Response Probability Based Decoding Algorithm for Large Vocabulary Continuous Speech Recognition

Zhanlei Yang, Hao Chao, Wenju Liu

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190

{zhanlei.yang, hchao, lwj}@nlpr.ia.ac.cn

Abstract

Acoustic space is made up of phonemes, and it can be modeled using universal background model (UBM). Therefore, there are some relations between the phonemes and Gaussian mixture components of the UBM. This paper represents these relations by proposing a response probability (RP) model, which describes the location information of speech observations within the whole acoustic space. At decoding stage, proposed RP model is fused with traditional acoustic model (AM) and language model (LM). After integrating RP, the decoder is guided to weaken or enhance different path candidates respectively and directed to extend the most promising paths. Experiments conducted on Mandarin broadcasting speech show that character error rate is relatively reduced by 9.15% when RP model is used and by 11.89% when an improved RP model is used.

Index Terms: speech recognition, decoding algorithm, pruning, response probability

1. Introduction

Conventional Large Vocabulary Continuous Speech Recognition (LVCSR) systems often encounter confusions when searching for optimal paths. Some of these confusions come from the uncertainty of the recognized hypotheses. Actually, because several AMs often obtain similar likelihoods on some speech observations, 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 can not guarantee to find the best state sequence [1][2][3].

To obtain accurate results, many decoders adopt several decoding passes. This approach utilizes utterance verification techniques to estimate the reliabilities of decisions made at an earlier decoding stages, by calculating some scores, such as confidence measures [4][5]. Then, the scores are applied to the final output of the decoder rather than being incorporated with the AM and LM probabilities during the decoding process.

In fact, if these scores are early incorporated, the decoder will be directed to extend the most promising paths, which may finally reduce the error rates. Sherif et al. [6] introduce a posterior probability-based confidence measure to provide guidance for the recognizer. Their experiments show that if the confidence scores are employed in the process of decoding, it will perform better than the conventional search approach.

Instead of the confidence measure, this paper proposes a Response Probability (RP) to distinguish path candidates when traditional AM and LM likelihoods become unable to precisely discriminate them in local search spaces [7]. The novel probability, RP, represents the location information of frames within the acoustic space. We integrate this probability at the decoding stage to enhance or weaken existing paths. Thus, the proposed method takes advantage of an auxiliary model to calculate a probability for every candidate besides AM and LM.

The rest of this paper is organized as follows. In Section 2, we show how to build RP model as well as improved RP model using UBM, and explain how to use the RP. Section 3 carries out experiments on Mandarin speech. In this section, the RP model is utilized for probability fusion. Moreover, we alter the weight of the RP to investigate the impact of the RP on system performance. In Section 4, we use an example to explain how the RP model works together with AM and LM and why error rate can be reduced when the RP model is employed. Finally, Section 5 concludes this paper.

2. Response probability model

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 acoustic space. Consequently, it is expected to utilize the unique location information of this observation to improve the decoding, aiming at reducing pruning errors.

2.1. Response probability model

The entire acoustic space is considered to be composed of plenty of speech data, including all phones of the phone set. Besides, this acoustic space is often modeled using UBM, especially GMM based UBM [8]. Thus, there are some relations 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 probability on some Gaussian components than on others. This paper uses a RP model to describe the relations between the phones and the components. RP indicates 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 the UBM as follows:

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

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:

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

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

In Formula (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. $P(q, m)$ 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.

In order to take advantage of the RP model, we define $P(q|O)$ for frame O as follows:

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

in which $I(O, m)$ equals to 1 if $m_O=m$ and to 0 otherwise. This likelihood contains the location information of O within the whole acoustic space, and is going to be fused at the following decoding stage. For the purpose of convenience, we use logarithmic form of the likelihood in this paper.

Conventional decoder needs to calculate AM and LM probabilities for path candidates. Then, these candidates compete with each other according to the 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 the history probability from 0 to t-1. P_{am} and P_{lm} are respectively the AM and LM probability, and α_1 and α_2 are their weights.

In our framework, $P(q|O)$ will be integrated into the conventional AM and LM probabilities. Hence, the previous formulation is 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.

