CORRECTING RECOGNITION ERRORS VIA DISCRIMINATIVE UTTERANCE VERIFICATION

Anand R. Setlur, Rafid A. Sukkar, and John Jacob

Lucent Technologies
2000 N. Naperville Rd., Naperville, IL 60566, USA
e-mail : anand.setlur@lucent.com

ABSTRACT

Utterance verification (UV) is a process by which the output of a speech recognizer is verified to determine if the input speech actually includes the recognized keyword(s). The output of the speech verifier is a binary decision to accept or reject the recognized utterance based on a UV confidence score. In this paper, we extend the notion of utterance verification to not only detect errors but also selectively correct them. We perform error correction by flipping the hypotheses produced by an N-best recognizer in cases when the top candidate has a UV confidence score that is lower than that of the next candidate. We propose two measures for computing confidence scores and investigate the use of a hybrid confidence measure that combines the two measures into a single score. Using this hybrid confidence measure and an N-best algorithm, we obtained an 11% improvement in word-error rate on a connected digit recognition task. This improvement was achieved while still maintaining reliable detection of non-keyword speech and misrecognitions.

1. INTRODUCTION

In many speech recognition applications, the N-best algorithm has been employed to produce multiple recognition hypotheses, resulting in improved performance. Our experience with connected digit recognition indicates that a majority of the recognition errors occur in a single digit position. Usually, the correct hypothesis is found within the top two or three candidates in an N-best search. The word error rate can typically be halved if the correct hypothesis can be picked via manual intervention from a list of the top two or three hypotheses. In applications, such as account number or credit card number recognition, it is perfectly acceptable to have a recognizer that has a good accuracy within the top N candidates since other knowledge sources such as a zip code are available to cross-validate the recognition hypotheses. On the other hand, in applications such as telephone number dialing, it may be useless to get a string correct in a subsequent hypothesis if the top hypothesis is wrong. Utterance verification techniques have been successfully employed to identify and reject strings that are likely to be in error, thereby improving overall post-rejection accuracy [1-4].

In this paper, we will extend the notion of utterance verification and utilize it to correct recognition errors. This is done by replacing the recognized string with the next best hypothesis when the verification algorithm produces a higher confidence score for the second candidate compared to the top one. This way, recognition accuracy is improved by correcting errors when possible, thereby rejecting fewer utterances, instead of rejecting all questionable utterances outright.

The organization of the paper is as follows. In the next section, we introduce two distinct measures that can be used as confidence scores for recognition hypotheses. We then propose to combine the two measures into a single confidence score that can be used to detect and subsequently correct errors. We then evaluate the proposed approaches using some experimental data.

2. HYPOTHESIS TESTING

We discuss two hypothesis testing measures based on likelihood ratio tests. One approach uses HMM models that are separate from the acoustic models used during recognition. These verification-specific models are discriminatively trained using the Minimum Verification Error (MVE) training method. The other approach does not require additional models.

2.1 MVE-Trained Likelihood Ratio

In this method, utterance verification is done as a post-processing step after the N recognition hypotheses and the associated word segmentations are available from the recognition phase. The problem is formulated at the word-level and subsequently extended to the string-level. It comprises a likelihood ratio test that is a function of two models (null and alternate hypothesis) which are both trained using a discriminative minimum verification error training (MVE) framework [1]. The null hypothesis is that the recognized word is correct while the alternate hypothesis is that the word is misrecognized. The alternate hypothesis covers two equally important categories : non-keyword speech and keyword speech that is misreconized by the recognizer.

We first define a word-based likelihood ratio statistic whose probability distribution parameters are determined

discriminatively. Let $S = w_{q(1)}\ w_{q(2)}\ w_{q(3)}\ \cdots\ w_{q(N)}$ be a keyword string hypothesis of length N produced by a Hidden Markov Model (HMM) recognizer with a vocabulary set of $\{w_k\}$, $1 \le k \le K$. The function $q(n)$, $1 \le n \le N$, maps the word number in the string sequence S to the index of the word in the vocabulary set. Let $\mathbf{O}_n$ be the observation vector sequence corresponding to the speech segment for word $w_{q(n)}$ in S, as determined by the HMM segmentation. The *word likelihood ratio* is written as,

$$WLR^{(mve)}(\mathbf{O}_n; w_{q(n)}) = \frac{L[\mathbf{O}_n | H_0(w_{q(n)})]}{L[\mathbf{O}_n | H_1(w_{q(n)})]}, \qquad (1)$$

