IEICE TRANS. INF. & SYST., VOL.E106–D, NO.1 JANUARY 2023

PAPER

Comparative Evaluation of Diverse Features in Fluency Evaluation of Spontaneous Speech

Huaijin DENG[†a)], ***Nonmember***, **Takehito UTSURO**[†b)], **Akio KOBAYASHI**[††c)], ***and*** **Hiromitsu NISHIZAKI**[†††d)], ***Senior Members***

SUMMARY There have been lots of previous studies on fluency evaluation of spontaneous speech. However, most of them focus on lexical cues, and little emphasis is placed on how diverse acoustic features and deep end-to-end models contribute to improving the performance. In this paper, we describe multi-layer neural network to investigate not only lexical features extracted from transcription, but also consider utterance-level acoustic features from audio data. We also conduct the experiments to investigate the performance of end-to-end approaches with mel-spectrogram in this task. As the speech fluency evaluation task, we evaluate our proposed method in two binary classification tasks of fluent speech detection and disfluent speech detection. Speech data of around 10 seconds duration each with the annotation of the three classes of "fluent," "neutral," and "disfluent" is used for evaluation. According to the two way splits of those three classes, the task of fluent speech detection is defined as binary classification of fluent vs. neutral and disfluent, while that of disfluent speech detection is defined as binary classification of fluent and neutral vs. disfluent. We then conduct experiments with the purpose of comparative evaluation of multi-layer neural network with diverse features as well as end-to-end models. For the fluent speech detection, in the comparison of utterance-level disfluency-based, prosodic, and acoustic features with multi-layer neural network, disfluency-based and prosodic features only are better. More specifically, the performance improved a lot when removing all of the acoustic features from the full set of features, while the performance is damaged a lot if fillers related features are removed. Overall, however, the end-to-end Transformer+VGGNet model with mel-spectrogram achieves the best results. For the disfluent speech detection, the multi-layer neural network using disfluency-based, prosodic, and acoustic features without fillers achieves the best results. The end-to-end Transformer+VGGNet architecture also obtains high scores, whereas it is exceeded by the best results with the multi-layer neural network with significant difference. Thus, unlike in the fluent speech detection, disfluency-based and prosodic features other than fillers are still necessary in the disfluent speech detection.

key words: *speech fluency evaluation, disfluency, acoustic features, multi-layer neural network, end-to-end*

1. Introduction

Speech disfluencies, such as "um," "un" and false start, occur frequently in spontaneous speech. They directly affect the quality of the presentation such as those in university lectures and those speaking skills also affect student understanding [1], [2]. Also, it is difficult for second language learners to recognize their disfluencies. Therefore, an automatic fluency evaluation method is necessary. Although disfluencies tend to be viewed as noisy or irregular events, previous studies have found that disfluencies show remarkable regularities in a number of dimensions [3], which indicates that the speech fluency can be measured properly through certain specific features.

Among existing studies related to automatic speech fluency evaluation, most of them concentrate on detecting disfluencies through lexical cues alone. Zayats et al. [4] utilized features of reparandum and correction to make disfluency detection. Bach and Huang [5] proposed a method in which words are tagged according to fluent/disfluent states and the disfluency detection was formalized as a sequence labeling task. Previous studies [6], [7] also focused on the effectiveness of word fragments and fillers. However, these studies did not consider acoustic features or prosodic cues which have been proved to be useful in combination with lexical cues [8]–[10]. Acoustic features can carry more information such as energy and pitch that are not represented in transcripts. Both Lin et al. [8] and Zayats and Ostendorf [10] used pitch-related features, while Zayats and Ostendorf [10] also used energy-related features. However, it was insufficient that these studies did not analyze how much contribution those diverse acoustic features have to speech fluency evaluation.

Other previous studies [11]–[13] reported the effects of the use of lexical (i.e., disfluency-based and prosodic) features* as well as frame-level acoustic features in speech fluency evaluation. Especially, Deng et al. [13] adopted multi-layer neural network, while the performance depended on the lexical features and it is not clear whether the performance is to be damaged or not if we do not utilize these features. Therefore, it is necessary to conduct the evaluation experiment without lexical features.

Overall, previous studies mostly utilized acoustic, disfluency-based, and prosodic features. However, previous studies did not report the results of comprehensive comparison on the effects of all of those acoustic, disfluency-based, and prosodic features. Thus, this paper first aims at

Manuscript received March 31, 2022.
Manuscript revised August 12, 2022.
Manuscript publicized October 25, 2022.

[†]The authors are with Graduate School of Science and Technology, University of Tsukuba, Tsukuba-shi, 305–8573 Japan.
[††]The author is with Department of Industrial Information, Tsukuba University of Technology, Tsukuba-shi, 305–8520 Japan.
[†††]The author is with Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Kofu-shi, 400–8511 Japan.
a) E-mail: denghuaijin@yahoo.co.jp
b) E-mail: utsuro@iit.tsukuba.ac.jp
c) E-mail: a-kobayashi@a.tsukuba-tech.ac.jp
d) E-mail: hnishi@yamanashi.ac.jp
DOI: 10.1587/transinf.2022EDP7047

*The term "lexical features" is used in this paper are those disfluency-based and prosodic features described in Sect. 3.1.

comprehensively evaluating the effects of all of those acoustic, disfluency-based, and prosodic features. As the model for evaluating those features, we first employed multi-layer neural network that we have already studied in our previous works [11]–[13]. In our preliminary experiment, in addition to those acoustic, disfluency-based, and prosodic features, we evaluated features such as mel-frequency cepstrum coefficients (MFCC) and mel-spectrogram that have not been studied in our previous works [11]–[13], while MFCC and mel-spectrogram damaged the performance of speech fluency evaluation. We suppose that this is mainly due to the limitation of multi-layer neural network, which is not strengthful enough to handle the disfluency related information involved in those high dimensional features such as MFCC and mel-spectrogram. Considering this limitation of multi-layer neural network, we decide to utilize the end-to-end approaches in our task, which are not adopted in the related research. We implement the state-of-art end-to-end models with only mel-spectrogram. One of the advantages of the end-to-end models is that, even with a single feature of mel-spectrogram, their performance is close to, or even higher than that by multi-layer neural network (details are to be discussed with the experimental evaluation results). Another advantage is that they are quite compatible when running under a multi task setting with end-to-end speech recognition models. In such a multi task setting, speech recognition error detection and correction modules are easily called when disfluent speech is detected, where disfluent speech is supposed to cause speech recognition errors.

