SUBWORD UNITS
FOR A MANDARIN KEYWORD SPOTTING SYSTEM

Chi-Yan CHOY and Hong C. LEUNG
Department of Electronic Engineering
Chinese University of Hong Kong
Shatin, Hong Kong
E-mail: hcleung@ee.cuhk.edu.hk

ABSTRACT

This paper is concerned with the problem of phonetic modeling in a Mandarin keyword spotting system. The task is to detect 20 keywords from continuous speech in the Call Home corpus from the Linguistic Data Consortium (LDC). Different speech units are explored, including whole word, syllable, and demi-syllable (INITIAL and FINAL). In our speaker-independent HMM-based Mandarin keyword spotting experiments, the keyword spotter based on base-syllable keyword models has achieved the best performance. The best spotting accuracy achieved is 83.8% with 9.8 FA/KW/H. In the second part of our study, keyword spotting with different numbers of general filler models (389, 182, 37 and 1 fillers) has been performed in an effort to reduce computation time and increase flexibility.

1. INTRODUCTION

Keyword spotting is the task of detecting the occurrences of certain predefined *keywords* in an unconstrained speech signal. Without having any a-priori knowledge of the nature of the non-keyword speech or the syntactic structure of the utterances, keyword spotters have the ability to detect or to spot the useful information embedded in natural conversational speech, which may be full of pauses, hesitations, coughing, background noises, out-of-vocabulary speech, etc.

Speech recognition systems are always subject to unexpected speech input. These include information retrieval, messages classification, detection of command words, automated operator queries systems and so on. Keyword spotting techniques can be applied to these systems so that the users can speak in a more natural manner. In recent years, a great deal of research has been carried out and many practical systems have been constructed. Most of the keyword spotters developed were hidden Markov model (HMM) based, continuous speech recognition systems [1,2,3,4]. In these systems, keyword models were trained to detect the keywords while the non-keyword intervals were represented by *garbage* or *filler* models.

Modeling of keyword and filler models is of the utmost importance in keyword spotting. In this paper, we will report a systematic study on how the performance of a speaker-independent Mandarin keyword spotter is affected by the amount of contextual information utilized in the acoustic modeling of the keywords: whole word, base syllable, tonal syllable, context-independent initial and final, context-dependent initial and context-independent final. We will also describe our investigation into the use of different general filler models to represent the non-keyword speech in an effort to reduce computation time and increase domain flexibility.

2. PHONETIC UNITS FOR MANDARIN

In Mandarin, all the characters are monosyllabic. There are basically five lexical tones and each syllable is assigned to different tones. The 1345 tonal syllables can be considered as the combinations of the 408 toneless base syllables and the five different tones. Conventionally, each Mandarin syllable is decomposed into an INITIAL and a FINAL, in which INITIAL means the initial consonant of the syllable while FINAL means the vowel or diphthong part, but including an optional medial or nasal ending. There are altogether 21 INITIALs (excluding the null INITIAL) and 37 FINALs [5].

3. RECOGNITION SYSTEM DESCRIPTION

3.1. Overall Architecture and Recognition Network for the keyword Spotters

The overall architecture of a keyword-spotter is shown in Figure 1 and the recognition network is shown in Figure 2. M keywords and N filler models are placed in parallel. The input speech utterance will be decoded into a sequence of keywords and non-keywords and this configuration allows any combination of keywords and fillers in any utterance.

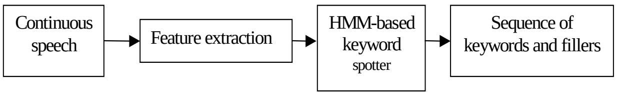

Figure 1: Overall architecture for an HMM-based keyword spotting system.

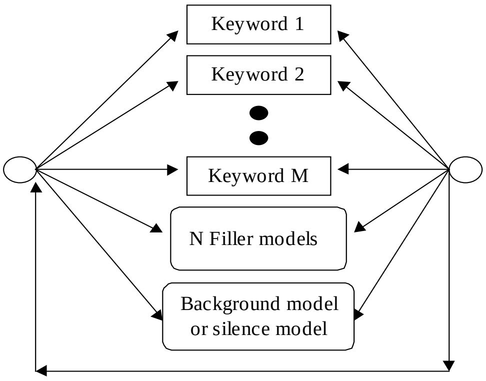

Figure 2: The HMM-based keyword spotting system

3.2. Signal Representation and Features

The feature vector of the speech signal is composed of 12 mel frequency cepstral coefficients (MFCCs), one normalized energy, 12 delta MFCCs and delta energy. The speech waveform window size is 25.6 ms and the analysis is performed once every 10 ms. Energy normalization is used to reduce the effect of short-term energy variations and cepstral mean normalization [6] is performed for channel compensation.

