A Large-Vocabulary Bilingual Speech Recognition System for Chinese and Japanese Language

Jyh-Shing Shyuu and Jhing-Fa Wang
Department of Computer Science and Information Engineering
National Cheng Kung University
1, Ta-Hsueh Road, Tainan, Taiwan, R.O.C.
wangjf@server2.iie.ncku.edu.tw

ABSTRACT

Bilingual or Multilingual speech recognition gradually becomes an attractive research topic because bilingual writings appear almost everywhere in present day. In this paper, we propose a continuous word-based speech recognition system to dictate the Mandarin and Japanese speech simultaneously. We find that there are about 62 basic phoneme like units(PLUs) among the mixed Mandarin and Japanese syllables. The 62 HMMs are used to decode the input speech into word hypotheses based on a fast tree-beam searching algorithm. In the language model, the bigram model and trigram model are used to select the most likely word from the word candidates. We also have a bilingual dictionary to deal with the cross language information. Our proposed system architecture can not only dictate Mandarin and Japanese speech simultaneously but also provide a possible solution to recognize any other bilingual speech.

I. INTRODUCTION

Bilingual or Multilingual speech recognition gradually becomes an attractive research topic because bilingual writings appear almost everywhere in present day. In this paper, a continuous word-based speech recognition system is proposed for dictating the Mandarin and Japanese speech simultaneously. In the acoustic processing model, we segment Mandarin syllables into phoneme like units (PLUs) instead of using traditional 21 consonants and 38 vowels in recognition. We find that there are totally 62 basic PLUs among the mixed Mandarin and Japanese syllables[1,2]. By using the

PLUs, it can not only increase the recognition speed but also increase the size of training patterns for each recognition unit. Because Chinese is a tonal language, tone recognition is also included in our recognition system by a proposed concatenated-tone recognition model. In the language processing model, two sets of bigram model and trigram model were used to dictate Japanese and Chinese language respectively. For the cross language information, if the cross language information is not available, a bilingual dictionary is used to translate the Japanese to Chinese or Chinese to Japanese. Hence, the bigram and trigram information can be obtained from the Chinese or Japanese language model respectively. There is also another alternative solution to deal with bilingual trigram and bigram information. We can assume that the bigram and trigram information are equal for transitions from Chinese to Japanese or from Japanese to Chinese. The assumption is based on the observation that most of the transitions are caused by the noun. The assumption can be eliminated or replaced when the bilingual information can be estimated from a bilingual text corpus. In the acoustic training phase, 400 Japanese balanced-sentences and 320 Chinese balanced-sentences were collected for training the 62 acoustic HMMs. The balanced-sentences are automatically generated by a previously proposed balanced-corpus generating algorithm[3]. In the recognition phase, a Viterbi tree-beam search algorithm is adopted to obtain the top-N most likely word candidates in frame synchronously[5]. By combining the acoustic scores with the language scores from the language decoder the most likely sentence can be obtained. The system block diagram is shown in Fig. 1.

In the following Sections, Section II describes the overall system block diagram of our proposed system. Section III gives a detailed description of the acoustic processing model, and Section IV presents the language processing model. Section V shows some experimental results. Finally, conclusion remark is given in Section VI.

II . SYSTEM ARCHITECTURE

The System block diagram is shown in Fig 1. The system block diagram can be divided into two parts, acoustic processing model and language processing model respectively. In the acoustic processing model, a fast frame-synchronous tree-beam search algorithm is adopted to recognize the input speech into word candidates. The word candidates are pushed into a FIFO word-candidate pool to form a word lattice for the language decoder. The language decoder use the bigram and trigram models to decode the word lattice into sentence. However, there is only one most likely word output from the word candidate pool.

III. ACOUSTIC MODEL

In the acoustic processing model, 62 PLUs are first extracted from Mandarin and Japanese Syllables. We use these 62 PLUs as a basis to generate two sets of balanced corpora based on an automatic generation algorithm[3]. Hence, we can use the balanced corpora to collect speech data for training the acoustic models. The speech features we used to train the HMMs are 14-order MFCC, 14-order delta-MFCC, frame energy and delta-energy. The MFCC parameters are derived from a 320-points FFT, and a Bark window is applied to each filter bank.

In the recognition phase, a fast frame-synchronous tree-beam searching algorithm as shown in Fig. 2 is developed so that the acoustic processing model can give real-time response for the input speech. The searching algorithm uses a forward searching method to predict the possible active states for the next frame. Only a small number of states on the lexicon tree will be examinated before the beam-width is applied to prune out impossible states. Besides, by using the forward searching method, the response time of the acoustic processing model is affected by the selection of the state beam-width and is not affected by the lexicon size.

