Final Report on ETRI Project

Sunghwan Shin and Biing-Hwang (Fred) Juang
Georgia Institute of Technology
01/31/2011

1. Project Description

Last year, we focused on the problems of proper detection and rejection in an isolated word recognition task consisting of a set of keywords and out-of-vocabulary (OOV) words for real application scenarios under mismatched conditions such as a new speaker or a new environment which is not matched with the original speakers or environments during training. Two main achievements have been reported in [1] and [2]. In [1], we presented an investigation of the minimum verification error linear regression (MVELR) method for discriminative linear-transform based adaptation. The minimum verification error (MVE) criterion is employed to estimate a set of discriminative linear transformations which achieve the smallest empirical average loss with the given adaptation data. The reported results confirmed the proposed MVELR method significantly reduces the total error rate over all categories of the detectors compared to the conventional adaptation method, maximum likelihood linear regression (MLLR). On the other hand, in [2] an adaptive and integrated utterance verification (UV) framework using the MVE training criterion was proposed as a new set of solutions for the entire utterance verification system in real application scenarios. In contrast to the conventional utterance verification framework in which the label information (segments/boundary) obtained from the recognition model is fixed throughout the training session, we proposed the adaptive reusability of the label information obtained from the MVE-trained target model at every iteration during the discriminative training of the verification models. In addition, in the context of the conventional utterance verification, the recognized hypotheses do not change regardless of the utterance verification models. On the contrary, we proposed the use of the target models updated in the MVE training for the recognition stage to obtain improved segmentation and duration in a way consistent with the verification models and the hypothesis testing. Under this new framework, remarkable enhancement in not only recognition performance, but also verification performance has been obtained.

As an extended research, the main goal of this year's project is to investigate the problem and issues related to the rejection of invalid input when applied to a continuous speech recognition (CSR) system. In the case of CSR, 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. This is in contrast to the case of isolated word recognition where only the substitution error is of concern. 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 because currently there is no evaluation guideline 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, separately or jointly. In this report, we summarize first our investigation of the minimization of individual types of error, and then the result of multi-objective minimization.

The main application of this year's project is a spoken dialog system for English education of a non-native speaker such as Korean. In the given application, each of these three errors is treated with a different degree of significance according to the learning level or under the certain situation for the English education. In addition, proper rejection for the insertion and substitution errors is also required to provide a user-friendly interface with flexibility in accepting/rejecting a wide range of user responses. Therefore, this year we have investigated a new training paradigm to selectively minimize each of the errors and at the same time to effectively reject insertion and substitution errors. For the ultimate goal, we follow the minimum error principle for acoustic modeling and formulate error objectives in insertion, deletion, and substitution separately for minimization during training.

Several Discriminative Training (DT) methods, such as maximum mutual information estimation (MMIE) [3], minimum classification error (MCE) [4], and minimum phone/word error (MPE/MWE) [5], have achieved success in various speech recognition tasks over years. Among them, the MCE and the MPE/MWE aim at direct minimization of mainly 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 rethink the recognition problem as a detection problem.

Table 1. Comparison between different problem descriptions

	Error Type	Alignment Errors	Training Criterion
Recognition Problem	Misclassification Error	Deletion Insertion Substitution	Minimize sub-errors only
Detection Problem	Type I/II (Miss/FA) Errors	Type I Type II Type I&II	Minimize Type I&II both

In this report, based on the above analysis, we propose a new multi-objective discriminative training method using minimum verification error criterion (MVE) [6,7] 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 string-level MCE [4], 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 named each objective criterion minimum deletion error (MDE), minimum insertion error (MIE), and minimum substitution error (MSE), respectively.