where $H_0(w_{q(n)})$ and $H_1(w_{q(n)})$ are the null and alternate hypotheses for verifying $w_{q(n)}$, respectively. We model the likelihoods in equation (1) using HMMs that are different than those used in the recognition process. Equation (1) is therefore rewritten as,

$$WLR^{(mve)}(\mathbf{O}_n; w_{q(n)}) = \frac{L[\mathbf{O}_n | \Lambda_{q(n)}]}{L[\mathbf{O}_n | \Psi_{q(n)}]}, \qquad (2)$$

where $\Lambda_{q(n)}$ and $\Psi_{q(n)}$ are the HMM model sets corresponding to the null and alternate hypotheses for word $w_{q(n)}$, respectively. The likelihood functions in equation (2) are considered to be normalized with respect to the duration of the word. In general, $\Lambda_{q(n)}$ and $\Psi_{q(n)}$ can each consist of one or more HMM models. In this study, $\Lambda_{q(n)}$ is represented by a single HMM model denoted by $\lambda_{q(n)}$. So,

$$L[\mathbf{O}_n | \Lambda_{q(n)}] = L[\mathbf{O}_n | \lambda_{q(n)}]. \qquad (3)$$

The definition of the alternate hypothesis model is motivated by our objective of reliably detecting both keyword misrecognitions as well as non-keyword speech. Accordingly, we define a *composite alternate hypothesis* model consisting of a set of two HMMs. Specifically, $\Psi_{q(n)} = \{\psi_{q(n)}, \phi_{q(n)}\}$, where $\psi_{q(n)}$ is an "anti-keyword model" modeling misrecognitions, and $\phi_{q(n)}$ is a filler model included to model non-keyword speech. The likelihoods of the anti-keyword and filler models are combined to result in the likelihood of the composite alternate hypothesis as follows:

$$L[\mathbf{O}_n | \Psi_{q(n)}] = \left[\frac{1}{2} \left[L[\mathbf{O}_n | \psi_{q(n)}]^{\kappa} + L[\mathbf{O}_n | \phi_{q(n)}]^{\kappa} \right] \right]^{1/\kappa}, \qquad (4)$$

where κ is a positive constant. We denote the verification-specific model set for a given keyword, $w_{q(n)}$, as $\mathcal{V}_{q(n)} = \{\lambda_{q(n)}, \psi_{q(n)}, \phi_{q(n)}\}$. Details of the discriminative training procedure used to train $\mathcal{V}_{q(n)}$ are given in [1].

Using the word likelihood ratio of equation (2), we define a *string-based likelihood ratio* as a geometric mean type function of the likelihood ratio of the words in the string, written as,

$$LR^{(mve)}(\mathbf{O}; S) = -\log \left[\frac{1}{N} \sum_{n=1}^{N} [WLR^{(mve)}(\mathbf{O}_n; w_{q(n)})]^{-\gamma} \right]^{1/\gamma}, \qquad (5)$$

where $\mathbf{O}$ is the observation sequence of the whole string and γ is a positive constant. The string likelihood ratio score, $LR^{(mve)}(\mathbf{O}; S)$, serves as a confidence measure and is compared to a threshold to make a verification decision. Defining the string likelihood score as given in equation (5) suggests that the keywords with low likelihood ratio scores tend to dominate the string score. This is a desirable property since this statistic will be used to isolate utterances likely to be recognized in error.

2.2 N-Best Based Likelihood Ratio

Another measure for hypothesis testing that has been used widely, is the duration normalized likelihood ratio or log likelihood difference between the recognition hypothesis and the next best hypothesis [5,6]. The confidence score for hypothesis k is the likelihood ratio of candidates k and k+1 and is represented by

$$LR^{(nbest)}(\mathbf{O}; k) = \frac{L[\mathbf{O} | h_k]}{L[\mathbf{O} | h_{k+1}]}, \qquad (6)$$

where h_k and h_{k+1} represent the k^{th} and $k+1^{th}$ hypotheses produced by an N-best algorithm and $\mathbf{O}$ is the observation vector sequence. Unlike the MVE method, this approach requires no additional models to perform utterance verification. This technique is effective especially if the recognition models are discriminatively trained. In this work, we used an N-best algorithm similar to the one described in [7].

2.3 Combined Confidence Score

The two likelihood ratio measures introduced in Sections 2.1 and 2.2 are derived from two distinct sets of HMMs. One uses a set of verification-specific HMMs, while the other reuses the HMMs used during the recognition phase. The natural question arises as to whether the two different likelihood ratio measures can be combined into a new hybrid confidence score that can outperform either measure when considered separately. This will be illustrated in the experimental results section. The hybrid score takes the form of a linear combination of $LR^{(mve)}$ and $LR^{(nbest)}$ and is written as

$$LR^{(hybrid)} = a\ LR^{(mve)} + b\ LR^{(nbest)}, \qquad (7)$$

where a and b are weighting factors determined discriminatively using Fisher's linear discriminant analysis as part of the training phase.

