Effects of End-to-end ASR and Score Fusion Model Learning for Improved Query-by-example Spoken Term Detection

Takumi Kurokawa* , Atsuhiko Kai* and Hiroki Kondo*
* Graduate School of Integrated Science and Technology, Shizuoka University, Japan
E-mail: kurokawa@spa.msys.eng.shizuoka.ac.jp
E-mail: kai.atsuhiko@shizuoka.ac.jp
E-mail: kondo_hiro@spa.msys.eng.shizuoka.ac.jp

***Abstract*—Query-by-example spoken term detection (STD) systems can make effective use of automatic speech recognition (ASR), especially in situations where the recognition accuracy is high. However, out-of-vocabulary (OOV) problem at the ASR stage has a significant impact on the performance of STD for speech retrieval and can often occur for query terms. Recent studies have shown that end-to-end (E2E) ASR systems can achieve competitive performance compared to conventional DNN-HMM-based ASR systems and reduce the impact of OOV problem by adopting output units of characters or subwords. This paper proposes to apply E2E ASR system in an STD method that considers acoustic similarity at sub-phone level, and to combine it with the DNN-HMM-based ASR and auxiliary information by a score fusion method. Experimental results on the NTCIR-12 SpokenQuery&Doc-2 task showed that the STD method using the hybrid CTC/Transformer E2E ASR improved the search performance over the STD method using the DNN-HMM-based ASR. The best detection performance was obtained using a score fusion model, demonstrating that combining E2E ASR and auxiliary information with DNN-HMM-based ASR is effective for both known and OOV word queries.**

I. Introduction

Spoken term detection (STD) is a task to search for a given search term from a large amount of speech documents. When the accuracy of automatic speech recognition (ASR) is reasonably high, the method of utilizing its output in an STD system has been shown to achieve high processing efficiency and detection accuracy [1], [2]. ASR-based STD approaches often use the dynamic time warping (DTW) algorithm to efficiently perform approximate matching (spotting) with a given query for searching very large target speech documents. Since a typical hierarchical ASR system can output a subword sequence corresponding to the decoded word string, subword-level DTW methods have been employed effectively to mitigate the effects of ASR errors. Previous studies have proposed subword-level or subphone-level DTW methods using various local distance metrics to reduce the impact of recognition errors [2], [3], [4].

Another challenge in the ASR-based STD approaches is the problem of out-of-vocabulary (OOV) words. In general, search tasks tend to involve OOV words in the query terms, and previous studies have shown that there remains a significant performance gap between the in-vocabulary (IV) and OOV word queries. A possible solution to this problem is to use feature-based acoustic matching methods, which is commonly used in low-resource STD task scenarios [5], [6]. The feature-based matching approaches have also shown to be effective when they were combined with the ASR-based STD system. In our previous study [7], a combined HMM state-level DTW from DNN-HMM-based ASR output and feature-level match along with a score fusion method could improve the STD performance for OOV queries. However, the issue of relative performance degradation for OOV word queries has not yet been resolved.

Recent ASR research has focused on end-to-end neural networks, such as encoder-decoders, which can learn input-output mappings without explicitly having predefined knowledge of lexicon or phonetic pronunciation of words. In [8], [9], the end-to-end (E2E) ASR approaches that use characters or subwords as output units has achieved performance comparable to conventional DNN-HMM-based hierarchical large-vocabulary ASR systems in recipes that utilize large training corpora. The character or subword level decoding capability of the E2E ASR system is also expected to improve the degradation of STD performance due to the OOV problem. Recently, a model structure called Transformer has been proposed. This method, proposed for a machine translation task [10], uses a self-attention approach. This approach achieved reduced learning time and improved performance. This method is also used in the speech recognition task and has shown performance as good as bi-directional LSTM based Encoder-Decoder [11]. It has also shown that hybrid CTC/Transformer ASR systems perform better than DNN-HMM-based ASR systems in rich resource scenarios [9].

In this paper, we extend our previous work [7] on ASR-based STD systems which have adopted a subphone-level local distance metric derived from acoustic model. We propose ASR-based STD methods to improve the performance by introducing the hybrid CTC/Transformer end-to-end ASR (E2E ASR). First, we present an ASR-based STD method that introduces the E2E ASR system by incorporating a grapheme-to-phoneme conversion pre-processor. We compare the impact

of the recognition performance of the DNN-HMM-based ASR or E2E ASR system alone on STD using DTW-based spotting with acoustic dissimilarity derived from acoustic models. Next, we propose multi-ASR based STD methods combined with a score fusion method using a logistic regression model. We also compare the effect of training the logistic regression model with additional auxiliary information of ASR output which has been shown to be effective in [7]. The experiments are conducted on the NTCIR-12 SpokenQuery&Doc-2 task [12] for a spoken document collection and the results are compared with conventional STD systems and a NTCIR-12 organizer's baseline result.

