A*-ADMISSIBLE KEY-PHRASE SPOTTING WITH SUB-SYLLABLE LEVEL UTTERANCE VERIFICATION

Berlin Chen[1,2]*, Hsin-min Wang* [2]*, Lee-feng Chien* [2]*, and Lin-shan Lee* [1,2]

[1]Dept. of Computer Science & Information Engineering, National Taiwan University
[2]Institute of Information Science, Academia Sinica,
Taipei, Taiwan, Republic of China
E-mail: berlin@iis.sinica.edu.tw

ABSTRACT

In this paper, we propose an A*-admissible key-phrase spotting framework, which needs little domain knowledge and is capable of extracting salient key-phrase fragments from an input utterance in real-time. There are two key features in our approach. Firstly, the acoustic models and the search framework are specially designed such that very high degree vocabulary flexibility can be achieved for any desired application tasks. Secondly, the search framework uses an efficient two-pass A* search to generate N-best key-phrase candidates and then several sub-syllable level verification functions are properly weighted and used to further improve the recognition accuracy. Experimental results show that the A*-admissible key-phrase spotting with sub-word level utterance method outperforms the baseline methods used in common approaches.

1. INTRODUCTION

In recent years, various spoken dialog systems have been widely investigated for the fast growing demand for real-world applications. It is difficult to recognize every portion of the unconstrained and spontaneous input utterances by a conventional large vocabulary continuous speech recognizer (LVCSR) with n-gram statistical grammar rules. On the other hand, the spotting-based approach, which extracts the semantically significant fragments of users' utterances and ignores unrecognized portions, has been found very useful in such problems, especially in dealing with the ill-formed spontaneous utterances including hesitations, repetitions, out-of-vocabulary words and so on. A spotting-based approach that does not need a large task-specific training corpus, and thus with high degree vocabulary flexibility and the portability to different application tasks is highly desired. In most of the spotting-based approaches, keywords are commonly used as the templates of spotting. However, these small templates are easily confused with each other or the local noise. Therefore, using key-phrases, each of which contains a few keywords and function words, as basic units has the ability to realize robust acoustic matching and language understanding of the unconstrained and spontaneous utterances.

In this paper we propose an A*-admissible key-phrase spotting framework, which needs little domain knowledge and is capable of extracting salient key-phrase fragments from an input utterance in real-time. There are two key features in our approach. Firstly, the acoustic models and the search framework for speech recognition are specially designed to be independent of the vocabulary, such that very high degree vocabulary flexibility and recognition accuracy can be achieved for any desired application tasks. Secondly, the search framework is based on an efficient two-pass A* search instead of some segmented or aligned beam search [1], in which the heuristic functions satisfying the search admissibility requirements are easily generated from the key-phrase lexicon structure without using any domain-specific knowledge [2]. The filler models, which contain a general acoustic model, a silence model and several syllable models, are used to absorb the noisy or out-of-vocabulary fragments of the input utterances to enhance the spotting efficiency. Besides, several sub-syllable level verification functions are included and weighted for utterance verification [3] to further improve the recognition accuracy. With all the key features described above, the key-phrase spotting system has been successfully applied to many related applications such as a bank telephone number query system, a speech-activated WWW browser, a Chinese text/speech information retrieval system and so on [4-5].

The rest of this paper is organized as follows: Section 2 presents an overview of our approach. In Section 3 and Section 4, we formulate the spotting problem and the verification problem, respectively. Some experimental results of key-phrase spotting are then discussed in Section 5 followed by the conclusion in Section 6.

2. OVERVIEW

The search framework proposed here is based primarily on a lexical network concatenated with a left filler model and a right filler model as shown in Figure 1. Each arc of the lexical network represents a sub-word unit (a syllable) thus the network is able to handle arbitrarily assigned vocabulary set for any desired application tasks without training on the specific words. There are 416 toneless syllables used here to compose the words/phrases of the vocabulary. In addition, the syllables are further decomposed into sub-syllabic units, which are INITIAL's and FINAL's [6]. The INITIAL is the initial consonant of the syllable while the FINAL is the vowel (or diphthong) part of the syllable but including an optional medial or nasal ending. This monosyllabic structure of the Chinese language actually becomes the key for the vocabulary-flexible key-phrase spotter here. Thus, in our approach, syllable is chosen as the sub-word unit and each syllable is composed of two sub-syllabic models. In addition, the left and right filler models that consist of a silence model, a general acoustic model, and a several syllable filler models are used to absorb out-of-vocabulary events and to handle ill-formed input utterances.