IV. LANGUAGE MODEL

The language decoder is used to select the most likely word from the word candidates that are output from the acoustic processing model. We setup a word-candidate pool to form a word lattice. The word-candidate pool is constructed as a first-in-first-out(FIFO) structure as shown in Fig. 3. Because the Viterbi scores are stored in the word pool, it is only necessary to compute the Viterbi scores for those word candidates that are last appended into the word pool.

Because trigram model needs much computational cost in the language decoder, a two pass Viterbi decoding algorithm is adopted to decode the word lattice into sentence. In the first pass, the bigram model is applied into the Viterbi decoder to pre-select the top N sentence hypotheses. The trigram model is then used to give scores to each sentence hypothesis. The two pass Viterbi decoding procedures are shown in Fig. 4.

To deal with the cross language information, a bilingual dictionary is used in the language decoder when the word transition from different language occurs. For example, considering the word sequence $w_1^J w_2^C w_3^J$, the Chinese word w_2^C will be converted into its corresponding Japanese word w_2^J so that the bigram and trigram information for word sequence $w_1^J w_2^J w_3^J$ can be estimated from the Japanese language model.

V. EXPERIMENTS

In the experiments, a dictionary that contains 32939 Chinese words and 44560 Japanese words is converted into a PLU-based lexicon tree. Two sets of text corpora which sizes are 24671492 Chinese words and 13904563 Japanese words respectively are used to estimate the language model parameters. We use a bilingual article which contains 323 Chinese words and 233 Japanese words for testing. The experimental result shows that the recognition rate is 90% in average. At the same time, 86.5% and 93.5% recognition rates are achieved for Chinese words and Japanese words respectively.

VI. CONCLUSION

In this paper, we proposed a bilingual speech recognition architecture to dictate Mandarin and Japanese speech simultaneously. Our proposed recognition architecture can not only be used to dictate mixed Mandarin and Japanese speech but also other languages. By changing the basic PLUs for acoustic model, the lexicon tree and the text corpus for learning language model parameters, our proposed system architecture provides an opened-environment for bilingual or multilingual speech recognition.

REFERENCES

[1] 蔡茂豐, " 日本語讀本, " 東吳大學日本文化研究所

[2] Jyh-Shing Shyuu and Jhing-Fa Wang, " A First Study on Converting a Mandarin Speech Recognition System into the Japanese Speech Recognition System, " ISMIP97 Proceeding, pp.69-74, Taipei, Taiwan.

[3] Jyh-Shing Shyuu and Jhing-Fa Wang, " An Algorithm for Automatic Generating of Mandarin Phonetic-Balanced Corpus, " To be published ICSLP98 Proceeding, Sydney, Australia..

[4] S. Austin, R. Schwartz, and P. Placeway, " The forward-Backward Search Strategy for Real-Time Speech Recognition, " IEEE ICASSP-91, Toronto, Canada, pp. 697-700, May 1991.

[5] P.S. Gopalakrishnan, L.R. Bahl and R.L. Mercer, " A Tree Search Strategy for Large-Vocabulary Continuous Speech Recognition, " IEEE, ICASSP-95, pp.572-575, Detroit, MI, May, 1995.

[6] D. Paul, " Algorithm for an Optimal A* Search and Linearizing the Search in the Stack Decoder, " IEEE, ICASSP-91, pp. 693-696, Toronto, Canada, May, 1991.

[7] S.K. Das, and M.A. Picheny, " Issues in Practical Large Vocabulary Isolated Word Recognition: The IBM TANGORA System, " Chapter 19, Automatic Speech And Speaker Recognition-

Advanced Topics, Kluwer Academic Publisher, 1996.

