EVALUATION OF PROPOSED MODIFICATIONS TO MPE FOR LARGE SCALE DISCRIMINATIVE TRAINING

Daniel Povey, Brian Kingsbury

IBM T.J. Watson Research Center
Yorktown Heights, NY, USA
{dpovey,bedk}@us.ibm.com

ABSTRACT

Minimum Phone Error (MPE) is an objective function for discriminative training of acoustic models for speech recognition. Recently several different objective functions related to MPE have been proposed. In this paper we compare implementations of three of these to MPE on English and Arabic broadcast news. The techniques investigated are Minimum Phone Frame Error (MPFE), Minimum Divergence (MD), and a physical-state level version of Minimum Bayes Risk which we call s-MBR. In the case of MPFE we observe improvements over MPE. We propose that the smoothing constant used in MPE should be scaled according to the average value of the counts in the statistics obtained from these objective functions.

Index Terms— Minimum Phone Error, Discriminative Training, Minimum Phone Frame Error, Minimum Divergence, Minimum Bayes Risk

1. INTRODUCTION

This paper reviews three different proposed modifications to Minimum Phone Error (MPE) for discriminative training, and evaluates them on two large vocabulary broadcast news tasks. Section 2 introduces MPE. Section 3 introduces the various proposed modifications and explains our implementation of them; Section 4 explains our approach to setting the I-smoothing constant τ for these objective functions. Section 6 explains the experimental conditions, and Section 7 gives the results and conclusions.

2. INTRODUCTION TO MPE

Minimum Phone Error (MPE) is an objective function for discriminative training of acoustic models [1, 2]. We try to minimize the phone-level Levenshtein distance between the training data when recognized with our acoustic models, and the correct transcript. Because this function would not be differentiable, we actually minimize the *expected* Levenshtein distance - a sum over all possible transcripts of the training data weighted by the likelihood of each transcript given the acoustic model. To improve generalization we also introduce an acoustic scale and use only a unigram language model to recognize the training data so that more of it is confusable. To get around the difficulty of computing the Levenshtein distance in a lattice framework we actually maximize an approximation to the raw phone accuracy (defined as the number of phones in the correct transcript minus the Levenshtein distance to the reference) where the contribution of each phone to the accuracy is based on the extent of its overlap in time with reference phone segments.

This work was funded by DARPA contract HR0011-06-2-0001

This local contribution to the approximated accuracy is:

$$\mathrm{PhoneAcc}(q) = \max_z \left\{ \begin{array}{l} -1 + 2e(q,z) \text{ if z, q are same phone} \\ -1 + e(q,z) \text{ if different phones} \end{array} \right\}, \quad (1)$$

where q is the phone arc (a phone instance in the lattice), z is a reference phone arc and $e(q,z)$ is the length of overlap between q and z, as a fraction of the length of z. It is easy to show that if the phones are time-aligned in the appropriate way this equals the number of phones in the reference minus the Levenshtein distance.

The MPE objective function is optimized using the extended Baum-Welch equations using numerator and denominator statistics accumulated from the training data. The process of going from the local function $\mathrm{PhoneAcc}(q)$ of each phone arc to the statistics that are required to be accumulated, involves two forward backward passes over the denominator lattice (or "recognition lattice"), as described in [1].

An important feature of MPE is "I-smoothing", which consists of adding τ points (sometimes we refer to this as τ^I) of average ML statistics to the "numerator" statistics of MPE. This can be thought of as a prior centered around the ML estimate of each Gaussian [2]. Without I-smoothing MPE does not give improvements over MMI. In the current implementation of MPE, we also accumulate statistics for an MMI update from the numerator (reference) and denominator (recognition) lattice. We use these with the Extended Baum-Welch equations to obtain MMI-updated means and variances on each iteration, and when we do I-smoothing we smooth back to the "dynamic MMI prior" formed by the MMI estimate rather than the ML estimate. This has previously been found to improve performance, and was mentioned in [3].

