Fusing Multiple Confidence Measures for Chinese Spoken Term Detection

Zejun Ma¹, Xiaorui Wang¹, Bo Xu¹,²

Digital Content Technology Research Center¹, National Lab of Pattern Recognition², Institute of Automation, Chinese Academy of Sciences, Beijing, 100190

{zjma,xrwang,xubo}@hitic.ia.ac.cn

1. Introduction

The ever growing volume of recorded speech data collected from telephones, cell phones, podcasts and internet conversations etc, has necessitated the development of spoken language processing technologies. Finding instances of a particular spoken term in audio archives has a long history and many names, including “keyword spotting”, “audio indexing”, “spoken term detection”, etc. Recently, approaches that couple speech-to-text(STT) technology with traditional text-matching technique have been more successful than other methods. However, such approaches severely suffer from rather high missing rate and are restricted by the vocabulary of STT system. Thus, the lattice instead of one-best recognition output is searched to compensate for poor recall [5][12]; the sub-word lattice instead of word lattice is used to avoid Out-Of-Vocabulary(OOV) issue [6].

In real-world STD applications, the confidence measures(CM) is vital to assess the reliability of searched terms. CMs are usually obtained during decoding procedure by collecting some information related to acoustics and languages as well as other features that can be useful to generate an indicator of correctness of putative terms. The most commonly used CM is based on the posterior probability computed from lattices using the efficient forward-backward algorithm. Due to the highly erroneous sub-word recognition, the CM estimated from the resulting lattice may not be reliable. To overcome this, we propose to improve the confidence scoring by using different knowledge: utilizing neural network to estimate acoustic posterior of term given acoustic observations and using phonetic duration model to estimate the similarity between the expected and actual duration distributions of term [9]. We believe these heterogeneous confidence estimation methods introduce complementary knowledge about the reliability of hypotheses and thus it is probable to obtain a more robust confidence if we use these knowledge appropriately. The main contributions of this paper include: firstly, we build a two-pass STD system. The termination confidence is re-scored at a backend module which is completely independent of indexing and searching module. Secondly, we discuss multiple confidence measures in detail and combine them to produce a more discriminative CM.

This paper is organized as follows: section 2 introduces the architecture of our STD system used in this paper. Section 3 discusses various confidence measures related to this paper. The basic definition of lattice confidence is simply recapitulated in subsection 3.1; the confidence based on acoustic posterior estimated by hybrid HMM/ANN is presented in subsection 3.2; the duration confidence based on phonetic duration distribution is introduced in subsection 3.3. In section 4, the detailed experimental setups and results are presented. Finally, our works are concluded in section 5.

2. Overview of system architecture

The architecture of our Mandarin STD system is as shown in figure 1. The system consists of four modules:

- The sub-word (syllable) decoder transcribes speech utterance to syllable lattice. The resulting lattice encodes the information about multiple hypotheses paths which survive during decoding.
- The indexer extracts information embedded in lattices and reserves distilled information in a rapidly accessible inverted index. The inverted index contains a sorted list of key-value pairs with the identity of index unit being the key and the occurrence details of current index unit being the corresponding value. After indexing operation, lattices can be discarded as all required information are already kept in indices.
- In query time, the searcher scans the processed indices to determine the putative occurrences of input query terms using text-matching techniques.
- The backend rescorer gives the estimation of acoustic and duration confidences. This module then comprehensively considers these measures with lattice confidence computed during searching procedure to make a final decision of the confidences for the putative terms.
3. Various confidence measures

3.1. Lattice based spoken term detection

In this section, the basic principle of lattice confidence based spoken term detection is reviewed. In speech recognition, the lattice is often used to keep the information about active hypotheses paths during decoding procedure.

A lattice is typically represented as a directed acyclic graph (DAG) which is comprised of a set of nodes $N$ and directed arcs $A$. Each lattice has an unique start node $n_{start}$ and unique end node $n_{end}$, denoting any word sequence before $t_s$ and after $t_e$, respectively, $W$ is computed utilizing the well-known forward-backward algorithm.

The term level acoustic confidences based on posteriors defined in Eq.(3) and Eq.(4) are compared in figure 2. The interesting observation is that the confidence based on double normalized posterior $PP_{wd}(W)$ obviously superior to the one based on usual $PP_{wd}$ as shown in DET curves. Thus if without specifica- tion, the acoustic confidence means the one based on double normalized acoustic posterior in the rest of this paper.

