Frame margin probability discriminative training algorithm for noisy speech recognition

Hao-Zheng Li, Douglas O'Shaughnessy

INRS-EMT University of Quebec
{lihz,dougo}@emt.inrs.ca

10.21437/Interspeech.2007-7

Abstract

This paper presents a novel discriminative training technique for noisy speech recognition. First, we define a Frame Margin Probability (FMP) which denotes the difference of score of a frame on its right model and on its competing model. The frames with negative FMP values are regarded as confusable frames and the frames with positive FMP values are regarded as discriminable frames. Second, the confusable frames will be emphasized and the overly discriminable frames will be deweighted by an empirical weighting function. Then the acoustic model parameters are tuned using the weighted frames. By this kind of weighting, the confusable frames, which are often noisy, can contribute more to the acoustic model than those without weighting. We evaluate this technology using the Aurora standard database (TIdigits) and HTK3.3, and obtain a 15.9% WER reduction for noisy speech recognition and a 13.13% WER reduction for clean speech recognition compared with the MLE baseline systems.

Index Terms: Hidden Markov Model, discriminative training, frame margin probability.

1. Introduction

In recent decades, discriminative training has been widely used in speech recognition and it has been proved to give significant improvement over the traditional maximum likelihood estimation (MLE) method on clean speech data. The widely used discriminative training methods are minimum classification error (MCE) [1] and maximum mutual information (MMI) [2, 3] training. However, there are few reports focusing on applying discriminative training for noisy speech recognition. The main reason for this is that the existing discriminative training are concerned with the overfitting problem and have poor generalization capability to unseen data, especially in noisy conditions.

This paper proposes a novel discriminative training algorithm that is suitable for noisy speech recognition. Unlike the MCE, which directly gears at minimizing an approximation word error rate function, and unlike the MMI, which targets at optimizing mutual information between an utterance and the correct string, we propose to try to decrease the number of the confusable frames. In isolated word speech recognition, the discriminable frames in a word can counteract those confusable frames' influence and thus make a correct recognition. However, in continuous speech recognition, there is no obvious boundary between speech units and the confusable frames may cause insertion, substitution or deletion errors. Moreover, in multi-condition training, i.e., the training utterances include clean and different SNR noisy utterances, the clean utterances are less variant and are predominant over the noisy utterances which may lead to the noisy utterance not appropriately modeled. This paper will focus on how to determine the confusable frames by using frame margin probability (FMP) and then trying to make the confusable frames become discriminable by applying a weighting function on them.

Margin probabilities is used for sequence learning in [4], and [5] tries to maximize a multi-class separation margin probability of a word for speech recognition. This paper proposes frame margin probabilities (FMPs) and uses it in a different way in that it is used to determine which frames are confusable and which frames are discriminable. By emphasizing those confusable frames and deweighting those overly discriminable frames, the number of confusable frames can be decreased and an acoustic model can learn more possible confusions with the hope that these confusions will be seen in the test data.

In the following section we briefly review the basics of the hidden Markov model and MLE. In Section 3 we define the FMP and show how to use it to reduce the confusable frames. We show our experimental results in Section 4, which is followed by a conclusion in Section 5.

2. Hidden Markov Models and MLE

In an HMM with N underlying states, a state sequence $S = (S_1, S_2, \cdots, S_T)$ generated by the Markov chain cannot be directly observed, but only through the observation sequence $Y = (Y_1, Y_2, \cdots, Y_T)$ result from the state sequence according to the observation distribution defined by $B = \{b_i(Y_t) : 1 \leq i \leq N\}$, with $b_i(Y_t) = P(Y_t|S_t = i)$. The transi-

tion from state i to state j is specified by an $N \times N$ matrix $A = [A_{ij}]$ with $A_{ij} = P(S_t = j|S_{t-1} = i)$. $\pi = [\pi_1, \pi_2, \cdots, \pi_N]$ is the initial state probability vector with $\pi_i = P(S_1 = i)$. λ is the compact notation for the model parameters in an HMM.

Maximum likelihood estimation (MLE) is to optimize A, B and π to maximize the probability of observation sequences in the training set. By defining the a posteriori probability variable

$$\gamma_t(i) = P(S_t = i|\lambda, Y)$$

which is the probability of being in state i at time t, given the observation sequence Y and the model λ, the Baum-Welch algorithm uses it to do soft alignment for the training utterance and assigns the aligned speech to the hidden states for adjusting the parameters in an iterative procedure [6].