II. Baseline spoken term detection system

In this study, we compares the STD approaches that utilize the output of ASR system. In the NTCIR-12 SpokenQuery&Doc-2 task, a baseline STD method provided by the task organizer (hereafter referred to as NTCIR-12 baseline) is based on a DTW-based STD method that performs approximate matching between subword sequences mapped from the ASR output of each of the spoken documents and a query [12]. We present the NTCIR-12 baseline performance as one of the compared results in section IV.

Fig. 1 shows an overview of STD systems, which illustrates the common components of our baseline and proposed STD methods. While the NTCIR-12 baseline uses subword-level edit distance as a measure of local distance, our previous works [3], [4] showed that employing subphone-level DTW with acoustic dissimilarity derived from an HMM-based acoustic model, even at the same ASR output, could significantly improve the STD performance. In [7], [13] the STD method combined with a score fusion model trained with a separate set of development data could achieve the best STD performance, and showed the effectiveness of additional ASR-derived features regarding the confidence measure and query length. Therefore, we incorporate most of the above features into our baseline STD system as illustrated in Fig. 1.

First, the ASR output is converted to the corresponding triphone sequence and then to a representation of the corresponding HMM state sequence (hereafter referred to as the triphone state sequence), omitting the loop transitions. Thus, spoken documents and spoken queries are converted to a representation of the state sequence that is finer than the phone units and shorter than the number of speech feature frames. Next, endpoint-free DTW processing is performed using a local distance metric between states, calculated by considering their state output distributions, as an acoustic dissimilarity. For the local distance metric between states, we compare two different methods that are calculated solely from the parameters of the acoustic model used in the ASR part. One is the GMM-based method used in our previous work [7], and the other is the DNN-based method proposed in [14], respectively, which are explained in the next subsections.

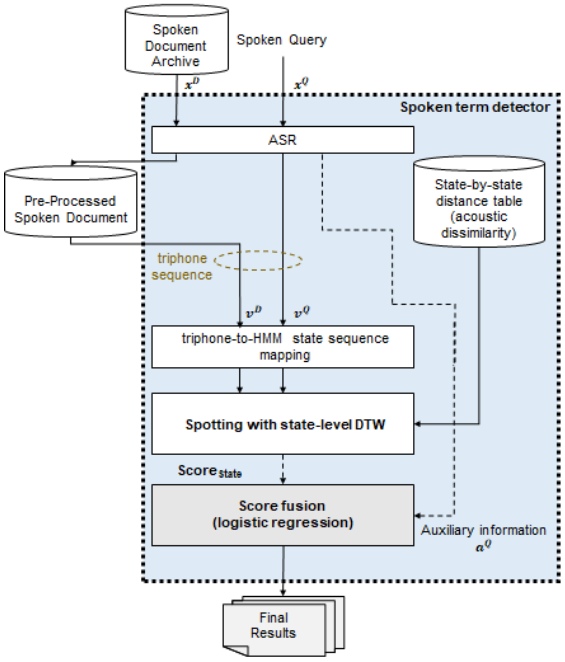

Fig. 1. Configuration of STD system common to both our baseline and proposed methods

A. Bhattacharyya Distance (BD) as Local Distance Metric

Conventional hierachical ASR systems construct a context-dependent acoustic model as stochastic generative model (typically triphone-unit GMM-HMM acoustic model with a Gaussian mixture model output distribution for each state), even when a DNN-based discriminative acoustic model is eventually trained and used. In this case, a distance metric between two output distributions have been used [3], [4], [7]. The Bhattacharyya distance-based acoustic dissimilarity between two GMM output distributions is defined as

$$BD(s,t) \equiv \min_{u,v} BD(s^{\{u\}}, t^{\{v\}}) \quad (1)$$

where $BD(s^{\{u\}}, t^{\{v\}})$ represents the Bhattacharyya distance between the u-th GMM component of the triphone state s and the v-th GMM component of the triphone state t.

B. Posterior-derived Distance (PD) as Local Distance Metric

Our baseline STD system uses the output of the DNN-HMM-based ASR system and the mismatch with the acoustic model (GMM) used for calculating local distance metric was a problem in previous work [7], [13]. Therefore, we introduces another local distance metric that is derived directly from the DNN-HMM acoustic model.

Since DNN-based discriminative acoustic models do not have explicit output distribution parameters, we employ a DNN-based distance estimation method which approximates the distance between output units (triphone states) by sampling DNN output values using the training dataset of the acoustic model [14]. The DNN-based acoustic distance (PD) is derived

by Bayesian rule to calculate $p(x|y=s)$, $p(x|y=t)$ from the posterior probability $p(y=s|x)$, $p(y=t|x)$ as follows.

$$PD(s,t) = -\ln \sum_l \sqrt{p(y_l=s|x_l)p(y_l=t|x_l)} + \frac{1}{2}\ln \sum_l p(y_l=s) + \frac{1}{2}\ln \sum_l p(y_l=t) \quad (2)$$

where x_l is the l-th sample of feature vector and y_l is the triphone state label (obtained as alignment information used to train DNN-HMM acoustic model).

