Target-Oriented Phone Selection from Universal Phone Set for Spoken Language Recognition

Rong Tong\textsuperscript{1,2}, Bin Ma\textsuperscript{1}, Haizhou Li\textsuperscript{1,2} and Eng Siong Chng\textsuperscript{2}

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

Abstract

This paper studies target-oriented phone selection strategy for constructing phone tokenizers in the Parallel Phone Recognizers followed by Vector Space Model (PPR-VSM) paradigm of spoken language recognition. With this phone selection strategy, one derives a set of target-oriented phone tokenizers (TOPT), each having a subset of phones that have high discriminative ability for a target language. Two phone selection methods are proposed to derive such phone subsets from a phone recognizer. We show that the TOPTs derived from a universal phone recognizer (UPR) outperform those derived from language specific phone recognizers. The TOPT front-end derived from an UPR also consistently outperforms the UPR front-end without involving additional acoustic modeling. We achieve an equal error rates (EERs) of 1.33\%, 1.75\% and 2.80\% on NIST 1996, 2003 and 2007 LRE databases respectively for 30 second closed-set tests by including multiple TOPTs in the PPR.

Index Terms: target-oriented phone tokenizer, spoken language recognition, parallel phone recognizer, vector space modeling, universal phone recognizer

1. Introduction

A spoken language recognition system typically consists of two main modules: one is the language characterization that extracts the information to represent a language. This module is also referred to as feature extraction; another one is the classifier that distinguishes one language from the others using the extracted features. The features that characterize spoken language fall into two broad categories: acoustic feature and phonotactic feature.

The acoustic feature is derived from the speech signal itself, such as Mel-frequency Cepstral Coefficients (MFCC). It is motivated by the observation that each language is made up of a variety of different sounds in a unique way. Therefore, the distribution of acoustic feature reflects the statistics of the sound distributions in a particular language. Acoustic features are shown to be useful in spoken language recognition with modeling techniques such as Gaussian Mixture Models (GMM) [1] and support vector machine (SVM) [2]. Although acoustic features can be easily obtained from the speech signal, the useful language information often corrupted by the distortion caused by the transmission channel or speakers. Many studies have been focused on improving the expressiveness of acoustic features for language characterization [1,3].

The phonotactic feature is motivated by the fact that languages differ in the arrangement of their sound tokens or phonemes, although they often share common sound inventory. The statistics of sound token sequences therefore characterize the languages. The sound tokens can be any acoustically meaningful segments, such as phones, syllables or lexical words. In practice, we need a sound tokenizer that decodes the input speech signal into a sound token sequence, from which we further derive the phonotactic features, such as unigram, bigram or trigram. Many studies have adopted phone units as the sound tokens, therefore, the sound tokenizer is also known as the phone recognizer, such as that in the parallel phone recognizer (PPR) front-end [4].

The success of PPR is attributed to the diverse statistics from multiple phone recognizers, each of which covers certain phonotactic aspect in the feature space. Whereas more phone recognizers help boost the performance [4], this also means that additional annotated speech data are needed as we train the new phone recognizers. We proposed a target-oriented phone tokenizer (TOPT) strategy to construct a set of new phone tokenizers without involving additional acoustic model training [5]. The phones to form each new phone tokenizer are extracted from the full phone set of original phone recognizer. They are chosen to yield the highest discriminative ability for the target languages. As smaller phone subset is used in the front-end, high order phonotactic features, such as trigram, become possible. Studies have shown that TOPT phone recognizers outperform the original phone recognizer [5].

Universal phone recognizer (UPR) [6] was introduced as an alternative to the PPR. It pools phones from multiple languages into a single sound inventory, thus covers a variety of sound patterns in multiple languages. In human perceptual experiments [7], listeners with a multilingual background often perform better than monolingual listeners in identifying unfamiliar languages. A UPR can be seen as a person who has knowledge of multiple languages. This paper extends the studies in [5] by comparing two phone selection methods for constructing TOPTs from universal phone recognizers.

This paper is organized as follows. In Section 2, we briefly introduce the PPR-VSM framework for language recognition. We describe two target-oriented phone selection strategies and two methods of constructing universal phone recognizers. In Section 3, we describe the experimental setup and report the experiment results of the TOPTs derived from universal phone recognizers and language specific phone recognizers. Finally we conclude in Section 4.