3. FMP discriminative training

To implement our proposed FMP discriminative training method, we define the frame a posteriori probability variable

$$P(Y_t|\lambda, Y) = \sum_{i=1}^{N} \gamma_t(i) b_i(Y_t)$$

which is the output probability of the tth frame, given the observation sequence Y and the model λ. Following [4], FMP is similarly defined as:

$$d(Y_t) = P(Y_t|\lambda, Y) - \max_{\lambda_j \in \Omega, \lambda \neq \lambda_j} P(Y_t|\lambda_j, Y)$$

where Ω denotes the HMM set. FMP denotes the difference of frame a posteriori probability on its right model and its competing model. In practice, we use log-likelihood instead of original likelihood in most HMM-based speech recognition system likelihoods for convenience. A frame with a high margin probability is discriminable while with negative margin probability is confusable. An utterance with all the frames having positive margin probability will be certainly correctly recognized. However, a large part of utterances may contain a portion of frames having negative FMP values, even if they can be correctly recognized. In isolated speech recognition, if we use whole utterance margin probability, i.e., the summation of FMP of all frames is positive, the speech unit can be correctly recognized. In this situation, the discriminable frames in a speech unit can counteract those confusable frames' influence. However, in continuous speech recognition, there no obvious boundary between speech units, and the confusable frames may cause errors. On the other hand, those frames with too large FMP values may pull the decision boundary too close to themselves to deviate the true decision boundary. So emphasizing the confusable frames and deweghting the too discriminative frames may push some confusable frames to the right decision region, with the objective of decreasing the recognition errors.

As for the frames that are far from the decision boundary and have very negative FMP values, these frames are perhaps misaligned, are trash frames, or are located in other speech units' decision regions, and emphasizing them may lead to new errors and thus should be avoided.

Based on the above discussion, the frames with high margin probability should be deweighted, the frames with negative margin probability in a certain range should be weighted and frames with very negative FMP values should be avoided. A function $f(x)$ that is very similar to the gamma function satisfies the above criterion:

$$f(x) = (x+a)^k \frac{e^{-(x+a)/b}}{C}, (x > -a, k > 1),$$

whose figure is shown in Fig 1, where k and b control how much each frame should be weighted; and $-a$ is a rejection threshold and the frames whose FMP values are less than it will be rejected; and C is a normalization factor.

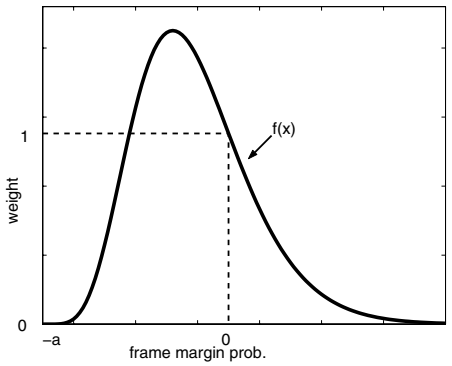

Figure 1: A weighting function

As for the HMMs parameters estimation, the only change required is to replace the $\gamma_t(i)$ with

$$\gamma_t'(i) = \gamma_t(i) f(d(Y^t)),$$

compared with the estimation equations in the literature [6]. While this is mathematically the only change required for an HMMs parameter estimation, one practical issue that needed to be considered is that the weight becomes too small when the FMP is large, which may result in that the frames with high FMP contribute too little to the model and some new confusable frames are generated; so a floor for the weight is needed.

As for the competing model in the HMM set Ω, if a task is isolated-word recognition, the number of competing models is limited. However, in continuous speech recognition, there are large numbers of possible competing models due to an unknown number of possible words in each utterance,

which makes it impossible to enumerate over them. In this paper, we use a lattice to represent the competing models in a compact way. In HTK [7], a lattice is made up of a set of nodes with arcs that connect them together, the nodes corresponding to time and arcs associated with word hypothesis and Fig. 2 is an example of it. A lattice can be generated by a recognition tool and at this rate it includes the correct transcription which is marked in bold line in Fig. 2. At a time slot, we select the word in the correct hypothesis as the correct model and select the words in incorrect hypotheses but having a different transcription from the the correct word as the competing model. For example, in Fig 2, for the word *one* in the correct hypothesis, its competing words are *three* and *four*; for the word *two* in the correct hypothesis, its competing words include several words that have time overlap with it but not include the word *two* in the incorrect hypothesis. Since the proposed discriminative training

