A Continuous Speaker-Independent Putonghua Dictation System

C. J. Chen, R. A. Gopinath, M. D. Monkowski, and M. A. Picheny

IBM Research Division, T.J. Watson Research Center
PO Box 218, Yorktown Heights, New York 10598, USA

ABSTRACT

We describe new methods for continuous putonghua speech recognition. We have augmented the IBM HMM-based continuous speech recognition system <1-3> with the following features: First, we treat tones in putonghua as attributes of certain phonemes, instead of syllables. We call those phonemes with tone tonemes. Second, instantaneous pitch is treated as a variable in the acoustic feature vector, in the same way as cepstra or energy. Third, by designing a set of word-segmentation rules to convert the continuous Chinese text into segmented text, the trigram language model works effectively. By applying those new methods, a speaker-independent, very-large-vocabulary continuous putonghua dictation system can be constructed.

INTRODUCTION

Chinese language speech recognition is particularly important due to the awkwardness of keyboard input. Voice input, or machine dictation, has been the "dream" input method for several decades. However, because of the unique features of Chinese language <4>, its progress has been slow:

1. Chinese languages (of various regions) are tonal languages, where syllables with the same consonant and vowel but different pitch contours are different phonetic syllables (representing different morphemes).

2. There are more than 4000 commonly used characters. The number of phonetically different syllables is much less. Even with tone distinction, the number is 1300. Some phonetic syllable may represent more than a dozen characters, creating a serious homonym problem.

3. Word boundaries are ill-defined, and do not exist in written text.

Chinese language speech recognition research has been pursued for more than 20 years by many institutions in China and elsewhere <5-7>. Because of the difference between Chinese and European languages, usually a completely different approach was taken. However, by studying the acoustic and grammatical structures of modern Chinese language, we found that by properly designing a phoneme system as well as a set of word-segmentation rules, Chinese languages can be mapped nicely into the mathematical structure of the speech recognition system initially developed for English <1-3>. By augmenting the IBM HMM-based speech recognition system to include Chinese-specific features, we have built a very-large-vocabulary continuous, speaker-independent putonghua speech recognition system.

In the following, we will briefly describe the basic concepts of the new approach. Then, we will present a series of recognition experiments with those methods.

TREATMENT OF TONES

Tone recognition has been the focal point of Chinese speech recognition for decades. A large number of research papers on this subject have been published <5-7>. The commonly used method is to recognize the base syllable (initials and finals) and tone separately: The base syllables are recognized by the conventional HMM-based method used (for example) in English. The tone of a syllable can be recognized by classifying the pitch contour of that syllable using discriminative rules. The recognition of toned syllables is a combination of the recognition of base syllables and the recognition of tones.
The above method, if possible in isolated-syllable speech recognition, is not applicable in the continuous case. First, in continuous speech recognition, the boundaries of the syllables are not well-defined. The boundaries are determined at the end of the entire recognition process. It is very difficult to provide syllable boundary information in the early stage of acoustic recognition. Second, the actual tone contour of a syllable with a given tone depends on the phonetic context. The rules to determine tones from the pitch contours, if possible, will be very complicated.

Our observations to the pitch contours of continuous Chinese language revealed the following facts: 1) The tone information is concentrated in the pitch behavior of the main vowel of a syllable. In other words, the pitch information of the main vowel alone determines the tone of the entire syllable. 2) The context dependence of pitch contour of a syllable can be expressed as the effect of the pitch contours of neighboring main vowels. 3) In continuous putonghua, both the average values of the pitch and the time-derivative of pitch near the center of the main vowel are equally important in determining the tone.

Based on the above observations, we found that the context-dependent quantization algorithm (1-2) used in recognizing phonemes (in languages without tones) provides an efficient way of recognizing tones in Chinese languages using the following recipe:

1. Treat the main vowel (or the latter part of a syllable including the main vowel) with different tones as different phonemes, called tonemes. The word toneme is not new. In Mario Pei's Glossary of Linguistic Terminology (8), a toneme is "a phoneme consisting of a specific tone in a tone language". For example, a, a, a, a, a (untoned) are five different tonemes.

