A Frame Level Boosting Training Scheme for Acoustic Modeling

Rong Zhang and Alexander I. Rudnicky

Language Technologies Institute, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, USA
{rongz,air}@cs.cmu.edu

Abstract

Conventional Boosting algorithms for acoustic modeling have two notable weaknesses. (1) The objective function aims to minimize utterance error rate, though the goal for most speech recognition systems is to reduce word error rate. (2) During Boosting training, an utterance is treated as a unit for re-sampling and each frame within the same utterance is assigned equal weight. Intuitively, the frames associated with a misclassified word should be given more emphasis than others. We propose a frame level Boosting training scheme that addresses these shortcomings and allows each frame to have a different weight. We describe a technique and provide experimental results for this approach.

1. Introduction

The Boosting algorithm [1] has become a very popular method in machine learning field over the last few years. The underlying idea of Boosting is to combine many "weak" classifiers to form an ensemble with improved performance. In the Boosting procedure, these base classifiers are learned in a fashion such that the examples misclassified by the current classifier will be given higher weights in the training of subsequent classifiers. Specifically, the Boosting algorithm maintains a probability distribution for the training data, and initially assigns equal weight to each example. In each training round, a new classifier is learned from the current distribution and then applied to classify every training example. After that, the probability distribution is updated in such a way that the weight of an example will be increased if it is misclassified, otherwise decreased. This enables the training of subsequent classifier to concentrate on the examples which are difficult to be correctly classified. In generalization, the base classifiers are composed to form the final classifier that outputs the hypothesis through majority voting.

The application of Boosting to speech recognition is at an early stage of development but appears to be highly promising. Successful examples include word and phone recognition [2], confidence annotation [3], speaker identification [4] and continuous speech recognition [5][6]. In Section 2, we will discuss a variant of Boosting algorithm for acoustic modeling which is widely accepted by speech community. However, we should point out that the original Boosting algorithm was designed for common classification problems without concern for the characteristics of speech recognition, especially large vocabulary real time continuous speech recognition.

In the current Boosting algorithm, utterance is the basic unit used for acoustic model training. Our analysis shows that there are two notable weaknesses in this setting. First, the objective function of current Boosting algorithm is designed to minimize utterance error instead of word error. Utterance error is such a metric that a hypothesis is judged as correct only when all the words within it are correct. Obviously, this is a very strict metric that doesn't have too much practical use. In speech recognition, word error rate is the most commonly used metric for measuring system performance, on the word level. Although there is strong correlation linking them, using a different training criterion may not result in an optimal model. Second, in the current algorithm, an utterance is treated as a unity for resample. This means all the words in the same utterance always have equal weights. However, intuitively, the misclassified words should obtain more emphasis than other words even though they exist in the same utterance.

To address these two problems, this paper proposes a frame level Boosting training scheme for acoustic modeling. Two characteristics distinguish our scheme from conventional utterance level training method: (1) the objective function of Boosting algorithm is modified to minimize a frame based word error rate; (2) each frame is assigned a weight and the resample is carried out on the frame level.

This paper is organized as follows. A brief overview of Boosting algorithm for acoustic modeling is given in Section 2. Section 3 introduces our new training scheme. Initial experimental results are presented in Section 4. Section 5 concludes this paper with a discussion of further research topics.

2. Boosting Algorithm for Acoustic Modeling

Suppose we have a training set $\mathbf{U} = \{(\mathbf{x}_i, h_i) \mid 1 \le i \le N\}$, where in continuous speech recognition, $\mathbf{x}_i$ is the sequence of feature vectors for the *i*-th training utterance, while h_i is the corresponding transcript. (Without confusion, the symbol h is also used to denote the hypothesis in this paper.) The goal of Boosting training is to minimize the following objective function.

$$L = \sum_{i=1}^{N} \sum_{h \neq h_i} \exp(P(h \mid \mathbf{x}_i) - P(h_i \mid \mathbf{x}_i)) \quad (1)$$

where $P(h \mid \mathbf{x})$ denotes the conditional probability of hypothesis h given input $\mathbf{x}$. Formula 1 could be interpreted as a loss function strongly related to classification error. The process minimizing the value of this function leads to $P(h_i \mid \mathbf{x}_i) >> P(h \mid \mathbf{x}_i)$ for $h \neq h_i$, and as a consequence, results in the reduction of classification errors.