As the speech fluency evaluation task, we evaluate our proposed method in two binary classification tasks of fluent speech detection and disfluent speech detection. More specifically, 201 speech data from Corpus of Spontaneous Japanese (CSJ) [14], [15] with the annotation of the three classes of "fluent," "neutral," and "disfluent" is used as the dataset for evaluation. Each of those 201 speech data is further divided into its constituent files, each of which is of around 10 seconds duration, obtaining 2,169 files in total. Then, according to the two way splits of those three classes, the task of fluent speech detection is defined as binary classification of fluent (494 files) vs. neutral and disfluent (1,675 files), while that of disfluent speech detection is defined as binary classification of fluent and neutral (1,916 files) vs. disfluent (253 files).

We then conduct experiments with the purpose of comparative evaluation of multi-layer neural network with diverse features as well as end-to-end models. Conclusions of the experiments can be summarized as below. In the fluent speech detection, lexical features only are the most appropriate to multi-layer neural network. More specifically, the performance improved a lot when removing all of the acoustic features from the full set of features, while the performance is damaged a lot if fillers related features are removed. Overall, however, the best performance is achieved by end-to-end Transformer+VGGNet models with mel-spectrogram. In the disfluent speech detection, the Transformer+VGGNet architecture with mel-spectrogram also achieves high scores in the evaluation metrics. It is close to but does not exceed the results of multi-layer neural network with acoustic features and lexical features without fillers. Therefore, we find out that lexical features other than fillers are still necessary to some extent in the disfluent speech detection.

This paper is organized as follows. Section 2 introduces the dataset used in the experiments. Section 3 describes the utterance-level disfluency-based and prosodic features as well as acoustic features and corresponding functionals. Section 4 introduces some model architectures employed in this paper, which include multi-layer neural network and end-to-end models. Section 5 introduces the experiment setup and shows the analysis of evaluation results and conclusions are described in Sect. 6.

2. Dataset for Evaluation

Corpus of Spontaneous Japanese (CSJ) [14], [15] is a large-scale database that includes spontaneous speeches (lecture etc.) in Japanese. It contains speech signal and transcription of about 7 million words along with various annotations like POS and phonetic labels. Table 1 shows an example of the transcription of CSJ, which is a portion of the transcript of an example of the "disfluent" class to be described later in this section. In CSJ, as shown in Table 1, recorded speech is transcribed in two different ways: orthographic and phonetic transcriptions. In orthographic transcription, speech is transcribed using Kanji (Chinese logograph) and Kana (Japanese syllabary) just like ordinary Japanese text. In phonetic transcription, on the other hand, its transcription is written exclusively in Kana letters so that the phonetic details of the utterance being transcribed can be traced. More detailed transcription as well as the description of tags used for the annotation are found on the Web site of CSJ[†]. Various tags were embedded in these transcriptions to mark phenomena specific to spontaneous speech like fillers, word fragment, reduced articulation, mispronunciation, etc.

In this paper, among the transcription of CSJ, we utilize the following information in the evaluation: the duration as well as pause or silence information, the mora length of utterances measured in terms of character length of the phonetic transcription, and fillers as well as word fragments information (as shown in Table 1) as the most important disfluency-related information.

The sources of speech data of CSJ consist of 89 academic presentation speeches and 112 simulated public speeches. In CSJ, to each of those 201 speech data, rated impressions of public speaking such as *"liking," "skillfulness," "speech rate," "activity,"* and *"formality"* are annotated [15]. Among those rated impressions of public speaking, this paper utilizes that of 7-ranks rating of fluency-disfluency out of the *"skillfulness"* ratings. 10 annotators rated each speech data according to 7-ranks rating of fluency-disfluency, where we utilize their average over 10

[†]https://pj.ninjal.ac.jp/corpus_center/csj/misc/preliminary/5.html

Table 1 An example of the CSJ transcription (an example of the "disfluent" class).

ID	start ~ end (sec.)	orthographic transcription		phonetic transcription			⟨ filler⟩ ⟨word -fragments⟩ tags
		in Kanji/Kana letters	English translation	in Kana letters	in syllable representation	# morae	
—	268.338 ~ 268.869	⟨pause⟩					
0127	268.869	(F あのー)	uh	(F アノー)	a no o	3	⟨ filler⟩
	~ 271.138	何か	well	ナニカ	na ni ka	3	—
—	271.138 ~ 271.791	⟨pause⟩					
0128	271.791	(F ま)	so	(F マ)	ma	1	⟨ filler⟩
		郵便配達の	a mail delivery	ユービンハイタツノ	yu u bi N ha i ta tsu no	9	—
		人は	person	ヒトワ	hi to wa	3	—
		あれでしょうけど	might be probably	アレデショーケド	a re de sho o ke do	7	—
	~ 273.823	(D ん)	what?	(D ン)	N	1	⟨word -fragments⟩
—	273.823 ~ 274.268	⟨pause⟩					

Table 2 Statistics of the data set.

class	# speech data (# speakers)	# files (each around 10 sec. duration)	mean opinion score of fluency-disfluency rating (MOS, 7-ranks score averaged over 10 annotators)
fluent	54	494	5.0 ~ 6.2
neutral	109	1,422	3.2 ~ 4.9
disfluent	38	253	1.9 ~ 3.1
total	201	2,169	1.9 ~ 6.2

annotators. Then, as shown in Table 2, we classify the total 201 (speech and its transcription) data into the following three classes: "fluent" whose average ratings range from 5.0 to 6.2, "neutral" whose average ratings range from 3.2 to 4.9, and "disfluent" whose average ratings range from 1.9 to 3.1. Finally, we divide each speech and its transcription data into its constituent files, each of which is of around 10 seconds duration, obtaining 2,169 files in total (as shown in Table 2)[†]. In the experimental evaluation of this paper, we pursue the following two way splits of those three classes, i.e., (i) fluent (494 files) vs. neutral and disfluent (1,675 files), and (ii) fluent and neutral (1,916 files) vs. disfluent (253 files). Then, we evaluate our proposed method in two binary classification tasks of fluent speech detection and disfluent speech detection[††].

