



Selecting Phonotactic Features for Language Recognition

Rong Tong^{1,2}, Bin Ma¹, Haizhou Li^{1,2} and Eng Siong Chng²

¹ Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore 138632

² School of Computer Engineering, Nanyang Technological University, Singapore 639798
 {tongrong, mabin, hli}@i2r.a-star.edu.sg, aseschn@ntu.edu.sg

Abstract

This paper studies feature selection in phonotactic language recognition. The phonotactic feature is presented by n -gram statistics derived from one or more phone recognizers in the form of high dimensional feature vectors. Two feature selection strategies are proposed to select the n -gram statistics for reducing the dimension of feature vectors, so that higher order n -gram features can be adopted in language recognition. With the proposed feature selection techniques, we achieved equal error rates (EERs) of 1.84% with 4-gram statistics on the 2007 NIST Language Recognition Evaluation 30s closed test sets.

Index Terms: feature selection, spoken language recognition, support vector machine, separation margin, Chi-Squared test

1. Introduction

Modeling a language with phonotactic features has been one of the successful techniques in spoken language recognition [1-5]. The phonotactic features are generally obtained in two steps: i) an input speech sample is firstly decoded by one or more phone recognizers such that a phone sequence is obtained from each of the phone recognizers; ii) the n -gram phone statistics are then derived as phonotactic features from the decoded phone sequences.

Several techniques have been developed to enhance the capability of phonotactic features. Parallel phone recognizers (PPR) are adopted as the decoding front-end [1][2] to achieve improved statistics over a single phone recognizer (PR). With the same phone recognizer, we can improve the performance significantly by utilizing the phonotactic statistics from lattice instead of 1-best phone recognizer [4]. The improvement is attributed to the rich information available in the lattice. It has also been shown that a higher order n -gram statistics which contain more discernible phonotactic information will yield a better language recognition result [2][3]. While these techniques improve the performance of language recognition, they often lead to a much higher feature dimension. Language modeling in a huge feature space is often at high computational cost, and requires a large amount of training data, which are not always available.

Feature selection is a strategy to keep a subset of features by removing those features that are less informative. A common practice for the phonotactic feature selection in language recognition system is to eliminate those phone n -grams with low occurrences by using a threshold of counts. This method is straightforward but it does not take into account the discriminative capability of phonotactic features. Thus the threshold is often conservatively set to a small value and only relatively small amount of features can be removed [3].

Recently a recursive feature elimination method was proposed to select discriminative keywords for language recognition [3]. Features are ranked by their contribution to the language separation and those low ranking features are removed. The remaining features are expanded to higher order n -grams, and the process runs iteratively.

In this paper, we study a different way of selecting phonotactic features with the criteria of separation margin and Chi-squared test. Both methods select phonotactic features by their contribution to the language separation. Experiment results show that the proposed feature selection methods greatly reduce the dimension of the phonotactic feature vectors while at the same time maintain the language recognition performance.

This paper is structured as follows: we describe the architecture of a phonotactic language recognition system in Section 2. In Section 3, we describe the feature selection methods based on the separation margin and Chi-squared statistics. We report the language recognition experiments based on the 2003 and 2007 NIST Language Recognition Evaluation in Section 4. Finally we conclude in Section 5.

2. Vector Space Modeling for Language Recognition

In a phonotactic language recognition system, one or more phone recognizers serve as the tokenization front-end to convert a speech utterance into one or more phone sequences.

Suppose that we have f phone recognizers. Let n_i denotes the number of phones in i -th phone recognizer. A speech utterance is decoded by these phone recognizers into f independent sequences of phone tokens. With vector space modeling, each of the token sequences can be expressed by a high dimensional phonotactic feature vector with n -gram probability attributes. We can then form a composite feature vector of d dimension [2] by stacking up the f n -gram feature vectors to represent the input utterance. Let p stand for the order of n -gram, we will have d be the sum of the dimensions of all the f n -gram feature vectors:

$$d = \sum_{i=1}^f \sum_{j=1}^p (n_i)^j \tag{1}$$

For each target language, a SVM is trained 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 can build L such SVMs. The output scores of these L SVMs can be further used to produce a L -dimensional discriminative vector. In this way, we project the high-dimensional feature vectors into a much lower dimension L . The L -dimensional discriminative feature vectors are then used as the input feature vectors for language modeling with Gaussian mixture models [2].

