AN IMPROVED STEADY SEGMENT BASED DECODING ALGORITHM BY USING RESPONSE PROBABILITY FOR LVCSR

Zhanlei Yang, Wenju Liu, Hao Chao

National Laboratory of Pattern Recognition (NLPR), Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100190
{zhanlei.yang, lwj, hchao}@nlpr.ia.ac.cn

ABSTRACT

This paper proposes a novel decoding algorithm by integrating both steady speech segments and observations' location information into conventional path extension framework. First, speech segments which possess stable spectrum are extracted. Second, a preliminarily improved algorithm is given by modifying traditional inter-HMM extension framework using the detected steady segments. Then, at probability calculation stage, response probability (RP), which represents location information of observations within acoustic feature space, is further incorporated into decoding. Thus, RP directs the decoder to enhance/weaken path candidates that get through the front end steady-segment-based decoding. Experiments conducted on Mandarin speech recognition show that character error rate of proposed algorithm achieves a 4.6% relative reduction when compared with a system in which only steady segment is used, and run time factor achieves a 10.0% relative reduction when compared with a system in which only RP is used.

Index Terms— *Decoding algorithm, path extension, steady segment, response probability, probability fusion*

1. INTRODUCTION

Frame based acoustic modeling method focuses on a short span of time, which leads to tractable and efficient implementations while restricts the usage of segmental level information on the other hand. In contrast, segment based modeling method outperforms the popular frame based modeling method while often introduces a higher computing complexity [1], [2]. Fortunately, the properties of the two modeling approaches are complementary. Hon et al. [3] developed a framework by combining frame based and segment based acoustic modeling method. Experiments carried out on LVCSR task show that incorporated system achieves better performance than individual acoustic modeling method. Similar work has also been done in [4], in which segmental level information is incorporated by modifying conventional path extension framework instead of constructing segmental level acoustic models.

Besides modeling aspects, difficulties also exist at decoding stage, which would further reduce the performance of ASR systems. One example is that a decoder often encounters confusions when distinct acoustic models obtain similar likelihoods on some particular speech observations. In this situation, the decoder becomes unable to distinguish these models thus error happens. Although beam search or N-best search can be employed to extend more than one hypothesis, both of the two algorithms are suboptimal, which cannot guarantee to find the best state sequence [5][6][7].

To obtain accurate results, many decoders adopt several decoding passes. This approach utilizes utterance verification techniques to estimate reliabilities of decisions made at an earlier decoding stages, by calculating some scores, such as confidence measures [8][9]. Besides, Sherif et al. [10] introduce a posterior probability-based confidence measure to provide guidance for the recognizer to extend the most promising paths. Their experiments show that if the confidence scores are employed in the process of decoding, it will perform better than the conventional search approach. In addition, work [11] proposes a response probability (RP) to distinguish path candidates when traditional acoustic and linguistic likelihoods become unable to precisely discriminate them in local search spaces [12]. The novel probability, RP, which represents the location information of frames within the acoustic space, is integrated into the decoding metric to enhance or weaken existing paths.

In this paper, an improved steady segment based decoding algorithm is proposed by integrating RP into path probability calculation. In the novel decoding framework, steady segments, speech pieces without remarkable change in spectrum, are detected and then utilized to modify conventional decoding metric by deleting inter-HMM extensions. Then, the proposed algorithm further modifies the probability calculation of decoder. For paths that get through the front end steady-segment-based decoding, their total probabilities are supplemented with RPs on the basis of acoustic and linguistic probabilities. Thus, the decoder is induced to search the most promising candidates according to both the steady status and the location information of frames. Consequently, this improved decoding algorithm is expected to be able to take advantage of segmental level knowledge to supplement the shortage of frame based modeling method, and local level information to avoid error pruning when traditional acoustic and linguistic likelihoods become unable to distinguish path candidates in a local area of acoustic feature space.

The rest of this paper is organized as follows. Section 2 and Section 3 respectively give a brief introduction to steady segment based decoding algorithm and response probability based decoding algorithm. In Section 4, an improved steady segment based decoding algorithm is proposed by incorporating response probability. Section 5 shows the experimental results and then gives analyses. Conclusions are drawn in the last section.

2. STEADY SEGMENT BASED DECODING ALGORITHM

This section briefly introduces steady segment based decoding framework. First, steady segment is extracted. Then, the decoding

978-1-4673-2507-3/12/$31.00 © 2012 IEEE

framework is described and two distinct path extension modes are explained. The last subsection integrates detected steady segments into decoding.