The rest of this report 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 MVE criterion. In Section 3, we will show the effectiveness of the individual criterion in minimizing its target error for continuous speech recognition task. In particular, the experimental results will be presented separately in the continuous phone recognition (CPR) task on TIMIT corpus and the large vocabulary continuous speech recognition (LVCSR) task on WSJ corpus. 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 [4] in 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:

$$d(X_k|\Lambda) = -g(X_k, S_r|\Lambda) + G(X_k, S_n|\Lambda) \quad (1)$$

where Λ 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 [8]. 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 and the one best decoded string from ASR is W^d. After a DP-based string alignment procedure, one deletion

error W_2^r and one insertion error W_2^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 as a miss error in the detection problem since W_2^r has to exist on the decoded string but it is missed with respect to the decoded output sequence. On the other hand, W_2^d has to be rejected but it is inserted on the decoded output sequence. Thus, the insertion error W_2^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_2^r and the insertion error W_2^d are trained by the first mis-verification measure $d_I(X_k, W_2^r\,|\Lambda)$ and the second mis-verification measure $d_{II}(X_k, W_2^d\,|\Lambda)$, respectively, as follows:

$$d_I(X_k, W_2^r\,|\Lambda) = -g_t(X_k, W_2^r|\Lambda_t) + g_a(X_k, W_2^r|\Lambda_a) \tag{2}$$

$$d_{II}(X_k, W_2^d\,|\Lambda) = +g_t\left(X_k, W_2^d\middle|\Lambda_t\right) - g_a\left(X_k, W_2^d\middle|\Lambda_a\right) \tag{3}$$

where d_I and d_{II} are the type I and type II mis-verification measures, respectively. In Eq. (2) and (3), g_t and g_a are the normalized log likelihood and Λ_t and Λ_a are the parameter set of the target and the anti model for the given segment, respectively.

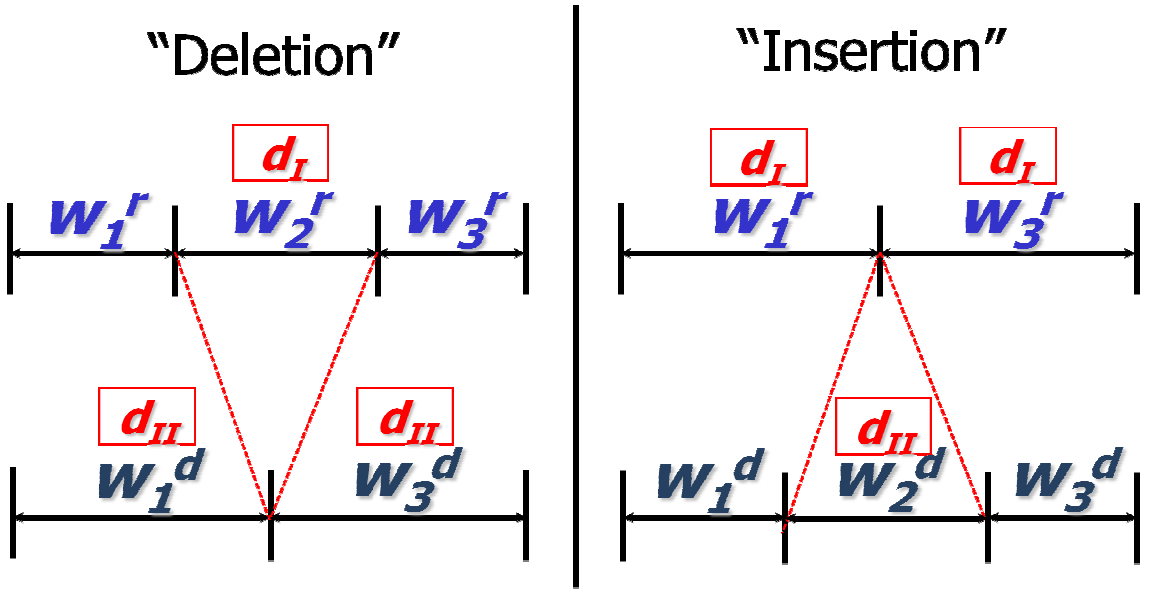

Fig. 1. Error count and corresponding mis-verification measures (d_I and d_{II}) under MVE criterion

