J. of Computer Science and Technology, 14(1): 18-26, Jan., 1999

HARKMAN - A VOCABULARY-INDEPENDENT KEYWORD SPOTTER FOR SPONTANEOUS CHINESE SPEECH

Zheng Fang(郑方), *Xu Mingxing*(徐明星), *Mou Xiaolong*(牟晓隆), *Wu Jian*(武健),
Wu Wenhu(吴文虎), *Fang Ditang*(方棣棠)

Speech Laboratory, Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, P. R. China
Tel.: +86-10-62784141, FAX: +86-10-62771138
fzheng@sp.cs.tsinghua.edu.cn

ABSTRACT

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

Keywords: keyword spotting, keyword spotter, vocabulary independent, acoustic modeling, continuous speech recognition

1. INTRODUCTION

Keyword spotting (KWS) has a wide range of application, such as message classification, topic identification, and speech content addressed applications. Moreover, the technologies in KWS are also useful for a spontaneous dictating or spoken language understanding system, where many meaningless words may be mixed in the sentences and should be deleted.

Unlike in the continuous speech recognition (CSR) system, a keyword-spotter (KWS) needs not recognize all the uttered words, only the specified keywords are concerned. So the structure of a KWS system is different from that of a CSR system. The basic structure of KWS can be filler-based whole-word spotting or phoneme-based vocabulary-independent (VIND) spotting. In the first structure, the figure-of-merit (FOM) [1] value is often higher, but the system is less flexible, so the second type is preferable for the VIND tasks.

Solutions for CSR can be applied to KWS, but there are still many special technologies dedicated to KWS, this is what will be focused on in this paper.

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 in each model 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 search steps and strategies adopted in *HarkMan* are described in details. In Section 5, our solutions to the robustness issue are given, the background noise, accent, gender, context, channel and so on are covered briefly. In Section 6, the rejection methods are given. In Section 7 the experimental results of *HarkMan* that adopted the described technologies are given.

2. ACOUSTIC REPRESENTATION

2.1 Database Description

The training and testing data were taken from a real-world telephone network. Speakers were asked to talk to each other over the local, DDD or IDD telephone network in a spontaneous manner, where the Signal-to-Noise Ratio (SNR) was about 25dB.

Speech signals were digitized at 8kHz sampling rate, they were compressed into A-law codes because of the specified hardware. The KWS should first expand the A-law codes to 13-bit linear PCMs. Expanded linear signals were pre-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 [2] 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 sets of coefficients was constructed as a vector in a *D*-dimensional Euclidean space.

The 4GB real-world telephone database consists of speech data uttered by 200 people. In this database, utterances were spoken very fast, the average Chinese syllable length was 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 acoustic modeling method used here is based on Center-Distance Continuous Probability Model (CDCPM) [3] 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 the observation feature space is described by Center-Distance Normal (CDN) [3] distributions instead of normal distributions. The CDN distribution is used to describe the distance between a normal random vector and its statistical mean vector. 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) $\vec{\mu}_{xnm} = (\mu_{xnm}^{(d)})$: 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; (6) g_{nm}: the weight of the *m*'th density component in the *n*'th state. Here $1 \le n \le N$, $1 \le m \le M$, $1 \le d \le D$.

Normally, mixed Gaussian densities (MGD) are used to describe the feature space. Similarly the mixed CDN densities can be used based on CDN distributions. Our experiments show that [4] mixed CDN densities are not so good as the Nearest Neighbor (NN) based Embedded Multi-Model (EMM) scheme [3] when scoring the feature vectors. An 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 practically proved robust for gender-dependent, accent-dependent, and context-dependent models and so on.

2.3 The selection of SR Units

Choosing the speech recognition units (SRUs) is a very important issue in CSR. SRUs should have the following characteristics: (1) they are flexible to make up any grammatical unit such as words or phrases, and (2) their corresponding acoustic models are robust. Considering the first factor results in choosing as small units as possible while considering the second factor results in choosing as big units as possible.

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. The Chinese syllables as the SR units can meet the previously stated factors so the Chinese syllables are the best choice. This has been proved in our previous experiment [5].

