An Adaptive Utterance Verification Framework Using Minimum Verification Error Training

Sung-Hwan Shin, Ho-Young Jung, and Biing-Hwang Juang

This paper introduces an adaptive and integrated utterance verification (UV) framework using minimum verification error (MVE) training as a new set of solutions suitable for real applications. UV is traditionally considered an add-on procedure to automatic speech recognition (ASR) and thus treated separately from the ASR system model design. This traditional two-stage approach often fails to cope with a wide range of variations, such as a new speaker or a new environment which is not matched with the original speaker population or the original acoustic environment that the ASR system is trained on. In this paper, we propose an integrated solution to enhance the overall UV system performance in such real applications. The integration is accomplished by adapting and merging the target model for UV with the acoustic model for ASR based on the common MVE principle at each iteration in the recognition stage. The proposed iterative procedure for UV model adaptation also involves revision of the data segmentation and the decoded hypotheses. Under this new framework, remarkable enhancement in not only recognition performance, but also verification performance has been obtained.

Keywords: Utterance verification, minimum verification error (MVE) training, adaptive framework.

Manuscript received Aug. 16, 2010; revised Nov. 11, 2010; accepted Dec. 22, 2010.

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.

Sung-Hwan Shin (phone: +1 404 895 1817, email: shshin@ece.gatech.edu) and Biing-Hwang Juang (email: juang@ece.gatech.edu) are with the Center for Signal and Image Processing, Georgia Institute Technology, Atlanta, USA.

Ho-Young Jung (email: hjung@etri.re.kr) is with the Software Research Laboratory, ETRI, Daejeon, Rep. of Korea.

doi:10.4218/etrij.11.0110.0489

I. Introduction

Conventional automatic speech recognition (ASR) systems are generally task specific with a fixed system construct, such as vocabulary and grammar, which does not provide a user-friendly interface with flexibility in accepting a wide range of user responses. Oftentimes, the performance of these systems is seriously degraded by out-of-vocabulary (OOV) words (improper input utterances) spoken by the user or mismatched operating designs, such as different training and testing conditions. To enhance the ASR performance for a friendlier voice user interface, it is necessary to provide a mechanism for verifying the level of confidence in the recognition results. Such a mechanism should reject OOV utterances as well as potentially misrecognized utterances (to allow, for example, reconfirmation from the user) to avoid detriments caused by senseless recognition errors. This is often called utterance verification (UV) [1]-[3].

UV is considered a hypothesis testing problem [4]-[6]. In this paper, UV refers to the ability to accept or reject a hypothesized word corresponding to a correctly decoded keyword, an incorrectly decoded keyword, or an OOV word. This capability, different from the conventional formulation of speech recognition, is implemented as a likelihood ratio-based hypothesis testing procedure for verifying individual subword units in a decoded word as a result of ASR decoding. That is, the verification is performed as post-processing after the recognition, and thus the UV performance conventionally shows how well the hypothesis testing can be done on the given ASR output. However, during testing, if we consider an utterance from a new speaker or a new environment which is not matched to the original speakers or environments during training, not only the recognition performance but also the

verification performance would degrade rapidly. In real-world applications, such mismatched scenarios are unavoidable. In order to overcome serious performance degradation in these scenarios, adaptation methods, for example, maximum likelihood linear regression [7] and minimum classification error (MCE) linear regression [8], have been investigated in recent years, and significant progress has been made for speech recognition. However, the issue of adaptation as a method to mitigate performance degradation in UV due to condition mismatch has not been established. Hence, in this paper, we propose minimum verification error (MVE) training [9]-[11] as the training method for UV in various adaptation scenarios. The essence of MVE is to directly minimize the total verification errors from both type I errors (miss) and type II errors (false alarm) with the given adaptation data. Based on the MVE principle, we propose a new adaptive and integrated UV framework as a new set of solutions to enhance the overall system performance in such mismatched scenarios.

In real application scenarios, conventional UV is commonly considered to be an add-on component in a modular approach consisting of two modules: recognition and verification. In the stage of recognition, knowledge sources, such as the computed likelihoods and the segmented durations, are used for finding all hypotheses. In the stage of verification, knowledge sources, such as likelihood ratios computed on the segments provided from the recognizer, are used as post-processors for accepting or rejecting the hypotheses. Although the two stages may jointly affect the overall verification performance, many researchers have been considering the first stage (recognition stage) and the second stage (verification stage) separately. Integrating speech recognition and UV in a single decoding scheme is believed to be able to offer substantial performance improvement, particularly for speech signals that contain OOV words, ill-formed words, or ill-modeled utterances. Past attempts at such integration include the hybrid decoder of Koo, Lee, and Juang [12] and the one-pass likelihood ratio-based decoder of Lleida and Rose [2]. Although these proposals take advantage of the information from anti-models and likelihood ratio testing, the benefits in general do not materialize simultaneously in terms of recognition and verification performances.

In this paper, we propose an integrated solution for the two stages. By adapting and sharing the target model based on the MVE method instead of the acoustic model for ASR, at each iteration in the recognition stage, we obtain an improved decoder in which a much reduced recognition error rate and more accurate segmentation (boundaries) on the hypotheses can be accomplished. This revision of recognition hypotheses helps to increase the consistency between the data offered for verification and the models employed for the test. Moreover, an updated transcription with the new segmentation realigned by the current-stage target model is used for the next discriminative training stage. In this paper, we report our study on the incorporation of this new adaptive strategy (involving both the hypotheses, together with segmentation and the models) in the MVE training. We call this new modeling strategy adaptive-MVE (A-MVE). We show that with this new strategy, remarkable enhancements in both the recognition performance and the verification performance can be obtained.

This paper is organized as follows. In the next section, we describe details of the adaptive UV framework using MVE training, and we review the MVE training method in section III. Experimental setup and results are presented in section IV. Finally, a conclusion is provided in section V.

II. Adaptive UV

In this section, an introduction to the basic framework for the conventional UV is first provided. We then present the new adaptive and integrated methodology as a solution for enhancing the entire UV performance in adaptation scenarios.

