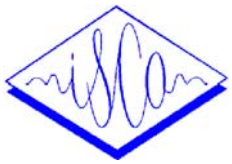

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

NEW METHODS IN CONTINUOUS MANDARIN SPEECH RECOGNITION

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

IBM Thomas J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA
*IBM China Research Laboratory, 26 6th Street, Shangdi, Beijing 100085, China

ABSTRACT

We describe new methods for speaker-independent, continuous mandarin speech recognition based on the IBM HMM-based continuous speech recognition system [1-3]: First, we treat tones in mandarin as attributes of certain phonemes, instead of syllables. 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, an effective trigram language model is trained[4]. By applying those new methods, a speaker-independent, very-large-vocabulary continuous mandarin dictation system is demonstrated. Decoding results showed that its performance is similar to the best results for US English.

1. INTRODUCTION

Mandarin Chinese has the largest number of speakers among all languages. Research and development of mandarin speech recognition has been conducted for decades [5,6]. It is characterized by the following general features:

1. Tone is a phonemic element of a syllable. In words, syllables with the same consonant and vowel but different pitch contours represent different morphemes.

2. The written text of Chinese does not have word boundaries.

Therefore, to recognize continuous Chinese efficiently, we must resolve both problems of tone recognition and word segmentation. In this paper, we will describe new methods of recognizing mandarin. Based on those new methods, we have demonstrated a continuous, speaker-independent, very-large-vocabulary mandarin dictation system. The measured error rate is comparable to the best system for US English.

2. 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 [5,6]. 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 new approach treats tone at the phonemic level[4]. Pitch is treated as an acoustic parameter, the same way as energy or a cepstrum. In the following, we describe the new methods of acoustic processing:

2.1. Pitch as a Continuous Acoustic Variable

Typically, the acoustic parameters extracted from the speech signal consists of a number of cepstra, with an optional instenteneous energy. In addition to their instenteneous values, the first and second derivatives of those parameters are also taken as components of the acoustic feature vector[1-3]. To make pitch as one of the acoustic parameters, special treatment must be made. According to the general understanding, pitch can only be defined for voiced frames of speech[7,8]. For silence and unvoiced sections, pitch does not exist. Figure 1A shows the measured pitch contour of a four-syllable phrase, using a conventional autocorrelation approach, after using an interpolation algorithm to remove certain jumps. As seen from Fig. 1A, during

silence frames and the frames of unvoiced consonants, the pitch is undetermined. At those frames, the derivatives of pitch would become zero. At the boundaries of a voiced section and an unvoiced section, the derivatives will become infinity. In both cases, serious problems will occur during training and decoding. We use a continuation algorithm to solve this problem[4]:

1. A running average is calculated based on all valid points.

2. At the beginning of an utterance, the pitch value is defined as the running average plus a random noise.

3. When the speech proceeds from a voiced section to an unvoiced section, the pitch is defined as an exponential decay function towards the running average, plus a random noise.

4. The entire signal is passing through a low-pass filter to remove spikes.

The addition of random noise to the unvoiced section is found to be necessary for avoiding zero variance in frames where pitch is not a significant variable.

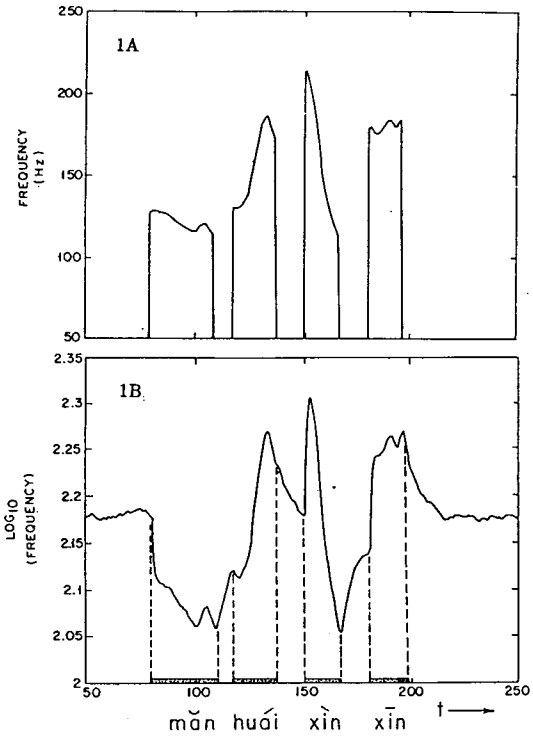

Figure 1A. A pitch contour obtained from a conventional autocorrelation function approach, after removing singular points. 1B, after applying a continuation algorithm. See text.

