ROBUST UTTERANCE VERIFICATION FOR CONNECTED DIGITS RECOGNITION

Mazin G Rahim, Chin-Hui Lee and Biing-Hwang Juang
AT&T Bell Laboratories, Murray Hill, NJ 07974

ABSTRACT

Utterance verification represents an important technology in the design of user-friendly speech recognition systems. This paper addresses the issue of robustness in utterance verification. Four different approaches to robustness have been investigated: a string based likelihood measure for the detection of non-vocabulary words and "putative" errors, a signal bias removal method for channel normalization, on-line adaptation technique for achieving desirable trade-off between false rejection and false alarms, and a discriminative training method for the minimization of the expected string error rate. When these techniques were all integrated into a state-of-the-art connected digit recognition system, the string error rate was found to decrease by up to 57% at a rejection rate of 5%. For non-vocabulary word strings, the proposed utterance verification system rejected over 99.9% of extraneous speech.

1. INTRODUCTION

During recent years, it has become increasingly essential to equip speech recognition systems with the ability to accommodate spontaneous speech input. Although this provides a friendly user-interface, it poses a number of new problems, such as the inclusion of out of vocabulary words, false starts, disfluency and acoustical mismatch. In these situations, a speech recognition system must be able to detect and recognize the "keywords" and reject the "non-keywords." Recognizers equipped with a keyword spotting capability allow users the flexibility to speak naturally.

Significant progress has been made in keyword spotting for unconstrained speech using hidden Markov modeling (HMM). The issue of constructing a suitable filler (or garbage) model has been extensively studied by Rose and Paul [5], Wilpon *et al* [7] and others. A filler is usually modeled with a structure of a whole word, a monophone, a triphone, or a broad phonemic class.

Several studies have focused on the selection of appropriate features for *keyword spotting* (e.g., [1]). Conventionally, the likelihood score corresponding to the best Viterbi path is typically the most important feature for verification especially when normalized by the score of the filler model. Further enhancement and feature reduction can be achieved by linear transformation or discriminative feature analysis [6].

As a generalization to keyword spotting, *utterance verification* attempts to reject or accept part or all of an utterance based on a computed confidence score (e.g., [6]). This is particularly useful in situations where utterances are spoken without valid keywords or when significant confusion exists among keywords which may result in a high substitution error probability. In general, to deal with these types of problems, recognizers must be equipped with both a keyword spotting capability to correctly recognize keywords embedded in the utterance, and with an utterance verification capability to reject utterances that do not contain valid keywords and utterances that have low confidence scores.

An important task in keyword spotting and utterance verification is the selection of an appropriate *operating point* (or critical threshold) to provide a desirable combination of *Type I error* (false rejection) and *Type II error* (false alarm). In this paper, we will demonstrate that utterances recorded under different environmental conditions require different operating points in order to satisfy an optimality criterion. Further, it will be shown that a small deviation in the selected operating point could result in severe changes in the overall performance of the verification system. This critical issue raises the important question of how to maintain "robustness" during verification. Robust verification is a subject that demands considerable attention and is the focus study of this paper.

2. DATABASES AND BASELINE SYSTEM

In order to assess the robustness of the various rejection techniques presented in this paper, the performance of two connected digits data sets was evaluated. Both data sets were recorded over telephone lines, using two electret and two carbon button microphones. Speech was transmitted over a long-distance telephone network which was either all analog, all digital or a mix, depending on the region.

- The first data set was collected from two regions, namely, Long Island and Boston, over a digital T1 interface. Digit strings of lengths 10 to 16 digits were collected from 250 adult talkers. Approximately half of the speakers were used for training the HMMs and the other half for testing. The testing data set which consists of 2842 strings will be referred to as DB1.
- The second data set was collected from five regions within the United States, namely, Long Island, Chicago, Boston, Columbus and Atlanta. Each region consisted of 100 adult talkers (50 males and 50 females), each speaking 66 connected digit strings from a predefined list (11 digit strings for each of the length two through seven). A subset of this database consisting of 7073 strings was assigned for testing. This data set will be referred to as DB2.

In order to provide non-keyword utterances for training and verification, phonetically-rich sentences were used. About 3000 sentences were used for training and another 3000 to 6000 sentences were added to DB1 and DB2 for testing.

The front-end process of the recognition system utilized 12 LPC liftered cepstral coefficients, along with a normalized energy feature. The combined feature vector was augmented with its first and second order time derivatives, resulting in a vector of 39 dimensions per frame.

