Individual Error Minimization Learning Framework and its Applications to Speech Recognition and Utterance Verification

Sunghwan Shin[1], *Ho-Young Jung*[2], *Biing-Hwang (Fred) Juang*[1]

[1]Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA, USA
[2]Speech Language Processing Team, Electronics and Telecommunications Research Institute, Daejeon, South Korea

{shshin,juang}@ece.gatech.edu, hjung@etri.re.kr

10.21437/Interspeech.2011-189

Abstract

In this paper, we extend the individual recognition error minimization criteria, MDE/MIE/MSE [1] in word-level and apply them to word recognition and verification tasks, respectively. In order to effectively reduce potential errors in word-level, we expand the training token selection scheme to be more appropriate for word-level learning framework, by taking into account neighboring words and by covering internal phonemes in each training word. Then, we examine the proposed word-level learning criteria on the TIMIT word recognition task and further investigate individual rejection performance of the recognition errors in utterance verification (UV). Experimental results confirm that each of the word-level objective criteria results in primarily reducing the corresponding target error type, respectively. The rejection rates of insertion and substitution errors are also improved within MIE and MSE criteria, which lead to additional word error rate reduction after the rejection.

Index Terms: discriminative training, individual error minimization, utterance verification

1. Introduction

The technology of automatic speech recognition (ASR) by machine has advanced substantially in the last two decades, thanks to the mathematical formalization of the statistical modeling approach that forms the foundation of the ASR system design methodology. One critical component of the methodology is the link between the system parameters and a prescribed system performance objective, which in its most essential and simplistic form would be the error probability resulted from a recognition test.

In more sophisticated tasks such as transcribing a continuous speech signal into a sequence of linguistic codes (phonemes, words, etc.), the recognition errors can be classified into three types after alignment between the true (or manual) transcription result and the recognized code string by a dynamic programming (DP) procedure. They are deletion, insertion, and substitution errors. In most ASR applications, they are minimized regardless of the subjective significance of individual errors and system performance is measured by the total error rate computed as the sum of the errors. However, a level of significance for each error is oftentimes scaled according to the task-specific direction and target. In other words, depending on the application of the system, these three types of errors may have a varying degree of significance. 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 [2, 3] 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/modeling algorithm which can directly minimize each of these three types of errors.

In [1], we proposed a discriminative training formulation for direct minimization of the deletion, insertion, and substitution errors. We re-interpreted the commonly known three recognition error types from an event detection viewpoint and introduced 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(erification)E criterion [4, 5] is generalized to MD(eletion)E, MI(nsertion)E, and MS(ubstitution)E, as the objective functions for direct minimization of each of the three types of errors. This viewpoint is one of the essences of the detection-based ASR [6, 7]. In the TIMIT continuous phone recognition task we showed results of each objective criterion of MDE, MIE, and MSE, as the primary target error to minimize, respectively.

In this paper, we extend the individual error minimization training algorithm, MDE/MIE/MSE, to direct minimization of the individual error types in continuous word recognition. In order to appropriately apply the previous learning framework to the word-level training, an extended word-level training event selection scheme and the accompanying learning framework are proposed since in the word recognition task the training and recognition decisions may not be on the same linguistic level in the performance measure. For example, the training and recognition may be performed on the phone level, but the system evaluation measure is defined on the word level as word error rate (WER). Hence, we propose a word-level learning framework to cover internal phonemes in each aligned word after word-level DP-matching is performed and an extended training event selection scheme that includes the words just before and just after the error word to effectively reduce potential deletion and insertion errors as shown in Figure 1.

