# Literature Study on Spoken Language Identification

**Suping Shi**
199401172862
suping@kth.se

**Yufei Zhu**
199408046085
yufeizhu@kth.se

## Abstract

This paper introduces three methos of language identification, which are Parallel Phoneme Recognition followed by Language Modeling(PPRLM), Phoneme Recognition followed by Language Modeling(PRLM), and GMM method. PPRLM system is based on the HMM model with extracting phonotactic information. The difference between PRLM and PPRLM is that PRLM only uses one phoneme recognizer. The last method is using GMM to classify languages based on acoustic information. GMM method needs less computational work and shorter time for identification. Comparing of HMM methods, we can find that PPRLM has an advantage in both accuracy and correction rate.

## 1 Introduction

The language identification(LID) task is an important branch of automatic speech recognition. Building of language model helps with narrowing the range of searching and then increase the accuracy of speech recognition. The purpose of LID is to identify the language spoken in a specific utterance.

The language identification meets some challenges. First is the feature extraction. The language system is grant and various with dialects and different accent. The speaker identity also affects the results. Hence it is significant to extract the general character for one language. Second is the vocabulary size. With limited computer resource and computing time, it is impractical to build an complete model of a language, which means if we test the model with some rare and special materials, the percentage of accuracy will decrease obviously. The third one is the unknown content of test material. The test material may not be included in the corpora. If the content and the way of pronounce is similar with another language, it will be difficult for the model to recognize[1].

Before recognition, we need to know the differences among different languages. In the conventional methods, the speech material is first converted into a sequence of phonemes that can characterize the language with phonemes recognizer. A phoneme is the smallest unit in a language here. Then an n-gram language model need to be applied to estimate the probability of the appearance of a particular phoneme sequence. Parallel Phoneme Recognition followed by Language Modeling (PPRLM) and Phoneme Recognition followed by Language Modeling(PRLM) are both using this difference to characterize different languages.

In this paper, we introduce two methods based on phonotactic information, which are PPRLM and PRLM. And we introduce one method which is based on acoustic information and is relied on GMM. Then we discuss the difference between three methods in the discussion and conclusion part.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

## 2 PPRLM

PPRLM is based on HMM model and extract the phonotactic information of one language. Figure 1 is the block diagram to show the principle of PPRLM system. With a language-dependent phoneme recognizer, the input speech signal is converted to a sequence of phonemes and with the statistics of the sequences, the language model can be trained.

Zissman and Singer used a single English phoneme recognizer and proved that it was feasible to model phonotactic character with the information of phoneme sequences from one language. Zissman used a single language-dependent phoneme recognizer to convert the input speech signal to a sequence of phonemes and used the statistics of the resulting symbol sequences for language identification. Then in the further research, he extended it using multiple language-dependent phone recognizers in parallel[1,2].

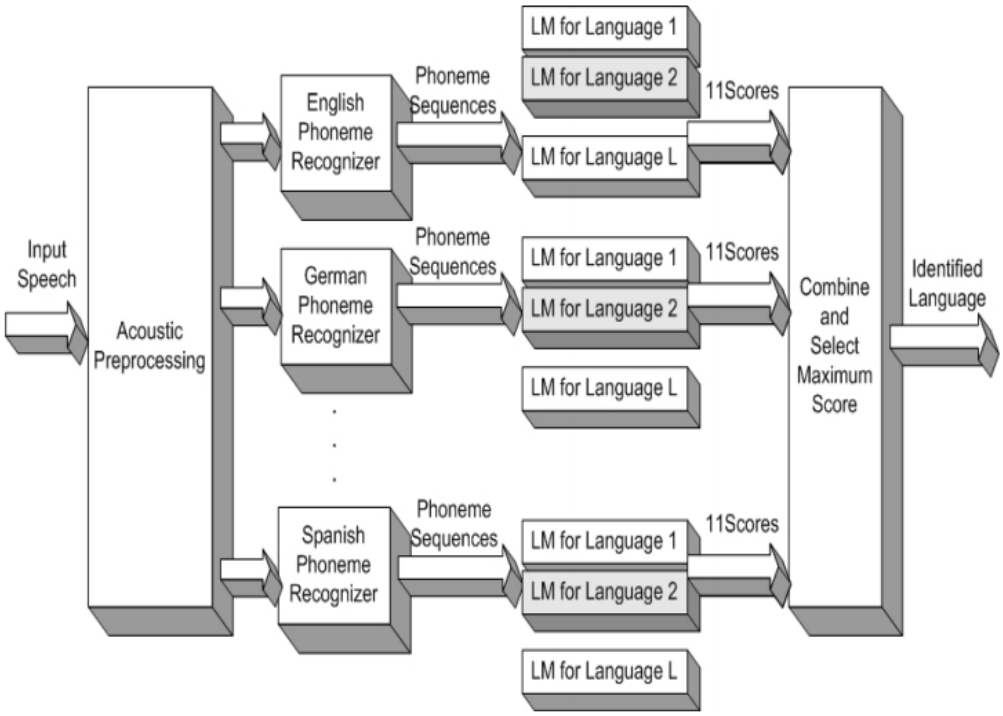


Figure 1: The process of training the PPRLM model[3]

### 2.1 Phoneme recognizer

Phoneme recognizer is the most important part in PPRLM LID system. It converts the speech utterance to a sequence of phoneme symbols. In PPRLM, phoneme recognizer is trained based on one language. OGI stories is widely used in training phoneme recognizer. Results of these phoneme recognizers is shown in terms of phoneme error rate[3].

