Improved Discriminative Training Using Phone Lattices

Jing Zheng *Andreas Stolcke*

Speech and Technology Laboratory, SRI International
333 Ravenswood Ave. Menlo Park, CA 94025
{zj,stolcke}@speech.sri.com

Abstract

We present an efficient discriminative training procedure utilizing phone lattices. Different approaches to expediting lattice generation, statistics collection, and convergence were studied. We also propose a new discriminative training criterion, namely, minimum phone frame error (MPFE). When combined with the maximum mutual information (MMI) criterion using *I*-smoothing, replacing the standard minimum phone error (MPE) criterion with MPFE led to a small but consistent win in several applications. Phone-lattice-based discriminative training gave around 8% to 12% relative word error rate (WER) reduction in SRI's latest English Conversational Telephone Speech and Broadcast News transcription systems developed for DARPA's EARS project.

1. Introduction

In recent years, discriminative training algorithms, such as maximum mutual information (MMI) [1, 2] and minimum phone error (MPE) [3], have achieved great success in producing more accurate acoustic models for large vocabulary continuous speech recognition tasks than the conventional maximum likelihood (ML) training algorithm, although they generally have greater computational complexities. In the EARS project, which aims for high-accuracy Conversational Telephone Speech (CTS) and Broadcast News (BN) transcription systems, discriminative training is used by all participating systems. Recently, one of our research efforts has been to leverage very large amounts of imperfectly transcribed data for acoustic model training. For example, both the English CTS and the BN task have more than 2000 hours of data available for acoustic training. The vast amount of training data poses a challenge to the efficiency of the discriminative training procedure, given limited computational resources and development time.

In contrast to ML training, discriminative training needs to consider not only the correct transcriptions, but also competing hypotheses, which are obtained from decoding the training corpus. Parameter estimation typically is carried out in an iterative manner, which makes both statistics collection and convergence rate important factors for overall training time. Recent research [3] shows that combining different training criteria under the framework of *I*-smoothing generally produces better results than use of a single criterion, so the training objective function is still worth exploring. We investigated all these aspects, trying to improve both efficiency and quality of discriminative training.

The rest of the paper is organized as follows: Section 2 introduces our approach for improving training speed on several fronts; Section 3 proposes a new discriminative training criterion, the minimum phone frame error (MPFE), and compares it with the standard MPE criterion; Section 4 shows experimental results based on SRI's latest 20xRT English CTS evaluation system; Section 5 summarizes the paper.

2. Improved training efficiency

The major computation load of a typical discriminative training procedure lies in two parts: decoding the training data and collecting statistics for parameter estimation. One or both parts need to be performed iteratively.

2.1. Lattice generation and statistics collection

In the standard MMI and MPE training procedure [2, 3], competing hypotheses are represented as word lattices, in which phone boundaries are marked in each word arc to constrain the search during the statistics collection. The word lattices are typically rendered with unigram language model scores. To ensure that the lattices have enough richness of competing hypotheses, which is important for training effectiveness, the pruning threshold should not be set very tight when lattices are generated. This makes decoding computationally expensive, especially when a large amount of training data must be processed. For large corpora, the lattices usually occupy a large amount of disk space. The advantage of using these lattices is that statistics collection can be very fast, especially with the "exact boundary approach" [2], which assumes that phone boundaries do not move over multiple training iterations. Although this assumption is generally not true, almost no loss was observed with this approximation [2].

To address the lattice storage issue, Huang et al. [4] proposed an "implicit-lattice" method for MMI training. They observed that if only a unigram language model (LM) is needed, word symbols are not necessary in decoding a graph to encode LM probabilities. They dropped word symbols, and compiled the pronunciation dictionary into a highly compact decoding graph, with LM probabilities embedded. This compact decoding graph allows fast lattice generation in memory. Statistics for MMI training can be collected on the fly from these lattices, which do not need to be stored. Therefore, disk space can be saved. The internal lattices are regenerated during every training iteration with updated model parameters, thus making the training procedure more optimal. It also allows the use of a tighter pruning beam width for a faster training speed. Implicit-lattice MMI training reportedly gives results comparable to those of the standard MMI training. However, the statistics collection is much slower than the "exact boundary" approach based on pregenerated lattices, since search is less constrained. In addition, a generalization of implicit-lattice MMI to MPE remains to be developed.