For all application tasks, the key-phrases can be top-down manually determined to match semantic representations, or be bottom-up trained and selected from the text corpus or the

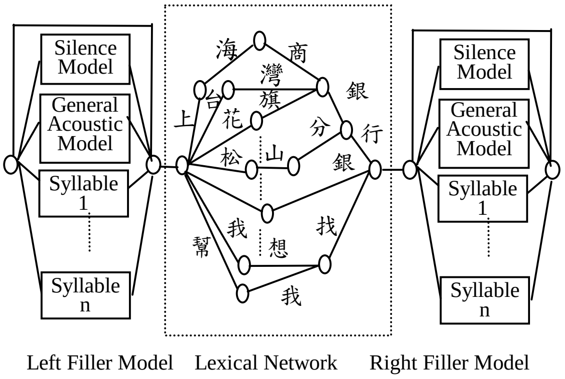

Figure 1: The search framework of key-phrase spotting.

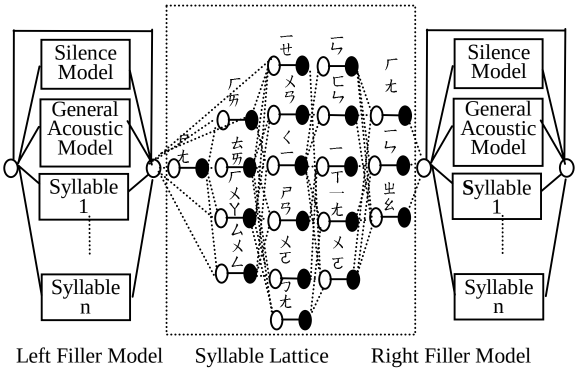

Figure 2: The structure of the compact syllable lattice and the filler models in the first pass.

transcription of the speech corpus through the mutual information criterion [7],

$$MI\,(w_1, w_2) = \log_2 \left[\frac{P(w_2 | w_1)}{P(w_2)} \right], \tag{1}$$

where w_1 and w_2 could be keywords or function words and the word pair $w_1 w_2$ is considered as a candidate key-phrase if it has high mutual information. This measure can be iterated to recursively construct longer units. Finally, the key-phrase set for a specific application task can be obtained.

3. KEY-PHRASE SPOTTING

The key-phrase spotting process here is based on a two-pass A* search strategy. In the first-pass search, the left and right filler models are decoded left-to-right and right-to-left respectively, with their Viterbi scores stored at every time t, i.e.,

$$f(t) = a \cdot sil(t) + b \cdot \overline{syl}(t) + (1 - a - b) \cdot fil(t), \tag{2}$$

where $f(t)$ represents the score of the left filler f_L or the right filler f_R at time t, while $sil(t)$, $\overline{syl}(t)$, and $fil(t)$ are the silence model score, the average score of syllable filler models, and the general acoustic model score at time t, respectively. a and b are weighting constants that can be empirically tuned. At the same time, a compact syllable lattice with the respective lattice node scores evaluated left-to-right including the left filler model scores is derived from the lexical network based on the constraint grammar considering the key-phrase structure. A partial list of the simplified compact syllable lattice is shown in Figure 2. In the syllable lattice, each arc stands for a syllable and is corresponding to one or several arcs on the lexical network. The end node (marked on black) of each arc stores its cumulative score at every time t, which will be used as the heuristic function for its corresponding arcs (or nodes) in Figure 1. For each arc k, the heuristic function stored in its respective end node n_k at time t in the syllable lattice is then represented as,

$$h^*(n_k, t) = \underset{0 \le t_1 < t}{MAX} \left[f_L(t_1) + h(n_k, t_1 + 1, t) \right], \tag{3}$$

where $f_L(t_1)$ is the cumulative score of the left filler model at time t_1 and $h(n_k, t_1 + 1, t)$ is the cumulative score of the best path which enters the syllable lattice at time $t_1 + 1$ and passes through the arc k at time t. That is, the heuristic function will satisfy the search admissibility requirements in the second-pass search.

