On Using Entropy Information to Improve Posterior Probability-based Confidence Measures

Tzan-Hwei Chen[1], Berlin Chen[1] and Hsin-Min Wang[2]

[1] Graduate Institute of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
{g93470018, berlin}@csie.ntnu.edu.tw
[2] Institute of Information Science Academia Sinica, Taipei, Taiwan
whm@iis.sinica.edu.tw

Abstract. In this paper, we propose a novel approach that reduces the confidence error rate of traditional posterior probability-based confidence measures in large vocabulary continuous speech recognition systems. The method enhances the discriminability of confidence measures by applying entropy information to the posterior probability-based confidence measures of word hypotheses. The experiments conducted on the Chinese Mandarin broadcast news database MATBN show that entropy-based confidence measures outperform traditional posterior probability-based confidence measures. The relative reductions in the confidence error rate are 14.11% and 9.17% for experiments conducted on field reporter speech and interviewee speech, respectively.

Keywords: confidence measure, entropy, posterior probability, continuous speech recognition.

1 Introduction

With the growing number of applications for automatic speech recognition (ASR) systems, the robustness and stability of a speech recognizer has become increasing important. The performance of ASR systems in real-world applications usually degrades dramatically compared to that of laboratory ASR systems. Therefore, verifying the recognition output of ASR systems is a critical issue. Confidence measures can be used to automatically label individual hypothesized words in the output of ASR systems as either *correct* or *incorrect*. This additional appraisal of word sequences has been adopted in unsupervised model training [1], and to improve the recognition accuracy [2].

Confidence measure algorithms can be roughly classified into three major categories [3] as follows:

1) Feature-based: These approaches assess the confidence based on some selected features, such as word duration, acoustic score, language model back-off, and part-of-speech.

2) Explicit model-based: These approaches treat confidence measures as hypothesis testing problems [4], and need to model extra alternative hypotheses.
3) Posterior probability-based: The posterior probability estimated according to the standard Maximum *a Posteriori* (MAP) framework is a good candidate for confidence measures, because it has a good bounded range between 0 and 1. The superior performance of the posterior probability has been demonstrated by using it as a confidence measure [5] [6].

In this paper, our objective is to improve the posterior probability-based approach by integrating entropy information into the posterior probability-based confidence measures of word hypotheses using a word graph. The experiments conducted on the Chinese Mandarin broadcast news database MATBN show that our approach can effectively reduce the confidence error rate.

The remainder of the paper is organized as follows. Section 2 describes traditional posterior probability-based confidence measures [5]. In Section 3, we explain how to combine the entropy information with the posterior probability-based confidence measures. Section 4 introduces the ASR system and the databases used in this paper. The experiment results are detailed in Section 5. Finally, in Section 6, we present our conclusions and indicate some future research directions.

2 Traditional Posterior Probability-based Confidence Measures

The fundamental rule in statistical speech recognition systems tries to find a word sequence $\{[w]_1^M\}_{opt}$ that maximizes the posterior probability, given a sequence of acoustic observations X of length T:

$$\begin{aligned}\{[w]_1^M\}_{opt} &= \arg\max_{[w]_1^M} p([w]_1^M \mid X) \\ &= \arg\max_{[w]_1^M} \frac{p(X \mid [w]_1^M)\, p([w]_1^M)}{p(X)}, \\ &= \arg\max_{[w]_1^M} p(X \mid [w]_1^M)\, p([w]_1^M)\end{aligned} \tag{1}$$

where $p([w]_1^M)$ and $p(X \mid [w]_1^M)$ denote the language model probability and the acoustic model probability, respectively; and $p(X)$ represents the prior probability of the acoustic observation sequence X. In practical implementations, the denominator term $p(X)$ is omitted because it is invariant for all possible word sequences. Apart from being used to select the most likely sentence, the sentence posterior probability also serves as a good confidence measure. However, the following question arises: How can we compute the sentence posterior probability?

