DISCRIMINATIVE TRAINING FOR DIRECT MINIMIZATION OF DELETION, INSERTION AND SUBSTITUTION ERRORS

Sungwhan Shin†, Ho-Young Jung‡ and Biing-Hwang (Fred) Juang§

†Center for Signal and Image Processing
Georgia Institute of Technology
{shshin, juang}@ece.gatech.edu

‡Speech Language Processing Team
Electronics and Telecommunications Research Institute
hjung@etri.re.kr

ABSTRACT

In this paper, we follow the minimum error principle for acoustic modeling and formulate error objectives in insertion, deletion, and substitution separately for minimization during training. This new training paradigm generalized from the MVE criterion can explain the direct relationship between recognition errors and detection errors by re-interpreting deletion, insertion, and substitution errors as miss, false alarm, and miss/false-alarm errors happening together. Under the MVE criterion, by applying two mis-verification measures for miss and false alarm errors selectively along with the types of recognition error definition, we developed three individual objective training criteria, minimum deletion error (MDE), minimum insertion error (MIE), and minimum substitution error (MSE), of which each objective function can directly minimize each of the three types of recognition errors. In the TIMIT phone recognition task, the experimental results confirm that each objective criterion of MDE, MIE, and MSE results in primarily minimizing its target error type, respectively. Furthermore, a simple combination of the individual objective criteria outperforms the conventional string-based MCE in the overall recognition error rate.

Index Terms— discriminative training, minimum verification error, continuous speech recognition

1. INTRODUCTION

In continuous speech recognition, the recognition errors can be classified into three types after alignment between the transcription and the recognized string by a dynamic programming (DP) procedure. They are deletion, insertion, and substitution errors. In various automatic speech recognition (ASR) applications, a level of significance for each of the errors is oftentimes scaled according to the task-specific direction and target. For example, a deletion error by the ASR system may be regarded as more serious than a substitution error in an automatic dialog-enabled language learning system because currently there is no evaluation guidelines for deletion errors and the system does not know how to respond to such errors. Thus, it is desirable to formulate a training algorithm which can directly minimize each of these three types of errors.

Several Discriminative Training (DT) methods, such as maximum mutual information estimation (MMIE) [1], minimum classification error (MCE) [2], and minimum phone/word error (MPE/MWE) [3], have achieved success in various speech recognition tasks over years. Among them, the MCE and the MPE/MWE aim at direct minimization of the substitution error on the chosen unit class, say a word, either on the same level as the unit, or at a level above (e.g., a string of words) or below (e.g., a string of phonemes) “word”. It is considered very hard to present a natural solution for directly minimizing the deletion and insertion errors. However, if we re-interpret the three types of recognition error in the context of a detection problem, the deletion, insertion, and substitution errors can be respectively explained as miss, false alarm, and miss/false-alarm errors happening together. Then, each of the errors can be minimized under the framework of the detection theory. Table 1 compares the difference between the two problem descriptions. First, in error type, the recognition problem is associated with only one misclassification error, while the detection problem is associated with both Type I error (miss) and Type II error (false alarm). Second, in alignment error, the recognition problem produces deletion, insertion and substitution errors and each of the errors in the recognition problem can be viewed as miss, false alarm, and both in the detection problem. Last, in the training criterion, normally during recognition, only the substitution error can be minimized to solve the recognition problem, whereas the training criterion for the detection problem minimizes a combination or the total of the detection errors associated with miss and false-alarm, respectively. As a result, we may re-think the recognition problem as a detection problem.

Table 1. Comparison between different problem descriptions (adopted from [4])

