IMPROVING UTTERANCE VERIFICATION USING ADDITIONAL CONFIDENCE MEASURES IN ISOLATED SPEECH RECOGNTION INTERFACES

Graham Greenland, Willy Wong, Hans Kunov

Edward S. Rogers Sr. Dept. of Electrical and Computer Engineering
Institute of Biomaterials and Biomedical Engineering
University of Toronto

ABSTRACT

A major problem with speech recognition interfaces is the detection and correct rejection of words that lie outside of the command vocabulary. Given the limited resources provided to an interface, any method that is used to detect out-of-vocabulary (OOV) words should ideally require minimal computational resources. A new utterance verifier is proposed which detects OOV words efficiently using confidence measures calculated for additional words in the N-Best list. This approach can be used with any existing confidence measure. We implement this utterance verifier with three different confidence measures and obtain a 3-16% improvement over their standard implementation.

1. INTRODUCTION

Speech recognition interfaces allow hands free operation of electronic devices. Many current and potential applications for such systems are consumer electronics, and accessibility aids for the elderly or disabled. However, the performance of a speech recognition interface can degrade significantly with presence of out-of-vocabulary (OOV) words. Efficient, accurate detection and rejection of these words is therefore required for robust interpretation of these words.

Present techniques for dealing with this problem can be loosely grouped into the categories (A) confidence measures and (B) utterance verification. Confidence measures (CM) assign a single confidence value to an utterance based on some heuristic or probabilistic algorithm. In isolated recognition systems only acoustic confidence measures can be created since there is typically no language information to exploit. By far the most common way of approaching this problem is by estimating the normalization probability from the Bayesian probability equation. Numerous approaches have attempted to approximate this probability using filler models[1], garbage models[2], on-line garbage models[3] and pseudo-filler models[4]. Additionally, heuristic confidence measures have been created using decoding statistics, or likelihood scores among other features[5].

The second approach, utterance verification, attempts to model the OOV utterance directly. Models for in-vocabulary (IV) and OOV words are used in a likelihood ratio test to optimally separate the two classes[6].

When only a small vocabulary is required, such as in a command interface, it is appropriate to use whole word models in the recognition system to save complexity and computation costs. As a result many filler models and utterance verification techniques are not applicable. In this work we construct an utterance verifier for whole word model systems using a new approach. We explore the use of CMs calculated from other less likely words in the N-Best list. Traditionally, a CM is calculated for only the top entry in the N-Best list. The extra CMs used in our approach can be calculated with very little additional cost since the decoded scores for the models have already been evaluated in the recognition stage. Present techniques have made use of scores from other HMMs to form a single CM for the top utterance [7]. However, to our knowledge no one has calculated CMs for these other words and explored their efficacy in an utterance verifier.

In the next section we outline the small vocabulary word recognizer used in this work. Following that, we present a new confidence measure in addition to two standard CMs that will be used in the proposed utterance verifier. We then develop the generalized utterance verifier that incorporates CMs from any number of words in the N-Best list. We then present the performance of the proposed verifier and compare it to the traditional single CM approach.

2. SPEECH RECOGNITION INTERFACE

An isolated speech recognizer was implemented using word-based HMMs based on the first 13 coefficients of an MFCC representation. Eight state HMMs were used as they provided an appropriate trade-off between recognition rate and computational complexity.

We used the TI-46 Word speech corpus for training and testing of the speech recognizer and utterance verifier. Specif-

The authors would like to acknowledge the support of NSERC.

ically, the TI-20 vocabulary subset was used, which contains 26 utterances of 20 words from 8 females and 8 males. The 20 words are all short command-like utterances similar to what might be used in a command control interface. Speaker independent models were trained using 5 utterances of each word from each individual. 16 utterances of each word from each individual were used for testing. Using the eight state models described above the speech recognizer achieved a recognition rate of 99.9% when tested on the same closed set of speakers. This recognition rate is comparable to existing small vocabulary recognition systems.

