INTERSPEECH 2009 BRIGHTON

Target-Aware Language Models for Spoken Language Recognition

Rong Tong[1,2], Bin Ma[1], Haizhou Li[1,2],Eng Siong Chng[2] and Kong-Aik Lee[1]

[1] Institute for Infocomm Research, Singapore 138632
[2] School of Computer Engineering, Nanyang Technological University, Singapore
{tongrong,mabin,hli,kalee}@i2r.a-star.edu.sg, aseschng@ntu.edu.sg

Abstract

This paper studies a new way of constructing multiple phone tokenizers for language recognition. In this approach, each phone tokenizer for a target language will share a common set of acoustic models, while each tokenizer will have a unique phone-based language model (LM) trained for a specific target language. The target-aware language models (TALM) are constructed to capture the discriminative ability of individual phones for the desired target languages. The parallel phone tokenizers thus formed are shown to achieve better performance than the original phone recognizer. The proposed TALM is very different from the LM in the traditional PPRLM technique. First of all, the TALM applies the LM information in the front-end as opposed to PPRLM approach which uses a LM in the system back-end; Furthermore, the TALM exploits the discriminative phones occurrence statistics, which are different from the traditional *n*-gram statistics in PPRLM approach. A novel way of training TALM is also studied in this paper. Our experimental results show that the proposed method consistently improves the language recognition performance on NIST 1996, 2003 and 2007 LRE 30-second closed test sets.

Index Terms: spoken language recognition, parallel phone tokenizer, target-oriented phone tokenizer, target-aware language model, universal phone recognizer

1. Introduction

The phonotactic approach, which makes use of a phone tokenization process using one or multiple phone tokenizers, has been widely applied in spoken language recognition. Although common sounds are shared considerably across spoken languages, the statistics of these sounds, such as *n*-gram, can differ considerably from one language to another. The success of the phonotactic features is that they can capture the lexical constraint of admissible phonetic combination in a language. The token sequences derived from tokenization are used to build language classifiers with statistical *n*-gram language modeling (LM) [1-4] or vector space modeling (VSM) method [5,6] by converting the phone *n*-gram statistics into high dimensional feature vectors. It is a common practice to apply phone recognizers in the tokenization process. 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. A PPR based language recognition system outperforms systems with single phone recognizer front-end [1,4] as each phone recognizer covers certain phonotactic aspects in the feature space. Whereas more phone recognizers help boost the performance, this also means that additional annotated speech data are needed for training new phone recognizers. In this paper, we are interested in constructing new phone tokenizers for the PPR front-end by using the same acoustic model but with different language models. Specifically, each language model of each target language is to capture the specific uniqueness of a target language phonotactic and hence contributes to discriminating the language even though the acoustic model is the same.

One of the important issues in phonotactic language recognition is how to identify the most discriminative features that are good in differentiating one language from others. In human listening experience, we observed that human listeners distinguish one language from another by spotting unique sound patterns. This suggests that some phones provide more salient discriminative information than others. A keyword selection method was proposed in [7] to find *n*-gram components that are more discriminative in language recognition. In our previous work, we proposed an discriminative feature selection method that derives a set of target-oriented phone tokenizers (TOPT) for language recognition [8], where we construct parallel target-oriented phone tokenizers from a phone recognizer using subset of phones that are uniquely important to the target languages. The approach improves language recognition system performance tremendously without requesting for additional labeled training data and acoustic modeling.

Despite its promising results [8,9], the TOPT method can be further improved: (i) The TOPT approach uses only a subset of phones from original phone inventory and hence has limited phone coverage; (ii) As TOPT makes hard decision on selecting a subset of phones with high discriminative ability, the choice of phones will affect performance. Refining the idea of phone selection, this paper proposes a method to address the above two concerns. We propose to derive parallel phone tokenizers with target-aware language models (TALM) from an existing phone recognizer. With TALM, we do not make hard decision on the choice of phones to be presented in a phone tokenizer, but use all the phones with a language model. Each language model is trained to reflect the discriminative ability of individual phones for a specific target language. If we interpret the TOPT method as making hard decision on the choice of phones, the TALM method can be interpreted as making soft decision. For simplicity, we only derive unigram language model to incorporate into the acoustic model decoding.

