論文 / 著書情報
Article / Book Information

Title	Speech Modeling Based on Committee-Based Active Learning
Author	Yuzo Hamanaka, Koichi Shinoda, Sadaoki Furui, Tadashi Emori, Takafumi Koshinaka
Journal/Book name	Proc. ICASP2010, Vol. , No. , pp. 4350-4353
発行日 / Issue date	2010, 3
権利情報 / Copyright	

SPEECH MODELING BASED ON COMMITTEE-BASED ACTIVE LEARNING

Yuzo Hamanaka, Koichi Shinoda, Sadaoki Furui*

Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku,
Tokyo, 152-8552, Japan

Tadashi Emori†, Takafumi Koshinaka

NEC Corporation
1753 Shimonumabe, Nakahara-ku,
Kawasaki, 211-8666, japan

ABSTRACT

We propose a committee-based active learning method for large vocabulary continuous speech recognition. In this approach, multiple recognizers are prepared beforehand, and the recognition results obtained from them are used for selecting utterances. Here, a progressive search method is used for aligning sentences, and voting entropy is used as a measure for selecting utterances. We apply our method not only to acoustic models but also to language models and their combination. Our method was evaluated by using 190-hour speech data in the Corpus of Spontaneous Japanese. It proved to be significantly better than random selection. It only required 63 h of data to achieve a word accuracy of 74%, while standard training (i.e., random selection) required 97 h of data. The recognition accuracy of our proposed method was also better than that of the conventional uncertainty sampling method using word posterior probabilities as the confidence measure for selecting sentences.

Index Terms— acoustic model, language model, active learning, progressive search, voting entropy

1. INTRODUCTION

Model parameters in statistical speech recognition are estimated from a large amount of speech data that are manually transcribed. Since it is expensive to manually transcribe speech data, many studies have attempted to reduce the cost of this transcription. Active learning is one of these attempts, where utterances are selected from untranscribed training data by using various criteria, manually transcribed, and then used as training data. The goal of active learning is to improve recognition accuracy more than that with standard model training with fewer transcribed training data.

There have been many studies on active learning in speech recognition [1, 2, 3, 4, 5, 6]. The key issues in active learning are the criteria for selecting utterances. Many approaches [1, 4, 6] have used *uncertainty sampling* based on *confidence measures*. The initial recognizer in these approaches, which is prepared beforehand, is first used to recognize all the utterances in the training set, and those utterances that have recognition results with less confidence are then selected. The word posterior probabilities (WPPs) for each utterance have often been used as confidence measures (e.g., [1, 4]). Varadarajan *et al.* [6] used entropy in a word lattice for each sentence, produced by a recognizer.

This paper proposes a novel method of active learning based on *query by committee* (QbC) for large vocabulary continuous speech recognition (LVCSR). Multiple speech recognizers are prepared beforehand in this approach, and those utterances with a high *degree of disagreement* between the recognition results are selected to be manually transcribed. QbC-based active learning was first proposed in the field of machine learning [7, 8] and Dagan *et al.* confirmed its effectiveness in a part-of-speech tagging task [9]. Tur *et al.* proved that it was also effective for call-type classification in telephone service, where they used transcribed text data [2]. Three issues need to be determined in applying this approach to speech recognition: 1) How to prepare and update the multiple recognizers. 2) What is the optimal number of recognizers? 3) How to measure the degree of disagreement between the recognition results. We address these three problems in the following sections.

Until now, active learning in speech modeling has been evaluated using rather small tasks where there have been around 10,000-30,000 training utterances (e.g. [4, 6]). We evaluated our method by using 224,434 utterances (190 h) in this study to assess the effectiveness of our method in real applications. With this large amount of training data, we not only evaluated the effectiveness of our method for acoustic-model training, but also for language-model training.

The rest of the paper is organized as follows. Section 2 outlines our active-learning scheme. Section 3 briefly summarizes the theoretical aspects of the QbC-based approach for active learning. Section 4 describes in detail how it is applied to speech recognition. Section 5 presents the results of our evaluation and Section 6 concludes the paper.

