Utilizing State-level Distance Vector Representation for Improved Spoken Term Detection by Text and Spoken Queries

Mitsuki Makino\textsuperscript{1}, Naoki Yamamoto\textsuperscript{2}, Atsuhiko Kai\textsuperscript{2}

\textsuperscript{1} Faculty of Engineering, Shizuoka University, Japan
\textsuperscript{2} Graduate School of Engineering, Shizuoka University, Japan
makino@spa.sys.eng.shizuoka.ac.jp, yamamoto nao@spa.sys.eng.shizuoka.ac.jp, kai@sys.eng.shizuoka.ac.jp

Abstract

In spoken term detection (STD) systems, approximate subword-level matching of query term and automatically transcribed spoken documents is often employed for its reasonable accuracy and efficiency. However, high out-of-vocabulary (OOV) rate often degrades the subword-level recognition accuracy and affect the STD performance. This paper describes the usage of new expanded acoustic representations of subword sequence for improved scoring between OOV query term and subword-unit transcription. Each subword is expanded in corresponding subword’s HMM states and each state is represented as a new acoustic structural feature, a distribution-distance vector (DDV). The proposed DDV representation and scoring is easily combined with two typical baseline STD approaches: a DTW-based approximate matching with subword-level acoustic dissimilarity measure and a lattice-based confidence scoring of subword n-grams. The experimental result showed that the proposed DDV-based scoring method significantly outperforms the simple DTW-scoring baseline with very little increase in the required search time. The combination of the DDV-based scoring with the confidence-based scoring showed the complementary effect and attained the best STD performance compared with the NTCIR-10 SpokenDoc2(SDPWS) submitted results when only the NTCIR reference automatic transcript is used. A preliminary experiment with spoken query terms also showed that the significant improvement for OOV queries.

Index Terms: Spoken term detection, distance between two distributions, distance measure between two structures, acoustic similarity, spoken query

1. Introduction

Spoken term detection (STD) is a task which locates a given search term in a large set of spoken documents. To deal with out-of-vocabulary (OOV) problems and recognition errors, many approaches using a subword-unit based speech recognition system have been proposed\cite{1, 2, 3, 4}. The approximate matching and spotting approaches based on dynamic time warping (DTW)-based matching or n-gram scoring of subword sequences have shown the robustness for recognition errors and OOV problems\cite{5, 6}.

In this paper, we investigate the STD approaches which introduce a new feature representation and acoustic dissimilarity measure which are derived solely from a set of subword-unit acoustic models for robust subword-level approximate matching. The STD system based on this new representation, which we call distribution-distance vector (DDV), has shown a significant improvement compared with simple DTW-based matching approaches\cite{7}. However, our previous study has not been compared or combined with another n-gram and confidence-based approach which also seems to robust for subword-level errors\cite{8}.

Related works using the acoustic similarity for STD task are roughly divided into two types: STD systems for text query input (e.g. \cite{9}) and those for spoken query input or unsupervised spoken keyword spotting (e.g. \cite{8, 11, 12}). Typically, the former systems use certain information about confusability between subwords. In \cite{6}, a syllable-level distance measure based on the Bhattacharyya distance derived from syllable-unit HMMs is used. Though our proposed acoustic measures is also based on subword-unit HMMs, the state-level DDV feature representation and local distance instead of subword-level one is used for evaluating the match between query and subword sequences. The DDV feature representation is related to the idea of using an invariant structural feature for removing acoustic variations caused by non-linguistic factors\cite{13, 14} and it is expected that the proposed feature is effective for erroneous transcriptions. Recently, similar idea of using structural feature for acoustic dissimilarity estimation is effectively applied to the systems of latter type\cite{12}.

In this study, the experiments were conducted on the NTCIR SpokenDoc and SpokenDoc-2 STD subtasks\cite{9, 10} which target two document collections: the Corpus of Spontaneous Japanese (CSJ) and the Corpus of Spoken Document Processing Workshop (SDPWS). This paper presents the experimental result for several STD systems which includes the combination of two approaches: DDV-based scoring and the n-gram confidence-based scoring. The result shows that the combination approach yields complementary effect and attain the best STD performance compared with the NTCIR-10 SpokenDoc2(SDPWS) submitted results when only the NTCIR reference automatic transcript is used. Also, a preliminary result for spoken query terms is presented and reveals that the proposed approach is effective for OOV queries and degraded documents with many recognition errors.