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_2^r and W_2^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_I 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 [1,10] and rescoring hypotheses [11,12] from 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 $\{X_1^r, X_2^r, \dots, X_k^r\}$ for the reference transcript and $\{X_1^d, X_2^d, \dots, X_k^d\}$ for the decoded output. From the samples and error assignments of the decoded output, the empirical average loss is defined by

$$L(\widetilde{\Lambda}) = \frac{1}{K}\sum_{k=1}^{K} l_{total}(X_k^d|\,\Lambda) \tag{4}$$

where $l_{total}(X_k^d|\Lambda)$ is the composite loss function which combines four different types of the recognition output from the general DP-based string error assignment. For the multi-objective discriminative learning, the composite loss function can be described as

$$\begin{aligned} l_{total}(X_k^d|\Lambda) = {} & l_{Del}(X_k^d|\Lambda)1(X_k^d \in \text{"}Del\text{"}) + l_{Ins}(X_k^d|\Lambda)1(X_k^d \in \text{"}Ins\text{"}) \\ & + l_{Sub}(X_k^d|\Lambda)1(X_k^d \in \text{"}Sub\text{"}) + l_{Hit}(X_k^d|\Lambda)1(X_k^d \in \text{"}Hit\text{"}) \end{aligned} \tag{5}$$

where $l_{Del}(\cdot)$, $l_{Ins}(\cdot)$, and $l_{Sub}(\cdot)$ are respectively represented for an individual objective function: MDE, MIE, and MSE. First, the objective function for MDE can be written as

$$\begin{aligned} l_{Del}(X_k^d|\Lambda) = {} & PW_I \sum_{i=1}^{M} l\left(d_I(X_k^r|\Lambda^i)\right) 1(X_k^r \in C_i) \\ & + PW_{II} \sum_{j=-1,1} \sum_{i=1}^{M} l\left(d_{II}(X_{k+j}^d|\Lambda^i)\right) 1(X_{k+j}^d \in C_i) \end{aligned} \tag{6}$$

where PW_I and PW_{II} are the penalty weights for type I and type II errors and $l(\cdot)$ is a smoothed loss function normally defined by a sigmoid function [4]. Note that the two kinds of mis-verification measures are separately assigned to the reference segment X_k^r and decoded segment X_{k+j}^d as defined by

$$d_I(X_k^r|\Lambda^i) = -g_t(X_k^r|\Lambda_t^i) + g_a(X_k^r|\Lambda_a^i) \tag{7}$$

$$d_{II}(X_{k+j}^d|\Lambda^i) = +g_t(X_{k+j}^d|\Lambda_t^i) - g_a(X_{k+j}^d|\Lambda_a^i). \tag{8}$$

Unlike Eq. (1), in Eq. (7) and (8), g_t and g_a are the segment-based normalized log likelihood and $\Lambda_t^i$and Λ_a^i 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

$$l_{Ins}(X_k^d|\Lambda) = PW_I \sum_{j=-1,1} \sum_{i=1}^{M} l\left(d_I(X_{k+j}^r|\Lambda^i)\right) 1(X_{k+j}^r \in C_i) + PW_{II} \sum_{i=1}^{M} l\left(d_{II}(X_k^d|\Lambda^i)\right) 1(X_k^d \in C_i). \quad (9)$$

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 we did above, the objective function of MSE can be formulated as

$$l_{Sub}(X_k^d|\Lambda) = PW_I \sum_{i=1}^{M} l\left(d_I(X_k^r|\Lambda^i)\right) 1(X_k^r \in C_i) + PW_{II} \sum_{i=1}^{M} l\left(d_{II}(X_k^d|\Lambda^i)\right) 1(X_k^d \in C_i). \quad (10)$$

Last, as the conventional segment-based MVE, the *hit* tokens can be optionally trained either on the reference transcript or on the decoded output. In this report, we will derive for only reference transcript as follows:

$$l_{Hit}(X_k^r|\Lambda) = PW_I \sum_{i=1}^{M} l\left(d_I(X_k^r|\Lambda^i)\right) 1(X_k^r \in C_i) + PW_{II} \sum_{i=1}^{M} \sum_{j,j\neq i}^{M} l\left(d_{II}(X_k^r|\Lambda^j)\right) 1(X_k^r \in C_i). \quad (11)$$

Finally, the minimization of each objective function can be accomplished through the generalized probabilistic descent (GPD) method [4,6] w.r.t. all parameters. In the following experiments, we used uniform penalty weights for both PW_I and PW_{II}. Moreover, the experiments in this report 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 [17] over the proposed learning framework.