The conventional UV framework consists of a recognition stage and a verification stage as shown in Fig. 1. In the recognition stage, the decoder produces a tentatively recognized output for the verification stage. The decoder produces the output using generally trained acoustic (recognition) models such as context independent (CI/monophone) models or context dependent (CD/triphone) models [13]. With the output from the decoder, the verification system considers them as hypotheses and verifies the confidence level for the provided tentative decisions or if they correspond to legitimate input to the system. The UV system determines the scores of the hypotheses by using the corresponding target models and anti-models, a set of the verification models, on the segments of the hypotheses provided by the decoder. Finally, in the evaluation stage, a ratio of the scores is compared to a pre-specified operating threshold. Based on the threshold, a final decision is made as to either accept or reject the hypothesis. The conventional hypothesis testing in the second stage is based on the Neyman-Pearson lemma [4]-[6] which teaches the use of likelihood ratio to accept or reject a proposed hypothesis as defined in

$$LR(k) = \frac{p_k(\mathrm{O}|H_0)}{p_k(\mathrm{O}|H_1)} \gtrless \tau_k; \text{ accept or reject.} \qquad (1)$$

A generalized likelihood ratio is computed when testing data O is observed, and then compared against a decision threshold to decide which one of two hypotheses is to be accepted. The two hypotheses are the null hypothesis H_0 corresponding to the target model and the alternative hypothesis H_1 corresponding

to the anti-model. The hypothesis testing is performed by comparing the likelihood ratio $LR(k)$ to a pre-specified operating threshold τ_k. If the two likelihood functions of $p_k(\mathrm{O}|H_0)$ and $p_k(\mathrm{O}|H_1)$ are known exactly, the above likelihood ratio test is the most powerful test [4]-[6]. However, the true likelihood or distribution functions are unknown in a real application.

The lack of knowledge in the data distribution manifests itself in two different perspectives. First, the most fundamental one is the form of the distribution function. Often the choice of the form of the distribution function is made out of convenience, for example, a Gaussian distribution or a mixture distribution [13]. The second pertains to the specific parameter values that define the chosen distribution function. These parameter values, for example, the mean and the covariance in the case of Gaussian distributions, are estimated from a labeled data set, which is often finite in size. When the chosen distribution form does not really match the real data distribution (for example, a Gaussian distribution is assumed while the data is uniformly distributed), statistical estimation methods do not lead to any meaningful numerical results. A successful alternative to statistical estimation is discriminative training [14]-[16], such as MVE training, which aims at direct minimization of the verification error, rather than fitting of the distributions. Nevertheless, an additional level of uncertainty needs to be addressed; namely, the potential mismatch in the statistical behaviors of the training data and of the field data. Such a mismatch situation includes but is not limited to a change of speaker population or the acoustic ambience. Since the pre-labeled data, normally consisting of phoneme boundaries, the start time and end time of each phoneme on a reference transcription, is at best a limited representation to support the given recognition models, the parameters optimized for a given training set often suffer significant degradation under mismatch operating conditions. In order to perform discriminative training to minimize errors arisen from mismatched conditions, a certain adaptation scheme must be incorporated, including realigning the initial labeled boundaries/segments. Moreover, in order to maintain the training consistency, segmentation should be obtained by the target model for verification rather than the original recognition model. In our implementation, the segments are sequentially updated along with the target model refinement in the discriminative training with the given adaptation data. That is, the verification models are being updated iteratively by discriminative training, using a matched set of data, associated with iteratively obtained labels and segmentations. We elaborate these points below.

In contrast to the conventional UV framework in which the label information obtained from the recognition model is fixed

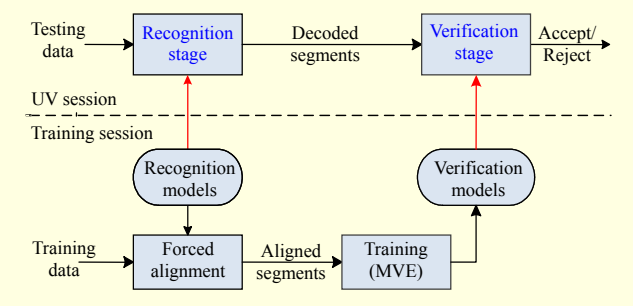

Fig. 1. Basic architecture of two-stage system in conventional UV.

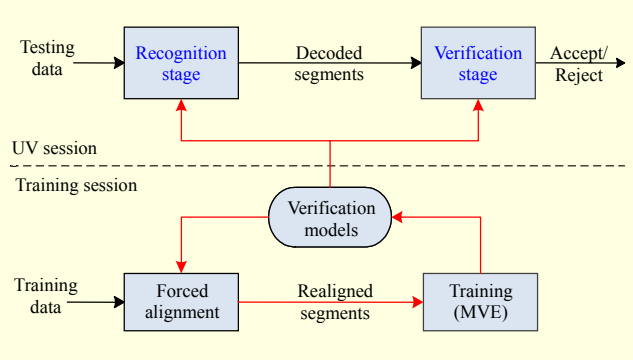

Fig. 2. Adaptive UV framework.

throughout the training stage, our experience indicates that the label information obtained from the target model can be advantageously utilized to adapt the model parameters to the field data for a substantial performance improvement. In this paper, we thus propose this new adaptive reusability of the label information on the transcription during the discriminative training of the verification models using the MVE training method with the given adaptation data.

Furthermore, in the context of the conventional UV as shown in Fig. 1, the recognized hypotheses do not change regardless of the UV models. This limitation of using only the recognized hypotheses carried out by the recognition models may substantially affect the entire verification framework. It is obvious that improved segmentation and duration in a way consistent with the verification models will directly affect the verification performance. Meanwhile, if the recognition error is improved, resulting in a reduced portion of the incorrectly recognized hypotheses, the entire verification framework will deliver a superior performance. Hence, as an integrated solution for the entire verification framework, we also propose the use of the target models updated in the MVE training for the recognition stage again as shown in Fig. 2.