<table>
<thead>
<tr>
<th>Recognition Problem</th>
<th>Error Type</th>
<th>Alignment Errors</th>
<th>Training Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Misclassification</td>
<td>Deletion</td>
<td>Minimize sub-errors only</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>Insertion</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Substitution</td>
<td></td>
</tr>
<tr>
<td>Detection Problem</td>
<td>Type I/II</td>
<td>Type I</td>
<td>Minimize Type I &amp; II both</td>
</tr>
<tr>
<td></td>
<td>(Miss/FA) Errors</td>
<td>Type II</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this paper, based on the above analysis, we propose a new multi-objective discriminative training method using the minimum verification error criterion (MVE) [4,5,6] to directly not only deal with each type of the recognition errors in the detection viewpoint, but also minimize each of the errors and the overall recognition error rate as well. Under the MVE criterion which is a special version of MCE for detection and verification problems, by applying two mis-verification measures for miss and false alarm errors selectively along with the types of recognition error definition, we developed a multi-objective learning framework, of which the objective functions are designed to directly minimize each of the three types of recognition errors and all together. In contrast to a string-level MCE [2], the proposed learning framework is performed only on error segments between the
transcription and the recognized string after DP matching. This learning framework provides a more direct measure of each type of the three errors and significantly reduces the computational complexity compared to the string-level MCE. Hence, we name each objective criterion minimum deletion error (MDE), minimum insertion error (MIE), and minimum substitution error (MSE), respectively.

The rest of this paper is organized as follows: In Section 2, we will describe the details of the multi-objective learning framework in the detection viewpoint, followed by a derivation of the proposed method under the MVE criterion. Experimental setup and results are then presented in Section 3. Finally, conclusions are drawn and the planned future works are discussed in Section 4.

2. MULTI-OBJECTIVE LEARNING FRAMEWORK

USING MINIMUM VERIFICATION ERROR

2.1. Recognition error from a detection viewpoint

Although the conventional MCE has shown how to minimize the total empirical errors on the training data, the MCE objective function was designed to mainly reduce the empirical substitution errors on the training data. For every training utterance $X_k$, a string-level misclassification measure [2] in the MCE criterion, $d(X_k | \Lambda)$, compares two discriminant functions, $g(X_k, S_r | \Lambda)$ for the known reference string $S_r$ and $G(X_k, S_n | \Lambda)$ for the competing $N$-best strings $S_n$, which can be formulated as:

$$d(X_k | \Lambda) = -g(X_k, S_r | \Lambda) + G(X_k, S_n | \Lambda)$$

(1)

where $\Lambda$ is the HMM parameter set and $G(X_k, S_n | \Lambda)$ is a weighted sum over the competing $N$-best strings. Given the misclassification measure, only the local accumulation of the string-level errors can be minimized. However, as argued in the introduction, it is not appropriate to ignore a direct measure of deletion and insertion errors in the discriminative training.

As an alternative, the enhanced minimum classification error (E-MCE) was proposed in [7]. By generating three sets of competing strings from constrained $N$-best search, MDI, MIE, and MSE were constructed following the conventional MCE training. However, E-MCE is not a direct individual error minimization method, but a balanced method for the three types of recognition errors. Furthermore, since it explicitly follows the conventional string based MCE framework based on the misclassification measure in Eq. (1), the objective function of the E-MCE still focuses on minimizing the empirical average loss of the three errors in the given competing string.

In order to construct direct objective functions for deletion and insertion errors, we propose a new training framework presented in Fig. 1. Suppose the reference string is $W_r^T$ and the one best decoded string from ASR is $W_d^T$. After a DP-based string alignment procedure, one deletion error $W_d^T$ and one insertion error $W_d^d$ are counted as shown in Fig. 1. If we interpret the two recognition errors from a detection viewpoint, the deletion error $W_d^T$ can be regarded as a miss error in the detection problem since $W_d^T$ has to exist on the decoded string but it is missed with respect to the decoded output sequence. On the other hand, $W_d^d$ has to be rejected but it is inserted on the decoded output sequence. Thus, the insertion error $W_d^d$ can be viewed as a false alarm error in the detection problem. Then, from the MVE criterion, the segments of the deletion error $W_d^T$ and the insertion error $W_d^d$ are trained by the first mis-verification measure $d_f(X_k, W_r^T | \Lambda)$ and the second mis-verification measure $d_{II}(X_k, W_d^d | \Lambda)$, respectively, as follows:

$$d_f(X_k, W_r^T | \Lambda) = -g_f(X_k, W_r^T | \Lambda_f) + g_f(X_k, W_d^d | \Lambda_a)$$

(2)

$$d_{II}(X_k, W_d^d | \Lambda) = +g_f(X_k, W_d^d | \Lambda_f) - g_f(X_k, W_d^d | \Lambda_a)$$

(3)

where $d_f$ and $d_{II}$ are the type I and type II mis-verification measures [4,10,11], respectively. In Eq. (2) and (3), $g_f$ and $g_a$ are the normalized log likelihood and $\Lambda_f$ and $\Lambda_a$ are the parameter sets of the target model and the anti-model [4,8,10,11,16] for the given segment, respectively.

