A PROMISING SYLLABLE DECOMPOSITION METHOD FOR TONAL LANGUAGES' SPEECH RECOGNITION

Haiping LI, Liqin SHEN, GuoKang FU
IBM China Research Lab, Beijing
lihp, shenlq, fugk@cn.ibm.com

C.J. CHEN
IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.
juchen@us.ibm.com

ABSTRACT
A new syllable decomposition method, which uses the tone information of the main vowel in a syllable to distinguish the tone of the whole syllable, is proposed in this paper. Compared to the scheme, in which a syllable is decomposed into an initial and a final, and the tone information is carried on by the final, the new scheme reduces the number of phonemes in the phone set of a recognition system. It handles the syllabic languages especially the ones with complicated tonal phonology such as Cantonese successfully, and also can be generalized to other tonal languages. Experiments on both Cantonese and Mandarin to compare the performance of systems using these two schemes, lead to that the new scheme got a little bit better accuracy than the old one while reduces the number of phonemes dramatically in recognition system. Such a method is promising to be used in more real speech recognition system or product.

1. INTRODUCTION
The tonal language is an important category of languages, which includes varied kinds of Chinese languages (such as Mandarin, Cantonese, Taiwanese), South Asia languages (such as Thai, Vietnamese), Swedish, and Norwegian. Compared with English and some other European languages, tonal languages have a unique but important property, tone information. Speech Recognition of tonal languages depends not only on the phonetic composition but also on the lexical tone pattern. For example in Chinese language, words with same syllable sequences but different tones have different meanings. So without tone information, the speech recognition accuracy will not be good, especially in the above cases where tones are the only distinguishing features.

The tone information can be represented with various pitch contours and the relative pitch level. Nowadays, in most of the successful tonal language recognition systems of continuous speech, tonal phonemes are used as the composition of phone set, taking the advantage that the intrinsic synchronization and the probability normalization of syllable and tone can be embodied easily in this way. In such systems, pitch is added as one additional feature above other selected speech features.

Different speech recognition systems have different phone sets. A popular method to decompose tonal syllable, which is used in syllable based tonal language especially in Mandarin, is splitting a syllable into a toneless initial and a tonal final, which carries the tone information of the whole syllable. Such a scheme leads to successful Mandarin speech recognition systems. [1] But when generalizing it to other tonal languages with more finals and tones than Mandarin, such as Cantonese, a much larger phone set is induced. The larger the phone set is, the much more training data is required.

A promising new scheme to decompose syllable is proposed in this paper. Based on the above scheme, it can be supposed that the tone information in the main vowel is sufficient to determine the tone of the whole syllable. Then the final semi-syllable can be further decomposed into a main vowel and a coda. Using this syllable decomposition method, the phonemes are much less than that using the former.

In the next section, these two schemes will be described in detail and compared. In section 3, the case of Cantonese is analyzed in order to show the advantages of this new scheme when it's used in a tonal language with complicated phonology. In section 4, experiments on Cantonese are carried out to compare the performance of systems using these two schemes. In addition, test results on Mandarin are also given there to confirm the conclusions in section 5.

2. THE SYLLABLE DECOMPOSITION METHODS

2.1 the Scheme Based on Using Tone Information of Final Semi-Syllables
Generally the tonal systems are composed of toneless consonants and tonal vowels. Accordingly, a syllable can be decomposed into a toneless initial and a tonal final in the syllable based acoustic structure of Chinese, including Mandarin and Cantonese. The initial is usually a consonant or sometimes a compound of consonant and a glide. The final is usually a compound of main vowel and an ending consonant or sometimes a single main vowel. It can be seen that the initials carry little tone information, and the finals carry almost all the tone information. Thus a
syllable is decomposed to a toneless initial and a tonal final, which results in a phone set comprising tonal phonemes. For example in Mandarin, the syllable “Wang2” is decomposed as following.

\[
\text{Wang2} \rightarrow W + \text{ANG2}
\]

Here, syllable “Wang2” is decomposed into the non-tonal initial “W” and the tonal final “ANG2”. It can be seen that each final will produce N various phonemes (tonal finals), where N is the number of tone variations. In Mandarin, there are 21 finals, including single, compound and with nasal-suffix ones. If each has 5 possible tones, there are 105 tonal phonemes.

Though there’re some successful Mandarin recognition systems using this scheme, many difficulties will rise when generalize it to a tonal language with complicated phonology such as Cantonese, the dialect commonly used in South China including Hong Kong. For Cantonese, because of more vowels and tones, the phone set generated by this decomposition method is very large. If 9 or 11 actual tone variations in Cantonese are merged into 6 ones, there still are more than 200 finals.

