International Symposium on Chinese Spoken Language Processing (*ISCSLP'98*), RSR-3, pp.202-206, Dec.7-9, 1998, Singapore

A VOCABULARY-INDEPENDENT KEYWORD SPOTTER FOR SPONTANEOUS CHINESE SPEECH

ZHENG Fang, XU Mingxing, MOU Xiaolong, WU Jian, WU Wenhu, FANG Ditang
Speech Laboratory, Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084
Tel. /Fax: +86 10 62772001, E-mail: fzheng@sp.cs.tsinghua.edu.cn

ABSTRACT

HarkMan keyword-spotter was designed so that it can be used in a real-world environment to automatically spot the given words of a vocabulary-independent (VIND) task in unconstrained Chinese telephone speech. In this spotter, the speaking manner and the number of keywords are not limited. This paper focuses on a novel technique that addresses acoustic modeling, keyword-spotting network, search strategies, robustness, and rejection adopted in *HarkMan*. The underlying technologies used in *HarkMan* given in this paper are not only for keyword spotting but also for continuous speech recognition, which had been proved very efficient. It achieved the figure-of-merit (FOM) value over 90%.

1. INTRODUCTION

Keyword spotting (KWS) has a wide range of application, such as message classification, topic identification, and speech-content-addressed applications. Moreover, KWS is also useful for a dictating system.

The basic structure of KWS can be filler-based whole-word spotting or phoneme-based VIND spotting. In the first structure, the FOM [10] value is often higher, but the system is less flexible, so the second type is preferable. Solutions for the continuous speech recognition (CSR) can be applied to KWS, but there are still many special issues for KWS like rejection. So this paper will focus on a novel technique dedicated to KWS in the *HarkMan* system.

The paper is organized as follows. In Section 2, the acoustic modeling is addressed, including the choosing of speech recognition units (SRUs), number of states and the number of densities in each state. In Section 3, the KWS network as well as the use of Bi-gram is described. In Section 4, the searching strategies adopted in *HarkMan* are described in details. In Section 5, our solutions to the robustness issues are given, the background noise, accent, gender, context, channel and so on are covered briefly. In Section 6, the rejection methods are given. The experimental results of *HarkMan* are given in Section 7.

2. ACOUSTIC REPRESENTATION

HarkMan keyword-spotter was designed to automatically spot the given words of a vocabulary-independent (VIND) task in unconstrained Chinese telephone speech under the real-world environment. For the real-world applications, the robustness issue is more important, which should be paid more attention to during the acoustic modeling. In this phase, the choosing of the training database, the acoustic modeling approach, and the speech recognition units is the first thing to do regarding the robustness issue.

2.1 Database Description

Because *HarkMan* is to be used in the real world, the speech data used to train the acoustic models were taken from the real world.

Speech signals over the telephone network where the Signal-to-Noise Ratio (SNR) was about 25dB were digitized at 8kHz sampling rate, they were compressed as A-law codes by the specified hardware. After expanded, the 13-bit linear PCMs were emphasized using a simple first-order digital filter, and then blocked into frames of 32 msec in length spaced every 16 msec. Having been weighted by the Hamming Window, each frame was represented by D-order (where D=10) LPC cepstral coefficients. Regression analysis [3] was applied to each time function of the cepstral coefficients over 5 frames every 16 msec and the regression coefficients were obtained then. Each of the two set of coefficients is constructed as a vector in a D-dimensional Euclidean space.

The database consists of speech data uttered by 200 people, and the amount is about 4GB. In this database, utterances were spoken very fast, the average Chinese syllable length is about 10 half-frames, i.e., 160 msec. This made the labeling and the modeling more difficult than in other applications.

2.2 Acoustic Modeling

The Center-Distance Normal (CDN) [16] distribution instead of the normal distribution was used here to describe the distance between a normal random vector and its statistical mean vector in the feature space.

The acoustic modeling here was based on Center-Distance Continuous Probability Model (CDCPM) [16] other than the Hidden Markov Model (HMM). A left-to-right CDCPM is in some sense similar to the HMM except that the CDCPM ignores the probability transition matrix, and is based on the CDN distribution instead of normal distribution. A mixture density CDCPM can be described by the following parameters: (1) N: number of states each model; (2) M: number of densities each state; (3) D: number of dimensions each feature vector; (4) $\mu_{xnm}=(\mu^{(d)}_{xnm})$: mean vector of the m'th density component in the n'th state; (5) μ_{ynm}: mean center-distance of the m'th density component in the n'th state. Here $1\le n\le N$, $1\le m\le M$, $1\le d\le D$.

