A SYNCHRONOUS METHOD FOR AUTOMATIC SCORING OF LANGUAGE LEARNING

Bin Dong, Yonghong Yan

ThinkIT Speech Laboratory
Institute of Acoustics, Chinese Academy of Sciences, Beijing
{bdong, yyan}@hcl.ioa.ac.cn

ABSTRACT

In this paper, a synchronous method based on state graph is proposed to calculate the evaluation feature for automatic scoring in computer-assisted language learning (CALL). The posterior probabilities of states are selected as the main feature. The score of hypothesized phonemes and words are estimated using the information of corresponding states. Traditional systems use two passes and two different models for decoding and computing posterior probabilities respectively. In this new algorithm, the posterior probabilities are calculated during the decoding of the state graph constructed from grammar. And in this new algorithm, the same acoustics model is used during the process of decoding and posterior probabilities computing. The old and new computing algorithms are compared through experiments, and the result shows that performance of the new algorithm is effectively improved. The scoring accuracy of new synchronous algorithm is improved, while the computing complexity reduces 16%.

Index Terms—CALL, Scoring, Posterior probability

1. INTRODUCTION

With the wave of globalization, more and more people go abroad. There are more contacts among institutions and persons from different countries. As a mean of communication, language is the premise. More people begin to learn the second language amid the request from their work and life.

As a result, the traditional teaching/learning model in classroom and the subjective evaluation methods can no longer satisfy the growing needs. And classroom gives more restricts to modern people who lives in a fast-paced society. Now, speech technology makes it possible to learn language with computer. Learners can keep their learning plan flexible by themselves and not be constrained by fixed time and fixed place no longer.

Computer-assisted language learning is based on the speech technology especially on speech recognition technology. In general, the posterior probability of phoneme during decoding is selected as the main scoring feature except some super-segmental features [1]. There are two hypotheses

with the feature used that the acoustics model used in CALL system can be regarded as the model of the perfect pronunciations and the posterior probability can represent the accessibility of learner's pronunciation to standard pronunciation.

The calculation of posterior probability is the key in the CALL system as scoring feature. There are some calculating methods in references. In [2] and [3], the same acoustics model is used in the process of decoding and calculating it on the basis of word graph. In [4] and [5], it is calculated with same acoustics model based on the N-best list. A phoneme recognizer is used in [6] for it whose model is different from the model for decoding.

In this paper, a synchronous method based on state graph is introduced. Posterior probability is calculated based on runtime decoding of the state graph. During this process, the same acoustics model is used and only one pass decoding is needed. The performance of new algorithm is effectively improved, while the computing complexity reduces.

The rest of the paper is organized as following. Section 2 describes the feature for scoring. Section 3 describes the detailed method of calculating scoring feature. Section 4 describes the experimental results and Section 5 presents conclusions and future directions.

2. ESTIMATION OF SCORING

2.1 Feature of scoring

In this paper, generalized posterior probability of each state is used as the feature for automatic scoring.

The following notation is used in this paper:

$x_t, t = 1, 2, \dots, T$ is acoustics observation sequence for spoken utterance;

$W_i, i = 1, 2, \dots, N$ is the set of output word sequence from recognition system;

$PH_k, k = 1, 2, \dots, L$ is the set of phonemes sequence of corresponding words;

$S_j, j = 1, 2, \dots, M$ is the set of state sequence of corresponding phonemes;

$p(x_t | s_j)$ is the conditional probability of the t^{th} frame speech data on the conditional of the j^{th} state.

The posterior probability of each state which constitutes the Hidden Markov Model (HMM) of each phoneme is the geometric mean of the posterior probability of each frame.

For each frame of speech data $x_t, t = 1, 2, \dots, T$, the posterior probability $p(s_{m(t)} | x_t)$ of corresponding state is defined as:

$$p(s_{m(t)} | x_t) = \frac{p(x_t | s_{m(t)})p(s_{m(t)})}{p(x_t)} = \frac{p(x_t | s_{m(t)})p(s_{m(t)})}{\sum_{s_j \in D} p(x_t | s_j)p(s_j)} \quad (1)$$

Where, $m(t)$ is the state which x_t belongs to, D is the set of all states.

During decoding, the time information of duration at each state can be retrieved. In the other word, we can get the start frame and end frame of each state from decoded output, the generalized posterior probability of each state is defined as the posterior probability normalized by time information:

$$p(S_j) = \frac{1}{\tau_e - \tau_b + 1} \sum_{t=\tau_b}^{\tau_e} \log p(S_j | x_t) \quad (2)$$

Where, τ_b, τ_e is the start frame and end frame of the state respectively.