3. Features

We investigate multi-domain features including disfluency-based, prosodic, and acoustic features as well as their combinations. Disfluency-based and prosodic features are computed in utterance level, which are extracted through the statistic information of transcription. Acoustic features excluding mel-spectrogram are also extracted in utterance level, which are obtained by open source toolkit OpenSMILE[†††] through the frame-level features and statistical functionals. They are merged into the multi-layer neural network model. The summary of these features are shown in Table 4, where codes are given to the disfluency-based and prosodic features in Table 4(a) and Table 4(b) for indicating the results of feature ablation studies in Table 5 and Table 6. We also extract mel-spectrogram for end-to-end architectures. Although it belongs to acoustic features according to the strict definition, there exists clear distinction in this paper between the mel-spectrogram for end-to-end architectures and the acoustic features for the multi-layer neural network model, since the latter acoustic features are hardly utilized in the end-to-end architectures.

3.1 Disfluency-Based and Prosodic Features

Table 4(a) and Table 4(b) list disfluency-based and prosodic features used in this paper. The employed prosodic features, namely, are the speech rate, number of pauses per mora[††††], and the ratio of the contiguous silence to the duration of the

[†]More specifically, within each of the total 201 speech data, 7-ranks rating of fluency-disfluency is annotated to its constituent portion of around 50 seconds or more duration, but not to the whole speech duration. Thus, sometimes it can happen that one of the 201 speech data has both a "fluent" rated portion and a "neutral" rated portion, or both a "disfluent" rated portion and a "neutral" rated portion. In those cases, we remove those "neutral" rated portions and only keep "fluent" rated or "disfluent" rated portions.

[††]In our preliminary experiment, we first applied the regression model, while it did not fit to the task of fluency evaluation due to the imbalanced data between the disfluent speech data and the neutral speech data, and the predicted scores tend to be the mean score among all of the training data. Then, we formalize the task as the classification task, while the number of the training data is relatively small, and furthermore, the task is still has the problem of imbalanced data split. Considering this situation, we decided to evaluate our proposed method in two binary classification tasks of fluent speech detection and disfluent speech detection rather than a single task of classifying the three classes (fluent/neutral/disfluent).

[†††]http://www.audeering.com/opensmile/

[††††]The number of morae is measured as the Kana length of the phonetic transcription of the CSJ corpus.

Table 3 Feature values and mean opinion scores of fluency-disfluency rating: averages of fluent/neutral/disfluent classes. (total # (Ps), total # (Fs), total # (WF), total mora length (MrFs), and total mora length (MrWF) represent total numbers of pauses, fillers, word fragments, total mora lengths of fillers and word fragments, respectively.)

class	prosodic features				disfluency-based features								mean opinion score (MOS, 7-ranks score averaged over 10 anno-tators)
		pauses			fillers				word fragments				
	SpR	total # (Ps)	Ps /Mr	SilR	total # (Fs)	Filler1	total mora length (MrFs)	Filler2	total # (WF)	WF1	total mora length (MrWF)	WF2	
fluent	7.94	3.3	0.038	0.125	15.0	0.0294	27.9	0.0547	2.1	0.0041	3.1	0.0062	5.3
neutral	6.84	3.8	0.052	0.187	14.8	0.0334	28.8	0.0653	2.9	0.0067	4.5	0.0105	4.1
disfluent	5.53	4.4	0.075	0.285	15.9	0.0443	31.2	0.0883	4.1	0.0115	6.7	0.0187	2.8

speech. On the other hand, the disfluency-based features employed in this paper are obtained from the transcription of the CSJ. As shown in Table 1, the number and the mora length of fillers as well as word fragments are available in the phonetic transcription of CSJ, which we utilize as the disfluency-based features in this paper[†].

Table 3 shows average values of those seven features and the mean opinion scores in each of the fluent / neutral / disfluent classes.

3.2 Acoustic Features for the Multi-Layer Neural Network Model

We utilize the utterance-level feature vectors derived by the projection of frame-level acoustic features, such as pitch or energy by the descriptive statistical functionals [16]. In detail, the eight sorts of frame based low-level features chosen are: root mean square (RMS) from energy, zero crossing rate (ZCR), voicing probability (VP), fundamental frequency (F0), harmonics-to-noise ratio (HNR), jitter local, jitter difference of difference of periods (jitter ddp) and shimmer local. We choose the five most common higher-order statistics to process frame level features, which are *maximum*, *minimum*, *mean*, *variance* and *standard deviation*. If we consider all of them, the total feature vector per audio data contains $5 \times 8 = 40$ dimensional features, which are extracted by open source toolkit OpenSMILE. Those acoustic features and functionals used in our experiments are summarized in Table 4(c).

More specifically, utterance-level acoustic features employed in this paper are categorized into energy related and voicing related ones. Energy related features are RMS and ZCR, while the rest are voicing related features. These features are selected from the standardized feature set that were used for the INTERSPEECH 2011 Speaker State Challenge [17]. We do not use spectral LLDs features in multi-layer neural network model. This is mainly because, in our preliminary experiment, features such as mel-frequency cepstrum coefficients (MFCC) and mel-spectrogram have damaged the performance. One of our hypotheses of the reason for this is that high dimensional features such as MFCC and mel-spectrogram are not appropriate for the simple architecture like multi-layer neural network.

