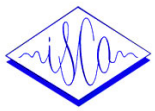

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

A COMPARATIVE STUDY ON VARIOUS CONFIDENCE MEASURES IN LARGE VOCABULARY SPEECH RECOGNITION

Gang Guo[2*], *Chao Huang*[1], *Hui Jiang*[3*] *and Ren-Hua Wang*[2]
[1]Microsoft Research Asia, 5F, Beijing Sigma Center, Beijing, China
[2]University of Science and Technology of China, Hefei, Anhui, China
[3]Dept. of Computer Science, York University, Toronto, Ont. M3J 1P3, Canada
paulguo@ustc.edu, chaoh@microsoft.com, hj@cs.yorku.ca and rhw@ustc.edu.cn

ABSTRACT

In this paper, we have conducted a comparative study on several confidence measures (CMs) for large vocabulary speech recognition. Firstly, we propose a novel high-level CM that is based on the inter-word mutual information (MI). Secondly, we experimentally investigate several popular low-level CMs, such as word posterior probabilities, N-best counting, Likelihood Ratio Testing (LRT), etc. Finally, we have studied a simple linear interpolation strategy to combine the best low-level CMs with the best high-level CMs. All of these CMs are examined in two large vocabulary ASR tasks, namely the Switchboard task and a mandarin dictation task, to verify the recognition errors in baseline recognition systems. Experimental results show: 1) the proposed MI-based CMs greatly surpass another existing high-level CMs which are based on the LSA technique; 2) Among all low-level CMs, word posteriori probabilities give the best verification performance; 3) When combining the word posteriori probabilities with the MI-based CMs, the equal error rate is reduced from 24.4% to 23.9% in the Switchboard task and from 17.5% to 16.2% in the mandarin dictation task.

1. INTRODUCTION

In an automatic speech recognition (ASR) system, we inevitably have a variety of recognition errors no matter how much we improve the technology. Thus, it becomes a very important issue how to make a better use of the error-prone ASR outputs if we want to design a successful ASR system for any real-world application. This requires an ASR system to automatically assess reliability of every recognized word or utterance. Apparently, one direct way to achieve this is to compute a score (preferably between 0 and 1), called confidence measure (CM), to indicate reliability of any recognition decision made during the ASR process. It is well known that good CMs will largely benefit a variety of ASR applications, e.g., to smartly reject non-speech noises, detect/reject out-of-vocabulary (OOV) words, detect/correct some potential recognition mistakes, clean up human transcription errors in a large speech database, guide the system to perform unsupervised learning, provide side information to assist high-level speech understanding and dialogue management, as so on.

In large-vocabulary speech recognition, confidence measures (CMs) can be calculated based on two levels of information. In the first level, which we call the low-level CMs, the CMs can be computed by using the similar information sources as that used in the recognition process, e.g., acoustics and language models. Generally speaking, the low-level CMs can be classified into three categories as described in [3]. The first one is to calculate CMs as a combination of various predictor features. Usually, these predictor features are derived from information collected in the recognition process. The second CM is the posterior probability as indicated in the maximum *a posteriori* (MAP) decision rule. Among many approximate methods, the word posteriori probability derived from a word graph based on the forward-backward algorithm is the most effective method. The third method is to formulate confidence measuring as a Statistical Hypothesis Testing problem, called utterance verification (UV), and calculate CMs based on the likelihood ratio testing (LRT). Besides, in the second level, which we call the high-level CM, the CMs can be calculated from the information sources beyond the acoustics and language models. The basic motivation for high-level CMs is that human can easily identify a certain portion of recognition errors in recognizer outputs on purely semantic grounds [2] in many large vocabulary speech recognition tasks since many misrecognized words tend to be "semantically incoherent" in the context. In [2], the authors propose to use latent semantic analysis (LSA) to annotate confidence scores for recognized words.