Furthermore, although the proposed learning method yields discriminatively trained anti-models as well as target models at the same time, we reported only the recognition performance by the target model in [1]. As the discriminatively trained target model can be directly used for the recognition task, the simultaneously trained anti-model with the target model as a set of detectors can also be used for detection and verification tasks. Therefore, we further investigate individual rejection performance of the recognition errors in utterance verification (UV) [8, 9] by using the target and anti models under the framework of likelihood ratio testing. In contrast to the conventional two-stage UV, in this paper UV is considered as one integrated

Copyright © 2011 ISCA

stage in which the recognition and verification share the same set of discriminatively trained models as proposed in [10]. In this paper, the verification experiments are performed between correctly recognized words versus misrecognized words (only including insertion and substitution errors).

The remainder of this paper is organized as follows: In Section 2, we describe details of the proposed word-level individual error minimization learning framework and its applications, followed by a derivation of the proposed method. Experimental setup and results are then presented in Section 3. Finally, conclusions are drawn and the planned future work is discussed in Section 4.

2. Individual Error Minimization Learning Framework and its Applications

2.1. Word-level Event Selection for MDE and MIE

Figure 1 illustrates the principle of our training event selection scheme for word-level MDE and MIE. Suppose the reference string is W^r and the one best decoded string from ASR is W^d. If all words except for w_2^r and w_4^d are exactly matched to each other, after a DP-based string alignment procedure, one deletion error w_2^r and one insertion error w_4^d are counted as shown in Fig. 1. If we interpret the two recognition errors from a detection viewpoint, the deletion error w_2^r can be regarded a miss (Type I) error in the detection problem since w_2^r has to exist between w_1^d and w_2^d on the decoded string but it is missed with respect to the decoded output sequence. On the other hand, w_4^d has to be rejected but it is inserted on the decoded output sequence. Thus, the insertion error w_4^d can be viewed as a false alarm (Type II) error in the detection problem. Then, from the MVE criterion, the segments of the deletion error w_2^r and the insertion error w_4^d are trained by the two different types of mis-verification measures, d_I for miss error and d_{II} for false alarm error, respectively.

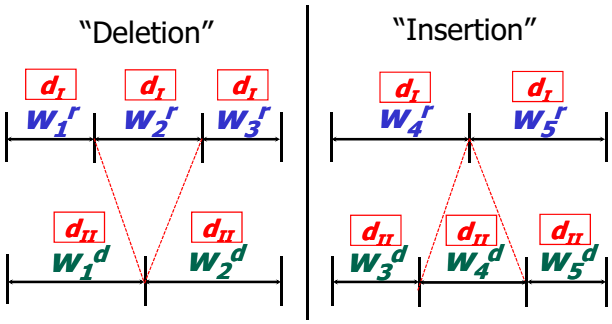


Figure 1: Word-level training event selection scheme and corresponding mis-verification measures (d_I and d_{II})

However, as discussed in [1], it is intuitively obvious that selecting only error words, w_2^r and w_4^d, as training events 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 words. Therefore, we propose an extended word-level training token selection scheme, as shown in Figure 1, including neighboring words of the error word on both reference and decoded strings to effectively reduce potential deletion and insertion errors in word-level. In addition, each word of w_k^r and w_k^d contains a phoneme sequence as $ph_k^1, ph_k^2, \cdots, ph_k^{N_k}$. The internal phonemes in each word are also trained by either d_I or d_{II} according to the corresponding word-level error assignment.

2.2. Application to Utterance Verification

Since the proposed individual error minimization learning criteria are essentially generalized from the MVE criterion, we may expect the proposed method has intrinsic nature of the MVE criterion and thus a viable application using the proposed method may not be limited to the recognition, but extended to detection and verification. The MVE criterion has been extensively applied to event verification applications such as speaker verification [11] and utterance verification (UV). In this paper, we will investigate whether the proposed method can be directly applied to UV task without any alteration and additional training. In particular, as an extended application of the individual error minimization framework, we will focus on individual rejection performance of the recognition errors (here only considering insertion and substitution errors) in UV using the target and anti models discriminatively trained under the proposed learning framework.