3. Experimental Setup and Results

The experiments reported in this section are carried out on TIMIT corpus and WSJ corpus, respectively. The TIMIT corpus was mainly used for continuous phone recognition (CPR) task and the WSJ corpus was used for large vocabulary continuous speech recognition (LVCSR) task. In TIMIT, we use the standard experimental setup as specified in [13,14]. We first present word recognition results as a preliminary experiment and then show phone recognition results. Note that the word recognition results are not intended to the general discriminative training scenario under separate training and testing sets. The preliminary experiment is based only on a testing set for both discriminative training phase and testing phase. From this experiment, we could observe how much the proposed method directly reduced the errors on the testing set. In addition, this

experiment is performed by a simple combination of the multi-objective training methods such as D+S+I and H+D+S+I. On the other hand, for the continuous phone recognition, we followed the general discriminative training scenario and will present the detailed performance by each objective training criterion as well as the performance by the combined objective training criteria. In WSJ, the baseline system is built following the recipe (http://www.inference.phy.cam.ac.uk/kv227/htk/) with WSJ0 database and we followed the general discriminative training scenario with the training corpus, the SI-84 set, and the test corpus, standard Nov92 set. For LVCSR task on WSJ, we will first present recognition results and then show verification performance as a hypothesis testing of the individual recognized words.

3.1. Preliminary Experiment - Continuous Word Recognition (CWR) Task on TIMIT corpus

In continuous word recognition task on TIMIT, the total number of words in the lexicon is 6,078 and for the ML models as a baseline the standard TIMIT training set 3,696 utterances were used. As a comparative method in the recognition, we will present MCE performance. For MCE and E-MVE training and testing, the standard TIMIT testing set 1,344 utterances were used in both. As the baseline performance by the ML models, the word error rate on testing set is shown in Table 2. With 41 CI monophones and a bigram language model, the word error rate of 46.89 % was observed. Based on the comparison between testing set reference transcription and 1-best decoding output, we set up two kinds of Extended-MVE (E-MVE) training scenarios as a simple combination of the individual objective training criteria. The first scenario attempts to use all tokens consisting of all error tokens including hit tokens on testing set. The second scenario attempts to use only error tokens assigned into deletion, insertion, and substitution.

Table 2. Detailed recognition performance with verification

	Hit	Del	Sub	Ins	WER
ML	6,114	1,135	3,776	259	46.89 %
MCE	7,624	792	2,609	226	32.89 %
E-MVE (H+D+S+I)	7,700	511	2,814	428	34.04%
	6,930 (10%)	511	**1,822 (35.22%)**	**268 (37.15%)**	**32.56 %**
E-MVE (D+S+I)	7,299	836	2,890	222	35.81%
	6,570	836	**2,127**	**168**	**34.91 %**

In the Table 2, each row shows the detailed recognition performance by each learning method. The second row shows the MCE performance and the third row, E-MVE by HDSI, means all the tokens on the testing output were used for the E-MVE training. The fourth row, E-

MVE by DSI, means only error tokens were used for the E-MVE training. In the performance analysis, both MCE and E-MVE significantly reduced the number of word errors on testing set. First, in the Correct (%), E-MVE shows better performance than MCE. However, in total word error rate, MCE is slightly better than E-MVE because of the number of increased insertion errors in E-MVE. In particular, as shown in the table, we could see a trade-off in the deletion and insertion errors between the two E-MVE training scenarios. In E-MVE including Hit tokens, the number of deletion error drops from 1,135 to 511 but the number of insertion error was a lot increased compared to MCE. One possible reason is that MVE training with Hit tokens may produce over-estimated models. As we saw in the derivation of E-MVE with Hit tokens, one true class is computed by the first misclassification measure, d_I, and at the same time, given the same observation (segment), all other classes excluding the true class are computed by the second misclassification measure, d_{II}. That might impact overestimation. On the other hand, without hit tokens during the training phase all types of errors were reduced compared to ML baseline. In particular, MCE mainly reduced the substitution errors as following its training criterion. However, E-MVE reduced the deletion and insertion errors more than MCE.

