Fusion of phonotactic and prosodic knowledge for language identification

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

Over the last few decades, language identification systems based on different kinds of linguistic knowledge had been studied by many researchers. Most of systems utilize one kind of linguistic knowledge only, i.e. phonotactic, phonetic repertoire, or prosody. It is possible to get the improvement by combining several linguistic knowledge. However, the combination of two systems based on different kinds of linguistic knowledge is not a trivial task. This paper presents a method where local identification results made by two individual systems, i.e. prosody-based and phonotactic-based systems, are fused in a Bayesian framework. Under this framework, local decisions, the associated false-alarm and miss probabilities are fused via Bayesian formulation to make the final decision. Experiments conducted on OGI-TS corpus demonstrate the effectiveness of this decision-level fusion strategy.

Index Terms:language identification, Bayesian formulation, fusion.

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. 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. The system based on phonetic repertoire utilizes the statistics of phone frequencies of occurrence. Many languages may share a common subset of phones, but the frequency of occurrence of a common phone may differ among these languages. The system based on phonotactics demonstrates the best language discrimination so far. The state-of-the-art systems like phone recognition followed by language modeling (PRLM) and its extension, parallel-PRLM (PPRLM), belong to this category. Prosody-based LID systems, on the other hand, capture the duration, the pitch pattern, and the stress pattern in a language.

It is believed that human beings use several different kinds of information to identify a language. For example, adults can use their phonotactic knowledge to discriminate languages. On the other hand, infants, who surely have no phonotactic knowledge at all, can use prosodic information to discriminate languages [1]. In order to further improve the discriminative ability, different linguistic knowledge should be utilized at the same time. Not many works deal with the language identification via the combination of different linguistic knowledge. Hazen [2] utilized phonotactic, acoustic-phonetic, and prosodic information within a unified probabilistic framework. Gutiérrez [3] calculated the performance confidence indexes deriving from each LID system and applied theory of evidence to do the fusion process. Moreover, Obuchi [4] combined the scores derived from phonetic HMM and prosody HMM with linear discriminant analysis. In this paper, two LID systems using different kinds of linguistic knowledge are implemented. The prosody-based system mainly uses the information extracted from pitch contours, and the PRLM system employs the phonotactic information. Then the Bayesian formulation is adopted to combine these two systems.

The remainder of this paper is organized as follows. In section 2, the prosody-based system using information extracted from pitch contours is introduced. In section 3, the state-of-the-art phonotactic-based system, phone recognition followed by language models (PRLM), will be reviewed. The fusion strategy derived from Bayesian formulation is described in Section 4. Section 5 presents some experiment results and Section 6 gives the conclusion.

2. Prosody-based LID System

2.1. Pitch Contour Extraction and Segmentation

The pitch contour extraction method we adopted here is the one proposed by Boersma [5]. This method utilizes the autocorrelation function to detect vocalic segments and find pitch candidates. Then dynamic programming is used to find the most possible path. However, the vocalic portion of speech signal may cross syllable or word boundaries in spontaneous speech. Some extracted pitch contours are too long and should be segmented. To do so, we utilize information from energy contour. First, we align the pitch contour with energy contour. The candidates of boundaries are those aligned with valleys on energy contour. The duration constraint is also set in order to avoid making segments too short. Duration constraint used here is 50ms. Detail of the above procedure was described in [6]

2.2. Pitch Contour Approximation

Different from representing a F_0 contour by polygonal lines, we suggest using Legendre polynomials to approximate F_0 contour instead. For a given F_0 contour f_k, we scale it to the interval $[-1, 1]$. Then the scaled version, $\hat{f}_k$, is approximated by first $M+1$ Legendre polynomials over this interval in the sense of minimum mean square error.

$$\hat{f}_k \approx \tilde{f}_k = \sum_{i=0}^{M} a_{ik} P_i \tag{1}$$

, where k is the index of F_0 contour, M is the polynomial order, a_{ik} is the coefficient corresponding to i-th order Legendre polynomial for k-th F_0 contour, and P_i is i-th order Legendre polynomial. The reason for scaling is that Legendre polynomials demonstrate the orthogonality relationship over $[-1, 1]$. Then coefficients a_{ik}, $0 \le i \le M$, can be easily derived. From our previous study, three parameters are good enough for capturing the language characteristics [6, 7]. These parameters are coefficients of first- and second-order Legendre polynomials, says a_{1k} and a_{2k}, and duration of F_0 contour, D_k. Therefore a feature vector $\vec{v}_k = [D_k, a_{1k}, a_{2k}]^T$ is formed for further manipulation. Notice that in geometry point of view, a_{1k} stands for the slope of F_0 contour and a_{2k} stands for the curvature of F_0 contour.