In this paper, we first investigate how to calculate the above-mentioned high-level CMs for large vocabulary speech recognition by using the high-level information sources alone. Based on the observation that many mis-

[*] This work has been done when this author was visiting MSR Asia.

recognized words are "semantically" incoherent with other surrounding words in the context, as opposed to LSA in [2], we propose to use the inter-word mutual information (MI) as another high-level confidence measure. From the transcriptions of all training data, we estimate mutual information between each pair of words in vocabulary. To solve the data sparsity problem, we also propose to use a simple smoothing technique for those word pairs where the mutual information cannot be reliably estimated. Experimental results in the Switchboard task and a mandarin dictation task clearly show that the proposed MI-based CMs yield much better verification performance when compared with the LSA-based CM technique. Secondly, we also experimentally study several popular low-level CMs, such as word-graph-based word posteriori probability, counts in N-Best list, LRT, etc. Finally, we examine a simple linear interpolation strategy to combine the best low-level CMs, i.e., word posteriori probability, with the best high-level CMs, i.e., inter-word mutual information. Experiments show that the combined CMs give better performance than either of them and consistently outperform the best single CM of using the word posteriori probabilities in both Switchboard and the mandarin dictation tasks.

2. HIGH-LEVEL CONFIDENCE MEASURES

2.1. The Existing Approach Based on LSA

Latent semantic analysis (LSA) is based on the assumption that words co-occurring across documents are semantically linked (see [6] for details). In LSA, each word is represented by a low-dimension vector in a reduced subspace, which is derived from the SVD of the word-document co-occurrence matrix. Therefore, "semantic coherence" between any two words can be calculated as the cosine of the angle between the two vectors representing the two words in the reduced space. In speech community, LSA has been applied to language modeling as in [6]. Besides, in [2], the author first applies LSA to confidence measure estimation, where three different methods are proposed to derive CMs from LSA scores. In this paper, we adopt the so-called *PSS* as our standard way to combine LSA scores:

$$PSS(w_i) = \prod \Pr_{w_i}(L \le K(u_i, u_j)) \quad (1)$$

where u_i and u_j specify the representing vectors for words w_i and w_j and $K(u_i, u_j)$ denotes cosine of the angle between u_i and u_j. In [2], the authors approximate the each distribution $Pr_{wi}()$ by five component Gaussian mixtures while we approximate the distributions by cumulative distribution functions.

2.2. The Novel CMs based on Mutual Information

2.2.1 Inter-word Mutual information as CM

Assume $N(x,y)$ being the co-occurrence times of word x and word y in all training documents, the joint probability of co-occurrence of word x and word y in any a document is calculated as

$$P(x,y) = \frac{N(x,y)}{\sum_{x,y} N(x,y)} \quad (2)$$

The marginal probabilities of co-occurrence of word x and word y in a document is calculated as follows:

$$P(x) = \sum_{y} P(x,y) \quad (3)$$

$$P(y) = \sum_{x} P(x,y) \quad (4)$$

Thus, according to the definition, mutual information between any two words x and y can be calculated as follows:

$$MI = \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (5)$$

In this approach, CM of each recognized word is calculated as the average mutual information of this word with all other recognized words in the same document, i.e., one sentence in this work. The above inter-word mutual information has been used for word (document) clustering, etc. To our best knowledge, this paper is the first effort to apply mutual information to the confidence measuring in speech recognition.

2.2.2 Smoothing

Due to the inherent sparseness of training document, mutual information between many word pairs may not be estimated reliably; therefore a smoothing technique is necessary to estimate the full MI matrix among all words. In this paper, we adopt a discounting strategy to smooth word co-occurrence counts as in n-gram language model training. Our smoothing method is operated in the following two steps:

Firstly, we derive the discounted co-occurrence times between any two words x and y as:

$$N'(x,y) = N(x,y) + C \quad (6)$$

where C is a constant to discount word frequencies.

Secondly, the smoothed jointed probability of any two words x and y is calculated based on a linear interpolation as follows:

$$P_{smooth}(x,y) = \frac{p(x,y) + \alpha \cdot p(x) \cdot p(y)}{1+\alpha} \quad (7)$$

Where α is another parameter to balance two terms in the interpolation. In our experiments, both C and α are experimentally chosen according to cross-validation.

3. SEVERAL POPULAR LOW-LEVEL CONFIDENCE MEASURES

3.1. Word Posterior Probability from word-graph

Given an acoustic observation sequence X and its word sequence W, the posterior probability is defined as follows:

$$P(W \mid X) = \frac{P(W)P(X \mid W)}{P(X)} = \frac{P(W)P(X \mid W)}{\sum_{all\ W} P(W)P(X \mid W)} \quad (8)$$

where *P(W)* is the language model score, *P(X|W)* is the acoustic model score. The *P(W|X)* is word posteriori probability, which can serve as a good confidence measure. However, the normalization term *P(X)* can not be easily computed. Several approximation strategies must be used.

