Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6429–6437
Marseille, 11–16 May 2020

Integrating Disfluency-based and Prosodic Features with Acoustics in Automatic Fluency Evaluation of Spontaneous Speech

Huaijin Deng[1], Youchao Lin[1], Takehito Utsuro[1], Akio Kobayashi[2], Hiromitsu Nishizaki[3], Junichi Hoshino[1]
[1]Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan
[2]Department of Industrial Information, Tsukuba University of Technology, Tsukuba 305-8520, Japan
[3]Graduate School of Interdisciplinary Research, University of Yamanashi, Kofu 400-8510, Japan
{s2020815, s1730186}_@_s.tsukuba.ac.jp, {utsuro, jhoshino}_@_iit.tsukuba.ac.jp
a-kobayashi_@_a.tsukuba-tech.ac.jp, hnishi_@_yamanashi.ac.jp

Abstract

This paper describes an automatic fluency evaluation of spontaneous speech. In the task of automatic fluency evaluation, we integrate diverse features of acoustics, prosody, and disfluency-based ones. Then, we attempt to reveal the contribution of each of those diverse features to the task of automatic fluency evaluation. Although a variety of different disfluencies are observed regularly in spontaneous speech, we focus on two types of phenomena, i.e., filled pauses and word fragments. The experimental results demonstrate that the disfluency-based features derived from word fragments and filled pauses are effective relative to evaluating fluent/disfluent speech, especially when combined with prosodic features, e.g., such as speech rate and pauses/silence. Next, we employed an LSTM based framework in order to integrate the disfluency-based and prosodic features with time sequential acoustic features. The experimental evaluation results of those integrated diverse features indicate that time sequential acoustic features contribute to improving the model with disfluency-based and prosodic features when detecting *fluent* speech, but not when detecting *disfluent* speech. Furthermore, when detecting *disfluent* speech, the model *without* time sequential acoustic features performs best even without word fragments features, but only with filled pauses and prosodic features.

Keywords: automatic fluency evaluation, disfluency, filled pauses, prosody, acoustics, LSTM, spontaneous speech

1. Introduction

Presentation ability has become important in various scenarios. For example, relative to a university lecturer's talk, speaking skill affects student understanding (Nishizaki et al., 2007; Kobayashi et al., 2008), and the evaluation of speech fluency of second language learners is important relative to measuring proficiency in second language learning (Mao et al., 2019). Therefore, an automatic speech fluency evaluation method is required. To date, several studies examined the evaluation of speech fluency; however, most of these studies focused on second language learning (Mao et al., 2019; Fontan et al., 2018; van Dalen et al., 2015).

Among the numerous issues related to automatic fluency evaluation, this paper focuses on disfluency (Mahesha and Vinod, 2012; Kjellgren and Nordstrom, 2016; Deshmukh et al., 2009; Mao et al., 2019; Tohyama and Matsubara, 2006; Rasipuram et al., 2016). Although we regularly observe a variety of different disfluencies in spontaneous speech, we focus on two types of phenomena, i.e., filled pauses and word fragments. With respect to filled pauses and word fragments, previous studies into automatic fluency/disfluency evaluation (Mahesha and Vinod, 2012; Kjellgren and Nordstrom, 2016; Deshmukh et al., 2009; Mao et al., 2019; Tohyama and Matsubara, 2006; Rasipuram et al., 2016) have been limited. Mahesha and Vinod (2012) and Kjellgren and Nordstrom (2016) did not consider those two issues. van Dalen et al. (2015) focused on disfluency features; however, they did not investigate the effectiveness of each disfluency feature type on fluency evaluation. Deshmukh et al. (2009) used features on filled pause and word repetition for automatic speech fluency evaluation; however, they did not consider a word fragment-related feature. In addition, it was insufficient in that it is not clear how much contribution those disfluency-related features actually have with respect to fluency evaluation. Overall, the limitation of these previous studies is primarily the lack of consideration of various word fragment phenomena.

Considering the limitation of previous automatic fluency evaluation approaches, this paper explores yet another finding on automatic evaluation of fluency and disfluency. In the task of automatic fluency evaluation, we integrate diverse features of acoustics, prosody, and disfluency-based ones. Then, we attempt to reveal the contribution of each of those diverse features to the task of automatic fluency evaluation. To this end, we first conducted a series of SVM classification experiments on a Japanese spontaneous speech corpus. The experimental results demonstrate that the disfluency-based features derived from word fragments and filled pauses are effective relative to evaluating fluent/disfluent speech, especially when combined with prosodic features, e.g., such as speech rate and pauses/silence. Next, we employed an LSTM (Long Short-Term Memory) based framework in order to integrate the disfluency-based and prosodic features with time sequential acoustic features, e.g., root mean square (RMS) frame energy, zero-crossing-rate (ZCR) from the time signal, fundamental frequency (F0), and voicing probability by autocorrelation function. The experimental evaluation results of those integrated diverse features indicate that time sequential acoustic features contribute to improving the model with disfluency-based and prosodic features when

ID	start ∼ end (sec.)	orthographic transcription		phonetic transcription			⟨filled -pause⟩ ⟨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	⟨filled -pause⟩
	∼ 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	⟨filled -pause⟩
		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 1: Example of Corpus of Spontaneous Speech (CSJ) transcription (example of "disfluent" class)

detecting *fluent* speech, but not when detecting *disfluent* speech. Furthermore, when detecting *disfluent* speech, the model *without* time sequential acoustic features performs best even without word fragments features, but only with filled pauses and prosodic features.

