Acoustic model training using committee-based active and semi-supervised learning for speech recognition

Takuya Tsutaoka* and Koichi Shinoda*
* Department of Computer Science, Tokyo Institute of Technology, Japan
E-mail: tsutaoka@ks.cs.titehc.ac.jp, shinoda@cs.titech.ac.jp

***Abstract*—We propose an acoustic model training method which combines committee-based active learning and semi-supervised learning for large vocabulary continuous speech recognition. In this method, each untranscribed training utterance is examined by a committee of multiple speech recognizers, and the degree of disagreement in the committee on its transcription is used for selecting utterances. Those utterances the committee members disagree with each other are transcribed for active learning, while those they agree are used for semi-supervised learning. Our method was evaluated using the Corpus of Spontaneous Japanese. It was shown that it achieved higher recognition accuracy with lower transcription costs than random sampling, active learning alone, and semi-supervised learning alone. We also propose a new data selection method called middle selection in semi-supervised learning.**

Index Terms: active learning, semi-supervised learning, LVCSR, query by committee

I. Introduction

A large amount of transcribed speech data are required for training acoustic models in statistical speech recognition systems. However, it is usually costly to transcribe speech data manually. This is a serious problem, especially when we develop a speech recognition system for a resource-deficient language because its market may be too small to afford such high a cost. Active learning and semi supervised learning have been studied as ways of reducing the transcribing cost.

In active learning, we select data to be transcribed and use them for training. The transcribing cost may be reduced when we can successfully select more informative data, which is useful in training, than the others. There have been many studies on active learning for speech recognition [**?**], [**?**], [**?**], [**?**], [**?**]. Many studies have used *uncertainty sampling* based on confidence measures [**?**], [**?**], [**?**]. In these studies, untranscribed utterances with lower confidence of recognition results are selected and transcribed for the model training. As confidence measures, the word posterior probabilities (WPP) [**?**], [**?**] or the entropy in a word lattice [**?**] for each utterance have been used. Another study called committee-based active learning [**?**] selects the untranscribed utterances which have a higher degree of disagreement between multiple recognizers. The degree of disagreement is measured by *Vote Entropy* of the committee.

In semi-supervised learning, the most probable recognition hypothesis of an untranscribed utterance obtained by a speech recognizer is used as its transcription in the model training. No manual transcription is provided for the utterance. Similar to active learning, confidence measures are often used to select utterances for semi-supervised learning [**?**], [**?**], [**?**]. For example, WPP [**?**] and the difference of likelihood between 1-best and 2-best results [**?**] were used as the confidence measure.

Combining active learning and semi-supervised learning has been also studied [**?**], [**?**]. In this combination, utterances with lower confidence are selected to be transcribed for active learning, and ones with higher confidence are used for semi-supervised learning without transcription. It achieved the same recognition accuracy with less transcribing costs than that for active learning alone and semi-supervised learning alone.

In this paper, we propose a combination of active learning and semi-supervised learning using a committee-based approach [**?**]. In this method, each untranscribed training utterance is examined by a committee of multiple speech recognizers, and the degree of disagreement in the committee on its transcription is used for selecting utterances instead of the confidence measures such as WPPs. In semi-supervised learning, it was reported that using the utterances with high confidence for training often deteriorated the recognition performance [**?**]. To avoid this problem, we also propose an alternative way of the utterance selection, middle selection, in which the data with the middle degree of disagreement are chosen for semi-supervised training. We evaluated our methods by simulation experiments using a fully transcribed database, Corpus of Spontaneous Japanese (CSJ) [**?**].

This paper is organized as follows. Section II describes our method using committee-based learning. Section III investigates how to select utterances in semi-supervised learning. Section IV reports our evaluation experiments using CSJ, and Section **??** concludes the paper.

II. Committee-Based Learning

Committee-based learning uses the degree of disagreement between multiple recognizers, called a *committee*, for data selection. We previously proposed a committee-based active learning method for speech recognition [**?**]. In this paper, we apply this method not only for active learning, but also for semi-supervised learning of acoustic models.