2.2. Improved RP model

Sometimes, only computing one PGC for each frame is unable to exactly determine the location of current frame within the acoustic space. In fact, for a frame O, it may have high probability on several Gaussian components of UBM. Therefore, we further define several PGCs for each frame to precisely describe its location. Multiple PGCs are calculated using a method like Beam algorithm: for a frame, components whose probabilities are larger than a lower bound are defined as PGCs. This lower bound is calculated by subtracting a constant beam value from the maximum probability. Then, these PGCs are ranked according to their probabilities, named 1^{st}, 2^{nd},…k^{th}…K^{th} PGC, respectively. We further mark these components as m_{O1}, m_{O2}, … m_{OK}, where K is the total PGC number of O. The improved RP is given as follows.

$$P(q,m,k) \triangleq \frac{\sum_{O} I(O,q,m,k)}{\sum_{O} I(O,q)} \tag{6}$$

in which $I(O, q, m, k)$ equals to 1 if $q_O=q$ and $m_{Ok}=m$ and to 0 otherwise.

When decoding, $P(q|O)$ is redefined:

$$P(q \mid O) \triangleq \sum_{m} \sum_{k} P(q,m,k) I(O,m,k) \tag{7}$$

in which $I(O, m, k)$ equals to 1 if $m_{Ok}=m$ and to 0 otherwise.

In the formulas above, $m_{Ok}=m$ means that O obtains the k^{th} highest probability on m within all components of UBM, that is, m is the k^{th} PGC.

3. Experiments

In this section, we carry out experiments and analyze the results. First of all, we give a brief introduction to the experimental setup as well as the baseline system. Then, the RP model is built by making use of an UBM. In the following subsection, we adjust the weight of RP to investigate the impact of RP on system error rate. Finally, we give a further experiment to study whether an improved RP model, which defines more than one PGC at every frame, is helpful for further decreasing the error rate.

3.1. 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 modeled by 16 Gaussian mixtures. Context-dependent triphone and bigram LM with 48188 words are used as acoustic model and language model, respectively.

We build a time-synchronous Viterbi decoding framework for baseline [9]. This framework constructs search space by using dynamic lexicon tree copy method. In this decoding process, the search space is formed gradually according to position that path extends to. Besides, pruning strategy is employed in time to prevent rapid expansion of "active" paths. Finally, LM look-ahead technique proposed by [10] is utilized when tokens stay within a tree. With this technique, LM likelihood can be applied before word identities are confirmed.

The baseline system achieves 12.78% Character Error Rate (CER) for the test set.

3.2. Build RP model

This subsection first builds a UBM, and then uses it to gather statistics for phone-Gaussian component pairs. Besides, every frame of data corpus should be assigned to a certain phone. The whole RP modeling process can be divided into the following three steps.

- Train a UBM. In order to avoid the unbalance of amount of data belonging to individual phones, we first model every phone using a Gaussian distribution, and then pool the Gaussians together. The pooled GMM is further trained using EM algorithm to form the UBM, which totally contains M=573 mixtures in our experiment.
- Assign a phone to every frame. This paper uses the acoustic model of baseline to align the training database [11]. Thus, every frame is labeled with a certain phone.
- Gather the statistics. By now, every frame has been tagged with a phone. Further, we calculate the PGC of every frame according to equation (1). When all frames have calculated their PGCs, we can obtain RPs of every Phone-Gaussian pair according to (2).

In fact, lots of Phone-Gaussian pairs do not coexist frequently. Thus, corresponding RPs are very small or even zero. This results in directly pruning without considering AM and LM probabilities. To avoid this pruning, we set a threshold T=1/M to replace the RPs which are smaller than T, where M is the number of mixtures of UBM mentioned above. The calculated RPs are shown in the following figure.

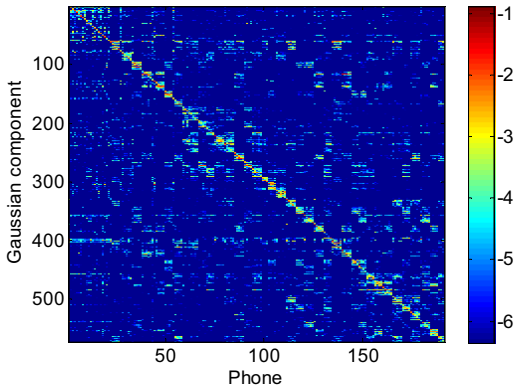

Figure 1: *Response probabilities of Phone-Gaussian pairs.*

As mentioned above, there are 191 phones and 573 Gaussian components in all. In this figure, pairs are ranked: pairs whose RPs are large are place on the diagonal. As is shown, the 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.

