Investigations on Ensemble Based Semi-Supervised Acoustic Model Training

Rong Zhang, Ziad Al Bawab, Arthur Chan, Ananlada Chotimongkol, David Huggins-Daines, Alexander I. Rudnicky

School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213, USA
{rongz, ziada, archan, ananlada, dhuggins, air}@cs.cmu.edu

Abstract

Semi-supervised learning has been recognized as an effective way to improve acoustic model training in cases where sufficient transcribed data are not available. Different from most of existing approaches only using single acoustic model and focusing on how to refine it, this paper investigates the feasibility of using ensemble methods for semi-supervised acoustic modeling training. Two methods are investigated here, one is a generalized Boosting algorithm, a second one is based on data partitions. Both methods demonstrate substantial improvement over baseline. More than 15% relative reduction of word error rate was observed in our experiments using a large real-world meeting recognition dataset.

1. Introduction

For many classification applications, collecting a large number of labeled training examples is a time-consuming and expensive process. For example, in our own meeting recognition research, it usually takes a skilled transcriber more than one week to generate and double-check the transcripts of a one-and-half hour meeting. One the other hand, massive amount of unlabeled raw data is relatively easy to obtain. Thus there exists the need for an automatic learning procedure that can use both labeled and unlabeled data for model training. This kind of approach is usually referred as semi-supervised learning, since it still uses a small fraction of labeled data. In contrast, unsupervised learning is carried out exclusively on unlabeled examples.

Semi-supervised learning approaches have aroused growing interests in various research fields, i.e. machine learning [1]. Many proposed methods rely on an extended EM algorithm to handle unlabeled data, and train the classifier in a bootstrap fashion [2]. In the case that two independent feature sets are available, Co-Training has demonstrated as an effective method to incorporate unlabeled data [3]. Please note that semi-supervised learning is not a risk-free strategy. Some negative experiments have been described in which degradation of performance is caused by adding unlabeled data. [4] analyzes the possible reasons for such degradation, pointing out that an increase in the number of unlabeled examples may lead to a larger estimation bias and classification error when the model assumption isn't correct.

As a potential solution to reduce system development cost, semi-supervised acoustic model training also attracts extensive attention from the speech community [5, 6, 7, 8, 9, 10, 11]. The basic idea is to train a seed model from a small portion of transcribed data, use it to decode a much larger amount of un-transcribed data, treat the hypotheses as an approximation of correct transcripts, and then train a new model with both transcribed and recognized data. Confidence measures are can be used to decide what kind of data is more suitable for semi-supervised training. However, researchers are divergent on the effectivness of using data with high confidence score,. For example, [5] suggests that this kind of data can't add substantial new information to the existing recognizer, while [7] shows that the good performance is mainly due to the contribution of this portion.

Most of the proposed semi-supervised learning approaches aim to train a single model or classifier, and work in a bootstrap fashion. Ensemble methods, i.e. Boosting algorithm were applied to semi-supervised learning problems, and reported to outperform other methods in a NIST evaluation [12, 13]. Inspired by this result and our previous work on supervised Boosting training, we investigated the feasibility of ensemble method for semi-supervised acoustic model training. The experimental results of two ensemble methods will be reported in this paper: one is a generalized Boosting algorithm updated for acoustic model training, and another one is a method based on data partition using confidence scoring. Both methods achieved substantial improvement on word error rate, demonstrating the effectiveness of ensemble methods for learning from unlabeled data. In addition, we will also report the experimental result for clarifying the role of high confidence data in semi-supervised acoustic model training.

2. Experiment Settings

This section describes the dataset and system settings as well as confidence scoring techniques used in our experiments.

2.1. Dataset

Our research is carried out in the context of a meeting recognition task [14]. There are a total of 75 meetings in the dataset, accounting for 60 hours of raw speech data. We use 10 meetings as the labeled set for initial acoustic training, 61 meetings as the unlabeled set for semi-supervised learning, 3 meetings as the hold-out set for recognizer tuning, and 1 meeting as the test set (which contains about 7500 words). The sampling rate is 11025Hz, and the frames rate is 105 per second. A 13-dimension MFCC feature vector is computed for each frame and then converted to a 39-dimension acoustic feature vector by adding delta and delta-delta coefficients.