A. Framework

Let T be a set of *transcribed* utterances, U be a set of *untranscribed* utterances. We first prepare a small set of transcribed utterances as the initial T. Then committee-based learning is carried out in a five-step process described below. Fig. 1 provides its schematic view.

1) Divide the training data, T, randomly and equally into K data sets, T_k $(k = 1, \ldots, K)$.
2) Train the k-th recognizer, M_k, using the k-th data set, T_k, for $k = 1, \ldots, K$.
3) Recognize each utterance in the untranscribed training data, U, with each of the K recognizers, M_k $(k = 1 \ldots, K)$, to generate K different recognition hypotheses for the utterance.
4) Select utterances according to the degree of disagreement between K recognizers until the amount of the selected utterances reach N (hours). In active learning, select those utterances with high degree of disagreement. In semi-supervised learning, select those with low degree of disagreement. How to calculate the degree of disagreement is described in the following subsection.
5) Transcribe the utterances selected for active learning. For semi-supervised learning, use the most probable recognition hypothesis of each utterance is used as its transcription
6) Move the selected utterances from U to T, and go to Step 1.

B. Degree of Disagreement

For each utterance, we first apply the progressive alignment method, which is often used in Bioinformatics (e.g., [**?**]), to align the K recognized sentences from the K recognizers. The result of this alignment is represented by a $K \times C$ matrix, where C is the number of words in the longest sentences among the K sentences. Here a *null* word is defined to represent a *deletion* in the alignment for short utterances. This alignment process was explained in more detail in our previous paper [**?**].

Then we measure the degree of disagreement for the utterance using the $K \times C$ matrix. Let P_c be the number of unique words in the c-th column, w_{cp} $(1 \leq p \leq P_c)$ be a unique word in the same column, and N_{cp} be the number of w_{cp} in the c-th column. Then, the vote entropy, $V(c)$, for the c-th column is:

$$V(c) = -\sum_{p=1}^{P} \frac{N_{cp}}{K} \log \frac{N_{cp}}{K}.$$

The vote entropy D for the whole utterance is calculated as the average of $V(c)$ over all the columns:

$$D = \frac{1}{C} \sum_{c=1}^{C} V(c).$$

As the recognition hypotheses of the committee recognizers become more different, this D becomes larger. We call D as the degree of disagreement for this utterance.

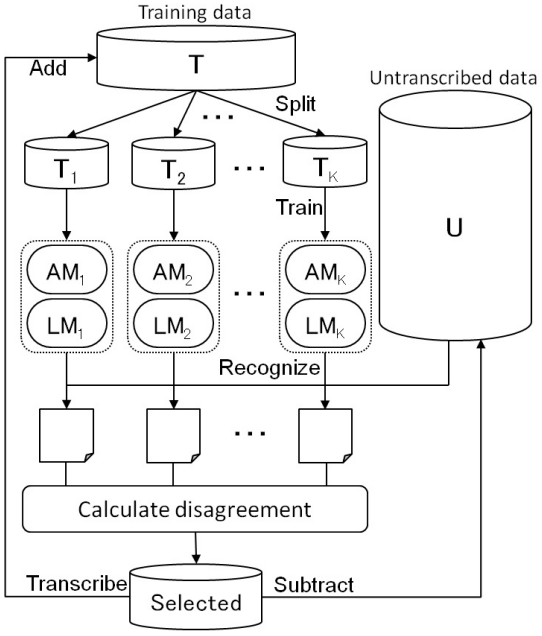

Fig. 1. *Scheme of committee-based learning.*

III. Data Selection in Semi-Supervised Learning

In semi-supervised learning, it was reported that using data with high confidence for training often deteriorated the recognition performance [**?**]. Utterances with high confidence are often very short and simple. Their vocabulary may be task-dependent, and its size is rather small. If we train a speech model only using those utterances, the resulting model may show poor performance in large vocabulary speech recognition. The same problem may occur also in our committee-based method. In order to avoid this problem, we propose to use utterances with a *middle* degree of disagreement for semi-supervised learning.