Figure 2 presents a schematic of the proposed solution, which produces significant performance gains in both recognition and verification for the entire verification framework. Although we still present two stages for the

reader's understanding in Fig. 2, it can be considered essentially as one integrated stage associated with only the verification models (the MVE target and anti-models) in contrast to the conventional rigid two stages associated with the inconsistent recognition models and verification models as shown in Fig. 1. In this new framework, at every iteration during the discriminative training, not only the label information for the next MVE training but also the recognized output for the hypothesis testing is sequentially updated by the current-stage MVE target model. Hence, throughout the adaptive UV framework with the MVE training, we can obtain improved decoding results and discriminatively trained verification models for the adaptation data simultaneously. It is obvious that the updated decoder would produce a possibly better set of hypotheses for the verification stage by the MVE-trained verification models. In section IV, we conduct a comparison between the conventional UV framework and this new adaptive UV framework. Remarkable performance enhancement by the proposed framework has been obtained compared to the conventional framework.

III. MVE Training

The MVE training method can be viewed as a special version of the MCE method [14], [17], [18] for detection and verification problems. Similar to the MCE criterion, the objective of the MVE training is to directly minimize the empirical average loss. In contrast to the conventional string-based MVE [9], [10], here we will derive the segment-based MVE [11], [19], [20]. We note that the string-based MVE was initially designed to minimize the empirical average loss in the given strings when a pair of detectors is used as a recognizer. Hence, it still focuses on minimizing the recognition errors rather than the verification errors. Alternatively, the segment-based MVE directly minimizes the total verification errors as the weighted sum of type I and type II errors not in the given strings but in the given segments. An obvious advantage of the segment-based MVE is that the intrinsic properties of the speech signal, which is based on segments during the recognition and the verification, can be directly embedded into the training phase, and accordingly, the total verification errors latent in every given segments are efficiently minimized. In this section, we will review the theoretical framework of the segment-based MVE.

Suppose there are M classes and K training tokens (segments) in a training set. For a given training set $\{O_1, O_2, \ldots, O_k\}$, the empirical average loss is defined by

$$L(\tilde{\Theta}) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{M} l_{\text{total}}(O_k \,|\Theta^i)\,1(O_k \in \text{class } i), \tag{2}$$

where $1(\cdot)$ is an indicator function that returns 1 when the given token O_k belongs to the certain class i among the total M classes, and 0 otherwise, and $l_{\text{total}}(O_k \,|\Theta^i)$ is the composite error estimation function which combines two different kinds of verification errors: type I error (miss) and type II error (false alarm). The composite error estimation function can be described as

$$l_{\text{total}}(O_k \,|\Theta^i) = PW_{\text{I}} l_{\text{I}}(O_k \,|\Theta^i) + PW_{\text{II}} \sum_{j=1, j\neq i}^{M} l_{\text{II}}(O_k \,|\Theta^j), \tag{3}$$

where PW_{I} and PW_{II} are the penalty weights for type I and type II errors, respectively, and l_{I} and l_{II} are smoothed loss functions to approximate the empirical verification error on each training sample O_k defined as

$$l_{\text{I}}(O_k \,|\Theta^i) = \frac{1}{1+\exp\{-\alpha d_{\text{I}}(O_k \,|\Theta^i)\}}, \tag{4}$$

and

$$l_{\text{II}}(O_k \,|\Theta^j) = \frac{1}{1+\exp\{-\alpha d_{\text{II}}(O_k \,|\Theta^j)\}}, \quad j = 1,2,\ldots,M, \quad j \neq i, \tag{5}$$

where α is a constant which controls the slope of the smoothing function, and $d(O_k)$ is the misclassification measure for the two types of detection errors. The two misclassification measures for each incoming training token O_k labeled as the i-th class event can be formulated as

$$d_{\text{I}}(O_k \,|\, \Theta^i) = -g_t^i(O_k \,|\, \Theta_t^i) + g_a^i(O_k \,|\, \Theta_a^i), \tag{6}$$

and

$$d_{\text{II}}(O_k \,|\, \Theta^j) = +g_t^j(O_k \,|\, \Theta_t^j) - g_a^j(O_k \,|\, \Theta_a^j), \quad j = 1,2,\ldots,M, \quad j \neq i, \tag{7}$$

where d_{I} and d_{II} are the type I and type II misclassification measures, respectively. In (6) and (7), g_t^i and g_a^i are the normalized log likelihood, and Θ_t^i and Θ_a^i are the parameter set of the target and the anti-model for the i-th class, respectively. In hidden Markov models (HMMs) [13], $g^i(O;\Theta^i)$ can be described as the maximum log likelihood of the state sequence obtained by Viterbi alignment [13]. For example, a set of the class discriminant functions $g^i(O;\Theta^i)$, $i = 1,2,\ldots,M$, can be expressed by

$$\begin{aligned} g^i(O;\Theta^i) &= P(O \,|\, \Theta^i) = P\left(O, \boldsymbol{q} \,|\, \pi^i, a^i, \{b_j^i\}_{j=1}^{N}\right) \\ &= \pi_{q_0}^i \prod_{t=1}^{T} a_{q_{t-1}q_t}^i b_{q_t}^i(o_t), \end{aligned} \tag{8}$$

where $\boldsymbol{q}=(q_1, q_2, \ldots, q_T)$ is any state sequence being generated by the Markov chain, and the parameter set Θ^i is associated with an initial state probability π, a state transition probability a,

and state observation distribution b_j, j=1, 2,…, N states. In this paper, we choose the maximum joint observation-state probability for the discriminant function $g^i(\mathrm{O};\Theta^i)$ such that

$$g^i(\mathrm{O};\Theta^i) = \log\{\max_q g^i(\mathrm{O},q;\Theta^i)\} = \log\{g^i(\mathrm{O},\bar{\boldsymbol{q}};\Theta^i)\} = \sum_{t=1}^{T}[\log a^i_{\bar{q}_{t-1}\bar{q}_t} + \log b^i_{\bar{q}_t}(o_t)] + \log \pi^i_{\bar{q}_0}, \quad (9)$$

