Proceedings of the Seventh IASTED International Conference
SIGNAL AND IMAGE PROCESSING
August 15-17, 2005, Honolulu, Hawaii, USA

THE DYNAMICALLY-ADJUSTABLE HISTOGRAM PRUNING METHOD FOR EMBEDDED VOICE DIALING

Defeng CHEN, Fang ZHENG, Jian LIU, Jing DENG, Wenhu WU, Zhanjiang SONG[1], and Xunyi ZHOU[1]
Center for Speech Technology, State Key Laboratory of Intelligent Technology and System,
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[chendf, fzheng, liuj, dengj, wuwh]@cst.cs.tsinghua.edu.cn, http://cst.cs.tsinghua.edu.cn
[1]Beijing d-Ear Technologies Co. Ltd., [zjsong, xyzhou]@d-Ear.com, http://www.d-Ear.com

ABSTRACT
Memory and speed are two key factors that must be faced when applying voice dialer to Pocket PCs. To provide a solution, a novel decoding method integrated with the score differences of token paths is proposed, named as "Dynamically-Adjustable Histogram Pruning". Additionally, the computation of likelihood score is accelerated by means of dynamic score lookup table. Furthermore, a new acoustic modeling method based on Extended Initial/Final (XIF) with less dimensioned acoustic feature is proven suitable for embedded speech command recognition. By adopting the methods developed above, we implement a speaker-independent, user definable voice dialing speech recognition system with good performance on a real PDA device. For a 200-Chinese-word vocabulary, its recognition accuracy reaches 97.80%. Meanwhile, it obtains better recognition speed by 80 times and saves decoding space by 30% in comparison to the baseline system using standard Viterbi decoding method.

KEY WORDS
Speech Recognition, Voice Dialing, User definable Vocabulary, Dynamically-Adjustable Histogram Pruning

1. Introduction

Speech is considered as one of the most important, straightforward and natural ways of communication between human and machine. With the rapid development of wireless communications and personal digit assistant (PDA), research on embedded speech recognition technologies has become so hot that various companies and organizations, such as IBM, Microsoft, d-Ear Technologies, InfoTalk Co. Ltd., Chinese Academy of Sciences, and *etc*, have laid much emphasis on the development of embedded speech recognition engine.

Though speaker-independent large vocabulary continuous speech recognition with high performance has been successfully implemented in lab, it still fails to deliver satisfactory performance in the embedded mobile communication device (PDA) due to following limitations:

(1) Low computing speed. The instruction microprocessor runs is mainly MIPS, which is merely comparable to the capability of PC 486.
(2) Small storage space. Its size is often 16M or 32M, which is incomparable to that in PC nowadays.
(3) Different recording channel. Sound Noise Rate (SNR) of the built-in microphone in PDA is rather low, whose channel type is apparently different from those in PCs or in telephones, and usually the corpus for model training is not collected in real PDA environment.

Therefore, the key to embedded speech command recognition application in PDA is to reduce the time and space complexities to meet the demands of embedded system under acceptable degradation of performance[1]. We note that these could be realized through two ways: (1) Applications on DSP processor chips. It improves the recognition speed mainly at the level of chip microprocessor. Good examples include: ADSP based Voice Command and Control development by Center for Speech Technology, Tsinghua University[2], Fixed-point DSP Chip based on System-on-Chip by Department of Electronic Engineering, Tsinghua University[3][4], and so on. In this way, speech recognition system is built in the chips, which is widely adopted due to poor hardware performance before. However, the insufficiency results from speaker-dependence, on-line training models, low recognition precision, low vocabulary size, poor flexibility and unsatisfactory recognition accuracy. (2) Applications based on algorithm. They are mainly focused on the reduction of model size and the optimization of decoding space and decoding speed; meanwhile, it could accelerate the computing by means of fixed-point computing on software layer[5][6]. With the advancement of hardware, i.e., the capacity of computing and memory, the applications in this way have become popular in present. Not only could user train model on line to carry out speaker-dependent speech recognition, but also load models trained in advance to realize speaker-independent speech recognition. Its merits were high recognition precision and flexibilities. However, the

disadvantage lies that it would have to trade off a lot between scalability and precision, and need to strictly control over the decoding cost and to improve decoding speed.