In this middle-selection method, We first sort all the utterances in the descending order of disagreement. Then we select the utterance closest to its average. Next we select an utterance which has the next lower disagreement than the previously selected utterance. We continue this process until we get the predetermined amount of utterances. We can use the similar utterance selection method for the WPP-based semi-supervised learning [**?**], [**?**] where the utterances with the middle confidence are chosen. While this middle selection method was significantly effective in our evaluation (see Section V.B.), we have not yet found any theoretical justification for this method. Further investigation in this direction is needed.

IV. Experiment

A. Experimental Conditions

We evaluated our method using Academic Presentation Speech (APS) and Simulated Public Speaking (SPS) of the Corpus of Spontaneous Japanese (CSJ) [**?**]. APS is live recordings of academic presentations in nine different academic societies held in 1999-2001. SPS is layman's "speech" on

everyday topics of about 10-12 minutes in front of a small friendly audience. Both APS and SPS are spoken by both male and female speakers. It should be noted that we assumed that they were untranscribed in the utterance selection experiments, while these utterances were actually fully transcribed. In APS dataset, there were 273,878 utterances (234.1h) as the training data set, and 2,410 utterances (2.1h) from another ten speakers for the test set. In SPS dataset, on the other hand, there were 367,137 utterances (284.8h) as the training data set, and 1,825 utterances (1.52h) from another ten speakers for the test set.

The frame period in speech analysis was 10ms and the frame width was 25ms. The speech-feature vector was 39 dimensional, consisting of 12-order mel-frequency cepstral coefficients (MFCCs) appended with energy, delta, and delta-delta coefficients. We applied cepstral mean subtraction to all utterances.

The acoustic model for a recognizer was a triphone hidden Markov model (HMM). Each HMM state had a Gaussian-mixture probability density function with 16 mixtures. 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. The same language models are used for training and testing.

In triphone HMMs, the decision-tree-based state tying is usually applied to decrease the number of states, where the resulting number should be controlled according to the amount of training data available. However, it is costly to optimize the number of states at each step of active and semi-supervised learning. In this study, therefore, we applied an automatic method using the MDL criterion [**?**] for this optimization.

We randomly selected 5 (h) of data as the initial transcribed training data from the training data, and used that to train the initial acoustic model and the initial 2-gram and 4-gram language models. The amount of data to be selected at one cycle of the learning process, N, was set to 5 (h). The number of acoustic models in a committee was set at 4. This number showed the best performance in our preliminary experiment.

B. Semi-Supervised Learning

We examined the effectiveness of the middle selection method in semi-supervised learning. Table **??** lists the recognition accuracy of the recognizers trained with 5 (h) of supervised utterances and 10.6 (h) semi-supervised utterances. We compared the four utterances selection methods, WPP, Com, WPP(mid), and Com(mid), where "Com" denotes our committee-based method, and "(mid)" denote the middle selection method. WPP(mid) and Com(mid) showed better recognition performance than the other two for the test data, while they had higher WER for the training data in semi-supervised learning. This result clearly indicates that the middle selection method was effective.

As an examination of the results of the middle selection method, we measured triphone coverages (Fig. **??**) of the utterances selected by the four methods. The coverage rates obtained by WPP(mid) and Com(mid) were significantly better than those of WPP and Com. This high coverage rates may be one reason for the higher performance.

TABLE I

Word accuracy (WA, %) of four utterance selection methods, where 10.6 (h) data was used for semi-supervised learning. "Com" is the committee-based method and "Com(mid)" is that with the middle selection. "WPP" is the WPP-based method and "WPP(mid)" is that with the middle selection. "Train" is WA for the data used for semi-supervised learning, and "Test" is that for the test data.