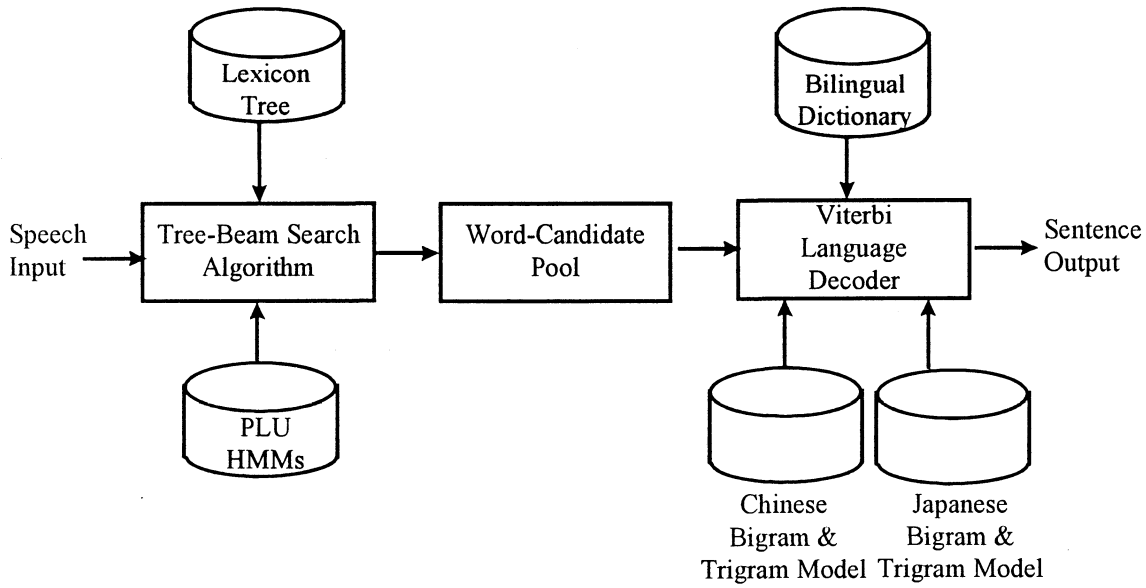

Fig. 1. System Block Diagram of the Bilingual Speech Recognition System

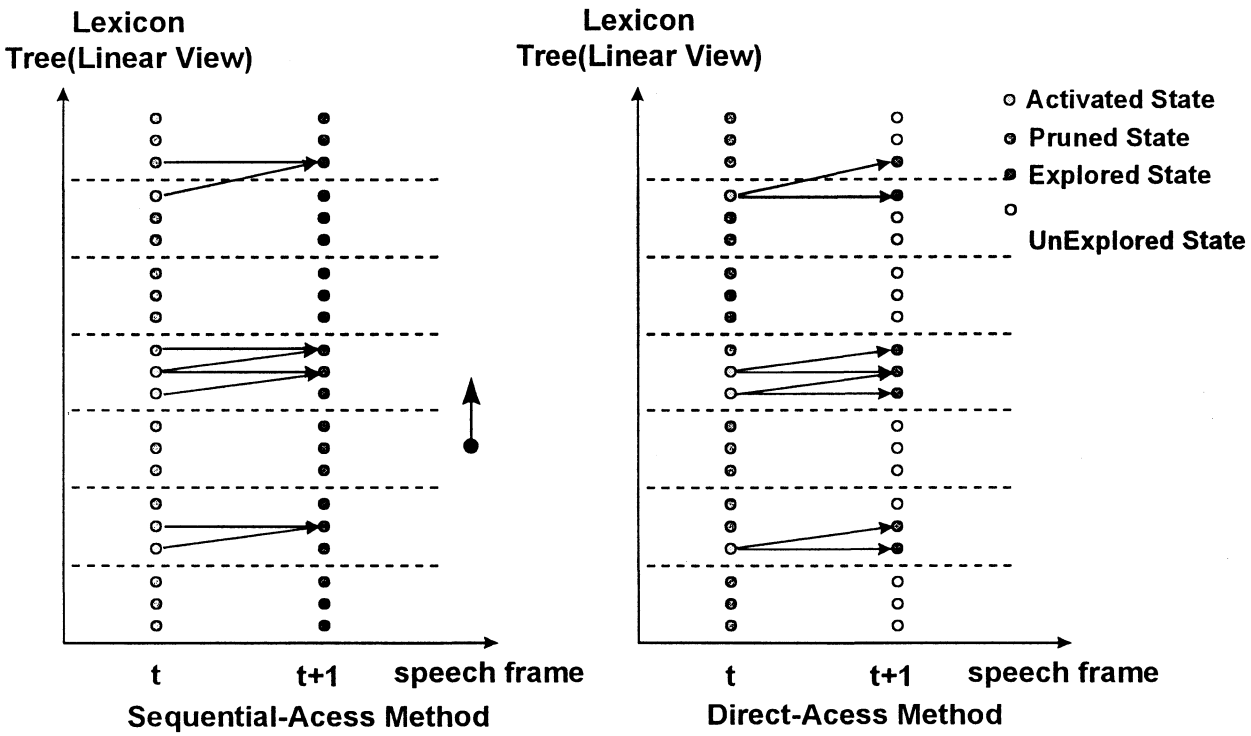