Figure 1B shows the pitch after continuation for the same utterance as in Fig. 1A. As shown, it is a well-behaved parameter which can be processed the same way as a cepstrum. Also, the logarithm of pitch is taken. The advantage is, by using the cepstra mean subtraction algorithm, the pitch contours become gender-independent.

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 mandarin, 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.

2.2. Concept of Tonemes

Based on the above observations, we found that the context-dependent quantization algorithm [1-3] used in recognizing phonemes (in languages without tones) provides an efficient way of recognizing tones in Chinese languages using the following recipe[4]:

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" [9], a toneme is "a phoneme consisting of a specific tone in a tone language". For example, ā, á, ǎ, à, 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 2D1 represents a typical result.

As shown in Fig. 2, 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 51 [10]. 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 mandarin 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.

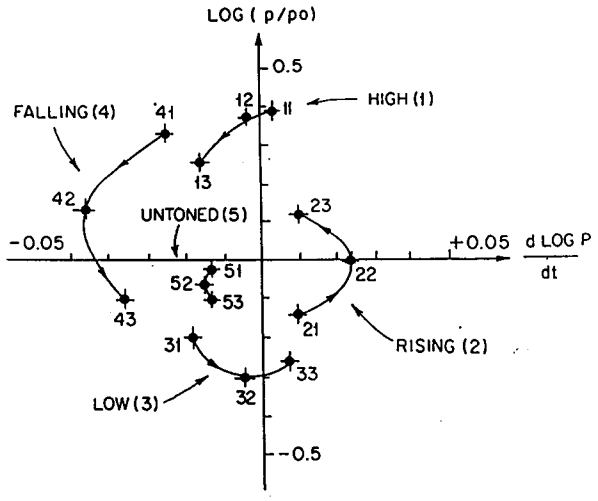

Figure 2. Typical means of instantaneous pitch and pitch derivative. The pitch and pitch derivative values of five tones (including the neutral), averaged over all tonemes (vowels or combinations of vowel and nasal ending) in mandarin (standard spoken Chinese). Each vowel, as a phoneme, is represented by three HMM states. The numbers 1, 2, and 3 indicate the sequence of HMM states in a single phoneme.

3. PHONEME SYSTEM FOR MANDARIN

Based on the new concept of tone recognition, we designed a phoneme set accordingly. This system differs from the conventional phoneme system for Mandarin [5,6,10] (21 initials and 39 finals) in two ways. First, each premes is a combination of the initial consonant with the glide if exists. Second, a basic toneme is defined as the latter part of the final, starting with the main vowel. By doing so, the number of tonemes are reduced from 195 (the complete set of toned finals) to 105 (21 basic tonemes). The cost of computation for tonal phonemes is thus substantially reduced. The number of premes is increased from 21 (initials) to 53. For example, the initial 'L' corresponds to four premes:

L	LI	LU	LYU

The mouth shapes of the four premes , even from the beginning, are very different, thus better to treat them as different phonemes. The four finals 'AN', 'IAN', 'UAN', and 'YUAN' share one basic toneme. With different tones, it splits into five tonemes:

AN1	AN2	AN3	AN4	AN5

Each syllable (with tone) is decomposed to a preme and a toneme, for example:

Syllable	Preme	Toneme
Lan3	L	AN3
Lian1	LI	AN1
Luan4	LU	AN4
Lyuan2	LYU	AN2

4. ACOUSTIC TRAINING PROCESS

Acoustic feature vectors are extracted from the 11kHz sampled data every 10 ms, which includes 12 mel FFT cepstra and instantaneous pitch, as well as their first-order and second-order derivatives. 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 [4]. Later, we collected speech from 300 speakers in China by IBM China Research Laboratory, and trained a set of acoustic model with over 3000 context-dependent leaves from an underlying mixture of some 30000 prototypes. Test results from the later acoustic model are reported here.