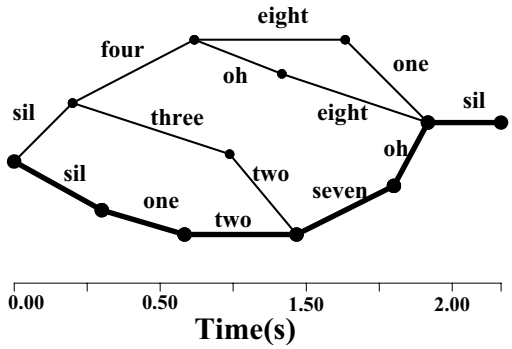

Figure 2: A lattice example

will be applied to a noise environment, it bootstraps directly from MLE models, which may provide relative good alignment in noise environment.

4. Experimental Results

We have carried out several experiments to test the proposed discriminative training method on a connected digit recognition task using the Aurora 2 database [8] and the HTK3.3 engine [7]. In the Aurora 2 database, speech is sampled at 8k Hz. Training and testing follow the specifications described in [8]. A word-based ASR system for digit string recognition where each HMM word model has 16 emitting states is adopted. A three-state silence model and a one emitting state short pause model are used. Training is done with 8440 multi-condition utterances under realistic background noise (subway, babble, car, exhibition hall) at various SNRs (clean, 20, 15, 10, 5 dB) from 55 male and 55 female adults. Testing data include eight types of realistic background noise (subway, babble, car, exhibition hall, restaurant, street, airport and train station noise) at various SNRs (clean, 20, 15, 10, 5, 0, and -5 dB) and are divided into two test sets. Set A contains the first four types of noise which are the same as that in the training set and Set B contains the other four types noise, which are different from that in the training set. Each of these two test sets has 4004 utterances. The baseline feature is MFCC_E_D_A, which contains 12 MFCCs and log energy, and their first and second derivatives, and is also computed with the program HTK3.3. Each feature vector thus contains 39 components.

Table 1 shows the word accuracy of the baseline and Table 2 shows the word accuracy and the relative error rate reduction of the proposed discriminative training algorithm. Note that the average accuracy and relative error rate reduction are computed with the results between 20 and 0 dB, as suggested in [8] for the Aurora 2 database. The average recognition results for Set A are improved from 88.37% to 90.22%, corresponding to a 15.9% WER reduction, which means the proposed algorithm is more robust in the matched noisy type condition. The improvement can be explained by the fact that the number of confusable frames is reduced after emphasizing the confusable frames and deweighting the overly discriminative frames, which can be seen in Fig. 3, where the solid line represents the histogram of FMP for MLE, and the dashed line represents the histogram of FMP for proposed discriminative training method on the 5 dB subway test utterances. It can be readily seen from this figure that the number of confusable frames (with negative FMP) is decreased after frame weighting.

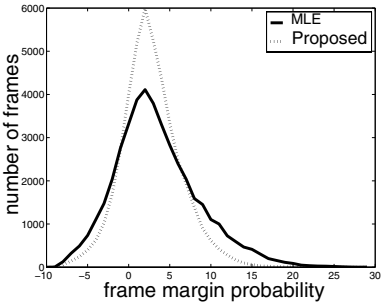

Figure 3: Frame Margin Probability histogram for MLE and the proposed discriminative training method.

We also test whether the proposed method can be generalized to unmatched noise type condition. For Set B, which has the different noise type from the training set, the improvement is observed from 87.49% to 88.39%, corresponding to a 7.17% WER reduction from Table 3 and Table 4.

It also can be seen from these tables that the proposed method can bring a 13.13% relative WER reduction in clean speech, which means the proposed training method is suitable for clean conditions as well as noise conditions.

