VERIFICATION OF MULTI-CLASS RECOGNITION DECISION USING CLASSIFICATION APPROACH

Tomoko Matsui[1], Frank K. Soong[2] and Biing–Hwang Juang [2]

[1]ATR Spoken Language Translation Research Labs, Kyoto, 619-0288 Japan
[2]Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, USA

ABSTRACT

We investigate various strategies to improve the utterance verification performance using a 2-class pattern classifier. They include utilizing N-best candidate scores, modifying segmentation boundaries, applying background and out-of-vocabulary filler models, incorporating contexts, and minimizing verification errors via discriminative training. A connected-digit database containing utterances recorded in a noisy, moving car with a hands-free microphone mounted on a sun-visor is used to evaluate the verification performance. The equal error rate (EER) of word verification is employed as the performance measure in our evaluations. All factors considered in our study and their effects on the verification performance are presented in detail. The EER is reduced from 29%, using the standard likelihood ratio test, down to 21.4%, when all enhancements are integrated together.

1. INTRODUCTION

The automatic speech recognition (ASR) systems in human-machine dialogue systems require a high word recognition accuracy. However, the performance of these systems can be seriously degraded, especially in noisy and hands-free environments. To enhance the ASR performance and to design a friendlier voice user interface, a procedure is often integrated into ASR systems to verify or reaffirm the recognition results. In [1]-[3], the recognizer itself is redesigned with a verification procedure. In [4]-[10], a verification procedure is used as a post-classification measure in the second stage before the recognition result is declared final.

This study investigates a verification procedure as a post-classification measure. Although recognition errors are inevitable, the verification procedure can potentially reduce the negative impact of an incorrect recognition decision or false triggering due to background interference, as found in hands-free environments.

Presently, utterance verification is performed based on the traditional framework of hypothesis testing. It employs a "standard" likelihood ratio (SLR) test to reconfirm a tentative decision that given token X_u is recognized as word w_u. The decision rule is expressed as follows:

$$f(X_u \mid w_u) = \log \frac{p(X_u \mid w_u)}{p(X_u \mid \overline{w}_u)} \mathop{\gtrless}_{\overline{w}_u}^{w_u} \rho \,, \tag{1}$$

where ρ is the prescribed threshold and $\overline{w}_u$ is related to an alternate decision (i.e., to reject the tentative decision).

The legitimacy of SLR based hypothesis testing depends on the conditional probability density function (pdf) being accurate for each word, and in the case of continuous speech recognition, it depends on acoustic token X_u being correctly segmented. In reality, however, the pdf needs to be modeled by an appropriate functional form and the parameters need to be accurately estimated from a limited amount of training data. Because the token is obtained as a byproduct of the recognizer, its segmentation may not be perfect due to noise in the observations. That is, the supporting probability space as implied in the traditional hypothesis-testing framework may not have a clear yet practical definition. One thus may wish to transform the verification problem as a separate 2-class classification problem potentially involving a new set of observations as outlined below.

In our new 2-class classifier approach to the verification problem, we use the likelihoods evaluated on all classes as the observations. This enables a more sophisticated discriminant function for classes w_u and $\overline{w}_u$ and provides several additional refinements, for example, tests at different levels (e.g., the phone and word-levels can be accommodated [5]-[7]); use of different features (e.g., the state duration can be incorporated [5]-[7]); use of non-linear classifiers; and integration of supplementary tests using N-best hypotheses. We believe that a 2-class classifier is a good alternative to the traditional SLR paradigm and allows the possible design of a robust test function.

2. CLASSIFIER DESIGN

As input parameters, the approach uses the likelihood value of each recognized word, which is a byproduct of the recognition process. To alleviate insertions, it also employs the likelihood values of the BG and OOV filler models. The traditional hypothesis testing is supplemented with the like-

B. –H. Juang has joined AVAYA Labs Research.

lihood ratios for the N-best candidates to make it robust to noise. To reduce the adverse impact of doubtful likelihood values due to outliers, the likelihood ratios are smoothed, compressed, and regulated by logarithm and sigmoid functions. The classifier is optimized using discriminative training to minimize the classification error.

