Telephone Speech Multi-Keyword Spotting Using Fuzzy Search Algorithm and Prosodic Verification

Chung-Hsien Wu, Yeou-Jiunn Chen, and Yu-Chun Hung

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.

ABSTRACT

In this paper a fuzzy search algorithm is proposed to deal with the recognition error for telephone speech. Since the prosodic information is a very special and important feature for Mandarin speech, we integrate the prosodic information into keyword verification. For multi-keyword detection, we define a keyword relation and a weighting function for reasonable keyword combinations. In the keyword recognizer, 94 INITIAL and 38 FINAL context-dependent Hidden Markov Models (HMM‘s) are used to construct the phonetic recognizer. For prosodic verification, a total of 175 context-dependent HMM’s and five anti-prosodic HMM‘s are used. In this system, 1275 faculty names and department names are selected as the keywords. Using a test set of 3595 conversional speech utterance from 37 speakers (21 male, 16 female), the proposed fuzzy search algorithm and prosodic verification can reduce the error rate from 17.64% to 11.29% for multiple keywords embedded in non-keyword speech.

1. INTRODUCTION

In recent years, many algorithms have been developed to spot keywords from continuous speech. There have been several attempts to spot keywords from syllable lattice. Most of them use dynamic programming algorithm based on acoustic scores of syllables obtained from the Viterbi algorithm. However, the acoustic score of syllable is a global event and will be affected by other speech segments. In Mandarin speech, all syllables are monosyllabic and the unvoiced part of syllable is difficult to be correctly recognized for telephone speech. Thus exact keyword search from syllable lattice is not a suitable way for syllable based keyword spotting. In this paper, a fuzzy search algorithm is proposed to deal with the recognition error problem.

For multi-keyword detection, we define a keyword relation table and a weighting function for combining multiple keywords. According to the positions of keyword candidates in a sentence and the keyword relation, we can find reasonable keyword combinations. On the other hand, Chinese is a tonal language in which the same phonetic syllable when pronounced in different tones gives quite distinct meanings. The five tones, four lexical tones and one neutral tone, in Mandarin Chinese have lexical meaning. Conventionally, there are 408 Mandarin base syllables, regardless of tones, which are composed of 21 INITIAL‘s and 37 FINAL’s. Prosodic information, such as pitch and spectral energy in fundamental frequency, plays an important role in Mandarin speech recognition. In order to improve the performance of our approach, we integrate the prosodic information into keyword verification.

The block diagram of the multi-keyword spotting system is shown in Fig. 1. First, the phonetic and prosodic features are extracted. Hidden Markov Models (HMM‘s) with continuous observation densities are adopted to model the phonetic and prosodic features. The N-best Viterbi-Parallel Backtracking algorithm (VPB) [1] is employed to find the scores of the syllable lattice and their corresponding subsyllable boundaries. Subsyllable boundaries are then used to extract the FINAL part of Mandarin syllables, which contains the prosodic information. A fuzzy search algorithm is proposed to spot keywords from syllable lattice. For multi-keyword detection, we define a keyword relation table and a weighting function for combining multi-keywords. Finally, a keyword verification function combining phonetic-phase and prosodic-phase verification functions is investigated and used to reorder the ranks of the N best multi-keyword candidates.

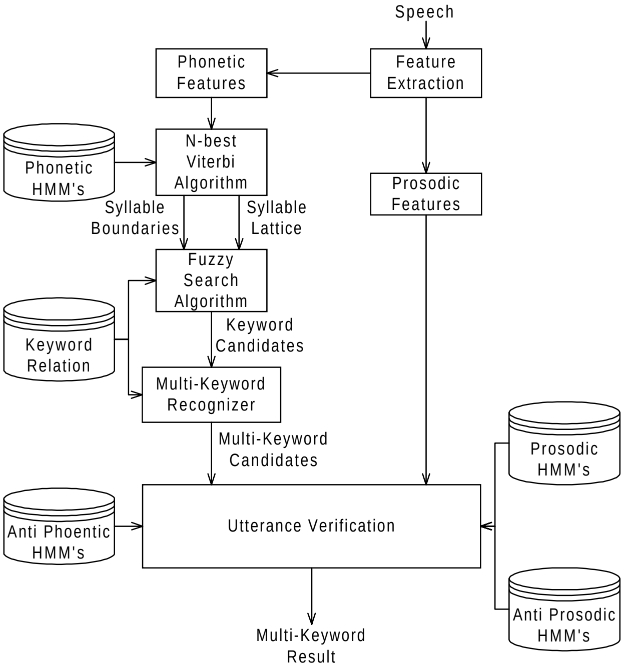

Fig. 1 The block diagram of the multi-keyword spotting system

2. FEATURE EXTRACTION AND HMM FOR SPEECH RECOGNITION

2.1 Feature Extraction

In the multi-keyword spotting system, first, the phonetic and prosodic features are extracted. For phonetic features, a 26-dimension feature vector is extracted. 12 Mel-Frequency Cepstrum Coefficient (MFCC), 12 delta MFCC, delta log energy, and delta delta log energy are adopted. The prosodic variations in human speech result from different speaking styles or emphasis of energies at different frequencies of the same utterance. These prosodic features are always encoded in duration, intensity, pitch contour, and spectral energy at fundamental frequency. A suitable representation of prosodic information is proposed to adequately model the different dynamic articulatory characteristics of that sequence when it is pronounced in different meaning contexts.

