Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

Mandarin Speech Recognition Using Keyword Spotting in a Real -Time System

Kuo-Chang Huang, Shin-Lun Tung and Yau-Tarng Juang
Department of Electrical Engineering
National Central University
Chung-Li, Taiwan, 32054
R.O.C

Abstract

In this paper, we applied the keyword spotting technique to Mandarin speech recognition system based on the semi-continuous Hidden Markov Models (SCHMMS) for automatic interpretation of phone operator. The technique separately deals with keywords and nonkeywords models in training phase. Ir the testing process, an HMM-based connect word recognition system is used to find the best sequence of keyword, nonkeyword for matching the actual input speech. Finally, we tasked with 35 keywords (names) in a telephone network system and developed an automatic phone operator system for speaker-independent case. The real-time recognizer was implemented on a PC-486 computer enhanced by only one digital signal processing board on which a TMS-320C30 chip operates as CPU. The best recognition accuracy for 35 keywords (names) is 95.72 % in the speaker-independent case.

I. Introduction

The problem of Keyword Spotting (KWS) is usually addressed by using template based [Bridle, 1973; Higgins & Wohlford, 1985] or Hidden Markov Model (HMM) based [Wilpon et al., 1990; Rose & Paul, 1990; Wilpon et al. 1991; Hofstettor & Rose, 1992] Continuous Speech Recognition (CSR) algorithms. In these cases, garbage templates or HMM models (also referred to as filler models) were used to match nonkeyword speech and non-speech sounds. When garbage models are defined, the input speech is then recognized by using a standard CSR algorithm in terms of an unconstrained sequence of garbage and keyword models. Furthermore, simple syntactical constrains are usually applied to the CSR system to reduce the error rate. As to research into speaker-independent speech recognition systems that perform well and are robust over telephone line has been carried out many years [1] [2]. The use of such technology will greatly enhance current telephone network-based services, for examples, automating phone operator or verifying credit card transactions. In addition, it can be used to create a wide variety of new services from telemarketing (voice response and auto-text services) to office automation (a voice calendar system).

Most speech recognition systems that exist today are constrained in that the speech to be recognized must consist only of words from a pre-defined vocabulary. For speech recognition applications in the telephone network, it is naive to assume that user will adhere strictly to some protocols, for example, when a user is asked to say only a name (楊麗花 (yang li hua)), they may say that, "Ah.... 你好,請問楊麗花在不在? "(Ah ...ni hao, qing wen yang li hua zai bu zai ?). It is obvious that only the name is important in that sentence, so keyword spotting technology is used to solve this problem.

In this paper, we will carry out the technology of keyword spotting and build a real-time Mandarin speech recognition system in continuous speech. The difficulty of keyword spotting is that we cannot know the nonkeywords that the user might said, so we use the HMM-based connect word models to present the relationship of keywords and nonkeywords. This method not only can reduce the memory but also speed up the processing time. According to the above concept, the continuous speech is divided into keyword and nonkeyword (garbage) and their models can be estimated, respectively. In the training phase, the keyword and nonkeyword models are established independently. And in the testing phase, they are connected into one recognition model. Then by using Viterbi algorithm, the best sequence of arbitrary continuous speech is determined. Finally, the keyword is spotted from the continuous speech. In our experiments, we tasked with a 35 Mandarin words (names) in a telephone network and developed an automatic recognition system of the phone operator for real-time speaker-independent case. The system is implemented on a PC-486 computer enhanced by only one digital signal processing board on which a TMS-320C30 chip operates as CPU. The best recognition accuracy is 95.72% for speaker-independent case.

In this paper, section II describes the overview of system. In section III, we describe the HMM-based keyword spotting algorithm for recognizing keywords in the continuous speech. In section IV, we present experimental results from a series of recognition experiments. Finally, conclusions and future works are provided in section V.

II System Overview

In our developed approach, the entire background environment, including silence, transmission noise and most importantly, nonkeyword speech is modeled. A given input is represented as an unconstrained sequence of background and nonkeyword speech followed by keywords and followed by another unconstrained sequence of background and nonkeyword speech. A grammar driven continuous word recognition system is then used to determine the best sequence of nonkeyword, background, and keyword speech signals. Given this structure for the recognition system, the garbage models match the nonkeyword speech and the trained vocabulary word models (keyword models) match the actual keyword that was spoken. Fig 1. shows a block diagram of the overall recognition system. The key elements of the system are described in the following.