2. Baseline spoken term detection system

The baseline system adopts a DTW-based spotting method which performs matching between subword sequences of query term and spoken documents and outputs matched segments. In NTCIR-9 SpokenDoc STD baseline system\cite{9}, a similar system with the local distance measure based on phoneme-unit edit distance is used. In our baseline system, the local distance measure is defined by a syllable-unit acoustic dissimilarity as used in \cite{6}. The distance between subwords $x$ and $y$, $D_{\text{sub}}(x, y)$, is calcu-
3. Proposed spoken term detection method

3.1. Proposed system overview

Overview of our proposed STD system is shown in Fig. 1. The system adopts two-pass strategy for both efficient processing and improved STD performance against recognition errors. One of the first pass methods simply performs the DTW-based query term spotting as described in Section 2. The second pass is a query term verifier which performs two kinds of detailed scoring (rescoring) for each candidate segment found in the first pass. The different approaches to scoring segments at the first and second passes and their combinations are described in the following sections.

3.2. N-gram confidence-based scoring

For finding the occurrence of certain subword sequence from the lattice, n-gram confidence-based scoring has been effectively used to deal with the recognition error problem[8]. We adopts the n-gram confidence-based scoring method as an alternative method of the baseline (first-pass only) system or as an additional filtering process which precedes or follows the two-pass spotting and rescoring passes mentioned in the previous section. In the latter case, the relevance score is compared with a threshold parameter to filter out unlikely speech segments before or after the two-pass match is performed.

Let the \( Q = \{w_1, \cdots, w_M\} \) be the subword sequence of a query term and \( \{w_i, \cdots, w_{i+n-1}\} \) denote partial n-grams of the query term. We define the relevance score \( R_n\)-gram of speech segment \( X \) and query term \( Q \) for each order of \( n \) as

\[
R_n\text{-gram} = \sum_{i=1}^{M-n+1} CM(W)C(W, \{w_i, \cdots, w_{i+n-1}\})
\]

(2)

where \( C(W, \{w_i, \cdots, w_{i+n-1}\}) \) is the occurrence count of n-gram \( \{w_i, \cdots, w_{i+n-1}\} \) in sentence hypothesis \( W \) which is included in subword lattice \( W(X) \), and \( CM(W) \) denotes the confidence score of sentence \( W \) as the posteriori probability in lattice \( W(X) \). The final relevance score is obtained by

\[
Score_{CM}(X, Q) = \sum_{n=1}^{N} a_n R_n\text{-gram}
\]

(3)

where \( a_n \) is a weight parameter. In practice, Eq. (2) is equivalently calculated by efficient forward-backward algorithm from a subword lattice[15].

3.3. Rescoring with state-level representation (2nd pass)

As described in Section 2, the first-pass query term spotting performs DTW-based matching by using the subword-level local distance metric \( D_{\text{dtw}}(x, y) \). The output is a set of aligned subword sequences which have the dissimilarity score below a threshold. The second pass first expands the aligned subword sequences into the corresponding HMM state sequences and calculates dissimilarity score based on a state-level local distance metric.

A simple approach to calculate dissimilarity score between HMM state sequences is the DTW matching based on the local distance measure defined in (1). The dissimilarity scores obtained for each candidate segments are compared with a threshold. We refer to this dissimilarity score as \( Score_{BD} \).

Our previous study introduced new acoustic dissimilarity score based on a distance-vector representation which is defined for each HMM state. Like a structural feature representation proposed in [13] and a self similarity matrix in [12], we can consider a feature representation for each HMM state based on the distances between a target state and all states in a set of subword-unit HMMs. It is expected that such structural feature can estimate more robust acoustic dissimilarity measure for comparing the subword sequences including recognition errors.

Let the \( P = \{P_s\} \) be a set of all distributions in subword-unit HMMs. We define a distance vector for the HMM state \( s \) as

\[
\phi(s) = (D_{BD}(P_s, P_1), D_{BD}(P_s, P_2), \cdots, D_{BD}(P_s, P_S))^T
\]

(4)

We refer to this vector representation as distribution-distance vector (DDV).