3. Phonotactic Feature Selection

We note from Eq. (1) that, if the number of phone recognizers f or the order of n -gram p increases, the feature dimension d can be increased dramatically. It is desirable to introduce a feature selection to control the dimension in practice. In this paper, we study two methods to measure the merit of features. A feature with high merit value means that this feature is more important for language recognition. In the first method, the features are measured by their contribution to the language separation margin. In another method, the Chi-squared value of feature existence and language classes is used to measure the importance of features.

3.1. Separation margin (SM)

We use the d -dimension feature vectors as in Eq. (1) to construct a one-versus-rest linear SVM hyperplane to separate a target language from other languages.

A SVM classifier learns a binary decision over the feature vectors x in the form of

$$f(x) = \alpha^T \varphi(x) + b, \quad (2)$$

where α is a weight vector, b is the offset, and $\varphi(\cdot)$ is a kernel function. The SVM learning is posed as an optimization problem with the goal of maximizing the separation margin, i.e. the distance between the separating hyperplane $\alpha^T \varphi(x) + b = 0$, and the nearest training vectors. Let the i -th element of α and x be denoted as α_i and x_i respectively. The aim is to estimate the importance of each feature element by examining how it influences the width of the margin in the resulting hyperplane. It was found that the margin is inversely proportional to $\|\alpha\|$, i.e. the length of α [6]. Hence for a target language l , a feature element i with higher weight $\alpha_{i,l}$ is more influential in separating language l with other target languages.

As there are multiple target languages, for certain feature, the average weight to each of the L target languages is used as the merit value of this feature.

3.2. Chi-squared measure (CM)

The Chi-squared (χ^2) statistic [7] measures the lack of independence between a feature and a target language. The Chi-squared (χ^2) statistics of a feature t and a language l is calculated by using the following two-way contingency table:

Table 1. Contingency table for a feature and a target language

	l	not l
t exists	A	B
t not exists	C	D

The Chi-squared (χ^2) value of a feature t and language category l is calculated as:

$$\chi^2(t, l) = \frac{N(AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}. \quad (3)$$

Here A stands for the number of times t and l co-occur;
 B is the number of times the t occurs without l ;
 C is the number of times l occurs without t ;
 D is the number of times neither l nor t ;
 N is the total number of training vectors;

A higher $\chi^2(t, l)$ value means the feature t and language l is more dependent, and a value close to zero means the feature and language are independent.

Similar to the separation margin, the average Chi-squared value of a feature element to each of the target language is used as its merit in the selection.

3.3. Efficient feature selection

With the two measures, the n -gram features can be ranked by their merits. As an example, Figure 1 shows the merit of features ranked by the separation margin criterion using 3-gram phonotactic features (the experiment setup will be described in section 4). A long tail in this figure suggests that many low ranking features have little difference as far as merit value is concerned.

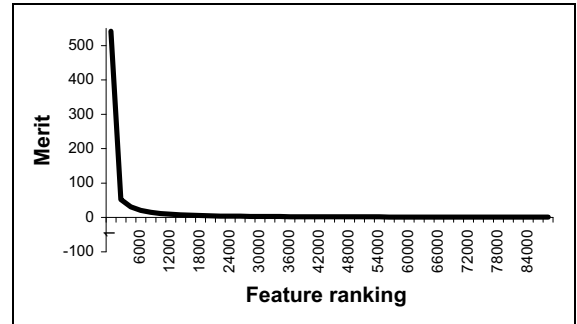


Figure 1. Features ranked by contribution to separation margin

In order to make an efficient feature selection based on the ranked merits, we apply a variation of relative mean difference (MD) to keep the most discriminative features. At any point of the ranked list, we calculate the arithmetic mean of its left n neighbors and right n neighbors as

$$\mu_{i,l} = \frac{1}{n} \sum_{j=1}^n m_{i-j} \quad (4)$$

and

$$\mu_{i,r} = \frac{1}{n} \sum_{j=1}^n m_{i+j} \quad (5)$$

Here m is the merit value. The mean difference between them is named as the mean difference of this point:

$$MD_i = \mu_{i,l} / \mu_{i,r}. \quad (6)$$

As the features are ranked from high to low, we have $MD_i \geq 1$. A small MD value means the difference between current n features and its left n neighbors are small, those features have similar contribution for the language separation. Excluding some features in small MD area may not affect the performance. We set the point with the minimum MD value as the feature selection decision point.