Assume the word lattice is given from recognition process, the posterior probability of a word W can be approximately calculated as follows:

$$P(W \mid Lat) = \frac{\sum_{W \in path,\, path \subset Lat} p(path \mid Lat)}{\sum_{path \subset Lat} p(path \mid Lat)} \quad (9)$$

The denominator and numerator of (9) can be computed via the forward and backward algorithm. When calculating the numerator of (9), all path scores through W with a slightly different starting time and ending time should be merged. The details for the computing of posterior probability can be found in [1]. In this paper, we adopt a merging method called MED in [1],

$$C_{med}(W_s^e) = \sum_{W_{s'}^{e'},\, s' \le (s+e)/2 \le e'} p(W_{s'}^{e'} \mid X) \quad (10)$$

where *s* and *t* specify starting time and ending time of word *W*, *s'* and *t'* specify starting time and ending time of *W'* in a hypothesis path.

3.2. N-Best counting

N-Best counting is a simple confidence measure based on the assumption that a correct word should appear more times in the NBest list than an incorrect word. We firstly align each of the N-Best list with the top hypothesis via a dynamic programming algorithm, then count the number of each word in the top hypothesis occurs in the same position in the N hypotheses. After normalized by *N*, these normalized counts can serve as CMs of all recognized words.

3.3. LRT based on positive/negative models

The LRT related CM is based on the statistical hypothesis testing theory. The log likelihood ratio is calculated as:

$$LLR_W = \frac{1}{T}\left[\log P(X \mid \lambda_W) - \log P(X \mid \lambda'_W)\right] \quad (11)$$

where λ_W denotes the positive model for word *W* in the *null* hypothesis and λ'_W the negative model for *W* in the alternative hypothesis and *T* is duration of *X*. The calculation of LLR is a two-pass process: We first obtain speech boundaries through speech recognition in the first process; LLR score is calculated for each speech segment and then combined for each word as its CM.

In this paper, we adopt the data selection strategy in [5] for training both positive and negative models. As in [5], the recognition process is performed in the whole training set. During the process, we collect the so-called true tokens and competing tokens for each phoneme. After that, positive models of all phonemes are trained from all corresponding true tokens while negative models of all phonemes from the corresponding competing tokens. When calculating LLR-based CM for each word, the word model is constructed by concatenating all its corresponding mono-phone models.

4. EXPERIMENT (I): SWITCHBOARD

5.1. Experimental Setup

In this part, all comparative experiments of various CMs are performed on the '2000 evaluation set of the Switcthboard task, denoted as *eva00*. The 39-D feature vector consists of 12 PLP, energy, and their delta and delta-delta coefficients in the baseline system. The acoustic models are state-tied tri-phone CDHMMs with 8 Gaussian mixtures per state. We use a small training subset (totally 23.5hr) of Switchboard for acoustic models estimation. The baseline system achieves 43.8% word error rate (WER) in *eva00*. In the following experiments, we attempt to use various CMs to detect recognition errors in the outputs of the baseline ASR system. In other words, different CMs are computed for all recognized words to verify the misrecognized words against the correctly recognized words.

5.2. Experimental Results (1): high-level CMs

We choose the 50,000 longest text sentences from the *Switchboard1* database to estimate the word-document co-occurrence matrix in LSA experiments. The dimension of the reduced space, where all of the representing vectors lie in, is of 150 here. By using LSA-based CMs, we get the 44.7% equal error rate (EER) in our verification experiments. For mutual information (MI) based CMs, we use all transcriptions in *Switchboard1* (roughly 250k sentences in total) to estimate the inter-word mutual information matrix. We use cross-validation to tune two smoothing parameters. In other words, we equally split all transcripts in *eva00* into three parts and use two parts to optimize *C* and α and calculate CMs in the remaining part. We rotate three times to get the CMs for all sentences in *eva00*. For MI based CMs, we get EER 41.4% without smoothing and EER 41% after smoothing. For the two examined high-level CMs, we can see that the proposed

MI-based CMs significantly outperform the LSA-based ones.