This paper briefly describes the research on speaker-independent speech command recognition system (Voice Dialer) based on embedded system in Pocket PC. To reduce the memory cost and to accelerate the computation, the following steps are taken. (1) By studying the dynamic interdependency between the rank of correct token path during Viterbi decoding and the input speech frames, and the interrelationship between the score differences of token paths and the input speech frames in Viterbi beam search, we propose a new search strategy named Dynamically-Adjustable Histogram Pruning with the integration of the difference scores of token paths. (2) By studying the repetition of likelihood scoring, we adopt the likelihood score lookup table and further accelerate Viterbi decoding. (3) By studying the less dimensioned acoustic feature for modeling that merely leads to slight performance degradation, we develop a new acoustic modeling method based on Extended Initial/Final (XIF), which proves suitable for embedded voice dialer.

Through sufficient experiments, we apply these methods proposed above to improve the performance of "d-Ear Voice Dialer".

The rest of this paper is organized as follows. The next section mainly covers the decoding method of "Dynamically-Adjustable Histogram Pruning" combined with the score differences of token paths to strictly control decoding cost; Section 2 describes the method of using score lookup table to accelerate Viterbi decoding. Section 3 introduces the acoustic modeling method based on Extended Initial/Final (XIF). In section 4, the application of "d-Ear voice Dialer" and its experimental results of performance are presented, which is followed by the conclusions.

2. Dynamically-Adjustable Histogram Pruning Strategy

Generally speaking, there are three factors that should be taken into consideration in speech recognition decoding [7].

(1) Accuracy. Various knowledge sources could be utilized to enhance the accuracy of recognition.
(2) Cost. Resources should be consumed as low as possible, including memory and hard disk storage.
(3) Efficiency. Recognition results could be obtained as soon as possible.

Given limited computational resources, it becomes important to find an appropriate approach to get these factors balanced. Previous experiments have proven that the computation of decoding obtains about 80% of overall computation in speech recognition[8], naturally speaking, the improvement of decoding speed is key to whole speech recognition system. Frame synchronous Viterbi search in HMM decoding demands intensive computation and too much memory cost, which is not suitable for embedded system, especially in large-vocabulary speech recognition, because the disadvantage is apparent. Therefore pruning methods are adopted to only permit the computation of most promising token paths. Whereas, inappropriate pruning methods may exterminate the potential optimal token path too early and undermine the accuracy. Thus it's important to find an appropriate pruning strategy to achieve the balance between the speed and the accuracy.

2.1 Traditional Pruning Strategy

Traditional pruning algorithms include two methods --- likelihood threshold based pruning and rank threshold based pruning, which are called as "Beam Pruning" and "Histogram Pruning" respectively [9].

Likelihood threshold based pruning allows only those token paths whose likelihood score is larger than the threshold, where the threshold of likelihood score is called "Beam Width". As the score threshold often is fixed during pruning, it's named as "Fixed Beam Pruning". Fixed Beam Pruning could significantly reduce the decoding cost when its threshold is small enough. However, the token paths could increase considerably when most paths have similar likelihood scores. Therefore, it's not effective enough to control decoding memory cost, for it fails to precisely predict and control the number of token paths that should be kept.

Rank based pruning, or, Histogram Pruning, could effectively avoid such complexion, for in any case, the number of token paths kept is affirmatory and fixed, so that the memory cost of Histogram Pruning is well predictable and under precise control, where the fixed number of token paths kept is named as "Beam Width". As for speech command recognition in embedded system, Histogram Pruning may be a good choice, but it's not easy to find the balance of accuracy and speed when the beam width is fixed, because high accuracy requires large beam width, which means heavy load of computation, then Histogram Pruning should be improved before being applied to voice dialer.

