

End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning

RYO IMAIZUMI¹, RYO MASUMURA², SAYAKA SHIOTA¹ AND HITOSHI KIYA¹

End-to-end systems have demonstrated state-of-the-art performance on many tasks related to automatic speech recognition (ASR) and dialect identification (DID). In this paper, we propose multi-task learning of Japanese DID and multi-dialect ASR (MD-ASR) systems with end-to-end models. Since Japanese dialects have variety in both linguistic and acoustic aspects of each dialect, Japanese DID requires simultaneously considering linguistic and acoustic features. One solution realizing Japanese DID using these features is to use transcriptions from ASR when performing DID. However, transcribing Japanese multi-dialect speech into text is regarded as a challenging task in ASR because there are big gaps in linguistic and acoustic features between a dialect and standard Japanese. One solution is dialect-aware ASR modeling, which means DID is performed with ASR. Therefore, the multi-task learning framework of Japanese DID and ASR is proposed to represent the dependency of them. We explore three systems as part of the proposed framework, changing the order in which DID and ASR are performed. In the experiments, Japanese multi-dialect ASR and DID tests were conducted on our home-made Japanese multi-dialect database and a standard Japanese database. The proposed transformer-based systems outperformed the conventional single task systems on both DID and ASR tests.

Keywords: Japanese multi-dialect automatic speech recognition, Japanese dialect identification, multi-task learning, transformer-based encoder-decoder, end-to-end model

I. INTRODUCTION

Recently, deep learning-based end-to-end (E2E) systems have demonstrated state-of-the-art performance on many tasks in speech processing, such as automatic speech recognition (ASR) and dialect identification (DID) tasks [1–4]. Since it is known that the performance of E2E systems depends on the amount of training data [5, 6], many databases are combined for use in many tasks [7, 8]. Using combined databases is sometimes necessary to deal with multi-condition modeling such as recording environments, speaking styles, and languages. Thus, many methods have been reported to consider multi-condition scenarios, for example, accented speech recognition [9, 10] and multi-lingual speech recognition [11, 12].

Combining multi-dialect Japanese and standard Japanese databases is also one multi-condition task. Since each Japanese dialect has lots of dialect-specific accents, vocabulary, and phrases, tasks related to Japanese dialects require considering variety in terms of both linguistic and acoustic aspects. To treat such a multi-condition task, in this paper, we focus on Japanese dialect identification and multi-dialect ASR (MD-ASR) tasks. Japanese DID and MD-ASR are highly dependent on each other because their linguistic and acoustic properties are different from those

of English and other accents. The existing studies (e.g. [10] and Viglino *et al.* [25]) are regarded the accent as subsidiary information, and not focus on both tasks as in equal contribution.

DID is the task of automatically identifying a dialect from a text or speech sequence. The conventional DID methods are categorized into two groups. One is using acoustic features such as accent and speaker-specific characteristics as input features [13, 14]. The other is using a text sequence as a linguistic feature for input features [15, 16]. Since Japanese dialects have variety in both the linguistic and acoustic aspects of each dialect, it is desirable to consider both linguistic and acoustic features simultaneously in Japanese DID. As a similar task to DID, a multi-lingual identification method has been reported by which ASR is performed on speech to obtain a text sequence for an identification task [17]. Inspired by this concept, we consider carrying out Japanese MD-ASR for Japanese DID with a one-model architecture.

Japanese MD-ASR involves robustly transcribing both multi-dialect and standard Japanese speech into text with a one-model architecture. One state-of-the-art MD-ASR system, E2E MD-ASR, has been actively researched [18, 19]. E2E MD-ASR systems basically use acoustic features as input. It is known that linguistic features are learned as hidden states in a network [20, 21]. Recognizing multi-dialect Japanese sequences is regarded as a challenging task for ASR because there are big gaps in terms of linguistic and acoustic features between a dialect and the standard language [13, 22, 23]. In [24, 25], a method of combining

¹Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo, 191-0065, Japan

²NTT Media Intelligence Laboratories, NTT Corporation, Japan

accent identification with accent speech recognition was reported to alleviate the big gap for acoustic features. Since the task is similar to Japanese MD-ASR, Japanese DID is also useful when combined with Japanese MD-ASR.

To joint model Japanese DID and MD-ASR, we propose E2E models with multi-task learning with Japanese DID and MD-ASR to utilize dialect information and text sequences, simultaneously. In this paper, we propose multi-task learning of Japanese DID and multi-dialect ASR (MD-ASR) systems with end-to-end models, changing the order in which DID and MD-ASR is performed. The first system, called “DID2ASR,” performs DID and MD-ASR in a series to influence an estimated dialect in word sequence estimation. It means DID2ASR performs ASR with a dialect-aware condition. DID2ASR is similar to performing DID then performing MD-ASR constructed for a dialect. The second system, called “ASR2DID,” involves reversing the order of MD-ASR and DID from the first system. Since DID of ASR2DID can consider the text information of MD-ASR more effectively than the first system, the performance of the DID should be higher than that of the first system. As a similar concept to ASR2DID, the third system, called “DID+ASR,” performs DID and MD-ASR simultaneously. DID+ASR is different in terms of the representation of the dialect label, which is represented by a probability distribution of registered dialects, while DID2ASR and ASR2DID select one dialect. Since the three proposed systems perform DID and MD-ASR in a multi-task manner, the dependency of DID and MD-ASR is maintained in modeling the network. Therefore, it is possible to improve the performance on both DID and MD-ASR tasks. In our experiments, a home-made database consisting of six Japanese dialects and a standard Japanese database were used for constructing transformer-based multi-task models. From the results, on the Japanese MD-ASR task, we demonstrate that DID+ASR outperformed the conventional multi-condition modeling and achieved an error reduction of 9.9%. On the Japanese DID task, ASR2DID improved the identification accuracy by 15.3% compared with the conventional system.

This paper is organized as follows. Section II describes the motivation for using multiple dialects, E2E Japanese DID, and E2E ASR. Section III explains the transformer-based network architecture. Then, the proposed methods are presented in Section IV. Experimental conditions and results are shown in Section V. Finally, Section VI concludes our work.