III. Proposed Spoken Term Detection Method

A. Overview of Proposed System

The overview of the proposed system is common to baseline system as shown in Fig. 1, except that an E2E ASR system is introduced as a replacement or addition to the ASR part and the input to the score fusion part is replaced or added accordingly. Fig. 2 illustrates the flow around the ASR part of the proposed system, with an example of adding the E2E ASR system to the ASR part. The E2E ASR system based on the hybrid CTC/Transformer model is used as a replacement or addition to the DNN-HMM-based ASR used in our baseline system to recognize spoken documents and queries. As described in the previous section, the ASR output is converted into a representation of triphone state sequence, but this process can be pre-processed for the entire spoken document. Also, a state-to-state distance table as a local acoustic dissimilarity information is calculated by using only the parameters of the acoustic model. Therefore, these processes are illustrated as "offline process" and the processes associated with the input spoken query are illustrated as "online process" in Fig. 1.

Both E2E ASR and DNN-HMM-based ASR systems are used to transcribe the speech segments $\{\mathbf{x}^D\}$ from a spoken document to a word/character sequence. In this case, a typical hierarchical DNN-HMM-based ASR can refer to its internal word/phoneme correspondence knowledge to generate the corresponding triphone sequence for ASR output, whereas an E2E ASR with word/character units of output is unable to generate it. Therefore, in order to apply the output of the E2E ASR to the aforementioned baseline system, it is necessary to convert the output of the E2E ASR into a triphone sequence. We use morphological analyzer and grapheme-to-phoneme conversion tools to perform the conversion (referred to as "Character to triphone sequence converter" in Fig. 2). These processes yield the triphone sequence $\mathbf{v}^D_{E2E}$ from the E2E ASR output and the triphone sequence $\mathbf{v}^D_{DNN-HMM}$ from the DNN-HMM-based ASR output for a given speech segment $\mathbf{x}^D$ of spoken document. Also, a spoken query $\mathbf{x}^Q$ is recognized by DNN-HMM-based ASR, and the corresponding triphone sequence $\mathbf{v}^Q$ is obtained. Both triphone sequences $\mathbf{v}^D$ (either of $\mathbf{v}^D_{E2E}$ and $\mathbf{v}^D_{DNN-HMM}$) and $\mathbf{v}^Q$ are mapped to triphone state sequences and the spotting is performed using the endpoint-free DTW algorithm with a state-to-state distance table as shown in Fig. 1.

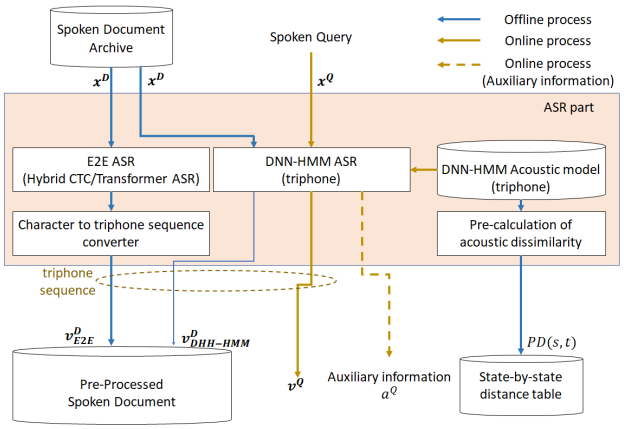

Fig. 2. ASR-related parts of proposed STD system (DNN-HMM-based ASR + E2E ASR)

We investigate the STD method using a score fusion method used in previous studies [7], [13] for combining the spotting scores obtained for each output of two ASR systems. Since the ASR error for each spoken query affects the matching score, certain confidence-related quantities derived from the ASR output are considered to be effective for normalizing the score. The auxiliary information $\mathbf{a}^Q$ derived from the ASR part is introduced as an additional feature of the logistic regression model to normalize the matching score for each query.

B. End-to-end Hybrid CTC/Transformer ASR (E2E ASR)

The End-to-End Neural Network-based Encoder-Decoder model, which has received much attention in recent years, can learn input-output mapping without explicit correspondence, and the performance of a speech recognition system can be determined by the performance of DNNs alone in speech recognition tasks. Also, the combination of Attention and CTC produces results comparable to those of conventional methods described in [8].

Transformer ASR does not use LSTM or bidirectional LSTM as a framework for encoder-decoder, but combines self-attention at the encoder layer and source-target-attention at the decoder layer [10]. Since attention assigns weight to the entire input, it is difficult to assign weight to the repeated sequences of the same phoneme at distant locations. Therefore, a hybrid CTC/Transformer ASR with Connectionist Temporal Classification (CTC) was proposed in comparison to the Transformer ASR (Fig. 3). CTC is a method of learning to map input and output to the same length. The combined use of the models allows us to incorporate the time constraints of CTC into the attention and to make it weighted [9].