A. LPC and Cepstral Analysis

The speech is sampled at 10kHz, pre-emphasized using the transfer function $1-0.95z^{-1}$. A Hamming window with a width of 45msec is applied every 15msec, 12th-order LPC analysis is implemented using the autocorrelation method. Next, a set of 14 LPC Cepstrum coefficients is computed from the LPC coefficients. When speech signals are obtained from the telephone network, the Cepstrum will be distorted by channel and handset noises, so the transitional Cepstrum is added to be the feature for the purpose of resistance of noise.

B. Generating Word Reference Models

In order to generate one or more word models from a training data set of the speech, a segmental k-means training algorithm is used. This word building algorithm (i.e., an estimation procedure for determining the parameters of the HMM's) is iterated for each model until convergence (i.e., until the difference in likelihood scores in consecutive iterations is sufficiently small).

This algorithm, based on the likelihood obtained from the current set of HMM's, separates from the set of training tokens those tokens whose likelihood scores fall below some fixed or relative threshold. That is, we separate all the tokens with poor likelihood scores and create a new model out of these so-call outlier tokens. Once the tokens have been clustered, the segmental k-means training algorithm is again used to give a (local optimal) set of parameters for each of the models. Further details of this algorithm can be found in [3].

C. Model Alignment Procedure

Because in Chinese interrogation system, people usually say one type of the syntax described by the graph in Fig. 2 to quaere. In our recognition model of the automatic phone operator system, the model alignment is based on the above concept. In this case, the syntax is described by leading silences (denoted by silence 0) and garbage models(denoted by garbage 0X), followed by the set of possible keywords(denoted by keyword 0X) and finally followed again by trailing silence (denoted by silence 1 which identify to the leading silence model) and garbage models (denoted by garbage 1X). As for the training process of keyword and garbage models, it will be described in the next section.

III Model Training and Keyword Searching

A. Training

The wordspotting technique uses an HMM network to model user-defined keywords in context of arbitrary continuous speech. Training the HMM network consists of two stages: 1) a static stage, in which unsupervised training is performed for the talker; and 2) a dynamic stage, in which keyword training is performed as the system is used. During the static training, an arbitrary segment of the talker's speech is used to learn the statistics for a pool of Gaussian distributions. The dynamic training stage uses a single repetition of the keyword by the users, as well as the Gaussians obtained in static training to create an HMM for the keyword. In keyword training, each keyword was used to train their correlative HMM-based keyword models, respectively. In nonkeyword training, three methods were used to train the nonkeyword models as follows:

Method 1: We choose a set of 5 lexicons of pre-nonkeywords and post-nonkeywords from the offscreen dialog in telephone, e.g., "請問一下(qing wen yi xia)", "麻煩你找(ma fan ni zhao)", "好嗎(hao ma)", "在不在(zai bu zai),"...etc. Total 10 nonkeyword HMM models were obtained in this method.

Method 2: We classify the data of pre-nonkeywords and post-nonkeywords from the first method with segmental k-mean's algorithm [4]. After carefully classification, a set of 5 lexicons including 3 lexicons of pre-nonkeywords and 2 lexicons of post-nonkeywords will be train to establish 5 nonkeyword models.

Method 3: We build an HMM model of pre-nonkeywords and post-nonkeywords to model the possible nonkeywords, respectively.

With the above descriptions, we hope that the

nonkeyword speech that does not appear in training data can be assimilated by our nonkeyword models. The training flow chart is described as Fig.3.

B. Viterbi Searching

In the Viterbi searching, a forward pass is performed through an entire speech, followed by a backward pass. The result is a series of intervals corresponding to keyword and nonkeyword speech. In order to perform wordspotting on the speech, Rose and Hofsetter [5] used the partial backtracking [6] to obtain results prior to the end of an utterance. In our method, backtracking is performed to locate the intervals to different states where all paths agree at a given time, thereby obtaining the optimal path. When the optimal path is obtained, the keyword is detected from the entire speech.

