Using Posterior Probability as Confidence Measure for LVCSR

Guan-Yu Menphis Chen

SLP Laboratory, NTNU
696470203@ntnu.edu.tw

Main Reference:
Outline

• Introduction

• Knowledge of Confidence Measure

• Experiment
Introduction

- Confidence measures are useful for improving performance of spoken language systems by assessing reliability of recognition output.

- There have been various approaches proposed for measuring confidence of speech recognition output. They can be roughly classified into three categories:
  1. Feature based
  2. Explicit model based
  3. Posterior probability based
Introduction

- Feature based approaches try to assess the confidence according to selected features using some trained classifiers.

- Explicit model based approaches employ a candidate class model with competing models and a likelihood ratio test is applied.

- The posterior probability based approach tries to estimate the posterior probabilities of a recognized entity given all acoustic observations.
Knowledge of Confidence Measure

- The fundamental rule in all statistical speech recognition systems is Bayes’ decision rule which is based on the posterior probability $p(w_i^M | x_i^T)$ of a word sequence $w_i^M = w_1, ..., w_M$, given a sequence of acoustic observations $x_i^T = x_1, ..., x_T$.

- The word sequence $\{w_i^M\}_{opt}$ which maximizes this posterior probability also minimizes the probability of an error in the recognized sentence.

$$\{w_i^M\}_{opt} = \arg\max_{w_i^M} p(w_i^M | x_i^T)$$

$$= \arg\max_{w_i^M} \frac{p(x_i^T | w_i^M) \cdot p(w_i^M)}{p(x_i^T)} = \arg\max_{w_i^M} \frac{\prod_{m=1}^{M} p(x_{s_m}^T | w_m) \cdot p(w_m | w_{m-1})}{p(x_i^T)}$$

where $p(w_i^M)$: the language model probability

$p(x_i^T | w_i^M)$: the acoustic model probability

$p(x_i^T)$: probability of the acoustic observations

National Taiwan Normal University
Knowledge of Confidence Measure – (posterior probability)

- In continuous speech recognition, the word posterior probability (WPP) can be computed by summing the posterior probabilities of all string hypotheses in the search space bearing the focused word, \( w \), starting at time \( s \) and ending at time \( t \), given as

\[
p([w; s, t] | x^T_t) = \sum_{\forall M, [w; s, t]^M \in M, 1 \leq n \leq M} \prod_{m=1}^{M} p(x^t_{s,m} | w_m) \cdot p(w_m | \hat{w}_i^M) \frac{p(x^T_t)}{p(x^T)}
\]

- \( p(x^t_s | w_m) \) is the acoustic likelihood; \( p(w_m | \hat{w}_i^M) \), the language model likelihood; \( x^t_s \), the sequence of acoustic observations; \( M \), the no. of words in a string hypothesis; \( \hat{w}_i^M \), the probability of the acoustic observations; \( T \), the length of the complete acoustic observations.
Knowledge of Confidence Measure – (posterior probability)

- We should be noted that the posterior probabilities of all parallel word graph edges hypothesized at a specific point in time $t$ always sum up to one.

$$
\sum_{[w; s', t']} p([w; s', t'] | x_t) = 1 \quad \forall t \in \{1, \ldots, T\}
$$
Knowledge of Confidence Measure – (posterior probability)

- First, summed up the probabilities of all hypotheses with an identical word index for which the intersection of the time intervals defined by the starting and ending times of the considered hypotheses is not empty.

\[
p([w, s, t] | x^T_i) = \sum_{\substack{[w, s', t'] \in \mathcal{H} \cap \mathcal{H}' \neq \emptyset \atop (s, t) \cap (s', t') \neq \emptyset}} p([w; s', t'] | x^T_i)\]

![Diagram with graph nodes and edges representing relationships between hypotheses and events.](image-url)
Knowledge of Confidence Measure – *(posterior probability)*

- Second, we can accumulate only the posterior probabilities of those hypotheses for word \( w \) which intersect the median time frame of the hypothesis under consideration.