Inspired by the finding in human perceptual experiments [11] that listeners with multilingual background often outperform monolingual listeners in identifying unfamiliar languages, in our previous study, we applied the TOPT method to derive parallel phone tokenizers from a universal phone recognizer, which shown superior performance to those derived from a language specific phone recognizer [9]. In this paper, we compare the performance of TALM derived phone tokenizers from a universal phone recognizer and TALM derived phone tokenizers from a language specific phone recognizer.

This paper is organized as follows. In Section 2, we present the TALM design method. In Section 3, we describe the experimental set up. In Section 4, we report the experimental results. Finally we conclude in Section 5.

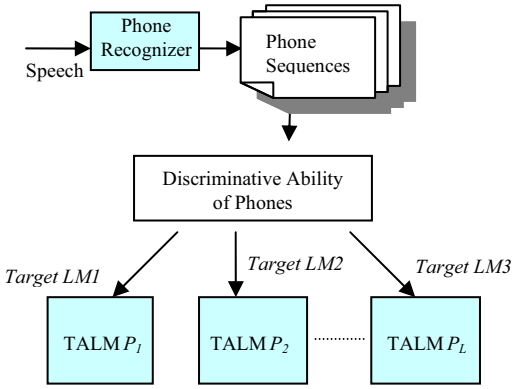
2. Target-Aware Language Models

Assuming the acoustic models of a phone recognizer are available, we are interested in reconfiguring the recognizer into multiple phone tokenizers that are suitable for language recognition in a target-aware manner. By doing so, we expect to improve the system performance without requiring for additional annotated data and acoustic model training.

In automatic speech recognition, the acoustic model encodes sources of acoustic variability such as speaker, channel, phonetic etc, while the language model imposes the lexical constraints. In the context of phone decoding, a proper language model provides prior knowledge of certain phones and phonotactics in the decoding. This is similar to a human listening test where the subject pays special attention to the phones and phonotactics of interest.

In language recognition with phonotactic features, a phone recognizer is used to generate phone sequences for all the target languages. It is a common practice that no language model is involved in the decoding processing since the speech data decoded by the phone recognizer are not in a specific language. We are interested in constructing multiple phone tokenizers from existing phone recognizer. The derived tokenizers will share a common set of acoustic models but each language tokenizer will have a unique target-aware unigram language model Compared with the original phone recognizer, we expect to see an improved language recognition performance by using parallel phone tokenizers constructed from these target-aware language models.

Figure 1: *A block diagram of the TALM modeling process*



It is generally agreed that each language has a unique phone and phonotactic composition. In other words, each phone has different discriminative ability in separating a target language from others. The target-aware language models (TALMs) will be generated by constructing a set of phone language models, each dedicated to a target language. Figure 1 shows the block diagram of the TALM modeling process. The training utterances in all the target languages are first processed by a phone recognizer to generate a set of phone sequences. Based on these phone sequences, we measure the discriminative ability of each phone in separating one target language from the others. For L target languages, we generate a new language model P_l for each target language $l = 1,...,L$ according to the discriminative ability of phones in separating the target languages. In this way, we obtain a set of new phone tokenizers which are sharing the same acoustic model, but they are differentiated by their language models.

2.1. Discriminative ability of phones

Based on the phone sequences from the training utterances of all the target languages, we make use of unigram statistics to construct a linear SVM hyperplane to separate a target language from other languages. The phone discriminative ability is measured by examining the discriminative property of each feature in the unigram vector.

Suppose that there is an inventory of v_f phones in the phone recognizer, each of the training utterances in the L target languages can be converted into a phone sequence consists of v_f phones. The 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 a 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. We 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. For the target language l, the features with higher $d_{i,l} = |a_i|$ are more influential in determining the width of the separation margin [12].

