A NOVEL EFFICIENT DECODING ALGORITHM FOR CDHMM-BASED SPEECH RECOGNIZER ON CHIP

ZHU Xuan, CHEN Yining

Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, P.R.China
Email: {zhuxuan98, chenyining99}@mails.tsinghua.edu.cn

LIU Jia, LIU Runsheng

Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, P.R.China
Email: {liuj, lrs}@mail.tsinghua.edu.cn

ABSTRACT

The efficiency of decoding algorithm is the main limitation of realizing a CDHMM-based large-vocabulary name dialing application on consumer electronic products. To solve this problem, a novel efficient decoding algorithm, TPVD (Two-Pass Viterbi Decoding), is represented in this paper. By coarse matching in the first pass, several high-confidence candidates are selected from hundreds. Then, fine matching in the second pass runs on base of those high-confidence candidates. Compared with traditional Viterbi beam search, the hardware resource requirement is significantly reduced. Meanwhile, the recognition accuracy and speed are satisfactory. This conclusion has been proven in an embedded mandarin 600-name dialing system.

1. INTRODUCTION

In recent years, embedded speech recognizer becomes a hot topic of speech recognition technique. In the field of consumer electronics, small-vocabulary speaker-dependent name dialing and speaker-independent command controlling have been implemented in mobile phones, personal digital assistants and toys. DTW (Dynamic Time Warping) and DHMM (Discrete Hidden Markov Model) algorithms are employed in these systems [1]. Because of time-consuming training process and poor robustness, this function is not convenient enough for users. On the contrary, introducing the speaker-independent large-vocabulary name dialing function based on CDHMM (Continuous-Density Hidden Markov Model) has more practical value, but the tradeoff between its performance and cost is still a major problem of realizing CDHMM-based systems [2][3][4].

To enhance the efficiency of a CDHMM-based name dialing system, a TPVD algorithm is proposed in this paper. Through coarse matching in the first pass, several candidates can be chosen from hundreds, thus the second decoding pass is reduced to a small-vocabulary isolated-word recognition task, in which the confidence measure is adopted as the criterion to select those candidates. When the on-chip memory size is limited, the recognition time with TPVD algorithm is much shorter than that with traditional Viterbi beam search. Meanwhile, the recognition accuracy doesn't decrease. Since the two-pass decoding is still based on Viterbi search, the efficient methods used in Viterbi search can also be applied. The similar idea has once been adopted in a LVCSR (Large-Vocabulary Continuous Speech Recognition) system on PC platform [5].

The contents of this paper are arranged as follows. Section 2 introduces the Viterbi beam search and two-pass decoding algorithms in details. Section 3 describes the hardware platform on which this algorithm is implemented. Experiments and results are illustrated in section 4. Finally, conclusions are made in section 5.

2. ALGORITHMS

2.1. Viterbi Beam Search

Viterbi search algorithm, a synchronous decoding algorithm, is the most popular decoding method in current HMM-based speech recognition systems. To decrease the search space and time, Viterbi beam search with tremendous pruning strategies has often been employed in the decoding process. In these systems a beam width is set at the beginning. If the accumulated score of a partial path is in beam, that partial path will be considered as an active path and permitted to be extended; otherwise the corresponding path will be cut out [6].

However, even with Viterbi beam search, the on-chip hardware resources can still hardly meet the requirement of a large-vocabulary name dialing system. From experiment 4.2.1 we can see that 300kByte on-chip memory is the minimum requirement of real time recognition with Viterbi beam search method. If it is not satisfied, much longer time must be waited for the results or only a low accuracy can be achieved.

In experiment 4.2.1, we employ a pruning strategy as follows to test the performance of beam search in our system.

$$BW(n) = \mu \times [LL_{\max}(n) - LL_{\min}(n)]$$

where $0 \le \mu \le 1$. $BW(n)$ is the beam width at the nth frame. $LL_{\max}(n)$ and $LL_{\min}(n)$ are the maximum and minimum accumulated scores of all the partial paths at the nth frame.

2.2. Two-Pass Viterbi Decoding

Since simply adopting the Viterbi beam search cannot meet the system demand, a more efficient algorithm is needed. In this paper we represent the TPVD algorithm to realize a CDHMM-based large-vocabulary name dialing system on chip. In TPVD,

the recognition process is divided into two decoding passes. In the first pass a coarse matching with simpler HMM is done in real time. According to the results of the first pass, the names with high confidence measures are selected as the candidates in the second pass. Then, the fine recognition result can be obtained with more complicated HMM. The details of this algorithm are described as follows.

2.2.1. The first step: coarse matching

The first step is a normal Viterbi decoding process. The most important thing in this step is to choose a proper set of HMM, which needs to fulfill two conditions. Firstly, this set of HMM must be simple enough to insure the first pass running in real time. Secondly, as the basis of second pass, this set of HMM must be fine enough to make almost all the right results acquiring high confidence with it.