\[
p([w; s, t] \mid x_i^T) = \sum_{[w; s', t'] \in \{w; s+t \leq t' \}} p([w; s', t'] \mid x_i^T)
\]
Third, the idea here is to determine the best-case probability for a given word to occur in a certain period of time.

\[
p([w; s, t] \mid x^T_i) = \max_{t_{\text{max}} \in \{s, \ldots, t\}} \sum_{[w; s', t'] \in \{w; s', t' \mid s' \leq t_{\text{max}} \leq t'\}} p([w; s', t'] \mid x^T_i)
\]
Knowledge of Confidence Measure – (Generalized WPP)

- Generalized word posterior probability is a generalization of WPP to take into account of three issues in computing WPP:

1. Reduced search space: A reduced search space (e.g., word graph or N-best list) is used when computing GWPP.

2. Relaxed time registration: the starting and ending time of a word is affected by various factors like the pruning threshold, noise, etc. In GWPP, words in the search space with the same identity and overlapping in time are considered as reappearances.
Knowledge of Confidence Measure – (Generalized WPP)

3. Reweighted acoustic and language model likelihoods:
   • Difference in dynamic range: Acoustic likelihoods obtained from pdf have an unbounded dynamic range, but language model likelihoods lie between 0 and 1.

   • Difference in frequency of computation: Acoustic likelihoods are computed every frame, while language model likelihoods are computed once per word.

   • Independence assumption: Neighboring acoustic observations are assumed to be statistically independent in computing acoustic likelihoods.

   • Reduced search space: The full search space is always pruned to a word graph (or an N-best list).
Knowledge of Confidence Measure – (Generalized WPP)

- The acoustic and language models weights are labeled as $\alpha$ and $\beta$, respectively.

- The corresponding GWPP is:

$$p([w; s, t] | x_T^1) = \sum_{\forall M, [w; s, t]_M^1} \prod_{m=1}^{M} p^\alpha(x_{s_m}^1 | w_m) \cdot p^\beta(w_m | w^M_{l}) \frac{p(x_T^1)}{p(x_T^1)}.$$
Knowledge of Confidence Measure – (Generalized UPP)

• Definition of the GUPP is similar to that of its word counterpart (GWPP). However, the time registration relaxation of beginning and ending of a utterance is no longer necessary since all string hypotheses share the same utterance boundaries.

• The generalized utterance posterior probability (GUPP) is defined as

\[
GUPP = \frac{p_a^\alpha \cdot p_l^\beta}{\sum_{\text{hypotheses}} p_a^\alpha \cdot p_l^\beta}
\]

where \( p_a \) is the acoustic model score and \( p_l \) the language model score of the hypothesized utterance; \( \alpha \) and \( \beta \), the acoustic and language model weights, respectively.

• The resultant GUPP is between 0 and 1 where a value close to 1 implies higher confidence on the correctness of an utterance.
Knowledge of Confidence Measure – (new UPP)

- The new way to measure the confidence of a recognized utterance is based on the joint confidence of all component words in the recognized string.

- GWPP of a word is a measure of its correctness, or a probability of a binomial distributed “word correct” event. The probability of an “utterance correct” event is then the product of all probabilities of component “word correct” events.

- The product of GWPPs of all recognized words in a recognized utterance is therefore proposed as a utterance level confidence as given below

\[ CF_{\text{sentence}} = \prod_{i=1}^{M} GWPP(w_i) \]

where \( M \) is the total number of words in the string hypothesis.
**Experiment – (Setup)**

- The speech corpus used for evaluation was a large vocabulary, continuous, read Chinese speech database in the Chinese Basic Travel Expression Corpus.