Table 4 Features summary

(a) Disfluency-based features

feature name	definition	code
fillers per mora,	ratio of total number of fillers to total number of morae,	Filler1,
mora length of fillers per mora	ratio of total mora length of fillers to total number of morae	Filler2
word fragments per mora,	ratio of total number of word fragments to total number of morae,	WF1,
mora length of word fragments per mora	ratio of total mora length of word fragments to total number of morae	WF2

(b) Prosodic features

feature name	definition	code
speech rate	average number of morae per sec.	SpR
pauses per mora,	ratio of total number of pauses to total number of morae,	Ps/Mr,
silence rate	ratio of contiguous silence to duration	SilR

(c) Acoustic features

acoustic features	functionals
RMS	maximum
ZCR	minimum
Voicing Probability	mean
F0	variance
HNR	standard variation
Jitter local, Jitter ddp	
Shimmer local	

[†]One of the major motivations to use the disfluency-based features obtained from manual transcriptions is to clarify whether or not these features are necessary to achieve the best performance for speech fluency evaluation. If it is confirmed that they are not necessary, that means that the best performance can be achieved fully automatically. Otherwise, further research effort should be inevitable so as to realize full automatic speech fluency evaluation by automatically extracting those disfluency-based features from the speech recognition results.

3.3 Mel-Spectrogram for End-to-End Models

The mel-spectrograms in frame level are compatible with deep convolution neural network model, however, due to the excellent feature extraction ability of convolution layer. We compute mel-spectrogram from speech recording for every 20 ms frame window shifted over 10 ms[†], using the short-time Fourier transform (STFT) with 64 frequency bins. Normalization is performed over frequency axis. Each utterance fed to the model is of size $T \times 64$, where T is the number of frames in a given speech.

4. Model Architectures

4.1 Multi-Layer Neural Network

We adopt multi-layer neural network architecture with two types of input to handle the disfluency-based and prosodic features as well as acoustic features. Figure 1 illustrates the neural network architecture of fluent/disfluent speech detection. In our experiments, the model consists of four fully connected hidden layers with 256 hidden units each and tanh activation are used for this experiment. We use the standard cross entropy objective function with L2 weight decay on parameters to prevent over fitting.

4.2 Deep Convolution Neural Network

Although features such as MFCC and mel-spectrogram do not perform well with the multi-layer neural network as we mentioned in Sect. 3.2, we are curious about its performance on end-to-end deep 1D convolution neural network. We conduct experiments with SpeakerNet [18] in this paper, which is based on the QuartzNet architecture comprising of an encoder and decoder structure. More specifically, the encoder consists of N blocks each with R sub-blocks, which is shown in Fig. 2. Each sub-block applies the following operations: a 1D convolution, batch norm, ReLU and dropout. The decoder consists of a statistic pooling layer and two feed forward layers. The statistic pooling layer computes the x-vector [19] using the encoder output. It computes the mean and standard deviation along the time axis, and it is necessary since we have to transfer the sequence-to-sequence model to the sequence-to-label model, which is similar to the statistical functionals we used for the acoustic features in Sect. 3.2. The two feed forward layers, on the other hand, execute the linear transformation. The model achieves the state-of-art result on speaker identification, which shows considerable strength on speech classification problem. Therefore, we adopt SpeakerNet in our disfluent/fluent speech detection task.

[†]This setting follows that examined in the speaker recognition task by SpeakerNet [18].

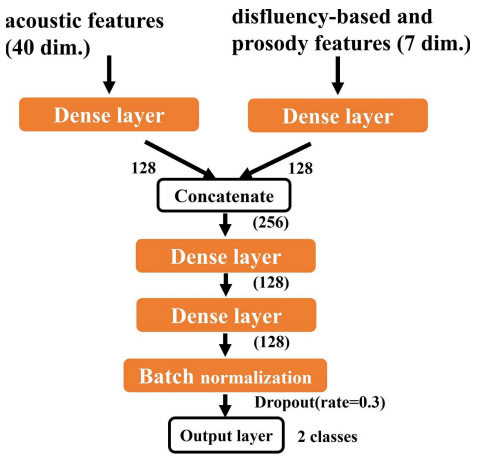


Fig. 1 Multi-layer neural network of fluent/disfluent speech detection

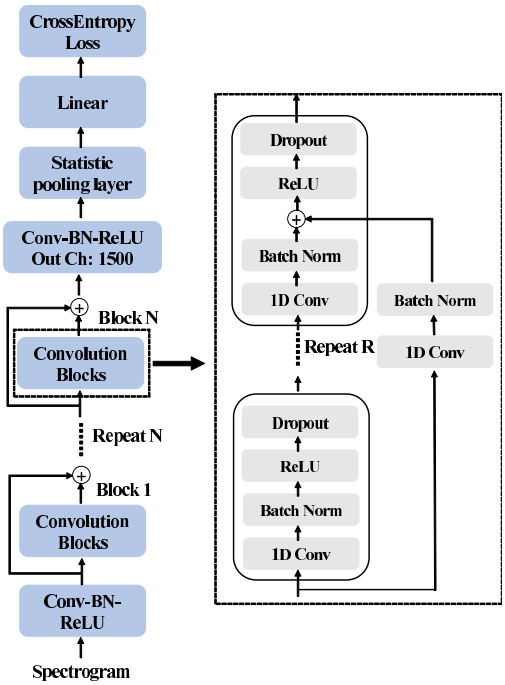


Fig. 2 SpeakerNet architecture of fluent/disfluent speech detection

4.3 Conformer and Transformer

In the recent research of Automatic Speech Recognition (ASR), Conformer outperforms the previous Transformer and convolution neural network (CNN) based models achieving state-of-art results [20]. Transformer models [21], [22] are good at capturing context interaction, while CNN extracts local features effectively. Conformer combines them to model both local and global dependencies of a speech sequence. Considering these advantages, we utilize the Conformer model in our task to investigate whether it could capture some important features which correspond to disfluency or fluency. The Conformer encoder consists of a convolution subsampling layer and a number of Conformer blocks as shown in Fig. 3, and the detail of its archi-

tectures is described in its original paper [20]. The decoder part is the same as QuartzNet, which involves x-vector extraction and linear transformation. Both the Conformer and the SpeakerNet architectures are trained with normal cross entropy loss. According to the results we show in the next section, we find that the Conformer architecture obtains better performance than SpeakerNet in both disfluent and fluent speech detection task. In order to further explore which part of the model is more effective, we conduct the experiment on the Transformer architecture, which is shown in Fig. 4. Specifically, following the Transformer's architecture [21], we modify the Conformer block shown in Fig. 3 through removing the convolution modules while keeping the multi-head attention module†. The subsampling layer is the same as that in the Conformer architecture.

