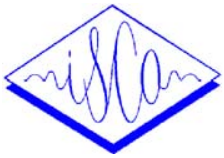

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

USE OF RECURSIVE MUMBLE MODELS FOR CONFIDENCE MEASURING

Qiguang Lin, David Lubensky, and Salim Roukos

IBM T.J. Watson Research Center
Computer Science Department
P. O. Box 218, Yorktown Heights, NY 10598, USA
qlin@us.ibm.com

ABSTRACT

In many speech recognition applications such as name dialing, it is necessary to have the ability to know when a recognition error has occurred so that undesired or unpredicted system behavior can be minimized. Confidence measure is usually used for detection of probable errors. In this paper, a new method for measuring confidence is presented. The method is based on use of recursive mumble models. During a regular decoding from which word hypotheses and word boundaries are known, the score of recursive mumble models is then determined. The (weighted) difference between the word detail-match score and the mumble score is used as the confidence measure. It is next compared to a predefined threshold to decide whether the decoded result is confidently correct or not. The method has been evaluated with two different databases. The results show that the new method outperforms our previous method solely based on the word detail-match scores. In particular, the results show that the new method is able to reduce the equal error rate from 32% to 23% and that it rejects far more (78% versus 35%) out-of-domain sentences at the fixed 5% false rejection rate.

Keywords: confidence measure, mumble model, equal error rate

10.21437/Eurospeech.1999-17

1. INTRODUCTION

Current speech recognition systems are not flawless. They are always associated with a certain amount of error rates. This is especially the case in real-world deployments of speech recognition technology such as telephony name dialing. Although the errors cannot be eliminated completely, for many applications it is desired to know when an error is likely to have occurred so that subsequent action for the uncertain words may be taken. Depending upon actual applications, the action may include: (1) to reject the word in a voice-operated system (such as command/control navigation) or to ask the user to repeat/confirm such that unpredicted behavior of the system can be minimized; (2) to exclude the uncertain words from being utilized in unsupervised speaker/environment adaptation; (3) to highlight uncertain words in a dictation scenario for easy location and repair of the errors.

Many different approaches have been proposed for confidence measure. Generally either a filler model or an anti-word (or anti-subword) is used. The difference lies in the fact whether the model has been discriminatively trained. Most of these approaches belong to the category of post-processing. Namely, the confidence is usually measured for hypothesized words following standard decoding [2, 5, 6, 8, 9, 10]. In [7], however, confidence measure is incorporated into search strategy and improved recognition performance is reported. In addition, instead of using a fixed threshold against which the confidence measure is compared, efforts have been made to use adaptive thresholding [4].

We have previously suggested three methods for measuring confidence [5, 6]. The first one is solely based on word detail-match (DM) score. If it is above a user-define threshold, the word is deemed to have been correctly decoded, and hence accepted. Otherwise the hypothesized word is rejected. The second method is to combine finite-state-grammar with trigrams in a parallel decoding. A task-specific grammar is constructed to cover all key-phrases, while the trigram statistics is estimated from some general-purpose corpora. When the grammar is triggered and successfully completed, the corresponding key phrase is deemed to have confidently detected. The trigrams work as a filler model to handle non-key phrases present in the conversation. The third method is based on rank-ordering subphone likelihood scores. The rank values are first derived at the subphone levels, then merged to the counterpart at the phone, word, and utterance levels. The method also incorporated selective weighting which deemphasizes contributions of phones known to have large acoustic variations and upper-bound limiting to ensure a rank computation not to be dramatically affected by a very bad segment.

In this paper, we present another method for measuring confidence. The method is based on use of recursive mumble models. During a regular decoding from which word hypotheses and word boundaries are known, the score of recursive mumble models is then determined. The (weighted) difference between the word detail-match score and the mumble score is used as the confidence measure. It is next compared to a predefined threshold to decide whether the decoded result is confi-

dently correct or not. The method has been evaluated with two different databases. The results show that the new method outperforms our previous method solely based on the word detail-match scores. In particular, the results show that the new method is able to reduce the equal error rate from 30% to 22% and that it rejects far more (78% versus 35%) out-of-domain sentences at the fixed 5% false rejection rate.

2. THE ALGORITHM

In this section, we first outline the recognizer used in the present study. We then describe in some detail how to compute mumble scores and how to determine confidence measure from given word DM scores and recursive mumble scores.