2.2. System configuration

All of our experiments, both training and test, were performed using the Carnegie Mellon Sphinx III system, which is a fully-continuous HMM recognizer designed for LVCSR [15]. The

language model was solely trained from the 10-meeting transcribed set, without any access to the unlabeled set. The language model is fixed in our experiments, since our main goal is to investigate suitable approaches to acoustic model training. We believe that semi-supervised learning technique could be applied to language model training as well, and this issue has been listed in our research plan. The dictionary adopted in the experiments was based on the CMU Dictionary (containing more than 125k words). The actual vocabulary used in decoding was the intersection between dictionary and language model lexicon (unigram), which consists of 4200 words.

The context independent phone set for acoustic model training contains 49 basic phonemes. In context dependent training stage, these phones are transformed to triphones and then tied together to make senones. The number of senones was set to 2000 for all the acoustic models. A 3-state left-to-right architecture is adopted to model each speech unit. For the initial acoustic model trained on manually transcribed data, each state was modeled using a mixture of 16 Gaussians. This number was determined through preliminary experiments in which we observed the performance began to deteriorate with larger mixture sizes.

The baseline for further experiments is the word error rate obtained by evaluating the initial acoustic model, 47.31%. Please note that the language model used in this decoding is solely trained on transcribed data, so unavoidably there are many OOV words in the test set that are not included in the language model lexicon.

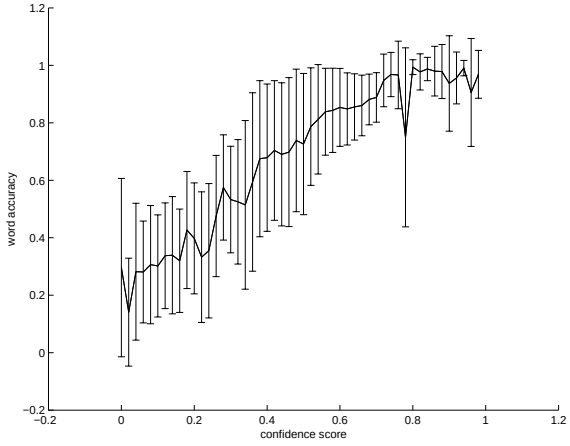

Figure 1: Performance of confidence scoring

2.3. Confidence measures

Our approach adopts confidence score as the criterion for data selection. In our experiments, data selection is performed on an utterance level so that an utterance is either kept or rejected as a whole depending on its confidence score. We use a neural network as the confidence annotator. The inputs consist of four features representing both language model and acoustic model information: *LM-backoff-mode*, *Utterance-level-posterior-probability*, *Word-level-posterior-probability* and *Frame-level-posterior-probability* [16], while the output is trained to approximate the word accuracy of each hypothesis.

The neural network-based confidence annotator is trained and tested on the transcribed data. Figure 1 illustrates its performance and shows the relationship between confidence score and the word accuracy of hypotheses (and also shows standard deviations). Figure 1 shows that the confidence score is generally proportional to word accuracy; that is, high confidence score indicates high accuracy and verse visa.

3. Analysis on Data Selection

There are two commonly used ways to handle unlabeled data: one is accepted by most machine learning researchers that use all of the unlabeled data for semi-supervised learning. The other is more common in the speech community;,the unlabeled data is incrementally added to the acoustic model training with the help of confidence scoring techniques. For the latter method, a confidence threshold is chosen such that the utterance whose confidence score is below it is rejected. Usually this threshold is first set to a strict one so that only the un-transcribed data with high confidence is selected, and then gets progressively relaxed to allow more data to be included in training. The assumption behind this strategy is that the high confidence data, usually corresponding to the data with high recognition accuracy (see Figure 1), will benefit, or at least not deteriorate, model training. However, there is counter example that suggests that high confidence unlabeled data does not help ([5]). These contradictory results raise for us the concern on how to properly select unlabeled data for semi-supervised training. This section will discuss several experiments we have conducted to clarify this question.

3.1. Training with all the un-transcribed data

In this experiment, all of the un-transcribed speech data, with their recognized hypotheses which are obtained by using the initial acoustic and language model, are combined with transcribed data for training new acoustic models. Table 1 presents the word error rates of the new acoustic models varying with different number of Gaussians per state.

