COMPLETE RECOGNITION OF CONTINUOUS MANDARIN SPEECH FOR CHINESE LANGUAGE WITH VERY LARGE VOCABULARY BUT LIMITED TRAINING DATA

Hsin-min Wang[1], Jia-lin Shen[1], Yen-ju Yang[2], Chiu-yu Tseng[3] and Lin-shan Lee[1,2,4]

[1]Dept. of Electrical Engineering, National Taiwan University,

[2]Dept. of Computer Science & Information Engineering, National Taiwan University,

[3]Institute of History & Philology, Academic Sinica,

[4]Information Science, Academic Sinica,

Taipei, Taiwan, Republic of China

ABSTRACT

This paper presents the first known results for complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but very limited training data. Although some isolated-syllable-based or isolated-word-based large-vocabulary Mandarin speech recognition systems have been successfully developed, a continuous-speech-based system of this kind has never been reported before. For successful development of this system, several important techniques have been presented in this paper, including acoustic modeling of a set of sub-syllabic models for base syllable recognition and another set of context-dependent models for tone recognition, a multiple candidate searching technique based on a concatenated syllable matching algorithm to synchronize base syllable and tone recognition, and a word-class-based Chinese language model for linguistic decoding. The best recognition accuracy achieved is 88.69% for finally decoded Chinese characters, while 88.69%, 91.57%, and 81.37% for base syllables, tones, and tonal syllables respectively.

1 INTRODUCTION

This paper presents the first known results for complete recognition of continuous Mandarin speech for Chinese language (i.e., recognition of the syllables, tones, Chinese characters, words, and sentences) with very large vocabulary but using only very limited training data. Input of Chinese characters into computers is still a difficult problem today because Chinese language is not alphabetic and many existing keyboard input systems are simply inconvenient. Voice input with large vocabulary is therefore highly desired, because in general the input materials or texts into computers are assumed to have very large vocabulary. Although some isolated-syllable-based [1, 2] or isolated-word-based [3, 4] large-vocabulary Mandarin speech recognition systems have been successfully developed, a continuous-speech-based system of this kind has never been reported before but highly desired. Only with continuous speech can the desired speed, convenience, and naturalness of man-machine communications be possibly achieved, though at the price of much more difficulties caused by many problems such as the very complicated context dependent co-articulation effects. This is why in the initial stage here in this paper only speaker dependent mode is considered. Also, because it is not feasible to request a new user to produce too much training speech before being able to use the system, relatively limited training data become a very natural constraint.

In Mandarin Chinese, there exist at least more than 80,000 commonly used words, each composed of from one to several characters, and more than 10,000 commonly used characters, all produced as mono-syllables. However, the total number of phonologically allowed different syllables is only 1345. This is why accurate recognition of all the 1345 Mandarin syllables is believed to be the first key problem in Mandarin speech recognition with very large vocabulary. Also, Mandarin Chinese is a tonal language, in which each syllable is assigned a tone and there are a total of 4 lexical tones plus 1 neutral tone. It has been found that the vocal tract parameters for Mandarin speech are only slightly influenced by the tones, and the tones can be separately recognized primarily using the pitch contour information [5]. When the differences among the syllables caused by tones are disregarded, the total of 1345 different tonal syllables is reduced to only 416 base syllables (i.e., the syllable structures independent of the tones). It is therefore helpful to recognize the tones and base syllables separately. Based on the above considerations, the block diagram of the present speech recognition system presented in this paper is shown in Figure 1. In the acoustic processor, the tones and base syllables are separately recognized using two different sets of models [5, 6] and a multiple candidate searching technique based on a concatenated syllable matching algorithm is used to achieve better tonal syllable recognition by integrating the two separate base syllable and tone recognizers into a synchronous and cross-referenced process. While in the linguistic decoder, a lexicon is used to construct the word lattice and a Chinese language model used to find the final output characters, words, and sentences [1, 2], because every tonal syllable in general is shared by many (e.g. 10,000/1345 in average) possible homonym characters. These are the basic features of the present research which are primarily due to the special characteristics of the Chinese language.