Each keyword (i.e., digit) was modeled by a left-to-right continuous density HMM. Training included estimating the

0-7803-2431-5/95 $4.00

mean, covariance and mixture weights for each state using maximum likelihood (ML) estimation. For each keyword model, a digit-specific anti-keyword model was also trained (details will be provided in the next section). Aside from keywords and anti-keywords, we also introduced a general acoustic filler model, trained on non-digit speech data, and a background/silence model trained on the non-speech segments of the signal. Therefore, a total of 24 models was used, each consisting of ten 16-mixture states with the exception of a single state silence model.

A two-pass strategy was adopted consisting of recognition followed by verification. In the first pass, recognition was performed via a conventional Viterbi beam search algorithm within an HMM framework. In the second pass, an utterance based confidence score was computed and applied for verification. The baseline string recognition performance when testing on DB1 with known length grammar and DB2 with unknown length grammar was 91.0% and 84.5%, respectively.

Following recognition, an input utterance was segmented into a string of keyword hypotheses. For each keyword segmentation, a likelihood score was also obtained for the corresponding anti-keyword and filler model. A confidence score based on a likelihood ratio test was then performed and the utterance was either accepted or rejected.

3. DIGIT BASED VERIFICATION

Starting with four sets of HMMs, namely, 11 digits $\{\lambda_k\}$, 11 digit-specific anti-keywords $\{\overline{\lambda}_k\}$, silence/background λ_s and filler λ_f, digit verification is carried out by testing the *null hypothesis* that a specific digit exists in a segment of speech versus the *alternative hypothesis* that the digit is not present. Based on a likelihood ratio test, the digit is accepted or rejected if the log likelihood ratio $L_k(O|\Lambda)$ lies above a specific verification threshold τ_k (here $\Lambda = \{\lambda_k\}, \{\overline{\lambda}_k\}, \lambda_s, \lambda_f$).

In this study, we considered several different formulations for the alternative hypothesis, two of which will be presented in this section. The first choice is simply to use the general acoustic filler model λ_f which is *digit independent*. This is trained using non-digit extraneous speech and is the same for all digits. The likelihood for the alternative hypothesis is defined as

$$G_k^{(1)}(O;\Lambda) = \log[p(O|\lambda_f)]. \quad (1)$$

This type of formulation is believed to improve discrimination between keywords and out of vocabulary words.

The second choice for the alternative hypothesis is to introduce a digit-specific anti-keyword model to provide better detection of near-misses in digit recognition. Clearly, there are many strategies for constructing such models, such as using the likelihood of all competing digits or constructing additional digit-specific anti-keyword models, $\{\overline{\lambda}\}_k$, is trained on all digits except for digit k. This paper will only discuss the latter type since it provided the best results.

In order to provide improved discrimination between keyword and non-keyword models as well as reasonable detection of putative errors, the likelihood of the alternative hypothesis based on the anti-digit model was formulated as

$$G_k^{(3)}(O;\Lambda) = \log\Big[\frac{1}{2}\exp\{\eta G_k^{(1)}(O;\Lambda)\} + \frac{1}{2}\exp\{\eta G_k^{(2)}(O;\Lambda)\}\Big]^{\frac{1}{\eta}}, \quad (2)$$

where $G_k^{(2)}(O;\Lambda) = \log[p(O|\overline{\lambda}_k)]$ and η is a constant. Figure 1 gives the equal error rates (i.e., Type I = Type II) for all the eleven digits when utilizing a likelihood ratio score based on either $G_k^{(3)}(O;\Lambda)$ or $G_k^{(1)}(O;\Lambda)$. Clearly, for almost all digits, it is safe to conclude that digit-specific anti-keywords are somewhat complementary to a general acoustic filler model. Combining the two measures in a geometric average has resulted in a reduced error rate. A similar trend was found when plotting the minimum total verification error (i.e., Type I plus Type II) for all digits.

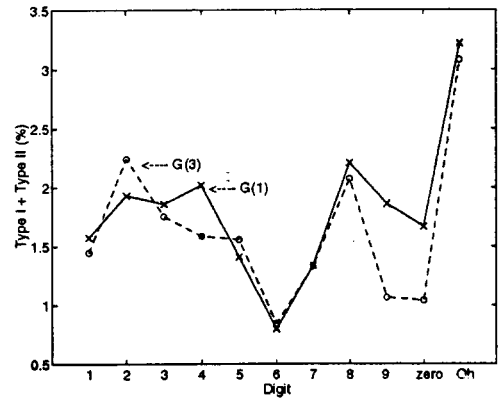

Figure 1. Equal error rates for the digits using likelihood ratio scores based on $G_k^{(1)}(O;\Lambda)$ and $G_k^{(3)}(O;\Lambda)$.