3.2. Duration confidence

In previous section 3.2, we go along the optimal HMM state sequence to obtain acoustic posterior. The phonetic durations can be also obtained during the forced alignment of state sequence with feature frames. Inspired by [9], we agree that the duration confidence based on double normalized posterior $PP_{wd}$ is computed utilizing a double normalization which takes both the number of frames in each phone and the number of phones in each term/word into account, yielding the following estimation.

$$
\log PP_{phn}(k) = \frac{1}{e-b+1} \sum_{n=b}^{e} \log PP(p_k^n|x_n) 
$$

(2)

We can compute a term or word level posterior using similar mechanism. For a term $W$ consisting of a sequence of $N$ phones $(p_1, p_2, ..., p_N)$, the term posterior is defined according to Eq.(3). The logarithm of global term posterior is normalized by the total frame numbers of term $W$.

$$
\log PP_{wd}(W) = \frac{1}{N} \sum_{k=1}^{N} \left( \frac{1}{e_k - b_k + 1} \sum_{n=b_k}^{e_k} \log PP(p_k^n|x_n) \right)
$$

(3)

At the term/word level, we also consider another posterior definition. The term posterior $PP_{wd}(W)$ is computed utilizing a double normalization which takes both the number of frames in each phone and the number of phones in each term/word into account, yielding the following estimation:

$$
\log PP_{wd}(W) = \frac{1}{N} \sum_{k=1}^{N} \left( \frac{1}{e_k - b_k + 1} \sum_{n=b_k}^{e_k} \log PP(p_k^n|x_n) \right)
$$

(4)

The term level acoustic confidences based on posteriors defined in Eq.(3) and Eq.(4) are compared in figure 2. The interesting observation is that the confidence based on double normalized posterior $PP_{wd}$ is computed utilizing a double normalization which takes both the number of frames in each phone and the number of phones in each term/word into account, yielding the following estimation.

$$
\log PP_{wd}(W) = \frac{1}{N} \sum_{k=1}^{N} \left( \frac{1}{e_k - b_k + 1} \sum_{n=b_k}^{e_k} \log PP(p_k^n|x_n) \right)
$$

(4)
right-context biphone, is considered. If all context-dependent triphone or biphone are not available, the duration of monophone is used. Supposing that term \(T\) contains \(N\) number of phones \((p_1, p_2, \ldots, p_N)\) and the actual duration distribution \(T(W) = \{t_1, t_2, \ldots, t_N\}\) is obtained during the forced alignment. The detailed procedure to obtain reference distribution \(R(W)\) is presented as follows: for each \(p_k \in W\), the left and right neighboring context of \(p_k\) are combined to construct a triphone \(T_k = \langle p_{k-1}, p_k, p_{k+1} \rangle\). The triphone has the form as shown in Eq.(5) when \(p_k\) is the first or last phone of term \(W\) where symbol \(\text{sil}\) represents the term boundary. Similarly, the left and right context biphone of \(p_k\) can be written as \(B_{\text{left}}^k = \langle p_{k-1}, p_k \rangle\) and \(B_{\text{right}}^k = \langle p_k, p_{k+1} \rangle\) respectively:

\[
T = \begin{cases} 
\langle \text{sil}, p_k, p_{k+1} \rangle & \text{if } k = 1 \\
\langle p_{k-1}, p_k, \text{sil} \rangle & \text{if } k = N 
\end{cases} \quad (5)
\]

It is supposed that \(d(.)\) and \(c(.)\) denote the duration and count of specific unit respectively; both statistics are obtained from training corpora. \(b(.)\) represents the backoff factor of unit which is set to 1 in whole paper. Then, the reference duration of triphone \(T_k\) can be computed as Eq.(6):

\[
r(T_k) = \begin{cases} 
d(T_k) & \text{if } c(T_k) > 0 \\
b(p_{k+1}) \cdot d(B_{\text{left}}^k) & \text{if } c(T_k) = 0 \text{ and } c(B_{\text{left}}^k) > 0 \\
b(p_{k-1}) \cdot d(B_{\text{right}}^k) & \text{if } c(T_k) = c(B_{\text{left}}^k) = 0 \text{ and } c(B_{\text{right}}^k) > 0 \\
b(< p_{k-1}, p_k, p_{k+1} >) \cdot d(p_k) & \text{if } c(T_k) = c(B_{\text{left}}^k) = c(B_{\text{right}}^k) = 0 
\end{cases} \quad (6)
\]

