TARGET-ORIENTED PHONE TOKENIZERS FOR SPOKEN LANGUAGE RECOGNITION

Rong Tong[1,2], Bin Ma[1], Haizhou Li[1,2] and Eng Siong Chng[2]

[1]Institute for Infocomm Research, Singapore
[2]School of Computer Engineering, Nanyang Technological University, Singapore
{tongrong, mabin, hli}@i2r.a-star.edu.sg, aseschng@ntu.edu.sg

ABSTRACT

This paper presents a new strategy for designing the parallel phone recognizers for spoken language recognition. Given a collection of parallel phone recognizers, we select a subset of phones from each phone recognizer for each target language to construct a target-oriented phone tokenizer (TOPT). As a result, the collection of target-oriented phone tokenizers is more effective than the original parallel phone recognizers. This approach improves system performance significantly without requesting for additional transcribed training samples. We validate the effectiveness of the proposed strategy within the framework of the parallel phone recognizer followed by vector space modeling backend, or PPR-VSM. We achieve equal-error-rate of 2.21% and 3.65% on the 2003 and 2005 NIST LRE databases, respectively, for 30-second trials.

Index Terms— Spoken language recognition, parallel phone tokenizer, target-oriented phone tokenizer

1. INTRODUCTION

Spoken language recognition is a process of determining the language spoken in an utterance, where an utterance can be seen as a sequence of sound units. The phonotactic features, representing the phonetic constraints in a language [1], can be extracted from an utterance using a phone tokenizer, also referred to as phone recognizer. Although the common sounds are shared considerably across spoken languages, the statistics of these sounds, such as phone *n*-gram, can differ considerably from one language to another. Since the introduction of parallel phone recognizers followed by language models (PPR-LM) [2], the study of phonotactic features [3, 4, 5, 6] has attracted much attention. In human perceptual experiments [7], listeners with a multilingual background often perform better than monolingual listeners in identifying unfamiliar languages. The parallel phone recognizers (PPR), which benefit from its multi-stream knowledge resources, provide an effective front-end mechanism that converts the input utterance into multiple phonetic token sequences. With the PPR as the front-end, both the phone *n*-gram language models [1] and the vector space modeling (VSM) [8] were adopted as the backend. In the PPR-VSM approach, for each of the phone sequences generated from PPR, a high-dimensional feature vector, also known as bag-of-sounds vector, of phone *n*-gram probability attributes is created. A composite vector is formed by stacking multiple bag-of-sounds vectors. Vector classification algorithms, such as support vector machine (SVM), can then be applied on the composite vector for classification.

In the PPR framework, the languages of parallel phone recognizers and target languages may not have to be the same languages. For example, an English phone recognizer may be regarded as a human listener with English background, who tries to extract the discriminative information from the spoken utterances of each target language from an English listener's perspective. The discriminative information is expressed in an English phone sequence.

The same English phone recognizer is used for all the target languages in the current PPR practice, but we believe that it would be more effective if one can design the phone tokenizers that are target-oriented, for example, Arabic-oriented English phone tokenizer, Mandarin-oriented English phone tokenizer, as Arabic and Mandarin each is believed to have its unique phonotactic features to a English listener. Note that not all the phones and their phonotactics in the target language may provide equally discriminative information to the listener, it is desirable that the phones in each of the target-oriented phone tokenizers (TOPTs) can be those extracted from the full phone set of a phone recognizer, and having highest discriminative ability in distinguishing the target language from other languages.

There are three major advantages to adopt the TOPT strategy. (i) the TOPT follows the intuition that each target language has its own unique, discriminative phonotactic information; (ii) more phone tokenizers can be made available without requesting for additional annotated speech data of new languages; (iii) with relatively smaller phone inventory in TOPTs, higher order *n*-gram phonotactic

statistics become feasible. In this paper, we derive TOPTs from PPRs that replace the original PPRs within the PPR-VSM framework for language recognition. We study the strategy to select the TOPTs and their phone inventories from a larger number of phone tokenizer candidates.

This paper is organized as follows. In Section 2, we describe PPR-VSM language recognition system. In Section 3, we study the TOPT design strategy for more efficient language recognition with the VSM approach. In Section 4, we report experimental results. Finally we conclude in Section 5.