3.3. RP integration

Three different probabilities, AM probability, LM probability and RP are fused for decoding according to equation (5). The weighted sum of the three probabilities is considered as the total probability of a path, and then employed to compete with other path candidates at extending and pruning stages. Here, we explore the impact of RP on CER.

First, AM probability and LM probability are balanced according to (4) by adjusting α_1 and α_2, aiming at the lowest CER of the baseline. Then we fix α_1 and α_2 and adjust α_3. The following curve gives the CER when α_3 is altered.

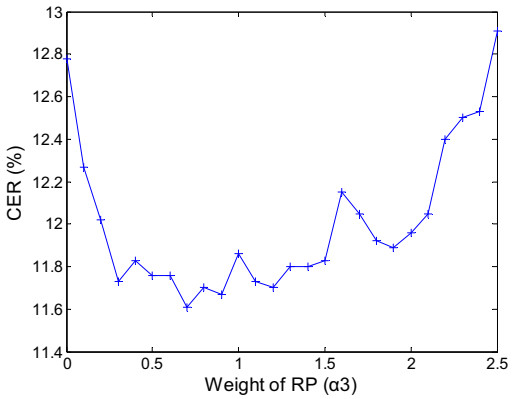

Figure 2: *Impact of RP on CER.*

It can be seen from the figure that when α_3=0, RP does not work, and the CER achieves 12.78%, which is equal to that of the baseline. As the increase of α_3, the CER is sharply reduced at the region of 0~0.3. Then, the CER keeps stable when α_3 ranges from 0.3 to 1.5, and achieves 11.61% at 0.7. Compared with the baseline, the CER has a 9.15% relative reduction. However, further increase α_3 is no longer helpful for reducing the CER. This can be analyzed as follows.

RP helps to decrease the CER when it is properly integrated. When decoding, RP coarsely classifies a frame *O* and tells the decoder which phones *O* most likely belongs to, and which phones it unlikely belongs to. For path candidates, if they are extending to or staying within the most likely phones of *O*, they will be enhanced. Otherwise, if these paths are extending to or staying within the phones that *O* unlikely belongs to, they will be weakened. Thus, the RP guides the decoder to pay more attention to the subspace that *O* lies in. As a result, path candidates are discriminatively strengthened or weakened and the CER is decreased.

However, the CER does not continue to decline when the RP is over-weighted. In fact, current frame may have large probabilities on several phones. Thus, the RP becomes unable to distinguish these phones precisely. In contrast, AM may be able to accurately discriminate these phones because it owns a much larger scale of parameters. In other words, the over-weighted RP would weaken the impacts of the AM and LM, and then excessively disturb the extension and pruning. Consequently, the CER rises when the RP is over-weighted.

3.4. Improved RP integration

The RP model mentioned above defines only 1 PGC at every frame. In fact, a frame may obtain high likelihoods on several Gaussian components when current frame is located in a shared acoustic region of several phones. Moreover, speech data of a certain phone are distributed in a broad acoustic space, thus only 1 PGC may not be able to describe this distribution precisely. Therefore, we need to define several PGCs at every frame. This improved RP model becomes able to describe Phone-Gaussian pairs more precisely. We build this improved model according to equation (6), and then employ it at either modeling or decoding stage, or both of them. Corresponding results are listed in Table 1.

Table 1. *CER comparison for several systems.*

	Modeling	Decoding	CER
Baseline	--	--	12.78%
System 1	1	1	11.61%
System 2	>1	1	11.38%
System 3	1	>1	11.35%
System 4	>1	>1	11.26%

System 1 has been described in the previous subsection, which defines only 1 PGC at either modeling or decoding stage. System 2~4 respectively define more than one PGCs at modeling stage, decoding stage, and both of the two stages.

It can be seen from the table that when we enlarge the number of PGCs, the CER is further decreased to 11.26% when multiple PGCs are used at both modeling and decoding stage. Compared with the baseline, the CER has a 1.52% absolute reduction and an 11.89% relative reduction.

4. Pruning analysis

This section will deeply analyze why the proposed model is helpful, by back tracing two optimal paths respectively derived from the baseline and System 1. We will see how the RP reduces pruning errors when AM and LM can not work well at local search space.