2.2 Dynamic Histogram Pruning

In general, when the Beam Width is small, the accuracy of Histogram pruning is low; and the accuracy accrues as the threshold becomes wider, but converges under the ceiling accuracy of Viterbi decoding. Therefore, the accuracy depends much on the threshold of Histogram Pruning. To investigate the relationship between them, we trace the rank of correct token paths during standard

Viterbi search in a 200-Chinese-word vocabulary including 150 three-character-length words and 50 two-character-length words, and obtained the statistical overall trend of correct token path rank shown in figure 1 below.

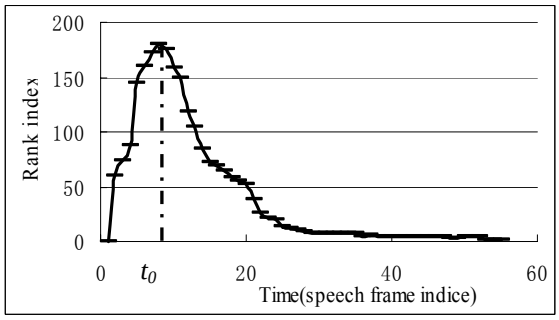

Figure 1. Correct token path rank trend

According to the experimental results demonstrated in Figure 1, the interdependence between the rank trend of correct token paths and the time (speech frame index) observes the following principles:

(1) At first, the rank severely decreases (rank index increases) as speech frame index increases, and the trend of decrease is random but would not exceed the size of vocabulary in average. The reason for this is that the early input speech is not long enough to discriminate candidates, so that the rank of correct one would not be surely higher than that of others.

(2) As input speech frame index increases, the rank starts to increase (rank index decreases), for more speeches help to distinguish different words, and the speeches start matching acoustic model better.

(3) In the latter phase, the token path of correct candidate occupies the top rank when more speech frames input. Moreover, the top paths remains stable, while the latter ones have rare chance to hit the top rank in this phase. Because after more speech are inputted, the extent that speech matches acoustic model increases, such that only those whose pronunciation are similar to the correct candidate have chance to reach the top ranks.

According to the experimental results, the following conclusions are drawn. (1) The beam threshold could dynamically drop by time during the pruning. (2) The pruning method should always ensure correct candidate always to be included in the threshold at the early phase, so that the initial beam threshold should be large enough, however, often not larger than vocabulary size. (3) The peak of correct candidate rank curve emerging time (t_0) linearly depends on the average word length (Len) and the total length of input speech (T), which can be formulated empirically as follows:

$$t_0 \approx 1.5 \times T/(Len+2) \quad (1)$$

Considering the result presented above, we propose a new pruning method, which is a time-dependent dynamic Histogram pruning firstly. Beam width threshold is wide enough to accommodate many token paths, whose size depends on the vocabulary size, and then the threshold dynamically decreases with speech frame input, which could be considered as a nonlinear time varying process. To simplify this interdependence, by means of piecewise linear method, we use linear difference equation to approximate the overall rank trend curve in Figure 1 in every piece of time. The rank beam width threshold at time t, $W_d(t)$, could be decided by following equation.

$$W_d(t) = g(t) \quad (2)$$

Where $g(t)$ is a piecewise linear function related to the input speech frame index t, for instance, $g(t)$ could be simply defined as follows:

$$g(t) = \begin{cases} g(t_0) & 0 \le t < t_0 \\ g(t_0) - a_1 \times (t - t_0) & t_0 \le t < t_1 \\ \quad \ldots & \quad \ldots \\ g(t_{m-1}) - a_m \times (t - t_{m-1}) & t_{m-1} \le t < t_m \end{cases} \quad (3)$$

Given that the overall trend curve is approached by using m pieces of division, variables $\{a_1 \ldots a_m\}$ can be approximately considered as constants, and set in needs of how fast or slow $W_d(t)$ decreases in practice. Therefore, the threshold of pruning is firstly adjusted by function $g(t)$, which saves much computation during the decoding.