2.4 The selection of the State Number and Density Number

In order to determine suitable *N* value, i.e., the number of states in the CDCPM, and *M* value, i.e., 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 [6]. Experiments also showed that given a fixed maximum *M* value, using suitable *M* values individually for different SRUs performed better than always using the maximal *M* value [7]. Taking the performance and the complexities into consideration, *N*=6 and *M*≤16 were chosen. Experimental results were satisfying, the syllable accuracy was 80.5% and the accuracy of top 10 candidates was over 95%.

3. KEYWORD SPOTTING NETWORK

3.1 Basic network

In general, the KWS systems are filler-based, the basic network is illustrated in Fig. 1 [8, 1], 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 between keywords and/or fillers, where $w_{Kp} (1 \le p \le P)$ is the keyword transition weights and $w_{Fq} (1 \le q \le Q)$ the filler transition weights. If the probability of keyword detection P_d is to be absolutely guaranteed, the keyword weights are often bigger than those of the fillers are.

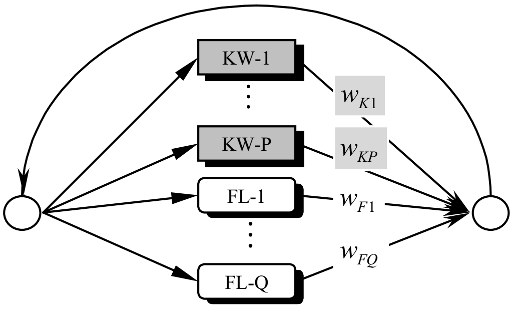

Fig. 1 The KWS network of *P* keywords and *Q* fillers

In *HarkMan*, each keyword was a catenation or a string of Chinese syllables. 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 Chinese KWS tasks.

3.2 Use of Language Models

Though KWS is different from CSR, some CSR technologies can be used in KWS. One is the language model. It is obvious that the performance will be improved if the language model is also used in the KWS system. The best way is to perform continuous speech recognition and then search for the desired keywords, but the time and space complexities become much bigger than expected. Because not the whole sentence needs be recognized and cared about, the word-level language model is not considered, only a syllable based *N*-Gram language model is adopted here. The language model is a syllable Bi-Gram [6].

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 corresponding acoustic-searching path. The syllable Bi-gram weights were different from the network transition weights. After considering the syllable Bi-Gram of two syllables across the boundary between a KW and a FL, the FL will be ignored no matter what syllable it is, this simplifies the CSR to KWS.

The use of syllable Bi-Gram was helpful for pruning some impossible syllable connection thus resulted in higher efficiency and speed.

4. SEARCH STRATEGIES

The frame-based searching algorithm is commonly used in CSR [9]. 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. 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 segmentation and search strategies together to make use of their advantages jointly, they form the two-stage search procedure.

4.1 Segmentation

The frame-energy, zero-crossing-rate, pitch, and/or cepstrum or their derivatives (such as difference, contour, linear regressive analysis) [6] are used to segment the speech. A series of putative separation points (PSPs) are obtained. These PSPs can be true separation points (TSP) or false separation points (FSP). In practice, we will choose a very confident threshold to ensure that the obtained PSPs are all TSPs.

Obviously, the higher the threshold is chosen, the fewer number of PSPs we can obtain, but the bigger ratio of TSPs to PSP we can get. The threshold should be carefully chosen so that as many PSPs as possible can be detected and almost all of the PSPs are TSPs [6].

The output in this stage is passed to the search stage.

4.2 Searching Inside Every Definite Segment (DS)

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 are 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 is then performed using the knowledge obtained previously. The obtained knowledge has been fully used 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.

During the search procedure, the path pruning procedure is very necessary and common, it is definitely unavoidable especially when searching in a relatively long DS. In our system, 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 five 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. (5) The accumulated path score is lower than a given threshold.

5. ROBUSTNESS ISSUE

In real-world applications, the robustness is a key issue to be faced. In KWS, 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 these problems theoretically and practically [1, 10-18]. We take both the performance and complexity into consideration and give our own solutions.

5.1 NIL (Noise Immunity Learning) Technology

A basic idea for noise cancellation is simple. Try to get the spectral characteristics of noises and establish an individual noisy-to-clean feature mapping or an individual model for each type of noise. Because there are many of kinds of noises in the nature, it is almost unsolvable to establish an individual mapping or model for each type of noise. Furthermore, some noises are unknown.