This new training paradigm generalized from the MVE criterion can explain the direct relationship between the recognition errors and the detection errors. Nevertheless, it is intuitively obvious that counting only error segments, $W_r^T$ and $W_d^d$, may not reflect the effective model separation and error minimization in the discriminative training phase since the deletion and insertion errors are directly related to the preceding and succeeding segments. In addition, there is a prominent need in identifying the part of speech data that contain potential deletion and insertion errors for the purpose of discriminative parameter optimization. Therefore, we propose a new training framework covering the segments right before and right after the error segment as shown in Fig. 1. One can further extend this framework by associating the preceding and succeeding segments with non-uniform error cost like [9] or by containing more connected segments with $d_f$ and $d_{II}$ than we propose.

2.2. Derivation of multi-objective discriminative training using the MVE criterion

Segment-based MVE has shown its effectiveness in constructing detectors [10,11] and rescoring hypotheses [12] from an ASR system for improved continuous speech recognition. In this section, we will derive the multi-objective discriminative training extended from the segment-based MVE criterion.

Suppose there are $M$ classes and $K$ training samples in a given training data set. After DP matching, the given $K$ training samples are assigned into $\{X_1^d, X_2^d, ..., X_K^d\}$ for the reference transcript and $\{X_1^d, X_2^d, ..., X_K^d\}$ for the decoded output. From the samples and error assignments of the decoded output, the empirical average loss is defined by

![Fig. 1. Error count and corresponding mis-verification measures ($d_f$ and $d_{II}$) under the MVE criterion](image-url)
where $l_{\text{total}}(X_k^d | A)$ is the composite loss function which combines four different types of the recognition outputs from the general DP-based string error assignment. For the multi-objective discriminative learning, the composite loss function can be described as

$$l_{\text{total}}(X_k^d | A) = l_{\text{del}}(X_k^d | A)1(X_k^d \in "Del") + l_{\text{ins}}(X_k^d | A)1(X_k^d \in "Ins") + l_{\text{sub}}(X_k^d | A)1(X_k^d \in "Sub") + l_{\text{hit}}(X_k^d | A)1(X_k^d \in "Hit")$$

(5)

where $l_{\text{del}}(\cdot)$, $l_{\text{ins}}(\cdot)$, and $l_{\text{sub}}(\cdot)$ denote respectively individual objective functions: MDE, MIE, and MSE. First, the objective function for MDE can be written as

$$l_{\text{del}}(X_k^d | A) = PW_{\text{I}} \sum_{i=-1,1}^M \sum_{j=1}^M l\left(d_i(X_k^d | A^j)\right)1(X_k^d \in C_i)$$

$$+ PW_{\text{II}} \sum_{j=-1,1}^M \sum_{i=1}^M l\left(d_j(X_k^d | A^i)\right)1(X_k^d \in C_j)$$

(6)

where $PW_{\text{I}}$ and $PW_{\text{II}}$ are the penalty weights for type I and type II errors, respectively and $l(\cdot)$ is a smoothed loss function normally defined by a sigmoid function [2]. Note that the two kinds of mis-classification measures are separately assigned to the reference segment $X_k^d$ and decoded segment $X_k^d$ as defined by

$$d_i\left(X_k^d | A^i\right) = -g_t\left(X_k^d | A^i\right) + g_a\left(X_k^d | A^i\right)$$

$$d_j\left(X_k^d | A^j\right) = +g_a\left(X_k^d | A^j\right) - g_t\left(X_k^d | A^j\right).$$

(7)

(8)