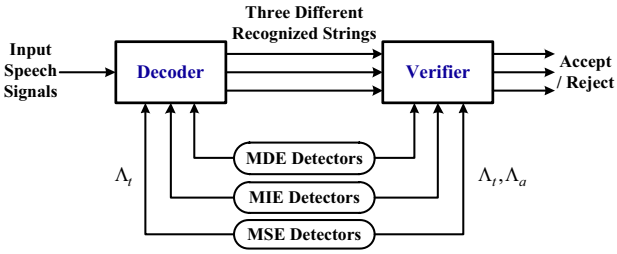


Figure 2: A schematic of an integrated utterance verification system sharing MDE/MIE/MSE detectors

Utterance verification is considered as a hypothesis testing problem and an important step in automatic speech recognition applications; it is used to reject improper, illegitimate, or potentially misrecognized input utterances. UV is traditionally considered an add-on procedure to ASR and thus treated separately from the ASR system model design. In contrast to this traditional two-stage UV approach, in this paper UV is constructed as one integrated stage in which the recognition stage and verification stage share the same set of discriminatively trained models as proposed in [10]. Using this integrated UV framework, the recognition and verification are effectively performed in a way consistent with the verification models and the hypothesis testing.

Figure 2 presents a schematic of the integrated utterance verification system under the proposed learning framework. First, in the recognition stage the decoder produces three different sets of recognized strings by using the target models of each MDE/MIE/MSE detectors. Then, for the hypothesis testing of the given recognized words on the three different sets of the recognized strings, the target and anti models are used to find subword-level acoustic verification scores based on the following log likelihood ratio LLR_p for the subword p

$$LLR_p = \log P(X|\Lambda_t^p) - \log P(X|\Lambda_a^p) \quad (1)$$

where X is the subword segment of the word w and Λ_t^p and Λ_a^p are the corresponding target subword and anti-subword models for subword p, respectively. The word-level confidence score CM_w is then defined by

$$CM_w = \frac{1}{N_w} \sum_n LLR_n \quad (2)$$

where N_w is the total number of subwords in the word w. The confidence measure score CM_w for each hypothesized word w is compared to a pre-specified operating threshold. Based on

the threshold, the final decision for the hypothesized word w is made as either acceptance or rejection.

2.3. Derivation of Word-level Individual Error Minimization Learning Algorithm

The segment-based MVE has shown its effectiveness in constructing detectors [5, 12] and rescoring hypotheses [13] from an ASR system for improved continuous speech recognition. In this section, we will derive the word-level individual error minimization learning criteria, MDE/MIE/MSE, extended from the segment-based MVE criterion.

Let X_k, k=1,$\cdots$,K, be the utterances in the training set, W^r be the labeled reference word sequence and W^d be the one best decoded word sequence from ASR for X_k, respectively. After DP matching between the word sequences W^r and W^d, each word in the sequences is represented by its DP-assignment result, word identity and word boundary starting from time t^s and ending at time t^e. In addition, positions of insertion and deletion errors are added and marked into reference and decoded strings but without no word name and time alignment, i.e., $W^{d'} = \{w_1^{d'}(hit, w_1^d, t_1^s, t_1^e), w_2^{d'}(del, \phi, \phi, \phi), w_3^{d'}(hit, w_2^d, t_2^s, t_2^e), \cdots\}$ about the deletion error case in Figure 1. That is to say, they have the same total number of words, L_k, from the DP matching and assignment. And all words, $w_{l_k}^r$ and $w_{l_k}^d$, contain a phoneme sequence as ph_k^1, ph_k^2, $\cdots$, $ph_k^{N_k}$. Therefore, $X_k = \{X_{k,l_k,n_k}\}_{n_k}^{N_k}$ is k^{th} training utterance that is segmented into N_k segments corresponding to the phoneme sequence. Then, from the word-level segments and error assignments, the empirical average loss is given by

$$L(\tilde{\Lambda}) = \sum_{k=1}^{K} \sum_{l_k=1}^{L_k} \ell_{total}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda) \tag{3}$$