2.1. Recognizer

The recognizer used in this study is the IBM stack decoder [1]. Its front-end process consists of (i) preemphasis; (ii) computing FFT spectra every 10 ms using a 25 ms Hamming window; (iii) converting the mel-band output of the spectra to 12-dimensional cepstral coefficients, MFCC's. (In some experiment only 9 cepstrum coefficients are used.); (iv) removing sentence-wise means of MFCC's and normalizing the energy term C_0; (v) computing first-order and second-order derivatives of MFCC's. The acoustic modeling includes 52 context-independent (CI) phones. Each of the phones are modeled with 3 left-to-right HMM states which are context-dependently (CD) trained. These states, representing subphone units, are actually terminal leaves of a decision tree. A speaker independent, large vocabulary speech recognition system typically contains 2,000 to 4,000 leaves. Each of the leaves is associated with a rank distribution histogram estimated from the training speech data. During recognition likelihood scores are computed using the histograms (cf. [1]).

2.2. Recursive Mumbles

Figure 1 depicts the network of the present recursive mumble model. Each path pertains to a context-independent phone also having 3 left-to-right HMM states. A backward loop, associated with a transition probability Pr, is provided to enable revisiting the network. The output of the network is a sequence of phones, as exemplified in Table 1.

As expected, the output resembles the baseform of the underlying word in some cases (e.g., row 1 of Table 1). It can also be seen from Table 1 that a small back transition probability Pr tends to produce fewer phones while a large Pr tends to insert additional phones.

2.3. Compute Mumble Scores

From a stack decoder, word boundaries and word scores (fast match score, detail match score, and overall likelihood) are readily available. As mentioned above, word

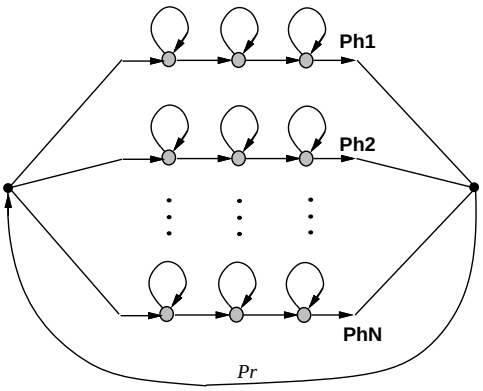

Figure 1. *Network representation of recursive mumble models.*

DM scores have been used as confidence measure in previous studies [5]:

$$CM_1 = \frac{1}{N}\sum_{n=1}^{N} D_{W_n} \qquad (1)$$

where N is the total number of nonsilence words in the utterance, and D_{W_n} is the DM score of the nonsilence word W_n. Eq.(1) will be referred to as the baseline in the following discussion.

Without better alternatives, we assume that the word boundaries estimated by the decoder is correct for the hypothesized word. Let us denote the word starts at time $T1$ and ends at $T2$. Over the segment $T2 - T1$, it is straightforward to compute the score of recursive mumble model using Viterbi algorithms. Different from conventional Viterbi search, the path does not necessarily begin at the first state $s1$ of the lattice representation in Figure 2, or terminate at the last state, sn.

Context-dependent acoustic prototypes are used to compute emission probabilities given an observation vector. The context-independency of mumble models is then achieved via a maximum-taking operation. This arrangement has the advantage that separate training of mumble models is avoided.

It is reasonable to assume that there is no silence token present over the $T2 - T1$ segment. Therefore, a feature is added to the recursive mumble models as that it may include or exclude silence phones in the network. Furthermore, a different probability can be given to the backward transition to the same phone, if desired. Otherwise, the mumble model is similar to that of [3].

We calculate the mumble model score right before new extension is to be made. This implementation is motivated by our plan to incorporate the score into the search strategy, as in [7]. For utterance verification applications, the mumble model score may also be computed only for the best hypothesized path. The latter significantly reduces the computational overhead.

	back transition probability	word	baseform	network output
1	0.02	printing	P R IH N T IX NG	P R IY N T IY N
2	0.02	shirts	SH ER TS	SH R TS Z
3	0.001	printing	P R IH N T IX NG	R IY D IY N
4	0.001	shirts	SH ER TS	SH R TS