2.3 Dynamically-Adjustable Histogram Pruning Integrating Score Differences of Token Paths

To control the dynamic pruning threshold in a precise way, we take further steps to investigate the likelihood score of token paths during the pruning. The experimental fact is that at the early phase of the decoding, the difference score of token path whose ranks differ much from each other is very small, and they are likely to be the first ones in later phase, such that they should be kept to participate in the further decoding. It happens frequently especially where there are many similar or same pronunciations in the vocabulary. On the contrary, in the late phase of decoding, the difference score of neighbor rank is so large that the smaller one has slight chance to become the top candidate. Therefore, we propose to combine the difference scores of token paths to adjust the dynamic pruning threshold at certain time, such that the decoding cost is effectively controlled and the decoding speed is enhanced.

At time t, the threshold of pruning $W_d(t)$ is precisely determined by means of the score differences between that of the top token path and the bottom one, i.e. $\Delta Score$. If $\Delta Score$ is large, which shows that the bottom token paths have less chance to be the final candidates, the pruning threshold could be more tightly contracted. Contrarily, if $\Delta Score$ is small, which promises good chance to be the final candidates, then the pruning threshold should be less contracted to avoid pruning the potential token paths too early. Therefore, the pruning

threshold $W_d(t)$ at time t is determined by the equation below,

$$W_d(t) = W_d(t-1) + f(\Delta Score(t)) \quad (4)$$

Where function $f(x)$ depends on the score difference of token path ($\Delta Score(t)$), that is, x. If x is larger than the score threshold, then narrow the pruning threshold $W_d(t)$ till $\Delta Score(t)$ is close to the score threshold. While x is smaller than the score threshold, then expand $W_d(t)$ till $\Delta Score(t)$ is close to the score threshold. However, the score threshold mentioned could be dynamic or static, and a dynamic one is proven better. By means of $f(x)$, $W_d(t)$ is moderated more precisely, thus the correctness of pruning is guaranteed in fast decoding.

Combining these two ways of dynamic control over the pruning threshold at time t, $W_d(t)$ is determined by functions $g(x)$ and $f(x)$, that is,

$$W_d(t) = W_d(t-1) + g(t) + f(\Delta Score(t)) \quad (2)$$

2.4 Comparison Experiments

To evaluate the performance of the decoding strategy (DHP) proposed above, the comparison experiments with standard Viterbi decoding (VTB) and fixed beam histogram pruning (FHP) are carried out as follows, where the fixed beam of FHP is 50. The experimental condition as follows. For the same speech input and the same 200-word-sized vocabulary, whose average word length is 2.7 characters, given the input speech length is T frames, then the peak point t_0 could be calculated by Equation 1: $t_0 = 0.32T$, for function $g(t)$, $m = 4$, $\{a_1,...,a_m\} = \{100/T, 400/T, 200/T, 120/T\}$, and $\{t_1,...,t_m\} = \{0.4T, 0.5T, 0.75T, T\}$; for function $f(\cdot)$, the dynamic thresholds are {80.00, 90.00, 100.00, 120.00}; the experimental result in Pocket PC (Intel StrongARM，206MHz，32MB RAM) is illustrated in Figure 2.

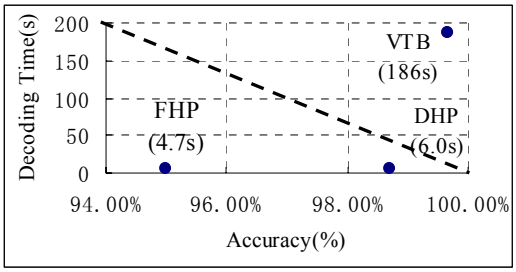

Figure 2 Comparison of Three Decoding Strategies

According to the experimental results, we find that the key factor affecting the decoding recognition accuracy is whether it could contain adequate token paths at the early phase. At the later phase, we just score the top few token paths to get the result, which could save much computation, meanwhile, would not compromise the accuracy too severely. Effectively as FHP saves the computation, its accuracy is unacceptably undermined, because the pruning threshold is fixed and too narrow at the early phase that the potential candidates are pruned incorrectly in advance. Combining the difference score of token paths to tune the dynamic pruning threshold at each time, DHP enhances the speed of decoding with acceptable accuracy by setting wide threshold at first and narrowing the threshold gradually.