II. CHALLENGES WITH MULTI-DIALECT JAPANESE

A) Motivation for using Multi-dialect Japanese Data

There are more than 100 different dialects across Japan [26]. Each of them has many dialect-specific accents, vocabularies, and phrases (see Appendix). For example, the

English word “I” is translated into “wa-ta-shi” in standard Japanese. In the case of the dialect of Sendai Prefecture, “I” is translated into “o-ra.” Like this example, there are many cases in which the meaning of a word is the same in a dialect and standard Japanese, but the pronunciations are completely different. In other cases, the majority of the sentence is the same as in standard Japanese, but only the end of the sentence has a unique feature in the dialect (see comparison of Hiroshima with standard Japanese in Appendix). Additionally, some vocabulary and phrases are found only in certain dialects. It is known that when dialects are mixed with standard Japanese for identification or recognition tasks, the performance deteriorates because dialects have different features from standard Japanese. Therefore, it is important to deal with acoustic and linguistic information adequately for tasks related to multiple Japanese dialect. Therefore, it is important to deal with acoustic and linguistic information adequately for tasks related to multiple Japanese dialects. Since the multi-task learning-based models for similar tasks such as accented speech recognition treat another task to ASR like subsidiary information, the importance both of acoustic and linguistic information is different point from the similar tasks.

B) E2E Japanese Dialect Identification

This section describes two conventional models of DID. One model predicts the probability of generating a dialect label l from a speech sequence $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$, where \mathbf{x}_m is the m -th acoustic feature in speech. The other model predicts the probability of generating dialect label l from a text sequence $\phi(\mathbf{X}) = \phi(\{\mathbf{x}_1, \dots, \mathbf{x}_M\})$. Let $\phi(\mathbf{X})$ be the text obtained by ASR $\phi(\cdot)$. M is the number of acoustic features in speech. In the case of using dialect speech as input, DID models using sequence-to-one neural networks define the identification probability of l as $P(l|\mathbf{X}; \Theta_{\text{did}})$, where Θ_{did} represents a model of DID parameter sets. In the case of using text as input, the identification probability of a dialect label l is defined as $P(l|\phi(\mathbf{X}); \Theta_{\text{did}})$. In the Japanese DID, a whole model parameter set can be optimized from speech-to-label paired data:

$$\mathcal{D} = \{(\mathbf{X}^1, l^1), \dots, (\mathbf{X}^T, l^T)\}, \quad (1)$$

where T is the number of utterances in a training data set. The objective function of the model for identification is defined as

$$\mathcal{L}_{\text{did}}(\Theta_{\text{did}}) = - \sum_{t=1}^T \log P(l^t|\mathbf{X}^t; \Theta_{\text{did}}). \quad (2)$$

Since Japanese dialects have variety in terms of both the linguistic and acoustic aspects of each dialect, it is not sufficient to use only linguistic or acoustic features for Japanese DID.

C) E2E Japanese ASR

This section describes a model for E2E Japanese ASR. This model predicts the generation probability of a text sequence

\mathbf{W} given a speech sequence \mathbf{X} , where w_n is the n -th token in the text. N is the number of tokens in the text. Auto-regressive generative models define the generation probability of \mathbf{W} as

$$P(\mathbf{W}|\mathbf{X}; \Theta_{\text{asr}}) = \prod_{n=1}^N P(w_n | \mathbf{W}_{1:n-1}, \mathbf{X}; \Theta_{\text{asr}}), \quad (3)$$

where $\mathbf{W}_{1:n-1} = \{w_1, \dots, w_{n-1}\}$ and Θ_{asr} represents a model of ASR parameter sets. In E2E ASR, a model parameter set can be optimized from the utterance-level labeled data (speech-to-text paired data) as

$$\mathcal{D} = \{(\mathbf{X}^1, \mathbf{W}^1), \dots, (\mathbf{X}^T, \mathbf{W}^T)\}, \quad (4)$$

where T is the number of utterances in the training data set. The objective function of ASR, based on maximum likelihood estimation, is defined as

$$\mathcal{L}_{\text{asr}}(\Theta_{\text{asr}}) = - \sum_{t=1}^T \sum_{n=1}^{N^t} \log P(w_n^t | \mathbf{W}_{1:n-1}^t, \mathbf{X}^t; \Theta_{\text{asr}}), \quad (5)$$

where w_n^t is the n -th token for the t -th utterance, and $\mathbf{W}_{1:n-1}^t = \{w_1^t, \dots, w_{n-1}^t\}$. N^t is the number of tokens for the t -th utterance.

Recognizing Japanese multi-dialect sequences is regarded as a challenging task for ASR because there are big gaps in terms of linguistic and acoustic features between a dialect and standard Japanese. Therefore, it is desirable to use the acoustic features of dialects obtained by DID for ASR.

III. TRANSFORMER-BASED NETWORK ARCHITECTURE

This section describes a general transformer-based E2E model [27, 28]. The transformer-based model has an encoder and a decoder that are composed of several transformer blocks. The DID system uses dialect identification instead of a text decoder to calculate $P(l|\mathbf{X}; \Theta)$ as a sequence-to-one model, while the ASR system uses a speech encoder and a text decoder as a sequence-to-sequence model to calculate $P(\mathbf{W}|\mathbf{X}; \Theta)$. For DID systems, the model of parameter sets Θ is split into a speech encoder θ_{enc} and dialect identification θ_{id} . For ASR systems, the model of parameter sets Θ is split into a speech encoder θ_{enc} and a text decoder θ_{dec} .

Speech encoder: The speech encoder converts input acoustic features into hidden representations $\mathbf{H}^{(I)}$ using I transformer encoder blocks. The i -th transformer encoder block composes the i -th hidden representations $\mathbf{H}^{(i)}$ from the lower layer inputs $\mathbf{H}^{(i-1)}$ as indicated by

$$\mathbf{H}^{(i)} = \text{TransformerEncoderBlock}(\mathbf{H}^{(i-1)}; \theta_{\text{enc}}), \quad (6)$$

where $\text{TransformerEncoderBlock}()$ is a transformer encoder block that consists of a scaled dot-product multi-head

self-attention layer and a position-wise feed-forward network. The hidden representation $\mathbf{H}^{(0)} = \{\mathbf{h}_1^{(0)}, \dots, \mathbf{h}_{M'}^{(0)}\}$ is produced by

$$\mathbf{h}_{m'}^{(0)} = \text{AddPositionalEncoding}(\mathbf{h}_{m'}), \quad (7)$$

where $\text{AddPositionalEncoding}()$ is a function that adds a continuous vector in which position information is embedded. $\mathbf{H} = \{\mathbf{h}_1, \dots, \mathbf{h}_{M'}\}$ is produced by

$$\mathbf{H} = \text{ConvolutionPooling}(\mathbf{x}_1, \dots, \mathbf{x}_{M'}; \theta_{\text{enc}}), \quad (8)$$

where $\text{ConvolutionPooling}()$ is a function composed of convolution layers and pooling layers. M' is the subsampled sequence length, which depends on the function.