2.1. Steady Segment Extraction

Steady speech segment is a speech piece without remarkable change in spectrum. To extract steady segments from continuous speech, three "actions" are needed: spectrum calculation, unsteady landmark detection, and steady segment locating [13]. For the spectrum calculation, several frequency bands are used for spectrogram division, and the energy of each band is calculated. For any a band, if its energy change sharply within a short period of time, corresponding frames are gathered to form unsteady segments. And the left segments are steady segments we are looking for. Detailed descriptions can be seen in [13].

2.2. Intra-HMM and inter-HMM extension

There are two distinct token spread forms at decoding stage: intra-HMM extension and inter-HMM extension. Intra-HMM extension means that paths extend inside HMMs, just between individual states. This kind of extension is constrained by transition matrix of HMM, which is calculated using training data with a given prototype (usually diagonal form). In contrast, inter-HMM extension is constrained by a higher level knowledge, such as the node position of lexicon tree that current model stays at, language model knowledge and so on. Inter-HMM extension happens at the last state of a HMM. For a certain token, if it stays in the last state of a HMM at current frame, at the following frame, this token will move out of current HMM and enter into a new HMM. All token information, such as survived paths, acoustic score, word sequence, is copied to the new HMM. The moment inter-HMM extension finished, paths extend forward to new acoustic models.

Before using steady segments, paths are extended in a normal way: first spread between states of a HMM until the last emitting state, and then enter the first state of a following HMM, followed by an interior HMM extension again. This process repeats until all frames of an utterance are recognized.

2.3. Steady Segment Based Decoding Algorithm

Under instruction of detected steady segments, original framework is modified by removing part of inter-HMM extensions. At steady frames, extension between HMMs is forbidden, and paths have to stay around. It means that at steady frames paths are only permitted to extend to states they are staying at. This is different from original extension mode in which paths can step across HMMs without any restriction. The reason why this kind of extension can be removed at steady frames can be explained as follows.

Steady frames, which are defined as pieces without remarkable change in spectrum, are stable segments of speech articulation. On the other hand, extension between individual HMMs indicates that both acoustic and articulatory parameters are changing significantly. Therefore, steady segment conflicts with inter-HMM extension and the coexistence of them cannot happen. In general, steady segment provide a new constrain on path extension at a long time interval level.

3. RESPONSE PROBABILITY BASED DECODING ALGORITHM

Traditional decoders do not directly take advantage of location information of frames when extending path candidates. In fact, for a speech observation, it is always located in a small region of the whole acoustic feature space. Consequently, it is expected to utilize the unique location information of this observation to improve the decoding, aiming at reducing pruning errors.

RP is employed to represent the unique location information of an observation [11]. RP based decoding algorithm can be decomposed into two parts: RP modeling module and RP integration module, which are described respectively in following subsections.

3.1. Response probability model

This subsection explains how to build a RP model. As is known, the entire acoustic space is composed of plenty of speech data, including all phones in phone set. What is more, this acoustic space is often modeled by using UBM, especially GMM based UBM [14]. Thus, there is a relationship between Gaussian components of UBM and phones: each Gaussian component is good at depicting a class of phones which possess unique similar property; in the meanwhile, data of a phone can obtain larger probabilities on particular Gaussian components than on others. To describe this relationship, a RP model is constructed to indicate whether a component of UBM has a descriptive ability to a phone and how strong it is.

Assume that O is an observation of phone q_O. We first calculate O's principal Gaussian component (PGC) m_O on UBM as follows:

$$\begin{aligned} m_O &= \arg\max_{m'} P(O \mid \lambda_{m'}) \\ &= \arg\max_{m'} N(O;\mu_{m'},\Sigma_{m'}) \end{aligned} \qquad (1)$$

where $\lambda_{m'}$ is a component with a mean $\mu_{m'}$ and a variance $\Sigma_{m'}$. The RP of a Phone-Gaussian component pair *(q, m)* is defined as:

$$P(q,m) \triangleq \frac{\sum_O I(O,q,m)}{\sum_O I(O,q)} \qquad (2)$$

where *I(·)* is a set of indicator functions. *I(O, q, m)* equals to 1 if $q_O=q$, $m_O=m$ and to 0 otherwise, whereas *I(O, q)* equals to 1 if $q_O=q$ and to 0 otherwise.

In (2), the numerator represents the number of frames who belong to *q* and take *m* as PGC, while the denominator represents the number of frames who belong to *q*. These statistics are gathered from the whole training database. The probability *P(q, m)* represents that for *q*, how much part of its data are responded by *m*. For example, if *P(q, m)*=0.3, it means that 30% of *q*'s data take *m* as PGC.

