Retrieval-oriented E2E ASR Modeling for Improved Query-by-example Spoken Term Detection

Takumi Kurokawa* and Atsuhiko Kai*
* 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

***Abstract*—Query-by-example Spoken Term Detection (STD) systems can make effective use of automatic speech recognizer (ASR), especially in situations with high recognition accuracy. However, the conventional DNN-HMM ASR has a problem of low recognition accuracy for out-of-vocabulary (OOV) words. This problem has a serious impact on the performance of STD-based speech retrieval and can lead to false retrieval. In a recent study, End-to-end (E2E) ASR has achieved comparable performance compared to DNN-HMM ASR. Since E2E ASR can learn direct speech-to-character mappings without knowledge of word or phonetic pronunciation, it can reduce the impact of the OOV problem. In this paper, we propose STD method using an ASR system adapted to spoken word retrieval and acoustic similarity at the sub-phone level. Our proposed method employs E2E ASR modeling with Kana as the output unit, which can be easily converted to phoneme sequence. To improve the recognition and retrieval accuracy of words that are likely to be searched for, we propose E2E modeling that adds optional tag information to the output unit. Experiments using the NTCIR-12 SpokenQuery&Doc-2 task show that the E2E ASR-based STD method significantly improves the retrieval performance over the STD method using DNN-HMM ASR. This is due to the fact that the proposed E2E ASR was able to reduce the OOV problem for spoken documents and spoken queries.**

I. Introduction

Spoken term detection (STD) is the task of retrieving a given spoken query term from a large number of spoken documents. In situations where automatic speech recognition (ASR) accuracy is high, such as for high-resource languages, it is effective to use ASR as a component of the STD system because its output provides rich clues for search. However, conventional DNN-HMM ASR is designed to decode the spoken word sequence based on hierarchical knowledge such as triphone-unit acoustic model (DNN-HMM) and phonetic pronunciation of vocabulary words. As a result, if out-of-vocabulary (OOV) words that are not included in the prior hierarchical knowledge are present in a query or document, it will greatly affect the recognition and retrieval accuracy. For such OOV problem, feature-based matching without ASR is a common method that has been shown to be effective [5]. However, the feature-based STD method is time-consuming and could not outperform the performance of the conventional ASR-based STD method, especially for query terms which do not contain OOV words. To solve these problems, a method employing post-ASR matching method to deal with subword-level acoustic similarity between documents and queries has been proposed [3]–[5], [9], [10].

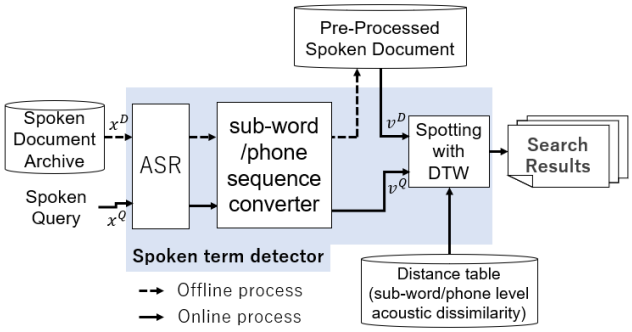


Fig. 1: Overview of the baseline STD system

In recent studies on ASRs, much attention has been focused on neural network-based end-to-end (E2E) ASR modeling. E2E ASR is able to learn input-output mappings without any pre-defined knowledge of linguistic/phonetic units and achieves performance comparable to conventional DNN-HMM ASR [6]. Recently, a model structure called Transformer has been proposed [15]. This method, proposed for a machine translation task, uses a self-attention approach to both reduce training time and improve translation performance. A hybrid approach with Connectionist Temporal Classification (CTC) and Transformer has attained a comparable or superior performance to DNN-HMM ASR System [1].