We use the hybrid CTC/Transformer model as an additional E2E ASR system to our baseline STD system. As in [9], it is used in combination with a language model trained independently of the hybrid CTC/Transformer model. In the decoding phase, the probability of the decoding hypothesis by the hybrid CTC/Transformer model is linearly interpolated with the probability of the hypothesis by the LSTM-based

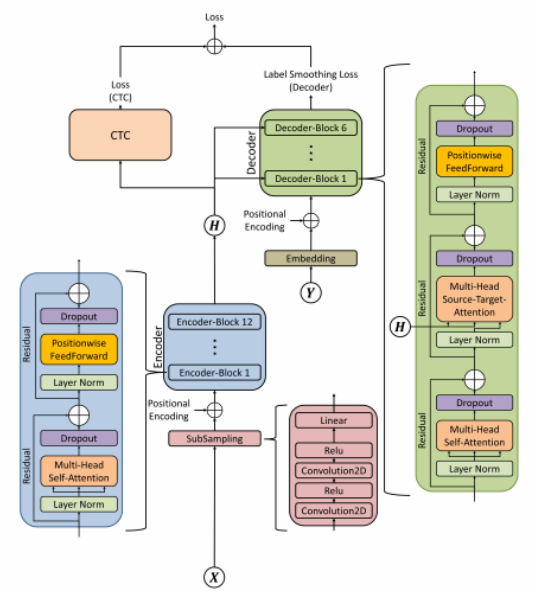

Fig. 3. Hybrid CTC/Transformer end-to-end model

TABLE I
EXAMPLE OF BINARY FEATURES FOR QUERY LENGTH
(IF THE NUMBER OF MORAE IN A QUERY TERM IS L_k AND LESS, THE k-TH FEATURE REPRESENTS 1, OTHERWISE 0)

Query term	Length (#morae)	Binary feature L_k			
		4	6	8	10
Ni ho n ji n (Japanese)	5	0	1	1	1
Shi zu o ka da i ga ku (Shizuoka university)	8	0	0	1	1
A ri ga to u go za i ma su (Thank you)	10	0	0	0	1

language model (LSTMLM), with a interpolation weight γ ($\gamma = 0.3$)

$$P_{decode} = P_{HybridCTC/TransformerASR} + \gamma P_{LSTMLM}. \quad (3)$$

C. Auxiliary Information on Spoken Queries (Query Length)

Another feature we introduced as an auxiliary information is the query length in terms of the number of morae included in the corresponding ASR output. Mora is a linguistic unit in Japanese language and often used as a convenient way to describe the length. We adopt the binary features on the query length as in the previous study [7]. As illustrated in Table I, if the number of morae in the query is less than or equal to a specific number of morae k ($k = 4, 6, 8, 10$), the feature L_k represents 1; otherwise, the feature L_k represents 0. In the following, we will refer to this auxiliary information as 'mora' for short.

D. Logistic Regression Model for Score Fusion (LR)

Since $Score_{state}$, which is the matching score at the spotting phase, is prone to change depending on ASR-derived conditions, such as the misrecognition of the queries at the ASR phase, it is difficult to determine an optimal threshold for score integration. Therefore, we adopt an approach that expects to train a score integration model to provide normalized scores from a large sample of correct and incorrect detections.

To train the logistic regression model, we use a development set of spoken documents and its manual transcript to automatically generate the ASR output and speech queries as follows [7]. The development data is divided into two parts: a target spoken document set for generating examples of matched scores by running STD and the other document set for extracting examples of spoken queries. The model is trained using a set of features and the corresponding class label: matching scores $Score_{state}$ obtained by spotting for two types of DNN-HMM-based ASR and E2E ASR transcriptions, respectively, and auxiliary feature on each spoken query (mora), and the correctness of the detection. We automatically generate spoken queries by using the results of forced-alignment (i.e. word-unit transcription and word boundary times in speech signal) which are used for training DNN-based acoustic model. The spoken queries are all noun phrases and those which only appear once in a single lecture are excluded. The trained logistic regression model is used to estimate the probability of correct detection and generate a new integrated and normalized detection score.

IV. EXPERIMENTS

A. Experimental setup

To compare the query-by-example STD performance for baseline and our proposed methods, STD experiments were conducted on a target document collection used in the NTCIR-12 SpokenQuery&Doc-2 tasks: the lecture of Spoken Document Processing Workshop (SDPWS, 107 lectures, about 29 hours). As with NTCIR-12 SpokenQuery&Doc-2 SQ-STD task evaluation, the inter-pausal units (IPUs) are used as the basic units of the search target, and the search results of IPUs are considered correct if the search term is included in the search results. Also, in the NTCIR-12 SpokenQuery&Doc-2 task, some queries consist of more than one term. To deal with such queries, we use automatic transcription of speech queries to divide the queries into units that do not contain silence intervals of 200 msec or more and treat them as independent queries, and perform searches for each query independently. In the SQ-STD task evaluation, we used 162 query terms spoken by 10 speakers used in the formal-run of the NTCIR-12 SpokenQuery&Doc-2 task. Note that our backend DNN-HMM-based ASR uses a pronunciation lexicon derived from a different morphological analyzer than the NTCIR organizer, hence a simple comparison with the Baseline [1] in Table VI is not possible. In our DNN-HMM-based ASR experimental conditions, 115 query terms contained only IV words and the remaining 47 query terms contained OOV words.