2 PRELIMINARIES

2.1 Speech Database and Initial Processing

The speech database used here is produced by 4 speakers, two male and two female. The results reported here are averages of the four. Each speaker uttered 3 sets of the 1345 isolated Mandarin syllables and 1 set of

0-7803-2431-5/95 $4.00

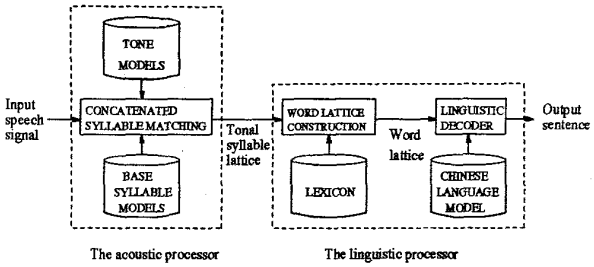

Figure 1: The block diagram of the continuous Mandarin speech recognition system.

phonetically balanced continuous Mandarin sentences covering all the 416 base syllables and different tones (including 200 sentences or 1514 syllables) for training, while another set of short paragraphs taken from different articles (including 142 sentences or 1451 syllables totally) for testing.

All the recorded materials are obtained in an office-like laboratory environment, and digitized with a sampling frequency of 16 KHz. For base syllable recognition, 14 cepstral and 14 delta-cepstral coefficients, delta-energy, and delta-delta-energy are used as feature parameters to form a feature vector with dimension 30, while for tone recognition the pitch period and the energy together with their first and second order delta-coefficients are used instead to form a feature vector with dimension 6.

2.2 CHMM Modeling

The hidden Markov models used in this paper are left-to-right continuous HMM's (CHMM's) with two transitions only. The partitioned Gaussian mixtures (PGM) [6] each with a diagonal covariance matrix are used as the observation density.

The training process here for the base syllable and tone models includes two stages [6]. In the first stage, the CHMM's are trained by the segmental K-means algorithm using the isolated training data. In the second stage, the CHMM's obtained from the first stage are used as the initial models, and further reestimated by the continuous training data using a modified segmental K-means training procedure, in which the CHMM's obtained after each iteration are linearly interpolated with the initial CHMM's obtained from the first stage.

The recognition algorithm used here is a N-best beam search algorithm. In the forward procedure, the beam search with dynamically chosen beam width to reserve at most 10 grammar nodes at each frame is applied, and the accumulated scores along with the arriving frames for all reserved grammar nodes are stored. At the end of the utterance, the backtracking procedure recursively searches through the saved lists of the grammar nodes to obtain N complete model string hypotheses.

3 RECOGNITION OF THE BASE SYLLABLES AND TONES

To find appropriate acoustic units for modeling is the first key problem. In this section, special efforts were made in selecting the most appropriate acoustic units for base syllable and tone recognition, considering the condition of very limited training data.

3.1 Acoustic Units for Base Syllable Recognition

Although the base syllable is a very natural recognition unit for Mandarin Chinese due to the monosyllabic structure of Chinese language, it suffers from inefficient utilization of the training data in the training phase and high computation requirement in the recognition phase. A set of appropriate sub-syllabic acoustic units for base syllable recognition is therefore highly desired, and the INITIAL/FINAL's are apparently good choices considering the basic structure of Mandarin syllables. Conventionally, each Mandarin base syllable can be decomposed into an INITIAL/FINAL format very similar to the consonant/vowel relations in other languages. Here INITIAL is the initial consonant part of a syllable, and FINAL is the vowel (or diphthong) part but including optional medial or nasal ending, and there exists a total of 22 INITIAL's and 41 FINAL's in Mandarin. An initial observation on continuous Mandarin speech is that the co-articulation effects within a syllable are much more significant than those across syllables, and within a syllable the acoustic characteristics of the INITIAL are certainly highly dependent on the FINAL, but those of the FINAL are much less dependent on the INITIAL. With these observations, the approach here is to assume both the co-articulation effects across syllables and the dependence of the FINAL on the preceding INITIAL within a syllable to be negligible. Furthermore, it can be further assumed that the dependence of an INITIAL on the following FINAL is primarily determined by the beginning phone of the FINAL only. For example, the INITIAL /s/ in syllables /sai/, /sau/, /san/, etc. are assumed the same, but different from the INITIAL /s/ in syllables /su/, /suo/, etc. In this way, the 22 INITIAL's can be expanded to 113 context-dependent(CD) INITIAL's. FINAL's, on the other hand, are just taken as context-independent units for our very limited training data condition. In this way, the co-articulation effects are much easier to model under our condition of very limited training data, and preliminary experimental results have shown these are reasonable assumptions [6], although it is always better to consider all possible co-articulation effects if enough training data are available.