2. Target Oriented Phone Selection

2.1. PPR-VSM

In PPR-VSM language recognition system [8], a collection of parallel phone recognizers (PPR) serve as sound tokenization front-end which is followed by a vector space modeling (VSM) back-end. The language classification is carried out based on the composite vector formed by stacking multiple
bag-of-sounds vectors from the PPR [8,9]. 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. For a \( n \) target languages recognition task, we can build \( n \) such SVMs. The output scores of those SVMs are used to produce \( n \) dimensional discriminative vectors \([9]\). In this way, we project the high dimensional composite feature vectors into a much lower dimension of \( n \). The generated discriminative feature vectors are used as the input feature vectors for the language recognition.

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) \( \{m^+, m^-\} \). Here \( m^+ \) is trained on the discriminative vectors of target language, while \( m^- \) is trained on those vectors of its competing languages. 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. The posterior odd is approximated by the likelihood ratio \( \lambda \) that is used for the final language recognition decision.

\[
\lambda = \log \left( \frac{p(O|m^+)}{p(O|m^-)} \right)
\]

(1)

### 2.2. Target-oriented phone selection

Assuming a phone recognizer is already trained, we are interested in selecting a subset of phones for each target language and reconfiguring a set of new tokenizers, we call them target-oriented phone tokenizers (TOPTs) [5]. In this way, we increase a number of sound tokenizers in a target-language and reconfiguring a set of new tokenizers, we call interested in selecting a subset of phones for each target language.

Assuming a phone recognizer is already trained, we are interested in selecting a subset of phones for each target language and reconfiguring a set of new tokenizers, we call them target-oriented phone tokenizers (TOPTs) [5]. In this way, we increase a number of sound tokenizers in a target-language and reconfiguring a set of new tokenizers, we call interested in selecting a subset of phones for each target language. The generated discriminative feature vectors are used as the input feature vectors for the language recognition.

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) \( \{m^+, m^-\} \). Here \( m^+ \) is trained on the discriminative vectors of target language, while \( m^- \) is trained on those vectors of its competing languages. 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. The posterior odd is approximated by the likelihood ratio \( \lambda \) that is used for the final language recognition decision.

\[
\lambda = \log \left( \frac{p(O|m^+)}{p(O|m^-)} \right)
\]

(1)

The target-oriented phone selection strategy is illustrated in Figure 1. Assuming we have a language recognition task of \( n \) target languages, given a phone recognizer with phone inventory \( V = \{v_1, v_2, \ldots, v_m\} \) which contains \( m \) phones, we estimate the discriminative power of each phone \( v_i \) in distinguishing a target language \( l_k \) from other target languages: \( l_j \) with \( j \in [1, n] \) and \( j \neq k \). The discriminative power of phones in \( V \) for distinguishing language \( l_k \) from others can be denoted as \( W_k = \{w_{v_1,k}, w_{v_2,k}, \ldots, w_{v_m,k}\} \). We select a subset of phones that have highest discriminative power to construct a new target-oriented phone tokenizer \( \text{TOPT}_k \). In this way, we can construct \( n \) new target-oriented phone tokenizers, one for each target language.

The target-oriented phone selection can be formulated as a standard feature selection problem. The objective of the feature selection is to find those phones that contribute the most to the correct language classification. In this paper, we study two discriminative power estimation methods, separation margin and mutual information, for the target-oriented phone selection.

#### 2.2.1. Separation margin (SM)

We adopt phone unigram statistics as the feature vector to construct a one-versus-rest linear SVM hyperplane to separate a target language from other languages. A SVM is a binary classifier in the form of

\[
f(x) = \alpha^T \phi(x) + b
\]

(2)

where \( \alpha \) stands for a weight vector, \( b \) is the offset, and \( \phi(\cdot) \) is a kernel function.