In the following experiments, a set of monophone model is adopted as the fast-matching model in a 600-name recognition task. The total memory consumption with this set of HMM in the first pass is 100kByte. The decoding time of the first pass is about 0.4X real time on a 100MIPS (Million Instructions per Second) DSP. Though the 1-best recognition accuracy is only 90.60%, but the 7-best recognition accuracy can achieve 99.02%. That is to say if we choose the 7-best results as the candidates in the second pass, the error brought out by the first pass will be no more than 1 percentage point and the first pass can be completed in real time.

2.2.2. The second step: candidates' selection

The target of the second step is to select the candidates in the second decoding pass. In this step, a tradeoff between recognition accuracy and users' waiting time must be made. Increasing the candidate number will benefit the final recognition accuracy, but will also prolong users' waiting time since the second-pass decoding running in non-real time becomes more complicated.

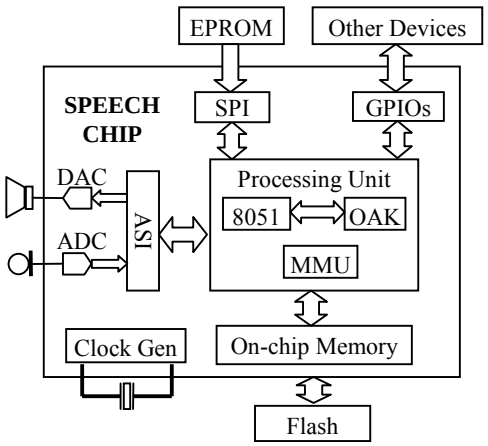
An efficient confidence measure NLLR (Normalized Log Likelihood Ratio) is introduced to choose the right candidates. The equation of this confidence measure is

$$NLLR_v = \frac{1}{T}\log\frac{LK_v}{LK_{\max}} = \frac{1}{T}\left(LL_v - LL_{\max}\right)$$

where T is the frame number of recognized speech; LK_v and LL_v are the likelihood and log likelihood of *vth* candidate; $LK_{\max}$ and $LL_{\max}$ are the maximum likelihood and log likelihood in all the candidates respectively. In our system a threshold Th is predefined. Only when $NLLR_v > Th$, the *vth* name will be considered as a candidate in the second pass [7]. Since the computation of NLLR criterion is very simple, it affects the users' waiting time very little. The good performance of NLLR criterion will be proven in experiment 4.2.3.

2.2.3. The third step: fine matching

The third step is also a normal Viterbi decoding process. Base on the candidates selected in the first and second steps, the third step is simplified to a small-vocabulary name dialing task. With a set of complicated HMM, the precise result can be achieved. It should be noticed that the third step cannot run in real time. Its recognition time is equal to the users' waiting time. Under the conditions of experiment 4.2.3, when the candidate number is no more than 15, the users' waiting time is acceptable. Therefore, an upper limit of 15 is set to the candidate number. In addition, when the candidate number is 1, we can attain the recognition result without this step and the recognition process will be completed in real time.



ADC：Analogue Digital Converter
DAC：Digital Analog Converter
MMU：Memory Management Unit
ASI：Audio Sample Interface
SPI：Single Program Initialization
GPIO：General Purpose Input/Output interface

Fig1. The block gram of hardware platform

2.2.4. Summary

From the detailed descriptions about TPVD, we know that:

- ✧ As two different sets of models and search networks are employed in the two decoding passes, upon the end of the first pass, the on-chip memory occupied in this pass can be released and reused in the second pass. Hence, the on-chip memory demand is much smaller than traditional Viterbi decoding algorithm.
- ✧ Since the first decoding pass runs in real time, the users' waiting time is determined by the second decoding pass, i.e., a small-vocabulary isolated-word recognition time.
- ✧ Although this algorithm is different from traditional Viterbi search, each pass of this algorithm is still based on Viterbi decoding. Therefore, almost all the efficient methods adopted in traditional Viterbi search can also be used to improve the performance.

3. HARDWARE PLATFORM

Since the algorithm is designed for an embedded large-vocabulary name dialing system, a proper hardware platform must be selected. Considering the factors of performance, cost and power consumption, we choose Infineon Unispeech 80D51 as the central processor of our hardware platform. The basic architecture of this chip is illustrated in figure 1.

There are two processing cores on this chip. One is 8051 series core, which is the master core and controls all the operations in this system. The other one is OAK, a 16bit fixed-point DSP core, who does mathematical computation in the system. The operation speed of OAK is 100MIPS. The on-chip memory is 104kByte. Two sets of ADC/DAC are integrated on this chip. The sample rate of both ADC/DAC is 8 kHz. The quantification precisions of them are 12bit and 11bit respectively [3].