Because of time and resource limitations, we want to speed up both the lattice generation and statistics collection procedures, and therefore propose phone-lattice-based

discriminative training, which is applicable for both MMI and MPE. Similar to implicit-lattice MMI, we compile the word-dropped pronunciation dictionary into a determinized and minimized [5] finite-state network, with both pronunciation and unigram language model probabilities embedded. On this finite-state network, we generate phone lattices in a very fast decoding pass, using an algorithm similar to that described in [6]. In a phone lattice, each arc represents a phone, with start and end time information. With a forward-backward search pass constrained by the timing of the phone arcs, statistics for both MMI and MPE can be collected in the standard manner [3].

The phone-lattice-based method has two advantages over the word-lattice-based method. First, phone lattice generation incurs less computation because of the highly optimized decoding network. In generating word lattices, word symbols cannot be omitted, which makes for a much larger decoding network and slower speed. Second, the phone lattice is more efficient in representing competing hypotheses in terms of phonetic difference, which is important in phone-based HMM training. Therefore, with the similar richness of phonetic variation, phone lattices can be a smaller size, and can be generated with a tighter beam width compared to word lattices. In practice, we used a much tighter pruning threshold to generate these phone lattices than in normal word decoding. To further reduce the space and computation needed for statistics collection, NULL arcs were introduced into the lattice to more efficiently represent the overlapping phone arcs. Compared to implicit-lattice MMI, the phone-lattice approach is much faster in statistics collection, and can be naturally used in both MMI and MPE training. However, phone lattices still require a large amount of storage space, albeit less space than the word lattices.

2.2 Training procedure

The standard MMI and MPE approaches use an iterative optimization algorithm to estimate model parameters; therefore, convergence speed plays an important role in training time. Below we discuss strategies to shorten convergence time. Recent research showed that *I*-smoothing with different prior models can help boost the effectiveness of discriminative training [3]. As we can collect statistics needed for ML, MMI, and MPE training simultaneously, at the end of each training iteration, we can update ML, MMI, and MPE models at the same time. With *I*-smoothing, we investigated the following combinations:

- MMI/ML: MMI model with ML prior.
- MPE/ML: MPE model with ML prior.
- MPE/MMI: MPE model with MMI prior, which itself is *I*-smoothed with ML model.
- MMI/MPE: MMI model with MPE prior, which itself is *I*-smoothed with ML model.
- MPE+MMI: Alternating MMI and MPE criteria during training. For odd-numbered iterations, estimate MPE model with MMI prior; for even-numbered iterations, estimate MMI model with MPE prior. The prior models themselves are *I*-smoothed with ML models.

Table 1 gives results with the different approaches for training a BN gender-independent crossword model on a 400-hour corpus consisting of BN and TDT4 data. The acoustic features were the perceptual linear prediction (PLP) cepstrum with heteroscedastic linear discriminant analysis (HLDA) projection. Intermediate models at all iterations were tested using a trigram language model on the Dev04 test set, consisting of about 3 hours of BN speech. Word error rate (WER) is plotted as a function of the training iteration until a minimal WER is reached, as shown in Figure 1.

Table 1: Results of different training methods

Method	WER (%)	Rel. reduction (%)
MLE	16.0	N/A
MMI/MLE	14.5	-9.4
MPE/MLE	14.8	-7.5
MMI/MPE	14.0	-12.5
MPE/MMI	14.0	-12.5
MPE+MMI	14.0	-12.5

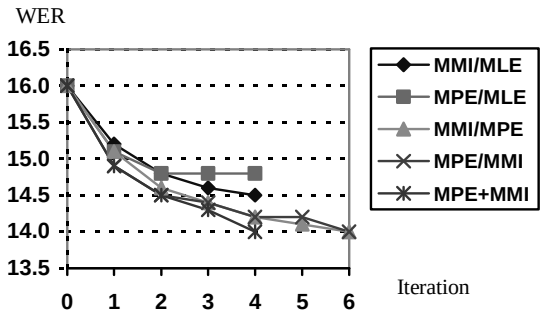

Figure 1: WER as function of training iterations.