The RP of every Phone-Gaussian component pair is shown:

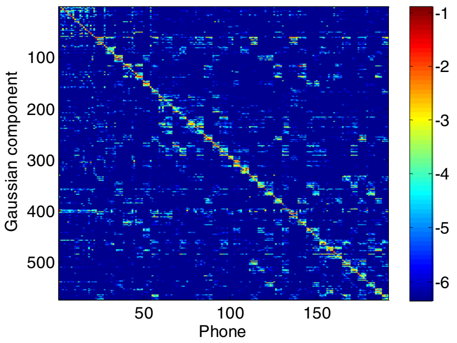

Fig. 1. Response probabilities of Phone-Gaussian pairs

In this figure, pairs are ranked: pairs whose RPs are large are place on the diagonal. As is shown in the figure, Gaussian components are good at depicting some particular phones. This relation is going to be used when recognizing a frame at its corresponding local acoustic space.

3.2. RP Integration

This subsection explains how RP model works together with acoustic model and language model. In order to take advantage of the RP model, *P(q|O)* of *O* is defined:

$$P(q \mid O) \triangleq \sum_{m} P(q,m) I(O,m) \tag{3}$$

in which *I(O, m)* equals to 1 if $m_O = m$ and to 0 otherwise. This probability reflects the location of *O* within the whole acoustic space, and is going to be fused at the probability calculation stage. For the purpose of convenience, the logarithmic form of probability is used.

Conventional decoder needs to calculate acoustic probability and linguistic probability for path candidates. Then, these candidates compete with each other according to a weighted sum of the two probabilities, which can be formulated as follows:

$$P(t) = P(t-1) + \alpha_1 P_{am} + \alpha_2 P_{lm} \tag{4}$$

where *P(t)* is the total probability at frame t. *P(t-1)* denotes history probability from 0 to t-1. P_{am} and P_{lm} are respectively probabilities of acoustic model and language model, and α_1 and α_2 are their weights.

In RP based decoding algorithm, *P(q|O)* is integrated into acoustic probability and linguistic probability. Hence, the previous formulation can be rewritten as:

$$P(t) = P(t-1) + \alpha_1 P_{am} + \alpha_2 P_{lm} + \alpha_3 P(q \mid O) \tag{5}$$

This modified probability is used for pruning when searching for optimal paths.

4. IMPROVED STEADY SEGMENT BASED DECODING BY INTEGRATING RESPONSE PROBABILIYT

This section gives an improved steady segment based decoding algorithm by integrating RP into acoustic and linguistic probabilities for path candidates.

For both steady segment based decoding algorithm mentioned in Section 2 and RP based decoding algorithm mentioned in Section 3, the total probability of path candidates can be represented in an unified framework:

$$P(t) = P(t-1) + \alpha_1 P_{am} + \alpha_2 P_{lm} + \alpha_3 P_{add} \tag{6}$$

in which P_{add} is an additional probability beside P_{am} and P_{lm}. For conventional steady segment based decoding algorithm, P_{add} equals to $-\infty$ if current frame O_t is steady and inter-HMM extension is conducted, and to 0 otherwise. For the RP based decoding algorithm, P_{add}=*P(q|O)*.

The improved steady segment based decoding algorithm is also represented in this unified framework, with a P_{add} who has a different meaning from that in conventional two algorithms. In this novel algorithm, P_{add} equals to $-\infty$ if current frame O_t is steady and inter-HMM extension is conducted, and to *P(q|O)* otherwise. Thus, all path candidates will extend according to the following rules.

First, for an unsteady observation, candidates are permitted to extend in both "intra-HMM" and "inter-HMM" modes. When calculating total probability, RP is integrated into conventional acoustic probability and linguistic probability for every path. Thus, the obtained probability contains location information of *O*, which induces the decoder to search within an interested subspace of the whole space.

Second, for a steady observation, candidates are only permitted to extend in the "intra-HMM" mode, while "inter-HMM" mode is limited. This is the same as is conducted in the conventional steady segment based decoding algorithm. However, for path candidates that get through this restriction, RP is further employed for probability fusion. This is different from the conventional algorithm which only takes advantage of acoustic probability and linguistic probability for probability fusion.

Compared with conventional RP based algorithm, the novel algorithm sets additional restrictions according to the steady condition of observations. This modification makes the decoder be able to take advantage of segmental level knowledge without constructing segmental level acoustic models. The improved decoding framework can be depicted in the following figure.