where $\bar{\boldsymbol{q}} = (\bar{q}_1, \bar{q}_2, ..., \bar{q}_T)$ is the optimal state sequence. In addition, the output likelihood $b^i_{q_t}(o_t)$ of the K-mixture Gaussian can be defined by

$$b^i_{q_t}(o_t) = \sum_{k=1}^{K} c^i_{jk}\mathcal{N}\left[o_t \middle| \mu^i_{jk}, R^i_{jk}\right] = \sum_{k=1}^{K} \frac{c^i_{jk}}{(2\pi)^{D/2}\left|R^i_{jk}\right|^{1/2}} \exp\left[-\frac{1}{2}\sum_{l=1}^{D}\frac{(o_{tl}-\mu^i_{jkl})^2}{(\sigma^i_{jkl})^2}\right], \quad (10)$$

where D is the dimension of o_t, and c^i_{jk}, μ^i_{jk}, and R^i_{jk} are the mixture weight, the mean vector, and the covariance matrix of the k-th mixture component in the j-th state for the i-th HMM model, respectively.

Finally, according to an iterative procedure with the given training data, all the parameters in Θ_t and Θ_a follow the update rule of the GPD algorithm [14], [17], [18] when minimizing (2) as defined by

$$\Theta_{k+1} = \Theta_k - \varepsilon_k \nabla l_{\text{total}}(\mathrm{O}_k \mid \Theta)\big|_{\Theta=\Theta_k}, \quad (11)$$

where ε_k is a learning rate, and k is the cumulative number of the processed training samples at time t. In our implementation, the optimization algorithm above is operated on a sample-by-sample basis update of four kinds of the parameters, $\Theta^i = \{\mu^i_{jkl}, \sigma^i_{jkl}, c^i_{jk}, a^i_{ij}\}$. For brevity, here we only derive the updating process for the mean vector in the parameter set. The discriminative adjustment of the mean vector in the target model parameter set Θ^i_t follows:

$$\tilde{\mu}^i_{jkl}(n+1) = \tilde{\mu}^i_{jkl}(n) - \varepsilon \frac{\partial l_{\text{total}}(\mathrm{O}_n \mid \Theta)}{\partial \tilde{\mu}^i_{jkl}}\bigg|_{\Theta=\Theta^i_{t_n}}, \quad (12)$$

where $\tilde{\mu}^i_{jkl} = \mu^i_{jkl} / \sigma^i_{jkl}$ satisfying the internal constraints [13], [14] in the HMMs. If $\mathrm{O}_n \in$ class i, then the partial derivative part in (12) is expressed in detail as follows:

$$\frac{\partial l_1(\mathrm{O}_n \mid \Theta^i)}{\partial \tilde{\mu}^i_{jkl}} = \alpha l_1(\mathrm{O}_n \mid \Theta^i)\left(1 - l_1(\mathrm{O}_n \mid \Theta^i)\right) \cdot \left(-\frac{\partial g^i_t(\mathrm{O}_n \mid \Theta^i_t)}{\partial \tilde{\mu}^i_{jkl}} + \frac{\partial g^i_a(\mathrm{O}_n \mid \Theta^i_a)}{\partial \tilde{\mu}^i_{jkl}}\right). \quad (13)$$

In (13), the mean vector $\tilde{\mu}^i_{jkl}$ is associated only with the output likelihood functions, and the gradient of $g^i_t(\mathrm{O}_n \mid \Theta^i)$ is therefore written as

$$\frac{\partial g^i_t(\mathrm{O} \mid \Theta^i)}{\partial \tilde{\mu}^i_{jkl}} = \sum_{t=1}^{T}\delta(\bar{q}_t - j)\frac{\partial \log b^i_j(o_t)}{\partial \tilde{\mu}^i_{jkl}}, \quad (14)$$

and

$$\frac{\partial \log b^i_j(o_t)}{\partial \tilde{\mu}^i_{jkl}} = \frac{c^i_{jk}}{(2\pi)^{\frac{D}{2}}\left|R^i_{jk}\right|^{\frac{1}{2}} b^i_j(o_t)}\left(\frac{o_{tl}}{\sigma^i_{jkl}} - \tilde{\mu}^i_{jkl}\right) \cdot \exp\left[-\frac{1}{2}\sum_{l=1}^{D}\left(\frac{o_{tl}}{\sigma^i_{jkl}} - \tilde{\mu}^i_{jkl}\right)^2\right], \quad (15)$$

where $\delta(\cdot)$ is the knonecker delta function. The last step is to convert $\tilde{\mu}^i_{jkl}$ back according to

$$\mu^i_{jkl}(n+1) = \tilde{\mu}^i_{jkl}(n+1)\sigma^i_{jkl}(n). \quad (16)$$

Similarly, the derivations for the variance vectors, the mixture weights, and the transition probabilities can be easily accomplished [14], [18], [21].

IV. Experimental Setup and Results

1. Experimental Setup

All of our experiments were conducted on distance-talking and noisy speech databases collected under four different remote talking conditions: 30 cm, 60 cm, 100 cm, and 150 cm corresponding to the distance between a talker and the microphone. Recordings were performed in a room with a realistic level of noise in a home-noise environment. In particular, the background noise components consisted of normal sounds of a refrigerator, television, audio playback, and people's conversations.

In all evaluation sets, the number of keywords and OOV words are chosen to be identical. Each of the distance-talking and noisy speech databases is comprised of 1,470 utterances recorded by 49 speakers, 30 utterances per speaker. Each utterance consists of an isolated word such as a command or point of interest for a voice control application of an in-car navigation system. For the keyword detection and OOV word rejection experiments, we set 130 keywords and 50 OOV words in the 1,470 utterances. Among the 1,470 utterances, 1,113 (75.71%) of them contain 130 keywords considered as legitimate inputs, and the other 357 (24.29%) contain 50 OOV words considered invalid inputs to be rejected by the system. In the recognition stage, a simple keyword-loop network with 130 keywords and a silence model is used in the decoding as shown in Fig. 3 (no language model used). For the isolated keyword recognition, the Viterbi algorithm [13] in the decoding is employed to find the most likely keyword through the keyword-loop network for each given observation.