3.3. Keyword Models

Different acoustic units were used for the modeling of keywords in the keyword-spotters. They were whole word, base syllable, tonal syllable, context-independent INITIAL and FINAL, context-dependent INITIAL and context-independent FINAL. Each keyword model was represented by a continuous density HMM (CDHMM). The HMM states were arranged in a left-to-right topology without state skipping. Embedded Baum-Welch re-esitimation algorithm was used for training the keyword models [6]. The details of each unit are shown in Table 1.

Acoustic Units	# states	# Gaussian mixtures per state
Whole-word	15	8
Base syllable	8	8
Tonal syllable	8	8
Context-independent INITIAL	3	8
Context-dependent INITIAL	3	8
Context-independent FINAL	5	8

Table 1: Details of different acoustic units

3.4. Filler Models

All filler models were base syllable HMMs with 5 states and 8 Gaussian mixtures per state. Similar to the keyword models, the HMM states were arranged in a left-to-right topology without state skipping. Embedded Baum-Welch re-esitmation algorithm was used for training the filler models. All the speech data, rather than just the non-keyword speech, was used to train the filler models. In this way, the filler models are vocabulary and task independent and do not need retraining if a new keyword is added to the system. We have found that using smaller sub-word units (e.g. INITIALs and FINALs) increases the insertion errors. Therefore base syllable fillers were used in the experiments for the representation of extraneous speech.

3.5. Language Modeling and Search

It is unlikely that the available training data has the same word sequence statistics as the testing data. Therefore null grammar, which allows a completely unconstrained syntax, was used in the experiments. Any keyword or filler can follow any other with equal probability. Null grammar is certainly more applicable for keyword spotting than general continuous speech recognition, as some keyword occurrences are comparative rare (e.g. only 9) in the training data.

Viterbi algorithm, with the pronunciation network and pronunciation dictionary, was used for decoding. No word-end pruning was set in the searching. The output was a continuous stream of fillers and keywords.

4. EXPERIMENTS

4.1. Tasks

In the first part of our study, we investigated how the performance of a speaker-independent Mandarin keyword spotter was affected by the amount of contextual information utilized in the acoustic modeling of the keywords: whole word, base syllable, tonal syllable, context-independent INITIALs and FINALs, context-dependent INITIALs and context-independent FINALs. We developed five different keyword spotters and compared their performance results.

For most typical systems, the non-keyword vocabulary is very large but the amount of training data available is limited. Thus more general filler model(s) should be used to model non-keyword speech. In general, most of the non-keyword utterances in the testing speech will not have been contained in the training set. Therefore in the second part of our study, we investigated the use of different general filler models in an effort to reduce computation time.

4.2. Speech Database

The Mandarin Call Home speech corpus from LDC was used in our study. The Call Home corpus [8] is a multilingual corpus of telephone conversations for use in spoken language research. The corpus was collected through T1 telephone lines. Each recording was 10 to 30 minutes long and the transcripts were time-stamped by speaker turn for alignment with the speech signal.

There were no specified topics for talkers to follow in the conversations. The spontaneous speech was embedded with lots of laugh, disfluencies, pauses, word slurring, background noises, which makes the spotting of vocabulary words difficult. In the testing set, 82.5% of the speech segments are extraneous speech. The sets for training and testing are shown in Table 2. Twenty bi-character keywords were chosen from the speech corpus for our experiments. Table 3 shows the twenty bi-character keywords.

	# keywords	# utterances	# speakers	# hours
Training set	2063	19,937	160	9.8
Test set	561	1,030	40	0.53

Table 2: Training and test sets in the experiments

北京	美國	中國	打工	大學
地方	電話	房子	工作	公司
國內	回來	機會	家裏	塊錢
特別	學校	晚上	地址	分鐘

Table 3: The list of the 20 bi-character keyword.