### 2.2 the Scheme Based on Using Tone Information of Main Vowel in Syllable

Based on the above scheme, it can be supposed that the tone information in the main vowel is sufficient to determine the tone of the whole syllable. Then the final semi-syllable can be further decomposed into a main vowel and a coda. Though the coda continues to carry the tone, the tone information embodied in the main vowel can be distinguishable from others.

So in this new method a syllable is decomposed into 3 parts, an onset (initial), a nuclear with tone, and a coda without tone. For the above example “Wang2”, the decomposition is,

\[
\text{Wang2} \rightarrow W + A2 + \text{NG}
\]

Here, syllable “Wang2” is decomposed into the onset “W”, the tonal nucleus “A2”, and the coda “NG”.

Suppose in a phonology of tonal language, there are x onsets, y1 nuclei, y2 codas, and N tones, we can predict the maximum number of phonemes S2 using this decomposition method.

\[
S_2 = x + y_1 \times N + y_2
\]

In the former method, the maximum number of finals y is,

\[
y = y_1 \times y_2
\]

And the maximum number of phonemes S1 is,

\[
S_1 = x + y \times N = x + y_1 \times y_2 \times N
\]

It’s seen from the above that the phonemes S1 are much more than S2.

In the Chinese Pinyin Scheme, which is the standard Mandarin phonology, there are 8 nucleus, 5 codas and 5 tones. Using the new scheme we get 45 phonemes corresponding to 105 tonal finals in the first scheme.

### 2.3 Advantages of the New Syllable Decomposition Method

The smaller phone set can bring much benefit. With the same amount of training data, the phonemes in smaller phone set can be trained more sufficiently.

When decompose syllables into phonemes, there are always some phonemes, which only exist in several syllables or even one. Such phoneme is called rare phoneme. The models for the rare phonemes are usually not so robust as the phonemes with high appearance frequency because it’s difficult to collect enough training data for rare phonemes. In the new scheme, the rare phonemes are much less or even not exist, which can bring great facility to design training scripts, training and decoding.

In addition, phonemes from the new method are closer to the ones of western languages on such property as duration. This makes it take advantage to be used in a multilingual speech recognition system.

### 2.4 Future Consideration

If the tone recognition accuracy is required to be more precise, the tone information of coda can be added. Thus a syllable is decomposed into such 3 parts, an onset, a tonal nuclear, and a coda with tone. Also for the example “Wang2”, the decomposition is,

\[
\text{Wang2} \rightarrow W + A2 + \text{NG2}
\]

Then the tonal codas are added into the phone set. Using the same assumption of phonology in section 2.2, the maximum number of total phoneme S3 is,

\[
S_3 = x + y_1 \times N + y_2 \times N = x + (y_1 + y_2) \times N
\]

Though S3 is larger than S2, it is still much smaller than the phone set size S1 got by using the first syllable decomposition method. So the merits of the new decomposition scheme will be kept.

### 3. IMPLEMENTATION ON CANTONESE DIALECT

Cantonese, an important dialect used in China, has the common property of the Chinese languages, being monosyllabic and tonal. It has much more finals than Mandarin and particularly is rich in tones. In this case, using the new scheme shows more advantages than using the old one.

#### 3.1 Phonology of Cantonese

“Cantonese” is originally referred to the Guangzhou dialect, and now the term used to represent the standard of “Yue” Dialect. There are varied kinds of romanization schemes for Cantonese. The phonology used in this paper is referred to the Cantonese Romanization Scheme of the Linguistic Society of Hong Kong [2], which is a very popular one used in Hong Kong. It consists of 20 initials and 59 finals. The initials include consonant, glide, nasal, and a zero initial. And the finals are made up of syllabic components, which are usually a vowel or syllabic nasal and an ending, which is a glide, a nasal or a glottalized stop consonant.
Compared with 21 finals in Mandarin, there are more phonetic combinations of initials and finals in Cantonese.

When the tones are introduced into Cantonese, the situation becomes more complicated. The nine tones theory is the most common one in Cantonese (Table 1). These nine tones can be categorized into two groups. One is called non-entering group including tone 1-6, another is called entering group including tone 7-9. The syllables with entering tones have a common property, ending with a glottalized stop consonant such as /p/, /t/, and /k/. Thus their duration is shorter than the syllables with non-entering tones. And from Table 1, it's found that the pitch values of the three entering tones 7-9 are respectively very similar to those of tone 1, 3 and 6.