Instead of using mixed CDN densities during the scoring procedure, an Embedded Multi-Model (EMM) scheme [16] was adopted. EMM scheme can expand a N-state M-density CDCPM to M^T T-state one-density CDCPMs when matched with any T-frame speech segment. It has been proved robust for gender-dependent, accent-dependent, and context-dependent models and so on.

2.3 SR Units

Choosing the speech recognition units (SRUs) is a very important issue in the continuous speech recognition (CSR). SRUs should have the following characteristics: (1) they are flexible to make up any grammatical unit, and (2) their corresponding acoustic models are robust.

Chinese is a syllabic language, each syllable consists of one initial followed by one final. An initial is often corresponding to one consonant phoneme while a final is made up of one, two, or three vowel phonemes. Phonemes or initials/finals are flexible but not robust, words are robust but not flexible. According to our previous experiments [15], the Chinese syllables are the best choice.

2.4 State Number and Density Number

In order to determine suitable N the number of states in the CDCPM, and M the number of densities in each state, a great deal of experiments were done across the database described in Section 2.1.

Experiments showed that the syllable recognition accuracy increases with N and/or M monotonously at a specified range [18]. Experiments also showed that given a maximum M value, using suitable M values individually for different SRUs performed better than always using the maximal M value [17]. In considerations of the performance and the complexity, choosing N=6 and $M\le$16 can get a satisfied result, the syllable accuracy is 80.5% and the accuracy of top 10 candidates is over 95%.

3. KEYWORD SPOTTING NETWORK

3.1 Basic network

The basic network for the filler-based KWS systems [5] is shown in Fig. 1 [10], where KW stands for keyword and FL for filler.

In such a network, the system operating point can be adjusted by changing the transition weights of the keywords and/or the fillers, where w_{Kp} $(1\le p\le P)$ is the keyword transition weights and w_{Fm} $(1\le q\le Q)$ the filler transition weights. In order to spot the keyworkds, the keyword weights are often bigger than those of the fillers.

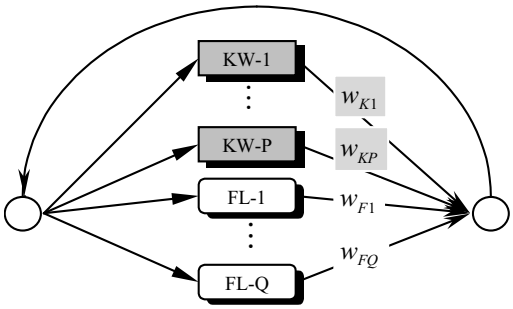

Fig. 1 The KWS network of P keywords and Q fillers

In *HarkMan*, each keyword was a catenation of several Chinese syllables, and according to the characteristics of Chinese and the above analysis the fillers were designed to consist of the following types: (1) single Chinese syllables, (2) silence (regarded as one special syllable), and (3) noise (regarded as another special syllable). Such a design was useful for acoustic modeling in VIND tasks.

3.2 Use of Language Models

The languages models used here in *HarkMan* were as simple as the syllable Bi-gram [18]. When transiting from one path to another through the KWS network, the connecting probability was considered. The connection could be KW-KW, KW-FL, FL-KW, or FL-FL. In such a situation, the probability of the two adjacent syllables across the transition point was calculated into the acoustic searching path. The syllable Bi-gram weights were different from the network transition weights. It was helpful for pruning some impossible syllable connection, thus resulted in higher efficiency and speed.

Although the syllable Bi-gram contains less or even no semantic information, it works well because of the concept of statistics.

4. SEARCH STRATEGIES

The frame-based searching algorithm is commonly used in the continuous speech recognition (CSR) [7]. Actually, the result of searching algorithm can be regarded as a segmentation of the input speech with maximum

likelihood, so it is natural to have an idea of finding the best segmentation method for a CSR system. But, it is not easy to find a method to exactly segment the speech into syllables because of the variety, complexity and co-articulation of the speech.