Acoustic Model	Word Error Rate (WER)
16 Gaussians / State	46.28%
32 Gaussians / State	44.41%
64 Gaussians / State	42.83%

Table 1: Experiment with all un-transcribed data

Table 1 shows that semi-supervised training achieves an encouraging improvement over the baseline, 47.31% WER provided by initial acoustic model. When we use 64 Gaussians for every state, the error rate falls to 42.83%, which represents a 9.5% relative reduction. However, we should note that the improvement is partly realized by increasing the number of parameters in acoustic model, since the change of word error rate is quite small if we keep the setting of 16 Gaussians/state dictated by the small size of the original training corpus.

3.2. High confidence vs. low confidence

In this experiment, the initial acoustic model is used to decode the un-transcribed dataset, and then the neural network based confidence annotator is used to compute the confidence score for each hypothesized sentence. On this basis of the confidence score, the un-transcribed data is divided into two subsets, high confidence set and low confidence set, with almost the same amount of speech data. The two subsets are then combined with the transcribed data separately for semi-supervised acoustic model training. Table 2 presents the training results, showing that the acoustic models trained from

low confidence data are better than those trained from high confidence data.

Dataset	Acoustic Model	Word Error Rate
High Confidence	32G / State	44.61%
	64G / State	44.49%
Low Confidence	32G / State	44.20%
	64G / State	43.76%

Table 2: High confidence vs. low confidence

3.3. Further analysis

The results in Table 2 indicate that high confidence data may not be the best choice for semi-supervised training. We did another experiment to further study this question. As before, initial acoustic model and neural network based confidence annotator are used to generate the hypothesis and confidence score for un-transcribed data. After that the un-transcribed data is chunked into 9 bins ($0 \le n < 8$) that the *n-th* bin contains the utterances which confidence scores are within the range from top (n*10%) to (n*10%+20%). Namely, each bin owns 20% un-transcribed utterances, and shifts 10% one by one. Adding these bins separately to the transcribed set and running the semi-supervised training, we obtain the results illustrated in Figure 2. Please note that bin 0 represents the data with highest confidence score while bin 8 represents the data with the lowest score.

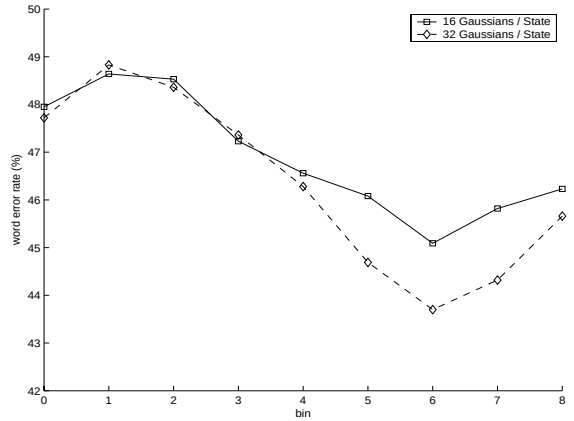

Figure 2: Performance of different part of un-transcribed data

The results show that the best performance is reached by using bin 6 for which the word error rate for the 32-Gaussian acoustic model is down to 43.7%. This indicates that the best portion of un-transcribed data for semi-supervised acoustic model training is the data which confidence score is neither too low nor too high. Too low means the hypotheses may contain many recognition errors, which is likely to cause the degradation of performance. On the other hand, high confidence implies the questioned data has been modeled well enough and is unable to provide extra information for improving model quality. This suggestes that one needs to balance two kinds of risks in semi-supervised training: the risk of using erroneous data and the risk of using less-informative data.

4. Ensemble Methods

Ensemble methods, especially Boosting algorithms, have demonstrated the ability to reduce classification errors in various supervised learning problems. In this section, we will investigate whether this kind of methods can be applied to semi-supervised acoustic model training in which the correct transcription is unknown for most of the data. Two methods are studied here; one constructs the ensemble by combining portions of un-transcribed data with different confidence score, and another one uses the semi-supervised Boosting algorithm proposed by [12, 13].