2.1 Calculating the Posterior Probability of a Hypothesized Word

Given a hypothesized sentence (or word sequence), the posterior probability of a hypothesized word $[w,s,e]$[1] in the sentence, $p([w,s,e]|X)$, is equal to the sum of the posterior probabilities of all the sentences that contain the word:

$$p([w,s,e]|X) = \frac{\sum_{\substack{[w_m,s_m,e_m]_{m=1}^{M} \\ w_m=w,s_m=s,e_m=e}} \left\{ \prod_{m=1}^{M} p(X_{s_m}^{e_m} | w_m) \cdot p(w_m | h_m)^{\kappa} \right\}}{\sum_{[w_n,s_n,e_n]_{n=1}^{N} \in \mathbf{W}_{\Sigma}} \left\{ \prod_{n=1}^{N} p(X_{s_n}^{e_n} | w_n) \cdot p(w_n | h_n)^{\kappa} \right\}}. \quad (2)$$

In Eq. (2), $p(X_{s_m}^{e_m} | w_m)$ is the acoustic likelihood; $p(w_m | h_m)$ is the language model probability given the preceding history h_m; $\mathbf{W}_{\Sigma}$ denotes the set of all possible word sequences belonging to the language; N denotes the number of words in an arbitrary word sequence; and κ is the language model scaling factor. Note that the denominator term in Eq. (2) is equal to $p(X)$ in Eq. (1). Practically speaking, it is infeasible to enumerate all the word sequences in the implementation of a speech recognition system. Thus, to calculate the denominator term in Eq. (2), a word graph is usually adopted to approximate the word sequence set $\mathbf{W}_{\Sigma}$. Let Ψ^X denote the word graph generated for an acoustic observation sequence X. The posterior probability of the hypothesized word $[w,s,e]$, $p([w,s,e]|X)$, can be approximated as $p(a:[w,s,e]|\Psi^X)$ computed by:

$$p(a:[w,s,e]|\Psi^X) = \frac{\sum_{\substack{[w_m,s_m,e_m]_{m=1}^{M} \in \Psi^X \\ w_m=w,s_m=s,e_m=e}} \left\{ \prod_{m=1}^{M} p(X_{s_m}^{e_m} | w_m) \cdot p(w_m | h_m)^{\kappa} \right\}}{\sum_{[w_n,s_n,e_n]_{n=1}^{N} \in \Psi^X} \left\{ \prod_{n=1}^{N} p(X_{s_n}^{e_n} | w_n) \cdot p(w_n | h_n)^{\kappa} \right\}}, \quad (3)$$

where $a:[w,s,e]$ denotes the word arc associated with the hypothesized word $[w,s,e]$ in the word graph. The word posterior probability can be computed efficiently by applying a forward-backward algorithm to the word graph [5]. The forward score $\alpha(a:[w,s,e])$ is recursively computed from the start time of the word graph to the start time of the word arc $a:[w,s,e]$, i.e., s:

$$\alpha(a:[w,s,e]) = p(X_s^e | w) \times \sum_{a':[w',s',s-1]} \alpha(a':[w',s',s-1])\, p(w | h_w)^{\kappa}, \quad (4)$$

[1] s is the start time and e is the end time of the hypothesized word w.

where the summation is conducted for all word arcs ending at $s-1$; $p(X_s^e \mid w)$ is the acoustic likelihood; and h_w is the preceding history of w. Similarly, a backward score $\beta(a:[w,s,e])$ is calculated from the end time of the word graph to the end time of $a:[w,s,e]$, i.e., e:

$$\beta(a:[w,s,e]) = \sum_{a':[w',e+1,e']} \beta(a':[w',e+1,e'])\, p(w' \mid w)^{\kappa}\, p(X_{e+1}^{e'} \mid w'), \tag{5}$$

where the summation is conducted for all word arcs starting at $e+1$. Obviously, the numerator in Eq. (3) is the product of $\alpha(a:[w,s,e])$ and $\beta(a:[w,s,e])$, while the denominator in Eq. (3) is the summation of the forward scores of all end-word arcs in the word graph. Therefore, Eq. (3) can be rewritten as:

$$p(a:[w,s,e] \mid \Psi^X) = \frac{\alpha(a:[w,s,e]) \times \beta(a:[w,s,e])}{\sum_{a':[w',s',T] \in \Psi^X} \alpha(a':[w',s',T])}. \tag{6}$$