Some experimental results for continuous speech base syllable recognition are listed in Table 1(a), where the recognition accuracy for base syllables is shown. Different choices of models including the 416 base syllable (BS) models, 22 context-independent (CI) INITIAL and 41 context-independent (CI) FINAL (22 CI-I/41 CI-F) models, and 113 context-dependent (CD) INITIAL and 41 context-independent (CI) FINAL (113 CD-I/41 CI-F) models, are all tested in the experiments. Each INITIAL model is represented by a CHMM with 3 states, each FINAL model with 4 states, and each syllable model with 7 states. It's obvious from Table 1(a) that the 113 CD-I/41 CI-F models have very good potential for base syllable recognition for continuous Mandarin speech when the training data are limited.

3.2 Acoustic Units for Tone Recognition

It has been well known that the tone behavior is very complicated in continuous Mandarin speech, although there are only 5 (4+1) different tones in Mandarin. It is

BS	22 CI-I/41 CI-F	113 CD-I/41 CI-F
69.84%	82.71%	88.25%

(a)

5 CI-T	23 CD-T	175 CD-T
86.92%	89.80%	88.25%

(b)

Table 1: The results for recognition of (a) base syllables and (b) tones.

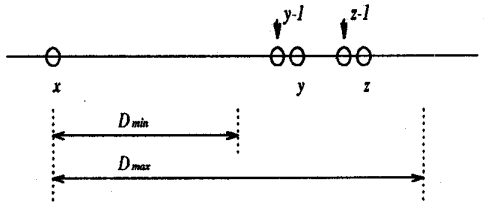

Figure 2: A section of an example utterance. x, y, and z are possible beginning points, while $y-1$ and $z-1$ are ending points corresponding to x.

therefore very important to select minimum number of appropriate context-dependent models such that they can adequately describe the very complicated tone behavior, but at the same time can be very well trained by the limited training data. If all possible tone concatenation conditions need different context-dependent models, a total of 175 models will be needed, i.e., 5^3 (in the middle of a sentence) + 5^2 (at the end of a sentence) + 4×5 (at the beginning of a sentence, because the neutral tone never appears at the beginning of a sentence) + 5 (isolated models). However, practically this number can be significantly reduced if special characteristics of the tone behavior can be carefully considered. For example, it was observed empirically that both Tones 1 and 2 end high at a similar level, therefore the influence of Tones 1 and 2 on the following tones is very similar and this can be used in merging context-dependent models, and so on. When all such phenomena are considered, many models can be merged and the total number of models used in the experiments here were reduced to 23 [5].

Some initial experiments on continuous speech tone recognition were performed. In these experiments, the 5 context-independent (CI) tone (5 CI-T) models and the 175 and 23 context-dependent (CD) tone (175 CD-T and 23 CD-T) models derived above were all tested, and each tone model is represented by a CHMM with 7 states. The results are listed in Table 1(b), where the recognition accuracy for tones is shown. It can be found that the 23 CD-T models have very good potential for tone recognition for continuous Mandarin speech when the training data are limited.