where $\ell_{total}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda)$ is the composite loss function which combines three different types of the word-level recognition errors. For the multi-objective discriminative learning containing the three types of the direct minimization criteria, the composite loss function can be described as

$$\begin{aligned}
&\ell_{total}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda) \\
&= \ell_{Del}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda)1(w_{l_k}^d \in \text{“}Del\text{”}) \\
&\quad + \ell_{Ins}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda)1(w_{l_k}^d \in \text{“}Ins\text{”}) \\
&\quad + \ell_{Sub}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda)1(w_{l_k}^d \in \text{“}Sub\text{”})
\end{aligned} \tag{4}$$

where $\ell_{Del}(\cdot)$, $\ell_{Ins}(\cdot)$ and $\ell_{Sub}(\cdot)$ denote respectively the word-level individual objective functions: MDE, MIE, and MSE. First, the objective function for MDE can be written as

$$\begin{aligned}
&\ell_{Del}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda) \\
&= PW_I \sum_{i=-1}^{1} \sum_{n_k=1}^{N_k} \ell\left(d_I(X_{k,l_k+i,n_k}|ph_{l_k+i}^{n_k}, w_{l_k+i}^r, \Lambda)\right) \\
&+ PW_{II} \sum_{i=-1,1} \sum_{n_k=1}^{N_k} \ell\left(d_{II}(X_{k,l_k+i,n_k}|ph_{l_k+i}^{n_k}, w_{l_k+i}^d, \Lambda)\right)
\end{aligned} \tag{5}$$

where PW_I and PW_{II} are the penalty weights for type I and type II errors, respectively and $\ell(\cdot)$ is a smoothed loss function normally defined as a sigmoid function. We note that the two kinds of mis-verification measures are separately assigned to the reference segment and decoded segment as defined by

$$\begin{aligned}
d_I(X_{k,l_k,n_k}|\Lambda^p) &= -g_t(X_{k,l_k,n_k}|\Lambda_t^p) + g_a(X_{k,l_k,n_k}|\Lambda_a^p) \\
d_{II}(X_{k,l_k,n_k}|\Lambda^p) &= +g_t(X_{k,l_k,n_k}|\Lambda_t^p) - g_a(X_{k,l_k,n_k}|\Lambda_a^p)
\end{aligned} \tag{6}$$

where g_t and g_a are the segment-based normalized log likelihood and Λ_t^i and Λ_a^i are the parameter set of the target and the anti models for the subword p, respectively. Similar to MDE, the objective function of MIE can be written as

$$\begin{aligned}
&\ell_{Ins}(X_{k,l_k}|w_{l_k}^r, w_{l_k}^d, \Lambda) \\
&= PW_I \sum_{i=-1,1} \sum_{n_k=1}^{N_k} \ell\left(d_I(X_{k,l_k+i,n_k}|ph_{l_k+i}^{n_k}, w_{l_k+i}^r, \Lambda)\right) \\
&+ PW_{II} \sum_{i=-1}^{1} \sum_{n_k=1}^{N_k} \ell\left(d_{II}(X_{k,l_k+i,n_k}|ph_{l_k+i}^{n_k}, w_{l_k+i}^d, \Lambda)\right).
\end{aligned} \tag{7}$$

For MSE, as was discussed, the substitution error can be regarded as miss and false alarm errors happening together for the given segment. As is done above, MSE criterion can be easily derived. Finally, the minimization of each objective function can be accomplished through the generalized probabilistic descent (GPD) method [4, 5] with respect to all parameters.

3. Experiments and Results

The experiments reported in this section were carried out on the TIMIT database. In this paper, we focus on word recognition experiments. For baseline Maximum Likelihood (ML) models, we trained context-dependent (CD) target model and context-independent (CI) anti model using the latest version of the HTK toolkit (http://htk.eng.cam.ac.uk/). The CI anti models consist of 41 monophones that are folded from the 48 monophone set defined in [14]. Separately, the cross-word tri-phone target models contain a total of 4328 physical triphone models with 1024 tied-states. In both CI and CD models, all the phones except for the short pause “sp” are modeled by 3-state strict left-to-right HMMs with each state having 8 mixture Gaussian components. The short pause model “sp” has only one state.