4.3. Performance Measures

Each putative keyword *w* is compared with the reference transcriptions. If the start and end times of *w* lie on either side of the mid-point of an identical label in the reference, then the *w* represents a *hit*, otherwise it is a *false-alarm* (FA) [6]. The keyword detection rate P_d is defined as the number of correct putative hits divided by the number of keyword occurrences in the original speech utterance. The false alarm rate (FAR) is defined as the number of false alarms per hour of speech normalized by the frequency of the keyword (FA/KW/H).

4.4. Details of Different Word-spotters

In the first part of our study, we developed 5 different HMM-based speaker-independent keyword spotters for Mandarin keyword spotting, and compared their performance results. Each keyword spotter was associated with the use a distinct set of speech units (see Table 4) for the keyword models. In order to determine the optimal phonetic units for the keyword models, the filler models used were the same for each spotter in this part of the experiment. All the 389 base syllables and one background model, which were available in the training set, were used as the filler models in each keyword spotter.

Details of the five keyword spotters are:

A. Keyword spotter using word-based keyword models.
B. Keyword spotter using base syllable keyword models.
C. Keyword spotter using tonal syllable keyword models.
D. Keyword spotter using context-independent INITIAL and context-independent FINAL models.
E. Keyword spotter using context-dependent INITIAL and context-independent FINAL models.

The acoustic units used in each keyword spotters, their frequencies for training and the average number of tokens per model are shown in Table 4.

Keyword Spotter	Phonetic units for keyword models	Avg. # of tokens/model
A	20 whole word models	103
B	30 base-syllable models	914
C	33 tonal-syllable models	485
D	17 context-independent INITIALs + 19 context-independent FINALs	7424
E	29 context-dependent INITIALs + 19 context-independent FINALs	3194
Filler models	389 base syllable filler models	483

Table 4: Phonetic units used in each keyword spotter and their frequencies in the training set of the speech data.

4.5. General Filler Models

In the first part of our study, all the 389 base syllables were used as the filler models in each keyword spotter. In the second part of our study, we chose the most effective units for modeling the keywords from the first part of the study and designed three sets of general fillers consisting of 182, 37 and 1 acoustic models for background representation (see Table 5). The advantage of using a smaller set of filler models is that less computation time is required for the recognition.

Keyword spotter	Phonetic units for keyword models	Phonetic units for filler models
B	30 base-syllable keywords	389 base-syllable fillers
B.182	30 base-syllable keywords	182 base-syllable fillers
B.37	30 base-syllable keywords	37 base-syllable fillers
B.1	30 base-syllable keywords	1 base-syllable filler

Table 5: Keyword spotters with different numbers of filler models

Details of the three set of general fillers:

A. The set consisting of 182 fillers:
Based on the manner of articulation, the 21 INITIALs can be divided into 5 broad phonetic classes. They are "*stops, affricates, fricatives, laterals and nasals*" (see Table 6). In Mandarin, a syllable can be decomposed into an optional INITIAL and FINAL. Syllables having the same FINAL and having INITIALs from the same broad phonetic class, for example /pian/, /tian/, /dian/, … , were combined together and a general filler was trained for this set of syllables. In this way, 182

general fillers were obtained for background representation.

B. The set consisting of 37 fillers:
Syllables having the same FINAL, for example /pian/, /jian/, /lian/, /tian/, … , all grouped into the same filler model.

C. The set consisting of 1 filler:
All the base syllables were combined together and a single filler model was trained as a universal extraneous speech model.

Manner of Articulation	INITIALs
Stops	p,t,k,b,d,g
Affricates	z,zh,j,c,ch,g
Fricatives	f,s,sh,x,h
Nasals	m,n
Laterals	l,r

Table 6: Broad phonetic classes for the INITIALs.

5. EXPERIMENTAL RESULTS

For the first part of our study, the performance results of the five keyword spotters at different false alarm rate are shown in Figure 3. It is found that Spotter B**,** which uses base syllable as the keyword models results in the best performance. The best spotting accuracy achieved is 83.8% with 9.8 FA/KW/H. Even at 0.2 FA/KW/H, 37.1% of the keywords can still be detected.

The performance of keyword spotters with different numbers of general filler models is shown in Figure 4. The performance of Spotter B acts as a baseline result. As the filler models become fewer and more general, the performance of Spotter B.182 (with 182 fillers) and Spotter B.37 (with 37 fillers), declines but only to a slight extent. At 5 FA/KW/H or more, the detection accuracy of Spotter B.182 and Spotter B.37 drops approximately only 2% and 5% respectively. However, the total computation time decreases by approximately a factor of 1.7 and 5 respectively as shown in Figure 5. Spotter B.1, with 1 general filler, shows the lowest performance among all. This indicates that the single general filler gives poor representation of the non-keyword speech.