An utterance from the test set is selected as an example. Transcriptions for this utterance are given as follows:

Label: 外商对此也没有信心。
Trans.1: 换上对斯也没有信心。
Trans.2: 外商对斯也没有信心。

"Label" is exactly transcribed by human, whereas Trans.1 and Trans.2 are obtained by the baseline and System 1, respectively. In this experiment, we will back trace Trans.1 and Trans.2, record variations of their probabilities, and discuss the impact of the RP on pruning. The traced results are shown in the following figure.

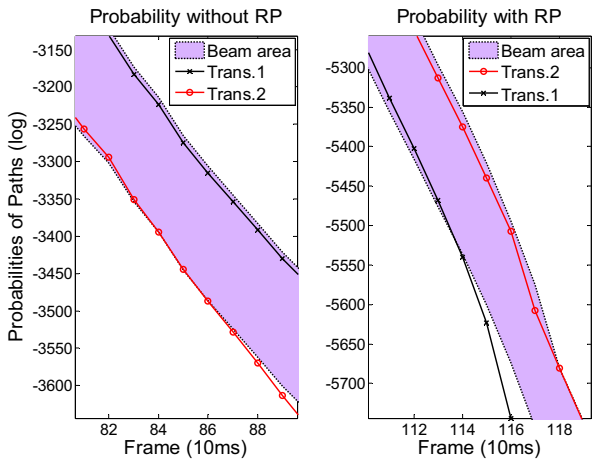

Figure 3: *Transcription probabilities without/with RP.*

The Beam area in the figure means variation range of the probabilities, which is bounded by an upper boundary and a lower boundary marked as dashed line. Paths whose probabilities fall into this area will be reserved and then extended in the following frame. Otherwise, if their probabilities are smaller than the lower bound, these paths will be pruned. In our experiments, Beam=180.0. The left and right half graphs give the probabilities of Trans.1 and Trans.2 without and with the RP, respectively. For each transcription, either integrating the RP or not does not affect the calculation of AM and LM probabilities.

In the left graph, paths compete against each other only using the AM and LM probabilities, whereas in the right graph, the RP is fused with the two probabilities for the competition. Before the fusion, Trans.1 is better than 2 according to their probabilities. At t=84, Trans.2 is pruned because its probability is smaller than the lower bound of beam area. At this frame, the difference between the probabilities of the two transcriptions is P_{trans1}-P_{trans2}=168.00. After integrating the RP, as shown in the right figure, Trans.2 has a higher total probability than 1 and the latter is pruned at t=115. At this frame, P_{trans2}-P_{trans1}=165.10. Competitive situation of the two transcriptions is completely changed after the RP is used.

After recording the probability at every frame, impacts of RP upon pruning can be interpreted. When the RP is integrated, at t=0~84, Trans.2 accumulates a larger RP than Trans.1, which keeps Trans.2 from being pruned. Experiments show that during t=0~84, the RP accumulated by Trans.1 and Trans.2 are -257.44 and -114.28, respectively. Advantage in RP makes Trans.2 gradually catch up Trans.1. At t=84, the difference between the total probabilities of the two transcription is P_{trans1}-P_{trans2}=24.84. During t=85~115, Trans.2 becomes superior to Trans.1 by further accumulating the RP. At t=115, the RP is -355.88 and -185.01 for the two transcriptions. Because of a lower RP accumulation during t=0~115, Trans.1 is finally pruned at t=115. During t=0~115, traditional likelihoods of the two transcriptions are respectively punished by a probability of -355.88 and -185.01. Relatively speaking, Trans.1 is weakened and Trans.2 is enhanced, which improves system performance finally.

5. Conclusion

This paper proposes a response probability model, which describes relations between phones and Gaussian components of universal background model. The RP model is used to reduce pruning errors when AM and LM are unable to describe local frames precisely. At the decoding stage, the RP directs the decoder to discriminatively weaken or enhance path candidates by making use of the location information of frames. This paper further explores an improved RP model by defining more than one principal Gaussian component per frame. Experimental results show that the CER has a 9.15% relative reduction when the RP model is used and an 11.89% relative reduction when the improved RP model is used.

6. Acknowledgements

This work was supported in part by the China National Nature Science Foundation (No.60675026, 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).

7. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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).

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

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

[10] S. Ortmanns, A. Eiden, H. Ney, N. Coenen, "Look-Ahead Techniques for Fast Beam Search," Proc. IEEE Int. onf. On Acoustics, Speech and Signal Processing, pp. 1783-1786, Munich, Germany, April 1997.

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