In all experiments, we represented the input speech using 39 dimensional feature vectors with 12MFCC, 12Δ, 12$\Delta\Delta$ and 3 log energy values. The standard 3696 training utterances excluding the “sa” utterances and 192 core-test utterances were used for training and testing, respectively. In the word recognition evaluation, we used a bi-gram language model over words estimated from the training set. The WER of the baseline system is 44.59% after 4 iterations with ML estimation using the bi-gram language model. For the proposed discriminative training, a word-loop network is used to generate competing strings in the training data. In addition, the number of training iterations for all MIE/MDE/MSE in Table 1 is fixed to be three.

3.1. Word Recognition Performance

Table 1 shows the performance comparison between the baseline ML and the proposed individual error minimization learning methods on the WER. It can be seen from Table 1, the proposed objective criteria of MIE, MDE, and MSE resulted in primarily reducing its target error type, respectively. Similar to the results in our previous work [1], compared to the ML baseline, MIE and MDE yield more deletion and insertion errors, respectively. As discussed in [1], one possible cause of the instability is the lack of modeling the anti models with a corresponding discriminability.

Furthermore, we tested the scheme of recognizer output voting error reduction (ROVER) [15] as a post-processing

Table 1: Detailed Recognition Performance Comparison

	Ins	Del	Sub	WER
ML	84	89	527	44.59%
MIE	**71**	95	527	44.14%
MDE	87	**76**	533	44.33%
MSE	83	88	**514**	43.63%
ROVER	79	81	522	**43.44%**

scheme for the multiple ASR system combination of the proposed MIE/MDE/MSE. The ROVER algorithm was originally proposed to improve the performance of speech recognition by combining multiple speech recognizers. The outputs of multiple ASR systems are aligned into a word transition network (WTN) by dynamic programming and then a majority voting is performed for each correspondence set. The consensus output yields word error rate (WER) of 43.44%, which is a slight reduction over the best single system MSE of WER of 43.63%. Note that this combination scheme of the individual recognition outputs using ROVER does not extensively explore the issue of sensitivity of individual error minimization since the essence of ROVER is to extract a consensus/unanimity hypothesis from multiple alternatives.

3.2. Rejection Performance on Utterance Verification

As discussed in subsection 2.2, we further investigated the performance of the error rejection rate in utterance verification by using the target and anti models based on likelihood ratio testing. Note that here utterance verification is performed between correctly recognized words versus misrecognized words. The verification performance was measured in three evaluation metrics as follows: (1) the rejection rates (Rej. rate%) of the insertion and substitution errors at a certain false rejection point of correctly recognized words, (2) the equal error rate (EER%), and (3) the re-evaluated WER(%) after the rejection. The confidence measure score CM_w for each hypothesized word w is computed by the Eq. (1) and (2). Based on the confidence score and threshold, the final decision for the hypothesized word w is made as either acceptance or rejection.

In Table 2, similar to the recognition results in the previous subsection, each objective criterion of MIE and MSE improved the rejection rate in its target error type at 7% false rejection point. In particular, both MIE and MSE yield a rejection rate of the insertion error at 40.85% and 42.17%, respectively, which is increased substantially over the ML baseline result. Furthermore, the EER and WER after the rejection of the errors are slightly better than the baseline. These experimental results confirm that the proposed individual error minimization learning framework can effectively improve the rejection rates in utterance verification and thus directly lead to additional word error rate reduction after the rejection.