In this paper, we propose E2E ASR-based STD system that employs a subphone-level local distance metric derived from an acoustic model. The E2E ASR is used to produce subphone-level representation of spoken documents and spoken queries. The STD system performs spotting on a converted subphone sequence with DNN-derived acoustic dissimilarity. In this paper, we focus on the Japanese language and perform an STD experiment. The experiments are conducted on the NTCIR-12 SpokenQuery&Doc-2 task [2] and the results are compared with the baseline results of the conventional STD systems and the result obtained with the NTCIR-12 organizer's ASR.

II. Overview of the baseline STD system

In this study, we consider the baseline STD approach which employs the output of ASR (Fig. 1). ASR part recognizes each segment of spoken document x^D and spoken query x^Q, and followed by a sub-word/phone sequence converter which obtains their respective sequences v^D and v^Q. We

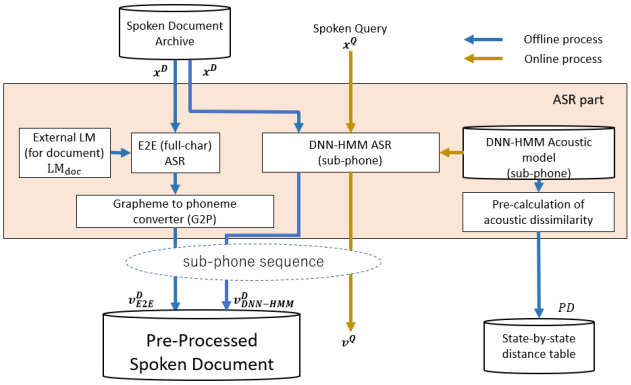


Fig. 2: ASR-related parts of STD system (E2E (full-char) + G2P)

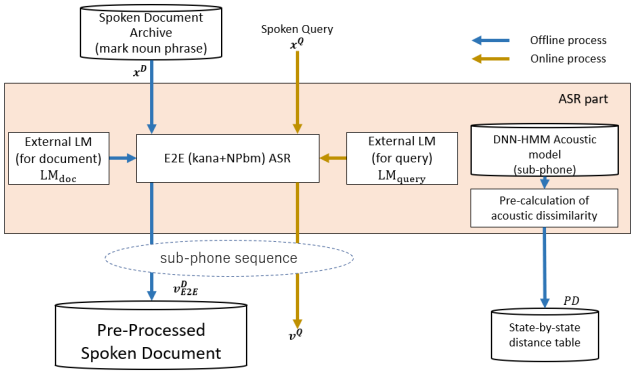


Fig. 3: ASR-related parts of STD system (E2E (kana+NPbm) ASR)

use a dynamic time warping algorithm (DTW) for spotting on sub-word/phone sequences v^D and v^Q. As for the local distance at subphone-level, our system uses Posterior-derived distance (PD, [8], [10]), which estimates subphone-level acoustic dissimilarity with triphone-unit DNN-HMM acoustic model [9], [10]. Spotting at the subword level has been shown to reduce the impact of ASR errors [3]. However, the search performance is reduced due to the OOV problem. To reduce the impact of such OOV problem, we have proposed STD method using subphone-level matching [4], [5] and also proposed an improved method using E2E ASR [10]. In the following, we introduce these two STD methods: method using DNN-HMM ASR only, and method using two ASRs, E2E and DNN-HMM ASRs, together.

A. STD method using DNN-HMM ASR with subphone sequence conversion

The conventional DNN-HMM ASR system can output a subword sequence corresponding to the output word sequence. In our conventional method, speech recognition was performed using DNN-HMM ASR on a spoken document x^D and spoken query x^Q, and spotting was performed using the sub-phone sequences corresponding to the output word sequences [4], [5], [9], [10]. In DNN-HMM ASR systems, OOV problem increases the number of false recognitions and degrades STD performance. A score fusion method was used to improve the retrieval performance by integrating the spotting score and auxiliary information using a logistic regression model [5], [9], [10]. This has been shown to reduce the effect of OOV problem. However, DNN-HMM ASR could not completely solve the OOV problem because it assumes prior information on the lexical and phonetic knowledge levels in addition to the character-level transcription at the training stage.