Different from traditional training methodology, such as Maximum Likelihood Estimation (MLE) that only outputs a single model which is optimized under certain criterion, Boosting algorithm generates a set of models $\{\lambda_1, \lambda_2, ..., \lambda_K\}$, which is called as ensemble in some literatures, by manipulating the distribution of training data. In

generalization stage, a weighted majority voting strategy is adopted to combine the predictions of individual models and makes the final hypothesis. For example, for a new utterance $\mathbf{x}$, the hypothesis is determined by

$$h^* = \arg\max_h \sum_{k=1}^{K} c_k * P_{\lambda_k}(h \mid \mathbf{x}) \quad (2)$$

where c_k is the importance weight for model λ_k.

Figure 1: Ada-Boosting Algorithm

Initialize:

- Let $\mathbf{U}^0 = \mathbf{U}$.
- Assign equal weight to each utterance $\mathbf{x}_i$ that $w_i^0 = 1$.

For k = 1 to K:

- Train new acoustic model λ_k from data set $\mathbf{U}^{k-1}$.
- Test model λ_k on the training set $\mathbf{U}^{k-1}$, generating N-best list for each utterance $\mathbf{x}_i$, and computing probability $P_{\lambda_k}(h \mid \mathbf{x}_i)$ for each hypothesis h in the N-best list of $\mathbf{x}_i$.
- Compute pseudo loss

$$\varepsilon^k = \frac{1}{2 \mid h \parallel \mathbf{U}^{k-1} \mid} \sum_{\mathbf{x}_i \in \mathbf{U}^{k-1}} \sum_{h \neq h_i^*} (1 - P_{\lambda_k}(h_i \mid \mathbf{x}_i) + P_{\lambda_k}(h \mid \mathbf{x}_i))$$

- Set $c_k = \varepsilon^k / (1 - \varepsilon^k)$
- Calculate weight for each hypothesis h ($h \neq h_i$) in the N-best list of utterance $\mathbf{x}_i$

$$w(\mathbf{x}_i, h) = c_k^{\frac{1}{2}(1 + P(h_i \mid \mathbf{x}_i) - P(h \mid \mathbf{x}_i))}$$

- Calculate new weight for each utterance $\mathbf{x}_i$

$$w_i^k = \sum_{h \neq h_i^*} w(\mathbf{x}_i, h)$$

- Resample training data according to normalized w_i^k, forming new training set $\mathbf{U}^k$.

In generalization, the hypothesis to a new utterance $\mathbf{x}$ is determined by $h^* = \arg\max_h \sum_{k=1}^{K} \log \frac{1}{c_k} P_{\lambda_k}(h \mid \mathbf{x})$, where h denoted the top-1 hypothesis of N-best list of each model.

Note that to make the Boosting algorithm work for acoustic model training, a couple of implementation issues need be considered. First, the number of possible hypothesis h for a given utterance $\mathbf{x}$ could be very large and it's impossible to calculate probability for every hypothesis. To solve this problem, we have to compress the hypothesis space into a subset with limited size, e.g. the N-best lists. Another issue to concern is how to compute $P_\lambda(h \mid \mathbf{x})$. Most speech recognizers only output the log-likelihood score or joint probability $P_\lambda(h, \mathbf{x})$ rather than the conditional probability. We use the following method converting the joint probability into conditional probability.

$$P_\lambda(h \mid \mathbf{x}) \approx \frac{P_\lambda(h, \mathbf{x})^\beta}{\sum_{h' \in \text{n-best list of } \mathbf{x}} P_\lambda(h', \mathbf{x})^\beta} \quad (3)$$

where β is the smoothing parameter whose value is empirically set. In the case that the correct hypothesis may not exist in the N-best list, one could run forced alignment for it to get the log-likelihood score, or just use a small default value. Figure 1 shows the pseudo code of a Boosting algorithm suitable for acoustic modeling.