IV. Experiments

Initially, speech signals are sampled at a 10kHz sampling rate and a Hamming window is used. A vector of 14 cepstral coefficients derived from 12-order LPC parameters is computed over 45ms frame shifted every 15ms. The overall feature vector consists of 14th-orded cepstrum and transitional cepstrum. In training phase, to train the nonkeyword models, a database(DB1) of 14 sets of 35 sentences uttered by 14 speakers (7 males and 7 females) was used. To train the keyword models, we use a database(DB2) of names (a set of 35 famous men in Chinese) uttered by 20 speakers (10 males and 10 females) of which each one repeats 3 times. In testing phase, a database(DB3) of 35 sentences including keywords and nonkeywords uttered by 14 speakers (out of in training phase) was used.

A. The Experimental Results of Telephone Network

Speaker-dependent case:

In the first experiment under telephone network, we measure the word spotting performance by 3 different techniques described in section III. The results of this experiment are shown in Table I. From Table I, it is seen that **Method 2** has the highest performance to model the nonkeywords. Therefore, **Method 2** is used to be the garbage models in the following experiments. In the second experiment that it runs in telephone network environment, we added transitional cepstrum as the new features in our recognition system to improve the recognition accuracy. The experimental results show that word recognition rate in continuous speech using SCHMM can improve over 5% by employing transitional cepstrum. This experiment proves that the transitional cepstrum can resist the noisy disturbance from telephone line and improve the recognition rate. So it is important to select the appropriate features to improve the system [7] [8].

Speaker-independent case:

In the third experiment, we analyzed about how the mixture number of state in the garbage HMM models will affect the recognition results. Table II shows this experiment results of changing the number from 1 to 8. The conception of mixture number is to utilize the N *pdf* to represent the N clusters. With more training data, it needs more clusters to classify the training data. From the Table, it is seen that the accuracy rate performs well when the mixture number of state is greater than 4. Consider the memory and speed, it is acceptable that the mixture number of garbage state is set to 5.

Besides, we examined the effects of varying the number of mixtures used for keyword models and nonkeyword models on the word recognition accuracy. Table II shows the experimental results. It is noted that when the mixture number of the keyword and nonkeyword models are selected to be five, the recognition accuracy achieves 95.72%. Finally, we used the different database consisting of different nonkeywords to test the accuracy rate in our recognizer. From the experimental results shown in Table III, we know that the recognition algorithm of this system has large tolerance of the nonkeywords.

B. Real - Time Implementation of the Speaker-Independent System

A PC 486 computer is used as the processor and control center for our real-time Mandarin speech recognition system. A DSP board is connected to the PC for speech signal preprocessing and likelihood score computing. The input speech signals from the telephone line are sampled by a 16-bit A/D converter at a rate of 10KHz. The end point detection for Chinese speech is performed by preprocessing procedure. Once the ending point of the speech is detected, all of the features of the speech have been obtained, then the process of searching and matching with reference models (HMMs) that are loaded to the DSP board will be started. The recognition procedure is processed at a DSP board and a list of candidates for recognition results is obtained and transferred to PC. And the processing center (PC) will show the results and provide a convenient man-machine interface. The framework diagram of the system is described as Fig. 4 and the flow chart of recognition system is described as Fig. 5.

V. Conclusions and Future Work

Automatic speech recognition can be used to greatly enhance current telephone network based services and to create a wide variety of new services. One example,

presented here, is automating phone operator services. Others include catalog order entry, credit card verification and repertory dialing. In this paper, we have presented an algorithm based on Hidden Markov Model technology, which shows capable of accuracy recognizing a predefined set of vocabulary item(keyword) spoken in the context of fluent unconstrained speech. From the experimental results, it is seen that when the transitional cepstrum is added, the recognition accuracy rises about 5% for speaker-dependent case. It means that the dynamic features can eliminate the noise from the speech signals and raise the accuracy rate. Moreover, it is shown that, for a specified vocabulary of 35 keywords(names) used in automatic phone operator services for speaker-independent case, the recognizer could correctly recognize 95.72% of the spoken keywords occurred in fluent speech spoken speech over the long-distance telephone network.

Furthermore by creating large number of garbage models, it will obtain good performance on recognition rate of the vocabulary words in unconstrained speech. But it may reduce the recognition speed and waste the memory space. So the goal of this paper is also to find the ways to reduce the number of garbage models and still maintain performance close to that of the system with a large number of garbage models. In the future work, we wish to enhance the capability of nonkeyword recognition (i.e., when the speech does not consist of the keyword that is in training database, the system will show out-of-keyword and not execute the user's order).

Acknowledgment

This work is supported by the National Science Council of the Republic of China under the contract NSC85-2213-E008-025.

