A MULTI-PHASE APPROACH FOR FAST SPOTTING OF LARGE VOCABULARY CHINESE KEYWORDS FROM MANDARIN SPEECH USING PROSODIC INFORMATION

Bo-Ren Bai[1], Chiu-Yu Tseng[2], and Lin-Shan Lee[1]

[1]Dept. of Electrical Engineering, National Taiwan University
[2]Institute of History & Philology, Academia Sinica
Taipei, Taiwan, R.O.C.
E-mail: white@speech.ee.ntu.edu.tw

ABSTRACT

This paper presents a multi-phase approach for fast spotting of large vocabulary Chinese keywords from a spontaneous Mandarin speech utterance using prosodic knowledge. Without searching through the whole utterance using large number of keyword models, the multi-phase framework proposed here including some special scoring schemes provides very good efficiency by considering the monosyllable-based structure of Mandarin Chinese. This approach is therefore very fast due to very good boundary estimations and the deletion of most impossible syllable and keyword candidates using context independent models, and also very accurate with the carefully designed scoring processes. A task with 2611 keywords was tested here. An inclusion rate of 85.79% for the top 10 candidates is attained, at a speed requiring only 1.2 times of the utterance length on a Sparc 20 workstation.

1. INTRODUCTION

Many different algorithms have been proposed for detecting a predefined set of keywords from continuous speech. Most of them use hidden Markov model (HMM) based continuous speech recognition techniques and require word/sub-word and filler models to decode an input speech utterance into a sequence of keywords and non-keywords[1][2]. To improve the performance, many different methods have been developed to adjust the scores of the keywords appearing in the Viterbi alignment[1][3][4]. However, it is always very difficult to train filler models for the non-keyword speech, and it is even more difficult to model the lower level events such as non-speech noise[5]. In addition, most reported keyword spotting techniques only consider the small vocabulary problems. Huang, *et. al.*[6] proposed an algorithm to deal with the large vocabulary problems, it needs to carefully design the filler models. In our previous research [7], we have presented a new strategy for large vocabulary Chinese keyword spotting without training the filler models, in which keywords were directly detected from the local segments of the speech utterance using modified very-large-vocabulary continuous Mandarin speech recognition techniques[8]. To further improve the performance of our approach, including the speed and accuracy, this paper presents a multi-phase approach which takes advantage of the prosodic information and carefully designs the word-spotting procedure such that the accuracy and the speed can be further improved.

In Chinese language all the characters are monosyllabic, a word is composed of one to several characters, and the total number of syllables is relatively small. Taking advantage of this monosyllabic structure, we have achieved a great success in the continuous speech Mandarin dictation technique. But when a keyword spotting problem is considered, the ability to deal with the spontaneous speech is required. In spontaneous speech, it is always full of lower level events such as pauses, filled pauses (e.g. "uh"), hesitation, laughter and other non-speech noises (inhalation, cough), etc., so it is difficult to recognize such an utterance[5]. After carefully considering the monosyllabic structure of Mandarin speech, the different phenomena between spontaneous speech and read speech, and efficiency and accuracy requirements for large vocabulary Chinese keywords spotting, a multi-phase approach with a special scoring method is developed. The first phase is to estimate the possible syllable boundaries as well as the obvious phrase boundaries on the speech utterance using the prosodic information. The utterance is then decoded into a syllable lattice via context independent sub-syllabic models in the second phase. A fuzzy search is then performed to pick up several keyword candidates in the third phrase, and the elaborate keyword models composed of context dependent sub-syllabic models are finally used in the fourth phase to re-score the keyword candidates. The log-likelihood score normalization techniques based on the background models[1], anti-keyword models[4] and some other special techniques are also adopted to improve the reliability and efficiency.

The rest of this paper is organized as follows. In Section 2, the multi-phase approach is first introduced in detail. In Section 3, some experimental results are shown. Finally, in Section 4, some conclusions are drawn.

2. MULTI-PHASE APPROACH

The block diagram of the multi-phase approach is shown in Figure 1. And here in this section, we will give the approach a detailed description phase by phase. The first phase is designed for possible syllable boundaries estimation, the second phase for acoustic recognition, including both base syllable and tone recognition, the third phase for keyword matching, while the fourth phase for re-scoring.