The definition is as follows.

$$
\begin{aligned}
&f(X_u \mid w_u^{(1)}) \\
&= \sum_{i=1}^{N} \lambda_i \delta\left[\log\left(\frac{p(X_u \mid w_u^{(i)})}{p(X_u \mid \overline{w}_u^{(i)})}\right)\right] + \\
&\qquad \lambda_\phi \delta\left[\log\left(\frac{p(X_u \mid \phi)}{p(X_u \mid \overline{\phi})}\right)\right] + \lambda_\varphi \delta\left[\log\left(\frac{p(X_u \mid \varphi)}{p(X_u \mid \overline{\varphi})}\right)\right] \\
&= \sum_{i=1}^{N} \lambda_i \delta\left[l_u(w_u^{(i)})\right] + \lambda_\phi \delta[l_u(\phi)] + \lambda_\varphi \delta[l_u(\varphi)]
\end{aligned} \tag{2}
$$

where $w_u^{(i)}$ indicates the i^{th} best word candidate, ϕ the BG model, and φ the OOV model. δ is the sigmoid function and l_u is the elliptical function of the log-likelihood ratio. The likelihood values of anti-models $\overline{w}_u^{(i)}$, $\overline{\phi}$, and $\overline{\varphi}$ are approximated by taking the geometric means of all likelihood values except those of $w_u^{(i)}$, ϕ, and φ, respectively, as follows.

$$
p(X_u \mid \overline{w}_u^{(i)}) = \left(\prod_{j=1, j\neq i}^{N} p(X_u \mid w_u^{(j)}) p(X_u \mid \phi) p(X_u \mid \varphi) \right)^{1/N+1} \tag{3}
$$

$$
p(X_u \mid \overline{\phi}) = \left(\prod_{j=1}^{N} p(X_u \mid w_u^{(j)}) p(X_u \mid \varphi) \right)^{1/N+1} \tag{4}
$$

$$
p(X_u \mid \overline{\varphi}) = \left(\prod_{j=1}^{N} p(X_u \mid w_u^{(j)}) p(X_u \mid \phi) \right)^{1/N+1} \tag{5}
$$

$\{\lambda_1,\ldots, \lambda_N, \lambda_\phi, \lambda_\varphi\}$ is discriminatively trained using the GPD method [11].

3. DISCRIMINATIVE TRAINING

The classifier coefficients $\{\lambda_1,\ldots, \lambda_N, \lambda_\phi, \lambda_\varphi\}$ are discriminatively trained using the GPD method [11] to minimize the classification error. In the hypothesis testing, the discriminative functions for null hypothesis (H_0) and alternative hypothesis (H_1) are defined as follow.

$$
\begin{aligned}
H_0 : g_0(X_u) &= \sum_{i_0=1}^{N} \lambda_{i_0} \delta\left[l_u(w_u^{(i_0)})\right] + \\
&\qquad \lambda_{\phi_0} \delta[l_u(\phi)] + \lambda_{\varphi_0} \delta[l_u(\varphi)]
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
H_1 : g_1(X_u) &= \sum_{i_1=1}^{N} \lambda_{i_1} \delta\left[l_u(w_u^{(i_1)})\right] + \\
&\qquad \lambda_{\phi_1} \delta[l_u(\phi)] + \lambda_{\varphi_1} \delta[l_u(\varphi)]
\end{aligned} \tag{7}
$$

The misclassification measure is defined as follow.

$$
\begin{aligned}
d(X_u) &= -g_0(X_u) + g_1(X_u) \\
&= \sum_{i=1}^{N} \theta_i \delta\left[l_u(w_u^{(i)})\right] + \theta_\phi \delta[l_u(\phi)] + \theta_\varphi \delta[l_u(\varphi)]
\end{aligned} \tag{8}
$$

$$
\theta_i = \lambda_{i_0} + \lambda_{i_1}, \quad \theta_\phi = \lambda_{\phi_0} + \lambda_{\phi_1}, \quad \theta_\varphi = \lambda_{\varphi_0} + \lambda_{\varphi_1}
$$

By minimizing the cost function in (9), which is defined as the sigmoid function of the misclassification measure, we obtain the optimal set of classification coefficients $\{\theta_1,\ldots, \theta_N, \theta_\phi, \theta_\varphi\}$.

$$
\delta[d(X_u)] = \frac{1}{1+\exp(\alpha \cdot d(X_u) + \beta)} \tag{9}
$$

θ_i is adjusted by a small amount $\Delta\theta_i$ according to (10) for the null hypothesis and (11) for the alternative hypothesis.