SVM learning is posed as an optimization problem with the goal of maximizing the separation margin, i.e., the distance between the separating hyperplane \( \alpha^T \phi(x) + b = 0 \) and the nearest training vectors. Assume we have a feature \( x_i \) with weight \( \alpha_i \) that indicates the contribution of the \( i \)th element in constructing the separation hyperplane. The idea is to look into the importance of each feature element by examining how it influences the width of the margin for the resulting hyperplane. It was found that the margin is inversely proportional to \( \| \alpha \| \), i.e., the length of \( \alpha \). Those feature elements with higher weight are more influential in determining the width of the separation margin.

#### 2.2.2. Mutual information (MI)

We can formulate language recognition as a series of 2-class separation problems. For each target language, a one-versus-rest classifier can be built. Given a task of classifying a target language \( l_k \) and its competing languages, the language category is a random variable that has two values \( L = \{l_k, l_\neq k\} \), where \( l_k \) stands for a positive language and \( l_\neq k \) stands for a negative language. The presence of phone \( v_i \) is another random variable that has two possible values \( T = \{t_i, l_k,t_i, l_\neq k\} \), where \( t_i \) denotes the phone \( v_i \) is present in an utterance that is spoken in language \( l_k \) and \( t_i, l_\neq k \) denotes the phone \( v_i \) is present in an utterance of language \( l_\neq k \). The mutual information of the phone presence \( T \) and language category \( L \) can be estimated as:

\[
I(T;L) = \sum_{t_i} \sum_{l_k} p(t_i,l_k) \log \frac{p(t_i,l_k)}{p(t_i)p(l_k)}
\]

(3)

where \( p(t_i,l) \) is the probability of phone \( t_i \) appears in language \( l \). Those phones with greater value of mutual information are more informative in separating target language from its competing languages.
2.3. Universal phone recognizer

The universal phone recognizer (UPR) is well motivated for multilingual speech processing. It offers many attractive properties as the front-end of a spoken language recognition system.

Ideally a UPR is trained from all the languages in the world. In practice, a UPR is often trained from data of several languages based on the assumption that common sounds are shared among languages. The key in developing a UPR is to have a unified set of phone inventory that covers all the existing languages. The International Phonetic Alphabet (IPA) provides a standard phonetic notation for most of the phonetic attributes in the existing languages. In acoustic model training, due to the limitation of the available training data, the universal phone set often obtained by mapping IPA notation to a smaller phone set in ASCII format. For example, Worldbset project [10] and the GlobalPhone project [11] proposed their universal phone sets and collected phonetic labeled corpus in multiple languages.

We study two methods of constructing a UPR for language recognition. The first method makes use of several existing language specific phone recognizers, by lumping together all the acoustic models from several language specific phone recognizers. This method is apparently straightforward. There is no need to train acoustic models and therefore no training data are required.

The second method of constructing a UPR requires a predefined universal phone set and phonetic labeled training data from several languages. Given a universal phone set, a UPR can be obtained by using a bottom-up training method: firstly bootstrapping the initial models with small amount of phonetic labeled data in several languages; secondly decoding new language data with initial models, and the decode results are used as transcription to refine initial models in an unsupervised manner. The advantage of this method is that the diversity of phones in different languages presented in a development database is observed. The acoustic models are calibrated through an unsupervised training process.

3. Experiment

3.1. Experiment Data

The language recognition experiments are conducted on the 30 second trials of the 1996, 2003 and 2007 NIST Language Recognition Evaluation (LRE). There are 12 target languages in 1996 and 2003 evaluation, 14 in 2007 evaluation. Table 1 shows the target languages and number of 30 second trials in each year’s evaluation. All the results presented in this paper are for closed-set language recognition.

We conduct language recognition experiments on 7 language specific phone recognizers: English, Korean, Mandarin, Japanese, Hindi, Spanish and German. There are 44 phones for English recognizer, 37 for Korean, 43 for Mandarin, 32 for Japanese, 56 for Hindi, 36 for Spanish and 52 for German [12].

Two universal phone recognizers are used in the experiments. The first universal phone recognizer, referred as UPR1, is constructed from the 7 language specific phone recognizers described above. All the phone models from 7 phone recognizer are combined to form a universal phone recognizer. There are 300 phones in total.