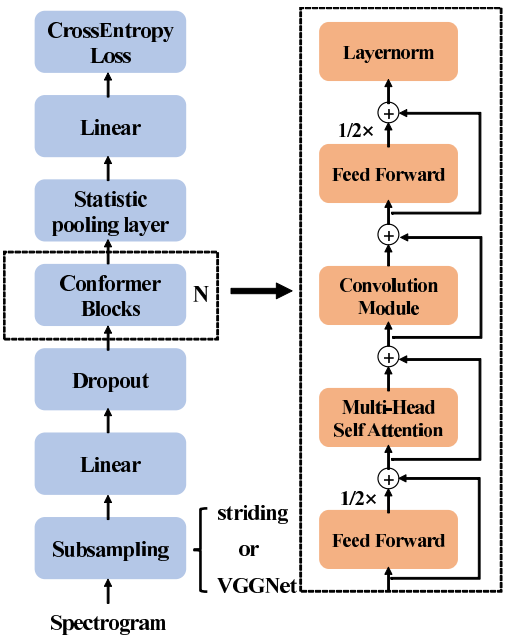


Fig. 3 Conformer architecture of fluent/disfluent speech detection [20]

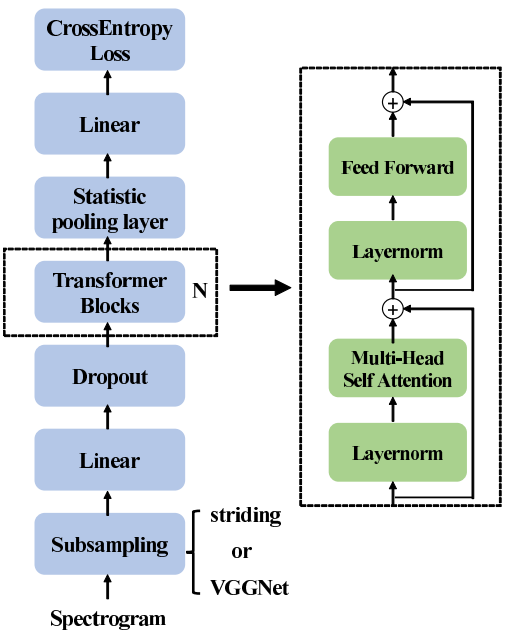


Fig. 4 Transformer architecture of fluent/disfluent speech detection [21]

5. Experiment

5.1 Settings

Throughout our experiment, the training procedure is performed through five-fold cross validation. In each of the five splits, 80% of the data set is further divided into the 80% training and the 20% development sets††, where the model with the number of epochs which minimizes the loss against the development set is evaluated against the test set (here, the maximum number of epochs is 300). The batch size is 32 for all the models. Other hyper parameters of the models examined in this paper are selected by consulting the performance against the development set in five-fold cross validation: initial learning rate as 0.0001 for the multi-layer neural network and 0.001 for other models, weight decay as 0.0001 for the multi-layer neural network and 0.001 for other models. All the models are trained with the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and a cosine annealing learning rate schedule [24].

In the multi-layer neural network architecture, the input of full features consists of seven dimensional disfluency-based and prosodic features, as well as eight dimensional acoustic features. In order to investigate the effect of diverse features, we conduct experiments through removing the single feature one after another.

In the SpeakerNet model, the inputs are mel-spectrograms which have been mentioned in Sect. 3.3. With respect to the encoder architecture, more specifically, we utilize QuartzNet architecture with the number of blocks $N = 2$ (optimized from the three candidates 2, 3, and 4, through five-fold cross validation on the development set) and the number of sub-blocks $R = 5$, and 512 channels.

For the Conformer and the Transformer architectures†††, the numbers of the blocks are set as $N = 8$ (optimized from the three candidates 4, 8, and 12, through five-fold cross validation on the development set) and the number of sub-blocks as $R = 1$. Following the implementation

†In the implementation of the Transformer block, we preliminarily examined existing architectures such as those found in the previous studies [21], [23] through tuning with five-fold cross validation on the development set of fluent/disfluent speech detection task of this paper. Then, we employ the architecture of the Transformer block studied in the previous work [23], which is shown in Fig. 4.

††Considering to make sure that the training, the development, and the test sets do not have overlap on speakers, we perform the data split with an open-speaker manner.

†††In the experiments of the Transformer model, we also tried the combination of mel-spectrogram and the acoustic features listed in the left half of Table 4(c) with their frame-levels, which have been mentioned in Sect. 3.2. However, it did not improve the results and mel-spectrogram only achieved better performance.

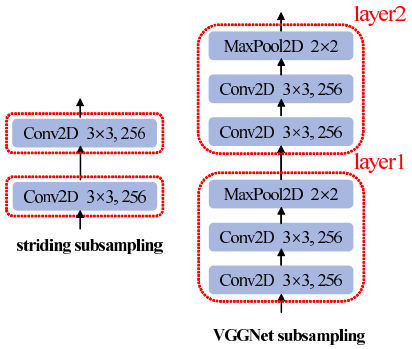


Fig. 5 Two types of convolution subsampling layer: striding and VGGNet. The subsampling factor is set to 4, so there are 2 subsampling layers in both striding and VGGNet approaches. In the striding approach, the convolution stride is set to 2, while in the VGGNet approach, the convolution stride is set to 1.

of the Conformer [20]†, as the types of convolution subsampling layers, we examined striding and VGGNet††. The difference between striding and VGGNet is shown in Fig. 5.

5.2 Results

The results are evaluated in the metrics of F-Score and Equal Error Rate (EER). When measuring F-Score, the lower bound of the probability of binary classification is selected so as to maximize the F-Score against the development set in five-fold cross validation. The experiments are conducted for disfluent speech detection and fluent speech detection respectively. For each sub-task, the results are summarized into two parts: comparing the effect of single features and investigating the effect of end-to-end approaches.