4. Experiment

4.1. Experiment setup

We use the PPR-VSM [2] system architecture in all the experiments. For simplicity, we use a single phone recognizer as the decoding front-end. The BUT Hungarian phone recognizer [8] is used as the front-end tokenizer to convert the input speech signals into phone sequences. There are 62 phones in the phone inventory. In all the experiments, the

phone lattices are derived in the decoding process and used to derive n -gram statistics features.

We conducted the experiments on the test trials of the 2007 and 2003 NIST Language Recognition Evaluation (LRE 07, LRE03) tasks in which the following 14 target languages are involved¹ in LRE07: Arabic, Bengali, Chinese, English, Hindustani, Spanish, Farsi, German, Japanese, Korean, Russian, Tamil, Thai, and Vietnamese. There are 12 target languages in LRE03: English, Arabic, Farsi, French, Mandarin, German, Hindi, Japanese, Spanish, Korean, Tamil, and Vietnamese.

The language recognition system for LRE07 is trained on the LDC CallFriend, the story data of OGI 22-language corpus [9], OHSU 2005² and LRE 2007 development data sets released by LDC. The language recognition system for LRE03 is trained on the LDC CallFriend data only. We also use the development data set to select the features. The performance is reported in terms of equal error rate (EER). The EER represents the operational point when the probabilities of the false acceptance and the false rejection are equal. We assume that the priors for target and non-target languages are equal.

4.2. Experiment Results

4.2.1 Effect of feature selection

In the first experiment, we compare the language recognition on different number of features derived from three different feature selection methods. The first one is feature selection by occurrence counts. For each feature, the average number of its occurrence in each of the target language is used as the merit of the feature element. Those features with low n -gram occurrence are eliminated from feature vectors. The other two methods are methods we proposed in Section 3.1 and 3.2.

According to Eq.(1), for single phone recognizer with 62 phones, the dimension of the possible 3-gram feature should be: $d = 62 * 62 * 62 + 62 * 62 + 62$. While in the recognition, there is 88292 features actually exists in LRE07 training sets. We selected the same amount of features using the 3 feature selection methods and use selected features for language recognition. Figure 2 compares the language recognition performance of these 3 different feature selection methods on LRE07 30 seconds data set.

Figure 2 shows that the two proposed feature selection method consistently outperform the counts threshold method on LRE07 30 seconds test set. The last data point (80k) indicates the performance of language recognition without any feature selection (baseline). The two proposed methods can achieve similar performance to the baseline while using only half of the features. The results also suggest that only a small amount of low occurrence features can be safely eliminated without affecting the language recognition performance.

When the number of selected features is small, the SVM-margin method gives better performance than Chi-squared method. This might be due to the fact that SVM margin method evaluates the contribution of features collectively, while the Chi-squared method evaluates the contribution of the feature individually. When selecting about half of the features for language recognition, the SVM-margin and Chi-squared methods achieve comparable performance.

¹ <http://www.nist.gov/speech/test/lre>.

² <http://www.ohsu.edu/>.