B. STD method using character-output E2E ASR with G2P conversion

E2E ASR system with output in character units can reduce the OOV problem because it learns input-output mappings using only pairs of speech waveforms and strings as training dataset, without using predefined lexical knowledge. For spoken document recognition, this baseline system used an E2E ASR that outputs Chinese/Kana characters for x^D. Hereafter, we call this ASR as **E2E (full-char) ASR**. For spoken queries x^Q, DNN-HMM ASR was used. E2E ASR output is converted to sub-phone level by using a grapheme to phoneme converter (G2P) (Fig. 2), and spotting is performed at the subphone level as in the DNN-HMM ASR-based method described above. In [10], recognition of spoken documents using E2E (full-char) ASR reduced the OOV problem and improved the retrieval performance. However, the G2P converter inherently produce conversion errors and affect the STD performance. Also, the use of DNN-HMM ASR for spoken queries did not reduce the OOV problem for spoken queries.

III. PROPOSED STD SYSTEM WITH IMPROVED E2E MODELING

We propose E2E ASR modeling for predicting characters in Japanese kana representation that is robust to the OOV problem. This E2E ASR is used for the recognition of both spoken documents and spoken queries (Fig. 3). Japanese is often represented by a string of mixed Kanji, Katakana, and Hiragana characters, and it is difficult to give correct phonetic representations from only string (grapheme) information. To solve this problem, we propose an E2E ASR that uses kana characters as the output unit, which can be easily converted to sub-phone level representation. When we do spotting, kana characters as the output unit is converted to sub-phone level representation. Kana can be used to represent any phoneme sequence in Japanese, so words that did not appear during learning can be represented. It can reduce the G2P conversion error derived from the converter, which could affect the STD performance with E2E (full-char) + G2P as described in section II-B. This is because the speech signal can be directly converted to a sub-phone level representation. Hereafter, this ASR is denoted as **E2E (kana) ASR**.

Since Kana characters can be easily associated with phonemes, E2E (kana) will reduce the impact of the OOV problem as a speech-to-pronunciation conversion system.

However, Kana characters increase linguistic ambiguity, especially in a language like Japanese that does not separate words, since it is expected that it will be difficult to acquire potential lexical and linguistic-level knowledge with only Kana characters. In contrast, E2E (full-char) ASR is expected to acquire more lexical and linguistic-level knowledge than E2E (kana) ASR, because more clues of such knowledge levels remain in a mixed string of Chinese and Kana characters. The lack of linguistic information, especially for words that are potentially query terms, is expected to lead to a decrease in recognition accuracy in such word segments. If E2E ASR can perform speech recognition with awareness of noun phrases, it can be expected to improve speech recognition accuracy for spoken queries and potential searched segments in spoken documents.

We propose to train E2E ASR with modified output label by marking the start (**S_** mark) and end (**_E** mark) of noun phrases (kana+Noun-Phrase-boundary-mark, kana+NPbm, Fig. 4). The ASR that can produce this output is called **E2E (kana+NPbm) ASR**. Fig. 4 shows an example of different definitions of output units. The word "機 能" is a noun in a phrase "機 能 と し て は". This phrase means "As for the function" in English. In case E2E (full-char) ASR + G2P method is used, if the "機 能" is an OOV word for G2P converter, it cannot be converted to a sub-phone level representation. E2E (kana) ASR converts the speech signal directly into kana, so it will not produce serious subphone level errors. It should be noted that the characters "キ ノ ウ" theirselves do not contain any linguistic information and common to no-existent noun phrase. In contrast, the E2E (kana+NPbm) ASR is expected to recognize noun phrases more accurately because it learns them with linguistic information related to the boundary of noun phrases.