Table 1. *Examples of network outputs.*

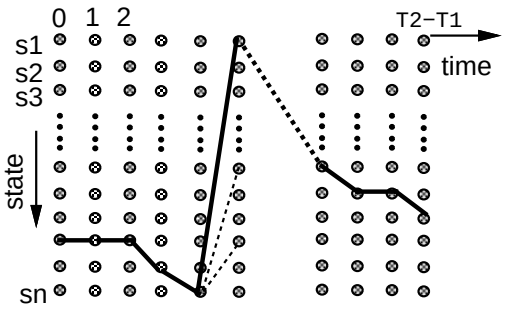

Figure 2. *Lattice representation for computing mumble scores.*

At the word level, the new confidence measure is:

$$CM_2^w = D_{W_n} - M_{W_n} \quad (2)$$

where M_{W_n} is the corresponding mumble model score for the given word W_n. At the utterance level, averaged CM_2^w augmented with the minimum CM_2^w is used:

$$CM_2 = \frac{1}{N}\sum_{n=1}^{N} CM_2^w + k \cdot min\,\{CM_2^w\} \quad (3)$$

where k denotes a weighting constant.

3. EXPERIMENTAL RESULTS

In this section, the proposed method for measuring confidence is evaluated using two telephony databases.

3.1. Experiment 1

This experiment is a telephony yellow-page application. The corpus contains 409 speakers and 2017 utterances. On average, each sentence has two words. There are approximately 5100 vocabulary words. The language model is based on trigram statistics. The word error rate is 6.84% and the sentence error rate is 9.42%.

Figure 3 compares the ROC curves between the new method and the method which uses only the word DM scores. The solid curves denote false acceptance and the broken curves denote false rejection. The intersection point between the curves denotes the "equal error rate" or ERR. Although the system performance can hardly assessed by comparing the ERR and the system usually does not operate at that particular point, a single quantity of ERR is easy to use and is able to indicate relative superiority between different systems. From Figure 3, it is shown that the new method reduces the ERR from 31.8% to 23.3%, a reduction of 26.7%. The following

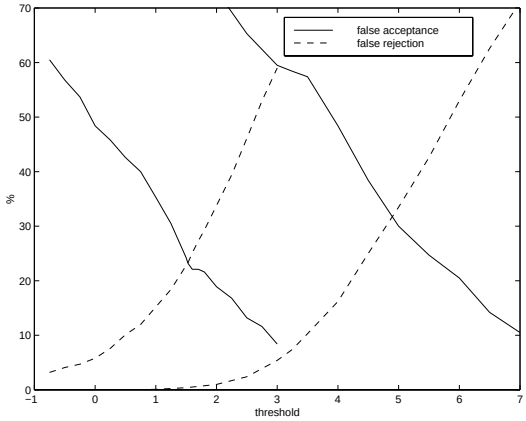

Figure 3. *ROC curves. Left: new method; Right: baseline. Yellow-page experiment.*

list shows how the ERR reduces when more features are added to the confidence measure:

- Use of recursive mumble models renders an ERR of 24.9%
- Excluding silence token in the recursive mumble models (Figure 1) reduces the ERR to 24.4%.
- Combining the averaged and minimum score, Eq. (3) (with a weight of $k = 0.2$) further lowers the ERR to 23.3%

3.2. Experiment 2

This experiment corresponds to a stocks price-quotes application. There are about 1000 queries with 60 being out-of-domain. The vocabulary has over 25000 words. A finite-state-grammar is used as the language model. The sentence error rate for in-domain queries is 11%.

From Figure 4, it is apparent that the new method significantly reduces the ERR when compared with the baseline. The reduction pattern is similar to that observed in Figure 3.

Figure 5 depicts the ROC curve in a different way, allowing more detailed comparisons between systems than the single quantity of ERR. The closer the ROC curve is to the origin, the better is the system. It is seen that the new method shifts the ROC curve toward the origin substantially, and that the shift is most around the ERR. If a more realistic point of operation is chosen, say, at a 5% false rejection rate, Figure 5 shows that a

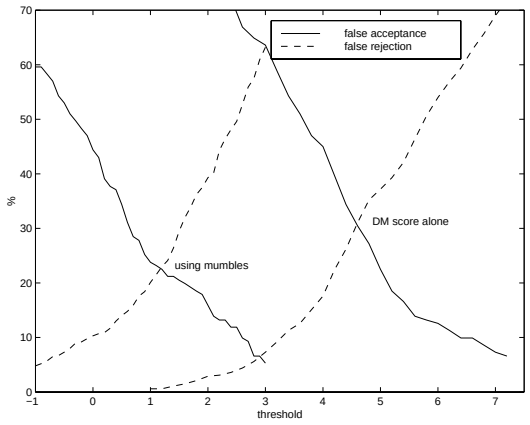

Figure 4. *ROC curves. Left: new method; Right: baseline. Price-quotes experiment.*