2. Treat pitch as one of the acoustic parameters, same as cepstra or energy. Following the standard procedure of treating the acoustic data, the time-derivative of the pitch is naturally included in the (extended) acoustic feature vector as a component.

In the training process, the acoustic feature vectors of different tonemes under different phonetic contexts (if there is an effect) are grouped together. After the training process is completed, we looked into the pitch and the time-derivative of pitch associated to different tonemes. Figure 1 represents a typical result.

As shown in Fig. 1, the pitch contours of different tonemes in continuous speech are different from those in isolated-syllable speech. In isolated-syllable utterances, the pitch contours of the four tones are usually described as 55, 35, 214, and 15 <4>. In a continuous speech, the pitch contours of syllables with different tones are simplified and look symmetric. The four tones can be simply characterized as high (1), rising (2), low (3), and falling (4). Especially, the pitch contour of syllables with a low (3) tone does not show the distinct downward turn as in the conventional description. This difference can be explained as the following: Originally, in continuous speech, the four tones in putonghua are simple and symmetric. However, if a low-toned syllable is uttered as an isolated syllable and pronounced with a flat pitch contour, then it is not distinguishable from a syllable with a high tone. Thus, a downward curve is created to make it distinct.

![Figure 1](image-url)  
**Figure 1.** Typical means of instantaneous pitch and pitch derivative. The pitch and pitch derivative values of five tones (including the neutral), averaged over all tonemes (vowels or combinations of vowel and nasal ending) in putonghua (standard spoken Chinese). Each vowel, as a phoneme, is represented by three HMM states. The numbers 1, 2, and 3 indicate the sequence of HMM states in a single phoneme.
ACOUSTIC MODEL TRAINING PROCESS

Acoustic feature vectors are extracted from the 11kHz sampled data every 10 ms, which includes instantaneous energy, 12 mel FFT cepstra, instantaneous pitch, as well as their first-order and second-order derivatives. The training started from an English prototype to build a Mandarin system without tone, and then a Mandarin system with tone. Also, we started with a speaker-dependent system, to a speaker-independent system using speech data of more than 30,000 sentences by 54 speakers from Beijing area. In one of the systems we built, there are 160 phonemes. Over 3000 context-dependent models <1,2> are constructed from an underlying mixture of some 30000 prototypes.

VOCABULARY AND LANGUAGE MODEL

In English and other European languages, words are well-defined. The white space defines word boundaries. To obtain a vocabulary from a text corpus, one can simply collect all strings bounded by white spaces (or hyphens and punctuations). An n-gram language model can be built upon a large text corpus easily. This is not true for Chinese.

Written Chinese has no word boundaries. Thus, the simplest idea is to take a characters as a word, and to construct an n-gram language model. The standard coding for Chinese characters in China (GB-2312-80), defines some 6700 characters. Statistics made from 4-years of People's Daily text corpus shows that the number of characters needed to make a decent coverage is much less than 6700:

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Number of characters in vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>6200</td>
</tr>
<tr>
<td>99.99%</td>
<td>5000</td>
</tr>
<tr>
<td>99.90%</td>
<td>3780</td>
</tr>
<tr>
<td>99.00%</td>
<td>2250</td>
</tr>
</tbody>
</table>

A trigram language model is built on a text corpus of 300 million characters (see below). The perplexity of a typical newspaper article, President Jiang Zemin's 1994 New Year Speech, is 32.9.

Next, we tried word-based language models. The definition of "word" in Chinese has been a difficult problem. Even the recently published government standard, GB-13715, "Word Segmentation Standard for Information Processing", contains a lot of ambiguities. However, since the word boundary does not appear in the final text from a recognition process, the details of definition of word boundaries could be solely based on the criterion of improving decoding accuracy and speed. By going through various speech recognition experiments, we found the following segmentation rules produce good results:

1. Treat all suffixes as individual words: such as De, Le, Guo, Zhao, Men, etc.
2. If a long compound word can be separated into two components, and the probability of occurrence of both components are high, then it is treated as two individual words.
3. Chinese numbers 1 through 99, 100 through 900, and 1000 through 9000 are considered as words: whereas the units "Wan4" (10000) and "Yi4" (1000000) are separated from the preceding numbers (for example, Di4 Shi2 Wu3 Wan4 San1 Qian1 Si4 Bai3 Er4 Shi2 Li34 Ge).
4. Arabic digits are considered as individual words.