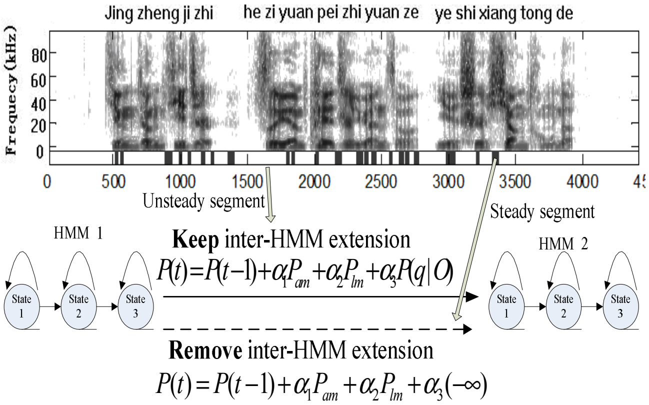
Jing zheng ji zhi
he zi yuan pei zhi yuan ze
ye shi xiang tong de
Frequency (kHz)
80
60
40
20
0
0
500
1000
1500
2000
2500
3000
3500
4000
45
Unsteady segment
Steady segment
HMM 1
HMM 2
State 1
State 2
State 3
Keep inter-HMM extension
$P(t)=P(t-1)+\alpha_1 P_{am}+\alpha_2 P_{lm}+\alpha_3 P(q|O)$
Remove inter-HMM extension
$P(t)=P(t-1)+\alpha_1 P_{am}+\alpha_2 P_{lm}+\alpha_3(-\infty)$

Fig. 2. The improved steady segment based decoding algorithm

To sum up, the novel decoding algorithm follows the same rules as conventional steady segment based decoding algorithm at extension stage. However, at probability fusion stage, all permitted path candidates will incorporate RP into the total probability. This is different from conventional steady segment based decoding algorithm which incorporate nothing beside acoustic and linguistic probabilities. Thus, the novel algorithm becomes able to take advantage of location information of observations. This information directs the decoder to focus on interested acoustic subspaces, and to extend the most promising path candidates. On the other hand, compared with conventional RP based algorithm, the novel algorithm becomes able to take advantage of segmental level knowledge: by adding some restrictions at path extension stage, some worthless extensions are discarded. This helps to compensate the shortage of frame based acoustic modeling method.

5. EXPERIMENTS

This section will introduce experiments and analyze results.

5.1. Experimental setup and baseline

The data corpus applied in experiments is provided by Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development. 83 male speakers' data are employed for training (48373 sentences, 55.6 hours) and 6 male speakers' for test (240 sentences, 17.1 minutes). Acoustic features are 12 dimensions

MFCC plus 1 dimension normalized energy and their 1st and 2nd order derivatives. There are 191 phones in our Mandarin phone set, which is composed of syllable initials and toned syllable finals.

Continuous density left-to-right HMM contains 5 states, 3 of which are emitting states. Each emitting distribution is modelled by 16 Gaussian mixtures. There are 4575 tied states for the acoustic model. Context-dependent triphone and bigram language model with 48188 words are used, respectively.

A time-synchronous Viterbi decoding framework is constructed and used as baseline system [15][16]. This framework constructs search space by using dynamic lexicon tree copy method. In decoding process, search space is formed gradually according to positions that paths extend to. Besides, pruning strategy is employed in time to prevent rapid expansion of “active” paths. Finally, language model look-ahead technique proposed by [17] is utilized when tokens stay within a tree. With this technique, linguistic probability can be applied before word identities are confirmed.

5.2. Results and analysis

The experiments are carried out on four different systems: the baseline system (Baseline), conventional steady segment based decoding system (SS decoder), conventional RP based decoding system (RP decoder), and improved steady segment based decoder by integrating RP (SS+RP decoder). For the four systems, both Character Error Rate (CER) and Running Time Factor (RTF) are investigated on the same task. The following table gives results of the four decoders on the test set.

Table 1. System performance of four decoders.

System	System Performance	
	CER	RTF
Baseline	12.78%	1.40
SS decoder	11.73%	1.27
RP decoder	11.61%	1.41
SS+RP decoder	11.19%	1.26

It can be seen from the table that the performance of the SS+RP decoder has a significant improvement compared with Baseline, and a moderate improvement compared with SS decoder and RP decoder. Comparing the CER of SS decoder and SS+RP decoder, it can be seen that when RP is incorporated, the CER is relatively reduced by 4.6%. This is achieved by paying more attention to search within subspaces where observations locate, and discriminatively enhancing or weakening path candidates according to corresponding RPs. Comparing the RTF of RP decoder and SS+RP decoder, it can be seen that when “inter-HMM” is properly limited, the RTF has a 10% relative reduction. This improvement shows that when segmental level knowledge is employed, the decoder will be directed to selectively delete some “inter-HMM” extensions according to the steady conditions of observations. Thus, impossible extensions are omitted and the decoding time is saved.