Obviously, the basic training unit in the conventional Boosting algorithm is the utterance: (1) the objective function is based on an utterance level metrics $P(h \mid \mathbf{x})$; (2) resampling is also carried out on utterance level. Our analysis shows that these two characteristics make the conventional Boosting algorithm deviate from the goal of acoustic modeling. First, minimizing the value of an utterance level objective function tends to increase utterance accuracy rather than word accuracy, the accepted standard for evaluating the performance of a speech recognition system. Although there is a strong relationship between utterance and word accuracy, from $P(h \mid \mathbf{x})$ only we can't figure out how many words are misrecognized. For example, we often observed that the top-1 hypothesis in N-best list, the one with the highest conditional probability, is not the one with lowest word error rate. Second, in the conventional algorithm, an utterance is considered as a unit for purposes of resampling. This means that all the words in the same utterance always have equal weights. However, intuitively, the misclassified words should be given more emphasis than other words even though they exist in the same utterance.

3. Frame Level Training Scheme

To address the problem described in the previous section, we propose a frame level Boosting training scheme. A metric to measure the correctness of the hypothesized word for a particular frame is investigated. Based on this metric, the objective function is designed to minimize frame based recognition error instead of the utterance error. Meantime, each frame is assigned a weight indicating how difficult this frame is to recognize, and correspondingly the frame becomes the unit for resampling instead of the utterance. We believe that the new scheme can at least partly address the considerable mismatch between training criterion and training target in the conventional Boosting algorithm for acoustic modeling.

3.1. Frame level conditional probability for hypothesized word

It is well known that segmentation information is usually unavailable for acoustic model training. Namely, we don't know exactly the starting and ending point for a particular word within a continuous utterance. Therefore, it's very difficult for us to build an objective function that could directly describe the word error rate. We provide a compromise solution that uses frame level conditional probability to measure the correctness of a hypothesized word.

Let $\mathbf{x}$ be the sequence of cepstrum vectors for an utterance with T frames, and h be one of the hypotheses in N-best list of $\mathbf{x}$. We write the hypothesis in the format like

$h = (a_1, a_2, ..., a_T)$ in which a_t is the hypothesized word at frame *t*. The metrics that we will use in Boosting training is the frame level conditional probability $P(a_t \mid \mathbf{x})$.

$P(a_t \mid \mathbf{x})$ describes how likely that word a_t is a correct decoding result at frame *t*. Moreover, the normalized frame level conditional probability (see Formula 4) could be interpreted as the expectation of frame based word accuracy for hypothesis h.

$$f = \frac{1}{T}\sum_{t=1}^{T} P(a_t \mid \mathbf{x}) \tag{4}$$

Please note that Formula 4 is only an approximation for the word accuracy used in speech recognition. Sometimes the value obtained in Formula 4 is far from the real word accuracy.

The calculation of $P(a_t \mid \mathbf{x})$ is as follows.

$$\begin{aligned} & P(a_t \mid \mathbf{x}) \\ &= P(a_t, \mathbf{x}) / P(\mathbf{x}) \\ &= \sum_{\substack{h \in \text{ n-best list of } \mathbf{x} \\ \text{and the word at frame } t \text{ is } a_t}} P(h, \mathbf{x})^{\beta} \Bigg/ \sum_{h \in \text{ n-best list of } \mathbf{x}} P(h, \mathbf{x})^{\beta} \end{aligned} \tag{5}$$

where β is an empirically set smoothing parameter.

Please note that the hypothesized word isn't the only choice for this frame based metric. One could define similar metrics to measure the correctness of hypothesized state and phoneme. For example, an all-phone decoder could return us a hypothesis like $h_q = (q_1, q_2, ..., q_T)$ in which q_t is the hypothesized phoneme for frame *t*. Similar to frame level word probability, we could measure the correctness of q_t by computing $P(q_t \mid \mathbf{x})$ and use it to approximate the phone accuracy.