In the second-pass search, a backward time-asynchronous A* search [8] on the right filler model and the lexical network is performed right-to-left with the aid of the heuristic functions obtained previously in the first pass. For a given partial path P extended to end node n_k of arc k, its evaluation function $E_p(n_k)$ is represented as,

$$E_p(n_k) = \underset{0 < t < T}{MAX} \left[d_p(n_k, t) + h^*(n_k, t - 1) \right], \tag{4}$$

where $d_p(n_k, t)$ is the score obtained by maximizing the combination of the right filler model score and the exactly decoded score of the extended partial path P, and T is the duration of the input speech utterance. $d_p(n_k, t)$ can be expressed as,

$$d_p(n_k, t) = \underset{t < t_2 < T}{MAX} \left[g_p(n_k, t, t_2 - 1) + f_R(t_2) \right], \tag{5}$$

where $g_p(n_k, t, t_2 - 1)$ is the score of partial path P extended to node n_k from time t to $t_2 - 1$, $f_R(t_2)$ is the cumulative score of the right filler model at time t_2. In the A* search process, at every iteration a partial path with the maximal evaluation function is popped, extended, and stored in a stack, and then the stack is sorted by the evaluation functions of all these extended partial paths. This process is terminated when the desired N-best complete paths (candidate key-phrases) are obtained sequentially.

4. SUB-SYLLABLE LEVEL UTTERANCE VERIFICATION

For each spotted key-phrase, the sub-syllable level verification is then performed. The sub-syllable level verification score of a specific sub-syllabic unit s is defined as a log likelihood ratio (*LLR*),

$$LLR_s = \log \frac{P(O|\lambda_0)}{P(O|\lambda_1)}, \quad (6)$$

where O is the observed speech segment obtained from Viterbi decoding, λ_0 is the corresponding sub-syllabic model (INITIAL or FINAL), and λ_1 is the competing model. To achieve the competing model, a specific anti-model for each sub-syllabic model is trained while the *m* most confusing sub-syllabic models obtained from Viterbi decoding on the speech segment are also used. Therefore, $P(O|\lambda_1)$ is obtained by properly weighting the above two competing models. Then, the log likelihood ratio is transformed into a range between 0 and 1 by a Sigmoid function ζ,

$$\zeta(LLR_s) = \frac{1}{1 + \exp(-\alpha \cdot (LLR_s - \beta))}, \quad (7)$$

where α and β are used to control the slope and the range of the sigmoid function. Furthermore, the confidence measure (*CM*) of a key-phrase candidate k is represented as,

$$CM_k = \frac{1}{S_k} \sum_s \zeta(LLR_s), \quad (8)$$

where S_k is the total number of sub-syllabic units contained in the key-phrase candidate k. Only the key-phrase candidates with the confidence measures upper than a predetermined threshold τ will be accepted while those with lower confidence measures will be rejected.

5. EXPERIMENTS

Two experiments on a bank telephone number query system were performed for evaluation, in which the key-phrase spotting was used to extract the key-phrases within input utterances. In the first experiment, there were a total of 1,300 testing utterances used for testing, which were recorded off-line by 15 male speakers and all utterances were grammatically valid. Each utterance contained only one key-phrase (bank name) or no key-phrase at all. Two vocabulary sets with different vocabulary size were used for evaluation: one with 450 key-phrases and the other with 2400 key-phrases. In the second experiment, 500 testing utterances collected from the on-line system were used for evaluation, they were spoken spontaneously by 7 male and 3 female speakers. Most of them were in-grammar utterances but some of them were out-of-grammar utterance. Every testing utterance contains about 2.1 key-phrases on the average, such as bank names, query phrases, and other phrases, and the vocabulary set contained about 2500 key-phrases.