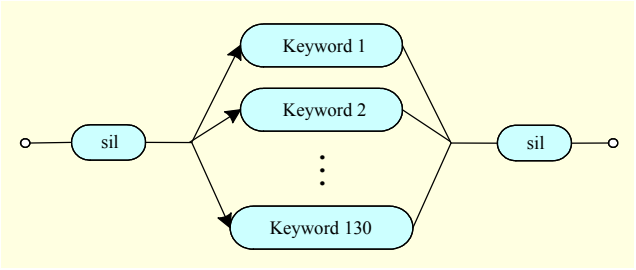

Fig. 3. Keyword-loop network in the decoding.

The feature extractor of the recognition and verification systems computes 39 components consisting of 12 mel-frequency cepstral coefficients plus normalized log energy and their first- and second-order time derivatives. For the baseline recognition models, a set of Korean 45 CI monophone acoustic models were used. All models are represented by 3-state strict left-to-right HMMs with 16 Gaussian mixture components per state. All the experiments in the training and testing were conducted on the abovementioned model description.

2. Training Process

For the baseline recognition models and verification models, a large vocabulary speech corpus consisting of 1,700,000 phone optimized word utterances, 40,000 sentence-based utterances, and 160,000 distant talking utterances were used for the initial maximum likelihood (ML) trained models. Then, we refined the ML-trained models with 1,500 utterances, which are a part of the data used for training the ML models, using the conventional MVE method. The refined-MVE models have been used for all adaptation experiments as the baseline models.

In the adaptation training side, we trained the baseline models based on the two MVE training scenarios: The first scenario is performed on the conventional MVE training under the two-stage conventional UV framework which does not update the transcription for the MVE training and produces the fixed recognition hypotheses carried out by the baseline recognition model. The second scenario is performed by the A-MVE training under the proposed framework, the adaptive UV framework, with the improved transcriptions sequentially updated for the next MVE training and re-decoded hypotheses produced by the MVE-trained target model so that a small portion of incorrectly recognized hypotheses is obtained, and improved knowledge sources on the hypotheses, such as segmentation and duration, efficiently affect the verification stage. Both are trained with 490 utterances (one third of the total 1,470 testing utterances) randomly chosen in the keyword utterances at each iteration. Since the experiments in this paper are intended for a speaker-independent UV system, we use the randomly chosen keyword utterances regardless of speakers at every iteration. Then, the discriminative training procedure is performed over 10 iterations. As previously discussed, at each iteration, the label information on the transcription is realigned by the current-stage MVE target model. Also, the updated label information is used for the next discriminative training stage. During the training process over 10 iterations, the discriminative adjustment of the parameter set follows the GPD algorithm as shown in section III. In terms of the parameter optimization by the GPD algorithm over the iterations, an important issue is how to design the learning rate ε_k in (11) and (12). We will discuss in detail this issue as well as the convergence issue in subsection IV.4.

3. Word-Level Hypothesis Testing

After the isolated keyword recognition, for the hypothesis testing of the given recognized word, we first consider subword-level (monophone-level) acoustic verification scores based on the following equation modified from (1):

$$LR_p = \frac{p(\mathrm{O}\,|\,H_0)}{p(\mathrm{O}\,|\,H_1)} = \frac{p(\mathrm{O}\,|\,\Theta_t^p)}{p(\mathrm{O}\,|\,\Theta_a^p)}, \tag{17}$$

where O is the speech segment of the word w, and Θ_t^p and Θ_a^p are the corresponding target subword and anti-subword models for subword p, respectively. By taking logarithm of (17), the log likelihood ratio LLR_p for the subword p can be expressed as

$$LLR_p = \log p(\mathrm{O}\,|\,\Theta_t^p) - \log p(\mathrm{O}\,|\,\Theta_a^p). \tag{18}$$

The word level confidence score CM_w is then defined by

$$CM_w = \frac{1}{N}\sum_n LLR_n, \tag{19}$$

where N is the total number of subwords in the word w. The confidence measure (CM) 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. In our experiments, all the UV results for the hypotheses are based on the CM in (19). We note that instead of the simple CM above, one can use enhanced CMs [22]-[24] using more knowledge sources and particular rule-based integration of the knowledge sources on the hypotheses. Better CMs may directly improve the verification performance and may have to be considered for the task containing extremely many keywords and OOV words. In this paper, we only focus on discriminative parameter separation and optimization using the MVE training with a part of the given keyword utterances to increase keyword detection rate and OOV rejection rate (REJ) simultaneously. One can further extend the proposed framework by associating the

enhanced CMs.

We then count the number of errors in keywords and OOV words and present the total word error rate (WER) and REJ of the OOV words, respectively, at the certain false REJ, for example, 7% false REJ and 15% false REJ. In practice, one usually has to pick a specific false acceptance or false REJ as part of the operating specifications on the system [25]. It is more desirable to optimize UV 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.

4. GPD-Based Optimization

As discussed, an important issue in the GPD-based optimization over the iterations is how to design the learning rate ε_k. Moreover, the hyper-parameter α of the sigmoid function is also related to the training performance and convergence. In this subsection, we focus on discussing a number of design techniques of these two parameters since the GPD algorithm requires the parameters to be properly set to converge [26].

The parameter α controls the slope of the sigmoid function and thus determines how the punishment is when an error occurs in the misclassification measures (6) and (7). For example, from (4) and (5), it can be shown that the sigmoid function curve varies with different values of α, and the curve becomes sharper as α increases. It means that a large value of α will make the convergence speed of the training process faster, but it may cause the over-fitting of the parameters. On the contrary, a small value of α may lead to a slow convergence. The conventional setup of the value of α is that it is fixed during the training with α>0.5. In the following all experiments, we set the value of α to 1.0 as a fixed global parameter.