Dialect identification: In dialect identification, a hidden representation \mathbf{H} , which is the output of the speech encoder, is used as input, and the generation probability of the dialect label l is calculated by the following equation as

$$P(l|\mathbf{X}; \theta_{\text{enc}}, \theta_{\text{id}}) = \text{Softmax}(\mathbf{s}; \theta_{\text{id}}), \quad (9)$$

$$\mathbf{s} = \text{Attention}(\mathbf{H}^{(I)}; \theta_{\text{id}}). \quad (10)$$

The attention function is a layer of the attention mechanism that tries and weighs the importance of hidden representations. A softmax function is applied to the output \mathbf{s} of the attention layer to calculate the predicted probability of a label.

Text decoder: The text decoder computes the generation probability of a token from preceding tokens and the hidden representations of speech. The predicted probabilities of the n -th token w_n are calculated as

$$P(w_n | \mathbf{W}_{1:n-1}, \mathbf{X}; \theta_{\text{enc}}, \theta_{\text{dec}}) = \text{Softmax}(\mathbf{u}_{n-1}^{(J)}; \theta_{\text{dec}}), \quad (11)$$

where $\text{Softmax}()$ is a softmax layer with a linear transformation. The input hidden vector $\mathbf{u}_{n-1}^{(J)}$ is computed from J transformer decoder blocks. The j -th transformer decoder block composes the j -th hidden representation $\mathbf{u}_{n-1}^{(j)}$ from the lower inputs $\mathbf{U}_{1:n-1}^{(j-1)} = \{\mathbf{u}_1^{(j-1)}, \dots, \mathbf{u}_{n-1}^{(j-1)}\}$ as

$$\mathbf{u}_{n-1}^{(j)} = \text{TransformerDecoderBlock}(\mathbf{U}_{1:n-1}^{(j-1)}, \mathbf{H}^{(I)}; \theta_{\text{dec}}), \quad (12)$$

where $\text{TransformerDecoderBlock}()$ is a transformer decoder block that consists of a scaled dot-product multi-head self-attention layer, a scaled dot-product multi-head source-target attention layer, and a position-wise feed-forward network. The hidden representation $\mathbf{U}_{1:n-1}^{(0)} = \{\mathbf{u}_1^{(0)}, \dots, \mathbf{u}_{n-1}^{(0)}\}$ is produced by

$$\mathbf{u}_{n-1}^{(0)} = \text{AddPositionalEncoding}(\mathbf{w}_{n-1}), \quad (13)$$

$$\mathbf{w}_{n-1} = \text{Embedding}(w_{n-1}; \theta_{\text{dec}}), \quad (14)$$

where $\text{Embedding}()$ is a linear layer that embeds an input token into a continuous vector.

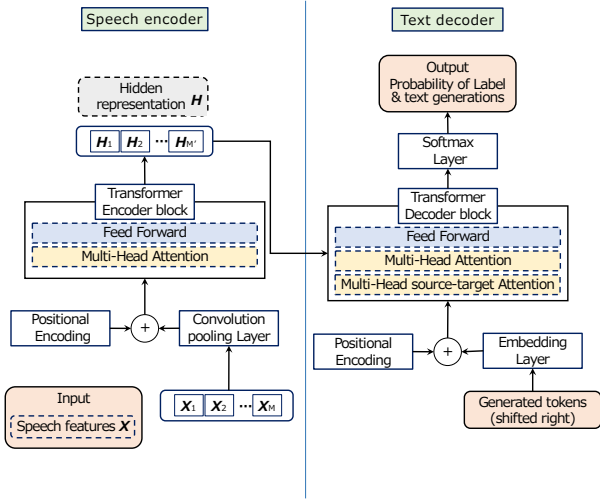


Fig. 1. Network architecture of transformer-based ASR model

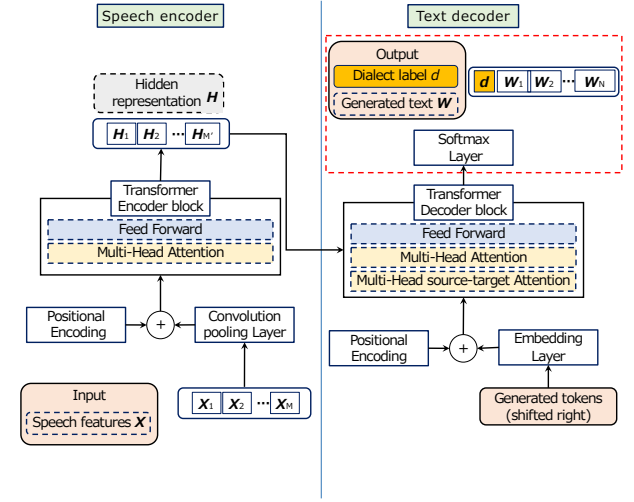


Fig. 4. Network architecture of multi-task learning systems performing DID and MD-ASR in series (DID2ASR)

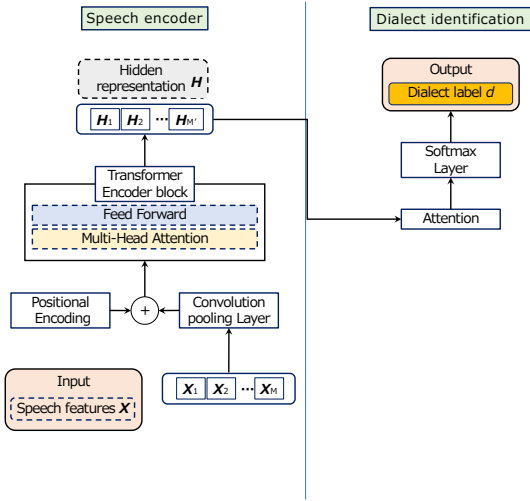


Fig. 2. Network architecture of transformer-based DID model using speech features as input

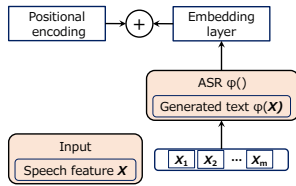


Fig. 3. Part of network architecture of transformer-based DID model using text features as input, applicable to red dashed-line box of Fig. 2

IV. MULTI-TASK LEARNING OF JAPANESE DID AND MD-ASR

A) Serial estimation (DID2ASR)

This section describes the transformer-based system in DID2ASR. This system focuses on estimating a word sequence by using estimated dialect information. DID2ASR is defined by adding a dialect label d and a text sequence \mathbf{W} to the conventional ASR model $P(l|\mathbf{X}; \Theta)$.