3. PROPOSED UTTERANCE VERIFIER

3.1. Confidence Measures

In this paper we make use of three confidence measures. The first confidence measure is our own proposed normalization procedure. This measure is obtained by dividing the top log-likelihood score (L_1) by the sum of the top N log-likelihood scores plus a filler or garbage model (L_{GB}). We refer to this CM here as the N-Best measure.

$$CM_{NBest} = \frac{L_1}{\sum_{i=1}^{V+1} L_i} \quad (1)$$

The garbage model was implemented similar to the one developed by Tsiporkova et al. [2] The garbage model is a two-state ergodic HMM. One state is used to model the average acoustic speech signal and the second state models the inter-word, start and end silences. The garbage model was trained with 2 of each word from all 16 speakers for a total of 640 utterances.

The second CM used is a standard approach based on the garbage model. It is produced by dividing the likelihood by the garbage model score. In terms of log-likelihoods this becomes.

$$CM_{GB} = L_1 - L_{GB} \quad (2)$$

Finally, we implement a pseudo-filler model similar to the ones proposed in as a secondary baseline measure. The pseudo-filler model is the average of the top N hypotheses scores. This average is then subtracted from the top model log-likelihood score as was done earlier[5][4].

$$CM_{NAVG} = L_1 - \frac{\sum_{i=1}^{N} L_i}{N} \quad (3)$$

Traditional OOV detection strategies involve calculating a single CM for only the top recognized model. This CM is then compared to a predetermined threshold. A single threshold can be set for all OOV words but traditionally word-dependant thresholds are used. In our standard single CM implementation word-dependent thresholds were obtained using a minimum classification error metric in training.

As was already mentioned, our own approach involves calculating the CMs for not only the top word but for all words included in the N-Best list. Therefore the CM Equations 1-3 which currently calculate the top CM are extended to calculate CMs for the remaining N words in the N-Best list. This is done by replacing L_1 by L_j where L_j is the log-likelihood score from the j^{th} word in the N-Best list. For example Equation 1 becomes the following.

$$CM_{NBest}(j) = \frac{L_j}{\sum_{i=1}^{V+1} L_i} \quad (4)$$

These CMs form a vector R which contains N CMs from the N words in the N-Best list.

$$R = \{CM_1, CM_2, ..., CM_N\} \quad (5)$$

We interpret the CM calculated for the top model as a measure of how well the utterance matches the model. We can interpret the other CMs from less likely words as measures of how well the utterance matches these other words. It is possible that the similarities or dissimilarities between the utterance and other words could provide discriminant information to aid the IV and OOV detection.

3.2. Utterance Verifier

Since we have more than a single CM a simple threshold can not be used. Therefore, the utterance verifier was constructed by considering the optimal separation between IV and OOV utterances based on the vector R. We define the conditional distribution for both IV and OOV classes as follows:

$$P(R|H_i) = P(R = \{CM_1, CM_2, ..., CM_N\}|H_i) \quad (6)$$

where H_0 is the outcome that the utterance is IV and H_1 the outcome that the utterance is OOV. These distributions are N dimensional where N is the number of CMs calculated from the N-Best list. Depending on how many entries we keep in the N-best list, the number of CMs we calculate can be varied to include only the first CM, as in the traditional approach, to all IV models. The nature of the distributions will vary depending on which CM is chosen. However it has been observed that for the CMs outlined above the distributions tend to be unimodal and relatively symmetric. It would seem reasonable to adopt a multivariate Gaussian approximation for the distribution given its simplicity and ability to capture at least the first and second moments. Therefore we approximate the distributions by multivariate Gaussians in the utterance verifier.

In order to separate the decision space optimally we want to minimize the Bayes risk associated with assigning utterances to each of the possible outcomes. In the binary case, we have four possible outcomes. The two outcomes

associated with correct detection of both IV and OOV utterances should not be penalized and so we only need to minimize the risk for the outcomes that produce errors. This occurs when the utterance is mislabelled (i.e. IV as OOV and vice versa). The risk equation is therefore:

$$\begin{aligned} \Re &= P_{OOV}\xi_{FA}\int_{z_0} p(R|H_1)dR \\ &+ P_{IV}\xi_M\int_{z_1} p(R|H_0)dR \end{aligned} \quad (7)$$

where ξ_M is the cost associated with a miss and ξ_{FA} is the cost associated with a false alarm. P_{OOV} is the a priori probability that the utterance will be OOV and P_{IV} is the a priori probability that the utterance will be IV. The total decision space z is divided between z_0 and z_1. How this space is divided determines the performance of the system. Classical Bayesian detection theory optimally separates the two regions and results in the following likelihood ratio.