Throughout the rest of the paper, digit verification will be conducted using a likelihood ratio test with the anti-digit function $G_k^{(3)}(O;\Lambda)$ $(= G_k(O;\Lambda))$.

4. UTTERANCE BASED VERIFICATION

There are several advantages is using utterance verification (or rejection) in connected digits recognition. The first is verifying whether the recognized digit string is a *valid* digit string. This enables rejection of strings which contain non-vocabulary words or noise. The second is verifying whether a valid digit string is a *correct* digit string.

Two approaches for utterance verification have been evaluated in the study. In the first approach, the utterance confidence measure is based on individual digit scores, such that an utterance is rejected if the test of any detected digit q

$$S^{(1)}(O;\Lambda) = LR_q(O;\Lambda) < \tau_q, \quad (3)$$

where $LR_q(O;\Lambda) = g_q(O;\Lambda) - G_q(O;\Lambda)$ and $g_q(O;\Lambda) = \log[p(O|\lambda_q)]$. (i.e., reject if any one detected digit falls below the operating point, τ_q).

The second approach for string-based verification computes an utterance based confidence score using a geometric average of all detected digits. Thus, for an N-digit string

$$S^{(2)}(O;\Lambda) = -\log\Big[\frac{1}{N}\sum_{q=1}^{N}\exp\{-\kappa \cdot LR_q(O;\Lambda)\}\Big]^{\frac{1}{\kappa}}, \quad (4)$$

where κ is a constant. There are two advantages of this measure compared to $S^{(1)}(O;\Lambda)$. First, it provides string verification statistics based only on one distribution rather than one per digit. This eases the computational effort when conducting on-line adaptation as will be discussed in section 5.2. Second, this measure acknowledges the contributions of all the digits within a given string based on the selected value of κ.

Experiments conducted using the two string-based verification functions defined in equations (3) and (4) have consistingly demonstrated a lower error rate when using $S^{(2)}(O;\Lambda)$ at any operating point. When testing on DB2, the equal error rates for $S^{(1)}(O;\Lambda)$ and $S^{(2)}(O;\Lambda)$ were at 2.5% and 2.3%, respectively. This amounts to a reduction

of 8%. The geometric function $S^{(2)}(O;\Lambda)$ $(= S(O;\Lambda))$ will be used in all remaining experiments.

5. ROBUST UTTERANCE VERIFICATION

In order to evaluate the robustness property of our rejection scheme, further experiments have been conducted with data set DB2. This data set can be considered as a case of "mismatched" condition since it was recorded in a completely different environment than the training corpus. Figure 2 shows a plot of the combined Type I and Type II errors (including non-vocabulary words) for the two data sets. Three remarks can be made. First, the minimum error rate point is different across the two data sets. For example, DB2 has a minimum at an operating point of 2.6 whereas DB1 has a minimum at about 3.2. Setting the operating point to 3.2 causes the total error rate for DB2 to become 7.8% rather than 6.8% (i.e., an increase of about 15%). Second, DB2 has a higher error rate than DB1. This is due to an environmental mismatch between the training model and the testing data. Third, it appears that DB2 is more affected by changes in the operating point than DB1. For example, if an original threshold at 3.2 is raised by, say, 25% to 4.0 then this causes the combined error rates for DB1 and DB2 to increase to 4.3% (i.e., +12%) and 11.7% (i.e., +50%), respectively.

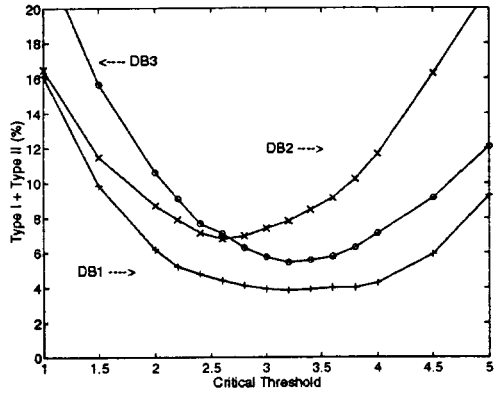

Figure 2. Total verification (Type I plus Type II) errors for DB1 and DB2.

The following sections will discuss three different approaches to robustness, namely, signal bias removal (SBR), maximum a posteriori (MAP) on-line threshold adaptation and discriminative training based on minimum classification error (MCE).