5.2.1 Fluent Speech Detection

The results of fluent speech detection with the multi-layer neural network are shown in Table 5. We can observe that the performance improved a lot when removing all of the acoustic features comparing to the full features. Thus, we regard the feature subset "remove acoustic features" as the best features in the fluent speech detection. It can be inferred that these acoustic features for the multi-layer neural network model, such as energy and voicing quality, are not compatible with our architecture. On the other hand, if Filler1+Filler2 are removed, the evaluation metrics are damaged a lot. It shows that filler related features are effective in this task, which is totally opposite to the results of the disfluent speech detection shown in the next section.

From Fig. 6, it is obvious that the Transformer+VGGNet architecture with mel-spectrogram perform better than the multi-layer neural network with the best features ("remove acoustic features"), which infers that deep end-to-end models with mel-spectrogram can classify the fluent speech and neutral/disfluent speech more effectively than lexical features and acoustic features for the multi-layer neural network model. In the comparison of end-to-end architectures, we can find that the Transformer architecture with VGGNet subsampling layer gives the best performance in both F-score and EER.

†https://github.com/NVIDIA/NeMo/blob/main/nemo/collections/asr/modules/conformer_encoder.py

††This implementation cited Dong et al. [23] for striding and Yeh et al. [25] for VGGNet [26]. In Yeh et al. [25], VGGNet [26] with causal convolution are adopted to incorporate contextual information into the Transformer networks.

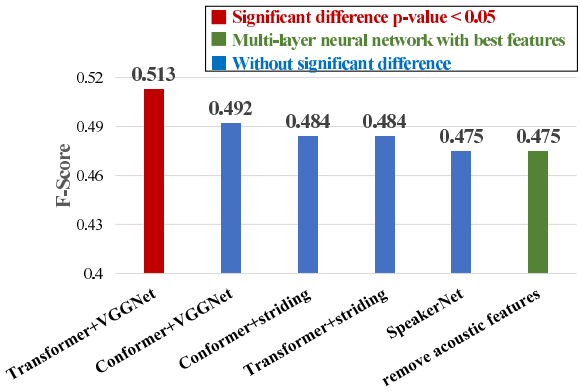
Table 5 The results of multi-layer neural network in fluent speech detection task. Each group corresponds to the set of remaining features after removing a single feature from the set of full features. **Bold faced and underlined group** means the features achieving the best results. † means that the result of the corresponding feature group has significant difference with p-value < 0.05 against **the best features** according to the t-test validation.

feature group	F-Score	EER
full features	0.447†	0.324†
remove acoustic features	**0.475**	**0.299**
remove Filler1+Filler2	0.429†	0.336†
remove WF1+WF2	0.443†	0.331†
remove SpR	0.440†	0.348†
remove Ps/Mr+SilR	0.437†	0.343†
remove RMS	0.436†	0.327†
remove ZCR	0.448†	0.313†
remove VP	0.453†	0.317†
remove F0	0.460†	0.329†
remove HNR	0.455†	0.330†
remove Jitter local + Jillter ddp	0.448†	0.321†
remove Shimmer local	0.445†	0.323†

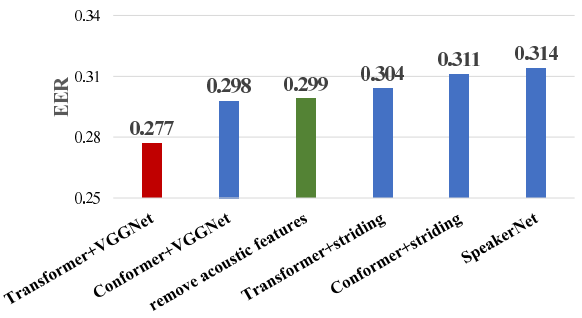
5.2.2 Disfluent Speech Detection

In Table 6, on the one hand, we can first observe that removing Filler1+Filler2 from the full features achieves the best results when considering both of F-Score and EER, and the significant difference against other feature subsets can also be explicitly verified. It shows that filler related features disrupt disfluent speech detection in someway. For example, from the results of carefully observing the transcription and the disfluency judgment annotated to the CSJ data, we can say that fillers occurring properly in a speech sentence will not affect the listeners' subjective impression. Although the F-Score of removing VP is the highest in Table 6, it has no significant difference against the full features in the metric of EER in Table 6. Therefore, we regard "remove Filler1+Filler2" as our best features in disfluent speech detection. On the other hand, we can notice that if we remove SpR or WF1+WF2, both F-Score and EER get worse than the results of full features. It means that speech rate and word fragments related features help to improve the performance of disfluent speech detection.

In Fig. 7, it can be noticed that in both F-Score and EER, the results of Transformer architecture with VGGNet subsampling layer do not exceed the multi-layer neural network with the best features ("remove Filler1+Filler2"), which shows that mel-spectrogram only are not enough to cover the disfluency related information comprehensively, even with the deep end-to-end models. Furthermore, in the comparison of end-to-end models, we can find that 1D con-



(a) F-Score



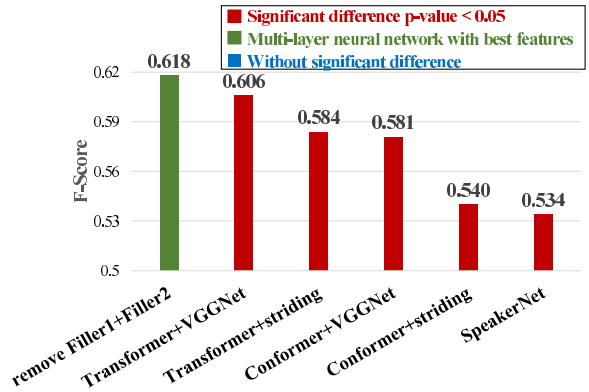
(b) EER

Fig. 6 The results of end-to-end models in fluent speech detection task. Each group corresponds to a model and the green bar refers to the result of the multi-layer neural network with the best features. The red bar means that the result of the corresponding model has significant difference against the multi-layer neural network with the best features according to the t-test validation.