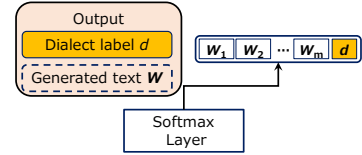


Fig. 5. Part of network architecture of multi-task learning systems performing MD-ASR and DID in series (ASR2DID), applicable to red dashed-line box of Fig. 4

The dialect label “ d ” is regarded as one token. The generation probability is redefined as

$$\begin{aligned} P(\mathbf{W}, d|\mathbf{X}; \Theta) &= P(\mathbf{W}|\mathbf{X}, d; \Theta)P(d|\mathbf{X}; \Theta) \\ &= P(\mathbf{Z}|\mathbf{X}; \Theta), \end{aligned} \quad (15)$$

$$P(\mathbf{Z}|\mathbf{X}; \Theta) = \prod_{n=0}^N P(z_n|\mathbf{Z}_{1:n-1}, \mathbf{X}; \Theta), \quad (16)$$

where \mathbf{Z} is a concatenated sequence that is defined as $\mathbf{Z} = \{d, w_1, \dots, w_N\}$. Figure 4 shows a network architecture for DID2ASR. The architecture of the speech encoder is the same as the general transformer-based E2E ASR described in Section III. As shown in Fig. 4, the text decoder of DID2ASR is essentially the same as the general transformer-based E2E ASR, although the softmax function is used to calculate the generation probability of the dialect label d and text \mathbf{W} . The model parameter set can be optimized from a set of speech, dialect label, and text as

$$\mathcal{D}_{\text{mtl}} = \{(\mathbf{X}^1, d^1, \mathbf{W}^1), \dots, (\mathbf{X}^T, d^T, \mathbf{W}^T)\}. \quad (17)$$

The objective function used in the proposed method is defined as

$$\begin{aligned} \mathcal{L}_{D2A}(\theta_{\text{enc}}, \theta_{\text{dec}}) &= - \sum_{t=1}^T \log P(\mathbf{W}^t, d^t | \mathbf{X}^t; \theta_{\text{enc}}, \theta_{\text{dec}}) \\ &= - \sum_{t=1}^T \sum_{n=1}^{|\mathbf{Z}^t|} \log P(z_n^t | \mathbf{Z}_{1:n-1}^t, \mathbf{X}^t; \theta_{\text{enc}}, \theta_{\text{dec}}). \end{aligned} \quad (18)$$

By optimizing with dialect labels, speech text, and generation probabilities, dialect-specific characteristics are clearly recognized, and this helps in estimating models without dialect-specific confusion. Since DID2ASR treats dialect labels uniquely, the MD-ASR performance of the direct modeling method depends on the accuracy of the DID.

B) Serial estimation (ASR2DID)

Next, we propose ASR2DID to reverse the order of MD-ASR and DID in DID2ASR. In ASR2DID, the generation probability is redefined as

$$\begin{aligned} P(\mathbf{W}, d | \mathbf{X}; \Theta) &= P(\mathbf{W} | \mathbf{X}; \Theta) P(d | \mathbf{X}, \mathbf{W}; \Theta) \\ &= P(\mathbf{Y} | \mathbf{X}; \Theta), \end{aligned} \quad (19)$$

where \mathbf{Y} is a concatenated sequence that is defined as $\mathbf{Y} = \{w_0, \dots, w_{N-1}, d\}$. The objective function used in the proposed method is defined as

$$\begin{aligned} \mathcal{L}_{A2D}(\theta_{\text{enc}}, \theta_{\text{dec}}) &= - \sum_{t=1}^T \log P(\mathbf{W}^t, d^t | \mathbf{X}^t; \theta_{\text{enc}}, \theta_{\text{dec}}) \\ &= - \sum_{t=1}^T \sum_{n=1}^{|\mathbf{Y}^t|} \log P(y_n^t | \mathbf{Y}_{1:n-1}^t, \mathbf{X}^t; \theta_{\text{enc}}, \theta_{\text{dec}}). \end{aligned} \quad (20)$$

This means that ASR2DID changes only a part of the text decoder of DID2ASR, applying Fig. 5 to the red dashed-line box in Fig. 4. The DID of ASR2DID should perform better than that of DID2ASR because the DID is performed with an estimated word sequence as a linguistic feature.

C) Joint estimation (DID+ASR)

As the third proposed method, we propose joint estimation of DID and ASR in one network, called ‘‘DID+ASR.’’ The concept of DID+ASR is similar to ASR2DID. However, the dialect information of DID+ASR is represented by a probability distribution of registered dialects, whereas DID2ASR and ASR2DID use one dialect label. In DID+ASR, a phoneme sequence and probability distribution of dialects can be predicted from input speech. The difference from Section IV.A IV.B is that the generation probabilities of the dialect label d and \mathbf{W} are independent, and dialect information is represented in a probability distribution. The generation probability of dialect label d and

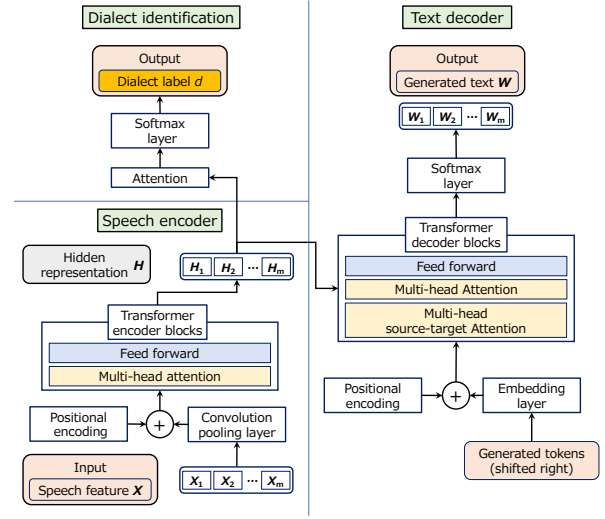


Fig. 6. Network architecture of multi-task learning systems with joint DID and ASR (DID+ASR)

\mathbf{W} is redefined as

$$P(\mathbf{W}, d | \mathbf{X}; \Theta) = P(\mathbf{W} | \mathbf{X}; \Theta) P(d | \mathbf{X}; \Theta). \quad (21)$$

The architecture of DID+ASR is a combination of the speech encoder, text decoder, and identification used in Section III.