External language model (LM) is used in combination with the E2E ASR. Our E2E ASR employs a score fusion with the following equation at decoding process [1].

$$P_{decode} = P_{E2E} + \alpha P_{LM} \tag{1}$$

where P_{LM} is a language model score calculated by LSTM LM, and α is a language model weight $(0 \leq \alpha \leq 1)$, respectively. In general, language models are trained to estimate P_{LM} for a single sentence. Since a spoken query is mostly a noun phrase, using a language model adapted to noun phrases is expected to improve speech recognition accuracy for spoken queries. We also propose to combine E2E ASR with a language model trained only on noun phrases for spoken queries (**$\mathbf{LM_{query}}$**). Such language model can be trained by simply using only the tagged segments as a subset of training data for E2E (kana+NPbm) ASR. External LM (**$\mathbf{LM_{doc}}$**) trained with full transcript of training data, as well as those used during E2E training, is used to recognize spoken documents, while $\mathrm{LM_{query}}$ is used to recognize spoken queries (Fig. 3). We discuss the effect of query-specific language model on the recognition of spoken queries in STD experiments.

E2E (full-char)	機 能	と	し	て	は
Part of Speech	noun	postpositional particle	verb	postpositional particle	postpositional particle
E2E (kana)	キ ノ ウ	ト	シ	テ	ワ
↓					
E2E (kana+NPbm)	S_キ ノ ウ_E	ト	シ	テ	ワ

Fig. 4: Example of different definitions of output units for E2E ASR modeling

IV. Experiment

A. setup

This study is based on the NTCIR-12 SpokenQuery&Doc2 task [2]. The spoken document used in this task consists of audio recordings from Spoken Document Processing Workshop (SDPWS, 107 lectures, about 29 hours). The spoken document is segmented in Inter Pausal Units (IPU) as defined in CSJ dataset [13], [14], and during evaluation, a search result is considered correct if a given query term is contained in the corresponding IPU. For the spoken queries, we used 162 query terms spoken by 10 speakers that are used in the formalrun of the NTCIR-12 SpokenQuery&Doc2 task. The distinction between in-vocabulary (IV) and OOV query terms in this study are confirmed by checking if all the words in the query term are included in the lexicon of our DNN-HMM ASR. The query set contained 115 IV terms and 47 OOV terms. Also, some queries contain more than one noun. It should be noted that these numbers are not correct under the condition of using the organizer's ASR because the vocabulary is different from that of our ASRs. In this study, we use this task as a test set.

ASR systems are described below. DNN-HMM ASR and E2E ASR (Hybrid CTC/Transformer ASR) are trained with 910 academic lectures from CSJ dataset. In the hybrid approach of E2E ASR, the following equation is used for linear interpolation.

$$Loss_{E2E} = \lambda Loss_{Transformer} + (1-\lambda) Loss_{CTC} \tag{2}$$

where $Loss$ is a loss score for transformer and CTC respectively, and λ is a weight $(0 \leq \lambda \leq 1)$ for multi-task learning. Language Model (LM) is trained with 2362 lectures from CSJ dataset. DNN-HMM ASR is trained with Kaldi toolkit [11], and E2E ASR is trained with ESPnet [12] using PyTorch as the backend. To train E2E (kana+NPbm) ASR, we prepare a kana-transcript with optional tags marking the start and end of the noun phrases that appear in the 910 lectures of CSJ dataset. The specification of model structure is shown in Table I.

When performing language model adaptation, it is necessary to determine the language model weight α. To determine the language model weight, we use the core set (177 lectures, about 44 hours) from CSJ dataset as the development set. We perform morphological analysis on the transcribed text of the development set. The results are used for collecting sample noun phrases, which are a sequence of up to five consecutive words those part-of-speech are assigned as noun. We automatically selected the phrases that has between 3 and

TABLE I: Specification of ASR Models

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

TABLE II: ASR performance in spoken query and query part in spoken document (test set)