5.1. Signal bias removal (SBR)

The SBR method is an iterative procedure which minimizes the effects of unknown adverse conditions that contaminate speech. It is based on a formulation that aims at separating two processes, one being the speech signal and the other is what we call a bias process. The reader is referred to [4] for further details of this method.

Figure 3 shows the string recognition accuracy as a function of rejection rate when testing on DB2 with and without the introduction of SBR. At a rejection rate of 5%, the string accuracy is improved from 87.4% to 89.2% (improvement from 93.6% to 93.9% was obtained for DB1).

5.2. On-line threshold adaptation

The performance of an utterance verification system is largely dependent on the selection of an appropriate critical threshold. In previous sections, thresholds have been set according to a predefined criterion, such as to minimize

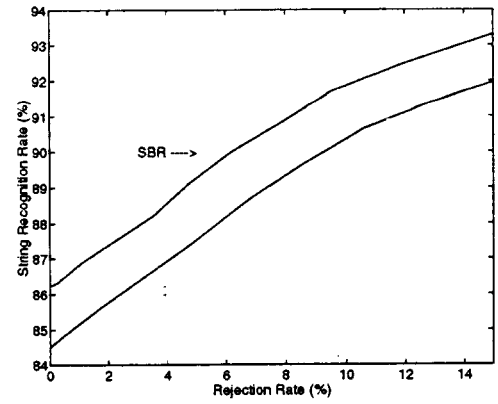

Figure 3. String recognition performance as a function of rejection rate for DB2 when introducing SBR.

the combined Type I and Type II errors or to achieve an equal error rate. Although this can almost be guaranteed for an evaluation set, e.g., DB1, the same criterion may not be satisfied when dealing with an environmental mismatch between the training and the testing data, such as when testing on DB2.

In this study, we have explored several different techniques for on-line adaptation based on neural networks, gradient descent and MAP estimation. Details of these methods and the results will be provided elsewhere. In this section, we will only discuss the MAP approach to demonstrate on-line adaptation.

The histograms for the string likelihood scores, $S(O;\Lambda)|_{O\in C_t}$ and $S(O;\Lambda)|_{O\notin C_t}$, are approximated by two single Gaussian densities with means, μ_{H_1} and μ_{H_0}, and variances, $\sigma^2_{H_1}$ and $\sigma^2_{H_0}$. Assuming that the mean of each histogram is a random variable with both a prior Gaussian distribution of a known variance σ_p^2, and a conjugate prior which is also Gaussian of mean μ and variance σ^2, then the MAP estimate of the mean, $\tilde{\mu}$, is

$$\tilde{\mu} = \frac{n\sigma^2}{\sigma_p^2 + n\sigma^2}\overline{x} + \frac{\sigma_p^2}{\sigma_p^2 + n\sigma^2}\mu, \qquad (5)$$

where n is the number of observations and $\overline{x}$ is the sample mean [3].

In our study, the prior variance σ_p^2 for the distributions was 10 times larger than their respective initial variance estimates. The mean of each a class of distribution was updated using equation (5). The update was conducted at every 40 string scores. It was established that about 30-50 scores were necessary to obtain stable parameter estimates. It was also found that applying the same formulations presented in equation (5) to update the variances helped in improving the overall performance of the verification system. Thus, the variance of each distribution was updated using a weighted average of the prior variance and the sample variance.

To explore robustness, the initial estimates of the means of the prior distributions were varied to allow the operating point to artificially change between 1 and 5. The operating point was adjusted following MAP adaptation to obtain a minimal total of Type I and Type II errors. Figure 4 shows the variation of the total error before adaptation and after supervised and unsupervised adaptation. Two interesting features can be concluded from this plot. One, the difference between supervised and unsupervised adaptation is very minimal. Two, the combined error rate is rather unaf-

fected by the setting of the initial parameters resulting in a robust verification system.

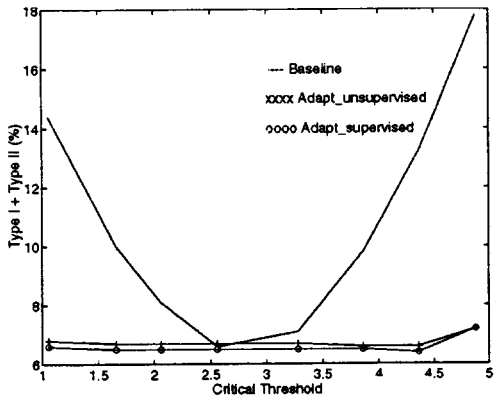

Figure 4. Combined Type I and Type II error for supervised and unsupervised threshold adaptation via MAP estimation using data set DB2.