2. OVERVIEW

Fig. 1 outlines the flow for our QbC-based active-learning framework for speech modeling. Let us assume we have training data, T, whose utterances are fully transcribed, and untranscribed training data, U. We determine the number of recognizers, K, for active learning, the amount of data N (h) to be selected in one active learning cycle, and the amount of transcribed data we would like to have, which are all done beforehand.

The active learning is carried out in the following process.

1. Divide the training data, T, randomly and equally into K data sets, T_k, $k = 1, \ldots, K$.
2. Estimate the parameters of the k-th recognizer, M_k, by using the k-th data set, T_k, for $k = 1, \ldots, K$.
3. Recognize all the utterances in the untranscribed training data, U, by each recognizer, M_k, $k = 1, \ldots, K$, to generate K different recognition results (sentences) for each utterance.
4. Select those utterances with a higher degree of disagreement between K recognizers than the others until the selected utterances reach N (h).
5. Subtract the selected data from U, add them to T, and go to Step 1.

*The contact address is hamanaka@ks.cs.titech.ac.jp
†The fourth author is now with NEC Information Systems, Ltd.

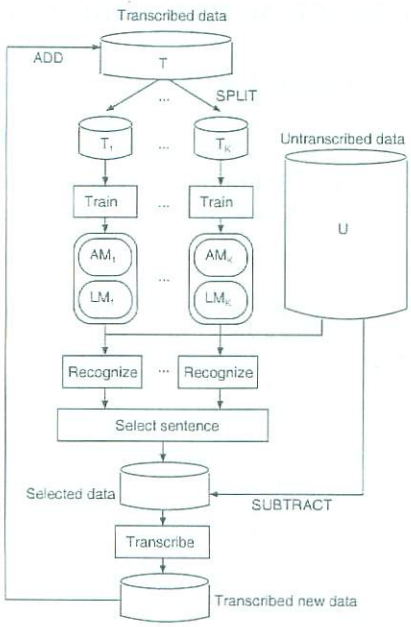

Fig. 1. QbC-based active learning scheme for speech recognition.

We repeat this active-learning cycle until the amount of transcribed data reaches our predetermined goal. Finally, all the transcribed utterances are used to make a single recognizer for speech recognition. The selection process in Step 4 is explained in detail in Section 4.

Active learning in speech recognition can be applied not only to an acoustic model but also to a language model. We apply the active learning process previously described to both of these in this study; we simultaneously update the parameters of both models.

3. QUERY BY COMMITTEE

The query-by-committee (QbC) paradigm was first proposed by Seung *et al.* [7] for active-learning problems in general, and applied to selective-sampling problems by Freund *et al.* [8], where the learner examined many unlabeled examples and only selected those samples that were more informative for learning than the others. The learner in this committee-based sampling scheme constructs a *committee* of classifiers using the training data currently available. Each committee member then classifies the candidate samples extracted from the unlabeled training data, and the learners measure *the degree of disagreement* among the committee members. Samples with larger degrees of disagreement are selected for labeling.

Early QbC studies [7, 8] considered their theoretical aspects within the context of binary-classification problems. They defined a *version space* as a set of concepts that labeled all the training examples correctly, and developed an algorithm to effectively restrict the version space as the number of examples increased.

4. SENTENCE SELECTION

It is rather difficult to directly apply the original QbC framework to speech recognition, since probabilistic classifiers are usually used. Probabilistic classifiers do not always assign the highest probability to the correct class for each training sample, and hence, the notion of version space cannot generally be applied. Furthermore, our classification problem, i.e., LVCSR, was much more complicated than simple problems with binary classification.

Instead, we simply assume that the degree of disagreement among classifiers accurately represents the uncertainty of transcription in the given utterance. For measuring the degree of disagreement, we first align word sentences from the K recognizers using an efficient search technique, and then calculate the *voting entropy* of the resulting word lattice. We will explain these steps in the following subsections. It should be noted that our approach is closely related to the entropy-based approach proposed by Varadarajan *et al.* [6]. While they measured the entropy of a word graph produced by a single recognizer, we measured that of a word lattice produced by the 1-best recognition results of many recognizers.