In this hardware platform, there are several important peripherals of Unispeech 80D51. EPROM (Erasable-Programmable Read-Only Memory) and FLASH memory store the program code and data separately. An electret receives the voice and passes it to ADC. A speaker radiates the speech prompt generated from DAC.

4. EXPERIMENTS AND RESULTS

4.1. Corpus, features and models

Training set: National 863 Mandarin Continuous Speech Corpus. It includes about 46 hours' continuous speech read by 83 male speakers. All those voice pieces are re-recorded with Unispeech 80D51. The sample rate is 8 kHz and the quantification precision is 12bit. The SNR (Signal Noise Ratio) of this corpus is about 30dB.

Test set: A name dialing speech corpus. It includes 10 male speakers and each speaker reads 600 names. It is directly recorded with Unispeech 80D51. The sample rate is 8 kHz and the quantification precision is 12bit. The SNR of this corpus is about 12dB.

Feature: 27 dimensions of feature vector. It consists of 12-dimension MFCC (Mel-Frequency Cepstral Coefficient), 12-dimension ΔMFCC, the normalized energy as well as its first and second derivatives. [8]

Coarse matching models: a set of monophone models. It involves 208 states and each state is fit with one Gaussion component. Its memory occupation is about 23kByte.

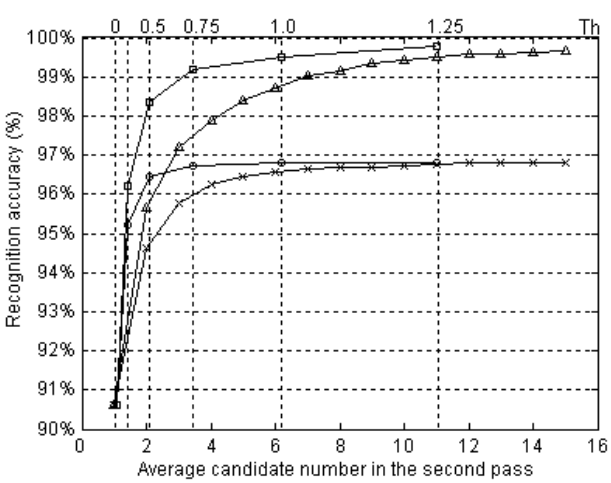
Precise matching models: a set of word-internal biphone models. It involves 358 states and each state is fit with the mixture of four Gaussian components. Its memory consumption is about 160kByte.

4.2. Experiments

4.2.1. Traditional Viterbi decoding with beam search

Experiment 4.2.1 is to test the performance of traditional Viterbi decoding method with beam search in the 600-name dialing system. The pruning strategy proposed in section 2.1 is applied in this experiment. The recognition accuracy and running time is listed in table1. While $\mu = 1$, the recognition accuracy is equal to that with full search.

From the result of experiment 4.2.1 we can see that if Viterbi beam search algorithm is adopted in our system, we can not acquire high performance and low cost at the same time, since the on-chip memory contributes much to chip cost. This problem has limited the implementation of CDHMM-based speech recognizers on chip.



+ Square-mark line and triangle-mark line are multi-candidate accuracies in the first pass with and without NLLR criterion respectively.
+ Circle-mark line and x-mark line are final accuracy of TPVD algorithm with and without NLLR criterion respectively.
+ The numbers on the top of this figure represents the threshold value of NLLR criterion.

Fig2. Recognition accuracy of TPVD

Tab1. The performance of Viterbi beam search algorithm

μ	Recognition Accuracy	Recognition time (X real time)	
		OCM>=300kB	OCM<300kB
0.5	36.38%	0.366	2.304
0.5625	54.02%	0.368	2.306
0.625	70.20%	0.394	2.331
0.6875	82.32%	0.419	2.356
0.75	91.10%	0.466	2.403
0.8125	96.03%	0.553	2.490
0.875	96.72%	0.662	2.600
0.9375	96.77%	0.789	2.727
1	96.80%	0.803	2.741

OCM: On-Chip Memory

4.2.2. TPVD without NLLR criterion

Candidates' selection in the second pass significantly affects the performance of this system. To test the efficiency of NLLR criterion, an experiment with static candidate number is done at first. When the candidate number is predefined from 1 to 15, the corresponding recognition accuracies are listed in figure 2. The two sets of models have been described in section 4.1. The scale of vocabulary is 600 names.

4.2.3. TPVD with NLLR criterion

The TPVD algorithm with NLLR criterion is employed in a mandarin 600-name dialing system. The recognition results reference to figure 2 and table 2.