3. EXPERIMENTAL RESULTS

3.1 Recognition Task and Database

We chose a connected digit recognition task for performance evaluation purposes. The database used consisted of a training set of 16089 digits strings and a testing set of 21723 strings. The string lengths ranged from 1 to 16 digits with an average length of 5.5. This database represents speech collected form many different trials and data collection efforts over the U.S. telephone network. To evaluate "out-of-vocabulary" performance, we used a second speech database that did not have any digit strings. It consisted of 6666 phonetically balanced phrases and sentences, where 3796 phrases were used for training and the remaining for testing.

The recognizer feature vector consisted of the following 39 parameters: 12 LPC derived cepstral coefficients, 12 delta cepstral coefficients, 12 delta-delta cepstral coefficients, normalized log energy, and the delta and delta-delta of the energy parameter. The recognition digit model set was similar to the one used in [8] and consisted of continuous density, context dependent subword HMMs that were trained in a task-dependent mode. The training of these recognition models was based on minimum classification error training using the generalized probablistic descent discriminative training framework [8,9]. A string accuracy of 95.15% with a null grammar was achieved with these models. The corresponding word error rate was 1.15%.

3.2 Error Detection

In this section we will discuss the relative performance of the approaches described in Section 2. We will primarily focus on the ability to detect and reject questionable utterances outright and improve overall recognition accuracy on the strings that are not rejected. In the next section, we will illustrate the error correction aspect. For our study, the MVE model set, $\mathcal{V}_{q(n)}$, is represented by context independent models that are discriminatively trained. Each keyword model, $\lambda_{q(n)}$ and anti-keyword model, $\psi_{q(n)}$ in the model set, $\mathcal{V}_{q(n)}$, is represented by a 10 state, 8 mixture HMM. A total of 11 sets corresponding to the digits *0-9* and *oh* are trained. A common filler model was used to represent the filler model, $\phi_{q(n)}$, for all 11 keywords. Figures 1 through 3 compare the relative performance of the MVE method, the N-best method and the combined method. Figure 1 shows the string accuracy as a function of the string rejection rate. Another way of viewing the improvement in recognition accuracy is shown in Figure 2. This figure represents an ROC curve showing the false alarm rate of valid digit strings that are incorrectly recognized versus the false rejection rate of strings that are correctly recognized. For example from Figure 1 we see that, at an operating point of 5% rejection of valid digit strings, the MVE approach results in a 97.45% string accuracy compared to 97.84% for the N-best approach and 98.06% for the combined approach. While the error rate is better for the N-best approach compared to the MVE method for valid strings, it is worse on the non-keyword database as illustrated in Figure 3. This figure shows an ROC curve of the false alarm rate of non-keyword strings versus the false rejection rate of correctly recognized strings. By rejecting 5.0% of the correctly recognized digit strings, the MVE method and the combined method are able to reject more than 99% of the non-keyword strings, while with the N-best method, the performance is worse.

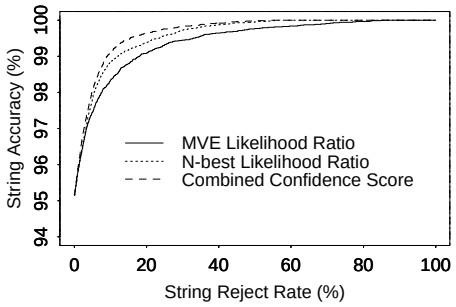

Figure 1. String accuracy versus rejection rate.

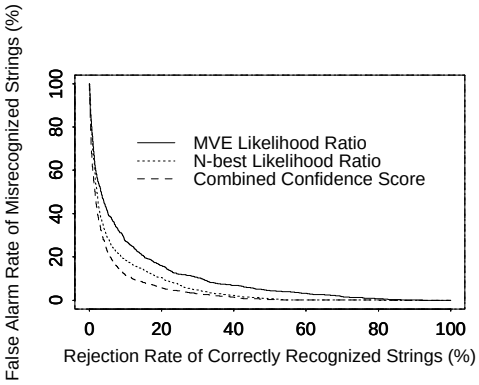

Figure 2. ROC curve for misrecognized valid strings.

