SPOKEN LANGUAGE IDENTIFICATION USING BIGRAM

CHENG Xuelin, WU Kaizheng, WANG Han, LI Zongge
Department of Computer Science Fudan University, Shanghai

hakulamutata@sohu.com

ABSTRACT

The task of automatic language identification (ALI) system is to distinguish the incoming utterances between different languages. In this paper the decoding bigram and extended bigrams of each language are exploited to interpret the characteristics of languages. In the final system which includes four languages, i.e. English, Mandarin, Japanese and Spanish, the phone sequences that are outputted by phone recognizers using viterbi algorithm over decoding bigrams are fed into extended bigrams, and based on the language scores the classifier makes a maximum decision. At last the system combined with decoding bigrams and extended bigrams shows an improvement of 21.2% over that with null grammar, especially the high identification rate of 96% between Mandarin and Spanish.

1. INTRODUCTION

Automatic language identification is usually used as the preprocessing in a wide range of applications such as multi-language translation system. The best way is to build large vocabulary speech continuous recognition system for each language [1][2]. However it needs a complete word recognizer, and the language models are based on the word and sentence level. Furthermore it is difficult to add a new language. Prosody modeling as well as duration information [10] can also acquire the inherent information of languages. For example the tones of Mandarin are special, and is discriminative to other languages. In [3][4], Statistical probabilities in phone sequences are another efficient way to demonstrate the discrimination between different languages. They are transition probabilities from one phone to another. Two kinds of statistical constraints -- decoding bigram and extended bigram are exploited to improve the performance of automatic language identification. Using Viterbi algorithm, the phone recognizers tokenize the incoming utterance with decoding bigrams. Then the phone sequences are fed into extended bigrams. They can be viewed as the mapping sets defined by one language through another language phone recognizer.

In section 2 a description of baseline language identification system with null grammar is given. The decoding bigram and extended bigrams of each language are dealt with in section 3. The results of experiments obtained by different systems are compared in subsequent section.

2. BASELINE SYSTEM

In the baseline system language-dependent phones are selected according to the phone transcriptions that can capture the allophonic variation. In purely phonemic transcription voiceless stop /t/ can not differentiate the phonetic variation included in contexts (like /t/ in “time” as opposed to /t/ in “too”). Basically there are two standards in multi-language transcription, IPA and Worldbet. The difference between them are purely academic and do not affect the training of spoken language system at all. In our system, the OGI Multi-language Telephone Speech Corpus [6] and the Worldbet are used.

In the phone recognizers of four languages 58 phones are selected for English, 47 for Mandarin, 33 for Japanese and 44 for Spanish. A tri-state left-to-right HMM architecture for the phones is shown on figure 1. Each model has entry state and exit state, which are designed for the convenience of linking different models in training and do not contribute to the probability of model.

Figure 1: Tri-state left-to-right HMM architecture

The extracted frame vectors from utterances are 39 dimensional, and composed of 13 Mel-frequency cepstral coefficients, their delta and acceleration coefficients. To reduce the channel effects the cepstral mean normalization is performed. There are three steps [5] to train the HMMs. In the first step, uniform segmentation initializes the parameters of HMMs, then Viterbi algorithm is used to find the most likely state sequence and HMM parameters are estimated. In the second step, using the Forward-Backward algorithm the Baum-Welch training finds the probability of being in each state at each time frame and update the parameters. At last, embedded training links all HMMs in terms of transcription and constructs a composite HMM, which spans the whole utterance. The embedded Baum-Welch re-estimation is then used to update the HMMs.
The baseline system architecture is shown on figure 2 where only two languages English and Mandarin are described. The incoming utterance is decoded by different phone recognizers, and the system chooses the language that has the maximum output probability as the output.