$$
\begin{aligned}
H_0 : \Delta\theta_i &= -\varepsilon \nabla \delta[d(X_u)] = -\varepsilon \nabla \delta'[d(X_u)] \frac{\partial d(X_u)}{\partial \lambda_{i_0}} \\
&= \varepsilon \frac{\alpha \exp(\alpha \cdot d(X_u) + \beta}{\{1+\exp(\alpha \cdot d(X_u) + \beta\}^2} \delta\left[l_u(w_u^{(i)})\right]
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
H_1 : \Delta\theta_i &= -\varepsilon \nabla \delta[d(X_u)] = -\varepsilon \nabla \delta'[d(X_u)] \frac{\partial d(X_u)}{\partial \lambda_{i_1}} \\
&= -\varepsilon \frac{\alpha \exp(\alpha \cdot d(X_u) + \beta}{\{1+\exp(\alpha \cdot d(X_u) + \beta\}^2} \delta\left[l_u(w_u^{(i)})\right]
\end{aligned} \tag{11}
$$

It should be noted that the notion of negative examples, as in the conventional discriminative learning paradigm, is not applicable here. We shall not attempt to use correctly recognized tokens for negative learning assuming they could have been misrecognized, for the recognizer already entirely defined by the data characteristics. These tokens do not appear in the testing to reduce uncertainties in misrecognized decisions.

4. CHANGING SEGMENTATION POINTS

Segmentation points are usually obtained based on the maximum likelihood (ML) criterion for a recognition unit, e.g., an utterance between silent pauses. It might not be optimal to verify each component word in an utterance. We therefore vary the segmentation points in an attempt to improve the verification result. This is motivated on the basis that neighboring speech events can overlap each other due to co-articulation effects. Moreover, in [6], it was reported that extending boundaries up to 50% to overlap with neighboring segments can lead to performance improvements. We expand the word segment by k frames at both ends and find the optimal k using the training data.

5. EXPERIMENTS

5.1 Database and System Description

All experiments are carried out using the car voice user interface (CARVUI) database, containing utterances recorded in a running car under typical car noise in the background. More specifically, the CARVUI database consists of speech data simultaneously recorded through multi-microphone channels, including a head-mounted, close-talking microphone and a 16-channel microphone array located on a sun-visor. 56 speakers including some non-native English speakers uttered phonetically-balanced TIMIT sentences, digit strings with 1 to 7 digits, and about 85 short commands for car application. The data was originally sampled at 24 kHz. In our experiments, hands-free speech data recorded through a channel of the microphone array is used, and all data is down-sampled to 8 kHz.

For the baseline recognizer, speaker-independent monophone acoustic models are built for 41 phones and three short/long/noisy silences using 3,984 utterances of digit strings and TIMIT sentences uttered by 45 speakers. The total number of mixture components is 2,055 and the averaged number per state is 15.8. The feature vectors of 39 components, consisting of 12th-order mel-frequency cepstral coefficients plus the normalized log energy term and both of their first and second derivatives, are derived every 10 ms over 20 ms Hamming windowed segments. The number of filters is 18. Cepstral mean subtraction is applied for each utterance both in the training and testing. A finite state grammar with digit strings of an unknown length is used as the language model. The lexicon size is 11 including /0/ to /9/ and /oh/.

The BG filler model is composed of a silence-loop model consisting of 3-state long, 1-state short, and 3-state noisy silence models in the above speaker-independent monophone models. The OOV filler model is composed of a phone-loop model consisting of 41 phone models in speaker-independent monophone models having less resolution with 483 mixture components in total and 3.9 per state.

In the GPD training, 7,481 correct segments in digit strings uttered by the same 45 speakers as those for the acoustic models are used. The number of digit speakers per speaker is 50.

In the testing, 965 correct, 125 substitution, and 72 insertion segments uttered by seven speakers, who are different from the training speakers, are used. The word correct rate with the speaker-independent models is 87.5% on average. We use the equal error rate (EER) of word verification as the performance measure in our evaluations. The verification threshold is set a posteriori and is digit-dependent. The classifier coefficients are estimated for each digit.