<table>
<thead>
<tr>
<th></th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># speakers</strong></td>
<td>4 M + 4 F</td>
<td>16 M + 16 F</td>
</tr>
<tr>
<td><strong># utterances</strong></td>
<td>841</td>
<td>3,437</td>
</tr>
<tr>
<td><strong># words</strong></td>
<td>4,030</td>
<td>16,781</td>
</tr>
<tr>
<td><strong># characters</strong></td>
<td>6,327</td>
<td>25,939</td>
</tr>
</tbody>
</table>

- Evaluation of verification performance was based on a normalized total verification errors – confidence error rate (CER).

\[
CER = \frac{\text{# false acceptance} + \text{# false rejection}}{\text{# recognized units}} \times 100\%
\]

- CER is 1 when all correctly recognized units are rejected and all incorrectly recognized units are accepted. A CER of 0 means that all units are correctly verified.
Experiment – (Results and Discussions)

• The figure shows the total error contours when word verification is carried out using the GWPP.

• In general, better verification performance is found near the lower left corner. The larger values of $\alpha$ and $\beta$ are used, more emphasis is put on higher ranked hypotheses. The smaller $\alpha$ and $\beta$ are, the more hypotheses are taken into account.

• So, more hypotheses are taken into consideration when computing the GPP making it more reliable.
Experiment – (Results and Discussions)

• The total error contours for utterance level verification are depicted in the Figure. The best verification performance is obtained when \( \alpha = 0.16 \) and \( \beta = 1.8 \).

• It is observed that the number of errors is very large along the y-axis where the language model weight is zero. Similar phenomenon is observed when the acoustic model weight is zero, or along the x-axis.

• These imply that neither the acoustic nor the language model score can be ignored when assessing the confidence of a recognized utterance using GUPP.
**Experiment – (Results and Discussions)**

- When the product of GWPPs of component words is used as the confidence measure for utterance verification, the total verification error contours are shown below.

- The optimal region lies close to the origin, where both $\alpha$ and $\beta$ are small. The best verification performance is obtained when $\alpha = 0.07$ and $\beta = 0.7$.
By applying verification using GPP, relative improvement in CER at utterance level using GUPP is higher (47.9%) than that of word level using GWPP (27.5%).

Furthermore, using product of GWPPs of component words in utterance verification can further reduce the already low CER. Comparing with the baseline, a 53.9% relative improvement in CER is obtained.
Experiment – (Setup and Results)

- Corpus: MATBN
- Total Length: 1.5hr

<table>
<thead>
<tr>
<th>Baseline-ML</th>
<th>Char</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>79.18</td>
<td>70.89</td>
</tr>
<tr>
<td>10</td>
<td>79.14</td>
<td>71.01</td>
</tr>
<tr>
<td>11</td>
<td>79.23</td>
<td>71.15</td>
</tr>
<tr>
<td>12</td>
<td>79.12</td>
<td>71.09</td>
</tr>
<tr>
<td>13</td>
<td>79.03</td>
<td>70.96</td>
</tr>
</tbody>
</table>
## Experiment – (Setup and Results)

<table>
<thead>
<tr>
<th>Cnormal</th>
<th>Char</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>79.29</td>
<td>71.05</td>
</tr>
<tr>
<td>10</td>
<td>79.22</td>
<td>71.08</td>
</tr>
<tr>
<td>11</td>
<td>79.32</td>
<td>71.27</td>
</tr>
<tr>
<td>12</td>
<td>79.24</td>
<td>71.22</td>
</tr>
<tr>
<td>13</td>
<td>79.16</td>
<td>71.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Csec</th>
<th>Char</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>75.94</td>
<td>66.75</td>
</tr>
<tr>
<td>10</td>
<td>76</td>
<td>66.91</td>
</tr>
<tr>
<td>11</td>
<td>75.98</td>
<td>67</td>
</tr>
<tr>
<td>12</td>
<td>75.94</td>
<td>66.99</td>
</tr>
<tr>
<td>13</td>
<td>75.8</td>
<td>66.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cmed</th>
<th>Char</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>76.33</td>
<td>67.27</td>
</tr>
<tr>
<td>10</td>
<td>76.4</td>
<td>67.48</td>
</tr>
<tr>
<td>11</td>
<td>76.37</td>
<td>67.54</td>
</tr>
<tr>
<td>12</td>
<td>76.3</td>
<td>67.47</td>
</tr>
<tr>
<td>13</td>
<td>76.21</td>
<td>67.33</td>
</tr>
</tbody>
</table>