For prosodic verification, we adopt four parameters, proposed in [2], namely normalized pitch period, delta logarithmic pitch period, spectral energy at the fundamental frequency, and spectral energy at the formant frequency extracted from the FINAL parts.

2.2 HMM for Multi-Keyword Spotting System

In the keyword recognizer, 94 INITIAL and 38 FINAL context-dependent HMM‘s are used to construct the phonetic recognizer. Each INITIAL HMM consists of 3 states and each FINAL HMM consists of 5 states, each with 10 Gaussian mixture densities. In general, for every subsyllable model in the model set, a corresponding anti-subsyllable model for every subsyllable is trained specifically for the verification task.

Earlier investigations showed that the tone behavior is very complicated in continuous Mandarin speech, although there are only 5 different tones in Mandarin. Therefore, we assume every kind of possible tone combination needs a context-dependent model, then a total of 175 prosodic context-dependent HMM‘s will be needed. Five anti-prosodic HMM’s each corresponding to one context-independent lexical tone, are constructed to enhance the discriminability among prosodic HMM‘s. An anti-prosodic HMM can be considered as a lexical-tone-specific model. It is based on similar concept to the cohorts in speaker verification [3]. Each prosodic HMM has 4 states and 6 mixtures.

3. MULTI-KEYWORD SPOTTING

3.1 Fuzzy Search Algorithm

In the continuous speech recognition process, the VPB algorithm is adopted to generate a syllable lattice. The grammar used in the VPB is called no-grammar where any syllable can follow any syllable. The syllable boundaries associated with the optimal syllable string can be generated with a simple backtracking procedure. Therefore, the work of segmentation can be done by the Viterbi search. According to the syllable boundaries, we can backtrack separately to find the N-best syllables to construct syllable lattice.

With the syllable lattice, a fuzzy search method is used to extract the possible keyword candidates. In the fuzzy search algorithm, each syllable is decomposed into two subsyllables, namely INITIAL part and FINAL part in Mandarin speech. The likelihood between every two subsyllables is calculated in the training process. The nearest neighbor for each subsyllable is determined and used to compensate the substitution, insertion, or deletion errors. Using the fuzzy search algorithm, we can find the most likely keyword *K(i)*, where

$$K(i) = s_1^i s_2^i \Lambda\ s_L^i \,. \qquad (1)$$

The subsyllable string $s_1^i s_2^i \Lambda\ s_L^i$ is the subsyllable lexical representation of the *i*-th possible keyword, *K(i)* , and *L* is the number of subsyllables which are included in *K(i)*.

3.2 Multi-Keyword Recognition

For extending the single keyword spotting to multi-keyword spotting, we define a keyword relation table, which is used to decide the combination with two or more keywords in an utterance. The structure of the keyword relation is denoted as (PK, SK) where PK is the primary keyword and SK is the secondary keyword. In an utterance, the primary keyword is unique and necessary and the secondary keyword is not necessary but helpful. Fig. 2 shows the diagram of possible keyword paths. According to the keyword relation table, we can find multi-keyword candidates.

After extracting multi-keyword candidates in an utterance, a weighting function is applied to order the rank of candidates. Given a multi-keyword candidate (PK_i, Sk_i), the weighting function for the primary keyword is defined as follows:

$$W_{PK_i} = \begin{cases} 1, & \text{if no SK} \\ \left(\prod_{j=1}^{N} \dfrac{c}{1 + e^{-\lambda \times (Dist_{SK_j} - DN)}} \right)^{1/N} + (1-c), & \text{otherwise} \end{cases} \qquad (2)$$

where c, DN, and λ are constant. N is the number of secondary keywords and $Dist_{SK_i}$ is the score, which is obtained by Viterbi algorithm for the secondary keyword SK_i. Thus the weighted distance for the primary keyword is defined in the following:

$$WD_{PK} = W_{PK} \times Dist_{PK} \qquad (3)$$

where $Dist_{PK}$ denotes the distance of the primary keyword.