5.2 SLR-Based Hypothesis Testing vs. 2-Class Classifier

The performance of the SLR-based hypothesis testing and that of our 2-class classifier are compared. The form for the SLR-based hypothesis testing is defined as follows.

$$f^{SLR}(X_u \mid w_u^{(1)}) = \delta\left[l_u(w_u^{(1)})\right] \tag{12}$$

For our 2-class classifier, in contrast, several variations can be considered. Here, we examine the following three variations.

$$f^{2CC\ N-best}(X_u \mid w_u^{(1)}) = \sum_{i=1}^{N} \lambda_i \cdot \delta\left[l_u(w_u^{(i)})\right] \tag{13}$$

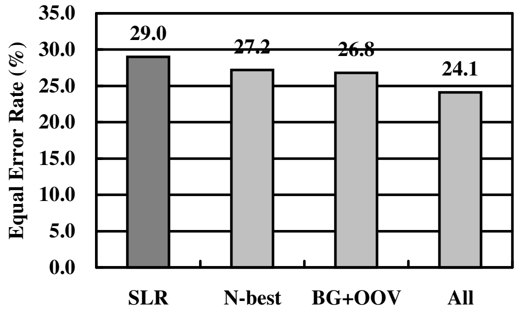

Fig. 1. SLR-based hypothesis testing vs. 2-class classifier.

$$\begin{aligned} f^{2CC\ BG+OOV}(X_u \mid w_u^{(1)}) = \lambda_1 \cdot \delta\left[l_u(w_u^{(1)})\right] + \\ \lambda_\phi \cdot \delta\left[l_u(\phi)\right] + \lambda_\varphi \cdot \delta\left[l_u(\varphi)\right] \end{aligned} \tag{14}$$

$$\begin{aligned} f^{2CC\ All}(X_u \mid w_u^{(1)}) = \sum_{i=1}^{N} \lambda_i \cdot \delta\left[l_u(w_u^{(i)})\right] + \\ \lambda_\phi \cdot \delta\left[l_u(\phi)\right] + \lambda_\varphi \cdot \delta\left[l_u(\varphi)\right] \end{aligned} \tag{15}$$

In "2CC N-best" of (13), only the likelihood ratios for the N-best candidates are used. In "2CC BG+OOV" of (14), the likelihood ratios for the best candidate, BG model, and OOV model are used. In "2CC All" of (15), all of the likelihood ratios (i.e., the N-best candidates, BG model, and OOV model) are used.

Figure 1 shows the equal error rates for word verification when comparing the performance of the SLR-based hypothesis testing and that of our 2-class classifier. The 11-best likelihood ratios are used in "2CC N-best" and "2CC All". As the figure illustrates, the 2-class classifier performs better than the SLR-based hypothesis testing. The performance improves as more information is used.

5.3 Likelihood vs. Likelihood Ratio Based Formulation

Our 2-class classifier is formulated based on not the likelihood but the likelihood ratio. The appropriateness of this is examined by comparing the performance of the likelihood based formulation and that of the likelihood ratio based formulation. The likelihood based formulation is expressed as follows.

$$\begin{aligned} f^{2CC\ All(L)}(X_u \mid w_u^{(1)}) = \sum_{i=1}^{N} \lambda_i \cdot \delta\left[\log\left(p(X_u \mid w_u^{(i)})\right)\right] + \\ \lambda_\phi \cdot \delta\left[\log\left(p(X_u \mid \phi)\right)\right] + \lambda_\varphi \cdot \delta\left[\log\left(p(X_u \mid \varphi)\right)\right] \end{aligned} \tag{16}$$

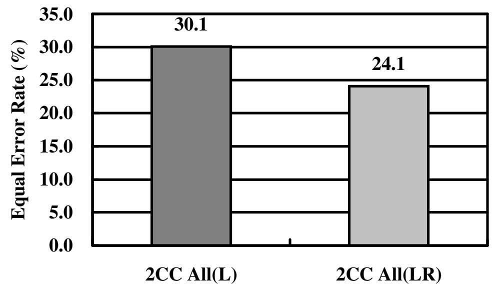

Fig. 2. Likelihood vs. likelihood ratio based formulation.

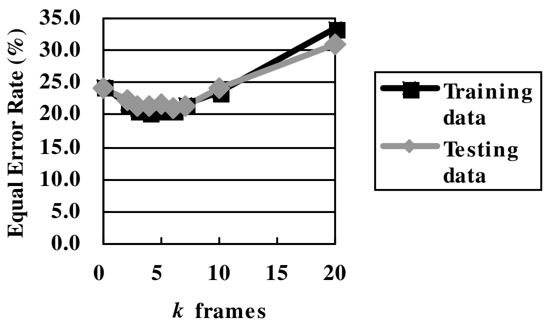

Fig. 3. Word verification performance for training and testing data as a function of *k* frames extended at both ends.