Table 1. Tones in Cantonese

<table>
<thead>
<tr>
<th>Groups</th>
<th>Tone Num.</th>
<th>Tone Name</th>
<th>Five Level System</th>
<th>Pitch Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Entering Tones</td>
<td>1</td>
<td>High Level</td>
<td>55 or 53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>High Rising</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Mid Level</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Low Falling</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Low Rising</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Low Level</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>High Entering</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Entering Tones</td>
<td>8</td>
<td>Mid Entering</td>
<td>3 or 33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Low Entering</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

When the new decomposition method is implemented on such syllables, the codas of syllables with tone 1, 3 and 6 are chopped, and the stop endings of the syllables with entering tones are also chopped. So the duration of the main vowels in the above two cases are closer. In addition, because the main vowels keep the tone information of the syllable, the tone patterns of the main vowels with tone 7, 8 and 9 should be identical to that with tone 1, 3 and 6 respectively. Thus the 9 tones in Cantonese can be merged to 6 tones, which reduces the number of phonemes and makes the tone situation simpler.

3.2 Phone Set Got by Scheme Based on Using Tone Information of Final

In the Cantonese Romanization Scheme of the Linguistic Society of Hong Kong [2], there are 20 initials and 59 finals. Using the formula (3) in section 2.2, the maximum number of the phonemes is 20+59*6, thus the size of phone set is 374.

In fact, in this romanization scheme, if the tone is ignored, there are about 600 combinations of initials and finals. If only 6 tones are introduced, the syllabary of Cantonese will be expanded to more than 2000 syllabic sounds. Thus, decompose syllables using this scheme, about 300 phonemes should be defined. This results in a large phone set for Speech Recognition.

3.3 Phone Set Got by Scheme Based on Using Tone Information of Main Vowel

Use the new scheme to decompose Cantonese syllables, the number of phonemes will be reduced much. In addition, the 9 tones can be merged to 6 tones to simplify the Cantonese tone situation.

In this scheme, the 20 initials are kept as onsets, and the 59 finals are decomposed as 11 nuclei and 8 codas. The nuclei also include 2 syllabic nasals which can independently act as finals and define the tone of the syllable. Among the 8 codas, there are 5 ones from the syllables with non-entering tones, and the other 3 are the stop endings (/p/, /t/ and /k/) from the syllables with entering tones.

Using the formula (1) in section 2.2, the maximum number of the phonemes is 20+11*6+8, thus the size of phone set is 94. In fact, not every nuclear has 6 tones, so the real phone set will be smaller.

4. PERFORMANCE EVALUATION ON THESE TWO METHODS

4.1 Speech Recognition System for experiments

The structure of the Speech Recognition System used in our experiments is showed in Fig. 1.
Sharing the same language model, the experiments on both Cantonese and Mandarin case compared the performance of two Acoustic Models built on two phone sets. One is generated by syllable decomposition scheme using the tone information of the finals, the other is generated by scheme using the tone information of the main vowel of the syllable. Since language model is used in this system, the recognition accuracy of the output text is calculated on the character level instead of syllable.

4.2 Experiments on Cantonese Case

These experiments aim at two goals here. One is to compare the performance of two systems built on different syllable decomposition schemes. The other is to find out the influence of amount of training data on systems built on different schemes.

For the second goal, two training data sets are collected. The Cantonese speech data in the first training set is collected from 38 speakers in Guang Dong province of China. It contains 18,857 utterance of continuous sentence. The second training set is collected from 300 speakers in Hong Kong, and it consists of 29,964 sentences and about 21,000 words. This training set is larger than the former.

Accordingly, two test set are prepared. Each test set consists of 8 speakers, 4 males and 4 females, and each test speaker has 100 utterances for decoding. For the two acoustic models built on the first training set, the test speakers are from Guang Dong province, and for those built on the second training set, the test speakers are from Hong Kong. The recognition error rate is listed in table 2.

<table>
<thead>
<tr>
<th>AMs based on different Schemes</th>
<th>Training/test data from Guang Dong</th>
<th>Training/test data from Hong Kong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old scheme</td>
<td>15.99</td>
<td>12.92</td>
</tr>
<tr>
<td>New scheme</td>
<td>14.89</td>
<td>12.71</td>
</tr>
</tbody>
</table>

There are 20 test speakers in test set, 11 females and 9 males, and each one has 100 sentences. The result is shown in table 3.

<table>
<thead>
<tr>
<th>AMs based on different Schemes</th>
<th>Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old scheme</td>
<td>13.65</td>
</tr>
<tr>
<td>New scheme</td>
<td>12.61</td>
</tr>
</tbody>
</table>

5. CONCLUSION

The new scheme based on using the tone information of the main vowel to distinguish the tone of the whole syllable results in a phone set with less phonemes than that which consists of initials and tonal finals generated by the old decomposition scheme. This makes the speech recognition system using the new scheme reach the same accuracy as the old scheme while fewer training data is required. In order to further improve the accuracy, the tonal codas can be added and at the same time not enlarge the phone set much. Experiments should be performed to prove it next step.

For English and other western languages, the phone set scheme is commonly agreed. For Chinese, it’s far from to reach common conclusion for good phone set design scheme let alone concrete phone set. We hope our research in this area is a good start to this goal.

6. REFERENCES