Unlike Eq. (1), in Eq. (7) and (8), $g_t$ and $g_a$ are the segment-based normalized log likelihood and $A^i_t$ and $A^j_a$ are the parameter set of the target and the anti model for the $i$th class, respectively. Similar to MDE, the objective function of MIE can be written as

$$l_{\text{ins}}(X_k^d | A) = PW_{\text{I}} \sum_{j=-1,1}^M \sum_{i=1}^M l\left(d_i(X_k^d | A^j)\right)1(X_k^d \in C_j)$$

$$+ PW_{\text{II}} \sum_{i=1}^M \sum_{j=-1,1}^M l\left(d_j(X_k^d | A^i)\right)1(X_k^d \in C_i).$$

(9)

For MSE, as discussed in the previous sections, the substitution error can be regarded as miss and false alarm errors happening together at the given segments. As is done above, the objective function of MSE can be formulated as

$$l_{\text{sub}}(X_k^d | A) = PW_{\text{I}} \sum_{i=1}^M \sum_{j=-1,1}^M l\left(d_i(X_k^d | A^j)\right)1(X_k^d \in C_j)$$

$$+ PW_{\text{II}} \sum_{j=1}^M \sum_{i=-1,1}^M l\left(d_j(X_k^d | A^i)\right)1(X_k^d \in C_i).$$

(10)

Finally, the minimization of each objective function can be accomplished through the generalized probabilistic descent (GPD) method [2,4,5] w.r.t. all parameters. In the following experiments, we used uniform penalty weights for both $PW_{\text{I}}$ and $PW_{\text{II}}$. Moreover, the experiments in this paper are conducted on each objective criterion and then a simple combination of the multi-objective criteria. One can investigate the non-uniform penalty weights and rule-based combinations of the multi-objective criteria with particular constraints such as [13] over the proposed learning framework.

3. EXPERIMENTS AND RESULTS

The experiments reported in this section were carried out on the TIMIT database and we used the standard experimental setup as specified in [14]. In this paper, we focus on phone recognition experiments. As a baseline, we trained both context-independent (CI) and context-dependent (CD) HMM phone recognizers using the latest version of the HTK toolkit (http://htk.eng.cam.ac.uk/). The CI system consists of 48 monophones defined in [14] and all the phones except for the short pause “sp” are modeled by 3-state left-to-right HMMs with 70 Gaussians per state. The short pause model “sp” has only one state. On the other hand, the CD system contains a total of 3420 physical triphone models with 986 tied-states and each state is modeled using a 16-component Gaussian mixture. Note that for the proposed discriminative training phase in both CI and CD systems, corresponding anti models are limited to 48 CI monophones.

In all experiments, we represented the speech using 39 dimensional feature vectors with 12MFCC, 12Δ, 12ΔΔ and 3 log energy values. The standard 3696 training utterances excluding the “sa” utterances and 192 core-test utterances were used. In the phonetic recognizer’s evaluation, we used a bigram language model over phones estimated from the training set and merged the 48 monophones into 39 monophones according to the standard mapping described in [14] and the confusion among the merged phones is not considered as errors. In addition, the number of training iterations for all the MCE and the proposed method in Table 2 and 3 is fixed to be five.

### 3.1. CI Results

Table 2 shows the performance comparison between the conventional string-based MCE and the proposed multi-objective discriminative training method. In particular, we present the detailed performance of each objective criterion and two kinds of simple combinations of the individual objective criteria such as “D+S+I” and “H+D+S+I”. Note that the combined multi-objective training methods mean the three error segments and “hit” segments are simply incorporated in the discriminative training phase.

It can be seen from Table 2, MCE mainly reduces the substitution error as intended. However, each objective criterion of MDE, MIE, and MSE resulted in primarily reducing its target error type, respectively. Furthermore, although we constructed the simple combinations of the individual objective criteria, the two combined multi-objective training methods still confirm the effectiveness of the proposed learning framework. A rule-based optimization method such as [13], unlike the simple combinations reported here, may bring about more overall error reduction.
Table 2. Detailed performance comparison on CI system