As can be seen, combining the MPE and MMI criteria led to a lower WER than using the MPE criterion or MMI criterion alone. MPE+MMI, which alternates the MPE and MMI criteria across iterations, reaches lowest WER two iterations faster than either MPE/MMI or MMI/MPE, although the final WER stays the same. In light of this result, we developed most of our acoustic models with the strategy of alternating different discriminative training criteria across multiple iterations.

3. Minimum Phone Frame Error (MPFE)

The objective function of the standard MPE criterion [3] is defined as

$$F_{MPE}(\lambda) = \sum_{r=1}^{R} \sum_{s} P_k(s \mid O_r, \lambda)\ RawPhoneAccuracy(s) \quad (1)$$

where $P_k(s \mid O_r, \lambda)$ is the posterior probability of hypothesis s for utterance r given observation O_r, current parameter set λ, and acoustic scale k. *RawPhoneAccuracy*(s) is a measure of the number of correctly transcribed phones in hypothesis s, which is typically computed as the sum of a phone accuracy measure of all phone hypotheses in s:

$$RawPhoneAccuracy(s) = \sum_{q \in s} PhoneAcc(q) \quad (2)$$

where

$$PhoneAcc(q) = \begin{cases} -1 + 2e & \text{if } q \text{ is correct in label} \\ -1 + e & \text{otherwise} \end{cases} \quad (3)$$

where e is the overlap ratio between q and its corresponding phone in the reference transcription.

We feel that the original MPE criterion has some shortcomings. First, it does not seem to sufficiently penalize deletion errors. Suppose that a reference transcription has 20 phones; a hypothesis with only one phone receives -1 *RawPhoneAccuracy* maximally, which seems to be too little. In general, the standard MPE objective function discourages insertion error more than deletion error. Second, the dynamic range of *RawPhoneAccuracy* is typically quite narrow, which makes MPE occupancies considerably lower than MMI occupancies [3]. This may lead to an MPE robustness problem when training data are not abundant.

To address these issues, we proposed a different phone accuracy definition:

$$PhoneFrameAcc(q) = \sum_{t=start(q)}^{end(q)} P(s_t \in S(q) \mid W, O) \quad (4)$$

where q is the phone hypothesis under study; $S(q)$ denotes the set of HMM states associated with this phone; $start(q)$ and $end(q)$ represent the start and end times of q in frame units; $P(s_t \in S(q) \mid W, O)$ is the posterior probability of the HMM state belonging to $S(q)$ at time t given observations O and transcription W, which can be obtained with the standard forward-backward algorithm that is widely used in HMM training.

Applying the accuracy measure defined by (4), we get the MPFE criterion:

$$F_{MPFE}(\lambda) = \sum_{r=1}^{R} \sum_{s} P_k(s \mid O_r, \lambda)\, FPhoneAccuracy(s) \quad (5)$$

where

$$FPhoneAccuracy(s) = \sum_{q \in s} PhoneFrameAcc(q) \quad (6)$$

is a measure of the number of frames having a correct phone label in hypothesis s. Maximizing Equation 6 leads to MPFE training. Because of the similarity in definition, MPFE can use the same algorithm as MPE except for the difference of measuring the hypothesis accuracy. Since all the competing hypotheses in a lattice have the same number of frames, MPFE does not have a systematic bias favoring deletion error. We also observed that the defined *FPhoneAccuracy* has a larger dynamic range than the *RawPhoneAccuracy* defined in Equation 2, which makes MPFE occupancies have values similar to those of MMI occupancies. This may make MPFE more robust than MPE when dealing with a small amount of data.