Query type	ASR system (speech-to-phone)	Document PER [%]	Query PER [%]
ALL	DNN-HMM (organizer)	13.86	12.19
	DNN-HMM (ours)	14.82	9.51
	E2E (full-char) + G2P	10.88	8.02
	E2E (kana)	14.88	4.33
	E2E (kana+NPbm)	8.37	4.24
IV	DNN-HMM (organizer)	15.03	13.22
	DNN-HMM (ours)	12.43	7.07
	E2E (full-char) + G2P	11.69	8.66
	E2E (kana)	15.26	4.26
	E2E (kana+NPbm)	9.20	4.22
OOV	DNN-HMM (organizer)	10.96	9.98
	DNN-HMM (ours)	20.77	14.72
	E2E (full-char) + G2P	8.87	6.63
	E2E (kana)	13.92	4.49
	E2E (kana+NPbm)	6.30	4.28

11 morae with high tf-idf values (923 queries). Mora is a linguistic unit of Japanese.

MAP is used as an evaluation measure of search performance. MAP is the mean of the average precision of all queries. The detection threshold is set to be 1000 candidate intervals per query, and the MAP is measured for them.

The ASR models to be compared in this experiment and the output units of those models are shown below.

- **DNN-HMM (organizer)** : Word
- **DNN-HMM (ours)** : Word
- **E2E (full-char)** : Chinese/Kana characters
- **E2E (kana)** : Kana
- **E2E (kana+NPbm)** : Kana and Noun-Phrase-boundary-mark

We also compare the effect of different units of phoneme/subphone and distance measures during spotting on retrieval performance. Two distance measure are compared: PD for triphone state unit [8], [10], and edit distance (ED) for phoneme unit. The latter is almost the same as one of the methods used in the NTCIR-12 organizer's baseline systems [2].

B. Comparison of ASR performance

Table II shows the phoneme error rate (PER) for spoken queries and the PER for the query part of the spoken document. PER for spoken queries shows a significant improvement when using E2E ASR compared to when using DNN-HMM ASR. Our proposed methods, E2E (kana) and E2E (kana+NPbm), show improvement over our conventional method, E2E (full-char) + G2P for all spoken queries. E2E (full-char) + G2P produces OOV related error as in the DNN-HMM ASR. E2E (kana) and E2E (kana+NPbm) directly output kana for speech signals and the effect of OOV could be reduced.

TABLE III: Details of PER improvement by E2E (kana + NPbm) compared to E2E (kana) ASR in the query part of spoken documents (test set)

PER changes	ALL (221)	IV (156)	OOV (65)
improved	150	100	50
no change (PER $= 0.0$)	37	32	5
no change (PER > 0.0)	16	13	3
worsened	19	11	7

For PER of all query parts of a spoken document, E2E(kana) does not improve significantly over DNN-HMM. This trend is the same for IV and OOV queries. This may be because E2E (kana) could learn less lexical features related to noun phrases than E2E (full-char) + G2P. In contrast, E2E (kana+NPbm) showed a significant improvement. E2E (kana+NPbm) also shows the best improvement for spoken queries.

Table III shows the details of PER change between E2E (kana) and E2E (kana+NPbm) ASRs in the query part of spoken documents. In Table III, we denote "improved" if there is an improvement, "worsened" if there is a worse result, "(PER $= 0.0$)" if both recognition results have PER $= 0$, and "no change(PER > 0.0)" if PER > 0 and no change. Note that the distinction between "IV" and "OOV" in Table III are based on the data used to train the E2E ASR (910 lectures of CSJ). Also, if a single query contains multiple nouns, they are counted as separate queries. As a result, the number of queries has increased from 162 queries to 221 queries. The results show that there is an improvement in many queries. The improvement is also seen regardless of IV and OOV queries. This shows that noun phrase features could be better learned by introducing additional NPbm marks.

Next, we discuss the relationship between the accuracy of predicted NPbm marks and PER. Table IV shows ASR performance improvements for the query part of the spoken document by E2E (kana+NPbm) compared to E2E (kana). When correctness of predicted NPbm marks is judged, we indicate "None" when both of "S_" and "_E" marks are missing. As shown in Table IV, it can be seen that the percentage of correctly predicted NPbm marks including start accounts for 80%, resulting in a large improvement in PER from E2E(kana). Even if the NPbm mark cannot be estimated accurately, an improvement in PER of more than 10% has been obtained. This result suggests that the introduction of NPbm marks may have helped in the acquisition of lexical features of the entire noun phrase, not just its beginning and end.