4.1. Sentence alignment

The accuracy of the alignment of multiple time sequences, which is called *multiple alignment*, is a critical issue in our method of selecting appropriate utterances. Since its computational costs increase exponentially as the number of sequences increases, some approximation should be introduced. A method of aligning sentences generated from multiple recognizers, which is called ROVER, has often been used for voting schemes in speech recognition [10]. This research on ROVER, however, did not focus much on the alignment algorithm itself. It repeats a pair-wise alignment from one sentence to each of the other sentences, but the result may be far different from the optimal solution. The same kinds of problems have been extensively studied in the field of bioinformatics. Here, we employ a *progressive search* [11, 12], which has proven to produce more accurate alignment than the search methods using only the pair-wise alignment such as that used in ROVER, while it needs more computational costs. In this approach, a *guide tree* is constructed by using the pair-wise alignment process, and it is used to align many sentences.

Let us assume we have K classifiers and there are thus K sentences as the recognition results for each utterance in untranscribed training data. We apply the following multiple-alignment process to a sentence set of K sentences. At first a guide tree is constructed in two steps.

1. Calculate the distance between all the pairs in a sentence set using the conventional dynamic time warping technique (pair-wise alignment).
2. Carry out bottom-up clustering using the distances. The distance between two clusters (sentence sets), X, Y, is calculated as an average of the distances of all the possible sentence pairs where one of them is in X and the other is in Y.

Then, multiple alignment is carried out for each node from the leaf of the guide tree. The objects to be aligned at each node are categorized into three types: A) A sentence and a sentence. B) A sentence and a sentence set consisting of more than one sentence (a cluster). C) A cluster and a cluster. The root node corresponds to the final result of the alignment. The alignment process in Type A is the same as the pair-wise alignment process used in constructing the guide tree. Let us define a *gap* as a word that corresponds to a blank generated by deletion/insertion, and define L as the maximum number of words in each sentence over the two sentences. Then, the alignment result is represented by a matrix where the number of rows is the number of sentences, M (here two), and the number of columns is L. Each element in this matrix is one word in one sentence. The alignment in Types B and C between two matrixes

Table 1. An example sentence-alignment result. The number of classifiers (the number of rows), K, is eight, and the maximum number of words over the eight recognition results (the number of columns), L, is nine. Each alphabet is a unique word in the vocabulary. The symbol "$-$" indicates a gap.

	1	2	3	4	5	6	7	8	9
1	A	B	D	F	G	F	–	P	T
2	A	C	E	F	H	L	M	Q	T
3	A	C	E	F	I	F	N	–	U
4	A	C	E	F	J	F	–	P	T
5	A	C	E	F	J	F	–	R	T
6	A	C	E	F	J	F	–	R	T
7	A	C	E	F	K	F	N	S	T
8	A	C	E	F	J	F	O	S	T

A_1 and A_2 is carried out as follows. Let the number of rows and that of columns for A_1 be M_1 and L_1, and those for A_2 be M_2 and L_2. (In Type B, a single sentence is regarded as a matrix with a single row.) In this alignment, the DP plane is an $L_1 \times L_2$ plane and the local distance, d_{ij}, at point (i, j) in this plane is calculated using M_1 words in the i-th column of A_1 and M_2 words in the j-th column of A_2; d_{ij} is calculated as the average distance between all the possible pairs between $(M_1 + M_2)$ words (the number of pairs is $(M_1+M_2)(M_1+M_2-1)/2$). The result of alignment is represented by a matrix with $(M_1 + M_2)$ rows and $\max(L_1, L_2)$ columns.

The local distance between word a and b in the pair-wise alignment is defined as

$$s(a,b) = \begin{cases} 2 & \text{if } a = b \neq -, \\ -1 & \text{else} \end{cases}$$

where "$-$" indicates a gap. This setting was selected from several settings in our preliminary recognition experiments.

The final result of this alignment for K sentences is represented by a matrix where the number of rows is K, and the number of columns is the maximum number of words in each sentence over the K sentences. Table 1 shows an example sentence-alignment result.