It should be noted that for both formulations, the input parameters of the likelihood values of the N-best candidates, BG model, and OOV model are the same.

Figure 2 shows the equal error rates for word verification when comparing the performance of the likelihood based formulation and that of the likelihood ratio based formulation, "2CC All(L)" and "2CC All(LR)", respectively. In the figure, the latter performs better than the former. This indicates that our formulation can be considered appropriate.

5.4 Effects of Changing Segmentation Points

This section evaluates the performance of changing segmentation point (CSP) optimization. Figure 3 shows equal error rates for training and testing data as a function of *k* frames extended at both ends. In Figure 3, the case of 0 frame corresponds to using the recognized segmentation. The equal error rates show almost the same tendencies for the training and testing data. From two frames to six frames, the curves of the equal error rates were relatively flat. With increasing number of frames, the performance degrades more. These results confirm that the optimal *k* can be found using training data. Here, we set *k* to 4.

Figure 4 shows the equal error rates of word verification with/without CSP optimization. The figure shows that CSP increases the performance. The relative error reduction from the SLR-based hypothesis testing to our classifier with CSP is 26.2%, while that from the SLR-based hypothesis testing to our classifier without CSP is 16.9%. In addition, the relative error reduction from our classifier without CSP to with CSP is 11.2%.

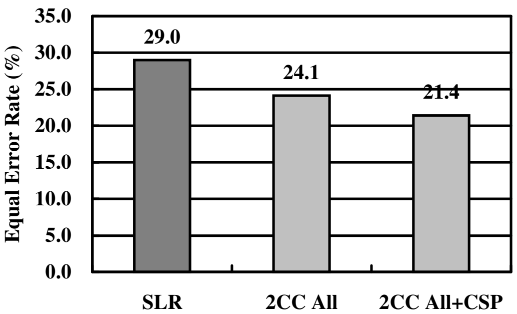

Fig. 4. Word verification performance with/without CSP.

6. CONCLUSIONS AND ACKNOWLEDGEMENT

This paper proposes a new confidence measure based on a 2-class classifier with GPD training. The traditional hypothesis testing is improved with N-best candidate scores. Segmentation point optimization improves the performance. In connected digit recognition experiments using distant-talking, hands-free speech data, the proposed method is shown to achieve an equal error rate of word verification of 21.4%. The relative error reduction is 26.2%, when compared with the standard likelihood ratio test.

We thank Dr. Seiichi Yamamoto and Dr. Satoshi Nakamura of ATR-SLT for supporting this study.

7. REFERENCES

[1] C. V. Neti, S. Roukos and E. Eide, "Word-based confidence measures as a guide for stack search in speech recognition," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Seattle, Munich, Germany, pp. 883-886,1997.

[2] E. Lleida and R. C. Rose, "Utterance verification in continuous speech: decoding and training procedures," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 126-139, 2000.

[3] M.–W. Koo, C.-H. Lee and B.-H. Juang, "A new decoder based on a generalized confidence score," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Seattle, WA, 1998, pp. 213-216.

[4] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 420-429, 1996.

[5] R. A. Sukkar, A. R. Setlur, C.-H. Lee and J. Jacob, "Verifying and correcting recognition string hypotheses using discriminative utterance verification," *Speech Commun.*, vol. 22, pp. 333-342, 1997.

[6] C. Garcia-Mateo, W. Reichl and S. Ortmanns, " On combining confidence measures in HMM-based speech recognizers," in *Workshop Automatic Speech Recognition Understanding (ASRU)*, 1999.

[7] C. Ma, M. A. Randolph and J. Drish, "A support vector machines-based rejection technique for speech recognition," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Salt Lake City, UT, 2001.

[8] S. O. Kamppari and T. J. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, Istanbul, Turkey, pp.1799-1802, 2000.

[9] T. Kawahara, C.-H. Lee and B.-H. Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 558-568, 1998.

[10] P. Ramesh, C.-H. Lee and B.-H. Juang, "Context dependent anti subword modeling for utterance verification," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998.

[11] S. Katagiri, C.-H Lee and B.-H. Juang, "New discriminative training algorithm based on the generalized probabilistic descent method," In *Proc. IEEE workshop, Neural Networks for Signal Processing*, pp. 299-300, 1991.