3. Optimization of Likelihood Computation

Pruning could be carried out in two layers practically: state-level pruning and model-level pruning [10]. The former only prunes the states with low rank, while the latter exterminates low ranked models containing certain states. Considering the efficiency and economization, model-level pruning is more suitable for embedded speech recognition.

To further improve the decoding efficiency and to reduce decoding memory space, we not only take advantage of lexical tree representation[11], but also use likelihood score lookup table to optimize decoding. Sufficient experiments reveal that there are plenty of repeated computation (on same models or states) during the course of state likelihood scoring for every frame of input speech. Generally speaking, statistics show that the computation of state likelihood scoring occupies over 80% of the overall decoding computation [12]. Therefore, the important factor for improving decoding speed is optimizing state likelihood scoring. According to this, we made use of data share technology. For every frame, we dynamically store the scores of every state in a table, i.e. score lookup table. When the state has not been scored before, and its score cannot be found in the lookup table, then the scoring is done and afterward stored in the lookup table, while when the same state has been scored, then fetch the score of this state from the lookup table directly. In this way, the computation of scoring is significantly reduced, and the decoding is accelerated correspondingly while the accuracy remains untouched.

When using Dynamically-Adjustable Histogram Pruning, we make comparison experiments to test the advantage of likelihood score lookup table technology and obtained the results that figure 3 illustrates below, where the percent represents the percentage of computation saved.

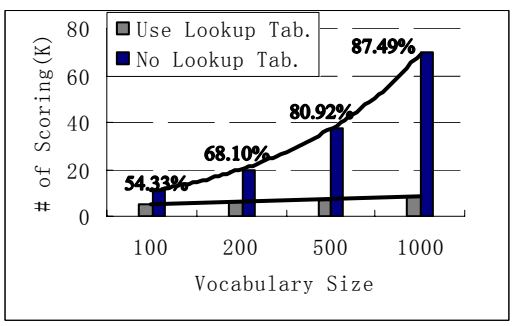

Figure 3 Optimizing decoding using data sharing tech.

As experimental results manifest, without the likelihood score lookup table technology, the computation of scoring increases in second-order polynomial, while with dynamic data sharing technology, when the vocabulary is small, scoring number increases at a low rate, and converges in a stable range when the vocabulary size accrues. Because the likelihood score lookup table technology ensures to compute every state only once for every frame of speech when the vocabulary covers all of the model states, then the sum of scoring merely depends on the number of model states and the length of the input speech.

4. Acoustic Modeling for Voice Dialer

As we know, the acoustic modeling plays an important role in speech recognition. There are two major acoustic modeling methods in speaker-independent speech command recognition: (1) Isolated word modeling. Its merits are small computation and high accuracy, but the apparent defect is its hard-to-custom vocabulary. (2) Lexical tree decoding[13] and Initial/Final based modeling[14]. Its merits are high precision and user-definable, large-sized vocabulary, no need of on-line training, which is practical. However, its shortcoming is high perplexity of algorithm.

In order to cater to the need of no on-line training, flexibility and user-definability for vocabulary, we should adopt a speech recognition unit (SRU) which is smaller than isolated word. In this paper, illuminated by isolated word modeling, lexical tree decoding and Initial/Final based modeling, we choose extended Initial/Finals[15] as SRUs, and train models off-line which would be loaded and connected into various isolated word during the decoding. In this way, the vocabulary would be easily defined and changed.