6. CONCLUSION

This paper proposes an improved steady segment based decoding algorithm by incorporating response probability into conventional probability calculation stage. In this novel framework, the decoder is first directed to extend path candidates in two distinct modes according to the steady conditions of observations. Then, for permitted extensions, response probability is incorporated into the total probability of every path candidate. When it is fused, the decoder is induced to search within interested subspaces, and to enhance or weaken path candidates discriminatively. Experiments compare CER and RTF for four algorithms: original decoding algorithm, conventional steady segment based decoding algorithm, conventional response probability based decoding algorithm, and the improved steady segment based decoding algorithm proposed in this paper. Experimental results show that both CER and RTF of the proposed algorithm are reduced when steady segment and response probability are properly incorporated into the decoding framework.

7. ACKNOWLEDGMENT

This work was supported in part by the China National Nature Science Foundation (No.91120303, No.90820303 and No.90820011), 863 China National High Technology Development Project (No.20060101Z4073, No.2006AA01Z194), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

8. REFERENCES

[1] Yun Tang, Hua Zhang, Wenju Liu, Bo Xu and Guo-Hong Ding, “One-pass Coarse-to-Fine Segmental Speech Decoding Algorithm”, *Proc. ICASSP*, 2006, Vol.1, pp.441-444.

[2] S. Gao, et al, ”Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR”, *ICSLP*, Beijing, 2000.

[3] Hsiao-Wuen Hon and Kuansan Wang, “Unified Frame and Segment Based Models for Automatic Speech Recognition”, in *Proc. IEEE ICASSP-2000*, Istanbul, Turkey, 2000.

[4] Zhanlei Yang, Wenju Liu, “A Novel Path Extension Framework Using Steady Segment Detection for Mandarin Speech Recognition”, in *Proceedings of Interspeech2010*, Makuhari, Japan, 2010, pp. 226-229.

[5] Xavier L Aubert, “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition”, *Computer Speech and Language*, vol. 16, no. 1, pp. 89-114, Jan. 2002.

[6] Haeb-Umbach, R. Ney, H., Improvements in beam search for 10000-word continuous-speech recognition，*IEEE Trans. Speech and Audio Processing*, 1994，Vol. 2, pp. 353-356.

[7] B. H. Tran, F. Seide, and V. Steinbiss, "A word graph based N-best search in continuous speech recognition", *Proc. ICSLP'96*, pp. 2127 – 2130.

[8] V. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition", *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 288 - 298, 2001.

[9] H. Jiang, "Confidence measures for speech recognition: A survey", *Speech Commun.*, vol. 45, pp. 455 - 470, 2005.

[10] Abdou, S. and Scordilis, M.S., “Beam search pruning in speech recognition using a posterior-based confidence measure”, *Speech Communication*, Vol. 42, pp. 409-428, 2004.

[11] Zhanlei Yang, Hao Chao, Wenju Liu, “Response Probability Based Decoding Algorithm for Large Vocabulary Continuous Speech Recognition”, in *Proceedings of Interspeech2011*, Florence, Italy, pp. 1929-1932, 2011.

[12] Demuynck, K., Duchateau, J., Van Compernolle, D., Wambacq, P., “An efficient search space representation for large vocabulary continuous speech recognition”, *Speech Commun*. 30(1), 37–53 (2000).

[13] Hua Zhang, Yun Tang, Wenju Liu, and Bo Xu, “Unvoiced Landmark Detection for Segment-based Mandarin Continuous Speech Recognition”, in *Proc. ISCSLP*, Singapore, 2006, Vol.2, pp.374-383.

[14] D. Povey, S. M. Chu, and B. Varadarajan, “Universal background model based speech recognition,” in *Proc. ICASSP*, 2008, pp.4561-4564.

[15] H. Ney and S. Ortmanns, "Progress in dynamic programming search for LVCSR", in *Proc. IEEE*, vol. 88, pp. 1224 - 1240, 2000.

[16] S. Young et al. The HTK Book for version 3.4.1, Cambridge, 2000.

[17] S. Ortmanns, A. Eiden, H. Ney, N. Coenen, “Look-Ahead Techniques for Fast Beam Search,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1783-1786, Munich, Germany, April 1997.