5.1 Database, Feature Extraction, and Modeling

In this paper, the acoustic modeling was based on the speaker-independent HMM's, which were trained by the training database that contains 5.5 hours of microphone speech materials recorded by 100 male and 40 female speakers. Each frame of a training utterance was conducted front-end feature extraction and represented by a 24-dimensional feature vector including 12 MFCCs (Mel-Frequency Cepstral Coefficients) and 12 Delta-MFCCs. The acoustic units chosen here were 112 right-context-dependent INITIAL's and 38 context-independent FINAL's. Each INITIAL was represented by a HMM with 3 states while each FINAL with 4 states. The state Gaussian mixture number ranged from 2 to 8. Therefore, every syllable unit was represented by a 7-state HMM. The general acoustic model was a 7-state HMM with 32 Gaussian mixtures, which was specially designed to capture the broad range acoustic events in Mandarin Chinese, and was trained on all of the speech data. The silence model was a 1-state HMM with 32 Gaussian mixtures trained on non-speech segments. Furthermore, in order to perform utterance verification, for each INITIAL/FINAL, a context-independent anti-model represented by a 3-sate/4-state HMM with 16 Gaussian mixtures was trained. There were a total of 58 anti-models in our experiments. The speech data was automatically labeled and then used to train all the above continuous density HMMs. Furthermore, Minimum Classification Error (MCE) training [3] was adopted to train these acoustic models. In this paper we adopted two approaches to implement MCE training: sentence-based MCE training and sub-word-based MCE training. The sentence-based MCE was applied on the whole utterance while sub-word-based MCE was applied individually on the sub-word units (INITIAL's/FINAL's).

At the same time of writing this paper, some preliminary experiments of key-phrase spotting based on the telephone speech data were under study.

5.2 Experimental Results

In the first experiment, the performance of single key-phrase spotting and the effect of the sub-word level utterance verification were evaluated. It was found that when only with a simple phrase duration constraint, the detection rates of 94.1% and 93.5% with respective to the false alarm rates of 12.8% and 16.5% for the 450-key-phrase task and the 2400-key-phrase task were obtained respectively. Here, the detection rate was defined as the correctly spotted key-phrases over the total key-phrases and the false alarm rate as the incorrectly spotted key-phrases over the accepted key-phrases, which contained correctly and incorrectly spotted key-phrases. Besides, the false reject rate was defined as 1 - the detection rate. If sub-syllable level utterance verification was further used, the best detection rates of 93.5% and 92.9% at the false alarm rates of 3.2% and 4.7% could be obtained for the 450-key-phrase task and the 2400-key-phrase task respectively. In Figure 3, we depict the ROC curves (Receiver Operating Characteristics) for these two tasks. It could be found that the performance only degraded slightly when the vocabulary size increased from 400 key-phrases to 2400 key-phrases. Furthermore, three verification mechanisms to obtain the final confidence measure were evaluated for comparison, including the sub-syllable level verification function, the syllable level verification function, and a simple verification function that normalized the key-phrase score with respective to the score obtained from free-syllable decoding on the key-phrase segment. Figure 4 shows the ROC curves of these three mechanisms on the 2400-key-phrase task. It can be found that sub-syllable level utterance verification significantly outperforms the other two mechanisms. In the second experiment, we evaluated the performance of multiple key-phrase spotting and the effect of MCE training. In the first stage, 30 key-phrase candidates, initially selected from the key-phrase spotting process, were used to achieve a higher detection rate at the expense of higher false alarm rates. Then utterance verification was used to fill out those key-phrase candidates. Figure 5 shows the ROC curves obtained by using the acoustic original models, the acoustic models further trained by sentence-based MCE training and the acoustic models further trained by sub-word-based MCE training. The acoustic models trained by the sentence-based MCE training are better than the original models while the acoustic trained by the sub-word-based MCE training are worse than the original models.

6. CONCLUSION

We have developed a vocabulary-flexible key-phrase spotting system for Mandarin Chinese, which needs little domain knowledge and is capable of extracting salient key-phrase fragments from an input utterance almost in real-time. Besides, several utterance verification mechanisms were developed to further improve the recognition accuracy. Experimental results has shown that the A*-admissible key-phrase spotting approach with sub-word level utterance method outperforms the baseline methods used in common approaches.