Similarly, the learning rate ε_k directly affects the discriminative adjustment of the parameter set. If the learning rate is too high, the parameter may be overstrained at the beginning, and thus the performance may degrade seriously. Alternatively, if the learning rate is too low, the parameter may be tuned little by little, and thus the convergence may be too slow. This property of the GPD-based optimization is well known, and a heuristic method to set the learning rate is commonly used. In our implementation, we apply various different values of the learning rate from 1.5 to 0.0005 to the individual MVE training at every iteration. Then, we choose the best model and proceed the next MVE training with the above various different values of the learning rate again. Therefore, there is no performance degradation over the iterations. In the following experiments, this procedure is repeated for 10 iterations. Figures 4 and 5 show the change of WER and OOV REJ by percent with increasing number of iterations. Note that both are measured at 7% false REJ. It is shown that there is no performance degradation iteration-by-iteration as we discussed above, and most of performance gains have been achieved largely in the first three iterations in both WER and OOV REJ. In addition, after 8 iterations, the performance change curves in both Figs. 4 and 5 are flat. It tells us that the GPD-based optimization converges to a local minimum of the empirical loss within the given training data after 8 iterations.

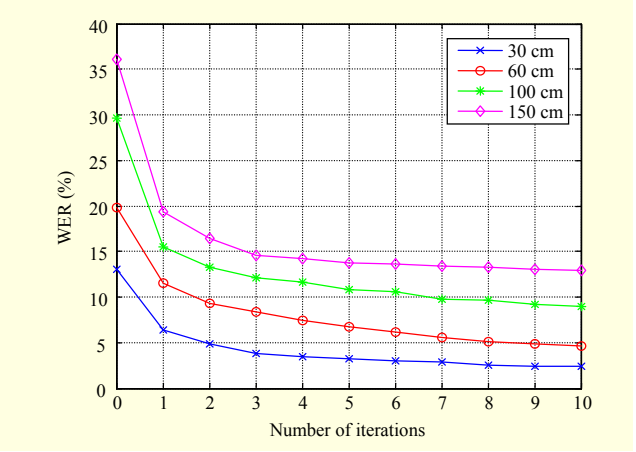

Fig. 4. Change of WER by percent at 7% false REJ over 10 iterations for four different databases.

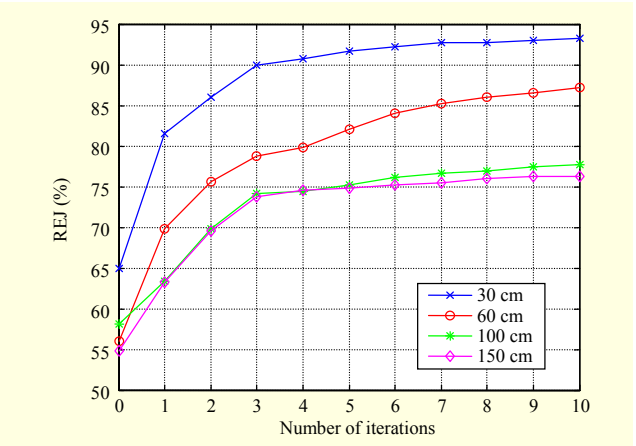

Fig. 5. Change of OOV REJ by percent at 7% false REJ over 10 iterations for four different databases.

5. Detailed Results on Four Different Databases

In this subsection, we will analyze the detailed overall performance between the proposed framework and the conventional framework with respect to each of the four

different databases. In particular, all the performance metrics are measured at specific iteration point 3. This is because most of the performance gains have been achieved largely in first three iterations, as shown in Figs. 4 and 5. Furthermore, in real applications, if there are too many iterations, the performance becomes labor-intensive and time-consuming.

A. 30 cm Database

The first testing set among the four different distance-talking and noisy speech databases is the "30 cm database" with 30 cm distance in terms of the remote talking condition between a talker and the microphone. An overall performance comparison of the three different methods, baseline, conventional MVE, A-MVE, respectively, is presented in Table 1.

From the second row in the table, with no rejection (that is, REJ=0.0%), the initial WER of 29.05% is observed by the baseline model. On the other hand, with the verification, the WER is reduced to 13.09% at 7% false REJ and 8.66% at 15% false REJ. Furthermore, the REJ of the OOV words is 64.99% at 7% false REJ and 79.27% at 15% false REJ, respectively, after the verification.

The third row (MVE) shows the overall performance by the conventional MVE-trained model and under the conventional UV framework. With the verification, the WER drops to 4.01%, and the OOV REJ is increased to 90.48% at 15% false REJ. Although the MVE method under the conventional UV framework produces substantial word error reduction rate and improved OOV REJ compared to the baseline performance, the proposed method, the A-MVE under the adaptive UV framework, confirms that considerable additional gains of performance can be achieved all over the performance metrics: In particular, the WER has been reduced to 3.77% and 1.48% at 7% false REJ and 15% false REJ, respectively. In addition, with respect to the OOV REJ, more remarkable performance improvement is observed. The OOV REJ of 89.92% and96.08% is achieved by the A-MVE method at 7% false REJ and 15% false REJ, respectively. Finally, we present EER

Table 1. Overall performance comparison on 30 cm database.

	WER at 0% rejection (%)	WER/OOV REJ at 7% false rejection (%)	WER/OOV REJ at 15% false rejection (%)	EER (%)
Baseline	29.05	13.09/64.99	8.66/79.27	17.08
MVE	29.05	7.98/78.99	4.01/90.48	12.49
A-MVE	**25.37**	**3.77/89.92**	**1.48/96.08**	**8.26**

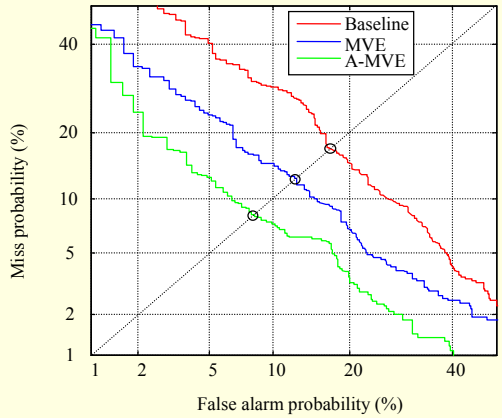

Fig. 6. DET curves of three different methods on 30 cm database. The circles on the diagonal line are EER points.

performance in the rightmost column of the Table 1. It can be shown that the EER of the A-MVE is significantly reduced compared to the baseline as well as the MVE. For details, Fig. 6 shows detection error tradeoff (DET) curves [27] of the three different methods on the 30 cm database.

