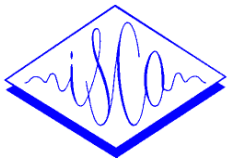
ISCA Archive
http://www.isca-speech.org/archive



International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

CONFIDENCE MEASURES FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Ping LV, Zuo-Ying WANG, Da-Jin LU

Department of Electronic Engineering,
Tsinghua University, Beijing
luping@thsp.ee.tsinghua.edu.cn

ABSTRACT

Estimation confidence of the output hypothesis of a speech recognizer can be used in many practical applications of speech recognition technology. In this paper, we propose to estimate the confidence of a hypothesized word directly as its posterior probability. There are tow methods to calculate posterior probabilities, one of which is based on word graphs and the other is based on N-best lists. We present experiment results on database provided by China National 863. And we also use confidence measures in unsupervised speaker adaptation.

1. INTRODUCTION

Confidence measures can be used to label individual words in the output of the speech recognition system with either correct or incorrect, thus enabling the system to spot the position of possible errors in the output automatically. A reliable measure of the confidence of a speech recognizer's output is useful in many circumstances, including supervised and unsupervised adaptation, recognition error rejection, out-of-vocabulary word detection and key word spotting.

Previous confidence measures have been mainly used in key word spotting and out-of-vocabulary word detection. The computation of confidence measures was based on acoustic score [1] or word lattices [2] and n-best lists [3]. [4] used generalized linear models for relating a confidence feature vector to calculate the probability of a word to be correct. [5] reported improvements of confidence performance based on the use of language model. In this paper we interpret confidence as posterior probabilities for hypotheses in the word graph. The posterior probability of a certain word hypotheses in sentence is estimated by summing up the posterior probabilities of all sentences, which contain this word at certain position.

Because of the special structure and various characteristic features of Mandarin Chinese [6], the speech recognition system is very different from that of the western language. For sake of ease of interpretation, the speech recognition system especially the acoustic model is first introduced in section 2 and the basic idea of confidence measure is described in section 3. Section 4 gives the experimental results and the conclusions are drawn in section 6.

2. DURATION DISTRIBUTION BASED HMM

Our system is a two-pass Mandarin speech recognition system. In first pass, the input speech data pass through the acoustic recognizer; the outputs are the multi-length multi-candidates spelling lattices. In second pass, the spelling lattices are conversed to Chinese characters by language decoder. The acoustic model used in system is a modified HMM called the duration distribution based hidden Markov model (DDBHMM)[7]. The language model is usually N-gram language model.

The DDBHMM is an inhomogeneous HMM which is based on the fact that the state duration distribution is relatively stationary. In DDBHMM, the duration distribution probability of the state is used instead of the state transition probability, which is used in the classical HMM.

$$\begin{aligned} W^* &= \arg\max_W P(X \mid W)P(W) \\ &= \arg\max_W \left\{ \max_{S_2,\cdots,S_N} \left[\prod_{i=1}^{N} P_i(\tau_i) \prod_{t=S_i+1}^{S_{i+1}} b_i(x_t) \right] \cdot P(W) \right\} \end{aligned} \quad (1)$$

3. CONFIDENCE METRICS

Current the speech recognizer, which achieves the expected minimum word recognition error rate, is the following maximum a posterior (MAP) decoder:

$$\begin{aligned} \hat{W} &= \arg\max_W P(W \mid X) \\ &= \arg\max_W \frac{P(X \mid W)P(W)}{P(X)} = \arg\max_W P(X \mid W)P(W) \end{aligned} \quad (2)$$

where, $\hat{W}$ is the recognition result.

$P(X \mid W)$ presents the probability of the acoustic sequence given a word string.

$P(W)$ presents the prior probability of the word string.

$P(X)$ represents the prior probability of the acoustic sequence.

Since $P(X)$ is the same for all utterances in a time synchronous decoding. Hence, the estimation of $P(X)$ will not change the relative order of word string hypotheses output by a speech recognizer. So the recognizer computes only $P(X \mid W)P(W)$. We know which utterance is most likely, but don't really know how good goodness of match and have no real means for evaluating accuracy of output word strings. Hence the likelihood of decoding of different utterances is not comparable and so not useful for building confidence-measure.

However we can use posterior word probability as a measure of confidence. The estimation of $P(X)$ is become main problem for the computation confidence.

3.1 Confidence based on word graph

Define the posterior word hypothesis probability of a word hypothesis w with starting and end time t_s and t_e is as the following:

$$P(W \mid X) = \frac{\sum_{W_s}\sum_{W_e} P(X \mid W_s, w, W_e) P(W_s, w, W_e)}{P(X)} \quad (3)$$

$$\begin{aligned} P(X) &= \sum_{W} P(X \mid W) P(W) \\ &= \sum_{w}\sum_{W_s}\sum_{W_e} P(X \mid W_s, w, W_e) P(W_s, w, W_e) \end{aligned} \quad (4)$$

where, W_s w W_e composed the complete word string W; W_s denotes all word hypothesis sequences proceeding w and W_e denotes all those succeeding w.

Our recognition system is a two-pass Mandarin speech recognition system. In first pass, the input speech data pass through the acoustic recognizer; the outputs are the multi-length multi-candidates spelling lattices. The acoustic model used in system is a modified HMM called the duration distribution based hidden Markov model (DDBHMM)[]. In second pass, the multi-length multi-candidate spelling lattices combine N-gram language prior probability to construct word graphs. The words of the optimum path in the word graph are recognized results.

Based on word graph, we can compute posterior probability of individual word hypothesis by standard forward-backward algorithm. The numerator of (3) can be computed as the summation probabilities of these paths that through the word w. (4) can be computed as the summation probabilities of all paths in the word graph.