2.2 Posterior Probability-based Confidence Measure of a Hypothesized Word

The word arc posterior probability calculated by Eq. (6) can be used directly as a measure of confidence for the word hypothesis $[w,s,e]$:

$$C_{normal}([w,s,e]) = p(a:[w,s,e] \mid \Psi^X). \tag{7}$$

Consider a word arc $a:[w,s,e]$ in a word graph. There usually exist some alternative word arcs whose word identities are identical to *w*, but the time marks are slightly different to [*s*,*e*]. The more such alternative word arcs exist, the more likely it is that [*w*,*s*,*e*] is correct and should be accepted. Based on this concept, Wessel *et al.* proposed three methods for calculating the confidence of a hypothesized word $[w,s,e]$ according to the word arc posterior probabilities [5]:

$$C_{\text{sec}}([w,s,e]) = \sum_{\substack{a':[w,s',e'] \\ \{s',\ldots,e'\} \cap \{s,\ldots,e\} \neq \phi}} p(a':[w,s',e'] \mid \Psi^X), \tag{8}$$

$$C_{med}([w,s,e]) = \sum_{\substack{a':[w,s',e'] \\ s' \le (s+e)/2 \le e'}} p(a':[w,s',e'] \mid \Psi^X), \tag{9}$$

and

$$C_{\max}([w,s,e]) = \max_{t \in \{s,\ldots,e\}} \sum_{\substack{a':[w,s',e'] \\ s' \le t \le e'}} p(a':[w,s',e'] \mid \Psi^X). \tag{10}$$

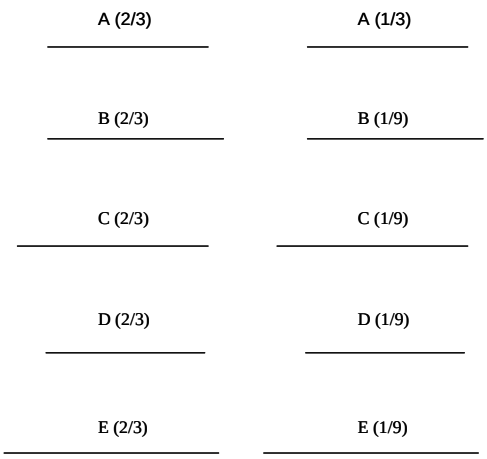

Fig. 1. Two examples of word arcs extracted from word graphs. The values in parentheses represent the confidence (e.g., C_{max}) of the word arc.

In Eq. (8), the summation is over all word arcs containing the same word identity that intersect with the word arc $a:[w,s,e]$. In Eq. (9), only word arcs that associate with the same word identity *w* and intersect the median time of the word arc $a:[w,s,e]$ are considered. In Eq. (10), for each time instance between *s* and *e*, the posterior probabilities of the word arcs whose word identities are *w* are accumulated. This process yields *e*-*s*+1 accumulated posterior probabilities. Then, the confidence of the word hypothesis [*w*,*s*,*e*] is the maximum of these accumulated posterior probabilities.

3 The Proposed Approach

To determine whether a recognized word is correct or not, it might be helpful to take all the other word arcs with time boundaries similar to the target word arc hypothesis into account, instead of just considering the word arcs whose word identities are identical to the recognized word to be evaluated. As illustrated in Fig. 1, the confidence of word 'A' on the left-hand side is not reliable in terms of alternative words because all the confidence measures are high. In contrast, the confidence of word 'A' on the right-hand side is trustworthy because it is relatively higher than those of the other words. Entropy, described as "a measure of the disorder", is a way to measure the amount of information in a random variable. To emphasize the reliability of confidence measure, we propose an entropy-based approach that evaluates the degree of confusion in confidence measures. By incorporating entropy information into traditional posterior probability-based confidence measures, the new entropy-based confidence measure of a hypothesized word is defined as:

$$C_{entropy}([w,s,e]) = CM([w,s,e]) \cdot (1 - E_{avg}([w,s,e])), \quad (11)$$

where $CM([w,s,e])$ denotes a traditional confidence measure of $[w,s,e]$, which can be estimated by Eqs. (7), (8), (9), or (10); and $E_{avg}([w,s,e])$ is the average normalized entropy defined as:

$$E_{avg}([w,s,e]) = \frac{1}{e-s+1} \sum_{t=s}^{e} E_f(t). \quad (12)$$

The larger the $E_{avg}([w,s,e])$, the greater the degree of uncertainty there will be about the confidence measure. Consequently, the originally estimated confidence of a recognized word is considered more unreliable. Based on this concept, we weight the conventional confidence measure by $1 - E_{avg}([w,s,e])$. In Eq. (12), $E_f(t)$ is computed by,

$$E_f(t) = -\frac{1}{\log_2 N} \sum_{[w,s,e], s \le t \le e} P_{CM}([w,s,e],t) \log_2 P_{CM}([w,s,e],t), \quad (13)$$

where N is the number of distinct word identities in frame *t*; and $P_{CM}([w,s,e],t)$ represents the normalized confidence in each frame *t*, calculated by

$$P_{CM}([w,s,e],t) = \frac{CM_{sum}([w,s,e],t)}{\sum_{[w',s',e'], s' \le t \le e'} CM_{sum}([w',s',e'],t)}, \quad (14)$$

and

$$CM_{sum}([w,s,e],t) = \sum_{\substack{a'=[w,s',e'] \\ s' \le t \le e'}} CM([w,s',e'],t). \quad (15)$$

When computing the entropy, we only consider the distribution of different words. We take the summation over the confidence of words with the same identity before calculating $P_{CM}([w,s,e],t)$. In [7], the entropy information was considered as one of predictor features in a confusion network. In this paper, we integrate entropy information into the posterior probability-based confidence measures directly.

4 The Speech Recognition System

The large vocabulary continuous speech recognition (LVCSR) system [1] and the databases used in this paper are described in this section.

Table 1. The statistics of two speech data sets used in this paper.

	Field reporter speech	Interviewee speech
Training data	25.5 h	8.80 h
Validation data	0.74 h	0.45 h
Test data	1.5 h	0.60 h

4.1 Front-end Signal Processing

Front-end signal processing is performed with the HLDA-based [8] data-driven Mel-frequency feature extraction approach, and further processed by MLLT transformation for feature de-correlation. Finally, utterance-based feature mean subtraction and variance normalization are applied to all the training and test utterances.

4.2 Speech Corpus and Acoustic Model Training

In this work, we use the MATBN (Mandarin Across Taiwan Broadcast News) speech database [9], which was collected by Academia Sinica and Public Television Service Foundation of Taiwan between November 2001 and April 2003. Approximately 200 hours of speech data was supplemented with corresponding orthographic transcripts. Our experiments are conducted on the "field reporter" and "interviewee" subsets. The statistics of these two subsets are summarized in Table 1.

The acoustic models used in our LVCSR system are comprised of 112 right-context-dependent INITIAL models, 38 context-independent FINAL models, and a silence model. The models are first trained by using the Baum-Welch training algorithm according to the ML criterion, and then optimized by the MPE-based [10] discriminative training approach.

4.3 The Lexicon and Language Model

The recognition lexicon consists of 72K words. The language models used in this paper consist of trigram and bigram models. They were estimated from a text corpus of 170 million Chinese characters collected from Central News Agency in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC) based on the ML criterion. For the interviewee task, we used the Mandarin Conversational Dialogue Corpus (MCDC) [11] to train an in-domain language model. The *n*-gram language models were trained with the SRI Language Modeling Toolkit (SRILM) [12]. We also employed the Katz back-off smoothing technique.

Table 2. Recognition results for the two test subsets.

	Field reporter speech	Interviewee speech
Character error rate	20.79%	49.56%

4.4 Speech Recognition

The speech recognizer is implemented with a left-to-right frame-synchronous Viterbi Tree search and a lexical prefix tree organization of the lexicon. The recognition hypotheses are organized into a word graph for further language model rescoring. In this study, the word bigram language model is used in the tree search procedure, while the trigram language model is used in the word graph rescoring procedure [1].

5 Experiments

5.1 Experiment Setup

The acoustic models for the field reporter and interviewee tasks were trained with approximately 25 hours and 9 hours of speech, respectively. The acoustic models for field reporters were trained by 150 iterations of ML training and 10 iterations of MPE training; while the acoustic models for interviewees were trained by 30 iterations of ML training and 8 iterations of MPE training. The character error rates of the two tasks are shown in Table 2.

