repertoire utilizes the statistics of phone frequencies of occurrence. Many languages share a common subset of phones, but frequency of occurrence of a common phone may differ among these languages. This idea was used in Muthusamy's [5] and Hazen's [6] LID systems. Prosody-based LID systems capture the duration, pitch pattern, and 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.

In this paper, we focus on the utilization of pitch information to do LID task. Very few papers deal with pitch information. Itahashi [7] used 17 parameters derived from pitch contours to achieve high identification rate in a small data experiment. Cummins [8] used Long Short-Term Memory (LSTM) model to utilize differenced log-F_0 and amplitude envelope information. He concluded that better performance could be achieved by using F_0 information only. His conclusion was also correspondent with Thymé-Gobbel's [9] work. Rouas [10] used fourth order statistics of pitch information in combination with rhythmic parameters for LID task. Here we propose a new method that utilizes pitch information on the LID task. A segment of pitch contour is approximated by a set of Legendre polynomials so that coefficients of polynomials form a feature vector to represent this pitch contour. The Gaussian Mixture Model (GMM) method based on feature vectors extracted from pitch contours is suggested for the LID. The length of segmented pitch contour is another useful feature for LID so that it can be included to further improve the performance.

2. DESCRIPTION OF THE SYSTEM

The block diagram of our proposed LID system is shown in Figure 1.

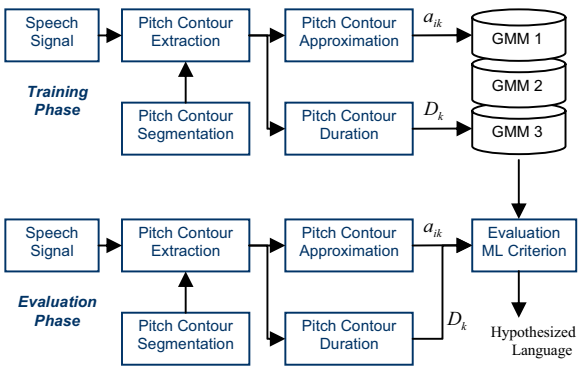

Figure 1. Proposed LID system architecture

In the training phase, the pitch extraction method proposed by Boersma [11] is applied to find the pitch contour. The advantage of this method is the embedded vocalic segment detection. After the extraction, pitch contours with long duration are further segmented into shorter ones. Then each segmented pitch contour is approximated by a Legendre polynomial. The shape of a pitch contour is captured by a set of polynomial coefficients of this representation. These coefficients, together with the length of segmented pitch contour, are used to form a feature vector. These feature vectors are then used for training a Gaussian mixture model for each language. In the evaluation phase, the similar procedure is performed to obtain the feature vectors. During pair-wise language identification, log-likelihood score are calculated for each language model. Then the one with higher score is the hypothesized language. In the following sections, each block will be described in detail.

2.1. Pitch contour extraction

Pitch contour extraction is mainly with help of Praat program[12]. The method we adopted is the one proposed by Boersma . This method utilizes autocorrelation function to detect vocalic segments and find pitch candidates. Then Viterbi algorithm is used to find the most suitable contour path. Some related parameter settings used in this paper are listed in Table 1. Detail description of each parameter is described in Boersma's paper.

Table 1. Pitch extraction parameter settings in Praat program.

Pitch extraction parameter settings	
Analysis window length	30 ms
Analysis window time step	10 ms
Pitch floor (Hz)	50
Pitch Ceiling (Hz)	500
Max. number of candidates	5
Silence threshold	0.03
Voicing threshold	0.6
Octave cost	0.01
Octave-jump cost	0.6
Voiced/Unvoiced cost	0.14

2.2. Pitch contour segmentation

Due to the spontaneous speech, 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 utilize the information from energy contour. First, we align pitch contour with energy contour as shown in Figure 2. The candidates of endpoints of a segment are those valley points of energy contours. Notice that the duration constraint must be set in order to avoid making a segment too short. Duration constraint used here is 50 ms. As being shown in Figure 2, two points are selected as candidates. Only the second candidate is chosen to be the segmentation point. The first candidate will make the first segment too short (less than 50ms). Therefore we ignore the first candidate.

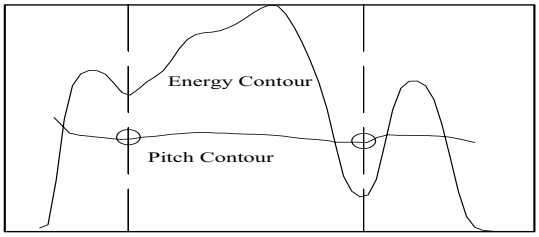

Figure 2. Pitch contour segmentation

2.3. Pitch contour approximation

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

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

where k is the pitch contour index, M is the highest polynomial order, a_{ik} is i-th order coefficient, and P_i is i-th order Legendre polynomial. In most cases, small value of M is sufficient so that we let M=3 here. Legendre polynomials P_i are illustrated in Figure 3.