2. PPR-VSM FRAMEWORK

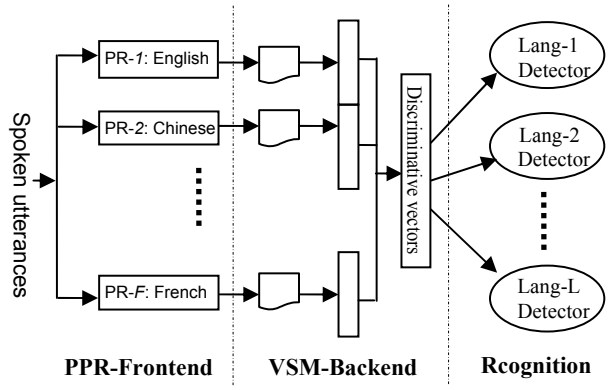

Fig. 1. PPR-VSM Language Recognition System

The PPR-VSM language recognition system is illustrated in Fig. 1. A collection of parallel phone recognizers (PPR) serve as voice tokenization front-end followed by vector space modeling (VSM) backend. The language classification is carried out based on the composite vector formed by stacking multiple bag-of-sounds vectors from the PPR [8].

Suppose that we have F phone recognizers with a phone inventory of $v = \{v_1, v_2, ... v_F\}$ and the number of phones in v_f is n_f. An utterance is decoded by these phone recognizers into F independent sequences of phone tokens. Each of these token sequences can be expressed by a high dimensional phonotactic feature vector with the n-gram probability attributes. The dimension of the feature vector is equal to the total number of n-gram patterns. If unigram and bigram are the only concerns, we will have a vector of $n_f + n_f^2$ phonotactic features, denoted as V_f to represent the utterance by the f-th phone recognizer. We concatenate all the F phonotactic feature vectors into a large composite vector

$$V = [V_1, ..., V_f, ... V_F]^t, \quad (1)$$

with a dimension of

$$S = \sum_f (n_f + n_f^2), \quad (2)$$

By using a single composite feature vector, we can effectively fuse the phonotactic features resulting from multiple phone recognizers and make the classification decision using a single decision hyperplane.

For each target language, a SVM is trained by using the composite feature vectors in the target language as the positive set and the composite feature vectors in all other languages as the negative set. With L target languages, we project the high dimensional composite feature vectors (with dimension of S) into a much lower dimension of [9]

$$Q = L \quad (3)$$

We formulate the language recognition as a hypothesis test. For each target language, we build a language detector which consists of two Gaussian mixture models (GMMs) $\{m^+, m^-\}$. m^+ is trained on the discriminative vectors of target language with dimension of Q, called positive model, while m^- is trained on those vectors of its competing languages, called negative model. We define the confidence of a test sample O belonging to language m^+ as the posterior odds in a hypothesis test under the Bayesian interpretation. We have H_0, which hypothesizes that O is language m^+, and H_1, which hypothesizes otherwise. The posterior odd is approximated by the likelihood ratio λ that is used for the final language recognition decision.

$$\lambda = \log\left(\frac{p(O|m^+)}{p(O|m^-)}\right) \quad (4)$$

3. TOPT DESIGN STRATEGY

The fundamental issue in spoken language recognition is to explore as many as possible discriminative cues for spoken languages, and to effectively organize these language cues in the classifier design. The success of PPR is attributed to the informative statistics from multiple phone recognizers, each of which covers certain phonotactic aspects in the feature space. Whereas more phone recognizers help boost the performance [2], this also means that additional annotated speech data are needed as we train the new phone recognizers.

Assuming that a collection of parallel phone recognizers are already trained, we are interested in reconfiguring the recognizers to increase the number of recognizers in a target-oriented manner. By doing so, we expect to improve the system performance without requesting for additional annotated speech data, nor additional acoustic modeling.

As we will only select a subset of phones from the original phone recognizer to serve in the TOPT, the smaller phone inventory also allows for the use of higher order n-gram statistics, such as trigram. It has been shown in our previous work [8] that trigram phonotactic features provide

a considerable improvement over the bigram features with VSM backend.

3.1. Selection of Phones

The phones in each of the TOPTs should be those with high discriminative characteristics to distinguish the target language from others. We adopt unigram phonotactic features to construct a linear SVM hyperplane to separate a target language from other languages. The phone selection is conducted by examining the discriminative property of each feature dimension.