We can replace the local distance measure used by the state-level DTW-based matching mentioned above with the L2 (Euclidean) norms of DDV representation in (4). The accumulated distance obtained by DTW matching is used as a dissimilarity score because it take a value closer to zero as two HMM state sequences become acoustically similar. We refer to this dissimilarity score as \( Score_{DDV,DL2} \).
To simplify the calculation of dissimilarity score using the DDV representation, we can utilize the alignment between two state sequences obtained as a result of calculating $Score_{BD}$. Let the $F = \{a_1, a_2, \ldots, a_i\}$ be the state-level alignment and the $c_k = (a_i, b_j)$ represents the correspondence between the $i$-th state in HMM state sequence $A = a_1, a_2, \ldots, a_i$ and the $j$-th state in HMM state sequence $B = b_1, b_2, \ldots, b_j$. We investigate the following two alternative definitions based on the DDV representation.

\[ Score_{DDV,L2} = \frac{1}{K} \sum_{k=1}^{K} \left( \frac{1}{S} \sum_{s=1}^{S} (\psi_s(c_k))^2 \right)^{1/2} \]  \hspace{2cm} (5)

\[ Score_{DDV,L2max} = \frac{\max \{ \sum_{s=1}^{S} (\psi_s(c_k))^2 \}^{1/2}}{S} \]  \hspace{2cm} (6)

where $\psi_s(c_k)$ is the $s$-th element of the vector $\phi(a_i) - \phi(b_j)$. $Score_{DDV,L2}$ represents a normalized score of accumulated L2 (Euclidean) norms (although not strictly L2 norm since a normalization term $1/S$ is included). On the other hand, $Score_{DDV,L2max}$ uses the maximum value of all L2 norms in a DDV sequence and thus it emphasizes the most dissimilar part in a subword sequence.

Finally, the above mentioned dissimilarity scores based on the state-level representations can be combined as

\[ Score_{fusion} = \alpha \cdot Score_{BD} + (1 - \alpha) \cdot \tau \cdot Score_{DDV} \]

where $0 \leq \alpha \leq 1$ is a weight coefficient and $\tau$ is a constant for adjusting the score range. To reduce the computational cost, the local distance values between states can be prepared beforehand by using a set of subword-unit HMM parameters.

### 4. Experiments

#### 4.1. Experimental setup

In the evaluation experiment of STD with text queries, we verify the robustness of the proposal method by using two target document collections used in the NTCIR SpokenDoc tasks[9, 10]: the CORE set of the Corpus of Spontaneous Japanese (CSJ-CORE, 177 lectures, about 44 hours) and the lectures of Spoken Document Processing Workshop (SDPWs, 107 lectures, about 29 hours). As with NTCIR-9 SpokenDoc STD evaluation[9], the Inter-Pausal Units (IPU) are used as the basic unit to be searched and a retrieval result of an IPU is regarded as correct if it includes the query term. For the task of CSJ-CORE target document collection, we use IV and OOV query terms (50 terms, respectively) which are defined as the dry-run of NTCIRC-9 SpokenDoc subtask[9]. The number of relevant documents for IV and OOV queries are 742 and 234, respectively. For the task of SDPWs target document collection, we use 100 query terms including OOV terms which were used for the formal-run in the NTCIR-10 SpokenDoc-2 SDPWs(moderate-size) task[10]. The number of relevant documents for IV and OOV queries are 443 and 459, respectively. We used both of word-based and syllable-based reference automatic transcriptions distributed at NTCIR-9 and NTCIR-10 evaluations. These reference automatic transcriptions include N-best results (N=10) and lattices using a triphone acoustic model and word/syllable n-gram language models. The syllable-unit accuracies for word-based automatic transcriptions were 83.0% for CSJ-CORE and 75.3% for SDPWs, respectively. The syllable-unit accuracies for syllable-based automatic transcriptions were 77.4% for CSJ-CORE and 67.7% for SDPWs, respectively.

#### Table 1: Effect of rescoring approaches with state-level representation (CSJ-CORE documents)[%]

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>IV</td>
<td>61.05</td>
<td>89.88</td>
</tr>
<tr>
<td></td>
<td>OOV</td>
<td>37.61</td>
<td>77.88</td>
</tr>
<tr>
<td><strong>Score_{BD}-only</strong></td>
<td>IV</td>
<td>64.29</td>
<td>91.73</td>
</tr>
<tr>
<td>((\alpha = 1))</td>
<td>OOV</td>
<td>46.15</td>
<td>78.26</td>
</tr>
<tr>
<td><strong>Score_{DDV,L2max}-only</strong></td>
<td>IV</td>
<td>62.53</td>
<td>91.88</td>
</tr>
<tr>
<td>((\alpha = 0))</td>
<td>OOV</td>
<td>37.61</td>
<td>73.95</td>
</tr>
<tr>
<td><strong>Score_{fusion}(Score_{DDV,L2})</strong></td>
<td>IV</td>
<td>64.15</td>
<td>92.97</td>
</tr>
<tr>
<td><strong>Score_{fusion}(Score_{DDV,L2max})</strong></td>
<td>IV</td>
<td>64.02</td>
<td>92.59</td>
</tr>
<tr>
<td></td>
<td>OOV</td>
<td>50.00</td>
<td>81.25</td>
</tr>
</tbody>
</table>