Table 2 compares two English CTS crossword gender-dependent models trained on about 1400 hours Switchboard and Fisher data, with MPE+MMI and MPFE+MMI, respectively. A 39-dimension speaker-adaptive training (SAT) [12] transform normalized Mel frequency cepstral coefficients (MFCC) and voicing feature with HLDA projection [7] was used for training. A bigram multiword language model was first used to generate lattices, and then a 4-gram language model used to rescore lattices. Final hypotheses were generated from consensus decoding [8, 9]. LM weight, word penalties, and so on were optimized in the Dev04 data set, and applied to Eval01, Eval02, Eval03, and Eval04 test sets, which were used for official NIST evaluation from 2001 through 2004. As can be seen, the MPFE+MMI approach has a small but consistent advantage over MPE+MMI on different test sets, ranging from 0.2% to 0.6% absolute.

Table 2: English CTS results in WER (%)

	Eval04	Eval03	Eval02	Eval01
MLE	24.5	25.3	26.7	26.1
MPE+MMI	22.6	23.7	24.9	24.4
MPFE+MMI	22.4	23.1	24.5	24.0

Table 3: Mandarin CTS results in CER (%)

MLE	MMI	MPE+MMI	MPFE+MMI
41.2	39.6	39.4	38.9

Table 3 compares two Mandarin CTS crossword models trained on about 100 hours of data using 42-dimension MFCC plus a 3-dimension pitch feature with SAT normalization. All models were first adapted some earlier decoding pass hypotheses using maximum likelihood linear regression (MLLR), and then used to rescore trigram lattices to produce 1-best hypotheses for character error rate (CER) scoring. The Mandarin CTS development data set Dev04 was used for testing. Results show little improvement from MMI to MPE+MMI, but a decent WER reduction of 0.7% absolute from MMI to MPFE+MMI. This may indicate that MPFE is more robust than MPE for this small training corpus.

4. Evaluation System Experiments

Figure 2 illustrates the flowchart of SRI's 20xRT English CTS system for the 2004 NIST CTS evaluation. We used two sets of front end configurations. The first is a standard MFCC-plus-voicing feature [7], HLDA projecting to 39-dimension space, and then concatenated with a 25-dimension multilayer perceptron (MLP)-based feature from ICSI [10], forming a 64-dimension feature vector. The resulting feature was then adapted with vocal tract length normalization (VTLN), speaker-level mean and variance normalization, and SAT. This front end is abbreviated as MEL in Figure 2. The second configuration is the standard 39-dimension PLP feature with VTLN, mean and variance normalization, and SAT, abbreviated as PLP. We trained three sets of models: within-word with MEL front end (MEL-WW), crossword with MEL front end (MEL-CW), and crossword with PLP front end (PLP-CW). All these models were trained on about 1400 hours of Switchboard and Fisher data in a gender-dependent manner. MEL and PLP models were trained on two disjoint subsets of Fisher data (half each) to increase difference and to shorten training time. Decoding was organized in three passes as indicated by the dashed lines in Figure 1. MEL and PLP models were cross-adapted with each other's decoding hypotheses to avoid error reinforcement.

Figure 2: Flow chart of SRI's 20xRT English CTS system for NIST-2004 evaluation.

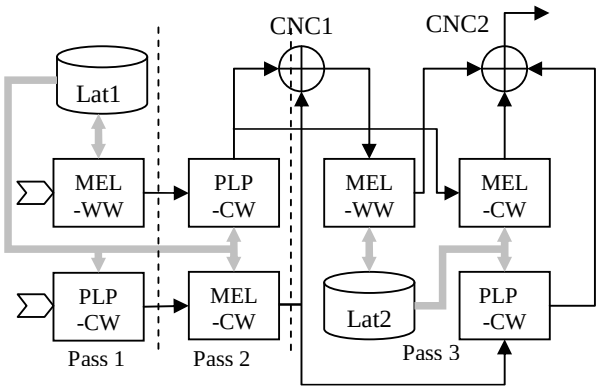

Table 4: English CTS system results on Dev04

Feature	Train	WER	Rel. Δ
With MLP Feature	MLE	16.7	-
	MPE+MMI	15.5	-7.8%
	CW-MPFE+MMI	15.3	-8.4%
Without MLP Feature	MLE	18.3	-
	MPE+MMI	16.8	-8.2%
	CW-MPFE+MMI	16.4	-10.4%

MEL-WW was used to generate two sets of lattices, Lat1 and Lat2, for crossword models to rescore. Language models of different sizes and N-gram orders were used in the different decoding stages. More detailed information can be found in our system description [12].