Table V shows examples of output results for E2E (kana) and E2E (kana+NPbm). The query is "チャセン" (name of the morphological analysis tool). In this example, E2E (kana+NPbm) could recognize query part where E2E (kana) has recognition error.

TABLE IV: Relationship between the correctness of predicted NPbm marks and the improvement of PER by E2E (kana+NPbm) for the query part of the spoken document

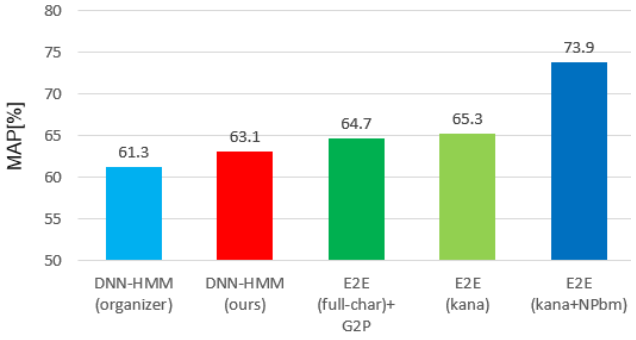
Correctly predicted NPbm marks for the query part (ratio [%])	PER [%]	Reduction rate from E2E (kana) [%]
Start and end (45.99)	**4.84**	60.16
Only start (35.43)	**13.74**	32.61
Only end (6.44)	**13.65**	14.69
None (12.13)	**27.66**	14.94
Total (100.00)	**8.37**	43.75

TABLE V: Examples of ASR output for E2E (kana) and E2E (kana+NPbm) (test set)

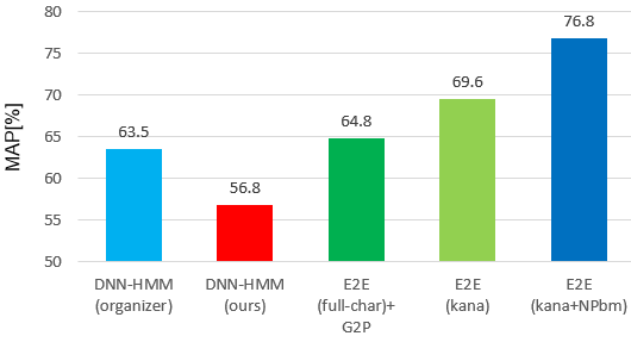
Query: **S_チ ャ セ ン_E** (“Chasen” in English)										
Reference	S_テ	キ	ス	ト_E	S_ジ	ョ	ー	ホ	ー_E	オ
	S_チ	**ャ**	**セ**	**ン_E**	オ	モ	チ	ー	テ	
	(“using chasen for text information” in English)									
E2E (kana)	テ	キ	ス	ト	ジ	ョ	ー	ホ	ー	オ
	タ	**ッ**	**セ**	**ン**	オ	モ	チ	ー	テ	
E2E (kana+NPbm)	S_テ	キ	ス	ト_E	S_ジ	ョ	ー	ホ	ー_E	オ
	S_チ	**ャ**	**セ**	**ン_E**	オ	モ	チ	ー	テ	

TABLE VI: Effect of query-specific language model for E2E (kana+NPbm) ASR (development set)

LM type	External LM weight	Query PER [%]
without LM		10.75
LM_{doc}	0.3	10.47
LM_{query}	0.1	10.27
LM_{query}	0.3	10.79
LM_{query}	0.5	14.43



(a) all queries in test set



(b) OOV queries in test set

Fig. 5: Comparison of STD performance for different ASRs and their output units (word unit is used for DNN-HMM ASRs)