Furthermore, the advantage of E-MVE is not limited only to the recognition stage, but can be further extended to the verification phase. In the Table 2, each E-MVE method has two rows and the second row is the performance after the verification. For the verification phase, each word is considered as an independent hypothesis and then with MVE-trained corresponding target and anti models we did hypothesis testing. Note that the specific false rejection rate at the verification stage is 10%. Thus, the number of Hit is decreased by ten percent for each E-MVE method. And the number of deletion error does not change at all since the verification was performed on the 1-best decoding output. As we discussed in the introduction, there is no evaluation measure for deletion errors in the given task as well as in many general ASR systems. However, with respect to the substitution and insertion errors, over 35 % rejection rate on both errors were obtained. Finally, in the total word error rate, we could achieve additional error reduction rate as follows in the table. Of course, this hypothesis testing can be re-defined on keyword detection and OOV rejection task as we did in [2].

3.2. Continuous Phone Recognition (CPR) Task on TIMIT corpus

In this section, we will report phone recognition results. Unlike the previous subsection, we followed the general discriminative training scenario and will present the detailed performance by each objective training criterion as well as the performance by the combined objective training criteria. As baseline, we trained both context-independent (CI) and context-dependent (CD) HMM phone recognizers using the latest version of the HTK toolkit. The CI system consists of 48 monophones defined in [13] 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. It is obvious that the CD anti-subword models with the same context as the CD target model would affect the direct performance improvement as shown in [15]. As a future work in the following year, we will develop the CD anti-subword models and directly apply them to the proposed training framework.

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 [13] 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 3 and 4 is fixed to be five.

3.2.1 CI Results

Table. 3 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.

Table 3. Detailed performance comparisons on CI system

	Del	**Ins**	**Sub**	**Acc %**
ML	678	170	1,289	70.57 %
MCE	674	179	**1265**	70.83 %
MDE	**655**	175	1279	70.95 %
MIE	687	**156**	1278	70.79 %
MSE	691	159	**1273**	70.76 %
D+I+S	687	159	1272	70.83 %
H+D+I+S	521	274	1278	71.45 %

It can be seen from Table 3, 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 [17], unlike the simple combinations reported here, may bring about more overall error reduction.

3.2.2 CD Results

Table 4 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 4.

Table 4. Detailed performance comparisons on CD system

	Del	**Ins**	**Sub**	**Acc %**
ML	444	246	1,271	73.26 %
MCE	448	250	**1,265**	73.23 %
MDE	**435**	266	1254	73.34 %
MIE	468	**222**	1257	73.45 %
MSE	453	243	**1245**	73.53 %
D+I+S	449	243	1255	73.45 %

Similar to the results in [16], 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 [15].

3.3. Large Vocabulary Continuous Speech Recognition (LVCSR) Task on WSJ corpus

We also evaluate the proposed discriminative training method on a large vocabulary continuous speech recognition (LVCSR) task, namely WSJ0 database. The training corpus is the SI-84 set with 7133 utterances from 84 speakers and the test set standard Nov92 with 330 utterances from 8 speakers. The baseline system is built following the recipe (http://www.inference.phy.cam.ac.uk/kv227/htk/) for WSJ database using the Hidden Markov Model Toolkit (HTK). Crossword tri-phone models with a total number of 2750 tied-states are trained, which are represented by 3-state strict left-to-right HMMs with each state having 8 mixture Gaussian components. Note that the anti models are currently limited to 40 CI monophones similar to CPR task on TIMIT corpus in the previous subsection and all the anti-phones are modeled by 3-state strict left-to-right HMMs with 16 Gaussians per state. In addition, the input feature is 12MFCCs + energy, and their first and second order time derivatives. The

WER of the baseline system is 7.17% after 5 iterations with maximum likelihood estimation (MLE) using a standard bi-gram language model. The parameter set estimated by the baseline system acts as the initial model for the all discriminative training experiments, which along with a weak uni-gram language model trained from the training corpus are used to generate the competing strings/segments to the reference transcription.

In order to appropriately apply the proposed learning framework to the LVCSR task, the extended error counting and learning framework should be considered since in the LVCSR task the training and recognition decisions are not on the same linguistic level of the performance measure. For example, the error count and learning framework for the CPR task shown in Fig. 1 take account into the each phoneme segment in the recognition decision and maps the segment directly to the training token in the discriminative training phase. However, in LVCSR task the training and recognition are performed on the *phone* level but the system evaluation measure is defined on the *word* level. Hence, we propose the extended error count and learning framework as shown in Fig 2.