4. Conclusions and Future Work

We studied the individual error minimization learning framework in word-level and as its applications continuous word recognition and utterance verification tasks were carried out under the proposed learning framework. We proposed the enlarged word-level training token selection scheme and the accompanying learning framework to effectively reduce potential word-level errors. The experimental results in both word recognition and utterance verification confirm that the proposed learning framework leads to the direct minimization of the word-level individual errors. Our future work involves developing the CD anti-subword modeling for better discriminability during training and embedding the proposed learning framework into lattice-based hypotheses for large-scale speech recognition tasks. Furthermore, we will investigate certain rule-based combination methods for the multiple ASR outputs from the MDE/MIE/MSE rather than the majority voting method such as ROVER shown in this paper.

Table 2: Detailed UV Performance Comparison

	Ins (Rej. rate)	Sub (Rej. rate)	EER	After Rej., WER
ML	54 (35.71%)	376 (28.65%)	37.51%	38.44%
MIE	**42** (40.85%)	370 (29.79%)	36.00%	37.70%
MSE	48 **(42.17%)**	**358 (30.35%)**	34.43%	36.76%
MDE	57 (34.48%)	374 (29.83%)	36.68%	37.81%

5. Acknowledgement

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), Rep. of Korea.

6. References

[1] Shin, S., Jung, H.-Y. and Juang, B.-H., "Discriminative Training for Direct Minimization of Deletion, Insertion and Substitution Errors," in *ICASSP 2011*, pp. 5328-5331, May 2011

[2] Eskenazi, M., "An Overview of Spoken Language Technology for Education," *Speech Communication*, vol. 51, pp. 832-844, 2009.

[3] Doremalen, J., et al., "Utterance Verification in Language Learning Applications," in *SLaTE 2009*, Sep. 2009.

[4] Rahim, M. and Lee, C.-H., "String-based minimum verification error (sb-mve) training for speech recognition," *Computer Speech and Language*, vol. 11, pp. 147-160, 1997.

[5] Fu, Q., "A generalization of the minimum classification error (MCE) training method for speech recognition and detection," Ph.D. thesis, Georgia Institute of Technology, 2008.

[6] Lee, C.-H. and Juang, B.-H., "A new detection paradigm for collaborative automatic speech recognition and understanding," in *SWIM 2004*, Jan. 2004.

[7] Lee, C.-H., Clements, M., Dusan, S., Fosler-Lussier, E., Johnson, K., Juang, B.-H. and Rabiner, L., "An overview on automatic speech attribute transcription (ASAT)," in *Interspeech 2007*.

[8] Rahim, M., Lee, C.-H. and Juang, B.-H., "Discriminative utterance verification for connected digits recognition," *IEEE Trans. on Speech and Audio Processing*, vol.5, pp.266-277, 1997.

[9] Lleida, E. and Rose, R., "Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures," *IEEE Trans. on Speech and Audio Processing*, vol.8, pp.126-139, 2000.

[10] Shin, S., Jung, H.-Y. and Juang, B.-H., "An Adaptive Utterance Verification Framework using Minimum Verification Error Training," *ETRI Journal*, vol. 33, no. 3, pp. 423-433, June 2011.

[11] Rosenberg, A., Siohan, O. and Parthasarathy, S., "Speaker verification using minimum verification error training," in *ICASSP 1998*, pp. 105-108, 1998.

[12] Shin, S., Jung, H.-Y., Kim, T.-Y. and Juang, B.-H., "Discriminative linear-transform based adaptation using minimum verification error," in *ICASSP 2010*, pp. 4318-4321, Mar. 2010.

[13] Fu, Q. and Juang, B.-H., "A study on rescoring using HMM-based detectors for continuous speech recognition," in *ASRU 2007*, pp. 570-575, 2007.

[14] Lee, K.F. and Hon, H.W.,"Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acoustics, Speech, & Signal Processing*, vol.37, no.11, pp.1641.1648, 1989.

[15] Fiscus, J., "A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER)," in *ASRU 1997*, pp. 347-354, 1997.