4 RECOGNITION OF THE TONAL SYLLABLES

When tones and base syllables are separately recognized as described above, a major problem is that the segmentation of the unknown utterances into syllables may become quite different in the two recognizers, and the problem gets even much worse when any insertion/deletion occurs in one of the recognizers, in which the resulting tone and base syllable sequences will be out of synchronization and such errors can propagate very long. A concatenated syllable matching (CSM) algorithm [7] summarized below is therefore developed for synchronizing base syllable and tone recognition. For a given test utterance, all possible syllable beginning frames can be first obtained by picking up all the dips in the energy contour, such as x, y, z in Figure 2. The possible ending frames such as $y-1$ and $z-1$ corresponding to each beginning frame such as x can then be found using estimated minimum and maximum duration of a syllable D_{min} and D_{max} as in Figure 2. With beginning and ending points of a syllable estimated as above, the accumulated score at an ending point such as $y-1$ in Figure 2 is then determined by the dynamic programming approach:

$$\begin{aligned} T[y-1] &= T[x-1] \\ &+ \max_{1 \le i \le 1345} [S_{bs_i}(x, y-1) + S_{t_i}(x, y-1)] \end{aligned} \quad (1)$$

where $T[u]$ is the accumulated score at a point u, $S_{bs_i}(u,v)$ and $S_{t_i}(u,v)$ are the scores for the corresponding base syllable and tone for a tonal syllable i (i.e., one out of the 1345, including a base syllable and a tone) matched with the utterance section (u,v). For any utterance section, the base syllable recognition was performed using the N-best beam search algorithm mentioned above because the sub-syllabic units were adopted, while for tone recognition the conventional Viterbi search algorithm was used. At the end of the utterance, the most probable tonal syllable sequence can then be easily obtained by backtracking the entire utterance. In this way, the recognition of base syllables and tones can be performed syllable by syllable in complete synchronization, with advantages that not only extra information such as energy contour dips have been applied, but the information in base syllable and tone recognition are now properly integrated in equation (1).

Previous experiences [1, 2] indicate that multiple candidates of tonal syllables are definitely needed for the linguistic processor to find out the most probable Chinese character or word sequences simply because the top 1 candidate is not necessarily correct. Here, a very efficient technique is developed to generate multiple candidates of the tonal syllables to construct a tonal syllable lattice for the linguistic processor to work on. In the syllable matching process mentioned above, for any possible utterance section, not only the most likely tonal syllable is reserved but all the N-best tonal syllable candidates are stored for backtracking. Therefore, as long as the most likely syllable sequence is obtained, the top N tonal syllable candidates for each utterance section will be obtained simultaneously. The obvious advantage of such a multiple candidate searching technique is that it's suitable for real-time implementation because no further computation is needed. In fact, as will be shown below, it also provides higher accuracy, i.e., a tonal syllable lattice including higher percentage of correct tonal syllables is produced.

The experimental results are listed in Table 2. The best recognition accuracy achieved here for tonal sylla-

tonal syllable	base syllable	tone
81.37%	88.69%	91.57%

Table 2: The results for recognition of tonal syllables, based on Concatenated Syllable Matching.

bles is 81.37%, while those for base syllables and tones are 88.69% and 91.57% respectively. Note that for both base syllables and tones, the recognition accuracies are higher than those obtained previously in Table 1, because here the information in base syllable and tone recognition are integrated. On the other hand, the top N recognition accuracies are listed in the first row of Table 3, in which the top N accuracy means the percentage that the correct tonal syllable is within the N tonal syllable candidates selected. These results show that such a multiple candidate searching technique is very efficient in finding N most likely candidates for the tonal syllables. Note that though the top 1 recognition accuracy achieved here is not very high (81.37%), the top 5 recognition accuracy of 97.12% is in fact reasonable as long as a good linguistic processor is applied.