3. MODIFICATIONS TO MPE

All of these modifications to MPE have in common the feature that they only change the very first stage of the MPE calculation, where we compute $\mathrm{PhoneAcc}(q)$ which is the contribution of each phone to the accuracy of complete sentences which include it. Although the modified expression will no longer be a phone accuracy, we still refer to it as $\mathrm{PhoneAcc}(q)$ to make it clear where it fits into the MPE calculations described in [1]. All of these modified versions of MPE have an expression for $\mathrm{PhoneAcc}(q)$ which can be computed as a sum over contributions from individual frames (unlike Equation 1) so it is no longer necessary to do the calculation on a phone level. Instead if we wanted we could define the arcs q in the lattice on a frame-by-frame or word-by-word level, according to convenience. It would even be possible to do a full forward-backward pass over the recognition HMM rather than fixing the phone boundaries in the lattice as is necessary for MPE. However, for clarity and compatibility with the old MPE implementation we retain the use of phone-level lattice arcs.

Objective Function	Expression for $\mathrm{PhoneAcc}(q)$	Notation
MPE	$\max_z \left\{ \begin{array}{l} -1 + 2e(q,z) \text{ if z and q are same phone} \\ -1 + e(q,z) \text{ if different phones} \end{array} \right\}$.	$\mathrm{q, z}$ are phone arcs in $(\mathrm{recognition, reference})$ lattice; $\mathrm{e(q, z)}$ is extent of their time overlap / z's length
MD	$\sum_{t=start_q}^{end_q} - \min_z D(s_q(t), s_z(t))$	$s_q(t)$ and $s_z(t)$ are acoustic states in arcs at time t $D(\cdot)$ is approx KL divergence of states, see text
g-MD	$\sum_{t=start_q}^{end_q} - \min_z D(s_q(t), s_z(t), \mathbf{x}(t))$	$s_q(t)$ and $s_z(t)$ are acoustic states in arcs at time t $D(\cdot)$ is approx KL divergence of top Gaussians
MPFE	$\sum_{t=start_q}^{end_q} \left\{ \begin{array}{l} 1 \text{ if } p(q) = p(z) \text{ for any z overlapping t} \\ 0 \text{ otherwise} \end{array} \right.$	start_q and end_q are start and end frames of arc q $p(q), p(z)$ are phone identities of arcs q, z
s-MBR	$\sum_{t=start_q}^{end_q} \left\{ \begin{array}{l} 1 \text{ if } \exists z : s_q(t) = s_z(t) \\ 0 \text{ otherwise} \end{array} \right.$	start_q and end_q are start and end frames of arc q $s_q(t)$ and $s_z(t)$ are acoustic states in arcs at time t

Table 1. Expressions for $\mathrm{PhoneAcc}(q)$ in the various objective functions

Previously proposed objective functions

Minimum Divergence was indroduced in [4] and intended initially for systems where there is no clear notion of phones (e.g. whole-word models). The phone accuracy measure is replaced by a sum over frames of a negated KL divergence between states, which should be more negative for different phones.

Minimum Phone Frame Error (MPFE) was introduced in [5] as an alternative to MPE where we replace a phone-by-phone accuracy with a frame-by-frame phone accuracy: basically, we include in the accuracy a 1 for each frame if the phone is correct and 0 if the phone is incorrect. We also report on an improved variant we call MPFE-nosil which is the same as MPFE except that for phone arcs q that are silences, we always assign a $\mathrm{PhoneAcc}(q)$ of zero as for MPE [1]

Minimum Bayes Risk is a general framework, revisited recently in [6], where the distance between the proposed and reference sequence can be computed in various ways (from word down to Gaussian level) but in [6] always on a frame-by-frame basis as for MPFE. One of the many variants described is the physical-state level MBR (we use s-MBR), in which the accuracy includes for each frame, a 1 if the physical state is correct (same as reference) and 0 if it is incorrect. The term "physical state" refers to the actual clustered state in a clustered context-dependent system rather than the notional state that depends on the triphone context. Note that in this framework MPFE would be something like p-MBR but we retain the earlier term.