Fig. 2 Fast Tree-Beam Search Algorithm

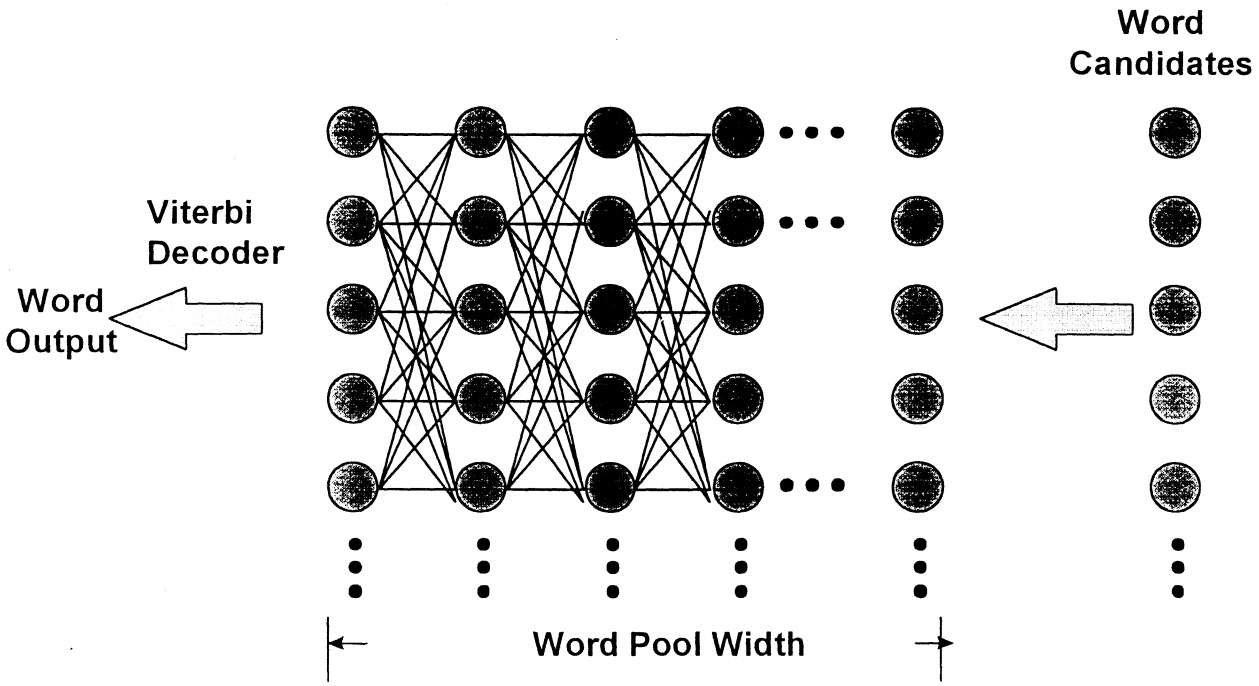

Fig. 3 FIFO Word-Candidate Pool

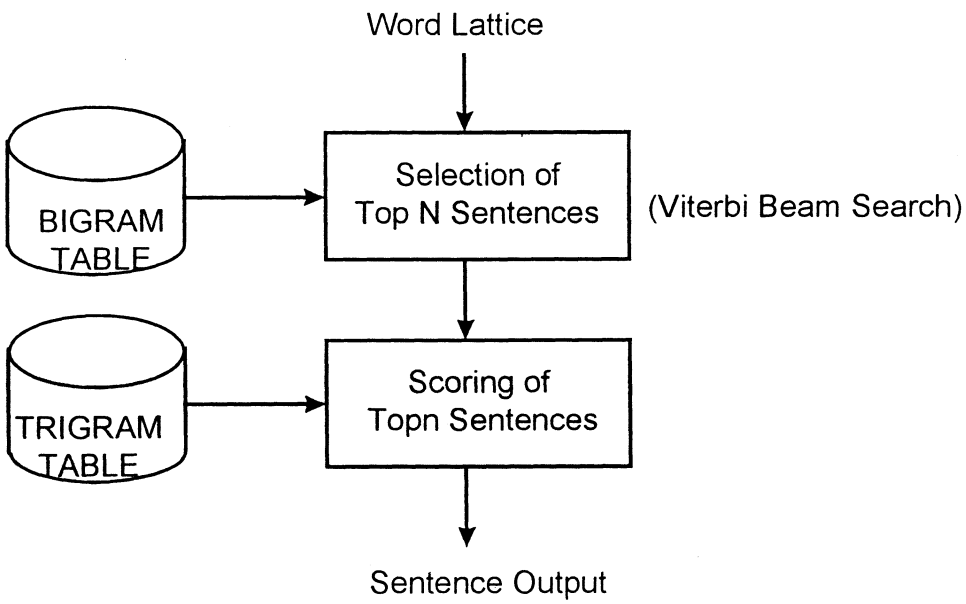

Fig. 4 Two-Pass Language Decoder