Anyway, it is well known that the continuous Chinese speech is intermittent at word boundaries. If the word-intermission information can be added into the acoustic searching procedure, not only the search efficiency and accuracy are improved but also the word-detection problem in the language processing is solved.

Our motivation here is to join the above search and segmentation strategies together to make use of their advantages jointly. The knowledge-based search is roughly as follows.

4.1 Segmentation

The frame-energy, zero-crossing-rate, pitch, and/or cepstrum or their derivatives (such as difference, contour, linear regressive analysis) [18] can be used to segment the speech, and then a series of putative separation points (PSPs) are obtained. These PSPs can be true separation points (TSP) or false separation points (FSP). In practical, we will choose a very confident threshold to ensure that the obtained PSPs are all TSPs.

Obviously, the bigger the threshold, the fewer PSPs we can obtain, but the bigger ratio of TSPs out of PSP. The threshold should be carefully chosen so that as many PSPs as possible can be detected and almost all of the PSPs are TSPs[18].

4.2 Searching Inside Every Definite Segment

The segmentation procedure gives several TSPs, any segment between two adjacent TSPs is refereed to as a definite segment (DS). A DS can also be a silence segment (SS). In each meaningful DS, there is possibly one or several Chinese syllables, i.e., a DS is often a syllable, a word or a phrase. According to the intermittence of speech, a DS will not contain too many syllables.

A frame-based searching procedure then can be performed using the knowledge obtained previously. The obtained knowledge is made full use of during the whole searching procedure. For instance, if a silence segment (SS) is detected, the obtained information is that it can not be a part of any syllable.

4.3 Pruning of Paths

The path pruning procedure is very necessary and common in CSR as well as in KWS system, it is definitely unavoidable especially when searching in a relatively long DS. Path pruning often occurs at such points as (1) the grammatical nodes in the KWS network, or (2) the TSPs.

Those paths that meet any one of the following conditions will be pruned: (1) there exists a state whose dwell is not inside a given range, (2) a TSP followed by a long SS is encountered but the path currently locates inside a keyword, (3) a TSP followed by a short SS is encountered but the path currently ends inside a syllable, (4) the accumulated path score is not among the top N candidates, or (5) the accumulated path score is lower than a given threshold.

5. ROBUSTNESS ISSUES

The robustness issue is a key issue to be faced in the real-world CSR applications as well as KWS ones. Robustness issue includes many items, such as background noises, different speakers (accents), different channels, different contexts (co-articulation), speed, and loudness. Many good approaches have been proposed to this problem [1][2][4][6][8] [9][10][11][12][13]. Here we will give our solutions.

5.1 NIL (Noise Immunity Learning)

The background noise issue is solved on the basis of Noise Immunity Learning (NIL) method [12][13]. Because there are many kinds of noises with different features, instead of building a kind of noisy-to-clean feature mapping or an individual model for each type, we use all the noisy speech data to train the models.

5.2 EMM (Embedded Multiple Model)

A direct approach to the speaker-dependent (including gender-dependent) and context-dependent items is to establish different set of models individually, such as Di-phone and Tri-phone models, gender-dependent models. But this kind of method will cost too much storage for models and time for searching. The Nearest Neighbor based EMM scheme for CDCPMs has been proved useful [16] and thus adopted here.

5.3 Arc-Splitting

Another item that can not be ignored is that in Chinese there are many different accents. It is something different from the different speaker issue. Different accents are due to different grown-up areas. Some syllables of this accent maybe the same as definitely different syllables of another accent. This problem is solved in the KWS network layer by the arc-splitting technique [18] instead of in the acoustic layer. For example, the ‘gui’ pronounced by a person from Sichuan Province and the ‘guo’ pronounced by a person from Beijing map to the same Chinese character, in the KWS network the ‘guo’ path is split into two parallel paths ‘guo’ and ‘gui’. The grouping of these

syllable is based on the known linguistic knowledge and the acoustic distance measure [18].

6. REJECTION METHODS

There are often two stages in a KWS system, (1) as many keyword candidates as possible are given so that the actual keywords will not be missed to guarantee the detection probability, and (2) a rejection/acceptation judgement is done to the given candidate list to reduce the false alarm rate (fa/h/kw).