	Train	Test
Com	68.3	58.8
Com(mid)	53.6	62.0
WPP	80.7	60.0
WPP(mid)	51.6	62.3

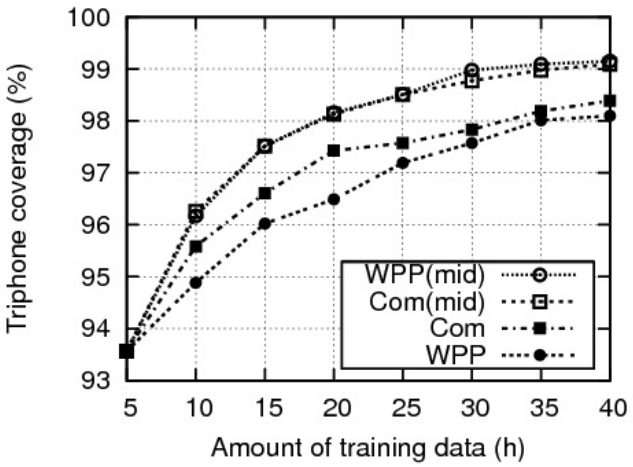

Fig. 2. *Triphone coverages of the four selection methods in semi-supervised learning on the APS dataset. WPP(mid): WPP-based method with the middle-confidence data selection, Com(mid): Committee-based method with the middle-disagreement data selection, WPP: WPP-based method, and Com: Committee-based method.*

C. Committee-based Learning

We compared committee-based active learning [**?**] (Active), committee-based semi-supervised learning (Semi-supervised), and the proposed combining method. It should be noted that the proposed method selects 5 (h) utterances for manually transcribing and another 5 (h) utterances to be automatically transcribed. Thus, transcription costs of 10 (h) training data in the active learning and ones of 15 (h) training data in the proposed method are the same. Fig. **??** shows the results on the APS dataset. The proposed method achieved higher recognition accuracy with lower transcribing costs than the active learning and the semi-supervised learning.

D. Combining Active and Semi-Supervised Learning

Finally, we compared the WPP-based method with the middle selection method (WPP(mid)), the proposed method with the middle selection method (Com(mid)), and random selection (Random). Fig. **??** shows the results on the APS dataset. Our proposed method outperformed the random selection and had almost the same performances as the WPP-

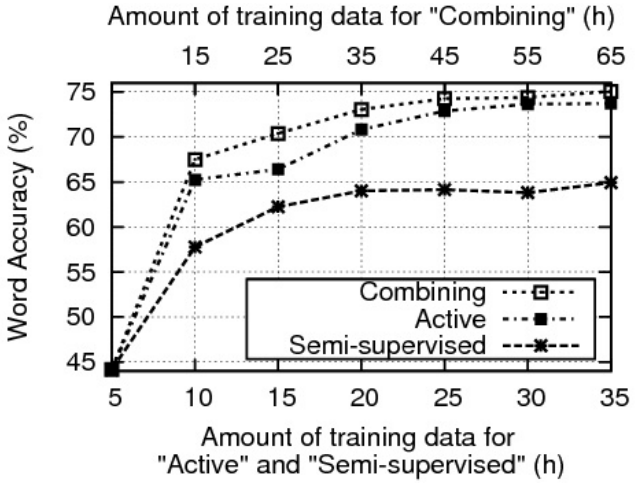

Fig. 3. *Recognition results on the APS dataset with different learning methods: proposed method (Combining), committee-based active learning (Active), and committee-based semi-supervised learning (Semi-supervised). "Combining" and "Active" are plotted so that they have the same transcription costs. For example, 15h training data for "Combining" consist of 10h supervised data which are transcribed, and 5h semi-supervised data without transcription.*

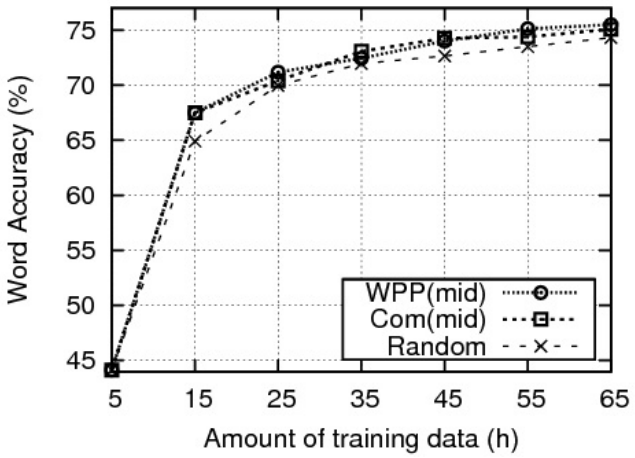

Fig. 4. *Recognition results for combining active and semi-supervised learning on the APS dataset with three selection methods: the WPP-based method with middle-confidence data selection WPP(mid), the committee-based method with middle-disagreement data selection Com(mid), and random selection Random.*