For the training of the DNN-HMM acoustic model and the E2E ASR model (Table II), we used 910 monologue recordings from the corpus of spontaneous Japanese (CSJ); for the training of the DNN-HMM acoustic model and speech recognition, we used the Kaldi toolkit [18]. ESPnet [19] was

TABLE II
SPECIFICATION OF ASR MODELS

DNN-HMM-based	ASR	7 layer(1024 unit)	
	LM	word 3-gram	
Hybrid CTC/Transformer (E2E)	ASR	Encoder	12 layer(2048 unit)
		Decoder	6 layer(2048 unit)
		Attention	4 head(256 dim)
	LM	LSTMLM	2layer(650 unit)

used to train the E2E ASR model and for speech recognition. As for the character-to-triphone sequence converter required for E2E ASR-based system, the morphological analysis tool MeCab [20] was used to assign a pronunciation to each word and convert it to a triphone sequence. E2E ASR and DNN-HMM-based ASR are used for transcribing documents to be searched, and only DNN-HMM-based ASR is used for transcribing speech queries.

For training the score fusion model as logistic regression model, the recordings of CORE set in CSJ (177 lectures, approximately 44 hours; hereafter referred to as CSJ-CORE dataset) were used as development document data for generating positive and negative search examples. The development data set is not used for acoustic model training. A set of artificial spoken query terms was automatically generated from a dataset for acoustic model training. First, the morphological analysis of the transcribed text was performed and a consecutive noun words of up to five words were considered as candidates for query term. Then, the candidate query terms were filtered by removing the words that were too long (more than 12 morae) or too short (less than 2 morae), followed by a tf-idf-based filtering for balancing the distribution of selected query terms. As a result, 923 query terms (325 terms containing IV words only and 598 terms containing OOV words) was selected as query terms for development document data.

F-measure(max) and MAP are used as search accuracy metrics, where F-measure (max) is the maximum value of F-measure when adjusting the threshold, and MAP is the mean of the average precision of all queries. The detection thresholds for the DNN-HMM-based ASR and the E2E ASR proposed in this paper are set to be 1000 candidates per query.

In the figures and tables in this section, the STD system using the DNN-HMM-based ASR and BD local distance is denoted as "state_spot(BD_DNN-HMM)", one using the DNN-HMM-based ASR and PD local distance is denoted as "state_spot(PD_DNN-HMM)", one using E2E ASR and PD local distance is denoted as "state_spot(PD_E2E)", and one using two ASR-based results by a logistic regression model is denoted as "LR(PD_DNN-HMM+PD_E2E)", respectively.

B. ASR Performance of DNN-HMM-based and E2E systems

Table III and Table IV show the ASR performance of document and spoken queries, respectively. As shown in Table III, the recognition performance of the E2E ASR outperformed that of DNN-HMM-based ASR for both datasets. This result indicates that E2E ASR has less impact on OOV words than DNN-HMM-based ASR and it can generalize more to different domain content.

TABLE III
ASR PERFORMANCE OF CSJ-CORE(DEVELOPMENT) AND SDPWS(NTCIR-12 FORMAL-RUN EVALUATION) TARGET DOCUMENTS

task	model	CER(%)
CSJ-CORE(Dev)	DNN-HMM-based ASR	13.0
	Hybrid CTC/Transformer ASR	8.5
SDPWS(Eval)	DNN-HMM-based ASR	19.0
	Hybrid CTC/Transformer ASR	12.2

TABLE IV
ASR PERFORMANCE OF CSJ-CORE(DEVELOPMENT) AND SDPWS(NTCIR-12 FORMAL-RUN EVALUATION) SPOKEN QUERIES

task	model	CER(%)	PER(%)
CSJ-CORE(Dev)	DNN-HMM-based ASR	50.3	17.6
SDPWS(Eval)	DNN-HMM-based ASR	30.1	11.7

Also, Table IV shows that the CER is higher for spoken queries than spoken documents. This may be due to a language model and lexicon which involve many OOV words for spoken queries in DNN-HMM-based ASR system. However, the PER will be more related to the STD performance in our STD methods since the spotting is done at a triphone state level which is smaller than the character level.

C. SQ-STD task result

The evaluation in the development set in Table V shows that the one using PD performed better than the one using BD for acoustic distance. This is because the use of PD accurately reflects the acoustic dissimilarities derived directly from the acoustic models used in the DNN-HMM-based ASR. In the experiment using DNN-HMM-based ASR and E2E ASR as the acoustic model, E2E ASR showed better performance than DNN-HMM-based ASR. The reason for the superior results of E2E ASR could be due to its higher ASR accuracy than DNN-HMM-based ASR for documents (Table III). Best performance was obtained by score integration of E2E ASR and DNN-HMM-based ASR spotting scores using logistic regression model.