4.2. Voting entropy

We measure the degree of disagreement among the recognizers by *voting entropy*. The result of the sentence-alignment process for each candidate utterance in unlabeled training data is represented by a word matrix as shown in Table 1. A *gap* in this word matrix is regarded as a word, and thus the number of words is the same for all the sentences generated from multiple recognizers.

Let K be the number of recognizers and L be the number of words in this utterance. Then, the voting entropy, $\mathrm{VE}(j)$, for the distribution of K words in the j-th column of the matrix is defined as

$$\mathrm{VE}(j) = -\sum_{w=1}^{W} \frac{V(w,j)}{K} \log \frac{V(w,j)}{K}, \qquad (1)$$

where W is the number of unique words in the j-th column and $V(w, j)$ is the number of occurrences of word w in the j-th column. The degree of disagreement D in this sentence is calculated as the average voting entropy, $\mathrm{VE}(j)$, over all the columns, $j = 1, \ldots, L$, of the matrix:

$$D = \frac{1}{L} \sum_{j=1}^{L} \mathrm{VE}(j). \qquad (2)$$

The utterances with larger D are selected to be transcribed.

5. EXPERIMENT

5.1. Experimental conditions

We evaluated our method using lecture-speech data obtained from male speakers in the Corpus of Spontaneous Japanese (CSJ) [13]. All the utterances in this database were fully transcribed. We used 224,434 utterances from 666 speakers as training data, and 2328 utterances from ten speakers as test data. The frame period in speech analysis was 10 ms and the frame width was 25 ms. The speech feature vector was 39 dimensional, consisting of 12-order mel-frequency cepstral coefficients (MFCCs) appended with the 0-th cepstrum, delta and delta-delta coefficients. We applied cepstral mean subtraction to all utterances.

The acoustic model for a recognizer was a hidden Markov model with 3000 states, each of which had a Gaussian-mixture probability density function. The number of mixtures in each state was 16. The structure of the acoustic model remained unchanged throughout all the experiments in this study. We applied a two-pass search for speech recognition. A 2-gram language model was used in the first pass and a 4-gram language model was used in the second.

We randomly selected 25.0 h (29,461 utterances) of data as the initial transcribed training data from CSJ, and used them to train the initial acoustic model and the initial 2-gram and 4-gram language models. The other data from the database were used as untranscribed data for active learning. The amount of data N to be selected at one cycle of the active learning process was set to 25 h.

We compared our proposed method with two other methods. The first was random selection, which corresponded to conventional model training. The second was the uncertainty sampling method using WPPs as a confidence measure [1] where utterances with low WPPs were selected.

5.2. Results

First we examined how to make the multiple recognizers in the committee. We tested three cases. In the first case, we used eight different models for both acoustic and language models (AM8-LM8). In the second case, one common language model is shared among eight recognizers (AM8-LM1), and in the third case, one common acoustic model is shared (AM1-LM8). The results was shown in Fig. 2. While the recognition performance of these three cases were similar, it was better to share the same language model for all the recognizers. This may be because language model training needs a larger number of samples than acoustic model training. The amount of training data was too small to estimate language model parameters precise enough to be used as the committee recognizers, when they were divided into eight.

This figure also plots the recognition accuracies obtained with the other two methods, the random selection method and the uncertainty sampling method. The proposed method was significantly better than that of random selection. For example, our method only required 63 h of data to achieve a word accuracy of 74%, while the random selection method required 97 h of data. Our method was also better than the uncertainty sampling method. Furthermore, the accuracy obtained by our method using 100 h data was as good as that of the case when all the 190 h data were used for training.