7. REFERENCES

1. A. S. Manos and V. W. Zue, ” A segment-based word spotter using phonetic filler models,” Proc. *IEEE-ICASSP97*, pp. 899-902.

2. T. Kawahara, N. Kitaoka, and S. Doshita, “Concept-based phrase spotting approach for spontaneous speech understanding,” Proc. *IEEE-ICASSP96*, pp. 291-294.

3. Rafid A. Sukkar, Anand R. Setlur, Mazin G. Rahim, and Chin-Hui Lee, “Utterance verification of keyword strings using word-based minimum verification error (WBMVE) training.,” Proc. *IEEE-ICASSP96*, pp. 518-521.

4. Hsin-min Wang, Bor-shen Lin, Berlin Chen, and Bo-ren Bai, “Towards a Mandarin voice memo system,” Proc. *ICSLP98*.

5. Hsin-min Wang, Yu-hsueh Chou, and Berlin Chen, “Surfing the Chinese Web pages by unconstrained Mandarin speech,” Proc. *IEEE-ICCE98*, pp. 84-85.

6. Hsin-min Wang *et al.*,” Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data,” *IEEE Trans. SAP*, Vol. 5, No. 2, pp.195-200, March 1997.

7. A.L. Gorin, G. Riccardi and J.H. Wright.” How may I help you,” *Speech Communication* 23 (1997) pp. 113-127.

8. P. Kenny, R. Hollan, V. N. Gupta, M. Lennig, P, “Mermrlstein, and D. O'Shaughnessy, “A*-Admissible Heuristics for Rapid Lexical Access,” *IEEE Tran. on ASSP*, Vol. 1, No. 1, January 1993.

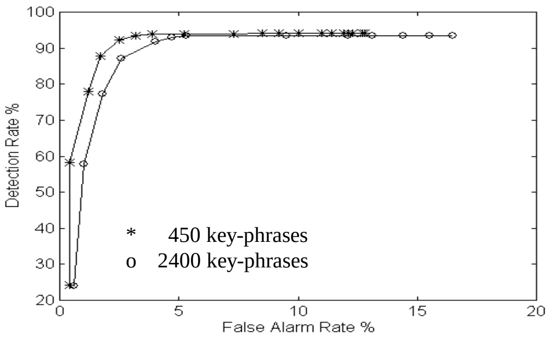

Figure 3: The ROC curves of the 450-key-phrase task and the 2400-key-phrase task.

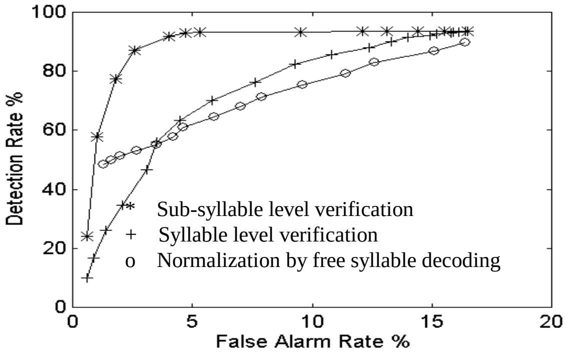

Figure 4: The ROC curves of three verification mechanisms on the 2400-key-phrase task.

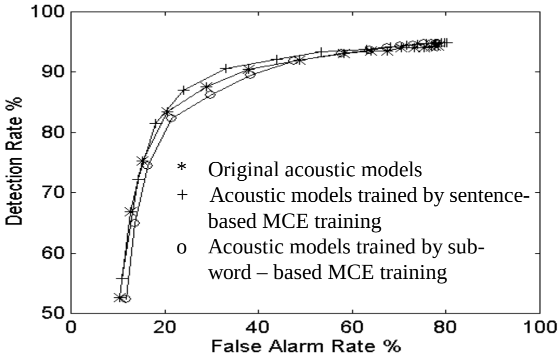

Figure 5: The ROC curves of three sets of acoustic models in the second experiment