Table VI shows the STD performance for evaluation set. We find that the E2E ASR-based system has less impact on STD performance improvement for IV queries compared to the development set results. In Fig. 4, the recall-precision curves for IV queries shows that the E2E ASR-based system has higher precision at low recall region. On the other hand, the DNN-HMM-based ASR-based system has higher precision when the recall is around 30〜60%. We believe that the post-processing of E2E ASR which convert ASR output into triphone state may have caused a mismatch between the phone-level representations derived from query and document, and this is the reason why the F-measure was worse for IV queries. We also compared the ASR accuracy for spoken query between two ASR systems. From Table VII, we can see that E2E ASR has a worse PER and WER even for IV queries than DNN-HMM-based ASR. In summary, the reason for the worse F-value of E2E ASR on IV queries can be attributed to

TABLE V
CSJ-CORE(DEVELOPMENT SET) STD PERFORMANCE

query	system	F(MAX)	MAP
IV	state_spot(BD_DNN-HMM)	65.7	81.3
	state_spot(PD_DNN-HMM)	70.3	82.1
	LR(PD_DNN-HMM+mora)	67.7	82.2
	state_spot(PD_E2E)	72.8	82.5
	LR(PD_E2E+mora)	70.6	82.5
	LR(PD_DNN-HMM+PD_E2E)	**75.2**	84.1
	LR(PD_DNN-HMM+PD_E2E+mora)	72.8	**84.2**
OOV	state_spot(BD_DNN-HMM)	19.5	47.0
	state_spot(PD_DNN-HMM)	21.6	48.5
	LR(PD_DNN-HMM+mora)	24.0	48.7
	state_spot(PD_E2E)	30.4	53.8
	LR(PD_E2E+mora)	**33.2**	53.8
	LR(PD_DNN-HMM+PD_E2E)	29.7	**54.7**
	LR(PD_DNN-HMM+PD_E2E+mora)	32.1	**54.7**
Total	state_spot(BD_DNN-HMM)	39.8	60.3
	state_spot(PD_DNN-HMM)	44.2	61.5
	LR(PD_DNN-HMM+mora)	45.3	61.6
	state_spot(PD_E2E)	52.3	64.9
	LR(PD_E2E+mora)	52.9	64.9
	LR(PD_DNN-HMM+PD_E2E)	54.2	**66.1**
	LR(PD_DNN-HMM+PD_E2E+mora)	**54.3**	**66.1**

TABLE VI
FORMAL-RUN(EVALUATION SET) STD PERFORMANCE

query	system	F(MAX)	MAP
IV	state_spot(BD_DNN-HMM)	56.7	63.9
	state_spot(PD_DNN-HMM)	**57.6**	65.4
	LR(PD_DNN-HMM+mora)	57.1	65.4
	state_spot(PD_E2E)	51.9	64.7
	LR(PD_E2E+mora)	55.3	65.0
	LR(PD_DNN-HMM+PD_E2E)	53.9	68.1
	LR(PD_DNN-HMM+PD_E2E+mora)	56.0	**68.3**
OOV	state_spot(BD_DNN-HMM)	41.4	54.2
	state_spot(PD_DNN-HMM)	38.2	55.9
	LR(PD_DNN-HMM+mora)	45.0	55.9
	state_spot(PD_E2E)	45.7	64.8
	LR(PD_E2E+mora)	**51.1**	65.9
	LR(PD_DNN-HMM+PD_E2E)	45.0	66.9
	LR(PD_DNN-HMM+PD_E2E+mora)	49.7	**67.1**
Total	NTCIR-12 Baseline [1]	47.1	54.1
	phoneme_spot(ED_DNN-HMM) [13]	43.5	58.2
	syll_spot(BD_DNN-HMM) [7]	38.3	64.4
	state_spot(BD_DNN-HMM)	53.2	61.1
	state_spot(PD_DNN-HMM)	53.1	62.6
	LR(PD_DNN-HMM+mora)	**54.4**	62.6
	state_spot(PD_E2E)	49.5	64.7
	LR(PD_E2E+mora)	53.5	65.2
	LR(PD_DNN-HMM+PD_E2E)	50.2	67.8
	LR(PD_DNN-HMM+PD_E2E+mora)	53.8	**67.9**

the worse ASR accuracy for the queries appeared in spoken document.

Next, Fig. 5 shows the recall-precision curve for OOV queries. We find the E2E ASR-based STD system's precision is reduced at low recall region. This can be attributed to the high number of false positives for some queries. One of the reasons for the false positives is the mismatch between phone-level representations as described above. Therefore, the decrease in the F-measure was caused by the false positive results. The logistic regression model was used to integrate the scores, but the false positives were not improved, and the results showed that the F-measures were worse than those of

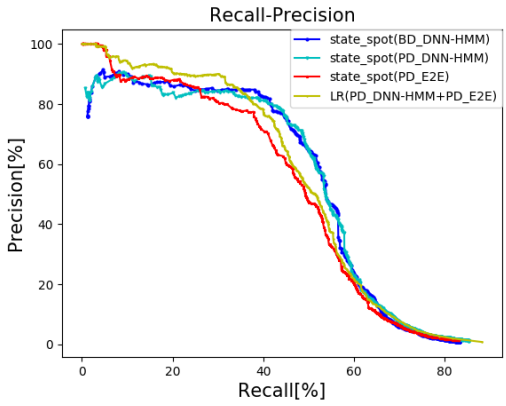

Fig. 4. Recall-Precision curves of different STD systems(IV queries, evaluation set)