<table>
<thead>
<tr>
<th></th>
<th>Del</th>
<th>Ins</th>
<th>Sub</th>
<th>Acc %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>678</td>
<td>170</td>
<td>1289</td>
<td>70.57 %</td>
</tr>
<tr>
<td>MCE</td>
<td>674</td>
<td>179</td>
<td>1265</td>
<td>70.83 %</td>
</tr>
<tr>
<td>MDE</td>
<td>655</td>
<td>175</td>
<td>1279</td>
<td>70.95 %</td>
</tr>
<tr>
<td>MIE</td>
<td>687</td>
<td>156</td>
<td>1278</td>
<td>70.79 %</td>
</tr>
<tr>
<td>MSE</td>
<td>691</td>
<td>159</td>
<td>1273</td>
<td>70.76 %</td>
</tr>
<tr>
<td>D+I+S</td>
<td>687</td>
<td>159</td>
<td>1272</td>
<td>70.83 %</td>
</tr>
<tr>
<td>H+D+I+S</td>
<td>521</td>
<td>274</td>
<td>1278</td>
<td>71.45 %</td>
</tr>
</tbody>
</table>

3.2. CD Results

Table 3 shows the performance comparison in the CD system. In particular, we ignored the training criterion with the “hit” tokens since in contrast to the CI system the CD system consists of a huge number of classes. Thus, we only show the detailed performance of MDE, MIE, MSE and “D+I-S” as presented in Table 3.

Table 3. Detailed performance comparison on CD system

<table>
<thead>
<tr>
<th></th>
<th>Del</th>
<th>Ins</th>
<th>Sub</th>
<th>Acc %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>444</td>
<td>246</td>
<td>1271</td>
<td>73.26 %</td>
</tr>
<tr>
<td>MCE</td>
<td>448</td>
<td>250</td>
<td>1265</td>
<td>73.23 %</td>
</tr>
<tr>
<td>MDE</td>
<td>435</td>
<td>266</td>
<td>1254</td>
<td>73.34 %</td>
</tr>
<tr>
<td>MIE</td>
<td>468</td>
<td>222</td>
<td>1257</td>
<td>73.45 %</td>
</tr>
<tr>
<td>MSE</td>
<td>453</td>
<td>243</td>
<td>1245</td>
<td>73.53 %</td>
</tr>
<tr>
<td>D+I+S</td>
<td>449</td>
<td>243</td>
<td>1255</td>
<td>73.45 %</td>
</tr>
</tbody>
</table>

Similar to the results in [15], we could not obtain performance gain from the conventional string-based MCE in the CD system. However, the proposed methods reconfirm that each objective criterion reduces its target error type and the combined multi-objective training method also reduces the overall recognition error. However, compared to the ML baseline, MDE and MIE yield more insertion and deletion errors, respectively. One possible cause of the instability is the lack of modeling the anti models with a corresponding discriminability. As mentioned, the anti models for the limited 48 CI monophones were employed during evaluation with the CD target models in the discriminative training phase. It is likely that use of the CD anti-subword models discriminatively trained with the corresponding CD target models would lead to improved performance as shown in [16].

4. CONCLUSION

In this paper, we interpret the commonly known three recognition error types, insertion, deletion and substitution, from an event detection viewpoint and introduce a new training paradigm aiming at direct reduction of these individual errors. By considering the deletion, insertion, and substitution errors as miss, false alarm, and simultaneous miss/false-alarm, the MV(efication)E criterion is generalized to MD(eletion)E, MI(insertion)E, and MS(substitution)E, as the objective functions for direct minimization of each of the three types of errors. Furthermore, we studied combined multi-objective training criteria by incorporating the three individual objectives and the “hit” segments. We carried out experiments in phone recognition on the TIMIT corpus. Results for both CI and CD systems confirm that each objective criterion of MDE, MIE, and MSE results in minimization of its target error, respectively. In addition, the combined multi-objective training methods outperform the conventional string-based MCE.

Our future work will involve developing the CD anti-subword modeling; we anticipate well-constructed CD anti-subword models will lead to further performance improvement as shown in [16]. Another possible improvement can come from associating the preceding and succeeding segments for the error-counted segment with non-uniform error cost like [9]. Finally, we plan to apply task-dependent non-uniform penalty weights for both PWl and PWl and further investigate certain rule-based optimization methods such as [13] for the improved multi-objective criteria.

5. ACKNOWLEDGEMENTS

This work was supported by the Industrial Strategic technology development program, 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded by the Ministry of Knowledge Economy (MKE, Korea).

6. REFERENCES