5. VOCABULARY, LANGUAGE MODEL

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, Zho, 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" (10000000) are separated from the preceding numbers (for example, Di4 Shi2Wu3 Wan4 San1Qian1 Si4Bai3 Er4Shi2Liu4 Get).

4. Arabic digits are considered as individual words.

The vocabulary is extracted iteratively from publicly available text corpa. 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 percent, 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 the text corpus again and constructed a trigram language model using the same algorithm for English developed at IBM Research [1-3].

6. 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.

The test scripts are 108 sentences randomly chosen from 1996 Peoples Daily, which is not part of the text corpus for training the language model and acoustic model. The perplexity against the langauge model is 130. The test speech is spoken by 16 speakers from different regions in China: B indicates Beijing area; N indicates North China; E indicates East China (Shanghai, Jiangsu, and Zhejiang); and S indicates South China (Guangdong, Guangxi, and Fujian). In the first column, 'm' indicates a male speaker, and 'f' indicates a female speaker. Character error rate in percentage is listed.

Speaker	Origin	Error rate
m419	B	5.42
f418	B	6.06
f440	B	7.40
m420	B	8.70
f417	E	8.77
f437	S	8.92
f413	N	9.24
f422	E	9.64
f412	E	9.64
m411	N	10.14
m427	S	11.95
m428	E	12.56
m423	S	13.75
f439	N	15.34
m433	E	15.85
m432	E	19.53

As expected, the speakers from Beijing area have the lowest error rates. The error rates are higher for speakers from other areas.

7. SUMMARY

We described new methods for mandarin speech recognition based on the IBM HMM-based continuous speech recognition system: Treating tones as attributes of phonemes, expanding instantaneous pitch into a continuous variable in the acoustic feature vector, and using a special set of word-segmentation rules for constructing vocabulary and langauge models. Those new methods were tested in a speaker-independent, very-large-vocabulary continuous mandarin dictation system. The results showed an performance similar to the best in US English.

8. REFERENCES

1. L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, M. A. Picheny: "Decision trees for phonological rules in continuous speech", Proc. IEEE ICASSP-91, Vol. 1, pp. 177-180, 1991.

2. L. R. Bahl, P. de Souza, P. Gopalakrisnan, M. Picheny, "Context-dependent vector quantization for continuous speech recognition", Proc. IEEE ICASSP-93, Vol. 2, pp. 632-635, 1993.

3. L. R. Bahl, S. Balakrishnan-Aiyer, J. R. Bellgarda, M. Franz, P. S. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. A. Picheny, S. Roukos, "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task". Proc. IEEE ICASSP-95, Vol. I, pp. 41-44, 1995.

4. C. J. Chen, R. A. Gopinath, M. D. Monkowski, and M. A. Picheny, "A Continuous Speaker-Independent Putonghua Dictation System", Proc. Int. Conf. Signal Processing, Oct. 1996, Beijing, pp. 821-824.

5. Y. Gao, B. Yuan, T. Tng, G. Loo, G. Loudon, and S. Yogan "Mandarin Chinese dictation system research and development", Proc. ICCC-94, pp. 59-101.

6. H. M. Wang, J. L. Shen, Y. J. Yang, C. Y. Tseng, and L. S. Lee, "Complete recognition of continuous Mandarin speech ...", Proc. ICASSP-95, Vol.1, pp. 61-64.

7. W. Hess, "Pitch Determination of Speech Signals: Algorithms and Devices", Springer-Verlag, Berlin, 1983.

8. W, Hess, "Pitch and Voicing Determination", in "Advances in Speech Signal Processing", edited by S. Furui and M. Sondhi, Marcel Dekker, New York, 1992, pp. 3-48.

9. M. Pei, "Glossary of Linguistic Terminology", Columbia University Press, New York 1966.

10. B. Yin and M. Felley, "Chinese Romanization: Pronunciation and Orthography", Sinolingua, Beijing, 1990, pp. 9-10.