We use the same training data described in eq. (17) to optimize the DID+ASR model. The objective function of the model for identification is defined as

$$\begin{aligned} \mathcal{L}(\theta_{\text{enc}}, \theta_{\text{did}}, \theta_{\text{dec}}) &= \\ &\alpha \mathcal{L}_{\text{mle}}(\theta_{\text{enc}}, \theta_{\text{dec}}) + \beta \mathcal{L}_{\text{id}}(\theta_{\text{enc}}, \theta_{\text{id}}), \end{aligned} \quad (22)$$

where α and β are the loss weights for ASR and DID. Figure 6 shows the network architecture of DID+ASR. The difference from the general transformer, E2E MD-ASR, is the added DID part. Since the dialect label is regarded as a probability distribution, the degree of dialect-specific information can be recognized, and both multi-dialect data and standard Japanese data can be utilized effectively.

V. EXPERIMENTS

A) Database

A home-made speech database of Japanese dialects [13] and a database of standard Japanese were used in all experiments. The dialect database consisted of six dialects: Aomori, Hiroshima, Kumamoto, Nagoya, Sapporo, and Sendai. The numbers of training and test utterances were 89,817 and 7,450, respectively. Details on the database are shown in Table 1. The OOV was replaced with an UNK label. The OOV rate was 0.00011%. Details on the database are shown in Table 1. The gender ratios of the speakers for each dialect were almost the same. Each dialect utterance was recorded by using an iPhone 5 or an Xperia Z1. The length of each dialect utterance was

Table 1. Number of utterances and times for each dialect and standard Japanese

		Train		Valid		Test		All
		Utt	Time	Utt	Time	Utt	Time	Utt
Dialect	Aomori	10741	21:42:17	676	1:04:54	676	1:03:43	12093
	Hiroshima	18670	35:30:57	566	1:03:25	567	1:02:44	19803
	Kumamoto	9328	19:20:09	719	1:13:08	719	1:13:25	10766
	Nagoya	18611	31:35:39	551	1:07:13	550	1:07:40	19712
	Sapporo	15955	32:11:43	678	1:11:24	678	1:11:35	17311
	Sendai	16512	35:54:01	535	1:09:13	535	1:09:53	17582
Total		89817	176:14:46	3725	6:49:17	3724	6:49:10	97267
Standard		162243	290	1292	2:20	2573	5:00	166108
Dialect+Standard		252060	466:14:46	5017	9:19:17	6297	11:49:10	263375

about 7 seconds, and the content of the dialect database was daily conversations. For the standard Japanese database, the Corpus of Spontaneous Japanese (CSJ) [29], consisting of academic lectures and simulated public speeches, was used. The numbers of training and test utterances were 162,243 and 3,865, respectively. The number of male speakers was about double that of female speakers. Both databases were sampled at 16 kHz and quantized to 16 bit. All transcriptions of both databases were hand-labeled.

B) Experiment conditions

The transformer-based E2E ASR and E2E DID described in Section III was used as a conventional model. Therefore, the main architectures of the proposed models and the conventional one were the same, and the common parts of the architectures were set as follows. The transformer network consisted of eight encoder blocks and six decoder blocks. All functions used in the transformer networks were implemented in accordance with [27]. Regarding the composition of the transformer blocks, the dimension of the continuous vector was 256, the dimension of the inner outputs in the position-wise feed-forward networks was 2,048, and the number of attention heads was set to 4. The parameters for the speech encoder and the text decoder were the same as in [13]. For the DID of DID+ASR, a seven-dimensional vector consisting of six dialects and standard Japanese language was outputted from the softmax layer. As the predict labels, six dialect labels; <aom>, <hir>, <kum>, <nag>, <sap>, <sen> and one standard language label; <jap> were settled. In the prediction part of the dialect label, one label was selected as a dialect token from the seven labels. The unit of tokens was represented as a character. In ASR decoding, all characters were used for the selection of a token. The objective function of multi-task learning was cross-entropy loss. The loss of the MD-ASR and DID weight, α and β , were set to 1 and 0.01, respectively. For DID2ASR and ASR2DID, dialect labels were put in the embedding layer and treated as a 256-dimensional vector. In the MD-ASR case, as a training set, three types of training data were prepared: dialect only, standard Japanese only, and joint data of dialect and standard Japanese. There were three sets of test data: dialect only, standard Japanese only, and joint data of dialect and

Table 2. ACCs (%) of conventional and proposed methods for DID

	ACC
Conventional method (speech)	71.2
Conventional method (text)	53.3
DID2ASR	83.1
ASR2DID	86.5
DID+ASR	81.8

standard Japanese. As methods to be compared, we adopted three types of training data for the conventional method, and only the joint data was adopted for the proposed methods. For DID2ASR, since dialect labels were estimated, the performance depended on the accuracy of the DID. Therefore, as a method for comparison, a case where all identifications were correct was prepared as an oracle. In the DID case, the character error rate (CER) was used to evaluate ASR, and accuracy (ACC) was used to evaluate the DID of the proposed methods.

$$\text{CER} = \left(1 - \frac{\text{COR} - \text{INS}}{\text{TOTAL}}\right) \times 100 (\%), \quad (23)$$

$$\text{ACC} = \frac{\# \text{ of correct files}}{\# \text{ of total files}} \times 100 (\%), \quad (24)$$

where COR and INS were the numbers of correct characters and inserted characters, respectively. TOTAL was the total number of characters. In the case of ASR+DID, the highest probability in the output of the softmax layer for dialect identification was selected to calculate ACC.