### 2.2 Language Model

With a given phoneme recognizer, an n-gram language model is employed to estimate the probability of the occurrence for a particular phoneme sequence. The n-gram language model estimates the probability of the occurrence of a particular phonetic event.

The identification of a language out of all the different languages through a single phoneme recognizer is performed by the maximum likelihood classifier decision rule[4].

### 2.3 Discounting Methods

Due to the constraints of limited data, discounting methods are used to make the probability distributions more uniform, by adjusting the probabilities. If the actual number of occurrences of an event $E$ is $r$, then the modified count is $r * d(r)$, where $d(r)$ is discount ratio.

Hence the revised probability estimates is $P(E) = r * d(r)/R$. The most suitable method in LID in Witten-Bell discounting scheme.

In Witten-Bell discounting method, the discounting ratio is not dependent on the event's count, but on the number of distinct events which follow a particular context. The formula of Witten-Bell discounting method is : $d(r, s) = \frac{R}{R+s}$, where s is the number of distinct n-grams in the model[3].

## 3 PRLM

The approach of PRLM is similar to PPRLM. We use single phoneme recognizer instead of paraller phoneme recognizers.

In PRLM, phoneme recognizer is trained base on single language. Multi-lingual PRLM is using the phoneme recognizer trained on several languages. The specific information of language is removed, such as the dialect label and specific and unique phoneme symbols. The phoneme recognition result(the rate of correction and accuracy ) for PRLM is worse than the one for PPRLM. Equal Error Rate(EER) is applied for evaluating LID performance. PPRLM outperform PRLM[3].

## 4 GMM

As we can see from the previous method of identifying the language. HMM is usually used to extract the phonotactic content of a speech signal[3]. And then we use the phoneme we recognized to compare with the language model that we trained. But the product from the phoneme recognizing step is not what we want. We just use it as a tool to compare with the language model. In the way of PRLM, even though we just use one phoneme recognizer to do the recolonization part. It is still a time consuming and money consuming thing. So we thought of only using GMM as our tool to recognize the general characteristic of a language. Here GMM is planed to be used to classify the language using acoustic content of the speech signal. In order to describe our method of identifying the language, we will illustrate our plan of this method with the following example: identifying the language of Enlish Russian and French.

### 4.1 Corpora

As the goal of the project is to accomplish the identification of the different languages, we choose three languages as examples for experiment
1. English
2. Russian
3. French
The total training time for each language is at least 30 minutes. The details are showed in Tables.

| | number | Time |
|---|---|---|
| Man | 90 | 961 |
| Woman | 90 | 842 |

Table 1: English

| | number | Time |
|---|---|---|
| Man | 90 | 965 |
| Woman | 90 | 877 |

Table 2: Russian

| | number | Time |
|---|---|---|
| Man | 90 | 991 |
| Woman | 90 | 890 |

Table 3: French

These is downloaded from the Voxforge. In order to guarantee the diversity of the training material. We will choose the utterances from both male and female. Also, we should take the accent of the language into account. Speaker variability and pronunciation variations have been a topic for researchers for some time. Take English as an example, we choose the utterances of American english and British english. Those variability can help to improve the accuracy of the training model. All the utterances are with the prompt of it, which means we know what exactly are those audios saying as a reference. The corpora for each language is divided into two groups. One group is for training the data, and the other group is for the testing. The rate of the number contained in the training and testing groups is about 8:1.

### 4.2 Feature extraction

After adaption of the utterances we get from the corpora, we will come to the step of features extraction. Mel frequency cepsrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power transform on a nonlinear Mel scale of frequency. MFCC are the coefficients that make up the MFC. We first convert the raw waveform data into vectors of features that are easy to be processed. We plan to start this step with extracting 13 MFCC features from input speech using predefined methods in HTK.

### 4.3 Model Training

The language models will trained with 3 languages we get from last step with the help of matlab. We will use the Expectation-Maximization (EM) algorithm for the training of the GMM model. This process is a iterative one. Just like K-means. Then there may be a problem that it may converges to one point. That may result in zero variance. So we need to initialize this process carefully with proper number of components. We choose 9 to be the number of the Gaussian components. Then we will train all language models with 9 Gaussian components for English, Russian and French. The algorithm runs for about 15 iterations with the MFCC vectors from the corpus for 3 languages respectively. After this step, the three languages will have their own models. In order to identify each model, we will remove the same component in the Gaussian mixture model for each languages and keep the different parts. Then the models we get can represent the uniqueness of each kind of language.