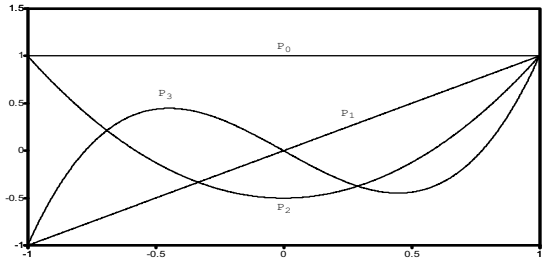

Figure 3. Illustration of Legendre polynomials

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 $\vec{v}_k$ is formed including the length of pitch contour D_k and four coefficients $a_{0k}, a_{1k}, a_{2k}, a_{3k}$. Further we will see that not all features are helpful.

2.4. Gaussian mixture model and evaluation

For each language ℓ, a GMM λ_ℓ is created. Under GMM assumption, the likelihood of a feature vector $\vec{v}_k$ extracted from model λ_ℓ is represented by a weighted

sum of multi-variant Gaussian densities:

$$p(\vec{v}_k|\lambda_\ell) = \sum_{i=1}^{N} w_i \cdot b_i(\vec{v}_k) \quad (2)$$

where $b_i(\vec{v}_k)$ are the component mixture densities and w_i are the mixture weights. The language model λ_ℓ is expressed by

$$\lambda_\ell = \{w_i, \mu_i, \Sigma_i\} \quad (3)$$

i is the mixture index.

During recognition, an unknown speech utterance is represented by a sequence of feature vectors. Then log-likelihood that language model λ_ℓ produced is calculated. The log-likelihood, L_ℓ, is defined as

$$L_\ell = \sum_{k=1}^{K} \log p(\vec{v}_k|\lambda_\ell) \quad (4)$$

where k is the pitch contour index, K is the total number of pitch contours in a utterance. Finally, a maximum-likelihood classifier hypothesizes $\hat{\ell}$ as the language of the unknown utterance, where

$$\hat{\ell} = \arg\max_{1\le\ell\le 2} L_\ell \quad (5)$$

3. EXPERIMENTS

The pair-wise LID experiment was performed using the Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus [13] which included 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 GMMs and 20 speakers in EvalSet were used to evaluate system performance. DevSet was not used here.

In the initial experiment, we try different combination of features to find out the most efficient features. Initial experiment was evaluated on 3-sec utterances with different mixture numbers. The results were shown in Table 2. ΔE in Table 2 means differenced log-energy calculated on the same segment by which pitch contour is extracted. It is interesting that not all features extracted are useful. Some features even degrade the overall performance. As Table 2 shown, useful features are D_k, a_{1k}, and a_{2k}. Remember that a_{1k} stands for the slope of pitch contour and a_{2k} stands for the curvature of pitch contour. Therefore we can concluded that pitch shape can be utilized efficiently with these 2 coefficients. In the following pair-wise experiments, we use 64-component GMM with $\{D_k, a_{1k}, a_{2k}\}$ as feature vectors.

Results of pair-wise LID task on five languages used in Cummins' paper are shown in Table 3. Evaluations are made using the 10-sec and 45-sec utterances. It is clear that our performance is superior to Cummins' work [8]. Evaluations of all 45 pair-wise LID tasks are given in Table 4. Results in the cell are made using 3-sec, 10-sec, and 45-sec utterances, respectively. Rouas' work on 45-sec utterances is also shown in square brackets. Compare to our results, our proposed method performs better on tonal languages and pitch accent languages, but a little bit worse on other cases.

Table 2. Result of initial experiment on 3-sec utterances. (Identification rate is averaged over 45 pair-wise experiments.)

Feature\Mix #	**4**	**8**	**16**	**32**	**64**	**128**
$a_0a_1a_2a_3$	44.3	44.6	42.9	42.5	42.9	43.1
$a_1a_2a_3$	45.2	56.6	56.9	56.2	53.1	54.1
a_1a_2	51.7	58.3	59.8	58.3	58.5	56.8
Da_1a_2	52.4	59.1	58.0	61.0	**62.9**	62.0
$Da_1a_2\Delta E$	57.5	59.4	56.1	59.8	59.1	60.1

Table 3. Confusion matrix of pair-wise LID task on five languages. (Results of Cummins' work using LSTM model are given in square brackets.)