Table 1 compares the definition of $\mathrm{PhoneAcc}(q)$ in MPE and the proposed alternative objective functions. The equations presented here are the versions implemented, which sometimes differ slightly from the original definitions as described below.

Reference lattice alternatives

One set of differences in our implementation concerns the fact that our reference transcription is a lattice covering alternate pronunciations, so the reference state is ambiguous. In all cases we choose the reference state, or phone, most favourable to the current hypothesis, which is different from the original MPFE in which a weighted average over reference states was used (a posterior, in their terminology [5]), and from MD and s-MBR where it seems to have been assumed that the reference state is unique. This approach to handling reference alternatives (take the best; do not average weighted by posterior) was found in previous unpublished experiments to be preferable for normal MPE training. All of our implementations assume a fixed state alignment within both recognition-lattice and reference-lattice arcs.

[1]The same treatment of silence has independently been adopted by the originator of MPFE, Jing Zheng, giving a 0.2% absolute improvement over MPFE (personal communication).

K-L divergence computation for MD

Another difference is the computation of the K-L divergence in MD. As shown in Table 1 we do two versions of MD. We compute a normal MD with state-level divergence as in [4], but with a simpler approximation: we compute a single merged Gaussian for each state and compute a simplified K-L divergence which is just a sum over all dimensions, of the squared difference in means times the average of the two Gaussians' inverse variance. We cache these merged Gaussians per state for speed. We also implement a Gaussian-level MD (g-MD) in which we compute the top Gaussian for each state on the current time and compute a simplified K-L divergence between them as above. In this case the distance $D(\cdot)$ is a function of $\mathbf{x}(t)$ because the top Gaussian is a function of the feature vector.

Divergence in MD not being optimized

In Minimum Divergence training, the objective function is affected by the Gaussian parameters in two ways: firstly, because the posterior of different paths in the denominator lattice changes depending on the parameters; and secondly, because the divergence measure itself can change. In the current work, and in [4], we optimize as if the second factor did not exist. This means that the objective function is not guaranteed to increase. Indeed, it would be inappropriate to corrrectly optimize this objective function, as it has its maximum possible value at zero when all the means and variances are the same.

4. I-SMOOTHING VALUES WITH DIFFERENT OBJECTIVE FUNCTIONS

The objective functions described here all have a much larger dynamic range than the MPE objective function, because they use per-frame measures rather than per-phone. The current writer's belief is that the statistics (as used in the Extended Baum-Welch equations) obtained from these objective functions behave for the most part like scaled-up MPE statistics, and therefore it is appropriate to scale up the smoothing constant τ by the same factor. Supporting this notion of the similarity of the statistics to scaled MPE statistics, Table 2 shows certain functions of the update statistics accumulated for a particular block of data (1/40 of the English broadcast news data), gathered for the various objective functions. The first column shows the sum of the numerator (or denominator) count $\gamma_{j,m}^{num}$ over all Gaussians (it is the same for both). The second and third column shows the "average number of points" for all the numerator and denominator statistics respectively. What this means for the numerator is, $\sum_{j,m} (\sum_t \gamma_{j,m}^{num}(t))^2 / (\sum_t \gamma_{j,m}^{num}(t)^2)$, which is a useful way of calculating an "equivalent" number of equally weighted data points corresponding to a set of data points that have different weights, i.e., equivalent in terms of the variance of a mean computed with these weights. While the counts vary by as much as a factor of 200 (be-

Objective function	Tot num or den count	Equiv #-points num	den	Objf, iter= 1	4
MPE	3.37e+5	8.68e+5	1.23e+6	0.767	0.842
MD	4.30e+7	8.62e+5	1.20e+6	-2.76	-1.86
g-MD	6.07e+7	8.93e+5	1.25e+6	-3.86	-2.83
MPFE	2.07e+6	8.39e+5	1.10e+6	0.879	0.920
MPFE-nosil	1.83e+6	8.08e+5	1.12e+6	0.763	0.808
s-MBR	3.14e+6	8.76e+5	1.18e+6	0.808	0.878

Table 2. Training statistics for 1/40 of the English BN data

.

tween MPE and MD), the equivalent number of points varies by only about 10% as we would expect if the statistics were similar to scaled-up MPE statistics. For MPE we use $\tau = 50$ and for the other objective functions we scale this by the ratio of counts (first column in Table 2), obtained on the first iteration. For reference, this comes out to around $\tau = 6000$ for MD, $\tau = 7500$ for g-MD and $\tau = 400$ for the other objective functions. In future work we may formulate the smoothing in a different way so that a universal value can be used.