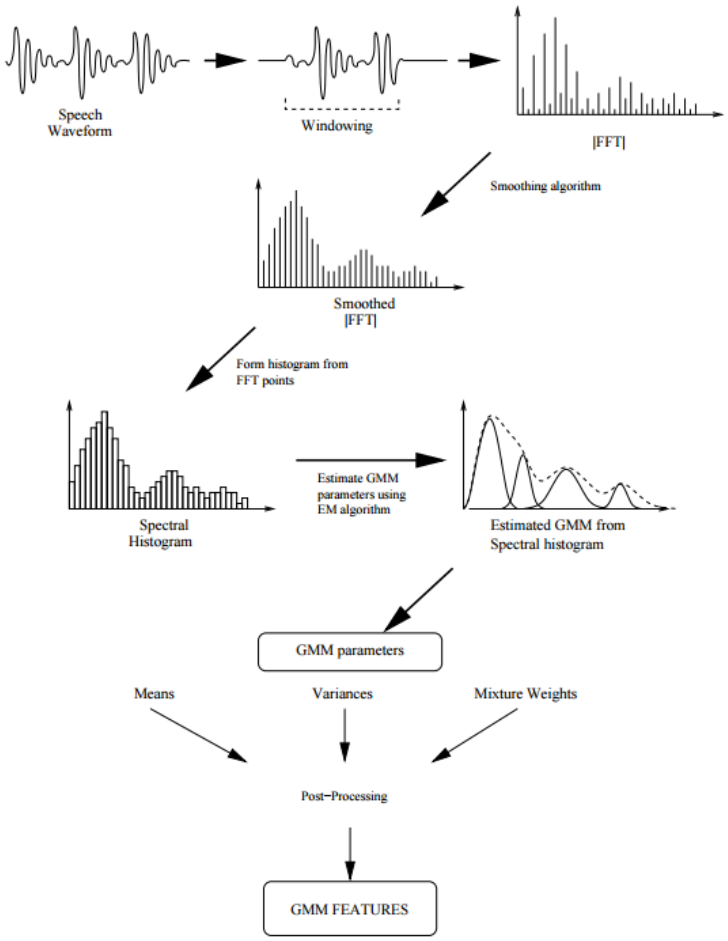


Figure 2: The process of training the Gmm model[5]

### 4.4 Language Identification

After all those steps above, we will use the testing input materials to test the models we get. And use the training models that we get from the 'model training step' to identify the languages. The log-likelihood of the test material with the model will be computed. And the MFCC vectors will be identified as the language who gives the largest likelihood.

## 5 Discussion and conclusion

In this paper we mainly study important ways of identifying the language PRLM and PPRLM. This two methods are mainly depend on the phoneme recognition. Which means in those two methods we have to first use the phoneme recognizer to extract the phonotactic content of a speech signal. And then we use the phoneme sequence that we get from this step to compare with the language model of each language. The language with the highest score will be chosen. In PRLM we only use one kind of phoneme recognizer to produce the phoneme sequence. Every input voice signal will run through one kind of language phoneme recognizer. Even the input language and the language of the phoneme recognizer are not matched. It will also produce a phoneme sequence to be compared. And then we use the phoneme to identify which language model will fit the best to decide which language it is.

While in the method of PPRLM. We use multi-lingual phoneme recognizers to recognize the phoneme of the input voice signal. And each phoneme recognizer will produce one phoneme sequence. Then we use all the phoneme sequence to compare with the language model. In PPRLM, the last step of the identifying process is more like a voting process other than just depend on the score. This step obviously increase the accuracy of the language identifying. Though several phoneme recognizers are used in the recognition part, we will not spend more time on the phoneme recognition part for the reason that we run them in a parallel way. We all know every language has a huge vocabulary and complicated grammar rules since they are formed through a tremendous long time. And one language can derive many branches with different accent. Hence the feature extraction can never be totally comprehensive.And a huge corpora is needed for this method. Based on the modelling of language mentioned above, we can also add some particular and apparent character of each language, such as the /r/ sound in Italian and /er/ in Chinese.So that we throw out the GMM method in the last part. But it is impossible to extract the general characteristics for a language with limited corpora. If we accomplish that work we think it may helps improve the speed and the accuracy of recognition.

## References


[1] K.Screenivasa Rao & V.Ramu Reddy &Sudhamay Maity Language Identification Using Spectral and Prosodic Features *Springer International Publishing 2005*

[2]M.A. Zissman Comparison of Four Approaches to Automatic Language Identification of Telephone Speech *IEEE Transactions on Speech and Audio Processing*,vol.4,no.1, 1996

[3] Liang Wang & Eliathamby Ambikairajah &Eric H.C. Choi(2006) Multi-lingual Phoneme Recognition and Language Identification Using Phonotactic Information.*The 18th International Conference on Pattern Recognition*

[4]Pavel Matejka & Petr Schwarz &Jan Cernocky & Pavel Chytil. Phonotactic Language Identification using High Quality Phoneme Recognition *Proc. Eurospeech, Sept. 2005.*

[5] Matthew Nicholas Stuttle A Gaussian Mixture Model Spectral Representation for Speech Recognition. *2003.University of Cambridge*