After traversing through term \(W\), the expected distribution \(R(W) = \{r_1, r_2, \ldots, r_N\}\) is obtained. The duration confidence can in turn be computed as a function of distance between distribution \(T(W)\) and \(R(W)\). Both \(T(W)\) and \(R(W)\) are normalized to compensate for different speaking rates. The distance metric used is the Jeffries-Matusita distance:

\[
D(T(w), R(W)) = \frac{1}{N} \sum_{i=1}^{N} (\sqrt{d_i} - \sqrt{r_i})^2 
\]

3.4. Combination of confidences

A set of confidence estimation methods were already discussed in previous sections. These confidences are interpolated in linear manner as shown in Eq.(8). For convenience, we use \(CM_{\text{lat}}, CM_{\text{ac}}\) and \(CM_{\text{dur}}\) to denote the lattice, acoustic and duration confidence respectively in the rest of this paper.

\[
CM_{\text{comb}} = w_1 \cdot CM_{\text{lat}} + w_2 \cdot CM_{\text{ac}} + w_3 \cdot CM_{\text{dur}} \quad (8)
\]

where \(w_1, w_2, w_3\) denote the weight of each confidence respectively. More specifically, the lattice and acoustic confidence are both computed from their counterpart posteriors normalized by term’s average duration \(\Delta_{\text{avg}}(\text{term})\) and phone number \(N(\text{term})\) where parameter \(\gamma\) is a scaling factor which is tuned on development set, like shown in Eq.(9). Finally, all confidences are scaled to range \([0-100]\) before interpolation.

\[
CM_{\text{lat/nc}} = \gamma P P_{\Delta_{\text{avg}}(\text{term})} N(\text{term}) \quad (9)
\]

4. Experiments

4.1. Databases

Three Chinese Mandarin databases namely the 863 evaluation, the CallHome Mandarin evaluation and a corpus called the TEST8h evaluation are used to assess our STD system performance. All utterances in three databases are conversational telephone speeches in spontaneous style. The details about three databases are presented in Table 1.

<table>
<thead>
<tr>
<th>Properties</th>
<th>863eva</th>
<th>CallHome</th>
<th>TEST8h</th>
</tr>
</thead>
<tbody>
<tr>
<td>durations (in hour)</td>
<td>1.0</td>
<td>1.0</td>
<td>8.0</td>
</tr>
<tr>
<td>keyword numbers</td>
<td>100</td>
<td>100</td>
<td>122</td>
</tr>
<tr>
<td>occurrence numbers</td>
<td>398</td>
<td>372</td>
<td>1184</td>
</tr>
</tbody>
</table>

4.2. Setups

In signal processing module, the analysis frame length and shift are 25ms and 10ms respectively. The speech frames are parameterized as PLP (Perceptual Linear Prediction) features. The 39 dimensional feature includes 12 PLP coefficients plus energy with their first-order and second-order derivatives. The cepstral mean and variance normalization are also applied.

The acoustic model is a context-dependent HMM model. Each tied-state triphone model takes the left-to-right topology with 3 emitting states and the pdf (probability distribution function) of each state consists of 32 diagonal-covariance Gaussian mixtures. The bigram syllable language model is used for the sub-word based decoder.

The system performance metric EER is defined as a point in DET curve where FA(False Alarm) rate equals to FR(False Reject) rate. The FA and FR are defined as:

\[
FA = \frac{\#fa}{\#kw \cdot \text{hour} \cdot \text{m}} \times 100\% \quad (10)
\]

\[
FR = \frac{\#fr}{N} \times 100\% \quad (11)
\]

where \(\#fa\) and \(\#fr\) stand for the number of false alarms and rejects respectively. \(\#kw\) is the number of detected instances.
4.3. Results

In this section, we apply the confidence measures discussed in this paper to spoken term detection task. The experiments are conducted on three evaluation sets and the DET curves are presented in Fig.(3). As shown in DET curves, the duration confidence has the worst discriminative ability and the reason might be the highly variable properties of duration distribution caused by the different speaking rates which are easily influenced by the age, gender, mood, and pronunciation fluency of speakers. On the other hand, it can be observed that the lattice confidence is substantially superior to the other two measures, i.e., the acoustic and phonetic duration confidences. The observation is consistent with the conclusion “ranking term by lattice posterior is theoretically optimal for word spotting task” in other works [5][12]. Although the acoustic and duration based confidences produce a DET curve significantly worse than the lattice confidence, however, it seems that these two confidences are complementary to the lattice one. The linear combination of the three confidences substantially improves the system confidence scoring mechanism and the combined one consistently outperforms the baseline lattice confidence on all three evaluation sets. At last, the comparison of the EER performance of our STD system using different confidence measures are summarized in Table 2.

6. References