Each phoneme is denoted by a HMM. The generalized posterior probability of each phoneme can be calculated and be defined as the arithmetic mean of posterior probability of each state [7]:

$$p(PH) = \frac{1}{N} \sum_{j=1}^N \frac{1}{\tau_e[j] - \tau_b[j] + 1} \sum_{t=\tau_b[j]}^{\tau_e[j]} \log p(S_j | x_t) \quad (3)$$

Where, N is the number of state in each HMM.

2.2 Traditional two-pass method for calculating posterior probability

Two-pass decoder method is used extensively for calculating posterior probability. In general, the acoustics models used in the two-pass decoder are different.

Figure 1 Two-pass decoder for scoring feature

As Figure 1 shows, the input speech is recognized in the first pass decoding process with one set of acoustics model.

A word sequence or a word graph is obtained from the process. Then the posterior probability is calculated in the second pass decoding with another acoustics model. This acoustics model is called catch-all model. This model has various forms. One of the most widely used is the so-called all-phone network, in which all the possible phonetic and non-speech HMMs are connected to each other, and with which any word sequence can be recognized.

Figure 2 Two-pass decoder with different acoustics models

In this method [6], the numerator and denominator in Equation (1) are both evaluated in the second pass decoding. The second pass decoding is working on the basis of the word graph from the output of first pass decoding. The performance of this method lies on whether the information of word graph from the first pass decoding is abundant enough.

Another method [8] also used two different acoustics models in the two-pass decoding processes. But the calculating method for equation (1) is different from the method shown in Figure 1. As shown in Figure 2, in this method, the numerator in Equation (1) is evaluated in the first pass decoding and the denominator is evaluated in the second pass decoding process. Of course, if Equation (1) is used in this manner, the two acoustics models must be homologous. Otherwise, the confidence from Equation (1) is not usable completely.

3. ONE-PASS SYNCHRONOUS CALCULATING ALGORITHM

3.1 Construction of the state graphs

In this paper, grammar composed of learning contents is used in the CALL system. The search space is constituted by the task grammar. Firstly, a word net is constructed based on the task grammar. This word net is the search space in which decoder will find the best possible state sequence. Then, with the pronunciation dictionary used, the word net is extended to a phone net. Each node of word is replaced by corresponding phone sequence. In the system, each phone is denoted by a HMM mentioned as above. The state graph is constructed by replacing each phone of phone net with the HMM state from the acoustics model used for decoding and confidence measure. A node of the state graph represents a

state of a HMM. Any path going through the state graph forms an alternative sentence hypothesis or a word hypothesis that is based on contents of the task grammar used.

The final state graph is constructed by merging all nodes with identical context node into or from a single node. During the process, forward merging and backward merging are performed. In the forward merging, the nodes with the same predecessor token will be merged into one node and in the backward merging, the nodes with the same successor token will be merged into one node.

3.2 The proposed one-pass algorithm

In this paper, one new method of computing scoring feature is proposed in contrast to the traditional method, in which posterior probability is calculated based on the state graph with same acoustics model at decoder run time shown in Figure 3.

During calculation, the prior probabilities of all states are assumed to be the same. So Equation (1) is simplified as:

$$p(s_{m(t)} | x_t) = \frac{p(x_t | s_{m(t)})}{\sum_{s_j \in D} p(x_t | s_j)} \quad (4)$$

Figure 3 One-pass decoder with one acoustics model on the state graph

3.3 Calculating numerator for generalized posterior probability of state

Pruning strategy for beam search is used in the decoding process of many recognition systems. In which, viterbi algorithm is a traditional method shown in Figure 4.

When a frame input speech feature is being decoded, the observation probabilities $p(x_t | S_j)$ of states which are in the state graph are computed. The accumulated probability is used for finding the best candidates and pruning. At the same time, the probability $p(x_t | S_j)$ is regarded as numerator in equation (4) for calculating the posterior probability of state j .

$p(x_t | S_j)$ is the acoustics score of the input speech feature x_t under condition of the corresponding state S_j of HMM, the model of state S_j is normal distribution $N(\mu_j, \Sigma_j)$.

3.4 Calculating denominator for generalized posterior probability of state

In Equation (4), D is the set of all states in the acoustics model. For the new algorithm, when calculating the scoring feature during the decoding process, only the posterior probabilities of states reserved after pruning are used for calculating denominator.

During the decoding process, a pruning strategy is employed for improving decoding speed and reducing the search space. In Figure 4, solid dot represents the reserved states and hollow dot represents the pruned states. As Figure 4 shows, some of states are pruned because they give few contributes to the input observation and their posterior probability is low. As each frame of input speech data is processed, the posterior probabilities of each state that are reserved are computed and stored in an array indexed by state number and frame number while other states are pruned and omitted during computing.

In the state graph used in our decoding system, a virtual node is set as the flag at the end of each phone. As soon as a virtual node is reached during searching in the state graph, a phone item is decoded.