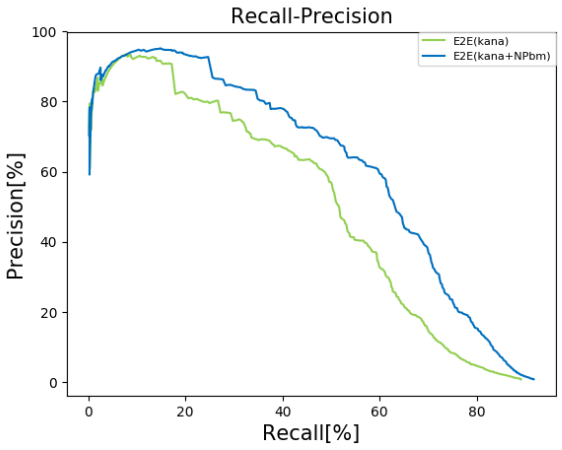


Fig. 6: Recall-Precision curves of different STD systems (all queries in test set, E2E (kana) and E2E (kana+NPbm))

C. STD task result

Fig. 5(a) shows STD performance for all queries in the test set. The detection performance of E2E (kana) and E2E (kana+NPbm) ASR-based systems was improved over DNN-HMM and E2E (full-char) + G2P ASR-based systems. Furthermore, the difference between MAP of E2E (kana) and E2E (kana+NPbm) in Fig. 5(a) is significant. Similarly, as for the STD performance in OOV queries (Fig. 5(b)), the improvement in E2E (kana) and E2E (kana+NPbm) is significant. The results are consistent with the trend of PER for the queries in Table II. Therefore, the improvement in MAP for OOV queries can be due to the improvement in PER. Fig. 6 shows the recall precision curves for E2E (kana) and E2E (kana+NPbm). In low recall region, the precision of E2E (kana) and E2E (kana+NPbm) is almost the same. In contrast, in high recall, E2E (kana+NPbm) has a higher precision.

Next, we compare the effect of query-specific language model. E2E (kana+NPbm) ASR is used for comparison. Table VI shows the PER when the language model weights are changed for the spoken queries in the development set. When the weight was set to 0.1, the performance improved the most. As the weight was increased, the performance deteriorated and ASR output for some queries were not remained. E2E ASR is trained with full transcripts including sentence unit, while LM_{query} is trained with only noun phrases. Therefore, the mismatch between latent language model and external language model may be the reason why the recognition performance becomes worse when the weights are increased. Based on this result, we set the weight of the LM_{query} to 0.1 in the test set.

TABLE VII: Effect of query-specific language model for E2E (kana+NPbm) ASR (test set)

Query type	External LM	Query PER [%]	MAP [%]
ALL	LM_{doc}	4.24	73.9
	LM_{query}	4.11	74.6
IV	LM_{doc}	4.22	72.8
	LM_{query}	4.14	73.6
OOV	LM_{doc}	4.28	76.8
	LM_{query}	4.04	77.2

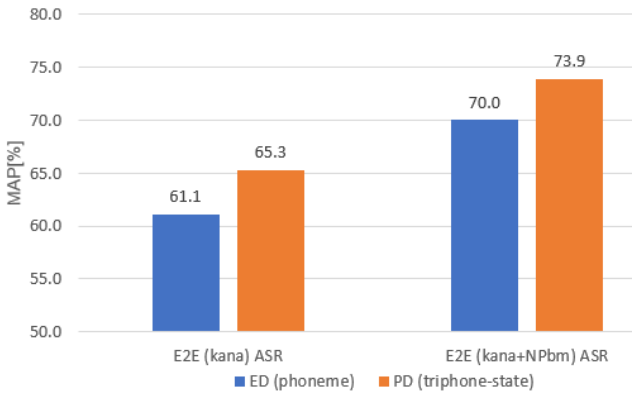


Fig. 7: Effect of different units and distance measures on STD performance with E2E ASR (all queries in test set)

The results of the PER and the STD experiment when using query-specific language model are shown in Table VII. PER was reduced and STD performance was improved when the language model was adapted for query. This result shows that even when using a common E2E ASR model changing the external language model for query is effective in improving PER for spoken queries and STD performance.