In the evaluation experiment of STD with spoken queries, we used the CSJ-CORE document collection. For the spoken query terms, 21 OOV query terms were selected from 50 OOV query terms which were used for the dry-run (CSJ-CORE set) in the NTCIR-9 SpokenDoc STD subtask. The number of relevant documents for OOV queries is 120. The speech intervals of the query terms are manually extracted from CSJ lectures which are not a part of CSJ-CORE lectures. The spoken query terms are automatically transcribed by speech recognition system with a syllable-unit n-gram language model trained by noun phrases. The syllable-unit accuracy for syllable-based automatic transcription was 63.2%.

As measures of search performance, we use recall, precision, and F-measure(max). F-measure(max) is the maximum value of F-measure when the threshold is adjusted.

#### 4.2. Comparison of rescoring approaches (2nd pass)

Table 1 compares the STD performance of the baseline system (first-pass only) and the two-pass systems with different rescoring methods for CSJ-CORE target document collection. All systems except baseline and Score_{BD}-only are based on the combined score described in Section 3.3. The parameters of the first-pass threshold and a weight coefficient for the combined score were optimized for the target document.

The result shows that the two-pass systems outperform the baseline system which uses only the first pass. The second pass which combines new acoustic dissimilarity measure based on the DDV representation could reject more unreliable candidates and improved the performance. Also, the results shows that two-pass method with Score_{DDV,L2max} attains the best performance to OOV query. Therefore, only the results for the second pass which combines Score_{DDV,L2max} are presented as the two-pass system with the combined score (referred to as BD-DDV method) in the following sections.

Fig.2 shows the effect of changing the weight coefficient $\alpha$. The result shows that the influence of the weight parameter $\alpha$ is small for IV queries. But for OOV queries, the combined weights of around $\alpha = 0.4$ improve the performance. Also, the result shows that the performance gain over the baseline is significant even for $\alpha = 1.0$. This indicates that the state-level representation is effective for matching a long subword sequence.

#### 4.3. Comparison of n-gram confidence-based method and its combination

Fig. 3 shows the result for all methods described in Section 3. CM denotes the system with n-gram confidence-based method
Precision [%] 

F-measure performance was higher for selections. In contradiction to the case of CSJ-CORE documents, the difference of recognition accuracy of target document collections indicates that CM method as a final decision stage is sensitive to the threshold.

On the other hand, the result of CM and BD-DDV methods indicates that BD-DDV scoring method is robust to the difference of recognition accuracy of target document collections. As a result, CM+BD-DDV attained the best performance. The result indicates that BD-DDV scoring method is robust to the difference of recognition accuracy of target document collections. On the other hand, the result of CM and BD-DDV+CM methods indicates that CM method as a final decision stage is sensitive to the difference of recognition accuracy of target document collections. In contradiction to the case of CSJ-CORE documents, F-measure performance was higher for $\alpha = 0$ than for $\alpha = 1.0$, which also indicates the robustness of the DDV-based scoring.

This result shows that CM method alone outperforms the baseline and comparable performance with BD-DDV method. This might be the relatively high accuracy of the automatic transcription of CSJ-CORE documents. As a result, CM+BD-DDV method which combines n-gram confidence-based and DTW-based matching attained the best performance.

Fig. 4 shows the result for SDPWS documents using the parameter which is optimized for the CSJ-CORE documents. Unlike the case of CSJ-CORE documents, BD-DDV method outperforms baseline and CM method. Also, the combined method CM+BD-DDV attained the best performance. The result indicates that BD-DDV scoring method is robust to the difference of recognition accuracy of target document collections. On the other hand, the result of CM and BD-DDV+CM methods indicates that CM method as a final decision stage is sensitive to the difference of recognition accuracy of target document collections. In contradiction to the case of CSJ-CORE documents, F-measure performance was higher for $\alpha = 0$ than for $\alpha = 1.0$, which also indicates the robustness of the DDV-based scoring.
6. References