So rejection plays a very important role in the two-stage keyword spotting system. The selecting of rejection methods should base on three factors: (1) the rejection quantity is different from that in the first stage, (2) the rejection quantity is a normalized one, (3) the rejection quantity is easy to calculate without extra training and modeling. Based on the above considerations, two kinds of rejection quantities were adopted in *HarkMan* system.

6.1 CAP: Percentage in Critical Area

The CDCPM is a modified version of HMM with left-to-right architecture [16], which eliminates the initial probability distribution and the probability transition matrix. The feature space of each state is divided into several sub-spaces described by one Center-Distance Normal (CDN) distribution [16]. These sub-spaces can be estimated by the clustering method according to a certain criterion [17].

For the Normal distribution $N(x; \mu_x, \sigma_x)$, about 95% samples fall into the critical area $[\mu_x-2\sigma_x,\mu_x+2\sigma_x]$. Similarly, for the normal-derived Center-Distance Normal (CDN) distribution $N_{CD}(y; \mu_y)$, about 95% samples fall into the critical area $[0, 2.5\mu_y]$, where y is the distance between normal vector $\boldsymbol{x}$ and its mean vector μ_x, μ_y is the mean value of y, and $\mu_y = \sqrt{2/\pi}\,\sigma_x$. CAP is based on the above discussion.

6.2 RSG: Recognition Score Gap

In the first-stage recognition module often outputs the K best candidates. The scores of top K candidates contain the information of the position of the correct answer. We found that the score differences between adjacent candidates are useful for the acceptation/rejection stage. There is often a large score gap between the candidates (including the correct one) and wrong ones, as shown in figure 2.

So a dynamic threshold is used to determine how many candidates should be reserved according to the score gaps.

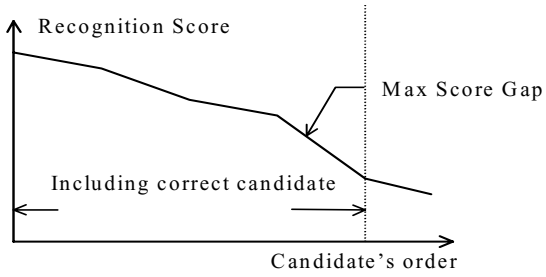

Fig.2 The Curve of Recognition Scores

6.3 Rejection Performance

In above rejection methods, CAPs calculate rejection score for each candidate independently. In other words the result score is only dependent on the utterance's feature vector sequence. Whereas the RSG method computes the rejection score according to the relation among candidates provided by recognition module at the first stage. Because CAP and RSG are based on different theories, there is less correlation between these two rejection methods. We can use them jointly, which gives better result than using individual one [14][18]. The experimental result is shown in figure 3.

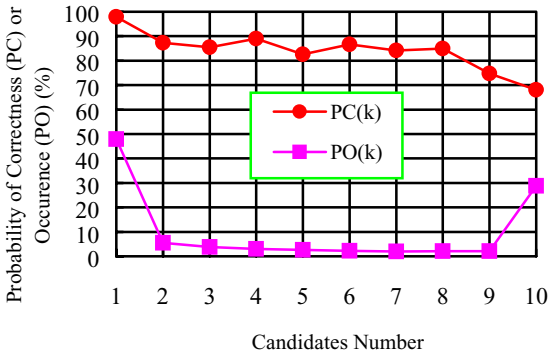

Fig.3 Rejection Results

Denote the total number of testing utterance samples by *TN*, the total number of utterance samples where k candidates are outputted in the acceptation/rejection stage by $T(k)$, and the total number of utterance samples where k candidates including the correct one are outputted by $C(k)$. Fig. 3 shows the curves of the Probability of Correctness $PC(k) = C(k) / T(k)$ and the Probability of Occurrence $PO(k) = T(k) / TN$.

Three quantities are defined to indicate the rejection performance. They are (1) the Total Rejection Accuracy (*TRA*) defined as the ratio of 'number of rejected candidates without correct candidates' to 'number of rejected candidates', (2) the Average Recognition Accuracy $ARA=\sum_k PC(k)*PO(k)$, and (3) the Average Candidate Number $ACN=\sum_k k*PO(k)$.

By combining the CAP and RSG, the rejection performance is TRA=99.73%, ARA = 86.33%, and ACN=3.46.