5.2 Evaluation Metric

The performance of the confidence measure is evaluated on the basis of the confidence error rate (CER) defined as:

$$CER = \frac{\#\,\text{falsely accepted words} + \#\,\text{falsely rejected words}}{\#\,\text{recognized words}}. \quad (16)$$

CER can be clarified as follows: Given the confidence of a hypothesized word and a rejection threshold, the word is labeled *correct* (i.e., accepted) or *incorrect* (i.e., rejected). If an incorrectly recognized word is labeled *correct*, it is a false acceptance; similarly, if a correctly recognized word is tagged *incorrect*, it is a false rejection. The baseline CER is calculated as the number of insertions and substitutions divided by the number of recognized words. Obviously, the CER is heavily dependent on the choice of the rejection threshold. In our experiments, the threshold was adjusted to minimize the CER of the validation set. Then, the threshold that yielded the minimal CER for the validation set was applied to the test set.

Table 3. Experiment results using the confidence error rate to evaluate traditional posterior probability-based confidence measures and entropy-based confidence measures.

Methods	Field reporter speech	Interviewee speech
baseline	24.52%	51.97%
C_{normal}	22.18%	31.31%
C_{sec}	21.47%	32.32%
C_{med}	21.47%	32.32%
C_{max}	21.47%	32.32%
$C_{entropy}(C_{normal})$	18.55%	31.08%
$C_{entropy}(C_{sec})$	18.44%	28.50%
$C_{entropy}(C_{med})$	18.44%	28.44%
$C_{entropy}(C_{max})$	18.45%	28.51%

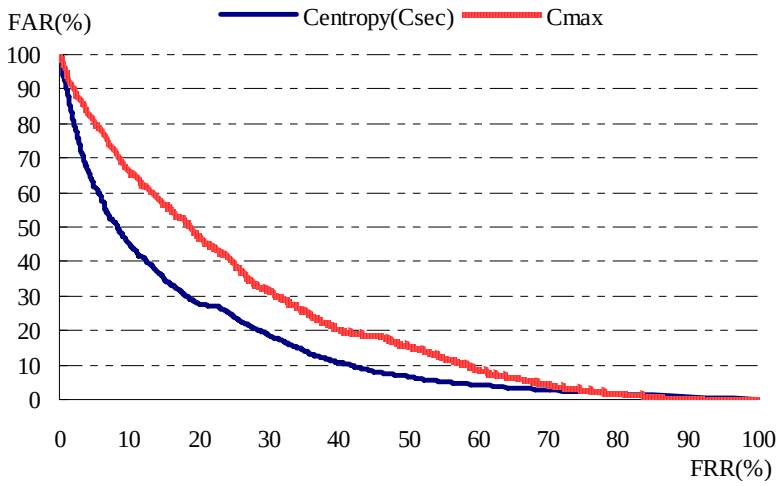

Fig. 2. DET curves for the test set of the field reporter task.

Another evaluation metric is the detection-error-tradeoff (DET) curve, which contains a plot of the false acceptance rate over the false rejection rate for different thresholds.

5.3 Experiment Results

The comparison of the proposed entropy-based confidence measures and the traditional posterior probability-based confidence measures is shown in Table 3. The language model scaling factor κ in Eq. (3) is set to 11. The third to sixth rows in the table show the results obtained when Eqs. (7), (8), (9) and (10) are used, respectively, and the seventh to tenth rows are the results obtained when entropy information is integrated into the listed methods. From Table 3, it is clear that the proposed entropy-