References

[1] J. G. Wilpon, D. M. DeMarco and R. P. Mikkilineni, "Isolated word recognition over the DDD telephone network-results of two extensive field studies," *Proc. IEEE Int. Conf. Acous. Speech. And Sig. Processing.* **1S.1.10**, pp. 55-57, New York City, New York, April 1988.

[2] L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition -- Theory and Sleeted applications," *IEEE Trans. On Comm.*, vol. Com-29, No.5, pp.621-659, May 1981.

[3] L. R. Rabiner, C. H. Lee, B. H. Juang, and J. G. Wilpon, "HMM clustering for connected word recognition system" in *Proc. ICASSP '89* (Glasgow, Scotland), May 1989, PP. 405-408.

[4] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A segmental k-means training procedure for connected word recognition based on whole word reference patterns," *AT&T. J.*, vol.65, no.3, pp.21-31, May 1986.

[5] R. C. Rose, E. M. Hofstetter, "Techniques for Robust Word Spotting in Continuous Speech Messages." *Proc.* Of Eurospeech, Genova, Italy, September 1991, PP. 1183-1186.

[6] F. Brown, J. C Spohrer, P. H. Hochschild, J. K. Baker, "Partial Traceback and Dynamic Programming." *Proc. Of the Int. Conf. on Acoustics, Speech and Signal Processing,* Paris, France, May 1982, PP. 1629-1632.

[7] Furui." Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing.* Vol. Assp-34. Pp.52-59, Feb 1986.

[8] F. Lee and H. W. Hon, "Speaker-independent phoneme recognition using Hidden Markov models," Dep. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, Pa, Tech. Rep. CMU-CS-88-121, Apr. 1988.