With any of the F phone recognizers shown in Fig. 1, each of the training utterances in the L target languages can be converted into a phone sequence within an inventory of v_f phones. The phonotactic feature vector of unigram statistics $x=[x_1,x_2,..,x_i,..x_{v_f}]$, in v_f dimension, is used to represent the utterance. A one-versus-rest linear SVM is built for each of the target languages, with the feature vectors in the target languages as the positive set and those from all other languages as the negative set. The SVM is binary classifier in the form of $f(x)=a^T\psi(x)+b$, described by a weight vector a, an offset b, and a kernel function $\psi(.)$.

SVM learning is posed as an optimization problem with the goal of maximizing the margin, i.e., the distance between the separating hyperplane, $a^T\psi(x)+b=0$, and the nearest training vectors. Thus, a feature x_i with the weight a_i indicates the contribution of the i^{th} dimension in constructing the hyperplane. The idea is to consider the feature important if it significantly influences the width of the margin of the resulting hyperplane. It was found that the margin is inversely proportional to $\|a\|$, the length of a. The features with higher $|a_i|$ are more influential in determining the width of the separation margin [10]. We choose those phones with highest influences to the margin width of SVM hyperplane to construct the TOPT for the specific target language.

3.2. Selection of Phone Tokenizers

F parallel phone recognizers and L target languages result in $F\times L$ phone tokenizer candidates. From practical point of view, it is desired to select a subset of phone tokenizers that are effective. We generate a binary code vector for each phone tokenizer candidate according to the selected phone set in Section 3.1. The tokenizers with the most distinctive binary code vectors are selected for language recognition.

For the l^{th} phone tokenizer candidate based on the f^{th} phone recognizer, a binary code vector in the dimension of v_f is generated as

$$c^l=[c^l_1,c^l_2,..,c^l_i,..,c^l_{v_f}],\ c^l_i\in[0,1], \tag{5}$$

where $c^l_i=1$ if the i^{th} phone of the f^{th} phone recognizer is included in the l^{th} phone tokenizer candidate, and $c^l_i=0$ otherwise.

The selection of phone tokenizers is to identify the most distinguishing tokenizers and to avoid the duplication. For each of the L binary code vectors related to the f^{th} phone recognizer, the pair-wise Hamming distances to other L-1 vectors are calculated and summed as the discriminative score. Those phone tokenizer candidates with highest discriminative scores are selected to serve as the TOPT front-end, as the PPR front-end does in Fig. 1.

4. EXPERIMENTS

4.1. Experiment Setup

We conduct the experiments on the 30-second test segments of the 2003 and 2005 NIST Language Recognition Evaluation (LRE) tasks. The evaluation is carried out on recorded telephony speech in 12 languages in the 2003 LRE and in 7 languages in the 2005 LRE. There are 80 test segments in each of the 12 languages in the 2003 LRE, and 3662 test segments in all the 7 languages in the 2005 LRE[1].

The PPR front-end described in Section 2 includes phone recognizers of seven languages, English, Korean, Mandarin, Japanese, Hindi, Spanish and German, with 44, 37, 43, 32, 56, 36, and 52 phones respectively [6]. The training sets of LDC CallFriend database are used to construct and select the TOPTs, the development sets of CallFriend are used to build the ensemble of SVMs for the dimensionality reduction in (3), and the evaluation sets of CallFriend are used to train the GMMs for the final decision.

4.2. Experiment Results

The first experiment is designed to study the suitable number of phones in each of the TOPTs. If the number is too small, the resulting phone tokenizer may not have sufficient discriminative characteristics to separate the target language from others. On the other hand, a large number of phones will hinder the VSM backend from deploying multiple TOPTs with higher order n-gram phonotactic features.

Without loss of generality, we conduct the experiments with English phone recognizer and the corresponding

[1] The test segments in India-accented English are removed due to insufficient training and development data in India English.

TOPTs for *L=12* target languages. The solid line in Fig. 2 shows the language recognition performance on the 2003 LRE 30-second task using TOPT-VSM approach, with different numbers of phones in each of the TOPTs. The bigram phonotactic features in all the 12 TOPTs are used to generate the discriminative vectors in VSM backend. The dotted line indicates the 2003 LRE result using English phone recognizer with bigram phonotactic features. We choose top 20 phones in each TOPT in the next experiments.