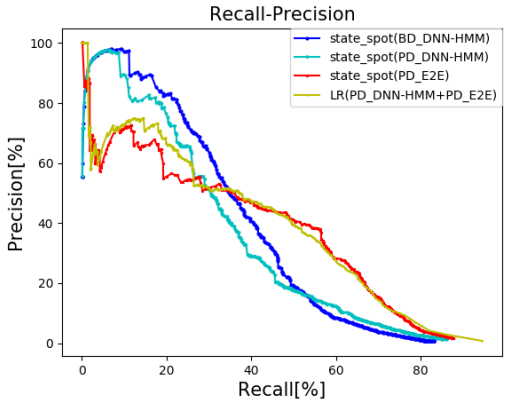

Fig. 5. Recall-Precision curves of different STD systems(OOV queries, evaluation set)

the conventional method and the E2E ASR alone. However, MAP showed improvement in E2E ASR-based system compared to the case where DNN-HMM-based ASR was used, and if we take into account the fact that there were more false positives for some of the queries, the E2E ASR-based system as a whole showed better performance. As a result, the score integration of two ASR-based spotting scores using a logistic regression model trained by development set led to a significant improvement of the STD system in MAP.

Finally, we discuss the effect of adding the feature "mora" as an auxiliary information mentioned in section III-C. In evaluation set, adding the "mora" feature improved the F-measure for both DNN-HMM-based ASR and E2E ASR-based systems. Particularly in OOV queries, we find improvements in both of the two ASR-based systems. The recall-precision curve of the OOV queries shows a significant improvement for low recall region (Fig. 6, 7). This indicates that the auxiliary information allows to reduce the impact of false positives on some queries. Based on these results, we combined the "mora" feature with the DNN-HMM-based ASR and E2E ASR scores.

TABLE VII
ASR PERFORMANCE OF NTCIR-12 FORMAL-RUN SPOKEN QUERY

query	model	PER(%)	WER(%)
IV	DNN-HMM-based ASR	9.5	35.1
	Hybrid CTC/Transformer ASR	11.0	42.2
OOV	DNN-HMM-based ASR	16.4	69.8
	Hybrid CTC/Transformer ASR	8.4	34.7
Total	DNN-HMM-based ASR	11.7	45.2
	Hybrid CTC/Transformer ASR	10.1	40.0

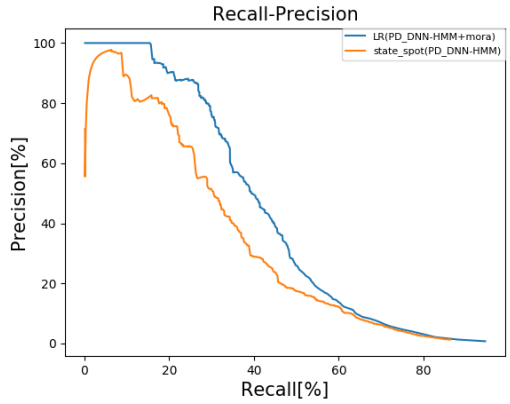

Fig. 6. Effect of additional auxiliary information "mora" on STD performance (OOV queries, evaluation set, DNN-HMM-based ASR)

In the evaluation set, the results showed the F-measure improvement compared to the results before integrating "mora" feature (Fig. 8), and attained the best MAP performance.

V. CONCLUSIONS

In this paper, we proposed STD methods for integrating DNN-HMM-based ASR and end-to-end neural network based hybrid CTC/Transformer ASR. Experimental results showed that the E2E ASR outperformed the traditional method of STD using DNN-HMM-based ASR. In addition, we integrated the

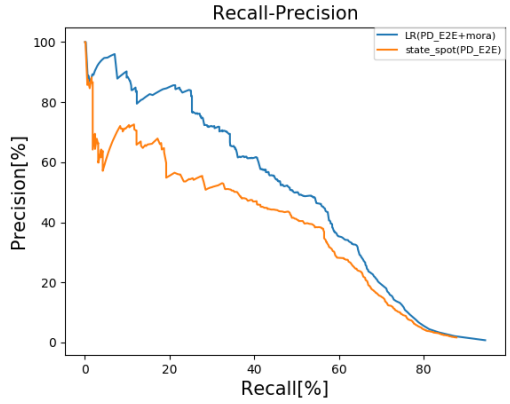

Fig. 7. Effect of additional auxiliary information "mora" on STD performance (OOV queries, evaluation set, E2E ASR)