The second universal phone recognizer, referred as UPR2, is trained from the 6-language stories data in OGI-MLTS corpus [13]. Each language has about one hour of phonetic labeled stories data. The phonetic symbols used in OGI-MLTS corpus is Worldbset. Around 300 phonetic symbols are presented in the 6 languages stories corpus. Consider the limited numbers of samples for some phones, only 164 phones with high occurrence are selected in UPR2 phone set. The initial acoustic models are trained using the phonetic transcribed data from OGI 6 languages. The stories data of OGI 22 languages corpus [14] are decoded with the initial models, and the decoding results are used to adapt the initial models to derive UPR2 acoustic models.

The training sets of LDC CallFriend database are used to perform phone selection and construct TOPTs. The PPR-VSM language classifiers are trained with the development sets of LDC CallFriend corpus, OHUS 2005 data and LRE07 development data released by LDC. As part of the LRE 2005 test sets are extracted from the OHUS 2005 data, the LRE 2005 test sets are not used in our experiments.

Table 1. Target languages and the number of 30 second trials in NIST LRE databases

<table>
<thead>
<tr>
<th>Year</th>
<th>English, Arabic, Farsi, French, Mandarin, German, Vietnamese, Hindi, Japanese, Spanish, Korean, Tamil</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>English, Arabic, Farsi, French, Mandarin, German, Vietnamese, Hindi, Japanese, Spanish, Korean, Tamil</td>
</tr>
<tr>
<td></td>
<td>1492</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>English, Arabic, Farsi, French, Mandarin, German, Vietnamese, Hindi, Japanese, Spanish, Korean, Tamil</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>English, Arabic, Farsi, French, Mandarin, German, Vietnamese, Hindi, Japanese, Spanish, Korean, Tamil</td>
</tr>
<tr>
<td></td>
<td>1280</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>English, Arabic, Farsi, Mandarin, Russian, Bengali, German, Spanish, Thai, Tamil, Hindustani, Vietnamese, Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>English, Arabic, Farsi, Mandarin, Russian, Bengali, German, Spanish, Thai, Tamil, Hindustani, Vietnamese, Japanese</td>
</tr>
<tr>
<td></td>
<td>2510</td>
</tr>
</tbody>
</table>

3.2. Experiment Results

A UPR can be seen as a person who has knowledge of multiple languages while a language specific phone recognizer can be seen as a person that has knowledge of single language. In the first experiment, we would like to compare the performance of UPR and language specific phone recognizers in language recognition. Table 2 shows the equal error rate (EER%) of systems using single phone recognizer as front-end and VSM as back-end. We report experiments using 2 UPRs and 7 language specific phone recognizers as the front-end with bigram phonotactic features.

Table 2. Language recognition performance (EER%) using single UPR and language specific phone recognizer as front-end tokenizer

<table>
<thead>
<tr>
<th>Sound Tokenizer</th>
<th>English</th>
<th>Hindi</th>
<th>German</th>
<th>Korean</th>
<th>Mandarin</th>
<th>Spanish</th>
<th>Japanese</th>
<th>UPR1</th>
<th>UPR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>5.63</td>
<td>8.33</td>
<td>7.26</td>
<td>8.62</td>
<td>5.37</td>
<td>8.93</td>
<td>5.89</td>
<td>4.14</td>
<td>4.41</td>
</tr>
</tbody>
</table>

1 http://www.arts.gla.ac.uk/IPA/
2 http://www.nist.gov/speech/test/lre
3 http://www.ldc.upenn.edu/
For three test sets, the two UPRs consistently outperform any of the 7 language specific phone recognizers. These results coincide with the finding in perceptual study that a subject knowing more languages helps in language recognition. The results of UPR1 have better accuracy than UPR2. This may be due to the fact that only small amount of labeled data are used in training UPR2. For example, the English phone models in UPR1 are trained from more than 100 hours labeled English data, while the labeled data used for UPR2 are only 1 hour for each of the 6 languages.