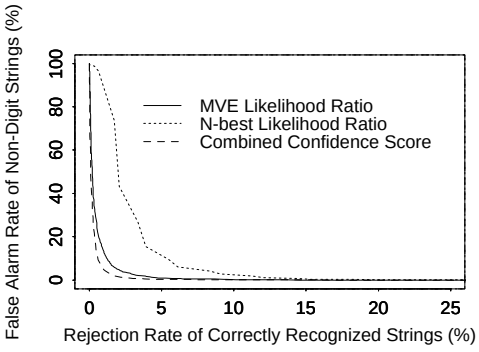

Figure 3. ROC curve for non-keyword strings.

Figure 4. Word error rate versus confidence score threshold.

From Figures 1-3, it is seen that the combined confidence score outperforms both the MVE and the N-best approaches under both scenarios.

3.3 Error Correction

Previously, utterances with low confidence scores were rejected outright. In these instances, we can reject fewer utterances and still improve recognition accuracy if we can selectively replace the top hypothesis with the next best hypothesis based on some criteria. Since it is crucial to not introduce too many new errors while trying to correct recognition errors, we restricted our space for error correction to only include substitution errors. The UV technique we propose lends itself best to this class of errors.

Figure 2 shows that we are able to reject 60% of all the errors at the cost of rejecting around 2% of the correctly recognized strings if we used the hybrid approach. Also from Figure 2 it can be seen that, we are able to reject nearly 100% of the errors if we use the hybrid approach and have the luxury of rejecting 40% of the correctly recognized strings. This means that all of the misrecognized strings have a hybrid confidence score below the median.

The criterion used for replacing the top candidate was as follows. If the string confidence score was higher than a threshold, it was deemed correct and not replaced. Only strings that had a string confidence score less than the threshold were considered for correction. The top hypothesis was replaced if the confidence score for the second-best hypothesis was larger than the threshold. Figure 4 shows a plot of the confidence score threshold versus the word error rate. At the threshold corresponding to the minimum word-error rate, we are able to lower the word-error rate by 11% from 1.15% to 1.02%. The corresponding string accuracy improved from from 95.15% to 95.70%. At higher thresholds, more errors are introduced than fixed.

4. CONCLUSIONS

We have proposed a UV method and used it to detect and possibly correct recognition errors. The UV confidence score is based on a linear combination of two distinct likelihood ratio scores. Using this combined confidence score, we were able to obtain a 11% improvement in word-error rate by selectively correcting errors detected by the UV method.

5. ACKNOWLEDGEMENTS

The authors would like to thank W. Chou and C. Mitchell for providing us with the N-best algorithm that we used for our experiments.

6. REFERENCES

[1] R. A. Sukkar, A. R. Setlur, M. G. Rahim, and C. H. Lee, "Utterance verification of keyword strings using word-based minimum verification error (WB-MVE) training," *Proc. ICASSP '96*, Vol. I, pp. 516-519, May 1996.

[2] M. G. Rahim, C. H. Lee, B. H. Juang, and W. Chou, "Discriminative utterance verification using minimum string verification error (MSVE) training," *Proc. ICASSP '96*, Vol. VI, pp. 3585-3589, May 1996.

[5] R. A. Sukkar and J. G. Wilpon, "A two pass classifier for utterance rejection in keyword spotting," *Proc. ICASSP '93*, Vol. I, pp. 451-454, May 1993.

[4] M. G. Rahim, C. H. Lee and B. H. Juang, "Robust utterance verification for connected digits recognition," *Proc. ICASSP '95*, Vol. I, pp. 285-288, May 1995.

[5] M. Weintraub, "LVSCSR log-likelihood ratio scoring for keyword spotting," *Proc. ICASSP '95*, Vol. I, pp. 297-300, May 1995.

[6] F. J. Caminero-Gil, C. de la Torre, L. A. Hernández-Gómez, and C. Martín-del-Álamo, "New N-best based rejection techniques for improving a real-time telephonic connected word recognition system," *Proc. Eurospeech '95*, Vol. III, pp. 2099-2102, Sept. 1995.

[7] F. K. Soong and E. Huang, "A tree-trellis based fast search for finding N-best sentence hypotheses in continuous speech recognition," *Proc. ICASSP '91*, Vol. I, pp. 705-708, May 1991.

[8] C. H. Lee, W. Chou, B. H. Juang, L. R. Rabiner and J. G. Wilpon, "Context-dependent acoustic modeling for connected digit recognition," *Proc. ASA '93.*

[9] W. Chou, B. H. Juang and C. H. Lee, "Segmental GPD training of HMM-based speech recognizer," *Proc. ICASSP '92*, Vol. I, pp. 473-476, Apr. 1992.