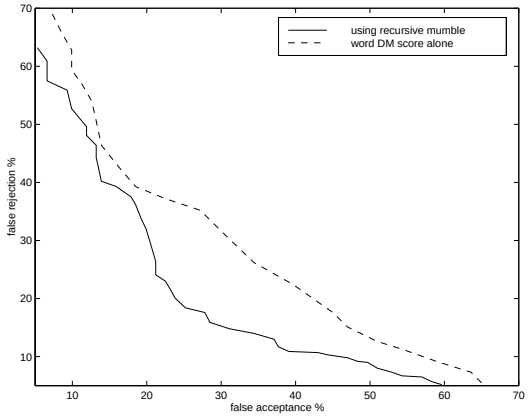

Figure 5. *ROC curves illustrating false rejection (type I error) as a function of false acceptance (type II error).*

reduction of 8% in false acceptance rate is achieved by the new method. At first, the reduction may not appear to be substantial. However, if we focus on those 60 OOD sentences, the new method rejects far more OOD sentences than the baseline does, see Table 2.

4. CONCLUSION

A new confidence measure has been proposed in the above and it has been favorably evaluated with experiments on two telephony speech databases. Similar improvement by the new method over the baseline has been observed for the both testsets. The improvement appears to be most in the vicinity of the ERR. For the yellow-page experiment, see Figure 3, a reduction of 26.7% in ERR is obtained. For other operation points, improvement by the new method is not as substantial. For example in Figure 5, the new method renders a reduction of 8% in the false acceptance rate at the 5% false rejection rate. However, the new method is effective in rejecting out-of-domain sentences, as illustrated in Table 2.

method	% rejected OOD sentences
baseline	35.0% (21/60)
new method	78.3% (47/60)

Table 2. *Rejection of OOD sentences.*

In this paper, the recursive mumble model score is used to determine whether an utterance has been correctly decoded or not, so as to decide whether to accept or reject the decoded result. To further improve our confidence measure, we plan to integrate our rank-based method [5] with the present method, to add discriminative training, and to incorporate various scores (language model score, fast match score, etc.). As mentioned earlier, our implementation allows the incorporation of the mumble model score into the search strategy. One might not need a sophisticated confidence measure as the recognition accuracy elevates.

REFERENCES

[1] Bahl, L., Souza, P., Gopalakrishnan, P., Nahamoo, D., Picheny, M., "Robust methods for using context-dependent features and models in a continuous speech recognizer," *ICASSP94*, pp. 533-536.

[2] Chase, L.: "Word and acoustic confidence annotation for large vocabulary speech recognition," *Eurospeech 1997*, pp. 815-818, Greece.

[3] Dharanipragada, S. and Roukos, S.: "A fast vocabulary independent algorithm for spotting words in speech," *IEEE-ICASSP98,* pp. 233-236, Seattle.

[4] Gupta, S. and Soong F.: "Improved utterance rejection using length dependent thresholds," *ICSLP 1998,* Sidney.

[5] Lin, Q., Lubensky, D., Picheny, M., and Rao, S.: "Key-phrase spotting using an integrated language model of N-grams and finite-state grammar", *Eurospeech 1997,* pp. 255-258 Greece.

[6] Lin, Q., Das, S., Lubensky, D., and Picheny, M.: "A new confidence measure based on rank-ordering subphone scores", *ICSLP 1998,* Sidney.

[7] Neti, C., Roukos, S., and Eide, E. "Confidence measures as a guide for stack search in speech recognition," *ICASSP96*, pp. 883-887, Germany.

[8] Rahim, M., Lee, C.-H., Juang, B.-H.: "Discriminative utterance verification for connected digits recognition", *IEEE-Trans SAP. 5*, pp. 266-277, 1997.

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

[10] Wilpon, J., Lee, C.-H., and Rabiner, L.: "Application of Hidden Markov models for recognition of a limited set of words in unconstrained speech," *ICASSP89*, pp. 254-257. Scotland.