3.2. Objective function

On the basis of frame level conditional probability, the objective function for our new training scheme is defined as follows.

$$L = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{a \neq a_t} \exp(P(a \mid \mathbf{x}_i) - P(a_t \mid \mathbf{x}_i)) \tag{6}$$

where T_i denotes the length of the *i*-th utterance and a denotes the possible hypothesized word appearing at the *t*-th frame. In Formula 6, $\sum_{a \neq a_t} \exp(P(a \mid \mathbf{x}_i) - P(a_t \mid \mathbf{x}_i))$ is the pseudo loss for frame *t*, which describes the degree of confusion of this frame for recognition. Apparently, minimizing the value of this objective function will lead to $P(a_t \mid \mathbf{x}_i) >> P(a \mid \mathbf{x}_i)$ for $a \neq a_t$, and increase the accuracy of recognition.

3.3. Training Scheme

Figure 2 shows the frame level training scheme.

Note that in this scheme the weight $w_{i,t}^k$ is linked to each frame instead of the hypothesis in the conventional Boosting algorithm. This leaves us a question that how to resample the training data. We use a very simple strategy to solve this problem. Suppose $\mathbf{x}_i$ is the feature vector sequence for utterance *i* with *T* frames that $\mathbf{x}_i = (\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,T})$, and the weight for frame $\mathbf{x}_{i,t}$ is $w_{i,t}$. Our method is to duplicate $\mathbf{x}_{i,t}$ for $\lfloor w_{i,t} \rfloor$ times and creates a new utterance for acoustic model training. It should be mentioned that this kind of resample method isn't optimal and further research is necessary. Another issue of concern is how to determine the correct word for each frame. In our experiments, we use forced-alignment to obtain the most likely word segmentation information and label each frame with such information.

Figure 2: Frame Level Training Scheme

Initialize:

- Assign equal weight to each frame *t* of each utterance $\mathbf{x}_i$ in training set $\mathbf{U}$ that $w_{i,t}^0 = 1$.

For *k* = 1 to *K*:

- Train new acoustic model λ_k from distribution w^{k-1}.
- Test model λ_k on the training set, generating N-best list for each utterance $\mathbf{x}_i$, and computing probability $P_{\lambda_k}(a_t \mid \mathbf{x}_i)$ for each hypothesized word a_t at frame *t*.
- Compute pseudo loss

$$\varepsilon^k = \frac{1}{2 \mid \mathbf{U} \mid \mid h \mid \mid T_i \mid} \sum_{\mathbf{x}_i \in \mathbf{U}} \sum_{t=1}^{T_i} \sum_{a \neq a_t} (1 - P_{\lambda_k}(a_t \mid \mathbf{x}_i) + P_{\lambda_k}(a \mid \mathbf{x}_i))$$

  where $\mid \mathbf{U} \mid$, $\mid h \mid$ and $\mid T_i \mid$ denote the size of training set, the size of N-best list and the length of utterance *i* respectively.
- Set $c_k = \varepsilon^k / (1 - \varepsilon^k)$
- Calculate new weight for each frame *t* of each utterance

$$w_{i,t}^k = \sum_{a \neq a_t} c_k^{\frac{1}{2}(1 + P(a_t \mid \mathbf{x}_i) - P(a \mid \mathbf{x}_i))}$$

4. Experiments

4.1. Data set and experiment configuration

The corpus used in our study was collected using the CMU Communicator system, a telephone based dialog system that supports planning in a travel domain [7]. The training set has 31,248 utterances, which were collected between April 1998 and November 2000. The test set consists of 1,689 utterances, which were collected during a NIST evaluation conducted in July 2000. In our experiments, the number of acoustic models trained by Boosting is set to 4.

Different from our previous experiments which used semi-continuous HMM, in this case we chose the Carnegie Mellon Sphinx-3 decoder, which is a continuous HMM system, for training and test. The decoding process consists of two passes: the first one is a Viterbi search, and the second one is an A* search that generates N-best list from the word lattice. Final hypothesis is determined by combining the top-1 hypotheses.