Table 1: Baseline MLE recognition results for Aurora 2 database Test A

Test A	subw.	babb.	car	exhi.	aver.
clean	98.62	98.73	98.48	98.61	98.61
20dB	97.97	97.76	97.91	97.41	97.76
15dB	96.90	97.25	97.70	96.76	97.15
10dB	95.24	95.71	96.00	93.61	95.14
5dB	89.93	88.15	88.25	86.55	88.22
0dB	69.91	62.94	55.62	65.87	63.59
-5dB	32.39	28.42	19.98	27.99	27.20
aver.	89.99	88.36	87.10	88.04	88.37

Table 2: FMP discriminative training method recognition results for Aurora 2 database Test A

Test A	subw.	babb.	car	exhi.	aver.	rela.
clean	98.62	98.76	98.93	98.86	98.79	13.13
20dB	98.13	98.28	98.36	97.25	97.98	10.84
15dB	97.57	97.79	97.88	96.67	97.30	11.41
10dB	96.07	96.07	96.54	93.74	95.39	9.57
5dB	91.89	89.60	90.84	88.95	89.88	17.83
0dB	75.50	64.12	65.82	73.34	70.36	16.78
-5dB	40.16	28.17	24.25	37.70	34.14	7.38
aver.	91.83	89.17	89.89	89.99	90.22	15.90
rela.	18.40	6.60	21.64	16.30	15.90	–

Table 3: Baseline MLE recognition results for Aurora 2 database Test B

Test B	rest.	stre.	Airp.	stat.	aver.
clean	98.62	98.73	98.48	98.61	98.61
20dB	97.61	97.79	97.76	97.32	97.62
15dB	95.49	96.86	96.75	95.96	96.27
10dB	92.39	95.13	94.63	93.67	93.96
5dB	85.14	87.24	88.91	85.07	86.59
0dB	62.88	63.15	68.27	57.76	63.02
-5dB	28.62	27.51	31.46	21.60	27.30
aver.	86.70	88.03	89.26	85.96	87.49

Table 4: FMP discriminative training method recognition results for Aurora 2 database Test B

Test B	rest.	stre.	Airp.	stat.	aver.	rela.
clean	98.62	98.76	98.93	98.86	98.79	13.13
20dB	98.00	97.94	97.70	97.69	97.71	8.93
15dB	96.75	96.92	96.42	96.11	96.42	7.63
10dB	94.11	95.41	93.65	93.43	93.69	3.23
5dB	86.74	88.88	87.89	85.87	87.12	5.63
0dB	65.03	68.29	69.22	61.68	65.03	8.62
-5dB	31.16	32.01	32.09	23.79	28.06	3.39
aver.	88.13	89.49	88.98	86.96	88.39	7.17
rela.	10.71	12.15	-2.68	7.12	7.17	–

5. Conclusions and Future Work

In this paper, we propose a discriminative training method that tries to reduce the confusable frames in order to increase the recognition accuracy. We use FMP to determine which are confusable frames and which are discriminative frames. By emphasizing the confusable frames and deweighting those overly discriminative frames, the number of confusable frames number can be decreased. Through a series of experiments on a connected digit recognition task using the Aurora 2 database, we find the proposed discriminative training method is more robust in noise conditions as well as in clean conditions than the MLE method.

Since the weighting function in this paper is critical and empirical, further investigation would include finding more effective ways to weight the frames. Also, more research and experiments on a sub-word (phone) based system for large vocabulary will be conducted in the future.

6. References

[1] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recogntion," *IEEE Tran. SAP*, vol. 5, no. 3, pp. 257–265, 1997.

[2] L. R. Bahl, P. V. de Souza, P. F. Brown and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. ICASSP*, pp. 49–52, 1986.

[3] Y. Normandin, R. Cardin and R. Demori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Tran. SAP*, vol. 2, no. 2, pp. 299–311, 1994.

[4] Y. Altun and T. Hofmann, "Large margin method for label sequence learning," *Proc. Europspeech*, pp. 993–996, 2003.

[5] X. Li, H. Jiang and C. Liu, "Large margin HMMs for speech recognition," *Proc. ICASSP*, pp. 513–516, 2005.

[6] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, pp. 257–286, 1989.

[7] S. Young, et. al. "The HTK Book for HTK V3.3," *http://htk.eng.cam.ac.uk*, 2005.

[8] H.-G. Hirsch and D. Pearce, "The AURORA experiment framework for the performance evaluation of speech recognition systems under noisy condition," *Proc. ASR*, pp. 182–188, 2000.