Tab2. Recognition results of two-pass decoding with NLLR

Th	Accuracy in the first pass	Final accuracy	Average candidate number
0	90.60%	90.60%	1.07
0.25	96.22%	95.23%	1.41
0.50	98.37%	96.45%	2.08
0.75	99.20%	96.72%	3.46
1.00	99.50%	96.80%	6.19
1.25	99.77%	96.80%	11.03

Th is the threshold of NLLR criterion.

Compared with the experimental results of TPVD without NLLR criterion, this algorithm can always attain higher performance at the same average candidate number. When $Th = 1.00$ and the average candidate number is 6.19, the final accuracy of this algorithm is equal to the recognition rate with Viterbi full search. But for the algorithm without NLLR criterion, such accuracy can be achieved when the static candidate number is 15.

Thus, the threshold of NLLR is set to 1.00 in our system. The on-chip memory demand is 100kByte and the average users' waiting time is 0.17X real time. Since the upper limit to candidate number is set to 15, the longest users' waiting time is about 0.4X real time.

4.2.4. The consistency of Th

Since the value of Th is set according to the results of 600-name speech recognition, this experiment will prove its consistency. TPVD with $Th = 1.00$ is invoked in 600-name, 200-name, 100-name, 50-name, 20-name and 10-name applications. The recognition results are shown as table 3.

Tab3. The consistency of Th

N-name	Accuracy with TVFS	Accuracy with TPVD	Average candidate number
600	96.80%	96.80%	6.19
200	97.72%	97.72%	4.43
100	98.58%	98.58%	2.86
50	99.15%	99.15%	1.92
20	99.58%	99.58%	1.43
10	99.78%	99.78%	1.19

TVFS: Traditional Viterbi decoding with full search
TPVD: Two-pass Viterbi decoding with NLLR criterion

From the table above we find that the recognition accuracy of two-pass decoding with NLLR criterion always equals to the accuracy of traditional Viterbi decoding with full search, whatever the vocabulary scale is. The average candidate number in the second pass reduces as the vocabulary scale decreases, which is just as we expect.

5. CONCLUSION

This paper describes the two-pass Viterbi decoding algorithm designed for CDHMM-based large-vocabulary name dialing and command controlling applications on fixed-point DSP.

Through TPVD algorithm, the whole recognition process is divided into a large-vocabulary coarse matching in real time and a small-vocabulary fine matching in non-real time. While the hardware resources are limited, the TPVD is much more efficient than traditional Viterbi beam search.

These two algorithms are compared in an embedded mandarin 600-name dialing system. When on-chip memory is about 100kByte, TPVD can achieve 96.80% accuracy with the average users' waiting time of 0.17X real time, while traditional Viterbi beam search spends 1.74X real time waiting time to attain the same accuracy.

6. ACKNOWLEDGEMENT

It is appreciated that Infineon Corp. supplies the chips and developing tools of Unispeech 80D51.

The research is supported by National Natural Science Funds of China (Item No. 69975007) and National High Technology R&D Program (863) of China (863-306ZD13-04-6).

7. REFERENCE

[1] Stefan Dobler, "Speech Recognition Technology for Mobile Phones," *Ericsson Review*, Vol.77, No.3, pp.148-155, 2000.

[2] M. Shozakai, "Speech Interface VLSI for Car Application," In *Proceedings of ICASSP*, Vol.1, pp.141-144, 1999.

[3] B. Burchard, R. Romer, and O. Fox, "A Single Chip Phoneme Based HMM Speech Recognition System for Consumer Applications," *IEEE Trans. On Consumer Electronics*, Vol.46, No.3, pp.914-919, Aug. 2000.

[4] X. Menendez-Pidal, L. Duan, J.W. Lu, et al, "Efficient Phone Based Recognition Engines for Chinese and English Isolated Command Applications," In *Proceedings of International Symposium on Chinese Spoken Language Processing*, Vol.1, pp.83-86, Aug. 2002.

[5] J. Zhao, J. Hamaker, N. Deshmukh, A. Ganapathiraju, J. Picone, "Fast Search Algorithms for Continuous Speech Recognition," In *Proceedings of IEEE SOUTHEASTCON*, pp.36-39, 1999.

[6] L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[7] M.J. Gardner, D.G. Altman, *Fundamentals of Speech Recognition Statistics with Confidence: Confidence Intervals and Statistical Guidelines*, British Medical Journal, London, 1989.

[8] X. Zhu, R. Wang, Y.N. Chen, et al, "Acoustic Model Comparison for an Embedded Phoneme-based Mandarin Name Dialing System," In *Proceedings of International Symposium on Chinese Spoken Language Processing*, Vol.1, pp.9-12, Aug. 2002.

[9] X. Zhu, Y.N. Chen, J. Liu, et al, "Feature Selection in Mandarin Large Vocabulary Continuous Speech Recognition," In *Proceedings of International Conference on Signal Processing*, Vol.1, pp.508-511, Aug. 2002.