Our solution is on the basis of Noise Immunity Learning (NIL) method [17, 18]. 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. This makes the training simpler and easier but more robust.

5.2 EMM (Embedded Multiple Model) Scheme

After determining the SR units, one acoustic model will be trained for each unit. But it is well known that there is co-articulation phenomenon in continuous speech. A solution to this problem is the well-known di-phone and tri-phone modeling, known as context-dependent (CD) modeling. The CD modeling has its disadvantage, many different CD models should be established for each individual unit. So the preciseness of models are declined if the training data are not enough or the training data do not cover all the di-phones and tri-phones of the desired units. Even if there are enough training data, the big number of CD models will cost too much storage for models and time consumption for recognition.

The above stated EMM scheme for CDCPMs has been proved useful [3], it can eliminate the above shortcomings appearing in the CD modeling and thus adopted here.

The EMM scheme does not only offer a solution to the co-articulation problem, but also is robust for speaker-dependent and gender-dependent cases.

5.3 Arc-Splitting Technology

With regard to the accent, the EMM scheme is only a solution for speakers with light accents. If the accents are too heavy, it does not work well because the accents are not modeled.

In Chinese there are many different accents. It is something different from the different speaker and different context issues. 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 [6] 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 splitted into two parallel paths ‘guo’ and ‘gui’. The grouping of these syllables is based on the known linguistic knowledge and the acoustic distance measure [6].

6. REJECTION METHODS

There are often two stages in a KWS system. (1) In stage one, as many keyword candidates as possible are given so that the actual keywords will not be missed and the detection probability is ensured. (2) In the second stage, 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.

According to our studies, 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 so that the threshold can be easily determined. (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 [3], 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 [3]. These sub-spaces can be estimated by the clustering method according to a certain criterion [7]. Unlike in the vector quantization (VQ) technique where a sub-space is represented by one centroid vector, the CAP takes the continuous valued distribution into consideration.

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

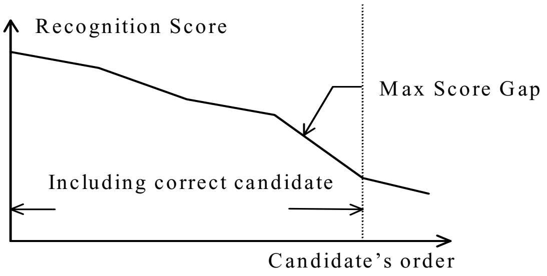

Fig.2 The Curve of Recognition Scores

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 [19, 6]. The experimental result is shown in figure 3.

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 Candidates 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.

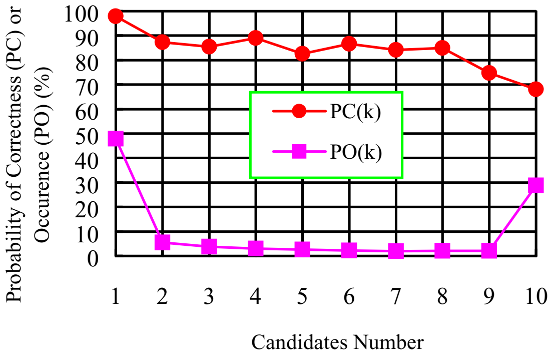

Fig.3 Rejection Results

7. EXPERIMENTAL RESULTS

By combining the above technologies, a keyword spotter for spontaneous Chinese telephone speech named *HarkMan* was implemented. In Fig. 4, its receiver operating characteristics (ROC) curve is shown, ROC curve reflects the probability of detection (P_d) of keywords as a function of permitted false alarm rate (i.e., fa/h/kw, the number of false alarms per hour per keyword). For a KWS system, figure-of-merit (FOM) value is a very important specification to indicate the spotting performance, FOM is the average value of P_d over the false alarm rate interval [0, 10]. From Fig. 4, we can get that the FOM of *HarkMan* is 90.4% and the probability of detection (P_d) 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.

The FOM of *HarkMan* proves that the above technologies are a great progress in the research of KWS. But many of them are still open issues, including the noise cancellation, accent problem and rejection. Further research is still in progress.