C) Results

Table 2 shows the ACCs of the conventional and proposed methods. There were two versions of the conventional method, one using speech as an input feature and one using text, and they both performed insufficiently, 71.2% and 53.3%, respectively. As described in Sections II.A and II.B, this is because Japanese dialects have unique information in terms of both acoustic and linguistic aspects, and using one side as input features did not lead to good performance. In comparison, the ACCs of three proposed methods were improved compared with two conventional methods. Compared ACC of DID2ASR, 83.1%, with that

Table 3. CERs (%) of conventional and proposed methods for each combination of databases for MD-ASR

	Training	Dialect	Standard	Dialect+Standard
Conventional method	Dialect only	52.9	100↑	100↑
	Standard only	52.5	14.3	64.7
	Dialect + Standard	8.0	14.9	11.1
DID2ASR	Dialect + Standard	8.4	14.2	11.0
DID2ASR (oracle)	Dialect + Standard	7.3	14.2	10.4
ASR2DID	Dialect + Standard	7.0	14.5	10.4
DID+ASR	Dialect + Standard	7.2	13.4	10.0

Table 4. CERs (%) of proposed methods in case of correct or incorrect DID

	DID2ASR	ASR2DID	DID+ASR
Correct	5.0	5.2	6.2
Incorrect	14.0	12.8	9.8

Table 5. CERs (%) of conventional and proposed methods trained with the joint training data (Dialect+Standard), except for Kumamoto data. Kumamoto dialect was regarded as an unknown dialect. The joint test data (Dialect+Standard) excluding Kumamoto was labeled as “Known” and the test data which include only Kumamoto data was labeled as “Unknown”

	Known	Unknown
Conventional method	10.9	16.9
DID2ASR	11.3	17.6
ASR2DID	11.3	16.8
DID+ASR	10.1	15.2

of ASR2DID, 86.5%, ASR2DID provided higher performance than DID2ASR. For predicting a dialect label, ASR2DID used the predicted ASR results effectively. However, in DID2ASR, since a dialect label was predicted first, it’s prediction could not utilize the prediction ASR results fully. These results show that it is useful to use the information of MD-ASR for DID. For ACC of DID+ASR, 81.8%, the performance was slightly lower than ACCs of DID2ASR and ASR2DID. We considered the joint estimation of DID and MD-ASR required to find adequate parameter settings between DID and MD-ASR carefully. Consequently, these results showed that it was useful to use the information of MD-ASR for DID.

To give detailed results of the DID tasks, Fig. 7 shows the confusion matrices of each dialect for each method in Fig. 2. Comparing conventional method (speech) with conventional method (text), the trends of the accuracy for each dialect were different. This means that the acoustic and linguistic features captured different characteristics of the dialects, respectively. Comparing the conventional methods with the proposed methods, the performances of the proposed methods were improved for almost all dialects. In particular, using ASR2DID, the identification rate of Aomori and Kumamoto exceeded 95%. This result suggested that phoneme sequences were important information for identifying Aomori and Kumamoto. On the contrary, the discrimination performance of Nagoya decreased, indicating that the recognition information was not effective

because the phoneme sequence of Nagoya was similar to other dialects.

Table 3 shows the CERs of the conventional and proposed methods for the multi-condition task for ASR. When the conventional methods were trained with only dialect data or standard Japanese, the CERs were over 50%, except for the case of the conventional method text speech trained using standard Japanese only. The reason for obtaining such a poor CER was that mismatches between dialect-specific characteristics were not considered as described in Section II.A. When the conventional method was trained with multi-condition data (Dialect+Standard), the CER of the conventional method trained with standard Japanese only on the dialect-only test had quite a large improvement from 52.5% to 8.0%. However, the CER worsened from 14.3% to 14.9% on the only-standard-Japanese test. These results indicate that dialect-specific characteristics caused model estimation confusion. Next, in the case of the proposed systems, the CER of DID2ASR with that of the conventional method (Dialect+Standard), in the dialect-only test case, the CER of DID2ASR, 8.4%, was higher than that of the conventional method, 8.0%. However, in the standard-Japanese-only test case, the CER of DID2ASR, 14.2% was lower than that of the conventional method, 14.9%. The reason of the degradation of DID2ASR in the Dialect test case could be considered some the effects of the DID errors. To prove the reason of the degradation in the CER, we performed DID2ASR with the oracle dialect label. In the oracle case, the CER for the dialect-only test went down to 7.3%. Additionally, the performance of ASR of DID2ASR (oracle) in the standard-Japanese-only test, was not affected by the prediction error because the prediction of the standard language was easy due to the difference in database and the amount of training data. From these results for DID2ASR, DID2ASR could be used for robust ASR, but the performances depended on the accuracy of DID. Next, for ASR2DID on the dialect-only test, the CER went down to 7.0%. In DID2ASR, the effect of DID errors is significant because the prediction of text tokens for ASR is performed considering with the result of the dialect label prediction. The influence of DID errors is shown from the results of DID2ASR and DID2ASR (oracle). On the other hand, ASR2DID predicts a dialect label considering with the result of the ASR prediction. Therefore, we consider that the CER of ASR2DID has a smaller effect from DID errors than that of DID2ASR. However, the CER of ASR2DID in standard-Japanese-only test,

14.5%, case slightly degraded from that of DID2ASR. It means that DID2ASR caused over adaptation to dialect. In the DID+ASR case, DID+ASR achieved the lowest CER on the standard-Japanese test, and the CER of DID+ASR on the dialect-only test was the second best. This result indicates that using the probabilistic distribution of dialect labels and controlling the effect of DID improved the reliability of the model in terms of ASR performance.

To investigate the effect of the identification error on the ASR performances, Table 4 shows the CERs for the proposed methods for cases in which the DID results were correct or incorrect. Comparing the correct case with the incorrect one, the CER was lower in the correct case for all proposed methods. From these results, the MD-ASR performance of DID2ASR was the most affected by the DID error because the difference between the correct case and the incorrect one was nine points. ASR2DID and DID+ASR had lower CERs for the incorrect case than DID2ASR. This result indicates that the effect of DID error was mitigated, and ASR2DID and DID+ASR showed robust MD-ASR performances even though the estimation of DID was not correct. Furthermore, the difference between the correct and incorrect cases for DID+ASR was the smallest at 3.6%. The results showed that it was possible to mitigate the effects of DID error by handling dialect labels in a probability distribution.