As a result, the A-MVE method reduces the WER without the verification and also provides benefits with the verification by producing improved knowledge such as segmentation for the hypothesis testing. All experimental results confirm that with the verification by the A-MVE, the WER is remarkably reduced, and a substantial improvement of the OOV REJ is achieved simultaneously.

B. 60 cm and 100 cm Databases

The second and third testing sets are the "60 cm database" and "100 cm database" with a longer recording distance than the 30 cm database. As we have observed in the 30 cm database, the A-MVE significantly reduces the WERs, both without the verification and with the verification, and notably improves the verification performance, the OOV REJ, and the EER on both databases. The details of the performance comparison on these databases are presented in Tables 2 and 3, respectively.

In particular, at 7% false REJ on 60 cm database and 100 cm database after 3 iterations, the proposed framework using the A-MVE training reduces the WER by further 6.86% and 7.70% and simultaneously increases the OOV REJ by further 12.88% and 3.38%, respectively, over the conventional framework using the MVE training.

From the experimental results on the 30 cm, 60 cm, and 100 cm databases, it is clear that under the proposed adaptive UV framework, the two types of false alarms (misrecognized keywords and OOVs) are minimized while the detection of

Table 2. Overall performance comparison on 60 cm database.

	WER at 0% rejection (%)	WER/OOV REJ at 7% false rejection (%)	WER/OOV REJ at 15% false rejection (%)	EER (%)
Baseline	37.69	19.79/56.02	14.88/70.87	19.03
MVE	37.69	15.24/65.83	9.64/81.79	14.97
A-MVE	**28.11**	**8.38/78.71**	**3.74/91.88**	**12.22**

Table 3. Overall performance comparison on 100 cm database.

	WER at 0% rejection (%)	WER/OOV REJ at 7% false rejection (%)	WER/OOV REJ at 15% false rejection (%)	EER (%)
Baseline	52.15	29.67/58.03	22.60/71.55	18.68
MVE	52.15	19.90/70.70	13.62/80.28	13.56
A-MVE	**33.06**	**12.13/74.08**	**5.44/89.58**	**12.60**

correctly recognized keywords is maximized.

C. 150 cm Database

The last testing set is the "150 cm database" with the longest distance between a talker and the microphone among all the databases (Table 4). However, it contains as much background noise as the others. Thus, the baseline performance is seriously degraded from 29.05% to 59.71% in terms of the WER compared to the 30 cm database. Even with the verification, the performance is limited to the WER of 26.82% and the OOV REJ of 74.01% at a 15% false REJ. With the verification by the conventional MVE, the WER drops from 26.82% to 15.91% and the OOV REJ improves from 74.01% to 83.90%, whereas by the A-MVE method, the WER rapidly drops from 59.71% to 39.06%, even without the verification. Furthermore, with the verification by the A-MVE, the WER is reduced to 7.66%, and the OOV REJ is increased to 88.14% at a 15% false REJ. These results reconfirm that the A-MVE method significantly reduces the WER and, at the same time, effectively enhances the function of the OOV rejection. We note that the EER in the A-MVE is slightly increased compared to the conventional MVE. The reason is that all information including label identities and their corresponding segments on the re-recognized hypotheses dramatically change. It means that the wide-ranging variation of the knowledge sources of the updated hypotheses may affect all the ratios of each subword and their average word-level confidence scores. However, as discussed, EER itself cannot be a conclusive measurement of the performance metrics for verifying the ASR system. To precisely measure the overall UV performance, we focus on the WER and the OOV REJ with the verification.

Table 4. Overall performance comparison on 150 cm database.

	WER at 0% rejection (%)	WER/OOV REJ at 7% false rejection (%)	WER/OOV REJ at 15% false rejection (%)	EER (%)
Baseline	59.71	36.09/54.80	26.82/74.01	17.79
MVE	59.71	23.00/73.16	15.91/83.90	12.88
A-MVE	**39.06**	**14.59/73.73**	**7.66/88.14**	**13.09**

V. Conclusion

In this paper, we have investigated the adaptive and integrated UV framework using the minimum verification error (MVE) training as a new set of solutions to the entire UV system in real applications. First, in order to mitigate serious performance degradation due to mismatched operating conditions in the real applications, in contrast to the conventional UV 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 target model at every iteration during the discriminative training of the verification models. Furthermore, in the context of the conventional UV, the recognized hypotheses do not change regardless of the UV models. 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 hypothesis testing. Consequently, for the entire UV system, the proposed framework can be considered as one integrated stage associated with only the verification models (the MVE target and anti-models) in contrast to the conventional rigid two stages associated with the inconsistent recognition model and verification models.

Throughout the proposed adaptive and integrated UV framework with the segment-based MVE training, we simultaneously obtained an improved overall system decoder with a much reduced recognition error rate and discriminatively trained verification models which significantly enhance the entire verification performance in such real application scenarios. All experimental results confirm that the proposed framework produces remarkable performance gains in both the recognition and verification. In particular, with the verification at both 7% and 15% false REJs, the WER was considerably reduced, and a substantial improvement of the OOV REJ was also achieved all over the distance-talking and noisy speech databases. The proposed framework shows

promise in user-friendly interface systems, such as in-car navigation and cell phones, and when aiming for proper detection of keywords and high rejection of OOV words in real-world applications.