4.2. Experimental results

We made a comparative study of conventional Boosting training and our new training scheme. Table 1 shows the performance of Boosting training as a function of the number of acoustic models. Note that the result for K=n means the word error rate obtained from hypotheses combination of n models. As we reported previously, cooperation of multiple models outperforms single model on recognition accuracy. When use 4 acoustic models, the word error rate is down to 13.27% from 14.99%, the number for one model, which represents 11.5% relative reduction. Special attention should be paid to the last two columns which shows that there is no improvement when increase the model numbers from 3 to 4. This indicates that a small set of acoustic models may be sufficient for some speech recognition task, and adding too many models to the system may lead to overfitting.

# models	K=1	K=2	K=3	K=4
W.E.R	14.99%	13.54%	13.31%	13.27%

Table 1 Result of utterance based Boosting training

Table 2 provides the experimental result of our new training scheme. We observe that the frame level scheme also achieves an encouraging improvement compared with single model. When use 4 acoustic models, the word error rate is down to 13.52% from 14.99%, which means 9.8% relative reduction. However, comparison between Table 1 and Table 2 shows that the performance of frame level training scheme is likely somewhat worse than that of conventional Boosting algorithm, even though the frame level performance does not appear to have asymptoted at 4 models. We have some speculations on this outcome in the next section.

# models	K=1	K=2	K=3	K=4
W.E.R	14.99%	14.21%	13.81%	13.52%

Table 2 Result of frame level training scheme

5. Discussion

We proposed a frame level Boosting training scheme for acoustic modeling which aims to address the weaknesses of the conventional Boosting training algorithm. The initial experimental results are encouraging that the new scheme achieves 9.8% relative reduction on Word Error Rate compared with single model. However, it doesn't demonstrate superiority over conventional method. Some of the reasons responsible for the underperformance are listed as follows and we believe that research development on these topics could improve the performance of the new training scheme.

First, the objective function used in the new scheme is based on a metric describing frame level classification error. However, our goal is to build a model with low Word Error Rate which is measured on word level without considering segmentation information. In other words, the mismatch between training criterion and target in principle still exists in the new scheme.

Second, the frame based resample method is exclusively depends on the weight calculated in the training process. As we have known, there is other information helpful for resampling, such as the type and duration distribution of a phoneme [8][9]. For example, to increase the importance of consonant, we could lengthen the duration of consonant while shorten the duration of vowel.

Third, in our experiments, forced-alignment is used to determine the correct word for each frame. We should point out that this method is far from perfect and there is no solid proof to show the obtained result is accurate.

Fourth, as we have analyzed, both utterance and frame based training schemes have their own shortcomings. A potential solution is to combine these two level information in the training. Namely, a new training scheme considering both utterance and frame level recognition errors may be more suitable for acoustic modeling.

Because we are still in the early stages of this research, the interpretation and conclusion made here are tentative and could be changed by subsequent findings.

Acknowledgement

This work was supported under DARPA grant NBCH-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.

6. Reference

[1] R. E. Schapire, "A brief Introduction to Boosting", Proc. of the 16th International Joint Conference on Artificial Intelligence, 1999.

[2] H. Schwenk, "Using Boosting to Improve A Hybrid HMM/Neural Network Speech Recognizer", Proc. of ICASSP 1999.

[3] P. Moreno, B. Logan and B. Raj, "A Boosting Approach for Confidence Scoring", Proc. of EuroSpeech 2001.

[4] S-W Foo and E-G Lim, "Speaker Recognition Using Adaptively Boosted Decision Tree Classifier", Proc. of ICASSP 2002.

[5] C. Meyer, "Utterance-Level Boosting of HMM Speech Recognizers", Proc. of ICASSP 2002.

[6] R. Zhang and A. I. Rudnicky, "Comparative Study of Boosting and Non-Boosting Training for Constructing Ensembles of Acoustic Models", Proc. of Eurospeech 2003.

[7] A. I. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, A. Oh, "Creating Natural Dialogs in the Carnegie Mellon Communicator System", Proc. of EuroSpeech 1999.

[8] J. P. Nedel and R. M. Stern, "Duration Normalization and Hypothesis Combination for Improved Spontaneous Speech Recognition", Proc. of Eurospeech 2003.

[9] Richard Cox, "Enhancing Speech Intelligibility Using Variable Rate Time Scale Modification", CMU ECE Distinguished Lecturer Series, March, 2004.