$$\Lambda' = \frac{p(R|H_0)}{p(R|H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \frac{P_{OOV}\xi_{FA}}{P_{IV}\xi_M} = \tau \quad (8)$$

The conditional probabilities of the IV and OOV distributions are approximated by multivariate Gaussians and can be substituted into the likelihood ratio. The OOV distribution is modelled as a single Gaussian with a mean of m_{OOV} and a covariance C_{OOV}. The IV distribution is modelled as a Gaussian mixture, with a component for each IV word, and the mixture weights are equated to P_W. This yields the following detector which functions as our utterance verifier.

$$\Lambda \underset{H_1}{\overset{H_0}{\gtrless}} \frac{P_{OOV}\xi_{FA}}{P_{IV}P_w\xi_M} = \tau \quad (9)$$

where

$$\Lambda = \sum_{i=1}^{V} |C_i^{-1}C_{OOV}|^{1/2} exp[-1/2[D_i - D_{OOV}]] \quad (10)$$

and D is the Mahalanobis distance defined by

$$D = (R^T - m)C^{-1}(R - m) \quad (11)$$

4. RESULTS

In the following experiments we partitioned the test data from the TI-20 corpus into two halves. All of the data was processed by the isolated word recognizer and the corresponding likelihoods scores used to calculated the vector of confidence measures R. The first half of the CMs were used to estimate the IV and OOV distributions. The second half of CMs were used to evaluate the utterance verifier performance.

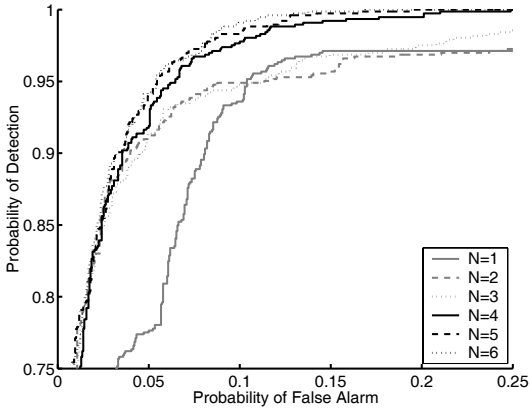

Fig. 1. ROC curves for proposed utterance verifier on a fixed vocabulary of six words

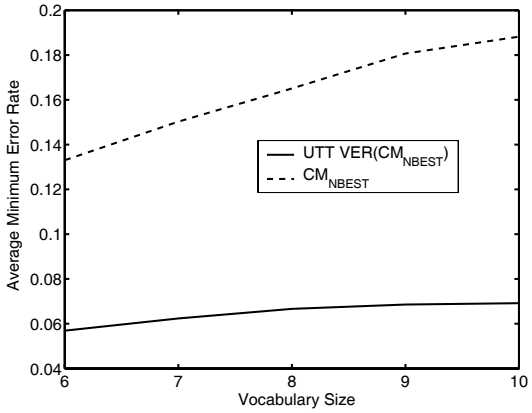

Fig. 2. Average minimum error rates for single CM approach and the proposed utterance verifier approach

Initially the utterance verifier was evaluated for a fixed vocabulary of 6 words. The N-Best CM was used and evaluated for the top model. Additionally the number of extra CMs were varied between 0-5 as is shown in Figure 1. This plot shows the ROC curves for the proposed utterance verifier using the N-Best CM with the number of extra CMs varied between 0-5. We can see that as the number of extra CMs is increased the performance of the verifier improves. From a single CM to all 6 available CMs we find a 5% reduction in the minimum error rate. This is not unexpected given that the additional information contributed by the extra CMs cannot impair the detection performance – a result that is easily proved mathematically.

