Word Error Rate Minimization Using an Integrated Confidence Measure

Akio Kobayashi, Kazuo Onoe, Shoei Sato and Toru Imai

NHK (Japan Broadcasting Corporation)
Science and Technical Research Laboratories, Japan
{kobayashi.a-fs, onoe.k-ec, satou.s-gu, imai.t-mq}@nhk.or.jp

Abstract

This paper describes a new criterion of speech recognition using an integrated confidence measure for minimization of the word error rate (WER). Conventional criteria for WER minimization obtain an expected WER of a sentence hypothesis merely by comparing it with other hypotheses in an n-best list. The proposed criterion estimates the expected WER by using an integrated confidence measure with word posterior probabilities for a given acoustic input. The integrated confidence measure, which is implemented as a classifier based on maximum entropy (ME) modeling, is used to get probabilities reflecting whether the word hypotheses are correct or incorrect. The classifier comprises a variety of confidence measures and can deal with a temporal sequence of them in order to attain a more reliable confidence. Our proposed criterion achieved a WER of 7.5% and a 2.6% improvement relative to conventional n-best rescoring methods in transcribing Japanese broadcast news under noisy field and spontaneous speech conditions.

1. Introduction

The recent progress with corpus-based spoken language processing has led to significant successes in real-world applications. NHK (Japan Broadcasting Corp.) developed a real-time large-vocabulary-continuous-speech-recognition (LVCSR) system for a closed-captioning service[1]. The system has definitely been success with read speech, but its accuracy in noisy field and spontaneous speech conditions has not been high enough for practical application. Various phenomena and reasons degrade speech recognition performance, and here, we especially focus on the criterion of speech recognition. The standard of speech recognition based on the Bayes’ rule yields a solution by way of maximizing a posterior probability of a sentence hypothesis given an acoustic input. It seems to diverge from the original purpose of obtaining the most accurate sentence, because the criterion does not exactly minimize the word error rate (WER) of the hypothesis, but maximizes the probabilities derived from both acoustic and language models. Thus, it is preferable that WER minimization is used as the criterion.

Several WER minimization techniques have been proposed from the perspective of Bayes risk minimization[2, 3]. In these studies, the methods estimate the WER of a sentence hypothesis simply by comparing it with other hypotheses in an n-best list and never use any other observations except word posterior probabilities. Therefore, we attempt to estimate the WER of the sentence hypothesis more directly and efficiently by calculating a statistical confidence from a variety of confidence measures.

In this paper, we propose a new criterion for WER minimization. The basic idea of our approach is to minimize a WER, which is explicitly computed on word hypotheses using the integrated confidence measure. The integrated confidence measure gives a probability reflecting whether a word hypothesis is correct or not. It is defined as a classifier based on ME modeling in order to estimate the WER efficiently. ME modeling is a powerful technique to integrate knowledge sources, and we use it to incorporate a variety of confidence measures such as the word posterior probability[4], the word hypothesis density[5], etc. The proposed measure can also deal with a temporal sequence of confidence scores across a word sequence. Finally, we show its classification ability and n-best rescoring results in experiments transcribing broadcast news.

2. Criterion for WER Minimization

2.1. WER Minimization

The established criterion of speech recognition on the basis of the Bayes’ rule is given by

\[ w^* = \arg \max_w P(w|z) \]  \hspace{1cm} (1)

where \( w \) is a sentence hypothesis and \( z \) is an acoustic input. The solution \( w^* \) is usually given by the combination of the acoustic model \( P(z|w) \) and the language model \( P(w) \) according to Eq. (2). Alternatively, the solution is computed from Eq. (1) using the forward-backward algorithm when a word lattice is given. The criterion does not always yield the solution with the minimum WER, because it does not compute the WER of a sentence hypothesis explicitly but maximizes its posterior probability.