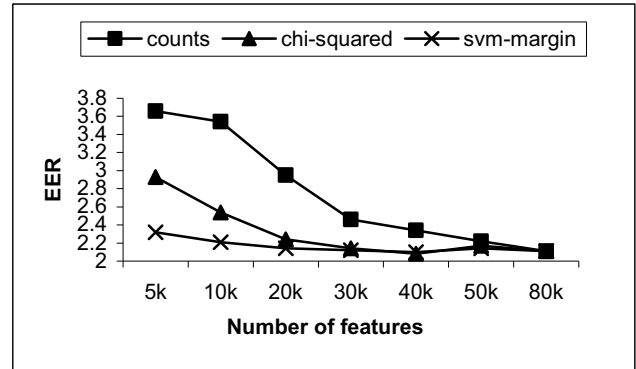


Figure 2. EER (%) with different number of features selected by different criteria on LRE07 30s test set

4.2.2 Efficient feature selection

In this sub-section, we evaluate our proposed Mean Difference method (MD) for efficient feature selection. We rank the features using two feature selection criteria described in Section 3 and conduct a series of language recognition experiments with different number of selected features. All the experiments are based on 3-gram phonotactic statistics on LRE 2007 30 second test set.

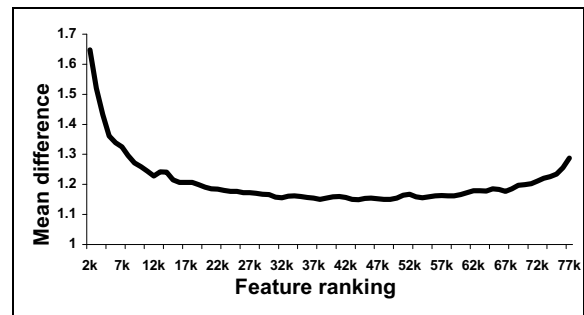


Figure 3. Mean difference of ranked features derived by separation margin criterion on LRE07 30s

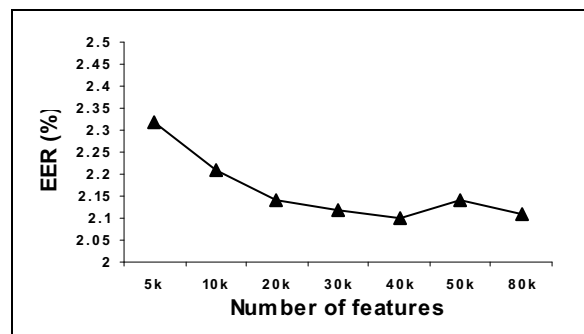


Figure 4. EER (%) with different number of features selected by separation margin criterion on LRE07 30s

Figure 3 shows the mean difference (MD) of features ranked by separation margin criterion on LRE07 30 seconds test set. The mean differences of $n=100$ are calculated as in Section 3.3. The minimum point is around 40000 features. Figure 4 shows the language recognition performance with different number of features selected by separation margin criterion on LRE07 30 seconds test set. Compare Figure 3 and 4, we can find out that the number of features for the best EER performance matches the feature selection based on MD.

Similar findings can be discovered with Chi-squared features selection. Figure 5 shows the MD of features ranked with Chi-squared criterion on LRE07 30 seconds test set. Figure 6 shows EER performance with different number of features selected by Chi-squared value on LRE07 30 seconds test set. The minimum point of MD is selected around 37000 features, we can also find in Figure 6 that using top 37k features selected by Chi-squared merit gives best EER.

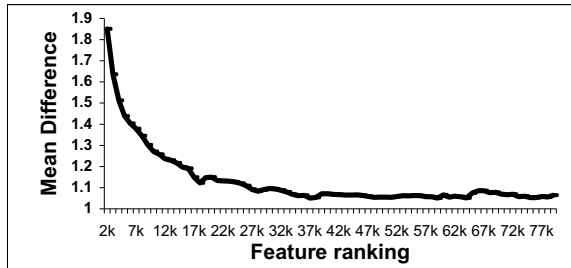


Figure 5. Mean difference of ranked features derived by Chi-squared value criterion on LRE07 30s

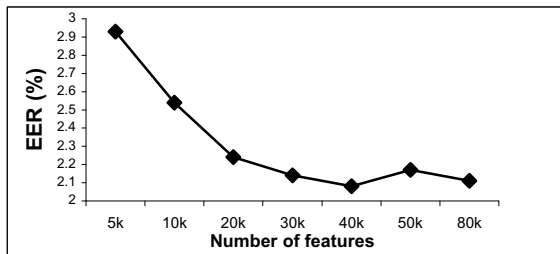


Figure 6. EER (%) with different number of features selected by Chi-squared value criterion on LRE07 30s