We have not done systematic experiments on the effect of τ, but for g-MD on the English test setup (described below) we used by mistake $\tau \simeq 2000$ instead of the value of around 9000 suggested by the count ratios for this setup. This reduced performance by 0.3%-0.4% on the fourth iteration, on the three test sets used (for the baseline see Tables 6 to 8 which used the correct value), and appears to confirm that g-MD requires a much larger τ than 50. Note that our approach to setting τ may not lead to an exactly equal amount of smoothing for all techniques: we observe that the percentage of Gaussians where the total numerator count (before I-smoothing) is smaller than τ on iteration 1 seems to be larger for MPFE and MPFE-nosil than for the others including MPE: in Arabic around 65% vs. 45% for the others, and in English 70% vs. 50%, suggesting stronger smoothing in MPFE and MPFE-nosil. These techniques also show less objective function improvement than MPE and s-MBR (Table 2, last two columns, divided by #-ref-phones for MPE and #-frames for the others), but neither this nor the better results from MPFE and MPFE-nosil (see Section 7) seem to be related to the stronger smoothing because when we decrease τ for MPFE-nosil in English from 420 down to 270 to bring this 70% down to the typical 50%, the objective function on iteration 4 only changes from 0.808 to 0.810 and the final WER averaged over the three test sets remains the same.

5. PREVIOUS RESULTS

Minimum Discrimination training [4] seems to have given a 0.2% absolute improvement in WER over MPE at around 38% WER, increasing the relative improvement from 5.6% to 6.1%, so a 0.5% relative improvement from MD versus MPE. However, it is not clear if this is a fair comparison since the same value of τ was used (400 in both cases) which according to the current writer's belief is around ten times larger than the typically optimal value of 50 for MPE [2] but probably around one twentieth the optimal value for MD based on the logic above. For unclear reasons the experiments reported in [4] contradict this; on the first iteration at least, the optimal value of τ for MPE is even larger than MD, at around 400.

In [6], what we call s-MBR is one of many approaches tried and appears experimentally to be among the best two (tied with phone-level MBR, which is equivalent to MPFE); however the improvements are very small, 1.7% relative versus ML, and the differences are probably not significant.

MPFE [5] has given reported improvements of somewhat larger magnitude than MD: 0.2% and 0.4% absolute improvement over MPE, at the 16.7% WER level, so 1.7% relative improvement versus MPE. The value of τ is 25 in both cases, and apparently the exact value did not make much difference [2], perhaps because for MPFE it is already so low as to be the same as zero.

6. EXPERIMENTAL CONDITIONS

We report results on two systems: a speaker-adapted Arabic broadcast news system with 750 hours of training data, and a speaker independent English broadcast news system with 450 hours of training data. Discriminative training is done with an acoustic weight of 0.1 and language model weight of 1.0, and $E = 2.0$ for both MMI (for backoff) and MPE. Further details on the baseline systems are given in the rest of this section.

Arabic system

The Arabic broadcast news system was the same as the vowelized component of IBM's 2006 submission for the GALE program, as reported in [7], except that feature space MPE (fMPE) [8] is omitted; instead we do model-space discriminative training only.

The acoustic model is trained on about 900 hours of data, consisting of 85 hours of FBIS+TDT4 data with transcripts provided by BBN, 51 hours of transcribed GALE data (first and second quarter releases), and 1800 hours of unsupervised data used for lightly-supervised training [9] of which only 750 hours are kept after pruning based on average word-posterior scores. The acoustic models have 4000 quinphone context-dependent states, with 400k Gaussians. Language-specific features such as the flat-start approach to vowelization and the use of 2-state HMMs for short vowels are described in [7]. The features are 19-dimensional PLP coefficients spliced across 9 frames and projected with LDA+MLLT. We adapt per speaker with cepstral mean normalization, VTLN and fMLLR, and train with feature-space SAT.