The vocabulary is extracted iteratively from publicly available text corpora. These include People's Daily, from 1991 to 1994; Market Daily, 1994; and the Collection of Selected Articles from One Hundred Chinese Newspapers and Journals, 1994: altogether 300 million characters. We started with a publicly available electronically readable vocabulary with some 50000 words. use it to segment the text, find the counts, improve the vocabulary manually according to the rules, resegment the text using the improved vocabulary, find the counts again, and edit the vocabulary manually again. After three iterations, we obtained a vocabulary of 29,000 words. The coverage of our vocabulary, 99.95%, is very high comparing with European languages. This is because almost all new words in Chinese are combinations of known words, and the word boundaries do not exist in written text.

Using the manually edited vocabulary, we segmented that 300 million character text corpus again and constructed a trigram language model using the same algorithm for English developed at IBM Research <3>. The perplexity of the 1994 New Year Speech of President Jinag is 104.3. This is substantially smaller than the perplexity of the character-based language model. Because on average, the length of a word in Chinese is 2, the perplexity of the character-based language model is 32.9 X 32.9, which is 1082, an order of magnitude higher than that of the word-based language model.
TYPICAL DECODING RESULTS

As in the case of IBM's speech decoder for English, the decoding process includes a rank-based labeling system and a stack decoder with an envelop search algorithm <3>.

Two sets of test scripts are used. The first one is the 1994 New Year Speech of President Jiang Zemin. This article represents the typical style of People's Daily text. The second one is 100 sentences randomly selected from 1994 Science and Technology Daily. It has a lot of technical terms and difficult sentences (even to read), and the perplexity is 154.6.

Typical results are listed in the following tables. The first one is for a male speaker from Beijing. The error rate is based on characters.

<table>
<thead>
<tr>
<th>Test script</th>
<th>Ins</th>
<th>Del</th>
<th>Sub</th>
<th>Tot</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pres. Jiang's speech</td>
<td>33</td>
<td>3</td>
<td>102</td>
<td>136</td>
<td>9.5%</td>
</tr>
<tr>
<td>Science and Tech. Daily</td>
<td>71</td>
<td>7</td>
<td>245</td>
<td>323</td>
<td>10.7%</td>
</tr>
</tbody>
</table>

Then, we tested the difference between the word-based language model and the character-based language model. For the character-based language model, using the 1994 New Year speech of President Jiang Zemin, we found an decoding error rate of 13.8%, poor in comparison with the word-based trigram language model, 9.5%. This is due to the higher perplexity of the character-based language model.

The error rate deteriorates substantially for speakers with accents. Using the 100 sentences from the Science and Technology Daily, the results for 6 speakers are:

<table>
<thead>
<tr>
<th>Name</th>
<th>Origin</th>
<th>Gender</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>yxw</td>
<td>Beijing</td>
<td>m</td>
<td>12.0%</td>
</tr>
<tr>
<td>chw</td>
<td></td>
<td>m</td>
<td>15.8%</td>
</tr>
<tr>
<td>lqs</td>
<td>Tangshan</td>
<td>f</td>
<td>16.5%</td>
</tr>
<tr>
<td>cxr</td>
<td>Jinhua</td>
<td>f</td>
<td>19.1%</td>
</tr>
<tr>
<td>cjf</td>
<td>Shanghai</td>
<td>m</td>
<td>28.3%</td>
</tr>
<tr>
<td>wyy</td>
<td>Shanghai</td>
<td>f</td>
<td>30.4%</td>
</tr>
</tbody>
</table>

For speakers with accents, it is clear that either speaker dependent adaptation or more sophisticated acoustic modelling will be necessary.

SUMMARY

In summary, we demonstrated that by using the context-dependent vector quantization method together with a phoneme system with tones, acoustic vectors with instantaneous pitch as a basic component, and a trigram language model based on segmented Chinese text, using the well-established IBM HMM-based speech recognition technology, a continuous speaker-independent putonghua speech recognition system can be established, with error rate similar to the results of Wall Street Journal task in English <3>.

REFERENCES