Fig. 3 shows the recognition performance of our proposed method when the number of recognizers (acoustic models) were changed. The improvements of our proposed method were not much different with different number of recognizers. Since the computational costs for recognition were propotional to the number of

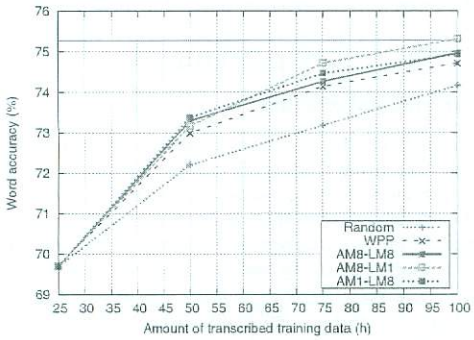

Fig. 2. Recognition results with different model combinations in the committee.In AM8-LM8, eight pairs of acoustic models and language models were used as the recognizers in the committee. In AM1-LM8, the number of recognizers was eight but they all used the same acoustic model trained by using all the training data transcribed until then. In AM8-LM1, the eight recognizers shared the same language model. They were compared with random selection (Random) and a method of selection using the WPP-based confidence measure (WPP).The horizontal solid line showed the recognition result (75.2%) obtained by using all the training data (190 h) we prepared for the experiment.

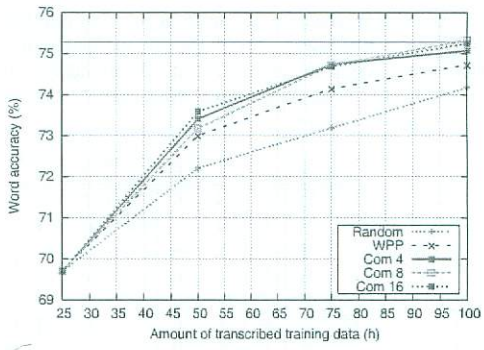

Fig. 3. Recognition results with different numbers (4,8,16) of recognizers (models) in the committee. Here the same language model was shared among the recognizers.

recognizers, the committee consisting of four recognizers should be chosen.

6. CONCLUSION

We proposed an active-learning framework based on the query-by-committee approach for speech recognition. A progressive search is used in our method to align sentences, and the degree of disagreement measured by voting entropy is used as a measure for selecting utterances. Our method was evaluated on CSJ. It proved to be significantly better than random selection and the conventional uncertainty sampling method using WPPs.

We constructed recognizers in the committee from randomly selected samples in the transcribed data in this study. In future, we plan to investigate how to construct recognizers that are different from each other. We are also interested in using many word graphs, each of which is generated by one recognizer, in our framework. We also plan to combine the proposed method and others using a confidence measure. Our method is expected to be effective in different tasks such as call routing in telephone applications and we plan to apply our method to these tasks.

7. ACKNOWLEDGMENTS

This study was partly supported by a Grant-in-Aid for Scientific Research (B) 20300063 from Japan Society for the Promotion of Science.

8. REFERENCES

[1] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. ICASSP*, 2002, pp. 3904–3907.

[2] G. Tur, R. Schapire, and D. Hakkani-Tur, "Active learning for spoken language understanding," in *Proc. ICASSP*, 2003, vol. 1.

[3] T. M. Kamm and G. G. L. Meyer, "Robustness aspects of active learning for acoustic modeling," in *Proc. ICSLP*, 2004, pp. 1095–1098.

[4] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," vol. 13, no. 4, pp. 504–511, 2005.

[5] H.-K. J. Kuo and V. Goel, "Active learning with minimum expected error for spoken language understanding," in *Proc. Interspeech*, 2005, pp. 437–440.

[6] B. Varadarajan, D. Yu, L. Deng, and A. Acero, "Maximizing global entropy reduction for active learning in speech recognition," in *Proc. ICASSP*, 2009, pp. 4721–4724.

[7] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. Workshop on Cmput. Learning Theory*, 1992, pp. 287–294.

[8] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," in *Machine Learning*, 1997, vol. 28, pp. 133–168.

[9] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proc. ICML*, 1995, pp. 150–157.

[10] J. G. Fiscus, "A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (rover)," in *Proc. IEEE Workshop on Automatic Recognition and Understanding*, 1997, pp. 347–354.

[11] D.-F. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *Molecular Biology and Evolution*, vol. 13, pp. 93–104, 1996.

[12] R. Durbin, S. R. Eddy, and A. Krogh, *Biological sequence analysis*, Cambridge University Press, 1998.

[13] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC*, 2000, vol. 2, pp. 947–952.