References

[1] M. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative Utterance Verification for Connected Digits Recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, May 1997, pp. 266-277.

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

[3] R.A. Sukkar, A.R. Setlur, and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 420-429, Nov. 1996.

[4] E.L. Lehmann, *Testing Statistical Hypotheses*, John Wiley & Sons, 1959.

[5] S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, NJ: Prentice-Hall, Englewood Cliffs, 1998.

[6] G. Casella and R.L. Berger, *Statistical Inference*, Duxbury Press, New York, 2001.

[7] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, 1995, pp. 171-185.

[8] J. Wu and Q. Huo, "A Study of Minimum Classification Error (MCE) Linear Regression for Supervised Adaptation of MCE-Trained Continuous-Density Hidden Markov Models," *IEEE Trans. Speech Audio Process.*, vol. 15, 2007, pp. 478-488.

[9] M. Rahim and C.-H. Lee, "String-Based Minimum Verification Error (sb-mve) Training for Speech Recognition," *Computer Speech and Language*, vol. 11, 1997, pp. 147-160.

[10] A.E. Rosenberg, O. Siohan, and S. Parthasarathy, "Speaker Verification Using Minimum Verification Error Training," *ICASSP*, 1998, pp. 105-108.

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

[12] M.-W. Koo, C.-H. Lee, and B.-H. Juang, "Speech Recognition and Utterance Verification Based on a Generalized Confidence Score," *IEEE Trans. Speech Audio Process.*, vol. 9, Nov. 2001, pp. 821-832.

[13] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[14] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, May 1997, pp. 257-265.

[15] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition," PhD thesis, Cambridge University, 2004.

[16] X. He, L. Deng, and W. Chou, "Discriminative Learning in Sequential Pattern Recognition: A Unifying Review for Optimization-Oriented Speech Recognition," *IEEE Signal Process. Mag.*, vol. 25, Sept. 2008, pp. 14-36.

[17] B.-H. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Process.*, vol. 40, Dec. 1992, pp. 3043-3054.

[18] W. Chou, C.-H. Lee, and B.-H. Juang, "Segmental GPD Training of HMM Based Speech Recognizer," *ICASSP*, Apr., 1992, pp. 473-476.

[19] Q. Fu and B.-H. Juang, "A Study on Rescoring Using HMM-Based Detectors for Continuous Speech Recognition," *ASRU*, Kyoto, Japan, Dec. 2007, pp. 570-575.

[20] S. Shin et al., "Discriminative Linear-Transform Based Adaptation Using Minimum Verification Error," *ICASSP*, Texas, USA, Mar. 2010, pp. 4318-4321.

[21] W. Chou, "Minimum Classification Error Approach in Pattern Recognition," *Pattern Recognition in Speech and Language Processing*, W. Chou and B.-H. Juang, Eds., Boca Raton: CRC Press, 2003, pp. 1-49.

[22] F. Wessel et al., "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. Speech Audio Proc.*, vol. 9, no. 3, Mar. 2001, pp. 288-298.

[23] T. Hazen and I. Bazzi, "A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring," *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Salt Lake City, Utah, May 2001.

[24] F.K. Soong, W.K. Lo, and S. Nakamura, "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words," *Proc. SWIM*, 2004.

[25] M.-H Siu, B. Mak, and W.-H. Au, "Minimization of Utterance Verification Error Rate as a Constrained Optimization Problem," *IEEE Signal Process., Letters*, vol. 13, Dec. 2006, pp. 760-763.

[26] J.A. Snyman, *Practical Mathematical Optimization*, New York: Springer, 2005.

[27] A. Martin et al., "The DET Curve in Assessment of Detection Task Performance," *Proc. European Conf. Speech Commun. Technol.*, 1997, pp. 1895-1898.

Sung-Hwan Shin received the BS in information engineering from Myong-Ji University, Rep. of Korea, in 2007, and the MS in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2009. He is currently pursuing the PhD at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. His research interests include speech recognition, utterance verification, discriminative training algorithms, and statistical signal processing.

Ho-Young Jung received the BS in electronics engineering from Kyungpook National University, Daegu, Rep. of Korea, in 1993, and the MS and PhD in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Rep. of Korea, in 1995 and 1999, respectively. His PhD dissertation was on robust speech recognition. He joined ETRI, Daejeon, Rep. of Korea, in 1999 as a senior researcher and has been with the Speech/Language Information Research Center from 2002. Since 2010, he has been a principle member of the research staff. His current research interests include speech recognition, noise-robust processing, blind signal separation, and machine learning. He has published or presented about 30 papers in speech processing.

Biing-Hwang Juang received his PhD from the University of California, Santa Barbara. He joined Speech Communications Research Laboratory (SCRL) in 1978 and Signal Technology, Inc. (STI) in 1979, working on a number of government-sponsored research projects. He was also a co-principal investigator for the project on cochannel separation of speech signals sponsored by the US Government. He subsequently joined Bell Laboratories in 1982, working in the area of speech enhancement, coding and recognition. Prof. Juang later became the director of Acoustics and Speech Research at Bell Labs, and at the turn of the century, the director of Multimedia Technologies Research at Avaya Labs (a spin-off of Bell Labs). Prof. Juang has published extensively, including the book "Fundamentals of Speech Recognition," coauthored with L.R. Rabiner, and holds about twenty patents. He has served as editor-in-chief for the IEEE Transactions on Speech and Audio Processing, and a number of positions in the IEEE Signal Processing Society, including chair of its Fellow Evaluation Committee. Prof. Juang has received a number of technical awards, notable among which are several best paper awards in the area of speech communications and processing, the Technical Achievement Award from the Signal Processing Society of the IEEE, and the IEEE Third Millennium Medal. He is a fellow of the IEEE, a fellow of Bell Laboratories, a member of the US National Academy of Engineering, and an Academician of Academia Sinica. Prof. Juang joined Gatech in 2002 holding the Motorola Foundation Chair Professorship and is an eminent scholar of Georgia Research Alliance.