Treating dialects as a probability distribution made it possible to represent unknown dialects by mixing the registered dialects. Table 5 shows the CERs in the case of recognizing utterances from an unknown dialect. In this experiment, the Kumamoto dialect was regarded as an unknown dialect and eliminated from the joint training data (Dialect+Standard). The joint test data (Dialect+Standard) separated into two groups; “Known” and “Unknown.” The “Known” group was the joint test data excluding Kumamoto data, and the “Unknown” test group was that the test data included only Kumamoto data. For known dialects, the results are almost the same as in Table 3. Comparing the conventional method with DID2ASR, the CER of DID2ASR was higher than the conventional method. The reason was that DID2ASR could not deal with the unknown dialect, so the effect of the DID error was serious as shown in Table 4. Since the results showed that ASR2DID estimated dialects after MD-ASR, it could relax the effect of error on unknown dialects. For DID+ASR, the CER of DID+ASR was the lowest. Since DID+ASR represented dialect information as a combination of the registered dialects, it could be considered that the performance for unknown dialects was more robust than DID2ASR and ASR2DID.

VI. CONCLUSION

In this paper, we propose multi-task learning of Japanese DID and multi-dialect ASR (MD-ASR) systems with end-to-end models. For serial estimation of two tasks in one model, DID2ASR and ASR2DID were described, and

DID+ASR involved the joint estimation of two tasks. The three methods were able to alleviate the differences between dialects and standard Japanese, and the accuracy of DID and dialect speech recognition was improved by using dialect-specific acoustic information and linguistic information. The experimental results showed that the three proposed methods outperformed the conventional method. From the prospect of the DID performance, ASR2DID had the best performance. When estimating dialect labels, it is useful to obtain linguistic and acoustic information in a series. Therefore, ASR2DID is useful to adopt as a pre-processing system for dialect-aware applications, such as spoken dialogue and entertainment. From the prospect of the MD-ASR performance, DID+ASR had the best CER. Additionally, in the case that the dialects of the input utterances were unknown, DID+ASR improved in performance.

As future work, we will experiment with other networks such as CTC/Attention hybrid system using other dialects. Also, we will experiment with combinations of the proposed methods.

REFERENCES

- [1] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in *proc. ICASSP*, 2017, pp. 4835–4839.
- [2] D. Amodei, S. Anantharayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, “Deep speech 2: End-to-end speech recognition in english and Mandarin,” in *proc. ICML*, 2016, pp. 173–182.
- [3] R. Gokay and H. Yalcin, “Improving low resource Turkish speech recognition with data augmentation and tts,” in *proc. SSD*, 2019, pp. 357–360.
- [4] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, “Speech recognition with augmented synthesized speech,” in *proc. ASRU*, 2019, pp. 996–1002.
- [5] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, “Back-translation-style data augmentation for end-to-end ASR,” in *proc. SLT*, 2018, pp. 426–433.
- [6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *proc. INTERSPEECH*, 2019, pp. 2613–2617.
- [7] T. Moriya, T. Ochiai, S. Karita, H. Sato, T. Tanaka, T. Ashihara, R. Masumura, Y. Shinohara, and M. Delcroix, “Self-distillation for improving ctc-transformer-based asr systems,” in *proc. INTERSPEECH*, 2020, pp. 546–550.
- [8] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *Transaction on speech and audio processing*, vol. 4, no. 1, p. 31, 1996.
- [9] K. Rao and H. Sak, “Multi-accent speech recognition with hierarchical grapheme based models,” in *proc. ICASSP*, 2017, pp. 4815–4819.
- [10] A. Jain, M. Upreti, and P. Jyothi, “Improved accented speech recognition using accent embeddings and multi-task learning,” in *proc. INTERSPEECH*, 2018, pp. 2454–2458.

		Correct label								
		Conv Speech	1	2	3	4	5	6		
Predict label	1	74.6	33.9	2.8	13.5	35.8	10.7			
	2	0.6	4.3	6.3	3.5	0.1	9.5			
	3	15.4	0.2	65.2	0.9	2.4	0.2			
	4	7.5	21.0	8.5	55.6	14.6	20.7			
	5	1.9	39.1	14.5	17.8	47.1	3.6			
	6	0.0	1.4	2.8	8.7	0.0	55.3			
		Conv Text	1	2	3	4	5	6		
		75.4	6.5	3.3	6.4	7.1	19.3			
		5.9	53.8	8.5	17.6	9.6	10.7			
		9.8	21.0	76.8	16.2	9.3	26.4			
		3.7	13.6	7.1	39.3	19.0	18.5			
		2.4	3.7	2.2	16.4	45.1	6.7			
		2.8	1.4	2.1	4.2	9.9	18.5			
		DID2 ASR	1	2	3	4	5	6		
		91.2	12.9	0.5	8.0	11.3	10.4			
		1.2	49.1	5.7	8.3	4.6	0.9			
		5.3	4.2	87.0	10.2	3.2	4.0			
		1.8	19.3	1.4	49.3	11.3	3.7			
		0.5	7.7	4.5	11.5	65.2	0.0			
		0.0	6.8	0.9	12.7	4.4	81.0			
		ASR2 DID	1	2	3	4	5	6		
		95.1	5.8	0.1	4.1	3.5	7.1			
		0.6	51.0	0.6	7.1	3.8	0.5			
		3.0	6.7	96.3	13.9	3.8	4.4			
		0.8	21.9	0.1	50.7	7.9	1.4			
		0.1	7.7	2.8	11.5	75.3	0.1			
		0.3	6.9	0.1	12.8	5.7	86.4			
		DID+ ASR	1	2	3	4	5	6		
		88.8	13.3	0.2	7.9	6.1	7.2			
		1.3	32.5	3.8	6.8	1.9	0.2			
		4.3	4.4	86.6	8.9	8.3	5.0			
		1.7	18.3	0.3	40.8	5.4	2.0			
		1.2	8.6	7.1	14.9	72.2	0.3			
		0.3	21.8	0.3	19.8	5.8	84.5			