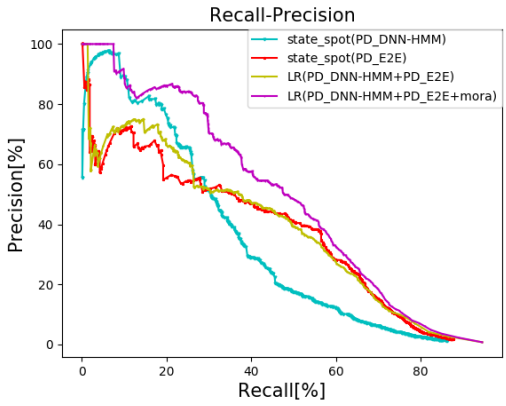

Fig. 8. Effect of integrating two ASR and additional auxiliary information on STD performance (OOV queries, evaluation set)

matching scores of DNN-HMM-based ASR and E2E ASR using a logistic regression model to improve the overall STD performance. The E2E ASR also increased the number of false positives for some queries. Since this was due to the phone-level mismatch between the query and the search document, adding an auxiliary information (mora) was proposed to reduce the false positives. As a result, the score integration of two ASR systems and the auxiliary information using a logistic regression model led to the best STD performance in MAP.

Our current STD system with E2E-based ASR employed a morphological analysis to convert the output of E2E ASR into phone sequence. This process may have caused a mismatch on phone-level representation between the spoken query and document when acoustic dissimilarity is calculated, resulting in an increase in false positives. Therefore, by directly converting the output of E2E ASR to the phone sequence, this mismatch is eliminated, and STD performance is expected to be improved. Further improvements can be expected by adding additional features, such as the search result of query expansion [7].

REFERENCES

[1] T. Akiba, H. Nishizaki, H. Nanjo and G.J.Jones. "Overview of the NTCIR12 SpokenQuery&Doc-2 task." Proc. of the NTCIR-12 Conference, Tokyo, Japan, 2016.

[2] S. Nakagawa, K. Iwami, Y. Fujii, and K. Ymamoto. "A robust/fast spoken term detection method based on a syllable n-gram index with a distance metric," Speech Communication, Vol.55, pp.470–485, 2013.

[3] N. Yamamoto and A. Kai, "Using acoustic dissimilarity measures based on state-level distance vector representation for improved spoken term detection," 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–4, 2013.

[4] Mitsuaki Makino, Naoki Yamamoto, Atsuhiko Kai: "Utilizing State-level Distance Vector Representation for Improved Spoken Term Detection by Text and Spoken Queries," Proc. INTERSPEECH, pp.1732–1736, 2014.

[5] H.Wang, et al.: "Acoustic Segment Modeling with Spectral Clustering Methods," IEEE/ACM Transaction on Audio, Speech, and Language Processing, Vol.23, No.2, pp.264–277, 2015.

[6] J. Tejedor, et al.: "Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations," EURASIP Journal on Audio, Speech, and Music Processing, 2016:1, 2016.

[7] S. Oishi, T. Matsuba, M. Makino and A. Kai. "Combining state-level spotting and posterior-based acoustic match for improved query-by-example spoken term detection," Proc. INTERSPEECH, pp.740–744, 2016.
[8] S. Watanabe, T. Hori. S. Kim, J. Hershey and T. Hayashi. "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," In IEEE Journal of Selected Topics in Signal Processing, Vol.11, No.8, pp.1240–1253, 2017.
[9] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix and A. Ogawa et al. "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," In Proceedings of INTERSPEECH, pp. 1408–1412, 2019.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit and L. Jones et al. "Attention Is All You Need," In Advances in Neural Information Processing Systems pp. 5999–6009, 2017.
[11] L. Dong, S. Xu, and B. Xu. "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," In Proceedings of ICASSP, pp. 5884–5888, 2018.
[12] National Institute of Informatics, http://research.nii.ac.jp/ntcir/ntcir-12/
[13] H. Kondo, A.Kai and S.Oishi. "Effect of score fusion model learning on spoken term detection from spoken query," Reports of the autumn meeting the Acoustical Society of Japan, pp. 989–992, 2018. (in Japanese)
[14] R. Konno, K. Kojima, S. Lee, K. Tanaka and Y. Itoh. "A Construction Method of an Acoustic Distance Using Output Probability of Deep Neural Network for Spoken Term Detection," Trans. IEICE, Vol.J100-D, No.8, pp.798–807, 2017.
[15] G. Mantena, and K. Prahallad. "Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios," Proc. of ICASSP, 2014.
[16] J. Tejedor, I. Szoke, and M. Fapso. "Novel methods for query selection and query combination in query-by-example spoken term detection," Proc. of SSCS, 2010.
[17] H. Wang, T. Lee, C. Leung, B. Ma, and H. Li. "Acoustic Segment Modeling with Spectral Clustering Methods," IEEE/ACM Transaction on Audio, Speech, and Language Processing, Vol.23, 2015.
[18] D. Povey, A. Choshal, G. Boulianne, L. Burget and O. Glembeket et al. "The Kaldi Speech Recognition Toolkit," Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.
[19] S. Watanabe, T. Hori, S. Karita, T. Hayashi and J. Nishitoba et al. "ESPnet: End-to-End Speech Processing Toolkit," In Proceedings of INTERSPEECH, pp. 2207–2211, 2018.
[20] "MeCab, " https://www.mlab.im.dendai.ac.jp/yamada/ir/ Morphological-Analyzer/MeCab.html .