10-sec	Ge	Sp	Ja	Ma
En	**61** [56]	47 [50]	**85** [63]	**74** [63]
Ge	—	47 [54]	**85** [69]	**71** [69]
Sp	—	—	**77** [60]	**66** [62]
Ja	—	—	—	**72** [50]

45-sec	Ge	Sp	Ja	Ma
En	**56** [55]	**53** [52]	**84** [62]	**76** [62]
Ge	—	49 [54]	**77** [72]	**84** [70]
Sp	—	—	**81** [71]	**71** [63]
Ja	—	—	—	**78** [44]

4. DISCUSSION

According to Ladefoged's book [14], languages can be classified into some categories by their rhythmic characteristics. English and German are called stress-timed languages. French and Spanish are called syllable-timed languages. Mandarin and Vietnamese are classified into tonal languages. Japanese belongs to a special category called mora-timed language. Tonal languages perform well in our system. The reason is pitch variation of this kind of language can be captured well with our method. High performance of Japanese and Farsi are due to their pitch accent pattern. Japanese is the most distinguishable one among these 10 languages. With further investigation, every mora in Japanese can either be pronounced with a high (H) or low (L) pitch. Pitch accent patterns in Japanese are always HL(L…) , LHL(L…), and LH(H…)L. This special pitch accent structure marks a downstep property which arises to the large minus value of a_2 (see Figure 4). In contrast, stress-timed and syllable-timed languages perform somewhat poor in our system. That means the linguistic information of this kind of languages is not conveyed within pitch contour.

3/10/45sec	Ge	Fr	Sp	Ma	Vi	Ja	Ko	Ta	Fa
En	49/61/56 [60]	48/44/**54** [52]	47/47/53 [68]	58/74/**76** [75]	62/58/**80** [68]	62/85/**84** [68]	58/59/75 [79]	53/53/64 [77]	39/54/62 [76]
Ge	—	44/44/42 [56]	53/47/49 [59]	58/72/**84** [62]	62/59/**69** [66]	63/85/**77** [66]	59/61/65 [71]	53/58/59 [70]	40/64/**73** [72]
Fr	—	—	51/57/57 [64]	64/69/**69** [61]	58/64/**76** [58]	62/81/**65** [56]	69/60/54 [55]	59/45/44 [60]	54/74/**87** [69]
Sp	—	—	—	66/66/71 [81]	65/65/61 [62]	64/77/**81** [63]	63/57/59 [76]	50/48/48 [65]	36/62/**73** [67]
Ma	—	—	—	—	64/75/**79** [50]	71/72/**78** [50]	66/67/**80** [74]	47/71/69 [74]	52/77/**82** [76]
Vi	—	—	—	—	—	77/85/**89** [69]	65/70/**73** [56]	57/69/**77** [71]	47/71/**69** [67]
Ja	—	—	—	—	—	—	61/80/**75** [66]	68/84/**79** [59]	55/85/**85** [67]
Ko	—	—	—	—	—	—	—	56/61/58 [62]	47/65/70 [75]
Ta	—	—	—	—	—	—	—	—	46/61/**71** [70]

Table 4. Confusion matrix of pair-wise LID task on ten languages. Rouas' work on 45-sec utterances is given in square brackets.

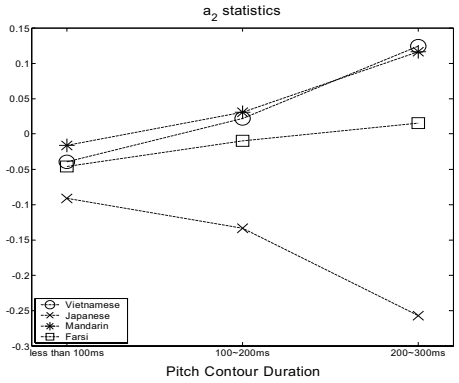

Figure 4. Mean value of a_2 in four languages.

5. ACKNOWLEDGEMENT

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-92-2213-E-007-036.

6. REFERENCES

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

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

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

[4] F. Ramus, and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," in *Journal of Acoustical Society of America*, January 1999, Vol. 105, pp. 510-521.

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

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

[7] S. Itahashi, J.X. Zhou, and K. Tanaka, "Spoken language discrimination using speech fundamental frequency," in *Proc. of ICSLP'94*, Yokohama, Japan, 1994, pp.1899-1902.

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

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

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

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

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

[14] P. Ledefoged, "A course in phonetics, 2nd edition," Harcourt Brace Jovanovich, Inc., 1982.