2.1 Phase 1 – Syllable boundaries estimation

For a given input speech utterance, prosodic information such as energy contour, zero crossing rate, pause duration are used in a dynamic program process to evaluate the probability for each frame to be a syllable boundary, and then all possible syllable beginning frames can be identified, such as x, y, z in Figure 2, in which the energy contour is an example input. Corresponding to a possible syllable beginning frame such as x in Figure 1, the possible ending frames for that syllable such as $y-1$, $z-1$ can then be found using the estimated minimum and maximum duration of a syllable, e.g. $D_{\min}$ and $D_{\max}$. The possible syllable boundaries estimated in this phase can be further divided into three kinds of boundaries. The hard syllable boundaries for those with relatively high probabilities to be syllable boundaries, the soft syllable boundaries for those with lower probabilities, and the phrase boundaries determined, which are of course also hard syllable boundaries. In this phase, the phrase boundaries can be roughly estimated by the pause duration, and in the later phase, phrase final lengthening[9] can help to estimate this kind of boundaries better. Distinguishing among these three kinds of boundaries is very helpful in improving the spotting speed.

2.2 Phase 2 – Acoustic Recognition

Given all possible syllable beginning frames and their corresponding ending frames, we then perform syllable recognition for each utterance segment which may include a syllable in the second phase. Only context independent sub-syllabic models, initial and final models are used to save the time, where "initial" is the initial consonant of a syllable, and "final" is everything in the syllable after the "initial", including the vowel or diphthong part plus optional medial or nasal ending[8]. In order to solve the problem in the Viterbi search that the likelihood scores over different time sequences may need different reference values, so it is difficult to compare them directly, the concept of background models[1] has been used here to reduce the score variation. In other words, in addition to evaluating the conditional probabilities $\log p(O|s_i)$, $\log p(O|s_f)$ of an observation sequence O given context independent initial and final models s_i and s_f, an extra conditional probability $\log p(O|g)$ of the observation sequence given the background model g is also evaluated for score normalization purposes. Three background models are used here, g_n for background noise, and g_i and g_f for initials and finals. The syllable recognition process including the scoring method for the background models is shown in Figure 3, i.e. the score for the initial part will be normalized by the score for g_n or g_i, whichever is larger, and the score for the final part will be normalized by the score for g_n or g_f, whichever is larger. And the score of a syllable will be the summation of the scores of the related initial and final.

When the syllable recognition process is performed, a syllable is not allowed to cross any hard syllable boundaries mentioned in the previous phase, this will help to save much time. To give better estimation of phrase boundaries, we can first get the mean and standard deviation of every syllable length from the speech database, and now more phrase boundaries can be estimated by comparing the recognized syllable length to the statistical syllable length. After this recognition phase, an asynchronous syllable lattice is then constructed in which N most possible syllable candidates are reserved for every segment which may include a syllable. The tone recognition and scores are handled similarly but separately. By this way, less storage memory is needed to store the acoustic recognition result, also in the later process, it is capable of getting more tonal syllables by combining these two kinds of acoustic scores.

2.3 Phase 3 – Keyword Matching

With the asynchronous syllable lattice obtained above including the acoustic scores for the syllable candidates, in this phase the most possible keyword candidates are found directly from the local segments of the speech utterance by searching through the syllable lattice. And to tolerate the information loss caused by the acoustic recognition error, here we adopt a fuzzy matching, instead of exact matching, together with a dynamically adjusted scoring method to pick out the keyword candidates form adequate location.

Since a possible ending frame of a syllable is also a possible ending frame of a keyword, starting from each ending frame, a backward search can be performed on the asynchronous syllable lattice to construct a possible keyword using the syllable candidates in the lattice. As shown in Figure 4, the root