4.1. Data partition based ensemble

In section 3, the un-transcribed data was partitioned into nine bins based on the confidence score. This suggests that we can construct an ensemble by combining the models trained from different bins. Given the experimental results showing the superiority of low confidence data, we decided to only use the models trained from bin 5, 6, 7 and bin 8. Each model takes the setting of 32 Gaussian/state. Table 3 presents the word error rate after ROVER combination [17] of these 4 models.

Ensemble	Word Error Rate
b5 + b6 + b7 + b8	40.54%

Table 3: Performance of data partition based ensemble

Compared with the baseline of 47.31% word error, the ensemble method realizes a 14.3% relative reduction for misclassifications, which also outperforms the result obtained by using all of the un-transcribed data (see Table 1).

4.2. Semi-supervised Boosting training

In supervised learning, Boosting algorithm tends to emphasize the "hard" examples responsible for classification errors by increasing their weights in the training of next model. [12, 13] extended the idea to semi-supervised learning, in which the unlabeled examples that the members of ensemble can't reach unanimity will be given higher weights. An updated version of this algorithm which is suitable for acoustic model training is as follows.

Let $\mathbf{U}_T$ be the transcribed dataset that $\mathbf{U}_T = \{(\mathbf{x}_i, y_i) | 1 \le i \le M\}$ where $\mathbf{x}_i$ is the sequence of acoustic feature vectors for utterance i and y_i is the corresponding transcript. Let $\mathbf{U}_N$ un-transcribed set that $\mathbf{U}_N = \{\mathbf{x}_j | M+1 \le j \le N\}$.

Initialize: train initial acoustic model λ_0 from $\mathbf{U}_T$, generate hypothesis h_j for each utterance $\mathbf{x}_j \in \mathbf{U}_N$ using λ_0, and form a new set $\mathbf{U}_N^* = \{(\mathbf{x}_j, h_j) | M+1 \le j \le N\}$. Let $\mathbf{U}_0 = \mathbf{U}_T \cup \mathbf{U}_N^* = \{(\mathbf{x}_k, h_k) | 1 \le k \le N\}$ where $h_k = y_k$ for $1 \le k \le M$. Assign equal weight to each utterance $\mathbf{x}_k \in \mathbf{U}_0$ so that $w_0(k) = 1$.

Training: For l = 1 to L

- Train new acoustic model λ_l from data set $\mathbf{U}_{l-1}$.
- Test model λ_l on the set $\mathbf{U}_{l-1}$, generating N-best list for each utterance $\mathbf{x}_k \in \mathbf{U}_{l-1}$, and computing probability $P_{\lambda_l}(h | \mathbf{x}_k)$ for each hypothesis h in the N-best list of $\mathbf{x}_k$.
- Compute pseudo loss

$$\varepsilon = \frac{1}{Z_l} \sum_{\mathbf{x}_k \in \mathbf{U}_{l-1}} \sum_{h \neq h_k} (1 - P_{\lambda_l}(h_k \mid \mathbf{x}_k) + P_{\lambda_l}(h \mid \mathbf{x}_k))$$

where Z_l is the normalization factor that $Z_l = 2 \mid h \mid \mid \mathbf{U}_{l-1} \mid$. Set $\beta = \varepsilon / (1 - \varepsilon)$.

- Calculate new weight for each utterance $\mathbf{x}_k \in \mathbf{U}_{l-1}$

$$w_l(k) = \sum_{h \neq h_k} \beta^{\frac{1}{2}(1 + P_{\lambda_l}(h_k \mid \mathbf{x}_k) - P_{\lambda_l}(h \mid \mathbf{x}_k))}$$

- Resample $\mathbf{U}_{l-1}$ according to normalized $w_l(k)$, forming a new training set $\mathbf{U}_l$. For un-transcribed utterance $\mathbf{x}_j \in \mathbf{U}_l$, set h_j with the hypothesis most agreed among existing ensemble $\{\lambda_1, \lambda_2, ..., \lambda_l\}$.

Our Boosting experiment was also carried on low confidence data as in Section 3.2 and 4.1. There are a total of 4 models generated by the Boosting algorithm, each of which adopts the architecture of 32 Gaussians per state. Table 4 presents the final result after ROVER combination of the 4 models.