4.1. Vocabulary Variation

In order to accurately assess the validity of the proposed technique with a small vocabulary set the following steps were taken. The number of IV words was varied between 6 and 10 to examine the effect of vocabulary size. Additionally 20 different randomly selected vocabulary combinations were tested to eliminate effects of a specific vocabulary combination. For these experiments the minimum error operating point was used as a measure of the performance. The minimum error operating point is found by equating the cost terms Equation 9. Using the same CM the standard implementation is compared to the utterance verification implementation where the maximum number of extra CMs are included. Figure 2 displays the minimum error rate of the two approaches for vocabulary sizes between 6 and 10. The error rate displayed is the average error rate for the 20 randomized vocabulary configuration. We can see in this figure that the new utterance verifier outperforms the single CM approach by around 16% over the various vocabulary sizes tested.

Similar results were obtained with the garbage model and pseudo-filler model CMs. Figure 3 shows the results obtained using the utterance verifier with the garbage model CMs and the standard garbage model CM. We see that the minimum error rate is reduced by around 3%. Figure 3 also shows similar results when the utterance verifier is implemented with the pseudo-filler CM. In this case the improvement in minimum error rate is around 18%. We can also see that both utterance verifiers outperformed the single CM approaches.

One of the assumptions of the Bayesian approach is that all the probabilities are known a priori. We have of course used a Gaussian approximation and so we should not expect the verifier to operate at the true optimal point. The deviation from optimality was investigated by comparing the experimentally obtained minimum error operating point to the apex of the ROC curves. On average these two error rates were found to differ by less than 0.5% when the N-Best CM is used. Therefore the Gaussian approximation does not significantly alter the optimal separation between the IV and OOV classes.

5. CONCLUSION

A new utterance verifier was proposed and explored to improve the detection of OOVs. The extra CMs used in this technique require very little added computational cost. Three types of CMs were used in this verifier including a new N-Best measure, a garbage model and a pseudo-filler model. The new verifier resulted in a 16, 3, and 18% improvement, respectively, over the standard single CM implementation. Future work on this method will include investigating the combination of different types of CMs and its use for verification in joint speech-speaker interfaces.

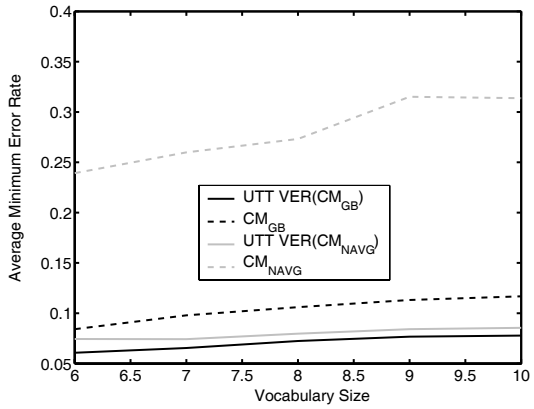

Fig. 3. Average minimum error rates of single CM approach and proposed utterance verifier approach for CM_{GB} and CM_{NAVG}

6. REFERENCES

[1] T.J. Hazen and I. Bazzi, "A comparison and combination of methods for oov word dectection and word confidence scoring," ICASSP, 2001, vol. 1, pp. 397–400, IEEE.

[2] E. Tsiporkova, F. Vanpoucke, and H. Van hamme, "Evaluation of various confidence-based strategies for isolated word rejection," ICASSP, 2000, vol. 3, pp. 1819–1822, IEEE.

[3] J. Caminero, C. de la Torre, L. Villarrubia, C. Martin, and L. Hernandez, "On-line garbage modelling with discriminant analysis for utterance verification," ICSLP, 1996.

[4] G. Hernandez-Abrego, X. Menendez-Pidal, and L. Olorenshaw, "Robust and efficient confidence measure for isolated command recognition," ICASSP, 2002, pp. 449–452, IEEE.

[5] A. Wendemuth, G. Rose, and J.G.A. Dolfing, "Advances in confidence measures for large vocabulary," ICASSP, 1999, vol. 2, pp. 705–708, IEEE.

[6] E. Lleida and R. Rose, "Utterance verification in continuous speech recognition: Decoding and training procedures," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 126–139, 2000.

[7] A. Gunawardana, "Word-based acoustic confidence measures for large-vocabulary speech recognition," ICSLP, 1998, vol. 3, pp. 791–794.