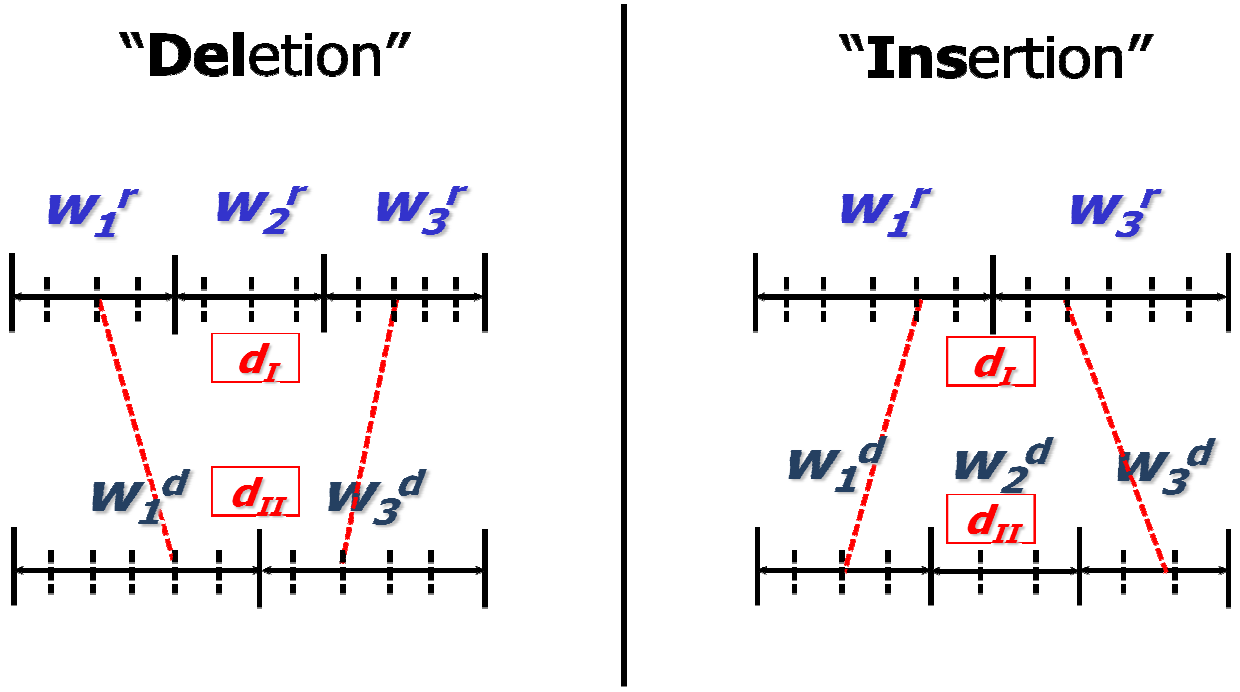

Fig. 2. Extended Error count and learning framework for LVCSR task

In contrast to the previous framework counting the proceeding and succeeding segments (phonemes) as the training tokens no matter what they are assigned from the DP-matching, more or less phonemes/segments can be selectively taken according to their word level decision and user's pre-defined setting. For example, here all word-level classes, W^r and W^d, determined by the recognition decision are internally represented as a sequence of phoneme segments and the number of the phonemes for the discriminative training is set to 2 in Fig.2. Of course, one can take whole phonemes in the preceding and succeeding words but it is not reasonable when they are assigned as *Hit* word from the recognition decision. Thus, we propose whole phonemes are chosen only if the preceding and succeeding words are assigned as errors (insertion/substitution/deletion). On the other hand, if they are assigned as *Hit* word, we use pre-

specified numbers for taking the preceding and succeeding phonemes. In our investigation, the best performance has been obtained when the number of phonemes is set to 2 as shown in Fig. 2.

In the following subsection, we first present recognition results and then show verification performance as a hypothesis testing of the individual recognized words.

3.3.1 Recognition results

Table. 5 shows the recognition performance comparison between the baseline MLE and the proposed multi-objective discriminative training method, MIE/MSE/MDE. In particular, we present the detailed performance of each objective criterion and recognizer output voting error reduction (ROVER) [18] as a post-processing scheme for the multiple ASR system combination of the proposed MIE/MSE/MDE. 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. Here, ROVER works at the output of the multiple ASR systems and thus no acoustic and language models are involved.