We use a 4-gram Kneser-Ney smoothed language model with 56M n-grams trained with a combination of transcripts of the audio data, the Arabic Gigaword corpus, and Web transcripts of broadcast conversations collected from news sites. We use in-line language-model rescoring of lattices to apply this in testing. The vocabulary contains 636K words with 3.18 pronunciation variants per word. Three test data sets are used: BCAD-05 (broadcast conversations), BNAT-05 (broadcast news), and RT-04 (the 2004 test set), whose lengths after silence removal are about 2:00, 5:10 and 1:00 hours respectively.

English system

The acoustic model for the English system is trained on 450 hours of speech comprising the 1996 and 1997 English Broadcast News Speech collections and the English broadcast audio from TDT-4. Lightly-supervised training [9] was performed on the TDT-4 audio because only closed captions were available. The recognition features are 40-d vectors computed via an LDA+MLLT projection of 9 spliced frames of 19-d PLP features. Utterance-based cepstral mean subtraction is used, but no speaker adaptation. The model has 6000 quinphone context dependent states and 250K Gaussians.

The language model used to build the decoding graph is trained on a 192M word corpus comprising the 1996 and 1997 English Broadcast News Transcripts, the 1996 CSR Hub4 Language Model data, the EARS BN-03 English closed captions, the English portion of TDT-4, and the GALE Y1Q1 and Y1Q2 English closed captions. The final language model is 4-gram, Kneser-Ney smoothed and has 3.2M n-grams. The vocabulary has 77K words with 1.08 variants per word.

[2] Personal communication from Jing Zheng

Objective function	Iteration 0	1	2	3	4
MPE	24.9%	24.6%	24.2%	24.1%	23.7%
MD	24.9%	24.4%	24.1%	24.1%	23.7%
g-MD	24.9%	24.3%	23.8%	24.0%	23.3%
MPFE	24.9%	24.4%	23.9%	23.4%	23.3%
MPFE-nosil	24.9%	24.0%	23.7%	23.4%	23.1%
s-MBR	24.9%	24.3%	23.7%	23.9%	23.0%

Table 3. Arabic WER on BCAD-05 (2:00 hours)

Objective function	Iteration 0	1	2	3	4
MPE	16.3%	15.7%	15.6%	15.6%	15.5%
MD	16.3%	15.7%	15.5%	15.5%	15.4%
g-MD	16.3%	15.6%	15.4%	15.4%	15.4%
MPFE	16.3%	15.6%	15.5%	15.4%	15.3%
MPFE-nosil	16.3%	15.6%	15.3%	15.2%	15.0%
s-MBR	16.3%	15.6%	15.4%	15.3%	15.2%

Table 4. Arabic WER on BNAT-05 (5:10 hours)

Objective function	Iteration 0	1	2	3	4
MPE	14.9%	14.4%	14.0%	14.0%	13.8%
MD	14.9%	14.3%	14.1%	13.9%	13.9%
g-MD	14.9%	14.3%	14.0%	14.1%	13.7%
MPFE	14.9%	14.4%	14.2%	14.1%	13.8%
MPFE-nosil	14.9%	14.2%	14.0%	13.9%	13.6%
s-MBR	14.9%	14.3%	14.2%	14.0%	13.8%

Table 5. Arabic WER on RT-04 (1:00 hours)

We use the test sets rt03, dev04f and rt04 as defined for the English portion of the EARS program, which after silence removal have lengths of 2:15, 2:00 and 4:00 hours respectively.