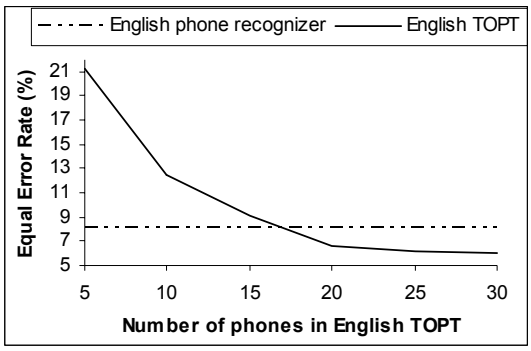

Fig. 2. Language Recognition using 12 English TOPTs with different numbers of phones in each of the TOPTs

The second experiment compares the discriminative capability between each of the 7 phone recognizers and the corresponding TOPTs. In Table 1, the second column shows the language recognition results on the 2003 NIST LRE 30-second task using single phone recognizer with the bigram phonotactic features. The third column shows the results of corresponding 12 TOPTs, each having 20 phones. The experiment results show that TOPT can provide better accuracy due to the increasing phonotactic resolution.

Table 1. Language recognition comparison between the phone recognizer and its TOPTs in VSM framework

EER (%)	Single Phone Recognizer	12 TOPTs
English	8.02	6.59
Korean	10.46	7.70
Mandarin	8.31	6.62
Japanese	9.59	7.17
Hindi	11.66	9.75
Spanish	10.48	10.18
German	11.12	8.58

In the third experiment, the top 5 distinguishing TOPTs are selected from each of the 7 phone recognizers based on the selection strategy described in Section 3.2. The trigram phonotactic features, which were too many to be included in PPR-VSM system, are now easily deployed in the TOPT-VSM system, with a smaller phone inventory. Table 2 shows the experiment results on the NIST 2003 and 2005 LRE 30-second tasks. PPR-VSM denotes the PPR language recognition system with 7 phone recognizers, while TOPT-VSM denotes the TOPT language recognition system with $7 \times 5 = 35$ TOPTs. More than 30% improvements have been achieved by replacing PPR with TOPT.

Table 2. Comparison between PPR-VSM and TOPT-VSM

EER (%)	NIST 2003	NIST 2005
PPR-VSM	3.16	5.61
TOPT-VSM	2.21	3.65

5. DISCUSSIONS

Target-oriented phone tokenizers method provides a solution to recruit more phone tokenizers without requesting for additional annotated speech data of new languages. Higher order phonotactic features can be deployed with the smaller phone inventory in each of the tokenizers. The increasing phonotactic resolution leads to a big improvement in the language recognition performance, tested on the NIST 2003 and 2005 LRE tasks.

In future works, we would like to study the selection of phone tokenizers and their phones from a universal phone recognizer, instead of parallel phone recognizers. The universal phone recognizer is expected to contain a large number of phones to represent the world's languages. In this way, we expect to derive TOPTs from a common set of universal phones.

6. REFERENCES

[1] T. J. Hazen and V. W. Zue, "Recent improvements in an approach to segment-based automatic language identification," in *Proc. ICSLP*, 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] C. Corredor-Ardoy, J. L. Gauvain, M. Adda-Decker, L. Lamel, "Language identification with language-independent acoustic models", in *Proc. Eurospeech*, 1997.

[4] 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*, 2003.

[5] J. L. Gauvain, A. Messaoudi, and H. Schwenk. "Language recognition using phone lattices", in *Proc. ICSLP*, 2004.

[6] R. Tong, B. Ma, D. Zhu, H. Li and E. S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *Proc. ICASSP*, 2006.

[7] Y. K. Muthusamy, N. Jain and R. A. Cole, "Perceptual benchmarks for automatic language identification," in *Proc. ICASSP*, 1994.

[8] H. Li, B. Ma and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech and Language Processing*, Vol. 15, No. 1, 2007.

[9] B. Ma, R. Tong and H. Li, "Discriminative vector for spoken language recognition," in *Proc. ICASSP*, 2007.

[10] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda and B. Scholkopf, "An introduction to kernel-based learning algorithm," *IEEE Trans on Neural Networks*, Vol. 12, No. 2, 2001.