Table 5. Detailed recognition performance comparisons on WSJ corpus

	Hit	**Del**	**Sub**	**Ins**	**WER %**
ML	5053	21	279	84	7.17 %
MIE	5059	21	273	**78**	6.95 %
MSE	5062	24	**267**	82	6.97 %
MDE	5056	**20**	277	86	7.15 %
ROVER	5064	21	268	80	**6.89 %**

It can be seen from Table 5, each objective criterion of MIE and MSE resulted in primarily reducing its target error type, respectively. However, MDE reduced only one out of twenty one in the deletion error. It is mainly because the number of the deletion errors is substantially small compared to the total word error rate (WER %) and most of them are either an article such as “a” and “the” or one short syllable-based word such as “in” and “on”. As we have already observed in TIMIT corpus, we anticipate that MDE criterion would effectively reduce the deletion error in another type of LVCSR corpus rather than WSJ corpus. Finally, we tested ROVER for combining the recognition outputs from the three different training criteria, MIE/MSE/MDE. The consensus output yields word error rate (WER) of 6.89%, which is a slight reduction over the best single system of WER of 6.95%.

3.3.2 Verification results

For the hypothesis testing of the given recognized word, after the recognition (decoding) with the MVE-trained target model, we first consider subword-level (monophone-level) acoustic verification scores based on the corresponding target subword and anti-subword models and then add up them for the word-level verification scores as same as we did in subsection 3.1. The verification performance was measured in two evaluation metrics, equal error rate (EER %) and rejection rate of substitution and insertion errors at the certain false rejection rate of hit words. In practice, one usually has to pick a specific false acceptance or false rejection rate as part of the operating specifications on the system [19]. It is more desirable to optimize verification performance at a particular operating point instead of equal error rate (EER %). Nevertheless, the EER has been widely used as one of the important performance metrics in the fields of detection and verification researches. Thus, we also include the EER performance in our overall performance evaluation.

Table 6. Detailed verification performance comparisons
in Equal Error Rate (EER %)

	Del	**Sub-EER**	**Ins-EER**	**Total EER**
ML	·	38.34 %	36.92 %	38.29 %
MIE	·	36.27 %	**31.18 %**	35.46 %
MSE	·	**34.83 %**	36.59 %	35.25 %
MDE	·	36.84 %	36.05 %	36.71 %

Table 7. Detailed verification performance comparisons
at 7% false rejection point

At 7% False Rejection Point	**Del**	**Sub-Rejection Rate**	**Ins-Rejection Rate**
ML	·	27.96 %	30.95 %
MIE	·	26.74 %	**42.86 %**
MSE	·	**29.09 %**	35.37 %
MDE	·	28.16 %	36.05 %

Table. 6 shows the verification performance in terms of EER % and Table. 7 presents the verification performance of the substitution and insertion errors at a 7% false rejection point of hit words. Similar to the recognition results in the previous subsection, each objective criterion of MIE and MSE mainly reduced the equal error rate in its target error type. For example, the EER in MIE drops from 36.92 % to 31.18 % and MSE yields the EER of 34.83 %, which is a lot reduced over the ML baseline of the EER of 38.34%. Furthermore, at 7 % false rejection rate, the rejection rate of the insertion error is significantly increased from 30.95 % up to 42.86 %. On the other hand, the rejection rate of the substitution error is slightly increased to 29.09 % at 7 % false rejection rate. As we discussed throughout Section 3, the anti models play a critical role in not only the proposed discriminative training phase, but also the verification phase (hypothesis testing). We expect that use of the CD anti-subword models discriminatively trained with the corresponding CD target models would directly lead to improved performance in both recognition and verification.

4. Conclusions and Planned Future Works

In this report, 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(erification)E criterion 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. Furthermore, we studied combined multi-objective training criteria by incorporating the three individual objectives and the "hit" segments and recognizer output voting error reduction (ROVER) as a post-processing scheme for the multiple ASR system combination of the proposed MIE/MSE/MDE. We first 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. Then, we also evaluated the proposed discriminative training method with the extended error count and learning framework which takes account into the variable number of the phonemes in preceding and succeeding words for LVCSR task on WSJ corpus. Similar to the results in CPR task on TIMIT corpus, each objective criterion of MIE and MSE resulted in primarily reducing its target error type, respectively. Furthermore, the verification performance both in terms of EER % and at 7 % false rejection rate was significantly improved.