Semi-supervised Boosting	Word Error Rate
Combination of 4 models	40.02%

Table 4: Performance of semi-supervised Boosting algorithm

The word error rate achieved by the Boosting training is the best one in our experiments. Compared to the baseline of 47.31%, it represents a 15.4% relative reduction of classification error. The encouraging result illustrates the potential of this approach as a useful method for improving the quality of acoustic models by exploiting un-transcribed data.

5. Conclusions

This paper describes a semi-supervised technique for acoustic model training. Our experiments on a real-word meeting recognition corpus demonstrate that the word error rate can be significantly reduced by using un-transcribed speech data. We have empirically shown that different portions of un-transcribed data can have different impact on semi-supervised learning. That is, low confidence data outperforms high confidence data in improving recognition accuracy. In addition, we discovered that increasing the number of model parameters can benefit the system performance in the case that large amount of un-transcribed dada is available, i.e. the best model settings in our semi-supervised experiments is 32 or 64 Gaussians/state, while these settings would cause overfitting when used for only the transcribed data. This paper also describes the application of ensemble methods for semi-supervised acoustic model training. The experimental results show that significant gains are achieved by ensemble methods, especially the Boosting algorithm. Together these suggest promising research directions for semi-supervised learning.

6. Acknowledgement

This research was supported by DARPA grant NB CH-D-03-0010. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

7. References

[1] M. Seeger, "Learning With Labeled and Unlabeled Data", Technical Report, University of Edinburgh, 2001.

[2] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM", Machine Learning, 39(2/3):103-134, 2000.

[3] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training", Proc. of 11th Conference on Computational Learning Theory, 1998.

[4] F. G. Cozman, I. Cohen and M. C. Cirelo, "Semi-Supervised Learning of Mixture Models", Proc. of 20th International Conference on Machine Learning, 2003.

[5] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments", Proc. of 6th Eurospeech, 1999.

[6] T. M. Kamm and G. L. Meyer, "Automatic Selection of Transcribed Training Material", Proc. of IEEE Workshop on ASRU, 2001.

[7] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Modeling for Large Vocabulary Continuous Speech Recognition", Proc. of IEEE Workshop on ASRU, 2001.

[8] D. Giuliani and M. Federico, "Unsupervised Language and Acoustic Model Adaptation for Cross Domain Portability", Proc. of ISCA-ITR Workshop, 2001.

[9] L. Lamel, J. Gauvain and G. Adda, "Unsupervised Acoustic Model Training", Proc. of ICASSP, 2002.

[10] P. J. Moreno and S. Agarwal, "An Experimental Study of EM-based Algorithms for Semi-Supervised Learning in Audio Classification", Proc. of ICML-2003 Workshop on Continuum from Labeled to Unlabeled Data, 2003.

[11] K. Visweswariah and R. Gopinath and V. Goel, "Task Adaptation of Acoustic and Language Models Based on Large Quantities of Data", Proc. of ICSLP, 2004.

[12] F. d'Alche-Buc, Y. Grandvalet and C. Ambroise, "Semi-Supervised MarginBoost", Proc. of 9th NIPS, 2001.

[13] K. P. Bennett, A. Demiriz and R. Maclin, "Exploiting Unlabeled Data in Ensemble Methods", Proc. of SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[14] S. Banerjee, J. Cohen, T. Quisel, A. Chan, Y. Patodia, Z. A. Bawab, R. Zhang, A. Black, R. Stern, R. Rosenfeld, A. I. Rudnicky, P. Rybski, and M. Veloso, "Creating Multi-Modal, User-Centric Records of Meetings with the Carnegie Mellon Meeting Recorder Architecture", Proc. of ICASSP 2004 Meeting Recognition Workshop.

[15] CMU Sphinx Open Source Speech Recognition Engines, http://cmusphinx.sourceforge.net/html/cmusphinx.php.

[16] R. Zhang and A. I. Rudnicky, "Apply N-Best List Re-Ranking to Acoustic Model Combinations of Boosting Training", Proc. of ICSLP 2004.

[17] J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", Proc. of IEEE Workshop on ASRU, 1997.