4. UTTERANCE VERIFICATION

Utterance verification can be treated as the problem of statistical hypothesis testing. The null hypothesis, H_0, represented by the input speech containing a given keyword K is tested against the alternative hypothesis, H_1, that K does not exist. According to the Neyman-Pearson Lemma [4], the optimal test is the likelihood ratio test such that the null hypothesis, H_0, is accepted if

$$LR(O;K)=\frac{L(O;H_0)}{L(O;H_1)}>\gamma \qquad (4)$$

where γ is the critical threshold of the test. In general, two types of errors can occur: false rejection (Type I) and false acceptance or false alarms (Type II) errors. In this verification process, a two-phase verification scheme is employed. An utterance verification combines the results of the phonetic-phase and prosodic-phase verification to make the final keyword acceptance/rejection decision.

4.1 Phonetic-Phase Verification

Given a subsyllable $s_n^{(k)}$, the normalized confidence measure is defined as

$$LR(O_{t_{n-1}}^{t_n};s_n^{(k)})=\frac{1}{T_n^{(k)}}\log[L(O_{t_{n-1}}^{t_n}\mid s_n^{(k)})]-\frac{1}{T_n^{(k)}}\log[L(O_{t_{n-1}}^{t_n}\mid \bar{s}_n^{(k)})] \qquad (5)$$

where $\bar{s}_n^{(k)}$ is the anti-subsyllable model of $s_n^{(k)}$, $T_n^{(k)}$ is the number of frames allocated for subsyllable $s_n^{(k)}$. For an N-subsyllable string $s_1^{(k)}s_2^{(k)}...s_N^{(k)}$ corresponding to the most likely keyword K, the whole word phonetic verification function is defined as follows:

$$L_s(O;K)=\frac{1}{N}\sum_{n=1}^{N}\alpha_n^{(K)}LR\left(O_{t_{n-1}}^{t_n};s_n^{(K)}\right) \qquad (6)$$

where $\alpha_n^{(k)}$ is the weight for subsyllable obtained from the subsyllable reliability. The value of $\alpha_n^{(k)}$ for the subsyllable $s_n^{(k)}$ is defined as

$$\alpha_n^{(k)}=\begin{cases}0.75 & if \;\; s_n^{(k)} is \;\; an \;\; INITIAL\\ 1.0 & if \;\; s_n^{(k)} is \;\; a \;\; FINAL\end{cases} \qquad (7)$$

The subsyllable weight for INITIAL is chosen smaller than that for FINAL. This is because that the INITIAL part in Mandarin syllable occupies just a short duration compared to the FINAL part and the recognition accuracy or reliability for INITIAL is lower than that for FINAL part.

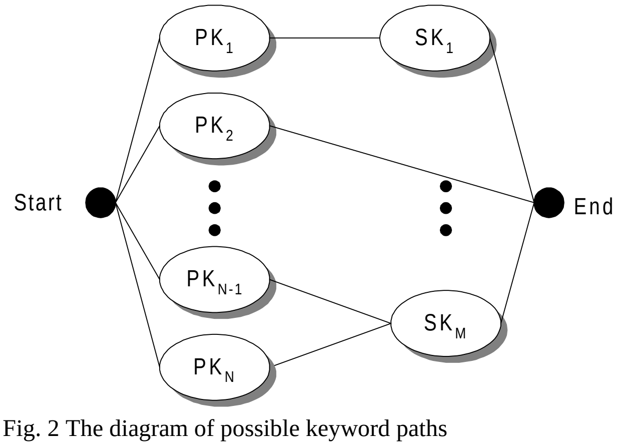

Fig. 2 The diagram of possible keyword paths

4.2 Prosodic-Phase Verification

For an N-subsyllable string $s_1^{(k)}s_2^{(k)}...s_N^{(k)}$ corresponding to the most likely keyword K, the corresponding lexical tone string T_K with respect to the keyword K is obtained using the sandhi rules [5] and written as

$$T_K=t_1^{(k)}t_2^{(k)}...t_N^{(k)} \qquad (7)$$

where N is the number of the FINAL subsyllables. Since most of the prosodic information is embedded in the FINAL part, the prosodic verification is only performed on the FINAL part. Given the prosodic feature vectors of a FINAL part corresponding to the lexical tone t_j, the prosodic confidence measure is written as

$$CM(P_j;t_j)=\log[G(p_{t_j};t_j)]-\log[G(p_{\bar{t}_j};\bar{t}_j)] \qquad (8)$$

where $P_j=[p_{t_j},p_{\bar{t}_j}]$ represents the verification feature vector, and $G(\bullet)$ is a Gaussian distribution of the verification feature vector. The parameters of the feature vectors p_{t_j} and $p_{\bar{t}_j}$ are obtained by processing the prosodic feature vectors of the segmented FINAL part through prosodic model t_j and anti-prosodic model t_j, respectively. Therefore, p_{t_j} forms a 21-dimensional vector consisting of the following:

1. Coefficients representing the contour of the prosodic features of the segmented FINAL part. To be more specific, each prosodic feature in V_j is represented by a smooth curve formed by orthonormal expansion with discrete Legendre polynomial [6]. The number coefficients used in this polynomial is up to the third order. The zero-th order coefficient represents the mean of the prosodic feature contour and the other three coefficients represent its shape. Given a 4-dimensional prosodic feature vector, the number of parameters is 16.