2.3. Model Description

A dynamic model in ergodic topology is proposed to model these information extracted from F_0 contours. In brief, the proposed dynamic model $\Lambda^{EMM,l}$ for each language l is composed of a set of states and a set of transition probabilities. Each state is modeled by a Gaussian mixture model, and transition probabilities are modeled by mixture of bigrams. This topology is the same as ergodic Markov model in speech recognition. In the training phase, feature vectors $\vec{v}_k$ for language l, denoted by $\vec{v}_k^l$ afterward, are pooled together and then clustered to N groups according to their duration component D_k^ls. These N groups corresponds to N states in the ergodic Markov model. Here we set N to be six and a global clustering criteria $R_D(\cdot)$ is defined as follows,

$$S_k = R_D(\vec{v}_k^l) = \begin{cases} S^{(1)} & \text{if } D_k^l \in [50ms, 100ms) \\ S^{(2)} & \text{if } D_k^l \in [100ms, 150ms) \\ S^{(3)} & \text{if } D_k^l \in [150ms, 200ms) \\ S^{(4)} & \text{if } D_k^l \in [200ms, 250ms) \\ S^{(5)} & \text{if } D_k^l \in [250ms, 300ms) \\ S^{(6)} & \text{if } D_k^l \in [300ms, \infty) \end{cases} \tag{2}$$

where S_k denotes the state which $\vec{v}_k^l$ belongs to. After clustering, feature vector $\vec{v}_k^l$ will be modified to $\vec{u}_k^l = [a_{1k}^l, a_{2k}^l]^T$ (D_k^l is removed). Then modified feature vectors belonging to the same state are used to train the observation probability for that state. As being mentioned above, the observation probability for each state $j, j = 1, 2, \cdots, 6$, is modeled by a Gaussian mixture model $\Gamma_{S(j)}^l$.

Transition probabilities between each state are modeled by mixture transition distribution model [8]. The main purpose of this method is to approximate a high-order Markov model with several low-order Markov models. In our case, a trigram probability can be approximated by sum of two bigrams as follows.

$$p(S_k|S_{k-1}, S_{k-2}) \approx \sum_{n=1}^{2} \beta_n p(S_k|S_{k-n}) \tag{3}$$

, where β_n is the mixture weight for bigram $p(S_k|S_{k-n})$ with constraint $\sum \beta_n = 1, 0 < \beta_n < 1$, and S_k denotes the state which $\vec{u}_k$ belongs to. Each bigram probability is estimated by maximum likelihood estimation,

$$p(S_k|S_{k-n}) = \frac{\text{\# of transitions emitted from } S_{k-n} \text{ to } S_k}{\text{\# of transitions emitted from } S_{k-n}} \tag{4}$$

In the evaluation phase, the log-likelihood $L^{EMM,l}$ of observing $\vec{u}_k$ for a given model $\Lambda^{EMM,l}$ is calculated as

$$\begin{aligned} L^{EMM,l} &= \sum_{k=1}^{K} \log p(\vec{u}_k|\Lambda^{EMM,l}) \\ &= \sum_{k=1}^{K} \left[\alpha \log p(\vec{u}_k|\Gamma_{S_k}^l) + (1-\alpha) \sum_{n=1}^{2} \beta_n p(S_k|S_{k-n}) \right] \end{aligned} \tag{5}$$

, where k is the index of F_0 contour, α is a balance factor between the observation log-likelihood and the transition log-likelihood. A maximum likelihood decision rule is applied to hypothesize $\hat{l}$ the language that input utterance belongs to.