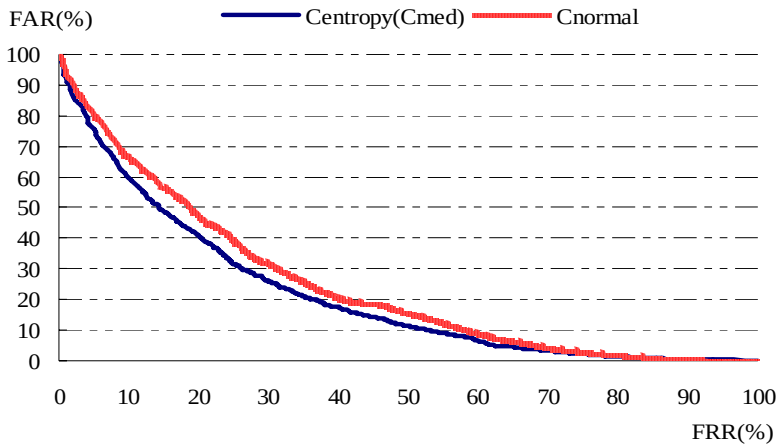

Fig. 3. DET curves for the test set of the interviewee task.

based confidence measures outperform traditional posterior probability-based confidence measures. The proposed approach achieves a relative CER reduction of 14.11% over the traditional approach ($C_{entropy}(C_{sec})$ or $C_{entropy}(C_{med})$ versus C_{sec}, C_{med}, or C_{max}) in the field reporter task; and a relative reduction of 9.17% ($C_{entropy}(C_{med})$ versus C_{normal}) in the interviewee task.

The DET curves of C_{max} and $C_{entropy}(C_{sec})$ for the field reporter task and the DET curves of C_{normal} and $C_{entropy}(C_{med})$ for the interviewee task are shown in Figs. 2 and 3, respectively. Again, we find that the proposed methods outperform traditional methods.

6 Conclusions

We have presented a new approach that combines traditional posterior probability-based confidence measures with entropy information to verify the output of large vocabulary continuous speech recognition systems. The proposed methods were evaluated on two speech recognition tasks: a field reporter speech set and an interviewee speech set. In the field reporter speech set, the proposed methods achieved a 14.11% relative reduction in the confidence error rate compared to traditional methods, while in the interviewee speech set, the proposed methods achieved a 9.17% relative reduction. In our future work, we will incorporate the entropy information into feature-based confidence measures or other confidence measures.

References

1. B. Chen, J.-W. Kuo and W.-H. Tsai, “Lightly Supervised and Data-driven Approaches to Mandarin Broadcast News Transcription,” International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 1, pp.1-18, 2005.
2. F. Wessel, R. Schlüter and H. Ney, “Using Posterior Probabilities for Improved Speech Recognition,” in Proc. ICASSP 2000.
3. W. K. Lo and F. K. Soong, “Generalized Posterior Probability for Minimum Error Verification of Recognized Sentences,” in Proc. ICASSP 2005.
4. R. C. Rose, B.-H. Juang and C.-H. Lee, “A Training Procedure for Verifying String Hypothesis in Continuous Speech Recognition,” in Proc. ICASSP 1995.
5. F. Wessel, R. Schlüter and H. Ney, ”Confidence Measures for Large Vocabulary Continuous Speech Recognition,” IEEE Trans. Speech and Audio Processing, Vol. 9, No. 3, pp.288-298, 2001.
6. W. K. Lo, F. K. Soong and S. Nakamura, “Generalized Posterior Probability for Minimizing Verification Errors at Subword, Word and Sentence Levels,” in Proc. ISCSLP, 2004.
7. J. Xue and Y. Zhao, “Random Forests-based Confidence Annotation Using Novel Features from Confusion Network,” in Proc. ICASSP 2006.
8. H. J. F. Gales, “Semi-tied Covariance Matrices for Hidden Markov Models,” IEEE Trans. on Speech and Audio Processing, Vol. 7, No. 3, pp. 272-281, 1999.
9. H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S Cheng, “MATBN: A Mandarin Chinese Broadcast News Corpus,” International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 2, pp. 219-236, 2005.
10. D. Povey, “Discriminative Training for Large Vocabulary Speech Recognition,” Ph. D Dissertation, Peterhouse, University of Cambridge, 2004.
11. S. C. Tseng and Y. F. Liu, “Mandarin Conversational Dialogue Corpus,” MCDC. Technical Note 2001-01, Institute of Linguistics, Academia Sinica, Taipei.
12. A. Stolcke, SRI language Modeling Toolkit version 1.3.3, http://www.speech.sri.com/projects/srilm/.