As for feature, the Mel-Cepstrum coefficient is proved to have good noise robustness, and is fit for dealing with the variation and noise that embedded devices may face. Considering that in speech spectrum, low frequency part contains much noise while high frequency part includes pitch and harmonics, which have less effect on the performance of speech recognition, we discard these two parts through filters when extracting speech features, in such a way that the feature dimension is downsized, then by means of Cepstrum Mean Normalization (CMN)[16] processing, we obtain a 20-dimensioned feature ({9MFCC + 1Engergy} + 1st-order derivative). To test its performance, we do the comparison experiments with 42-dimensioned feature ({9MFCC + 1Engergy} + 1st-order derivative + 2nd-order derivative) and 28-dimensioned feature ({9MFCC + 1Engergy} + 1st-order derivative), and obtain the results as below in figure 4 with the same training database and test database.

As experimental results demonstrate, the degradation of performance of 20-dimensioned feature is acceptable in speech command recognition. Furthermore, it leads to the reduction of model scalability and of computation of likelihood scoring, thus accelerating the overall speech recognition system.

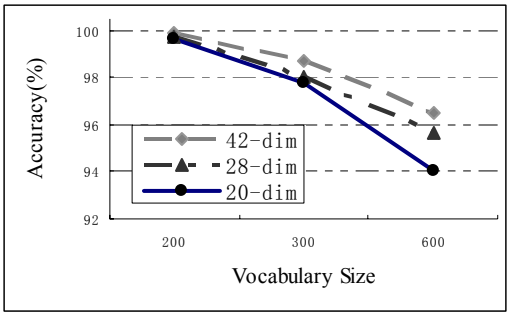

Figure 4 Comparison of 3 kinds of features

5. Design and Analysis of Voice Dialer

According to the analysis on decoding strategy and acoustic modeling, we design and implement a practical Chinese Voice Dialing System - "d-Ear Voice Dialer", which automatically retrieves names from the contact list of PDA, forms its vocabulary, and helps users quickly call someone via voice directly. Meanwhile, user can define the voice dialer vocabulary: customizing the pronunciation of one name, for instance, assigning a nickname for someone. Still, the recognition parameters can be defined as well. This voice dialer is a speaker-independent speech command recognition system without on-line training, or recording voice tags before hand, and is of high accuracy, high speed and small model size (only 290KB), and is suitable for embedded device.

To test the performance of the Voice Dialer, we make comparison experiments as below, the baseline system uses the standard Viterbi decoding, without using the dynamic likelihood score cache technology, while "d-Ear Voice Dialer" uses the Dynamically-Adjustable Histogram Pruning integrated with the score differences of token paths, with dynamic likelihood score lookup table technology (*abbr*. d-Ear VD).

The vocabulary for testing consists of 200 randomly selected Chinese names, including 50 two-character names and 150 three-character names. The model for testing is trained off line from plenty of telephone speech corpus using HTK [17]. The testing speech corpus is collected on Dopod Pocket PC from 6 persons (4 males & 2 females) with sampling rate of 8,000Hz and sampling precision of 16bits. Each of testers speaks the name in the testing vocabulary once and every speech lasts about 3 seconds, which adds up to 1,174 utterances in total. The experimental results are illustrated in table 1.

Table 1 Comparison of different systems

System	Acc./%	Time/s	Memory/KB
Baseline	99.66	186.0	190
d-Ear VD	98.70	2.3	120

As experimental results demonstrate, by synthesizing the new decoding strategy, the accuracy of "d-Ear VD" decreases slightly by 0.96% in comparison to the baseline system, which is acceptable in embedded applications. The decoding speed and memory cost are significantly improved by 80 times speed promotion and 30% memory cost reduction. Moreover, the application of "d-Ear VD" in practice proves the applicability and the effectivity of the decoding strategy we adopted.