$$\hat{l}^{EMM} = \arg\max_{l} L^{EMM,l} \tag{6}$$

3. Phonotactic-based LID System

A state-of-the-art PRLM [9] (phone recognition followed by language models) is implemented as our phonotactic-based LID system. The front-end English phone recognizer is trained on TIMIT database. 48 context-independent phones listed in Table 1 are selected to tokenize the input utterance , and each phone was modeled a 3-state left-to-right HMM. Notice that the speech data in TIMIT database were originally recorded in microphone quality with 16kHz sampling rate. In order to match the telephone-line condition of OGI-TS database, the techniques of downsampling and cepstral mean subtraction are applied before the training procedure.

Once the training procedure for phone models is completed, training utterances of language l in OGI-TS database are fed into the phone recognizer and tokenized to a sequence of phones. In general, null-grammar was applied during the decoding process. Then those produced sequence of phones are used to estimate the bigram language model for language l, λ_l^{BG}. Also, back off scheme proposed by Katz [10] is further used to smooth language models. During evaluation, a test utterance is fed into the English phone recognizer and then tokenized to a sequence of phones, $W = \{w_1, w_2, \cdots, w_M\}$. The log-likelihood that language model λ_l^{BG} produced while observing W is

$$L^{PRLM,l} = \sum_{m=1}^{M} \log p(w_m|w_{m-1}, \lambda_l^{BG}) \tag{7}$$

The classifier hypothesizes $\hat{l}$ the language of input utterance if $\lambda_{\hat{l}}^{BG}$ produces the highest log-likelihood.

$$\hat{l}^{PRLM} = \arg\max_{l} L^{PRLM,l} \tag{8}$$

4. Bayesian Fusion Formulation

The approach for dealing with multiple systems fusion is one of the most popular topics in the field of distributed detection and estimation [11, 12]. The fusion topology adopted here is depicted

Table 1: 48 CI phone units extracted from TIMIT database

Stops(6)	[b] [d] [g] [p] [t] [k]
Affricates (2)	[jh] [ch]
Fricatives (8)	[s] [sh] [z] [zh] [f] [th] [v] [dh]
Nasals (6)	[m] [n] [ng] [em] [en] [eng]
Semivowels & Glides (6)	[l] [r] [w] [y] [hh] [el]
Vowels (18)	[iy] [ih] [eh] [ey] [ae] [aa] [aw] [ay] [ah] [ao] [oy] [ow] [uh] [uw] [er] [ax] [ix] [axr]
Non-speech (2)	sil, non-phonetic(pau, epi, h#)

in Figure 1. Phenomenon H stands for the observed raw signal, $\mathbf{y}_i$ denotes the parameterized feature vectors for system S_i, u_i denotes the local decision made by system S_i, and u_0 is the final decision made by fusion center. Each u_i is a binary random variable characterized by the associated false alarm probabilities and miss probabilities. The fusion of all local decisions u_i, $i = 1, \cdots, N$, in Bayesian framework is suggested in the fusion center. From [13], the fusion formulation minimizing the Bayes risk is in the form of likelihood ratio test (LRT).

$$\frac{P(u_1, u_2, \cdots, u_N|H_1)}{P(u_1, u_2, \cdots, u_N|H_0)} \underset{u_0 = 0}{\overset{u_0 = 1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} = \eta \qquad (9)$$

where C_{mn} denotes the cost of decision being H_m when H_n is present, and P_m denotes the prior probability for hypothesis H_m. All u_is are assumed to be conditional independent so that the likelihood ratio test can be expressed in the following form by taking logarithm.

$$\begin{aligned} &\log \frac{P(u_1, u_2, \cdots, u_N|H_1)}{P(u_1, u_2, \cdots, u_N|H_0)} \\ &= \sum_{\text{all } u_i=1} \log \frac{P(u_i = 1|H_1)}{P(u_i = 1|H_0)} + \sum_{\text{all } u_j=0} \log \frac{P(u_j = 0|H_1)}{P(u_j = 0|H_0)} \\ &= \sum_{i=1}^{N} \left[u_i \log \frac{1 - P_{M_i}}{P_{F_i}} + (1 - u_i) \log \frac{P_{M_i}}{1 - P_{F_i}} \right] \underset{u_0 = 0}{\overset{u_0 = 1}{\gtrless}} \log \eta \end{aligned} \qquad (10)$$

where P_{M_i} and P_{F_i} are associated miss probability and false alarm probability for system S_i, respectively. Thus a weighted sum of local decisions is formed and is compared with a threshold $\log \eta$. The weights are functions of false alarm and miss probabilities in each local system. In other words, these weights are functions of the quality of local decisions.