4.2.3 Higher order n -gram features

With the above-mentioned feature selection methods, the dimension of the 3-gram feature vectors is reduced to less than half of its original size. We are able to expand the selected features into higher order n -gram, such as 4-gram. That is, if a 3-gram $x_1x_2x_3$ is selected in the feature selection, when we incorporate 4-gram features, the 4-gram feature $x_1x_2x_3x_i$ and $x_1x_2x_3x_j$ will be included in feature vector. (x_i denotes any phone in the phone inventory).

Table 2. EER (%) with different order of n -gram and number of selected features

n-gram order	LRE07		LRE03	
	EER(%)	# Features	EER(%)	# Features
Baseline 3-gram	2.11	88292	1.35	83165
SVM-margin 3-gram	2.10	40000	1.03	30000
SVM-margin 4-gram	1.84	279299	0.92	257312
Chi-squared 3-gram	2.08	37000	1.04	30000
Chi-squared 4-gram	1.88	264396	0.95	259871

The feature selection thresholds derived for LRE07 language recognition system are applied in LRE03 system. As there are fewer target languages in LRE03, less number of

features actually observed in LRE03 training and testing sets. Table 2 reports the performance of 3-gram and 4-gram features on the LRE07 and LRE03 30 seconds closed test sets. The numbers of selected features, which are the dimension of resulting feature vectors, are also shown. The feature selection is conducted with both SVM-margin and Chi-squared criteria. The 4-gram features are derived from those 3-gram features which have been selected due to their high merits. It is clearly shown that the performance is improved, as the higher order n -gram features are included. The proposed two feature selection methods achieved comparable performance in feature selection.

5. Conclusion

This paper studies the way to select phonotactic feature with the criteria of separation margin and Chi-squared measure for spoken language recognition. We propose to use minimum mean difference point to make the decision of the threshold setting in features selection. With the efficient and effective feature selection strategy, much lower dimensional features are obtained without affecting the language recognition performance. A higher order n -gram features can then be derived from the selected features and incorporated into the language modeling process, and thus a better accuracy can be achieved.

Features selected from SVM-margin and Chi-squared methods achieve comparable performance in language recognition. The Chi-squared method measures the individual feature's contribution to the language separation, while the SVM-margin method considers the contribution of each feature in the group of features. In the future works, we would like to study the way to combine two methods in efficient phonotactic feature selection for language recognition.

6. References

- [1] P. Torres-Carrasquillo, E. Singer, W. Campbell, T. Gleason, A. McCree, D. Reynolds, F. Richardson, W. Shen and D. Sturim, "The MITLL NIST LRE 2007 Language Recognition System", *Interspeech 2008*, pp. 718-722, 2008
- [2] R. Tong, B. Ma, H. Li and E. S. Chng, "A target-oriented phonotactic front-end for spoken language recognition", *IEEE Transactions on Audio, Speech and Language Processing*, Volume 17, issue 7, pp.1335-1347, Sept, 2009
- [3] F. S. Richardson, W. M. Campbell, "Language recognition with discriminative keyword selection", *ICASSP 2008*, pp. 4145-4148, 2008
- [4] J. L. Gauvain, A. Messaoudi and H. Schwenk, "Language recognition using phone lattices", *ICSLP 2004*, pp. 1283-1285, 2004
- [5] P. Matejka, P. Schwarz, J. Cernocky, P. Chytil, "Phonotactic Language identification using high quality phoneme recognition", *Interspeech 2005*, pp. 2237-2240, 2005
- [6] 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.
- [7] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization", *The Fourteenth International Conference on Machine Learning*, pp 412-420, 1997
- [8] P. Schwarz, P. Matejka and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition", *ICASSP 2006*, pp. 325-328, 2006
- [9] T. Lander and R. Cole and B. Oshika and M. Noel, "The OGI 22 language telephone speech corpus", *Eurospeech 1995*, pp. 895-898, 1995