5.3. Discriminative training

A different approach to improving the robustness of utterance verification systems is to employ discriminative training. This section describes the application of string-based MCE training, in the framework of the generalized probabilistic descent (GPD), to utterance verification. Further details of the MCE/GPD method is given in [2]

Throughout all our experiments, MCE/GPD was only applied in training the filler and the keyword models. The digit-specific anti-keyword models were not trained with this technique. Figure 5 shows the two histograms for the in-class/out-class string likelihood scores when applying ML training (dotted lines) alone and followed by MCE/GPD training (solid line) using DB1. Clearly, the discriminative training technique has provided a better separation of the two histograms, a feature which is more apparent in the left histogram representing the incorrect class.

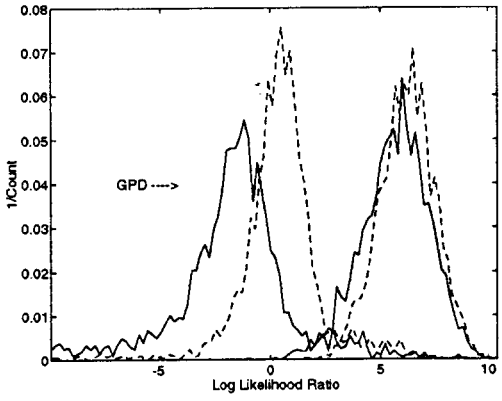

Figure 5. Histograms showing the distribution of the string likelihood scores before and after GPD.

Naturally, since the histograms of the string likelihood scores are less overlapped than those previously obtained with ML training, a decrease in the error rate would be expected. Introducing MCE/GPD training over SBR processing, at a rejection rate of 5%, resulted in an improvement in the string verification performance from 93.9% to 96.1% for DB1 and from 89.2% to 90.3% for DB2.

6. SUMMARY

This paper described a robust utterance verification system for connected digits recognition. A digit based likelihood ratio combining the scores of keywords, anti-keywords and filler models was found to be effective in detecting and rejecting non-vocabulary words and, in some instances, reducing putative errors.

To perform utterance verification, a string-based likelihood measure was proposed based on a geometric average of the digit likelihood ratios. This measure has the advantage of being able to deal with a single set of histograms and providing a small improvement over standard techniques for utterance verification.

Table 1. String recognition performance at a rejection rate of 5%. The verification system included the string based likelihood measure, SBR, unsupervised MAP adaptation and MCE/GPD training.

Dataset	Before (%)	After (%)	Improv. (%)
DB1	91.0	96.1	56.7
DB2	84.5	90.3	37.4

The issue of robustness in utterance verification was examined in this study. It is demonstrated that a verification system may not perform adequately under all environmental conditions. Different operating points were found necessary in order to establish a desired combination of false alarms and false rejection. Different methods were investigated for robustness, namely, SBR, MAP estimatation and MCE/GPD training. When all of these methods were integrated into the recognition system, a reduction in the string error rate was achieved up to about 57% (see table 1). The same system also achieved over 99.9% correct rejection of non-vocabulary sentences.

Acknowledgements

The authors acknowledge useful discussions with R. Rose and R. Sukkar. We also thank W. Chou for his code on string based MCE/GPD training.

REFERENCES

[1] Chigier, B. (1992) "Rejection and Keyword Spotting Algorithms for a Directory Assistance City Name Recognition Application," *Proc. ICASSP*, **II**, pp. 93-96.

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

[3] Lee, C.-H., Lin, C.-H., and Juang, B.-H. (1991) "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *Trans. ASSP*, **39**(4), pp. 806-814.

[4] Rahim, M and Juang, B-H. (1994) "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," *Proc. ICASSP*, **I**, pp. 445-448.

[5] Rose, R., and Paul, D. (1990) "A Hidden Markov Model Based Keyword Recognition System," *Proc. ICASSP*, **I**, pp. 129-132.

[6] Sukkar, R. (1994) "Rejection for Connected Digit Recognition Based on GPD Segmental Discrimination," *Proc. ICASSP*, **I**, pp. 393-396.

[7] Wilpon, J., Rabiner, L., Lee, C.-H., and Goldman, E. (1990) "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *Trans. ASSP*, **38**(11), pp. 1870-1990.