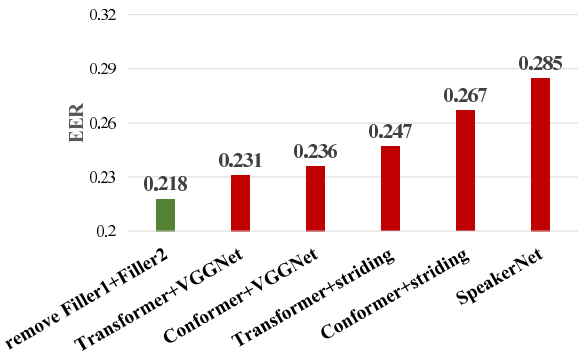
Table 6 The results of multi-layer neural network in disfluent speech detection task. Each group corresponds to the set of remaining features after removing a single feature from the set of full features. **Bold faced and underlined group** means the features achieving the best results (considering both of F-Score and EER). † means that the result of the corresponding feature group has significant difference with p-value < 0.05 against **the best features** according to the t-test validation.

feature group	F-Score	EER
full features	0.586†	0.241†
remove acoustic features	0.551†	0.254†
remove Filler1+Filler2	**0.618**	**0.218**
remove WF1+WF2	0.560†	0.292†
remove SpR	0.566†	0.257†
remove Ps/Mr+SilR	0.588†	0.242†
remove RMS	0.579†	0.256†
remove ZCR	0.597†	0.231†
remove VP	0.619	0.234†
remove F0	0.606†	0.233†
remove HNR	0.594†	0.226
remove Jitter local + Jillter ddp	0.577†	0.240†
remove Shimmer local	0.597†	0.241†

volution based SpeakerNet architecture performs worse than the multi-layer neural network in F-score and EER. Thus, it can be inferred that the local feature extracted by the convolution layer is not helpful to our task. And it can also ex-



(a) F-Score



(b) EER

Fig. 7 The results of end-to-end models in disfluent speech detection task. Each group corresponds to the result of multi-layer neural network with the best features. The red bar means that the result of the corresponding model has significant difference against the multi-layer neural network with the best features according to the t-test validation.

plain the results of the Transformer architecture performing better than the Conformer, which has the similar tendency with the fluent speech detection task. In other words, the contextual information detected by the multi-head attention is exactly what we need in the disfluent speech detection†. The results of the Transformer architecture performing better than the Conformer is quite opposite to the case of speech recognition evaluation results [20]. The major reason of this difference is that Conformer requires much more training data than Transformer, since the number of parameters of Conformer is much more than that of Transformer, where the size of the training data in our evaluation is relatively small. Furthermore, compared with speech recognition, the task of disfluent speech detection requires capturing global

†We can find certain evidence of this with x-vectors. The x-vector is obtained from the encoders, which compress the information of disfluency involved in the speech audio. With the powerful representation ability of the encoder, the disfluent speech's x-vectors could be clustered together in the space. According to the results of the rates of "inter-class covariance / intra-class covariance," it shows that the transformer block has better representation ability than the conformer block in that the former has higher rate than the latter.

dependencies of a speech sequence, while Conformer concentrates more on capturing local dependencies of a speech sequence with convolution neural network.

Finally, in both the Transformer and the Conformer architecture, we can observe that the VGGNet subsampling layer is explicitly more appropriate to the disfluent speech detection task. As the discussion about the effect of the VGGNet subsampling layer, we consider that subsampling means dropping some information of original data, while picking some important frames. Compared to the striding subsampling layer, the VGGNet subsampling layer consists of the combination of more convolution layers and max pooling layer, which will help selectively keep important frames, whose effect is what we exactly need for disfluent speech detection.

6. Conclusion

In this paper, we conducted experiments with the purpose of comparative evaluation of multi-layer neural network with diverse features as well as end-to-end models. We first utilize multi-layer neural network to investigate the effective disfluency-based and prosodic features as well as acoustic features for disfluent speech detection and fluent speech detection tasks respectively. We then evaluated the performance of end-to-end architectures with mel-spectrogram, in order to verify whether the lexical features are necessary in the fluency evaluation.

Conclusions of the experiments can be summarized as below. In the fluent speech detection, lexical features only are the most appropriate to multi-layer neural network. More specifically, the performance improved a lot when removing all of the acoustic features from the full set of features, while the performance is damaged a lot if fillers related features are removed. Overall, however, the best performance is achieved by end-to-end Transformer+VGGNet models with mel-spectrogram. Thus, it is quite important to note here that the best performance for the fluent speech detection is achieved without the manual transcription oriented features. This means that the best performance can be achieved fully automatically. In the disfluent speech detection, the Transformer+VGGNet architecture with mel-spectrogram also achieves high scores in the evaluation metrics. It is close to but does not exceed the results of multi-layer neural network with acoustic features and lexical features without fillers. Therefore, we find out that lexical features other than fillers are still necessary to some extent in the disfluent speech detection. Here, again, it is also quite important to note that it is necessary to use the manual transcription oriented features in order to achieve the best performance for the disfluent speech detection[†].

In the future work, in order to further verify the effect of end-to-end models, we will try to collect more speech data to extend the current dataset. Since there is no existing large corpus with disfluent and fluent labels, it might be necessary to explore some semi-supervised or self-supervised approaches such as wav2vec [28]. On the other hand, we have investigated some effective features such as filler, word fragment and speech rate. Therefore, combining them with end-to-end models through embedding layers is another interesting research direction. Especially, in order to overcome the limitation that the best performance for the disfluent speech detection requires the use of the manual transcription oriented features, further research direction should be fully automating the disfluent speech detection process by integration with end-to-end speech recognition models and extracting features for disfluent speech detection from the speech recognition results.

[†]Considering that the tasks such as disfluency detection within transcripts are performed only with text information within the transcript [27], it is quite reasonable that the best performance of our disfluent speech detection task is achieved with the manual transcription oriented features.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 21H00901.