2.2. Target-aware language modeling

For a phone recognizer with v_f phones, the discriminative ability of phones toward target language l can be obtained following the discriminative ability measure method described in Section 2.1. It can be denoted as: $D_l = \{d_{1,l}, d_{2,l}, ...d_{i,l}, ..., d_{v_f,l}\}$. Based on the phone discriminative ability, we can derive a unigram language model by approximating the phone's occurrence probability with its discriminative ability,

$$P_l = \{p_{1,l}, p_{2,l}, ..., p_{i,l}, ..., p_{v_f,l}\} \quad (1)$$

where the occurrence probability for phone i is computed as:

$$p_{i,l} = \log(d_{i,l}^2 / \sum_{i=1}^{v_f} d_{i,l}^2). \quad (2)$$

This language model gives higher probability to those phones having higher discriminative ability for a target language. During the decoding process, the language model works together with the acoustic model to constrain the phone occurrence in the output sequences.

This can be seen as a generalized version of the target-oriented phone tokenizer (TOPT) approach, where in TOPT,

some phones are removed during decoding and the selected phones are given equal probabilities.

To summarize, it should be noted that the proposed TALM is very different from the LM widely adopted in the PPRLM [1-4]. Firstly, TALM applies language model during the tokenization process in the front-end, while the PPRLM uses language models to derive statistics from the output of the tokenizer in the back-end. Furthermore, the target-aware language models are estimated from the discriminative ability of the phones, which are different from the traditional *n*-gram statistics [10].

2.3. Universal phone recognizer

A universal phone recognizer (UPR) can be seen as a person who has the phonetic knowledge of many languages. In human perceptual experiments, listeners with a multilingual background often perform better than listeners that only know single language in identifying unfamiliar languages [11]. A UPR pools phones from multiple languages into a single sound inventory, thus covers a variety of sound patterns in multiple languages. One can expect that the TALMs derived from UPR will achieve better performance than the TALMs derived from a language specific phone recognizer.

Ideally a UPR is trained from the samples of all the existing languages in the world. However, in practice, a UPR is often trained from the samples of several languages based on the assumption that common sounds are shared among languages. We study the UPR that takes advantage of several existing language specific phone recognizers, by lumping together all the phone models from them. It makes use of the current available language specific acoustic models that are normally well trained with sufficient training data.

3. Experiment

3.1. Experiment setup

We used the PPRVSM system architecture for all the experiments [6]. The systems were trained on the LDC CallFriend (http://www.ldc.upenn.edu/), OHSU 2005 corpus (OHSU: http://www.ohsu.edu/), and the development data released by LDC for the 2007 NIST Language Recognition Evaluation. From the LDC CallFriend corpus, the Train sets were used to construct and select the TOPT and TALM phone tokenizers, the Devtest sets were used to build the ensemble of SVMs for the dimensionality reduction [8], and the Evaltest sets (excluding those in LRE96) together with the OHSU and LRE07 development data were used to train the GMMs for the final language recognition decision [8].

We conducted the experiments on the 30-second test trials of the 1996, 2003 and 2007 NIST Language Recognition Evaluation (LRE) tasks. The test trials are grouped into three test durations of 30, 10 and 3 seconds. There are 12 target languages in the 1996 NIST LRE (LRE96) and 2003 NIST LRE (LRE03), and 14 target languages in 2007 NIST LRE (LRE07). As part of the 2005 NIST LRE (LRE05) test sets were extracted from the OHSU 2005 data, the LRE05 corpus was not used in our experiments. In the experiments, we report the results on the 30-second closed-set trials.