Now, we attempt to redefine the problem from the point of view of WER minimization, or

\[ w^* = \arg \min_{\hat{w}} \text{WER}(w|z). \] \hspace{1cm} (3)

The function \( \text{WER}(w|z) \) is the WER of \( w \) given \( z \). We try to solve this problem by employing a confidence measure, which determines the confidence of word hypotheses according to particular criteria. Using the confidence measure, we define \( \text{WER}(w|z) \) in Eq. (3) as follows:

\[ \text{WER}(w|z) \equiv 1.0 - \frac{1}{|z|} \sum_{i} \left \{ P(w_i = \text{correct}) \times P(w_i|z) \right \}, \] \hspace{1cm} (4)

where \( P(w_i|z) \) denotes the word posterior probability of the \( i \)-th word hypothesis in \( w \), and \( |w| \) is the total number of words in \( w \). The statistical model \( P(w_i = \text{correct}) \) represents the confidence measure and gives the probability that \( w_i \) is correct. The model is realized by a confidence classifier \( P(y_i|x_i) \), where \( y_i \) is a tag, which indicates 1 if \( w_i \) is correct or -1 if incorrect, and \( x_i \) is a vector of confidence scores computed with several confidence measures. The probability given by the model can be
regarded as an accuracy for every word hypothesis. The solution of Eq. (3) apparently depends on the ability of the model as a confidence measure. Therefore, constructing an efficient model \( P(y_i|x_i) \) is key to solving the problem.

### 2.2. Related Works

There is an alternative approach for the WER minimization technique[2, 3].

\[
\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_{i} \text{WER}'(\mathbf{w}, \mathbf{w}')P(\mathbf{w}'|z). \tag{5}
\]

The function \( \text{WER}'(\mathbf{w}, \mathbf{w}') \) is implemented as a loss function derived from the Levenshtein algorithm which measures an edit distance between a sentence hypothesis \( \mathbf{w} \) and another hypothesis \( \mathbf{w}' \). It estimates the WER by comparing hypotheses in an n-best list with each other and never use any other knowledge sources.

### 3. Integrated Confidence Measure

#### 3.1. Maximum Entropy Modeling

Maximum entropy models are often used in classification problems[6]. In this section, we propose an integrated confidence measure, which incorporates a set of measures by ME modeling.

The integrated confidence measure is regarded as a classifier which gives an optimum tag sequence \( y^* \) using the statistical model \( P(y_i|x_i) \):

\[
y^* = \arg \max_y \prod_{i=1}^{N} P(y_i|x_i). \tag{6}
\]

We define a feature function in order to represent relations between vectors of confidence scores and tags.

\[
\mathcal{F} = \{ f_k : (x, y) \rightarrow \{0, 1\}, k \in \{1, 2, \ldots \} \}. \tag{7}
\]

It returns 1 only if \((x, y)\) matches an adequate condition with regard to \( f_k \). The expectation of \( f_k \) on \( P(y|x) \) is defined by

\[
E_P[f_k] = \sum_x \sum_y P(x)P(y|x)f_k(x, y), \tag{8}
\]

and constrained by

\[
E_P[f_k] = E_{\tilde{P}}[f_k], \tag{9}
\]

where \( \tilde{P} \) represents the empirical distribution over given training data. The other constraint is

\[
\sum_y P(y|x) = 1. \tag{10}
\]

The entropy for the conditional model is given by

\[
H(P) = -\sum_{x, y} P(x, y) \log P(y|x). \tag{11}
\]

One of the models satisfying Eq. (9) and Eq. (10) is obtained from the maximum entropy principle:

\[
P_{ME}(y|x) = \arg \max_P H(P). \tag{12}
\]

\( P_{ME} \) is obtained by

\[
P_{ME}(y|x) = \frac{\exp \sum_k \lambda_k f_k(x, y)}{\sum_y \exp \sum_k \lambda_k f_k(x, y)}. \tag{13}
\]

\( \lambda_k \) is a parameter of the model \( P_{ME} \) estimated with the generalized iterative scaling (GIS) algorithm[7]. We use \( P_{ME}(y|x) \) as the integrated confidence measure represented by \( P(w_i = \text{correct}) \) in Eq. (4).