7. EXPERIMENTAL RESULTS

The receiver operating characteristics (ROC) curve of *HarkMan* is shown in Fig. 4. From Fig. 4, we can get that the FOM of *HarkMan* is 90.4% and the probability of detection (*Pd*) is 92.4% at the operating point fa/h/kw=5. In our experiments, the maximal keyword vocabulary size is 100 Chinese phrases, each phrase consists of 2 to 10 Chinese syllables.

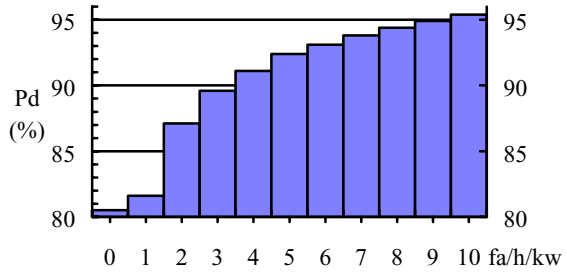

Fig. 4 The ROC Curve of *HarkMan*

REFERENCES

[1] Cox S.J., Bridle J.S., "Unsupervised speaker adaptation by probabilistic spectrum fitting," *ICASSP-89*, 3: 294-297

[2] Erell A., Weintraub M., "Spectral estimation for noise robust speech recognition," *Darpa Speech & Natural Language Workshop*, Cape Cod, MA, 1989

[3] Furui S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. on ASSP*, 34(1):52-59, Feb., 1986.

[4] Gish H., Chow Y.-L., Rohlicek J.R., "Probabilistic vector mapping of noisy speech parameters for HMM word spotting," *ICASSP-90*, 1: 117-120

[5] Higgins A.L., Wohlford Robert E., "Keyword Recognition Using Template Concatenation," *ICASSP-85*, 3: 1233-1236

[6] Juang J., Rabiner L.R., "Signal restoration by spectral mapping," *ICASSP-87*, 2368-2371

[7] Lee C.-H., Rabiner L.R., "A Frame Synchronous network search algorithm for connected word recognition," *IEEE Trans. on ASSP*, 37(11): 1649-1658, Nov. 1989

[8] Nadas A., Nahamoo D., Picheny M., "Speech recognition using noise-adaptive prototype," *IEEE Trans. on ASSP*, 37(10): 1495-1503, 1989

[9] Ng K., Gish H., Rohlicek J. R., "Robust mapping of noisy speech parameters for HMM word spotting," *ICASSP-92*, 2: 109-112

[10] Rohlicek J.R., Russel W., Roukos S., Gish H., "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting," *ICASSP-89*, 3: 627-630

[11] Rose R.C., Paul D. B., "A Hidden Markov Model Based Keyword Recognition System," *ICASSP-90*, 1: 129-132

[12] Takebayashi Y., Tsuboi H., Kanazawa H., "A Robust Speech Recognition System Using Word-Spotting with Noise Immunity Learning," *ICASSP-91*, 905-908

[13] Takebayashi Y., Tsuboi H., Kanazawa H., "Keyword-spotting in noisy continuous speech using word pattern vector sub-abstraction and noise immunity learning," *ICASSP-92*, 2: 85-88

[14] Xu M.-X., Zheng F., Wu W.-H., "Rejection in speech recognition based on CDCPMs," *'97 Int'l Conf. Research on Computational Linguistics*, 412-419, Aug. 1997

[15] Zheng F., Wu W.-H., Fang D.-T., "Speech recognition units in the Chinese dictation machines," *4th National Conf. on Man-Machine Speech Comm. (NCMMSC-96)*, pp.32-35, Oct. 1996, Beijing, P. R. China (in Chinese)

[16] Zheng F., Chai H.-X., Shi Z.-J., Wu W.-H., Fang D.-T., "A real-world speech recognition system based on CDCPMs," *'97 Int'l Conf. on Computer Processing of Oriental Languages (ICCPOL'97)*, 1: 204-207, Apr. 2, 1997, Hong Kong

[17] Zheng F., Xu M.-X., Wu W.-H., "Descriptions of the intra-state feature space in speech recognition," *'97 Int'l Conf. Research on Computational Linguistics*, 272-276, Aug. 22-24, 1997, Taiwan

[18] Zheng F., "Studies on approaches of keyword spotting in unconstrained continuous speech," Ph.D. Dissertation. Beijing: Dept. of Comp. Sci. & Tech., Tsinghua Univ., June 1997