6. CONCLUSIONS

In this paper, we presented a systematic performance comparison among various levels of acoustic modeling for speaker-independent Mandarin keyword spotters. In our experiments, base-syllable-keyword models gave the best performance of 83.8% detection rate with 9.8 FA/KW/H. More explicit modeling of the non-keywords (389 fillers) results in higher performance but requires more computation. With the use of general filler models (182 fillers and 37 fillers), there is only a little degradation of the detection accuracy at 5 FA/KW/H or more. However, the average computation time drops significantly.

7. REFERENCES

[1] J. G. Wilpon, L. R. Rabiner, Chin-Hui Lee, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans. on ASSP*, Vol. 38, No.11, pp. 1870-1878, Nov. 1990.

[2] J. G. Wilpon, L. G. Miller, P. Modi, "Improvements and Applications for Key Word Recognition Using Hidden Markov Modeling Techniques", *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Processing*, Vol. 1, pp. 309-312, May. 1991.

[3] R. C. Rose, D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System", *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Processing*, Vol. 1, pp. 129-132, April. 1990.

[4] J. R. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus, M. Siu, "Phonetic Training and Language Modeling for Word Spotting", *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Processing*, Vol. 2, pp. 459-462, 1993.

[5] Lin-Shan Lee, "Voice Dictation of Mandarin Chinese-Computer Data Entry Without a Keyboard via Speech Recognition", *IEEE Signal Processing Magazine*, pp. 63-101, July. 1997.

[6] Steve Young, V. Valtchev, J. Odell, D. Ollason, P. Woodland, "The HTK Book (for HTK version 2.1)", Cambridge University, 1997.

[7] A. S. Manos, V. Zue, "A Segment-Based Wordspotter Using Phonetic Filler Models", *IEEE Trans. on ASSP*, Vol. 2, pp. 899-902, 1997.

[8] R. Agarwal, B. Wheatley, Y. Muthusamy, T. Staples, "Diagnostic Profiling for Speech Technology Development: Call Home Analysis", *Proceedings of Speech Research Symposium*, pp. 131-137, June. 1995.