6. Conclusion

In all, a novel decoding method, Dynamically-Adjustable Histogram Pruning combined with the difference scores of token paths is proposed to improve decoding speed and memory cost, moreover, a method of dynamic score look up table is used to accelerate the computation of likelihood score. Still, a new acoustic modeling method based on Extended Initial/Final (XIF) with less dimensioned feature is proved suitable for embedded speech command recognition. By using the methods developed above, we implement a speaker-independent, user definable voice dialing speech recognition system, which delivers good performance on a real PDA device. In 200-Chinese-word vocabulary, its recognition accuracy reaches 97.80%. Meanwhile, it obtains better recognition speed by 80 times and saves decoding space by 30% in comparison to the baseline system using standard Viterbi decoding method.

References

[1] Min FANG, Jiantao PU, Study and implementation of embedded speech recognition system, *7th National Conference on Man-Machine Speech Communications*, 2003, Xiamen, China (in Chinese)

[2] Fang Zheng, Qixiu Hu, Xiang Deng, Wenhu Wu and Ditang Fang. An introduction to a kind of voice dialers for dummies, *4th National Conf. on Man-Machine Speech Communications*, pp. 165-168, Oct. 1996, Beijing, China (in Chinese)

[3] Chunyi ZHU, Jianhua LU, Runsheng LIU, Design and implementation of DSP based voice electronic notebook, *Electronic Technology and Application,* 2002 Vol. 9 (in Chinese)

[4] Jiamin JING, Jia LIU, Runsheng LIU, Application of HMM based speech recognition system in embedded system, *Application of Electronic Technique*, 2003 Vol.29 No.10 (in Chinese)

[5] Yonggang DENG, Bo XU, Taiyi HUANG, Speech recognition algorithm and implementation in Palm PC, *Computer Research and Development*, Aug. 2000 (in Chinese)

[6] Lingyun XIE, Limin DU, Bin LIU, Implementation of fast Gaussian computation in embedded speech recognition system, *Computer Engineering and Application, 2004 Vol. 23* (in Chinese)

[7] Guoliang ZHANG, Search algorithm in large vocabulary continuous speech recognition (Master Thesis) (Dept. Computer Sci.&Tech., Tsinghua University, Beijing, China, 2001) (in Chinese)

[8] J. Suontausta, Fast decoding in large vocabulary name dialing, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, 2000

[9] Dongbin ZHANG, Limin DU, Adaptive beam search method in speech recognition, *Electronic and Audio Engineering,* 2004 Vol. 8 (in Chinese)

[10] Janne Suontausta, Fast decoding techniques for practical real-time speech recognition system, *IEEE Automatic Speech Recognition and Understanding Workshop*，Keystone，Clorado, 1999

[11] Guoliang ZHANG, Mingxing XU, Jing LI, et al, "Cross-word search algorithm based on two-layer lexical tree in speech recognition," *J. Tsinghua Univ* (*Sci &Tech*), 43(7): 981-984, Jul. 2003 (in Chinese)

[12] Imre Kiss, Marcel Vasilache, Low complexity technique for embedded ASR system, *International Conference on Spoken Language Processing*, Denver, Colorado USA, 2002

[13] Guoliang Zhang, Fang Zheng, Wenhu Wu, "Two-layer lexical tree based search algorithm for LVCSR," *National Conference on Man-Machine Speech Communications*, Shenzhen, 2001, 239-242 (in Chinese)

[14] Feng LI, Jiantao PU, Study of Initial/Final concatenation and full word recognition based speaker-independent isolated word recognition, *7th National Conference on Man-Machine Speech Communications*, Xiamen, China, 2003 (in Chinese)

[15] Jing LI, Mingxing XU, Jiyong ZHANG, et al, Comparison on acoustic modeling unit in continuous Chinese speech recognition: syllable, phoneme and initial/final, *6th National Conference on Man-Machine Speech Communications*, Shenzhen, China, 2001, 267-271 (in Chinese)

[16] Fang Zheng, Guoliang Zhang. Integrating the energy information into MFCC, *International Conference on Spoken Language Processing*, Beijing, China, 2000, I-389~292

[17] Yong, S., Kershaw, D., Odell, J., Ollason, *et al*, The HTK book (for HTK version 2.2) (Cambridge University, 1999)