References

[1] H. Nishizaki, M. Somiya, K. Kobayashi, and Y. Sekiguchi, "The effect of filled pauses in a lecture speech on impressive evaluation of listeners," Proc. 8th Interspeech, pp.2673–2676, 2007.
[2] K. Kobayashi, M. Somiya, H. Nishizaki, and Y. Sekiguchi, "Is a speech recognizer useful for characteristic analysis of classroom lecture speech?," Proc. 9th Interspeech, pp.1341–1344, 2008.
[3] E. Shriberg, Preliminaries to a Theory of Speech Disfluencies, Ph.D thesis, University of California, Berkeley, 1994.
[4] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional LSTM," Proc. 17th Interspeech, pp.2523–2527, 2016.
[5] N. Bach and F. Huang, "Noisy BiLSTM-based models for disfluency detection," Proc. 20th Interspeech, pp.4230–4234, 2019.
[6] R.C. van Dalen, K.M. Knil, and M.J.F. Gales, "Automatically grading learners' English using a gaussian process," Proc. SLaTE, pp.7–12, 2015.
[7] O. Deshmukh, K. Kandhway, A. Verma, and K. Audhkhasi, "Automatic evaluation of spoken English fluency," Proc. 34th ICASSP, pp.4829–4832, 2009.
[8] C.K. Lin, S.C. Tseng, and L.S. Lee, "Important and new features with analysis for disfluency interruption point (IP) detection in spontaneous Mandarin speech," Proc. 4th DiSS, pp.117–121, 2005.
[9] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," IEEE TASLP, vol.14, no.5, pp.1526–1540, Sept. 2006.
[10] V. Zayats and M. Ostendorf, "Giving attention to the unexpected: Using prosody innovations in disfluency detection," Proc. NAACL-HLT, pp.86–95, June 2019.
[11] H. Deng, Y. Lin, T. Utsuro, A. Kobayashi, H. Nishizaki, and J. Hoshino, "Automatic fluency evaluation of spontaneous speech using disfluency-based features," Proc. 45th ICASSP, pp.9239–9243, 2020.
[12] H. Deng, Y. Lin, T. Utsuro, A. Kobayashi, H. Nishizaki, and J. Hoshino, "Integrating disfluency-based and prosodic features with acoustics in automatic fluency evaluation of spontaneous speech," Proc. 12th LREC, pp.6431–6439, May 2020.
[13] H. Deng, T. Utsuro, A. Kobayashi, and H. Nishizaki, "Comparison of static and time-sequential features in automatic fluency detec-

tion of spontaneous speech," Proc. 24th O-COCOSDA, pp.158–163, 2021.
[14] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," Proc. SSPR, pp.7–12, 2003.
[15] T. Kagomiya, K. Yamasumi, Y. Maki, and K. Maekawa, "Development and analysis of a psychological evaluating database of public speaking," Jpn. J. Lang. Soc., vol.9, no.2, pp.65–76, 2007. (in Japanese).
[16] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," Proc. 10th Interspeech, pp.312–315, 2009.
[17] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," Proc. 12th Interspeech, pp.3201–3204, 2011.
[18] N.R. Koluguri, J. Li, V. Lavrukhin, and B. Ginsburg, "Speakernet: 1D depth-wise separable convolutional network for text-independent speaker recognition and verification," arXiv:2010.12653v1, 2020.
[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," Proc. 43rd ICASSP, pp.5329–5333, 2018.
[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for speech recognition," Proc. 21st Interspeech, pp.5036–5040, 2020.
[21] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer Transducer: A streamable speech recognition model with Transformer encoders and RNN-T loss," Proc. ICASSP, pp.7829–7833, 2020.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Proc. 30th NIPS, pp.5998–6008, 2017.
[23] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," Proc. 43rd ICASSP, pp.5884–5888, 2018.
[24] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," Proc. ICLR, pp.1–16, 2017.
[25] C.F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M.L. Seltzer, "Transformer-Transducer: End-to-end speech recognition with self-attention," CoRR, vol.abs/1910.12977, 2019.
[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Proc. 3rd ICLR, pp.1–14, 2015.
[27] Q. Chen, M. Chen, B. Li, and W. Wang, "Controllable time-delay Transformer for real-time punctuation prediction and disfluency detection," Proc. 45th ICASSP, pp.8069–8073, 2020.
[28] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," Proc. 34th NeurIPS, pp.12449–12460, Dec. 2020.

Huaijin Deng received his B.E. degree in engineering from Northwestern Polytechnical University, China in 2018 and M.E. degree in engineering from University of Tsukuba, Japan in 2022. After graduating from the master's program in Intelligent and Mechanical Interaction Systems, Degree Programs in Systems and Information Engineering, Graduate School of Science and Technology, University of Tsukuba in 2022, he has been a member of Baidu, Inc.. His main research topics are speech processing and spontaneous speech fluency evaluation.

Takehito Utsuro received his B.E., M.E., and Ph.D. Eng. degrees in electrical engineering from Kyoto University in 1989, 1991, and 1994, respectively. He has been a professor at the Division of Intelligent Interaction Technologies, Faculty of Engineering, Information and Systems, University of Tsukuba, since 2012. His professional interests in natural language processing, Web intelligence, information retrieval, machine learning, and spoken language processing. He is also a senior member of the Institute of Electronics, Information and Communication Engineers (IEICE), and a member of Information Processing Society of Japan (IPSJ), and the Acoustic Society of Japan (ASJ).

Akio Kobayashi received his B.E. degree from Waseda University in 1991 and Ph.D. from Toyohashi University of Technology in engineering in 2012. He is an associate professor of Tsukuba University of Technology after his carrier as a research engineer at NHK Science and Technology Research Laboratories. His current research topics are spoken language processing and information accessibility technologies for challenged people. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustic Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), and IEEE.

Hiromitsu Nishizaki received his B.E., M.E., and Ph.D. degrees in engineering from Toyohashi University of Technology, Japan in 1998, 2020, and 2003, respectively. From August 2015 to March 2016, he was a visiting researcher at the National Taiwan University in the Republic of China. He has been a professor at Graduate Faculty of Interdisciplinary Research, University of Yamahashi, since 2022. His research interests include spoken language processing and image processing using deep learning. He is also a senior member of the Institute of Electronics, Information and Communication Engineers (IEICE) and IEEE.