Experiments are conducted with an English phone recognizer that has 44 phones and a UPR constructed by lumping all the phone models from seven language-specific phone recognizers, denoted as UPR-Merge. The phones of English phone recognizer mentioned above is included in the phone set of the UPR-Merge, the other phones are from six language specific phone recognizers, Korean, Mandarin, Japanese, Hindi, Spanish and German [6]. There are 37 phones for Korean phone recognizer, 43 for Mandarin, 32 for Japanese, 56 for Hindi, 36 for Spanish and 52 for German. There are 300 phones in the UPR-Merge phone inventory.

3.2. Language recognition with TALM

Three types of phone tokenizer front-ends are studied, English phone recognizer, English-TOPT and English-TALM.

The English-TOPT consists of a set of 12 target-oriented phone tokenizers (TOPTs), each contains a subset of 20 phones that have highest discriminative ability in separating one target language from others on the CallFriend training data [8]. No language model is applied in TOPTs. The language recognition output scores are obtained by taking the average of individual TOPT scores.

The English-TALM consists of a set of 12 phone tokenizers, each having a target-aware language model that emphasizes on those phones that has high discriminative ability, as described in Section 2. The discriminative ability of phones is estimated using CallFriend training data. Each TALM has the same 44 phones as in the English phone recognizer. The output scores are obtained by taking the average of individual TALM scores. Table 1 shows the equal error rates (EERs) of three tokenizers on NIST LRE96, LRE03 and LRE07 30-second test trials.

Table 1. *EER(%) of language recognition systems on 30-second test trials, with English phone inventory*

Recognizer/Tokenizer	LRE96	LRE03	LRE07
English Phone Recognizer	5.63	7.71	9.32
English-TOPT	4.63	6.25	7.67
English-TALM	3.60	5.65	6.63

Table 2. *EER(%) of language recognition systems on 30-second test trials with merged phone inventory*

Recognizer/Tokenizer	LRE96	LRE03	LRE07
UPR-Merge Phone Recognizer	4.14	6.46	8.39
UPR-Merge-TOPT	2.16	3.14	4.52
UPR-Merge-TALM	1.34	1.91	3.34

We also conduct experiments on the UPR-Merge phone recognizer which has a merged phone inventory from seven languges, and its derived TOPT and TALM. Table 2 shows the equal error rates using UPR-Merge phone recognizer and its derived UPR-Merge-TOPT and UPR-Merge-TALM phone tokenizers on NIST LRE96, LRE03 and LRE07 30-second test trials. The UPR-Merge-TOPT consists of 12 tokenizers and each has 50 phones that have highest discriminative ability in separating one target language from others. The UPR-Merge-TALM consists of 12 tokenizers, each has 300 phones. The 12 TALM tokenziers share the same acoustic models of 300 phones, are different only in language models. The language recognition output scores of UPR-Merge-TOPT and UPR-Merge-TALM are obtained by taking average on the output scores of individual tokenizers.

By constructing TOPT, which only uses a subset of high discriminative phones in the decoding process, we consistently improve language recognition performance [8]. This confirms the effectiveness of target-oriented phone selection. We also observe that TALM which has soft phone selection decision outperforms both the systems based on either the original phone recognizer or TOPT. This suggests

that the soft-decision outperforms the hard-decision phone selection through target-aware language modeling.

We study the prior probabilities for phones in UPR-merge-TALM tokenizers. Figure 2 shows the distribution of top 50 phones that have the highest discriminative ability for separating English from other 11 CallFriend languages. Note that not all the English phones are among the most informative phone sets in separating English from other languages. This can be explained by the fact that (i) human listeners can distinguish one language from others even he only has limited knowledge of that language; (ii) listeners with a multilingual background often perform better than monolingual listeners in language identification. It is also confirmed that a phone tokenizer with variety of phone coverage is generally give better performance in spoken language recognition.