node x_0 is a possible ending frame, and x_1^1, x_1^2 are the possible ending frames for the preceding syllable. Each of these two branches $(x_0 \rightarrow x_1^1)$ $(x_0 \rightarrow x_1^2)$ may include a syllable, and the N most possible syllable candidates with their scores and tone scores for each of such branch at x_0. Similarly, we can find all possible ending frames for the next preceding syllable starting from x_1^1 and x_1^2 and so on, until the length of this tree equals the maximum number of syllables in the desired keywords. Now every path on the tree, from the root node to a leaf node, can be matched to the desired keywords, i.e. the $i-th$ level branch is matched to the last $i-th$ syllable of the keywords, etc.. For every path within the tree, if the last $i-th$ syllable of the keyword is found among the N candidates of the $i-th$ level branch, the acoustic scores including the tone scores of the syllable will be accumulated to the keyword. If the desired syllable of the keyword is not one of the N syllables candidates stored, a relatively lower but dynamically adjusted score is also accumulated to the keyword to kept the path going on for further observation. When a leaf node is reached at, the accumulated score is then normalized by the duration and compared with those already stored in a stack. If this normalized score is high enough, the keyword with this score will be stored in the stack too. This keyword-matching process is repeated for every keyword starting at every possible ending frame. After all ending frames and all keywords have been searched through, the keywords with the highest scores will be picked up. Since a short keyword is seldom uttered separately as two parts, we can use the phrase boundaries estimated in the previous phrases to help reduce the search space.

2.4 Phase 4 – Re-scoring

The re-scoring process is now performed in this phase for each keyword candidate picked up in the previous phase. The acoustic models used in this phase are more elaborate models: the context dependent initial and final models plus anti-models trained by discriminative training, and the probabilities for the anti-models will be used for normalization[4]. Because the search space is now limited to the keyword candidates selected in the previous phase, use of the elaborate models won't take too much time. Also, since the models used in the second phase are relatively coarse, the segment boundaries for keywords obtained previously may not be very accurate, so we can reasonably relax the segment boundaries to a certain range in this phase. The keywords with the highest scores after re-scoring in this phase will be the output.

3. EXPERIMENTAL RESULTS

A task with 2611 keywords was tested with the algorithm presented in this paper. Each test utterance may include any number of keywords. Without the re-scoring process in phase 4, the background models used in phase 2 actually raised the inclusion rate for top 10 candidates from 73.34% to 79.56%. With the re-scoring process in phase 4 the inclusion rate was further improved to 85.79%, at speed requiring 1.2 times of the utterance length on a Sparc 20 workstation. Comparing to the results without using the prosodic information, whose speed may require more then 4 times of utterance length and it was impossible to achieve such good spotting result, the high speed of the presented approach was apparently achieved by the accurate estimation of syllable boundaries using prosodic information, and by the use of coarse context independent acoustic models in the second phase to delete many impossible syllables and keywords in the beginning.

4. CONCLUSION

In this paper, we present a multi-phase approach for spotting large vocabulary Chinese keywords from a spontaneous Mandarin speech utterance. Taking advantage of the mono-syllabic structure of Chinese language and prosodic information extracted from the speech utterance, a multi-phase framework is designed to spot keywords directly from the local segments within a speech utterance. A special scoring method and some techniques are also developed to further improve the efficiency and accuracy of the approach. Very attractive performance was demonstrated in the experiments.

References

[1] Richard C. Rose and Douglas B. Paul, "A Hidden Markov Model Based Keyword Recognition System", *Proc. 1990 IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Vol. 1, pp. 129-132, 1990.

[2] Rafid A. Sukkar, Anand R.Setlur, Mazin G. Rahim, and Chin-Hui Lee, "Utterance Verification of Keyword Strings Using Word-Based Minimum Verification Error (WB-MVE) Training", *Proc. 1996 IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Vol. 1, pp. 518-521, 1996.

[3] Mitchel Weintraub, "LVCSR Log-likelihood Ratio Scoring For Keyword Spotting", *Proc. 1995 IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Vol. 1, pp. 297-300, 1995.

[4] Mazin G Rahim, Chin-Hui Lee and Biing-Hwang Juang, "Robust Utterance Verification for Connected Digits Recognition", *Proc. 1995 IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Vol. 1, pp. 285-288, 1995.

[5] Ron Cole, et al, "The Challenge of Spoken Language Systems: Research Directions for the Nineties", *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 1, pp. 1-21, January 1995.

[6] Eng-Fong Huang, Hsiao-Chuan Wang, and Frank K. Soong, "A Fast Algorithm for Large Vocabulary Keyword Spotting Application", *IEEE Trans. On Speech and Audio Processing*, Vol. SAP-2, No.3, pp. 449-452, July 1994.