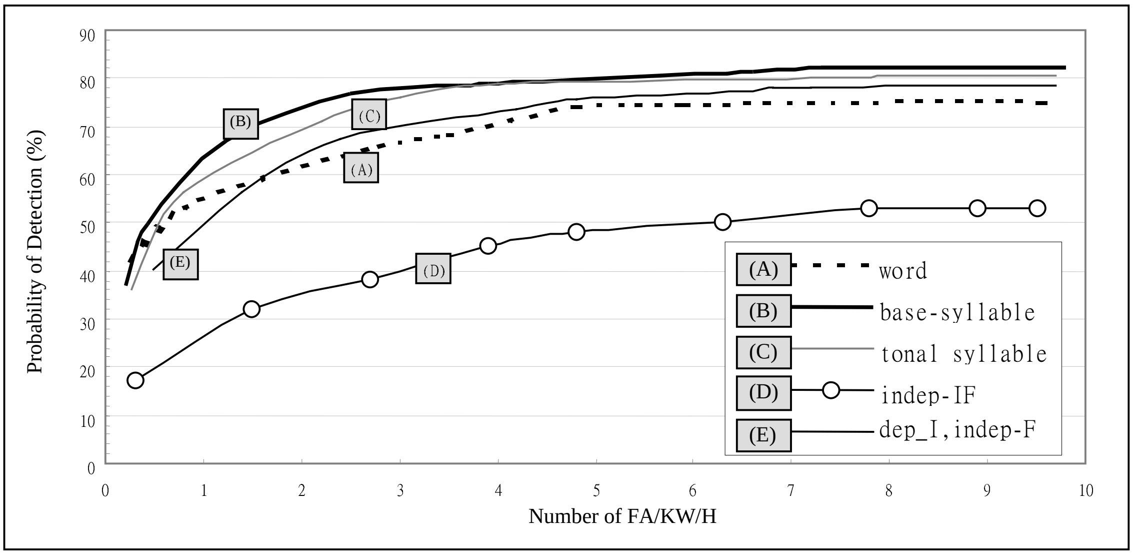

Figure 3: The performance of the five keyword spotters

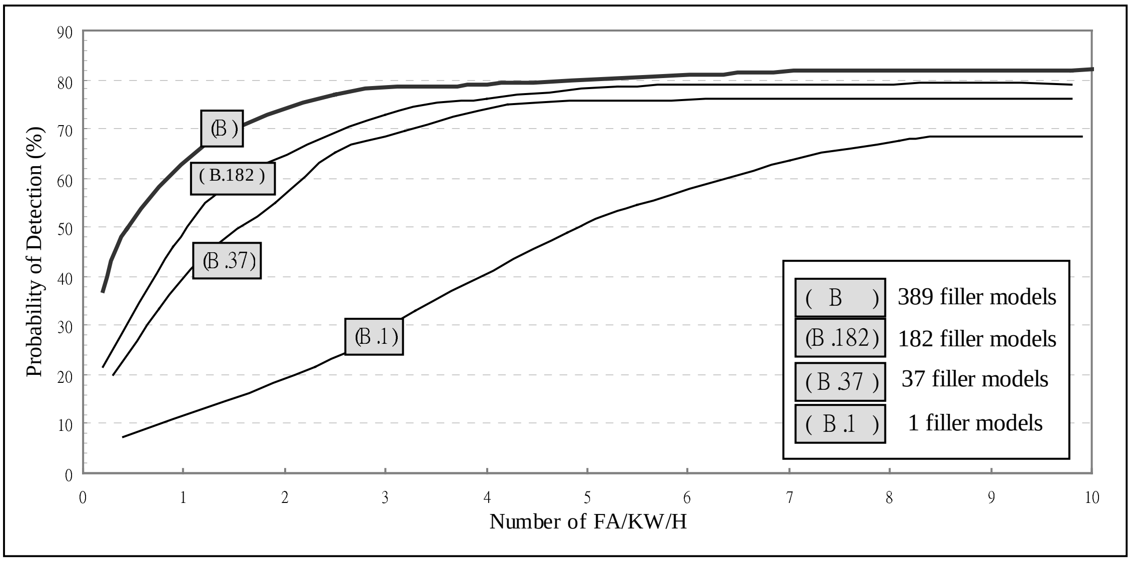

Figure 4: The performance of the keyword spotters with different number of general fillers

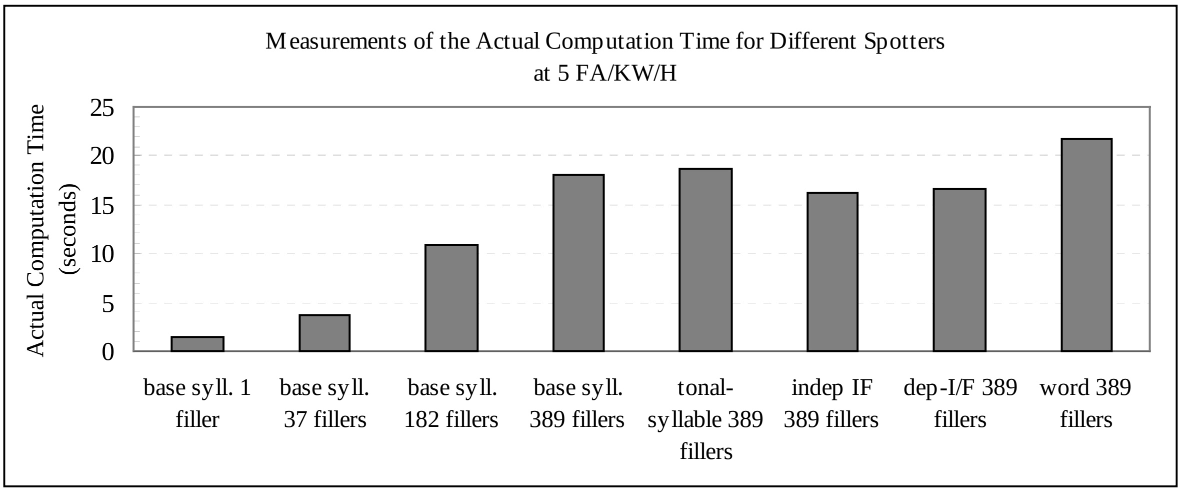

Figure 5: The average computation time of the keyword spotters.