Table 3. Language recognition performance (EER%) using UPR and TOPTs derived from UPR as front-end tokenizer

<table>
<thead>
<tr>
<th>Sound Tokenizer</th>
<th>1996</th>
<th>2003</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPR1</td>
<td>4.14</td>
<td>6.46</td>
<td>8.39</td>
</tr>
<tr>
<td>TOPT-UPR1 (SM)</td>
<td>2.20</td>
<td>3.23</td>
<td>4.64</td>
</tr>
<tr>
<td>TOPT-UPR1 (MI)</td>
<td>1.58</td>
<td>3.87</td>
<td>4.75</td>
</tr>
<tr>
<td>UPR2</td>
<td>4.41</td>
<td>6.55</td>
<td>8.52</td>
</tr>
<tr>
<td>TOPT-UPR2 (SM)</td>
<td>2.27</td>
<td>3.53</td>
<td>5.23</td>
</tr>
<tr>
<td>TOPT-UPR2 (MI)</td>
<td>2.46</td>
<td>3.70</td>
<td>5.56</td>
</tr>
</tbody>
</table>

In the second experiment, we evaluate the language recognition performance of the target-oriented phone tokenizers (TOPTs) derived from two UPRs using two phone selection strategies described in Section 2.2. Table 3 compares the system performance between the UPR front-ends and the UPR derived TOPT front-ends. The TOPT results are obtained by using top 5 TOPTs and trigram phonotactic features. As all the TOPTs are derived from the same phone inventory, we only select the distinctive TOPTs as PPR front-end. All the TOPTs are ranked based on their accumulated Hamming distance, those top TOPTs are selected from the ranked list [5]. Each TOPT contains 20 phones that have highest discriminative power based on the proposed phone selection criteria. The two phone selection methods achieve comparable performance. However, UPR derived TOPTs (parallel phone recognizer front-end) consistently outperform UPR (front-end of single phone recognizer), although the same set of phones are involved.

Table 4. Language recognition performance (EER%) using TOPTs derived from UPRs and from language specific phone recognizers as front-end tokenizers (all TOPTs derived using separation margin criteria)

<table>
<thead>
<tr>
<th>Sound Tokenizer</th>
<th>1996</th>
<th>2003</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOPT-English</td>
<td>3.75</td>
<td>5.92</td>
<td>5.64</td>
</tr>
<tr>
<td>TOPT-Hindi</td>
<td>4.52</td>
<td>6.23</td>
<td>7.74</td>
</tr>
<tr>
<td>TOPT-German</td>
<td>4.25</td>
<td>6.27</td>
<td>8.20</td>
</tr>
<tr>
<td>TOPT-Korean</td>
<td>4.40</td>
<td>6.74</td>
<td>8.13</td>
</tr>
<tr>
<td>TOPT-Spanish</td>
<td>5.54</td>
<td>8.06</td>
<td>9.04</td>
</tr>
<tr>
<td>TOPT-Japanese</td>
<td>3.66</td>
<td>5.47</td>
<td>6.36</td>
</tr>
<tr>
<td>TOPT-Mandarin</td>
<td>2.85</td>
<td>4.30</td>
<td>5.68</td>
</tr>
<tr>
<td>TOPT-UPR1</td>
<td>2.20</td>
<td>3.23</td>
<td>4.64</td>
</tr>
<tr>
<td>TOPT-UPR2</td>
<td>2.27</td>
<td>3.53</td>
<td>5.23</td>
</tr>
</tbody>
</table>

In the third experiment, we compare the system performance of the TOPTs between those derived from UPR and those derived from language specific phone recognizers. In both cases, the system has parallel phone recognizers as the front-end. Table 4 shows the language recognition results of using TOPTs derived from 7 individual language specific phone recognizers and 2 UPRs as the sound tokenizers. The TOPT phone selection is based on the separation margin criteria. Without surprise, the result shows that the target-oriented phone tokenizers derived from a universal phone recognizer (UPR) consistently outperform those derived from any language specific phone recognizers in language recognition. The language recognition performance has been further improved by fusing all the TOPTs with score averaging.

4. Conclusion

We study two target-oriented phone selection strategies for deriving a set of target-oriented phone tokenizers (TOPTs) from an existing phone recognizer. Without requiring model training, we synthesize multiple phone tokenizers that have smaller phone inventory, hence allowing for higher order phonotactic features. The system performance can be further improved by constructing TOPTs from universal phone recognizers. Our preliminary results show that UPR has better discriminative ability than language specific phone recognizer. The two phone selection strategies, namely separation margin and mutual information, work equally well in all three test sets.

5. References