For comparison, we trained all the models with the ML and MPE+MMI approach. We also retrained MEL-CW and MPE-CW with the MPFE+MMI approach. Table 4 compares the final system output on Dev04. As we can see, the MPE+MMI models gave about 7.8% relative WER reduction over ML models. And retraining crossword models with MPFE+MMI gave another 0.2% absolute, with an 8.4% total improvement over ML models.

As ICSI features were extracted with MLPs, which are themselves trained discriminatively, we suspected that this may reduce the benefit from model-based discriminative training. We therefore removed the ICSI features from the MEL front-end configuration, and reran all the experiments. The improvement from discriminative training indeed increased. MPE+MMI gave an 8.2% WER reduction, and MPFE+MMI gave 10.4%.

5. Conclusions

We described an efficient discriminative training approach using phone lattices, which can be generated more efficiently than word lattices. Alternating discriminative training criteria between training iterations was found to help accelerate convergence, and therefore reduce the necessary training time to reach minimal WER. We proposed an improved discriminative training criterion, Minimum Phone Frame Error. Experiments showed that MPFE+MMI gave a small but consistent win over MPE+MMI in several experiments. The proposed phone-lattice-based discriminative training gave more than a 10% relative improvement to SRI's English CTS system when no discriminative feature was applied.

6. Acknowledgment

This work was funded by DARPA under contract No. MDA972-02-C-0038. Distribution is unlimited. We thank our colleagues at SRI, ICSI, and U. Washington for help in building the recognition systems.

7. References

[1] L.R. Bahl, P.F. Brown, P.V. de Souza and R.L. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," *Proc. ICASSP'86*, pp. 49-52, Tokyo, 1986.

[2] P.C. Woodland and D. Povey, "Large Scale Discriminative Training of Hidden Markov Models of Speech Recognition," *Computer Speech & Language*, Vol. 16, pp. 25-47, 2002.

[3] D. Povey and P.C. Woodland, "Minimum Phone Error and *I*-Smoothing for Improved Discriminative Training," *Proc. ICASSP'02*, Orlando, 2002.

[4] J. Huang, B. Kingsbury, L. Mangu, G. Saon, R. Sarikaya and G. Zweig, "Improvements to the IBM Hub-5E System," *Proc. NIST RT-02 Workshop*, 2002.

[5] M. Morhri, "Finite-State Transducers in Language and Speech Processing," *Computational Linguistics*, 23(2): 269-312, 1997.

[6] A. Ljolje, F. Pereira and M. Riley, "Efficient General Lattice Generation and Rescoring," *Proc. EUROSPEECH'99*, Budapest, 1999.

[7] M. Graciarena, H. Franco, J. Zheng, D. Vergyri and A. Stolcke, "Voicing Feature Integration in SRI's Decipher LVCSR System," *Proc. ICASSP'04*, Vol. 1, pp. 921-924, Montreal, 2004.

[8] L. Mangu, E. Brill and A. Stolcke, "Searching for Consensus to Improve Recognition Output," Proc. Hub-5 Conversational Speech Recognition Workshop, Linthicum Heights, 1998.

[9] G. Evermann and P.C. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination," *Proc. Speech Transcription Workshop*, College Park, 2000.

[10] Q. Zhu, A. Stolcke, B. Chen and N. Morgan, "Incorporating TANDEM/HATS MLP Features into SRI's Conversational Speech Recognition System," *Proc. EARS RT04 Workshop*, Palisades, 2004.

[11] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech and Language*, Vol. 12, pp 75-98, 1998.

[12] SRI system description, *EARS RT04 Workshop*, Palisades, 2004.