#### 3.2. Feature Definition

The simplest feature, for example, is given by

\[
f_{\text{simple}}(x_i, y_i) = \begin{cases} 1 & \text{if } x_i > c_t \land y_i = 1 \\ 0 & \text{else} \end{cases} \tag{14}
\]

where \( c_t \) is a threshold for a particular confidence score (e.g. the word posterior probability). However, the feature is too much simple to represent relations between word hypotheses and their confidence scores, because the confidence scores are generally distributed on their domain continuously. Then, we define the features by using a series of thresholds \( c_t = \{c_{t0}, c_{t1}, \ldots, c_{tm}\} \) in order to describe the relations in detail. The first threshold \( c_{t0} \) for a confidence measure is determined at the point where the classification error rate becomes the minimum on development data. Then, we set a new pair of thresholds \( c_{t1} = c_{t0} + \Delta \) and \( c_{t2} = c_{t0} - \Delta \) for new features. The set of the thresholds and the interval \( \Delta \) are determined from preliminary experiments.

Confidence scores of predecessors and successors of word hypotheses are also available to get information about the confidence. For instance, a word hypothesis is likely to be incorrect if its predecessor and successor have low confidence scores. Such a feature is represented by

\[
f_{\text{seq}}(x_{i-1}, x_i, y_i) = \begin{cases} 1 & \text{if } x_{i-1} > c_{t_m} \land x_i > c_{t_{m'}} \land y_i = 1 \\ 0 & \text{else}. \end{cases} \tag{15}
\]

The ME model becomes more information-rich by having the features derived from a series of thresholds and temporal sequences of confidence scores. The ensemble of confidence measures is expected to make the model more reliable.

The procedure of the feature definition is summarized as follows:

**Step 1** Determine the threshold \( c_{t0} \), which gives the minimum classification error rate (CER) on development data for a particular confidence measure.

**Step 2** Set a new pair of features, which evaluates either threshold \( c_{t1} = c_{t0} + \Delta \) or \( c_{t2} = c_{t0} - \Delta \) according to Eq. (14), and train a model with ME modeling.

**Step 3** Repeat Step 2 until the CER on the development data does not change.

**Step 4** Make a new pair of features, which have confidence scores from either a predecessor or a successor according to Eq. (15), and re-train a model with ME modeling.

**Step 5** Repeat Step 4 until the CER on the development data does not change.

**Step 6** Do Step 1 to 5 for all kinds of confidence measures and train the final model.

According to Eq. (13), the contribution of the defined feature \( f_k \) for the ME model is represented by the parameter \( \lambda_k \).

#### 3.3. Related Works

Hazen[8] studied integration of confidence measures. This integration method aggregates a set of confidence scores with corresponding weights determined by the maximum a posteriori
A support vector machine (SVM) is commonly used in many classification problems and was also adopted for a confidence classifier in [9]. According to preliminary experiments, our ME-based classifier gave better classification results than the SVM-based classifier. In addition, it is very difficult to find the optimum parameters needed for the SVM-based classifier, and the computational cost of SVM training is very expensive. Therefore, our proposed WER minimization criterion employs the ME-based classifier as the integrated confidence measure, and compare with other conventional criteria.

4. Experiments

4.1. Setup

Table 1 shows training and evaluation data taken from NHK’s broadcast news, and they were uttered by male speakers under noisy field and spontaneous speech conditions. The training data were extracted from 10 days of NHK’s news, and data from the last day, totaling 18,106 words, were used as development data. The evaluation data were extracted from 45 different news programs, and consisted of two overlapped subsets: a noisy field condition set of 870 sentences or 20,739 words, and a spontaneous speech condition set of 286 sentences or 5,401 words. The WERs in Table 1 are results for a baseline system[1].

The integrated confidence measure was trained from transcriptions including errors, which were produced by the LVCSR system from the training data. The confidence measures to be incorporated by ME modeling are listed in Table 2. The features with regard to the measures were determined on the development data according to the feature definition. In Table 2, the word posterior probability[4] is defined on a trigram lattice produced by the LVCSR system. The acoustic stability[10] represents the number of particular word hypotheses, which are included in the sentence hypotheses generated by a cycle of rescoring with a set of different grammar weights. The word hypothesis density[5] is defined as the relative frequencies of the word hypotheses at a particular frame. The acoustic scores are normalized by the frame lengths of the word hypotheses. The back-off case is a value that indicates whether the word hypotheses are covered by trigrams, bigrams or unigrams. The number of activated HMMs represents the average number of HMM states activated by the system in word hypotheses. The phoneme duration denotes the average number of phoneme lengths in word hypotheses.