5 LINGUISTIC PROCESSING

With the top N tonal syllables recognized and a tonal syllable lattice obtained as described above, this lattice is first transformed into a word lattice via a lexical access process in the linguistic processor as in Figure 1. Because every tonal syllable is in general shared by many homonym characters, and it is possible for a character to form either a mono-character word or poly-character words after combining with adjacent characters as mentioned previously, such a word lattice can be very large and complicated, especially when top N tonal syllables are included and N is large. With such complicated word lattices, the linguistic decoder certainly requires a very powerful Chinese language model to find out the most probable output characters, words, and sentences. Because the words are the basic building blocks of a sentences, a preliminary word-class-based language model [2] is believed to be an efficient approach for the language model at least in the initial stage here. The training corpus used in this research includes texts from newspapers, articles from magazines, and parts from various novels, with a total of 5,303,554 characters (or 3,500,067 words). Some of the tone sandhi phenomena in continuous Mandarin speech have also been included here. For example, a simple tone sandhi rule is that when a syllable of Tone 3 is followed by another of Tone 3, the previous one will be pronounced as Tone 2. Therefore, when the obtained tonal syllable lattice includes any tonal syllable pair with a Tone 2 followed by a Tone 3, if the base syllable of the previous syllable of Tone 2 combined with Tone 3 is also one of the 1345 phonologically allowed tonal syllables but not within the N-best list selected, the tonal syllable with Tone 3 is then automatically added to the list.

The final results for character accuracy with different numbers of syllable candidates included in the tonal syllable lattice are listed in the second row of Table 3. It can be found that the best recognition accuracy for characters achieved here is 88.69% when 5 syllable candidates (or 97.12% of correct syllables) are included in the tonal syllable lattice. Although 98.67% of correct syllables can be included when 10 candidates are used, it's obvious that higher degree of ambiguity will reduce the recognition accuracy in that case.

number of candidates included in the tonal syllable lattice	2	3	4	5	10
% of correct tonal syllables included	90.47	94.90	96.45	97.12	98.67
character accuracy	83.15	86.25	88.03	88.69	88.03

Table 3: The final recognition accuracies for Chinese characters.

6 CONCLUSIONS

In this paper, a continuous Mandarin speech recognition system with very large vocabulary but trained by very limited training data is presented. This system integrates several important techniques including acoustic modeling of a set of sub-syllabic models for base syllable recognition and another set of context-dependent tone models for tone recognition, a concatenated syllable matching algorithm with an efficient multiple candidate searching technique to integrate the base syllable and tone recognizers, and a specially designed word-class-based Chinese language model for linguistic decoding based on the special property of Chinese words. The integration of these techniques gives the first known results for complete recognition of continuous Mandarin speech with very large vocabulary but very limited training data. The best recognition accuracy achieved here is 88.69% for finally decoded Chinese characters, while 88.69%, 91.57%, and 81.37% for base syllables, tones, and tonal syllables respectively.

References

[1] Lin-shan Lee, et al, "Golden Mandarin(I)-a real-time Mandarin speech dictation machine for Chinese language with very large vocabulary", *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 2, pp. 158-179, April 1993.

[2] Lin-shan Lee, et al, "Golden Mandarin(II)-an improved single-chip real-time Mandarin speech dictation machine for Chinese language with very large vocabulary", *ICASSP*, Minneapolis, Minnesota, USA, 1993, vol. 2, pp. 503-506.

[3] Jung-Kuei Chen, F. K. Soong, and Lin-Shan Lee, "Large vocabulary word recognition based on tree-trellis search", *ICASSP*, Adelaide, South Australia, 1994, vol. 2, pp.137-140.

[4] Hsiao-Wuen Hon, Kai-Fu Lee, et al, "Towards large vocabulary Mandarin Chinese speech recognition", *ICASSP*, Adelaide, South Australia, 1994, vol. 1, pp. 545-548.

[5] Hsin-min Wang and Lin-shan Lee, "Tone recognition for continuous Mandarin speech with limited training data using selected context-dependent hidden Markov models", to appear in *Journal of The Chinese Institute of Engineers.*

[6] Hsin-min Wang, Renyuen Lyu, Jia-lin Shen, and Lin-shan Lee, "An initial study on large-vocabulary continuous Mandarin speech recognition with limited training data based on sub-syllabic models", in Proc. *Int. Computer Symposium*, Hsin-chu, R.O.C., 1994, pp. 1140-1145.

[7] Jia-lin Shen, Hsin-min Wang, Bo-ren Bai, and Lin-shan Lee, "An initial study on a segmental probability model approach to large-vocabulary continuous Mandarin speech recognition", *ICASSP*, Adelaide, South Australia, 1994, vol. 2, pp. 133-136.