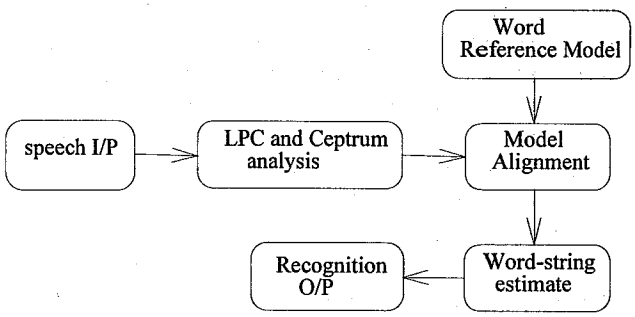

Fig.1 Block diagram of overall recognition system

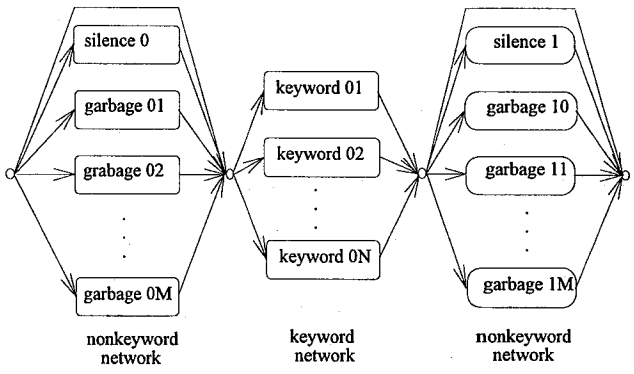

Fig.2 The syntactical diagram of keyword spotting

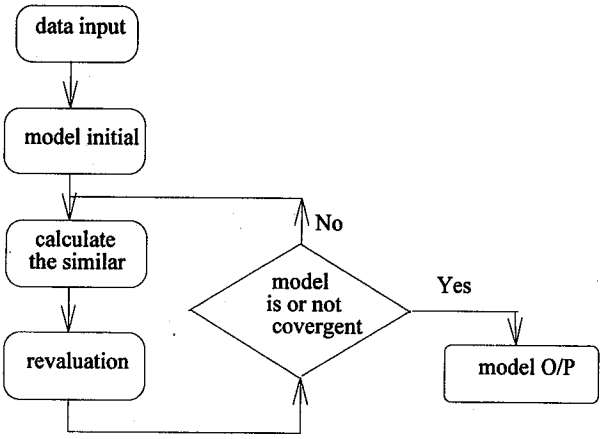

Fig.3 The trainig flow chat of keyword and nonkeyword

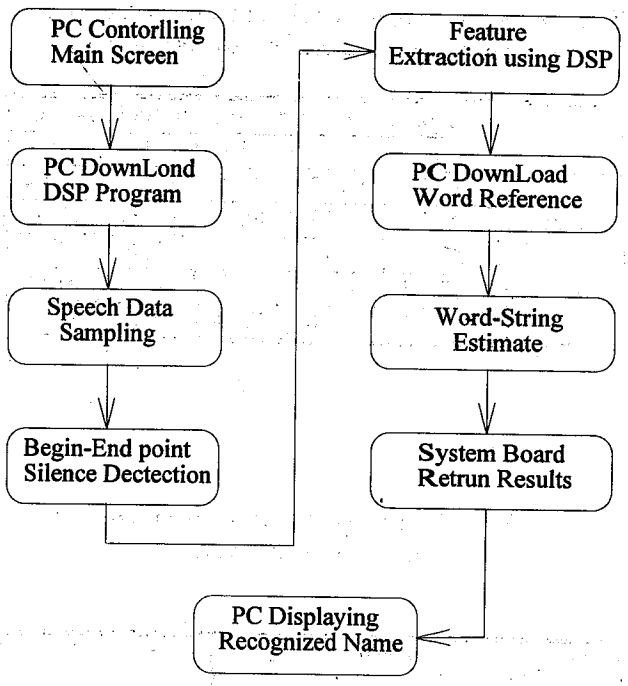

Fig.5 The flow chart of the recognition system

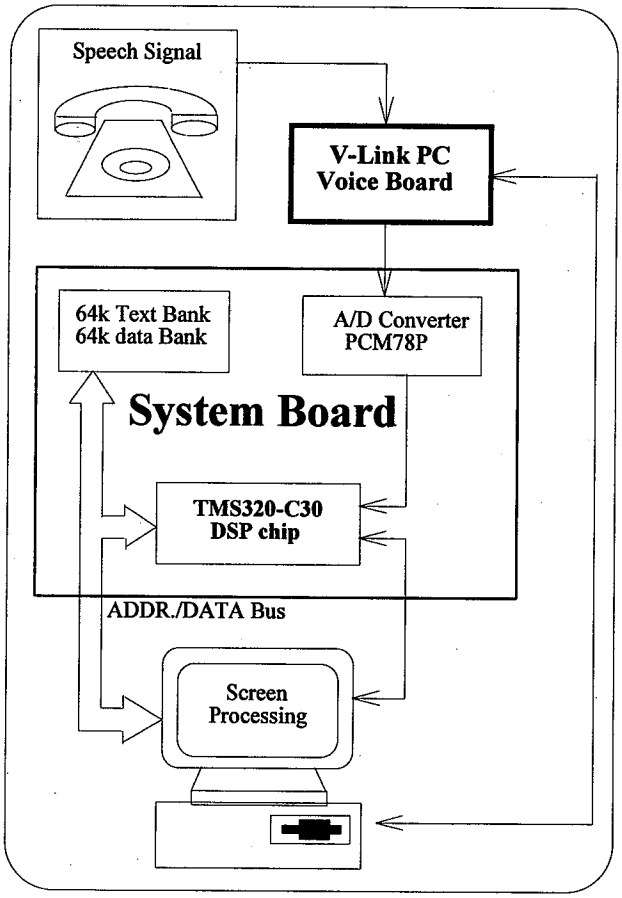

Fig. 4 The framework diagram of the dialogic system

Table I The results of using three methods to establish the nonkeyword models

Method of nonkeyword	Method 1	Method 2	Method 3
	90.86%	92.04%	89.71%

Table II The recognition results of changing the mixture number

	The number of mixture							
	1	2	3	4	5	6	7	8
	86.12	91.22	89.80	94.08	95.72	93.27	96.53	95.92

Table III The recognition results of changing the mixture number of keyword and nonkeyword models

		The number of mixture for nonkeyword				
		1	2	3	4	5
No. Mixture for keyword	1	86.12	86.33	84.49	87.96	86.94
	2	86.73	91.22	89.90	93.67	92.65
	3	87.96	91.43	89.90	93.06	93.06
	4	87.55	90.41	86.53	94.08	95.51
	5	90.41	93.06	92.24	91.22	95.72

Table III The recognition rate of using the different database

No of mixture	2	3	4
database 1	89.05	93.09	94.05
database 2	89.76	90.95	94.05