The LVCSR system decodes input acoustic feature vectors to a trigram word lattice using a tree lexicon and bigram language models. In the second pass, it rescors hypotheses in word lattices using the trigram language models and yields the 300 best sentence hypotheses. We estimated the parameters such as grammar weight and insertion penalty from the development data. Acoustic models were obtained from 118 hours of male speech from broadcast news. The acoustic inputs were processed by RASTA[11] and parameterized into 39 dimensional vectors: 12 MFCC with log-power and their first and second order differentials. We used a set of trigram language models, which were adapted to each news program[12]. The models were smoothed by the Good-Turing estimation, and cut-offs were set to 1 for the bigrams and 2 for the trigrams. The lexicons of the models depended on the program, and they had 60k words on average.

4.2. Classification Results

First, we investigated the ability of the proposed integrated confidence measure as a classifier. Its ability was compared with that of a conventional binary classifier based on the word posterior probability. We measured correlation coefficients between these probabilities and the tags indicating correct or not. The correlation coefficient of the proposed measure was 0.59 while that of the word posterior probability was 0.31. The integrated confidence measure had larger correlations with the tags than the word posterior probability had.

Next, we used both methods to classify the words in the evaluation data. We set a threshold of the classifier based on the word posterior probability to 0.76 by doing a preliminary test on the development data. The results in Table 3 were measured according to classification error rate (CER), false acceptance rate (FAR) and false rejection rate (FRR). The proposed measure achieved a relative improvement of 7.7% in CER compared with the conventional measure. On the other hand, the FAR of the proposed measure was higher than that of the conventional measure. This indicates that the proposed measure tends to identify word hypotheses as correct words because the transcriptions of the training data consist of many more correct words than incorrect ones. This tendency was similarly exhibited in the noisy field and the spontaneous speech conditions.

The normalized cross entropy[13], which is another standard for evaluating the classification power of the measures, improved from 0.31 to 0.43 by the proposed measure.

4.3. Rescoring Results

We compared the following several rescoring methods on the evaluation data.

Baseline : Rescoring conventionally on the trigram lattice.
FW-BW : Rescoring using the forward-backward algorithm on the trigram lattice according to Eq. (1).
BRM : Rescoring using the Bayes risk minimization technique according to Eq. (5).
Proposed : Rescoring using the proposed method according to Eq. (3) and Eq. (4).

The proposed rescoring achieved the lowest WER (7.5%) and a 2.6% relative reduction in WER from Baseline rescoring (Table 4). This is statistically significant according to the matched
Table 3: Classification Results

<table>
<thead>
<tr>
<th></th>
<th>all conditions</th>
<th>noisy field</th>
<th>spontaneous speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%CER</td>
<td>%FAR</td>
<td>%FRR</td>
</tr>
<tr>
<td>word posterior probability</td>
<td>5.2</td>
<td>46.4</td>
<td>2.1</td>
</tr>
<tr>
<td>integrated confidence measure</td>
<td>4.8</td>
<td>53.5</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Table 4: Rescoring results (WER,%)

<table>
<thead>
<tr>
<th></th>
<th>all conditions</th>
<th>noisy field</th>
<th>spontaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.7</td>
<td>6.1</td>
<td>19.1</td>
</tr>
<tr>
<td>FW-BW</td>
<td>7.6</td>
<td>6.0</td>
<td>18.9</td>
</tr>
<tr>
<td>BRM</td>
<td>7.7</td>
<td>6.0</td>
<td>18.9</td>
</tr>
<tr>
<td>Proposed</td>
<td>7.5</td>
<td>5.9</td>
<td>18.4</td>
</tr>
</tbody>
</table>

5. Conclusions

We proposed a new criterion of speech recognition based on WER minimization. The criterion employed an integrated confidence measure that incorporates a set of confidence measures by ME modeling. The integrated confidence measure’s CER was 7.7% relatively lower than the CER of the conventional classifier based on the word posterior probability. The proposed criterion of WER minimization gave the best result in n-best rescoring and achieved a relative improvement of 2.6% in WER compared with the conventional rescoring. It is considered that the advantage of our proposed criterion results from various observations through the integrated confidence measure. For further improvement, we will use not only the observations such as word posterior probabilities but also other knowledge sources such as syntactic and semantic information.

6. References