7. EXPERIMENTAL RESULTS AND CONCLUSIONS

Tables 3 to 5 and 6 to 8 show the results for Arabic and English respectively. We can focus on the last iteration (last column) since this always gives the best result for any given technique and test set. For five out of the six test sets, the best result is with MPFE-nosil, which was always better than MPE. The average improvement is 1.9% relative. In the remaining case, s-MBR is the best with MPFE-nosil close behind. If we consider which variants of techniques are best, MPFE-nosil is always better than MPFE and g-MD is always better than or the same as MD. Compared to MPE, g-MD and s-MBR show inconclusive results. In both cases, they are better than MPE or worse than MPE an equal number of times. Because MPFE-nosil appears to skew the system somewhat towards insertion over deletion errors, as would be expected because it penalizes silence in training, we examined for the English setup whether the improvement over MPE remained after tuning the language model scale in both cases (trying 16 and 20 as well as the default 18). The improvement actually increased slightly, by 0.0%, 0.1% and 0.1% absolute.

In conclusion, from the results presented here it appears that the most promising technique is MPFE-nosil (a frame-level version of MPE, with special treatment of silence, followed by MPFE. g-MD and s-MBR seem to give the same results as MPE and may be useful as equivalently good alternatives to MPE where it is not possible to assign states to phones. Future work is needed to simplify the handling of I-smoothing with MPFE-nosil and to examine the application to feature-space versions of discriminative training.

Objective function	Iteration 0	1	2	3	4
MPE	12.9%	12.4%	12.2%	11.9%	11.6%
MD	12.9%	21.4%	12.1%	12.0%	11.8%
g-MD	12.9%	12.5%	12.1%	12.0%	11.7%
MPFE	12.9%	12.4%	12.1%	11.9%	11.6%
MPFE-nosil	12.9%	12.2%	11.8%	11.5%	11.5%
s-MBR	12.9%	12.4%	12.0%	11.9%	11.7%

Table 6. English WER on rt03 (2:15 hours)

Objective function	Iteration 0	1	2	3	4
MPE	23.5%	22.5%	21.6%	21.3%	21.0%
MD	23.5%	22.2%	21.8%	21.6%	21.6%
g-MD	23.5%	22.2%	21.7%	21.5%	21.5%
MPFE	23.5%	22.2%	21.7%	21.4%	21.0%
MPFE-nosil	23.5%	22.1%	21.3%	21.1%	20.8%
s-MBR	23.5%	22.2%	21.6%	21.4%	21.1%

Table 7. English WER on dev04f (2:00 hours)

Objective function	Iteration 0	1	2	3	4
MPE	20.5%	19.4%	18.8%	18.5%	18.1%
MD	20.5%	19.3%	18.7%	18.5%	18.2%
g-MD	20.5%	19.2%	18.4%	18.4%	18.2%
MPFE	20.5%	19.4%	18.5%	18.3%	18.0%
MPFE-nosil	20.5%	19.1%	18.4%	17.9%	17.7%
s-MBR	20.5%	19.3%	18.5%	18.5%	18.1%

Table 8. English WER on rt04 (1:00 hours)

8. REFERENCES

[1] Povey D. and Woodland P.C., "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *ICASSP*, 2002.

[2] Povey D., *Discriminative Training for Large Voculabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[3] Soltau H., Kingsbury B., Mangu L., Povey D., Saon G., and Zweig G., "The IBM 2004 Conversational Telephony System for Rich Transcription in EARS," in *ICASSP*, 2005.

[4] Du J., Liu P., Soong F.K., Zhou J-L., and Wang R-H., "Minimum Divergence Based Discriminative Training," in *Interspeech*, 2006.

[5] Zheng J. and Stolcke A., "Improved Discriminative Training Using Phone Lattices," in *Interspeech*, 2005.

[6] Gibson M. and Hain T., "Hypothesis Spaces For Minimum Bayes Risk Training In Large Vocabulary Speech Recognition," in *Interspeech*, 2006.

[7] Soltau H., Saon G., Povey D., Mangu L., Kingsbury B., Omar M., and Zweig G., "The IBM 2006 GALE Arabic System," submitted to: *ICASSP*, 2007.

[8] Povey D., Kingsbury B., Mangu L., Saon G., Soltau H., and Zweig G., "fMPE: Discriminatively trained features for speech recognition," in *ICASSP*, 2005.

[9] H. Y. Chan and P. C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP*, 2004.