Fig. 7. Confusion matrices of conventional and proposed methods on DID task

- [11] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *proc. ICASSP*, 2018, pp. 4909–4913.
- [12] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *proc. ICASSP*, 2018, pp. 4904–4908.
- [13] R. Imaiuzmi, R. Masumura, S. Shiota, and H. Kiya, "Dialect-aware modeling for end-to-end japanese dialect speech recognition," in *proc. APSIPA ASC*, 2020, pp. 297–301.
- [14] T. Purnell, W. Idsardi, and J. Baugh, "Perceptual and phonetic experiments on american english dialect identification," *Transaction on language and social psychology*, vol. 18, no. 1, pp. 10–30, 1999.
- [15] K. Abe, Y. Matsubayashi, N. Okazaki, and K. Inui, "Multi-dialect neural machine translation for 48 low-resource japanese dialects," *Transaction on Natural Language Processing*, vol. 27, no. 4, pp. 781–800, 2020.
- [16] G.-E. Zaharia, A.-M. Avram, D.-C. Cercel, and T. Rebedea, "Exploring the power of romanian bert for dialect identification," in *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 2020, pp. 232–241.
- [17] T. Okamoto, A. Hiroe, and H. Kawai, "Reducing latency for language identification based on large-vocabulary continuous speech recognition," *Transaction on Acoustical Science and Technology*, vol. 38, no. 1, pp. 38–41, 2017.
- [18] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *Transaction on Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [19] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *proc. INTERSPEECH*, 2017, pp. 949–953.
- [20] R. Masumura, M. Ihori, A. Takashima, T. Moriya, A. Ando, and Y. Shinohara, "Sequence-level consistency training for semi-supervised end-to-end automatic speech recognition," in *proc. ICASSP*, 2020, pp. 7054–7058.
- [21] S. Li, D. Raj, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation," in *proc. INTERSPEECH*, 2019, pp. 4400–4404.
- [22] Y. Zhao, J. Yue, X. Xu, L. Wu, and X. Li, "End-to-end-based Tibetan multitask speech recognition," *Transaction on Access*, vol. 7, pp. 162 519–162 529, 2019.
- [23] S. Li, X. Lu, C. Ding, P. Shen, T. Kawahara, and H. Kawai, "Investigating radical-based end-to-end speech recognition systems for Chinese dialects and japanese," in *proc. INTERSPEECH*, 2019, pp. 2200–2204.
- [24] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *proc. ICASSP*, 2018, pp. 4749–4753.
- [25] T. Vigliano, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition," in *proc. INTERSPEECH*, 2019, pp. 2140–2144.
- [26] D. Long, "Geographical perceptions of japanese dialect regions," *Handbook of perceptual dialectology*, vol. 1, pp. 177–198, 1999.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *proc. NIPS*, 2017, pp. 5998–6008.
- [28] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *proc. ACL*, 2019, pp. 4593–4601.

Table 6. Examples of Japanese dialect in dialect database with Japanese transcription and alphabet representation

Dialect	Text (Japanese)	Sequence
Aomori	何もはやってね	na N mo ha ya q te ne
Hiroshima	いいや私はやっとなんよ	i i ya wa ta shi ha ya q to ra N yo
Kumamoto	いえ私はやっとりません	i e wa ta shi ha ya q to ri ma se N
Nagoya	いや私はやっとなんよ	i ya wa ta shi ha ya q to ra N yo
Sapporo	いいえ自分はやってません	i e zi bu N ha ya q te na i wa
Sendai	いやおらはやってねえ	i ya o ra ha ya q te ne e
Standard	いいえ私はやってません	i i e wa ta shi wa ya q te ma se N

[29] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, “Spontaneous speech corpus of Japanese.” in *proc. LREC*, 2000, pp. 947–952.

APPENDIX

Table 6 illustrates examples of utterances in the dialect database used in our experiments. All sentences have the same meaning. As described in Section II.A, the phoneme sequence of Aomori is different from that of standard Japanese. In addition, almost all parts of the Nagoya and Hiroshima dialects are similar to those of standard Japanese. However, in many cases, the accents or other acoustic features are different from standard Japanese.



Ryo Imaizumi received his B.E. degree in Tokyo Metropolitan University, Tokyo, Japan in 2020. His research interests include dialect speech recognition and dialect identification.



Ryo Masumura received B.E., M.E., and Ph.D. degrees in engineering from Tohoku University, Sendai, Japan, in 2009, 2011, 2016, respectively. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2011, he has been engaged in research on speech recognition, spoken language processing, and natural language processing. He received the Student Award and the Awaya Kiyoshi Science Promotion Award from the Acoustic Society of Japan (ASJ) in 2011 and 2013, respectively, the Sendai Section Student Awards The Best Paper Prize from the Institute of Electrical and Electronics Engineers (IEEE) in 2011, the Yamashita SIG Research Award and the SIG-NL Excellent paper award from the Information Processing Society of Japan (IPSJ) in 2014 and 2018, the Young Researcher Award and the Paper Award from the Association for Natural Language Processing (NLP) in 2015 and 2020, the ISS Young Researcher’s Award in Speech Field and the ISS Excellent Paper Award from the Institute of Electronic, Information and Communication Engineers (IEICE) in 2015 and 2018. He is a member of the ASJ, the IPSJ, the

NLP, the IEEE, and the International Speech Communication Association (ISCA).



Sayaka Shiota received her B.E., M.E., and Ph.D. degrees in intelligence and computer science, Engineering and engineering simulation from Nagoya Institute of Technology, Nagoya, Japan in 2007, 2009 and 2012, respectively. From February 2013 to March 2014, she had worked at the Institute of statistical mathematics as a project assistant professor. In April of 2014, she joined Tokyo Metropolitan University as an Assistant Professor. Her research interests include statistical speech recognition and speaker verification. She is a member of ASJ, IPSJ, IEICE, APSIPA, ISCA, and IEEE.



Hitoshi Kiya received his B.E and M.E. degrees from Nagaoka University of Technology, in 1980 and 1982 respectively, and his Dr. Eng. degree from Tokyo Metropolitan University in 1987. In 1982, he joined Tokyo Metropolitan University, where he became a Full Professor in 2000. From 1995 to 1996, he attended the University of Sydney, Australia as a Visiting Fellow. He is a Fellow of IEEE, IEICE and ITE. He currently serves as President-Elect of APSIPA, and he served as Inaugural Vice President (Technical Activities) of APSIPA from 2009 to 2013, and as Regional Director-at-Large for Region 10 of the IEEE Signal Processing Society from 2016 to 2017. He was also President of the IEICE Engineering Sciences Society from 2011 to 2012, and he served there as a Vice President and Editor-in-Chief for IEICE Society Magazine and Society Publications. He was Editorial Board Member of eight journals, including IEEE Trans. on Signal Processing, Image Processing, and Information Forensics and Security, Chair of two technical committees and Member of nine technical committees including APSIPA Image, Video, and Multimedia Technical Committee (TC), and IEEE Information Forensics and Security TC. He has organized a lot of international conferences, in such roles as TPC Chair of IEEE ICASSP 2012 and as General Co-Chair of IEEE ISCAS 2019. He has received numerous awards, including six best paper awards.