![Figure 2: Baseline system architecture](image)

### 3. BIGRAM MODELS

The probability of phone sequence is decided by

$$\hat{B}^i = \arg \max_{B^i} P(B^i | O)$$  \hspace{1cm} (1)

where $B^i = b^i_1, ..., b^i_T$ is the phone sequence decoded by language $i$, and $O$ is the observation vectors. According to Bayesian equation the above $P(B^i | O)$ is defined as

$$P(B^i | O) = \frac{P(B^i) P(O | B^i)}{P(O)} = P(B^i) P(O | B^i)$$  \hspace{1cm} (2)

where $P(O | B^i)$ is the probability of the observation vectors $O$ when the phone sequence is $B^i$. It is corresponding to acoustic model. $P(B^i)$ is the prior probability for language $i$. It is corresponding to language model. In the baseline system $P(B^i)$ is same to every phone sequence. In order to make better use of language models, the back-off bigram models are built to capture the information inherent in languages. The probability $p(i, j)$ of adjacent pair of phones $i$ and $j$ is given by

$$p(i, j) = \begin{cases} \frac{(N(i, j) - D)}{N(i)} & \text{if } N(i, j) > t \\ b(i) p(j) & \text{otherwise} \end{cases}$$  \hspace{1cm} (3)

where $N(i, j)$ is the total number of pair $(i, j)$ in transcriptions of training data, $\theta$ is the number of distinct labels, $D$ is a smoothing factor and $t$ is a threshold. When $N(i, j) \leq t$, the bigram is backed-off to unigram probability $p(j)$ scaled by a weight $b(i)$ which is given by

$$b(i) = \frac{1 - \sum_{j \neq h} p(i, j)}{1 - \sum_{j \neq h} b(i, j)}$$  \hspace{1cm} (5)

where $B$ is the set of all phones for which $p(i, j)$ has a bigram. It can be proved that $\sum_{i=1}^{\theta} p(i, j) = 1$ . The back-off bigram is shown on figure 3.

![Figure 3: Back-off bigram](image)

The dashed line implies that there is no bigram between phones $i$ and $j$, so it is replaced by multiplication of weight $b_i$ and unigram $p_j$. The back-off node on figure 3 is the transfer node between phones that have no bigram to connect.

The bigrams of language models are used to weight the transition probabilities from one phone to another in trellis searched by viterbi algorithm, so they can be called decoding bigrams. The figure 4 shows the architecture after they are added into the baseline system.

![Figure 4: Language identification system with decoding bigram](image)

If the incoming utterances recognized by above language identification system are from English, two sets of phone sequences are constructed. One is English-English set
represented by English phones, and the other is English-Mandarin set represented by Mandarin phones. If the phone recognizer and decoding bigram of a language form one pipe, they can be viewed as different mapping sets from the same input through different pipes. The extended bigrams are trained by them. The language identification system shown on figure 5 is the system added by decoding bigrams plus extended bigrams.

Figure 5: Language identification system with decoding bigrams plus extended bigrams

There are \( N^2 \) extended bigrams for \( N \)-language identification system. The score through the extended bigram given the language \( j \) is based on the following equation

\[
E_{i,j}(B' | L_j) = \frac{1}{T} \left\{ \log P(b'_1) + \sum_{i=2}^{T} \log Q(b'_i | b'_{i-1}, L_j) \right\}
\]

where \( B' = b'_1, ..., b'_T \) is the phone sequence recognized using decoding bigram of language \( i \). \( Q \) is an interpolated bigram score given by

\[
Q(b'_i | b'_{i-1}, L_j) = (1-\beta)P(b'_i | b'_{i-1}, L_j) + \beta P(b'_i | L_j)
\]

The score \( S_j \) calculated using the following equation for language \( j \) is fed into maximum likelihood classifier

\[
S_j = \sum_{i=1}^{N} E_{i,j}
\]

where \( N \) is the number of languages in the system.

100 utterances from each language are decoded by the system shown on figure 4 and the resulting phone sequences serve as the data for training the extended bigrams. The intersection of it and the data for training the phone recognizers is null.