3.2 Confidence based on N-best lists

We also study the estimation of posterior probabilities on N-best lists instead of word graphs. N-best lists are the N best sentence hypotheses and can be constructed on the basis of word graphs. Then denominator of (2) becomes the sum of the N-best path probabilities, the numerator of (2) becomes the sum of path probabilities in which word hypothesis w with starting and end time t_s and t_e .

3.3 Evaluation of Confidence Measures

In our experiments, the confidence is later on compare with a tagging threshold optimized on a cross validation corpus beforehand. Once the confidence has been computed, each word of the recognized sentence is simply tagged as either correct or false, depending on whether its confidence exceeds this threshold or not. Here, two different types of errors can occur. The first is a false acceptance, i.e. a false word is tagged as correct, and the second is a false rejection, i.e. a correct word is tagged as false. Obviously, there is a trade-off between the two types of errors, depending on the choice of the tagging threshold.

In this paper, confidence performance is measured in terms of a figure of merit (FOM) and the confidence error rate (CRE). The confidence error rate (CRE) is defined as the number of incorrectly assigned tags divided by the total number of recognized words. The baseline CER is given by the number of insertions and substitutions, divided by the number of recognized words.

In addition, we compare posterior probability based on word graph and N-best list with the acoustic confidence measure.

4. UNSUPERVISED ADAPTATION

In unsupervised adaptation, training utterances have not corresponding correct transcription. On the contrary, the recognition results of these training utterances are used as transcription usually. However, there always have some errors in the recognition results.

We can use confidence metrics to guide the adaptation process by selecting or emphasizing speech segments with high confidence. In our experiments, we reject all segments that correspond to word whose confidence is below a certain threshold.

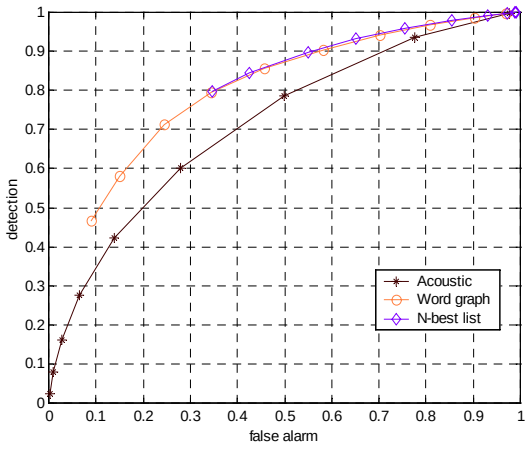
5. EXPERIMENTS

The baseline system is a speaker independent continuous density DDBHMM system. The speech was parameterized into a 45 dimensional feature vector that includes 14 MFCCs, the normalized log energy and the first and second differences of these parameters. The resulting system has approximately 28,000 Gaussians. We carried out experiments on database provided by National 863 Hi-Tech Project for large vocabulary continuous speech recognition.

For all of the following experiments the tagging threshold are optimized on a cross validation corpus.

We first investigate the performance of confidence in term of FOM and CER. We compare posterior probability based on word graph and N-best list with the acoustic confidence measure.

Figure 1: FOM of Confidence Measure



Form Figure 1, we can see the performances of confidence based on word graph and N-best list are preceded that of acoustic confidence. We can get same conclusion from Table 1.

Table 1: Confidence Error Rate

Baseline	Acoustic	Word graph	N-best list
49.62%	33.44%	18.82%	18.43%

The test data of unsupervised adaptation comprises of 10 sentences each from 6 speakers.

We apply the three confidence measures, which are posterior probability based on word graph, N-best list and acoustic score respectively, on unsupervised speaker adaptation. The adaptation algorithm is MLLR. The performance of unsupervised adaptation based on confidence measures is compared with that of unsupervised self-adaptation. These experiment results are shown in Table 2.

Table 2: The recognition error rate on different unsupervised adaptations

Baseline	45.62%
Unsup	37.55%
Acoustic confidence + unsup	39.06%
Word graph confidence + unsup	36.64%
N-best list confidence + unsup	36.83%

Confidence based on word graph and N-best lists improve the performance of adaptation than just unsupervised self-adaptation. But acoustic confidence reduces the effect of adaptation.

6. CONCLUSION

In this paper, confidence measures based on word graphs and N-best lists are presented and compared. Experimental results show that posterior probabilities outperform the conventional acoustic confidence measure. And confidence based on word graph has best performance either in the merit of confidence or the effect of unsupervised adaptation.

7. REFERENCES

[1] S. Cox, R. Rose. “Confidence Measures for the Switchboard Database,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Atlanta, May 1996.

[2] T. Schaaf, T. Kemp. “Estimating Confidence Using Word Lattices,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Munich, Germany, April 1997.

[3] M. Weintraub, F.Beaufays, Z. Rivlin, Y.Konig, and A. Stolcke. “Neural-network based measures of confidence for word recognition,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Munich, Germany, April 1997.

[4] L. Gillick, Y.Ito, and J.Yong. “A probabilistic approach to confidence measure estimation and evaluation,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Munich, Germany, April 1997.

[5] M. Weintraub. “ LVCSR Log-likelihood Ratio Scoring For Keyword Spotting,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Detroit, Michigan, May 1995.

[6] Hsin-min Wang, Jia-lin Shen, Yen-ju Yang, and Lin-shan Lee. “Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Detroit, Michigan, May 1995.

[7] WANG zuoying. “Duration distribution based HMM for speech recognition,” *National Chinese Characters Speech Recognition Conference-89. Chinese Info Proc Ass of China.* 1989. (in Chinese).