Although we have obtained various preliminary results with positive potential both in CPR and LVCSR tasks, there are still many challenging issues. First of all, the CD anti-subword modeling is very important in both discriminative training and hypothesis testing. We anticipate well-constructed CD anti-subword models will lead to further performance improvement as shown in [15]. Another possible issue can come from associating the preceding and succeeding segments for the error-counted segment with non-uniform error cost like [9]. Furthermore, we plan to apply task-dependent non-uniform penalty weights for both PW_I and PW_{II} and further investigate certain rule-based optimization methods such as [17] rather than the majority voting method such as ROVER for the improved multi-objective systems. Finally, for the long term goal of our collaboration in the following year, we propose the extension of detection-based ASR [20,21,22], an alternative paradigm to conventional ASR techniques, which constructs a bottom-up hypothesis testing framework based on the detection theory. The biggest advantage of the

detection-based ASR is its flexibility to incorporate different knowledge sources and the ability to fuse lower level information (e.g., attributes) into higher level hypotheses (e.g., events), and at the same time to neglect superfluous input that is not relevant to the speech decoding task.

5. References

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

[2] S. Shin, H.-Y. Jung, and B.-H. Juang, "An adaptive utterance verification framework using minimum verification error training," accepted to appear in *ETRI Journal*, vol. 33, no. 3, June 2011.

[3] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition, " in *ICASSP 1986*, 1986, vol. 1, pp. 49–52.

[4] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp.257-265, May 1997.

[5] D. Povey, "Discriminative learning for large vocabulary speech recognition, " *Ph.D. thesis*, 2004.

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

[7] A. E. Rosenberg, O. Siohan, and S.Parthasarathy, "Speaker verification using minimum verification error training," in *ICASSP-98*, 1998, pp. 105-108

[8] Y.-F. Liao, J.-J. Tu, S.-C. Chang, and C.-H. Lee, "An enhanced minimum classification error learning framework for balancing insertion, deletion and substitution errors," in *ASRU 2007*, Kyoto, Japan, Dec. 2007, pp. 587-590.

[9] Q. Fu, D. Mansjur, and B.-H. Juang. "Empirical system learning for statistical pattern recognition with non-uniform error criteria," *IEEE Transactions on Signal Processing*, vol. 58, pp.4621-4633, Sep. 2010.

[10] Q. Fu and B.-H. Juang, "Segment-Based Phonetic Class Detection Using Minimum Verification Error (MVE) training," in *Interspeech-2005*, Lisbon, Portugal, Sep. 2005.

[11] Q. Fu and B.-H. Juang, "Investigation on rescoring using minimum verification error (mve) detectors," in *Interspeech-2006*, Pittsburgh, PA, Sep. 2006.

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

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

[14] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 137–152, 2003.

[15] P. Ramesh, C-H. Lee, and B-H. Juang. "Context dependent anti subword modeling for utterance verification," In *Proceedings of ICSLP-1998*, pages 3233–3236, 1998.

[16] Q. Fu, X. He, and L. Deng, "Phone-Discriminating Minimum Classification Error (P-MCE) Training for Phonetic Recognition," in *Interspeech-2007*, Antwerp, Belgium, Aug. 2007.

[17] S. Yaman and C.-H. Lee, "A flexible classifier design framework based on multi-objective programming," *IEEE Trans. on Audio, Speech, and Language Processing*, vol.16, no.4, pp.779-789, 2008.

[18] J. Fiscus, “A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER),” in Proc. of ASRU, 1997, pp. 347–354.

[19] M.-H Siu, B. Mak, and W-H. Au, “Minimization of Utterance verification error rate as a constrained optimization problem,” *IEEE Signal Processing Letters*, vol. 13, pp. 760-763, Dec. 2006.

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

[21] I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Mattews, A. Moreno-Daniel, J. Morris, Y. Tsao, and Y. Wang, “Detection-based ASR in the automatic speech attribute transcription project,” in *Interspeech 2007*, pp. 1829-1832, 2007.

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