4. EXPERIMENTS

The OGI Multi-language Telephone Speech Corpus was used for training the systems that were built with the tool HTK [5]. The systems included four languages: English, Mandarin, Japanese and Spanish. The test set contained 200 utterances in each of the four languages whose lengths were between 7 and 10 seconds. The results of different systems are compared in table 1 to 4. At last the identification rate for four-language system was 83.6%.

Table 1: Identification rates for English utterances

<table>
<thead>
<tr>
<th></th>
<th>Jap/En</th>
<th>Jap/Man</th>
<th>Jap/Spa</th>
<th>Four Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>En</td>
<td>89.5%</td>
<td>77.5%</td>
<td>94.5%</td>
<td>77%</td>
</tr>
<tr>
<td>Jap/Man</td>
<td>88%</td>
<td>80%</td>
<td>94.5%</td>
<td>78.5%</td>
</tr>
<tr>
<td>Jap/Spa</td>
<td>90%</td>
<td>94%</td>
<td>98%</td>
<td>88%</td>
</tr>
</tbody>
</table>

Table 2: Identification rates for Japanese utterances

<table>
<thead>
<tr>
<th></th>
<th>Man/En</th>
<th>Man/Jap</th>
<th>Man/Spa</th>
<th>Four Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jap/En</td>
<td>72%</td>
<td>63.5%</td>
<td>75.5%</td>
<td>40%</td>
</tr>
<tr>
<td>Jap/Man</td>
<td>81%</td>
<td>70%</td>
<td>79.5%</td>
<td>48.5</td>
</tr>
<tr>
<td>Jap/Spa</td>
<td>95%</td>
<td>91.5%</td>
<td>93.5%</td>
<td>83%</td>
</tr>
</tbody>
</table>

Table 3: Identification rates for Mandarin utterances

<table>
<thead>
<tr>
<th></th>
<th>Spa/En</th>
<th>Spa/Jap</th>
<th>Spa/Man</th>
<th>Four Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>En</td>
<td>71%</td>
<td>80%</td>
<td>77%</td>
<td>57.5%</td>
</tr>
<tr>
<td>Spa/Jap</td>
<td>81%</td>
<td>77.5%</td>
<td>78%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Spa/Man</td>
<td>92%</td>
<td>88%</td>
<td>95.5%</td>
<td>83.5%</td>
</tr>
</tbody>
</table>

Table 4: Identification rates for Spanish utterances
The values in the first row of the tables are the identification rates for null grammar, the second for decoding bigrams and the third for decoding bigrams plus extended bigrams. Furthermore in each table the identification rates of two languages are compared in the first three columns from left, and those of four languages in last column. From the tables we can see that the system with decoding bigrams and extended bigrams outperforms the system with null grammar by 21.2%, and the identification rate between Spanish and Mandarin is 96%. Because there was the enough training data for English, the system performed best for it as a whole. However Mandarin was not better identified than other languages. The higher-level linguistic information the system uses, the better performance can be achieved. If the tones are considered, the performance is improved further. In table 2, after adding the extended bigrams the identification rate of Mandarin and Japanese for Japanese utterances markedly increased, while there was a decrease for Mandarin utterances in table 3. Thus, the whole identification rate of Mandarin and Japanese was still increased by adding extended bigrams.

5. CONCLUSION

In this contribution displacing the LVSCR based language identification approach, phonotactics proved to be an efficient way to capture the information inherent in languages. Back-off bigram is used to build the language models and extended bigram models. Finally, tested on the OGI Multi-language Telephone Speech Corpus by adding the decoding bigrams and extended bigrams the system including four languages shows an improvement of 21.2% over that with null grammar, especially the high identification rate of 96% between Mandarin and Spanish. Because the system is built on phone level, it needs less computational expense than LVSCR based language identification system, which broadens the range of applications.

In the future study the phonotactics, prosody and duration information can be combined to increase the identification rates between languages. Compared to bigram, the trigram contains more information in languages, so it will further improve the system performance.

6. REFERENCES