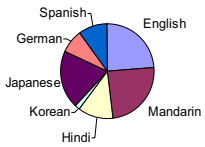


Figure 2. *Number of phones from original phone inventory in UPR-Merge-TALM 50 for English language recognition*

3.3. TOPT vs TALM

Note that the TOPT and TALM tokenizers shown in Table 1 and Table 2 have different number of phones. It would be good if we can compare the language recognition performance of TOPT and TALM approach when they are using same phone inventory. To this end, we can first select the phones using TOPT approach [8,9]. Then we train a TALM for the set of selected TOPT phones. Figure 3 shows the equal error rates (EERs) on LRE03 30-second evaluation data. The two curves show the performance of UPR-Merge-TOPT and UPR-Merge-TOPT-TALM with different number of phones in the phone inventory. For a given number of phones, both UPR-Merge-TOPT and UPR-Merge-TOPT-TALM are using the same phone inventory. The phone inventory consists phones which having highest discriminative ability in separating a specific target language from other languages. The UPR-Merge-TOPT consists of 12 phone tokenizers that use null-grammar while UPR-Merge-TOPT-TALM consists of 12 phone tokenizers that use target-aware language models constructed as in Section 2.2. The language recognition output scores are obtained by taking average on the output scores of individual tokenizers. It can be found that the UPR-Merge-TOPT-TALM consistently outperforms UPR-Merge-TOPT with different numbers of phones in the phone inventory.

When the number of phones increases over 50, the performance of UPR-Merge-TOPT in language recognition degrades. This can be explained by the fact that the TOPT treats each phone with equal importance, with the increasing of the phone numbers, the individual TOPT tends to become similar, thus providing little complimentary information with each other. In contrast, the UPR-Merged-TOPT-TALM, which benefits from the discriminative language models, is able to achieve a better accuracy in language recognition even with a large number of phones. The results suggest that TALM is more robust than TOPT in phonotactic front-end construction in language recognition.

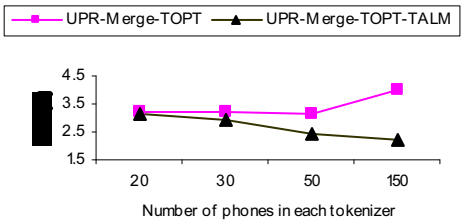


Figure 3. *EER(%) of different number of phones in UPR-Merge-TOPT and UPR-Merge-TOPT-TALM on LRE03 30-second test*

4. Conclusion

We extend the target-oriented phone tokenizer construction method by incorporating target-aware language models (TALM) for spoken language recognition. TALM offers a new way to construct parallel phone recognizers (PPR) for PPR front-end that provides a better performance in language recognition. We have shown that TALM outperforms target-oriented phone tokenizers that doesn't employ language model during phone decoding in the PPRVSM paradigm. The same technique is readily applicable to PPRLM and other phonotactic language recognition systems. It would be an interesting future work to study how to derive target-aware *n*-gram language models for language recognition.

5. References

[1] 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.

[2] 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.

[3] T. J. Hazen and V. W. Zue, "Recent improvements in an approach to segment-based automatic language identification," in Proc. *ICSLP*, 1994.

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

[5] W. M. Campbell, T. P. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation," in Proc. *IEEE Odyssey*, 2006.

[6] B. Ma, H. Li, and R. Tong, "Spoken Language Recognition Using Ensemble Classifiers," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 15, issue 7, pp. 2053-2062, 2007.

[7] F.S.Richardson, W. M. Campbell, "Language recognition with discriminative keyword selection", in Proc. *ICASSP,* 2008.

[8] R. Tong, B. Ma, H. Li and E. S. Chng, "A target-oriented phonotactic front-end for spoken language recognition," accepted for *IEEE Trans. on Audio, Speech and Language Processing, 2009*.

[9] R. Tong, B. Ma, H. Li and E. S. Chng, "Target-oriented Phone Selection from Universal Phone Set for Spoken Language Recognition," in Proc. *Interspeech,* 2008.

[10] Jirí Navrátil, Werner Zühlke, "Double Bigram-Decoding In Phonotactic Language Identification", in Proc. *ICASSP*, 1997.

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

[12] 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.