5.3. Experimental results (2): low-level CMs

In this part, we also investigate three popular low-level CMs, i.e., word posteriori probabilities, N-best counting and LRT based on positive/negative verification models. From the results in Table 1, we can see that word posteriori probability gives the best verification performance, i.e., 24.4% EER, which is the best single CM.

5.4. Experimental results (3): combination of CMs

Experimental results also show that even the best high-level CMs (i.e., the MI-based ones) perform significantly worse than the best low-level CMs (i.e., word posteriori probabilities). By definition, the high-level and low-level CMs are computed from two totally independent information sources. This calls for a combination of the best low-level CMs with the best high-level ones. Due to their independence in nature, we investigate the verification performance by combining the best low-level CMs, namely word posteriori probabilities, with the best high-level CMs, namely MI with smoothing. The combination is done by the simple linear interpolation strategy and all weights are tuned based on the above-mentioned cross-validation procedure. From the result in Table 1, we clearly see that the combined CMs yield the best verification performance, 23.9% of EER, which outperform the best low-level CMs, i.e., word posteriori probabilities.

CM	Switchboard	Dictation
LSA	44.7	38.5
MI (no smoothing)	41.4	35.8
MI (with smoothing)	41.0	33.7
Posterior probability	24.4	17.5
N-Best counting	28.3	21.1
LRT	41.4	-
MI+ Posterior	***23.9***	***16.2***

Tabel 1. Performance comparison (EER in %) of various CMs, where Switchboard means the Switchboard task, and Dictation means the mandarin dictation task.

5. EXPERIMENT (II): MADARINE DICTATION

In this part, we repeat our verification experiments in another large vocabulary task, i.e., a 65k word mandarin dictation task. In the baseline system, we use 39-D feature including 12 MFCC, pitch and their 1-order and 2-order delta coefficients. The acoustic models are state-tied tri-phone CDHMMs with 8 mixtures per state, estimated from hundreds of hours of training data. The tri-gram language model is trained from a large Chinese newspaper corpus. The evaluation set is a 50 people (25 male and 25 female), 1000 sentences (20 sentences per people) set. The baseline system gets 7.7% error rate in Chinese character level. Similarly, the verification experiments are conducted by using various CMs to verify recognition errors in the baseline outputs in level of Chinese characters. The final results are shown in the third column of Table 1. For LSA, we train the word-document matrix with 50k text sentences from a Chinese newspaper. The dimension of the reduced space in LSA is of 175 in this experiment. LSA-based CM gets 38.5% in EER. For MI, the mutual information matrix is trained from about 80k sentences in the training set, we get two EERs: 35.8% (without smoothing) and 33.7% (with smoothing). It is shown that the MI-based CMs outperform the LSA technique. In the dictation task, word posterior probabilities is still the best single CM, which yield 17.5% in EER. At last, when combining word posteriori probabilities with the MI (with smoothing), the verification performance is further improved to 16.2%, which indicates roughly 10% relative error reduction in EER.

6. CONCLUSION

In this paper, we propose a new high-level CM for large vocabulary speech recognition that is based on mutual information between each pair of words. In our verification experiment, it is shown that it significantly outperforms the LSA-based CMs. Moreover, we also experimentally study several popular low-level CMs. The experiments show that word posteriori probabilities give the best performance. At last, we investigate how to combine the best low-level CMs with the best high-level ones. The experimental results show that a linear combination of word posteriori probabilities and mutual information consistently improve the verification performance, which significantly outperforms the best single CM.

11. REFERENCES

[1] F.Wessel, R. Schluter, K. Marcherey and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol.9, no.3, 2001.

[2] S. Cox, S. Dasmahapatra, "High-level Approaches to Confidence Estimation in Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol.10, no.7, 2002.

[3] H. Jiang, "Confidence Measures for Speech Recognition: A survey," *Technical Report CS-2003-06*, York University, 2003. (*http://www.cs.yorku.ca/techreports/2003/CS-2003-06.html*)

[4] S. Cox and R. Rose, "Confidence measures for the Switchboard database", *Proc. of ICASSP '96*, pp.511-415, 1996.

[5] H. Jiang, F. Soong, C.-H. Lee, "A Data Selection Strategy for Utterance Verification in Continuous Speech Recognition," *EuroSpeech'2001*, 2001.

[6] J. R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling," *Proceedings of the IEEE* , vol.88, no.8, 2000.