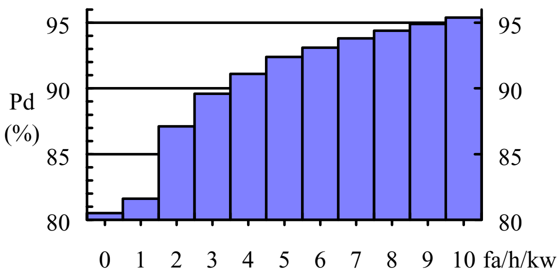

Fig. 4 The ROC Curve of *HarkMan*

8. REFERENCES

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

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

[3] 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

[4] Zheng Fang, Wu Wenhu, Fang Ditang, “Center-distance continuous probability models and the distance measure,” J. of Computer Science and Technology, Vol.5, 1998

[5] 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)

[6] 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

[7] 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

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

[9] 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

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

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

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

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

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

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

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

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

[18] 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

[19] 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

中 文 题 目：HarkMan – 一个词表无关的汉语自然语音关键词识别器

眉 题：HARKMAN - A VOCABULARY-INDEPENDENT KEYWORD SPOTTER

作者英文简介：

Dr. Fang ZHENG currently is an associate professor of Tsinghua University. He is an Associate Director of the Department of Computer Science & Technology, the Director of the Speech Lab, and also the Director of the Analog Devices Inc.-Tsinghua DSP Technology Research Center.

Dr. Zheng was born in Jiangsu Province, P.R. China, in 1967. He graduated from the Department of Computer Science & Technology of Tsinghua University and received his B.S., M.S. and Ph.D. degrees from Tsinghua University, in 1990, 1992 and 1997 respectively, in Computer Science and Technology, Computer Application and Computer Application, respectively. Dr. Zheng has been working in Speech Recognition at Speech Lab., Dept. of Computer Science and Technology, Tsinghua, since 1988. His research interest includes acoustic/language modeling, isolated/continuous speech recognition, keyword spotting, dictating, language understanding and so on.

Mr. Mingxing XU was born in 1973 in Hubei Province. He received his B.S degree from the Department of Computer Science and Technology, Tsinghua University in Computer Science and Technology in 1995. He currently is a Ph.D. candidate in Computer Application. His research interest includes speech recognition and language processing.

Mr. Xiaolong MOU was born in 1973 in Gansu Province. He received his B.S degree from the Department of Computer Science and Technology, Tsinghua University in Computer Science and Technology in 1996. He currently is a M.S. candidate in Computer Application. His research interest includes speech recognition and language processing.

Mr. Jian WU was born in 1975 in Hubei Province. He received his B.S degree from the Department of Computer Science and Technology, Tsinghua University in Computer Science and Technology in 1998. He currently is a M.S. candidate in Computer Application. His research interest includes speech recognition and language processing.

Prof. Wenhu WU was born in Beijing, P.R.China, in 1936. He studied in the Department of Electrical Engineering, Tsinghua University, from 1955 to 1958, and then in the Department of Automation, Tsinghua University, from 1958 to 1961.

Since then, he has been teaching at Tsinghua University and now a Full Professor in the Department of Computer Science and Technology and has been its director from 1990 to 1997.

He is devoted in researching Chinese speech recognition and understanding, especially the speaker-independent Chinese speech recognition. As a result, he has been awarded several times.

He is also devoted in the computer-spread education. He is the chairman of Computer Spread Education Commission of CCF (China Computer Federation). He has led the China Team to take part in the IOI'89 - IOI'95 (International Olympiad in Informatics) and won many golden medals.

Prof. Ditang FANG was born in Shanghai, P.R.China, in 1930. He received the B.S. degree from Jiaotong University and the M.S. degree from Tsinghua University, both in electrical engineering, in 1953 and 1956, respectively.

Since then, he has been teaching at Tsinghua University and now a Full Professor in the Department of Computer Science and Technology. In 1979, he founded the Laboratory for Human-Machine Speech Communications and has been its director from 1979 to 1990. The laboratory received the National Scientific Research and Technology Progress Award twice, in 1987 and 1989, respectively, the National Scientific Invention Award in 1990, and three other awards.

He is the Deputy Chief of the Artificial Intelligence and Pattern Recognition Committee of the Chinese Computer Science Society.