2. Four parameters representing the state durations in number of frames normalized by the total frame duration of segmented FINAL part.

3. The prosodic HMM likelihood $L(V_j\mid t_j)$.

Similarly, $p_{\bar{t}_j}$ is formed by processing V_j using the anti-prosodic model t_j and computing the corresponding 21 parameters. For the whole word verification, the verification function can be decomposed into a series of FINAL part verification functions. Assuming independence, the whole word prosodic-phase verification function is defined as follows:

$$L_P(P;K)=\frac{1}{N}\sum_{n=1}^{N}CM(P_n;t_n) \qquad (9)$$

The outputs of the phonetic-phase and prosodic-phase verification functions are then combined as follows.

$$L(O,P;K)=L_s(O;K)+\beta L_P(P;K) \qquad (10)$$

where β is a weighting. Finally, the keyword rejection/acceptance decision is made by comparing $L(O,P;K)$ with a predefined threshold.

5. EXPERIMENTS

In order to assess the multi-keyword spotting system performance, a faculty name inquiry system has been implemented. In our system, 1275 faculty names and department names in National Cheng Kung University, Taiwan were selected as the keywords. A continuous telephone-speech database was employed to train the system. The database is part of the MAT (Mandarin Speech Across Taiwan) speech database and is composed of short spontaneous speech, number, syllables, words, and sentences. The total number of files is 20,386. This database was pronounced by 295 speakers (192 males, 103 females). All speech database were recorded via public telephone lines in 8 kHz using a Dialogic D/41D telephone card.

We also recorded 3,595 utterances for testing spoken by a different group of 37 speakers (21 males, 16 females) responding to requests for a person‘s name in our vocabulary. Three speech categories used in the testing database are single keyword (WORD), single keyword embedded in non-keyword speech (S-KW), and multiple keywords embedded in non-keyword speech (M-KW). Table 1 shows the results of the error rates with different approaches. From table 2, we can see that the fuzzy search algorithm can significantly improve the error rates for WORD, S-KW, and M-KW. Eventually, using the fuzzy search algorithm and prosodic verification, the error rate can be reduced from 12.85%, 25.72%, and 17.64% to 9.55%, 17.75%, and 11.29% for WORD, S-KW, and M-KW, respectively.

	WORD	S-KW	M-KW
Baseline	12.85	25.72	17.64
Fuzzy Search	11.32	23.26	15.85
Fuzzy Search + Prosodic Verification	9.55	17.75	11.29

Table 1. Error rate (%) for different approaches

6. CONCLUSION

In this paper, we have demonstrated some achievements in continuous Mandarin speech multi-keyword spotting and verification. In this system, 132 context-dependent subsyllables are used as the basic recognition units. A fuzzy search algorithm is proposed to extract keywords from syllable lattice that is constructed by Viterbi-Parallel Backtracking Algorithm. For a multiple keyword spotter, we construct the keyword relation table. According the keyword relation table, we utilize a weighting function for combining keywords. A keyword verification function combining phonetic-phase and prosodic-phase verification is also investigated. Experimental results show that the fuzzy search algorithm and utterance verification with prosodic information outperforms the baseline system without prosodic information.

7. ACKNOWLEDGEMENTS

The author would like to thank the National Science Council Republic of China, for financial support of this work, under Contract No. NSC87-2622-E-006-008.

8. REFERENCES

1 E. F. Huang and H. C. Wang, “An efficient algorithm for syllable hypothesization in continuous Mandarin speech recognition,” IEEE Trans. On Speech and Audio Processing, July 1994.

2 C-H. Wu, and Y-J Chen, “Use of Prosodic information for Mandarin Word Post-Recognition,” in Proc. TENCON‘97, Brisbane, Australia, 1997.

3 A. E. Rosenberg, C.H. Lee, B.H. Juang, and F.K. Soong, “The use of cohort normalized scores for speaker verification,” in Proc. 1992 Int. Conf. Spoken Language Processing, 1992, pp.599-602.

4 K. Fukunaga, Intorduction to Statistical Pattern Recognition. New York:Academic, 1972.

5 L.S. Lee, C.Y. Tseng and M. Ouh-Young, “The synthesis rules in a Chinese text-to-speech system,” IEEE Trans. Acoustic Speech, and Signal Processing, Vol. 37, No. 9, pp. 1309-1319, September 1989.

6 S.H. Chen and Y. R. Wang, “Vector quantization of pitch information in Mandarin speech,” IEEE Trans. Commun., 38, pp. 1317-1320, 1990.