5. Experiments

The pair-wise LID experiments were conducted on Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus. There are ten languages in the corpus, they are English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. In the training set, utterance belonged to language l are used to train ergodic Markov model $\Lambda^{EMM,l}$ for prosody-based system and language models λ_l^{BG} in PRLM. Speech data in development set were used to estimate false alarm and miss probabilities for Bayesian formulation. 45-sec unrestricted domain utterances (STB) and 10-sec domain specific utterances (HTC, HTL, MEA, ROO) are chosen for the evaluation. The identification rate is calculated as the percentage of correctly identified utterances out of all evaluation utterances in each pairwise LID task.

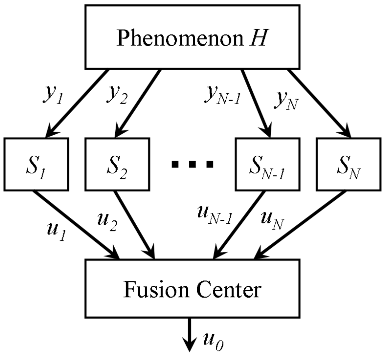

Figure 1: *Fusion system in parallel configuration.*

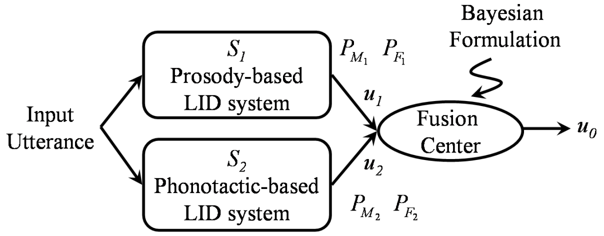

Figure 2: *Fuse prosody-based and phonotactic-based LID systems via Bayesian formulation*

Our experiment configuration is shown in Figure 2, S_1 and S_2 denote prosody-based and phonotactic-based LID systems respectively. Meanwhile, u_1 and u_2 are local decisions made by each system. The associated false alarm probabilities P_{F_i} and miss probabilities P_{M_i} for system S_i are determined from the experiments conducted on OGI-TS development set. Because the pair-wise language identification is conducted here, each language pair of P_{F_i} and P_{M_i} should be estimated. That is, there are forty-five pairs of $(P_{F_i}, P_{M_i})_{L_1,L_0}$ are needed for each system to discriminate between language L_1 and language L_0. While employing Equation (10), hypothesis H_1 denotes the input utterance belongs to language L_1, whereas H_0 hypothesizes L_0 the hypothesized language. Input utterance belongs to either language L_1 or language L_0 depends on the value of LRT. If LRT is greater than the given threshold, the fusion center hypothesizes L_1 the language that input utterance belongs to. Otherwise, language L_0 is the hypothesized language. With the assumption of equal priors, $P_1 = P_0$, and the uniform cost assignment, C_{mn} is a Kronecker delta function. The threshold η can be simplified to one in Equation (9) and becomes to zero in Equation (10) after taking logarithm.

Experimental results are listed in Table 2 and Table 3. In Table 2, each identification rate in the cell is obtained by averaging

45 pair-wise LID tasks. The results reveal the effectiveness of our suggested fusion technique. In average, the fused system achieves 10.87% error rate reduction for 10-sec utterances, and 14.88% for 45-sec utterances when comparing to the better individual system, PRLM. Furthermore, Table 3 demonstrates the detailed identification rate for each language. So each rate in Table 3 is obtained by averaging nine pair-wise LID tasks.