[7] Bo-Ren Bai, and Lin-Shan Lee, "A Local-Segment-Based Approach for Spotting Large Vocabulary Chinese Keywords from Mandarin Speech", *Proc. 1996 Int. Computer Symposium.*

[8] Hsin-min Wang, Jia-lin Shen, Yen-ju Yang, and Lin-shan Lee, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data", *Proc. 1995 IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Vol. 1, pp. 61-64, 1995.

[9] Donia R. Scott, "Duration as a Cue to the Perception of a Phrase Boundary", *Journal of the Acoustic Society of American*, Vol. 71, No.4 , pp. 996-1007, April 1982.

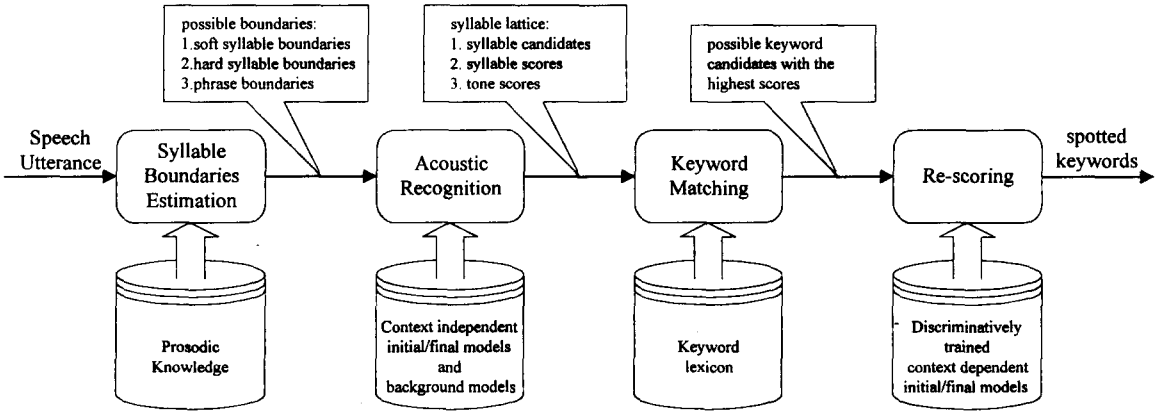

Figure 1. The block diagram of the multi-phase approach

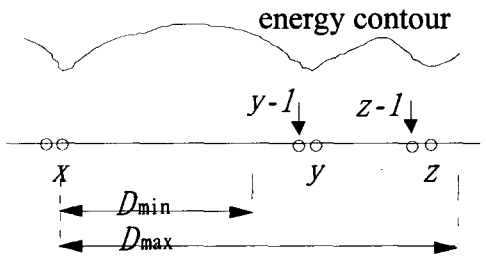

Figure 2. A section of an example utterance. *x*, *y*, and *z* are possible syllable beginning frames, while *y-1* and *z-1* are possible syllable ending frames corresponding to *x*.

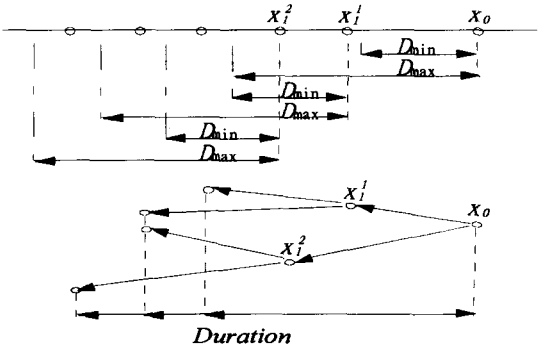

Figure 4. An example of a searching tree for a two-syllable keyword ending at x_0.

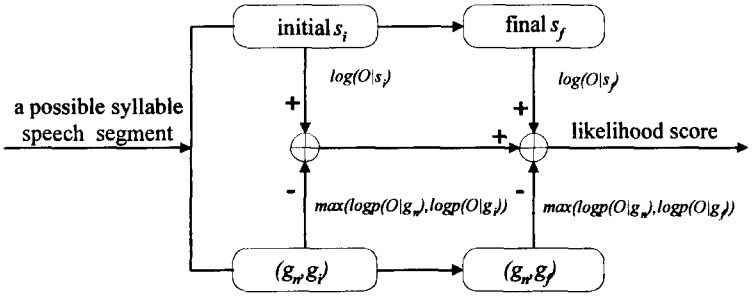

Figure 3. The block diagram for using the background models.