based method. Fig. **??** shows the results on the SPS dataset. The proposed method outperformed the random selection but fell below the WPP-based method. From these results, These results indicate that the characteristics of the sentences selected by WPP-based method and the committee-based method may be different. We plan to analyse this difference in future.

V. Conclusions

We proposed a combination of active learning and semi-supervised learning method using committee-based approach. A degree of disagreement in multiple recognizers was calculated by vote entropy and used for data selection. The performance of our method was significantly better than random selection, active learning alone, and semi-supervised

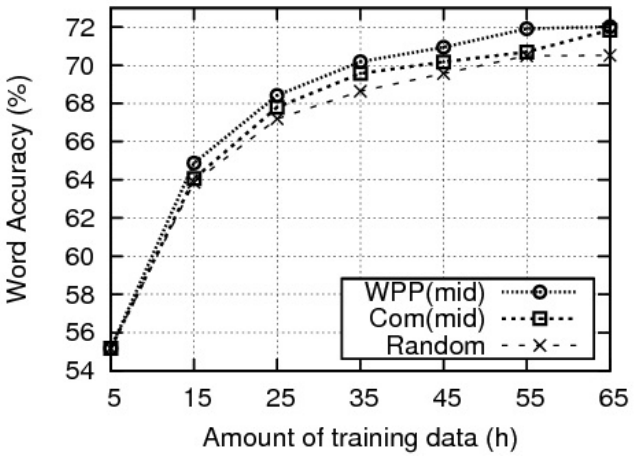

Fig. 5. *Recognition results for combining active and semi-supervised learning on the SPS dataset with three selection methods: the WPP-based method with middle-confidence data selection WPP(mid), the committee-based method with middle-disagreement data selection Com(mid), and random selection Random.*

learning alone. We also confirmed that the proposed middle data selection method was more effective than the conventional selection method.

In the future, we will further investigate the relationship between recognition accuracies and data selection methods in the semi-supervised learning. We will also try to find theoretical justification of the middle selection method.

References

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

[2] G. Riccardi and D. Hakkani-Tür, "Active learning: Theory and applications to automatic speech recognition," *Trans. IEEE*, Vol.13, No.4, pp.504-511, 2005.

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

[4] H. Lin, and J. Bilmes, "How to select a good training-data subset for transcription: submodular active selection for sequences," *Proc. Interspeech*, pp.2859-2862, 2009.

[5] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, "Speech modeling based on committee-based active learning," *Proc. ICASSP*, SP-L8.1, 2010.

[6] T. Kemp, A. Waibel, "Unsupervised training of a speech recognizer: recent experiments," *Proc. Eurospeech*, pp.2725-2728, 1999.

[7] D. Charlet, "Confidence-measure-driven unsupervised incremental adaptation for HMM-based speech recognition," *Proc. ICASSP*, pp.357-360, 2001.

[8] R. Zhang, "A new data selection Approach for semi-supervised acoustic modeling," *Proc. ICASSP*, pp.421-424, 2006.

[9] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," Journal of Speech Communication 45 (2), pp.171-186, 2005.

[10] D. Yu, B. Varadarajan, L. Deng, A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," Journal of Computer Speech and Language 24 (3), pp.433-444, 2010.

[11] DM. Mount, "Bioinformatics: Sequence and Genome Analysis 2nd ed.," Cold Spring Harbor Laboratory Press, 2004.

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

[13] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn. (E)*, vol.21, no.2, 2002.