Table 2: The average pair-wise identification rate for prosody-based, phonotactic-based, and fusion system

	10s Utts	45s Utts
Prosody System	70.02%	81.35%
PRLM System	84.45%	90.93%
Fusion	86.14%	92.28%

Table 3: Identification rate of each language

		EMM	PRLM	Fusion
EN-other	45s	81.84%	94.55%	95.17%
	10s	65.99%	85.70%	86.45%
FA-other	45s	85.05%	94.49%	95.36%
	10s	70.98%	84.05%	86.02%
FR-other	45s	71.51%	87.94%	89.71%
	10s	66.91%	83.38%	85.34%
GE-other	45s	84.65%	92.64%	94.12%
	10s	70.02%	85.32%	87.26%
JA-other	45s	86.08%	91.80%	93.41%
	10s	79.48%	86.51%	87.88%
KO-other	45s	82.71%	90.36%	92.17%
	10s	69.57%	84.50%	85.52%
MA-other	45s	83.41%	90.62%	92.32%
	10s	73.09%	87.16%	88.74%
SP-other	45s	73.31%	85.54%	86.43%
	10s	64.23%	79.13%	82.39%
TA-other	45s	76.75%	95.02%	95.62%
	10s	67.71%	87.19%	88.54%
VI-other	45s	88.21%	86.39%	88.52%
	10s	73.24%	81.57%	83.27%

6. Conclusion

Decision fusion via Bayesian framework is one of the popular methods in distributed detection and estimation. Under this framework, the result of fusion process is easy to be analyzed if all local decisions u_i are the same. The final decision u_0 is the same as u_is in this circumstance. On the other hand, if there is a contradiction between local decisions, says $u_1 \neq u_2$. Quality of local decisions, P_{F_i} and P_{M_i}, should be considered and treated as weighting terms during the information fusion. In fact, performance confidence indexes employed in Gutiérrez's work is conceptually the same as ours. The performance of each local system provides a more reasonable choice for weighting. Although Bayesian framework demonstrates its effectiveness in our work, such decision level or hypothesis level fusion may not fully utilize the information extracted from different linguistic knowledge. Thus, fusion technique in feature level which could provide a better result should still be worthy to investigate.

7. Acknowledgements

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC94-2213-E-007-098.

8. References

[1] R. Ramus, and J. Mhler, *”Language identification with suprasegmental cues: A study based on speech resynthesis”*, in Journal of Acoustical Society of America, Vol. 105, No. 1, pp. 512-521, Jan. 1999.

[2] 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, 1997.

[3] J. Gutiérrez, J. Rouas, and R. André-Obrecht, *”Fusing language identification systems using performance confidence indexes”*, in Proc. ICASSP 2004, Montreal, Canada, Vol. 1, pp. 385-388, 2004.

[4] Y. Obuchi, and N. Sato, *”Language identification using phonetic and prosodic HMMs with feature normalization*”, in Proc. ICASSP 2005, Philadelphia, USA, Vol. 1, pp. 569-572, 2005.

[5] 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.

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

[7] C.Y. Lin, H.C. Wang, *”Language identification using pitch contour information in the ergodic Markov model”*, in Proc. ICASSP 2006, Toulouse, France, Vol. 1, pp. 193-196, 2006.

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

[9] M.A. Zissman, and E. Singer, *”Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling”*, in ICASSP 1994, Vol. I, pp. 305-309, 1994.

[10] S.M. Katz, *”Estimation of probabilities from sparse data for the language model component of a speech recognizer”*, in IEEE Trans. Acoustic, Speech, and Signal Processing, vol. 35, no.3, pp. 400-401, 1987.

[11] R. Viswanathan, and P.K. Varshney, *”Distributed detection with multiple sensors: Part I-Fundamentals”*, in Proceedings of the IEEE, Vol. 85, No. 1, pp. 54-63, Jan. 1997.

[12] R.S. Blum, S.A. Kassam, and H.V. Poor, *”Distributed detection with multiple sensors: Part II-Advanced Topics”*, in Proceedings of the IEEE, Vol. 85, No. 1, pp. 64-79, Jan. 1997.

[13] Z. Chair, and P.K. Varshney, *”Optimal data fusion in multiple sensor detection systems”*, in IEEE Trans. Aerospace Elect. Syst., vol. AES-22, pp. 98-101, Jan. 1986.