Finally, we compare the effect of unit (phoneme vs. triphone-state) and acoustic distance (ED vs. PD) for spotting with E2E ASR output (Fig. 7). The retrieval performance was improved by using PD in both E2E (kana) and E2E (kana+NPbm). This improvement was also observed for E2E (kana+NPbm) which have high phoneme level speech recognition accuracy, indicating the effectiveness of spotting at the subphone level. This result support our findings in the conventional ASR-based systems [5].

V. Conclusions and Future Work

In this paper, we proposed STD method using an E2E ASR system adapted to speech retrieval and considering acoustic similarity at the subphone level. E2E ASR adapted to STD was used to reduce the effect of OOV and improved the retrieval performance. Our proposed E2E ASR modeling, especially with additional mark to E2E output label, could improve the recognition performance for spoken queries, and the retrieval performance was significantly improved. In addition, we showed an improvement in PER and STD performance with only separated language model for query. The results showed that spotting using acoustic similarity at subphone level outperformed spotting at the phoneme level even when using best performed E2E ASR at the phoneme level.

In this study, the sub-phone level acoustic dissimilarity is derived using the conventional DNN-HMM acoustic model. This may have caused a mismatch with the E2E ASR output. Future work is to perform subphone level spotting using only E2E ASR backend and further improvement in retrieval performance can be expected. Also, there is room for improvement in ASR modeling, which balances retrieval oriented and whole speech-to-text accuracy.

References

[1] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Morc Delcroix, Atsunori Ogawa and Tomohiro Nakatani, “Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration”, Proc. of INTERSPEECH, pp. 1408–1412, 2019.

[2] Tomoyasi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo and Gareth J. F. Jones, “Overview of the NTCIR12 SpokenQuery&Doc-2 task”, Proc. of the NTCIR-12 Conference, Tokyo, Japan, 2016.

[3] Seiichi Nakagawa, Keisuke Iwami, Yasuhisa Fujii and Kazumasa Yamamoto, “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.

[4] Mitsuaki Makino, Naoki Yamamoto and 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] Shuji Oishi, Tatsuya Matsuba, Mitsuaki Makino and Atsuhiko Kai, “Combining state-level spotting and posterior-based acoustic match for improved query-by-example spoken term detection”, Proc. INTERSPEECH 2016, pp.740-744, 2016.

[6] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey and Tomoki Hayashi, “Hybrid CTC/Attention Architecture for End-to-End Speech Recognition”, Proc of IEEE Journal of Selected Topics in Signal, Vol.11, No.8, pp.1240-1253, 2017.

[7] Linhao Dong, Shuang Xu and Bo Xu, “Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition”, Proc of ICASSP, pp. 5884-5888, 2018.

[8] Ryota Konno, Kazunori Kojima, Shi-Wook Lee, Kazuya Tanaka and Yoshiaki 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.

[9] Hiroki Kondo, Atsuhiko Kai and Shuji 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)

[10] Takumi Kurokawa, Atsuhiko Kai and Hiroki Kondo, “Effects of End-to-end ASR and Score Fusion Model Learning for Improved Query-by-example Spoken Term Detection”, Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp.654-661, 2020.

[11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek and Nagendra Goel et al.,“The Kaldi Speech Recognition Toolkit”, Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

[12] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba and Yuya Unno et al., “ESPnet: End-to-End Speech Processing Toolkit”, Proc of INTERSPEECH, pp. 2207–2211, 2018.

[13] Kikuo Maekawa, “Corpus of Spontaneous Japanese : its design and evaluation”, Proc of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 7-12, 2003.

[14] Kikuo Maekawa, “Corpus of Spontaneous Japanese : Design, annotation, and XML representation”, Proc of Large-scale Knowledge Resource, 2004.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones and Aidan N. Gomez et al., “Attention is All You Need”, Proc of the 31st International Conference on Neural Information Processing System, pp. 6000-6010, 2017.