Figure 4 Synchronous calculating method on the base of state graph

On the virtual node, the posterior probability of the phone will be computed. Firstly, the numerator of Equation (4) will be estimated as specified in 3.3 above. In fact, they are computed and stored during searching. Secondly, the stored probabilities of each state indexed by frame number and state number will be used. The sum of these posterior probabilities of states calculated with the same frame input speech data is regarded as the denominator in Equation (4). This way, the generalized posterior probability of each state which constitutes the phone is prepared, and then according to Equation (3), the posterior probability of the phone is achieved with them as well as the duration information from decoding.

Because the sum of these probabilities of states reserved is much higher than the sum of what are pruned, sum of them can be regarded as the denominator of generalized posterior probability of state with the same frame input speech data. So Equation (4) is changed into:

$$p(s_{m(t)} | x_t) = \frac{p(x_t | s_{m(t)})}{\sum_{s_j \in D^*} p(x_t | s_j)} \quad (5)$$

Where, D^* is the set of states which are reserved after pruning the search space on one frame.

4. EXPERIMENTS AND RESULTS

In order to test the performance of the proposed algorithm, experiments were conducted using a database comprising 140 Chinese words read by 600 non-native students. All of the utterances in the database are scored by five experienced raters with three-point rubric. The final score of each utterance is the combination of the scores from five experience raters.

In the experiment, two algorithms are adopted, one shown in Figure 1 (defined as two-pass) and the other shown in Figure 3 (defined as one-pass). In two-pass method, two different acoustics models are used. The model used for decoding includes 4800 states and 16 Gaussian mixtures and the other model is a small acoustics model which only covers all the phones for calculating confidence with 960 states and 8 Gaussian mixtures. In the one-pass method, the model with 4800 states and 16 Gaussian mixtures is used for decoding and calculating scoring feature. The performances of these two algorithms are shown in Table 1.

Calculating method	Scoring accuracy
One-pass decoder	90.2%
Two-pass decoder	87.6%

Table 1 comparison of performance of two algorithms

From Table 1, one can see that the performance of one-pass algorithm is better than which of two-pass algorithm. The scoring accuracy of one-pass method is 87.6% and scoring accuracy of two-pass method is 90.2%. Because the same acoustics model is used in one-pass algorithm and the model for computing posterior probability used is more subtle than in two-pass algorithm, the performance of one-pass algorithm is better despite that the denominator in equation (5) is approximate.

Calculating method	Run time
One-pass decoder	0.198s
Two-pass decoder	0.238s

The computing complexity of One-pass decoder reduced 16% than which of two-pass decoder.

Table 2 Comparison of speed performance between two algorithms

In addition, the computing complexities of two methods are different. Table 2 shows the comparison. The speed of one-pass method is about 16% faster than that of the two-pass method.

5. CONCLUSION

In this paper, a new method for calculating scoring feature of posterior probability is proposed. In this method, calculating posterior probability is synchronous to the decoding process on the basis of state graph and only one acoustics model is needed. Through this, the conflict between the size of catch-all model and the speed of calculating confidence can be solved. Experiment results show that the performance of the method is better than that of the traditional two-pass method and at the same time the speed is improved.

In this algorithm, posterior probability is also adopted as the scoring feature. Finding new and valid scoring features and improving the performance as well as running speed will be our future work.

6. REFERENCES

- [1] B Dong, Q Zhao, J Zhang, Y Yan. "Automatic Assessment of Pronunciation Quality". Chinese Spoken Language Processing, 2004 International Symposium on, Hong Kong, December 2004.
- [2] T. Kemp, T. Schaaf, "Estimating confidence using word lattices", in Proc. 5th Eur. Conf. Speech, Communication, Technology 1997, Rhodes, Greece, Sept. 1997, pp. 827-830.
- [3] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel, "The RWTH large vocabulary continuous speech recognition system," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Seattle, WA, May 1998, pp. 853-856.
- [4] B. Rueber, "Obtaining confidence measures from sentence probabilities," in Proc. 5th Eur. Conf. Speech Communication Technology 1997, Rhodes, Greece, Sept. 1997, pp. 739-742.
- [5] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, A. Stolcke, "Neural network based measures of confidence for word recognition," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 1997, Munich, Germany, Apr. 1997, pp. 857-890.
- [6] S. Young, "Detecting misrecognitions and out-of-vocabulary words," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 1994, Adelaide, Australia. Apr. 1994, pp. 21-24.
- [7] Ze'ev Rivlin, Michael Cohen, Victor Abrash, Thomas Chung. "A Phone-Dependent confidence Measure For Utterance Rejection", Proceedings of IEEE, 1996.

[8] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Spoken Language Processing, Prentice Hall PTR, Upper Saddle River, New Jersey.

Proof