2. Dataset for Evaluation

The Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003; Kagomiya et al., 2007) is a large dataset that includes spontaneous speeches (lectures, etc.) in Japanese. It contains the speech signals and transcriptions of approximately seven million words with various annotations, e.g, POS and phonetic labels. Table 1 shows an example of the transcription of CSJ, i.e., part of the transcript of an example of the "disfluent" class, which is described later in this section. In CSJ (Table 1), recorded speech is transcribed as orthographic and phonetic transcriptions[1]. In an orthographic transcription, speech is transcribed using Kanji (Chinese logograph) and Kana (Japanese syllabary) just like ordinary Japanese text. In contrast, a phonetic transcription is written exclusively in Kana letters such that the phonetic details of the transcribed utterance can be traced. More detailed transcriptions and descriptions of the tags used for annotation can be found at CSJ website of CSJ[2]. Various tags are embedded in these transcriptions to mark phenomena specific to spontaneous speech, e.g., filled pauses, word fragment, reduced articulation, and mispronunciation.

In this paper, among the CSJ transcriptions, we utilize the following information in our evaluation: duration and pause or silence information, mora length of utterances measured in terms of character length of the phonetic transcription, and filled pauses and word fragments information (Table 1). We consider these as representing the most important disfluency-related information.

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

Table 2: Statistics of the dataset

The sources of speech data in CSJ comprise 89 academic presentation speeches and 112 simulated public speeches. In CSJ, rated impressions of public speaking, e.g., such as *"liking," "skillfulness," "speech rate," "activity,"* and *"formality,"* are annotated to each of those 201 speech data (Kagomiya et al., 2007). Among the rated impressions, we utilized that of a seven-rank rating of fluency-disfluency, which is one of the *"skillfulness"* ratings. Ten annotators rated each speech data according to a seven-rank rating of fluency-disfluency, where we utilized the average over the 10 annotators. As shown in Table 2, we then classified the 201 (speech and its transcription) data into the following three classes: "fluent" (average ratings of 5.0 to 6.2), "neutral" (average ratings of 3.2 to 4.9), and "disfluent" (average ratings of 1.9 to 3.1). Finally, we divided each speech and its transcription data into its constituent files, each of which having a duration of approximately 10 s. We obtained 2,169

[1] Each line of a transcript of the CSJ corresponds to a Japanese *bunsetsu*, consisting of one or more content words such as nouns and verbs, followed by zero or more functional words such as particles, auxiliary verbs, and suffixes. In a transcript, several lines of bunsetsus are then delimited by a line representing a *pause*, where, in CSJ, duration of silence over 200 ms is considered as a pause.

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

class	prosodic features				disfluency-based features								mean opinion score (MOS, 7-ranks score averaged over 10 annotators)
		pauses			filled pauses				word fragments				
	SpR	total # (Ps)	Ps /Mr	SilR	total # (FP)	FP /Mr	total mora length (MrFP)	Mr FP /Mr	total # (WF)	WF /Mr	total mora length (MrWF)	Mr WF /Mr	
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

Table 4: Feature values and mean opinion scores of fluency-disfluency rating for prosodic and disfluency-based features (averages of fluent/neutral/disfluent classes)

(a) 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

(b) Disfluency-based features

feature name	definition	code
# filled pauses per mora	ratio of total number of filled pauses to total number of morae	FP /Mr
# word fragments per mora	ratio of total number of word fragments to total number of morae	WF /Mr
mora length of filled pauses per mora	ratio of total mora length of filled pauses to total number of morae	MrFP /Mr
mora length of word fragments per mora	ratio of total mora length of word fragments to total number of morae	MrWF /Mr

Table 3: Prosodic and disfluency-based features

files in total (Table 2)[3]. In our experimental evaluation, we examined the following two-way splits of these classes: (1) fluent (494 files) vs. neutral and disfluent (1,675 files), and (2) fluent and neutral (1,916 files) vs. disfluent (253 files). Then, we evaluated the method we propose in this paper in two binary classification tasks, i.e., of fluent speech and disfluent speech detection tasks.

3. Prosodic and Disfluency-based Features

Table 3 lists the prosodic and disfluency-based features employed in our evaluation. Table 2(a) lists three prosodic features, i.e., speech rate, number of pauses per mora[4], and the ratio of contiguous silence to speech duration. Table 2(b) lists the disfluency-based features, which were obtained from the CSJ transcription. As shown in Table 1, the number and mora length of filled pauses and word fragments are available in the CSJ phonetic transcription, which we utilized as disfluency-based features. Table 4 shows the average values of those seven sorts of features and the mean opinion scores for the fluent, neutral, and disfluent classes. As can be seen from the distribution of their averages, for each of those seven prosodic and disfluency-based features, its averages over fluent, neutral, and disfluent classes have certain consistency and seem to be useful in discriminating those three classes.

4. Evaluating Prosodic and Disfluency-based Features through Automatic Fluency Evaluation by SVM

We first evaluated the prosodic and disfluency-based features introduced in the previous section through automatic fluency evaluation with five-fold cross validation by SVM. We employed the SVM toolkit from the scikit-learn (Pedregosa et al., 2011) package for automatic classification in fluency evaluation. In each data split of the five-fold cross validation, we performed the grid-search[5] of the kernel functions (RBF, linear, and second degree polynomial) and hyper parameters C and γ, and we evaluate them against the test set.

Table 5 shows the results obtained by running SVM with each single feature, where the kernel functions and hyper-parameters were optimized through grid search by maximizing recall or f-measure. Note that Table 5 (a) and Table 5 (b) rank the individual features in descending order of optimized recall or f-measure. Here, one of the most important findings is that, roughly speaking, the prosodic and word fragment features outperformed filled pauses features in both binary classification tasks. In addition, it is obvious that those higher ranked features performed better when optimizing the *f-measure* of detecting *disfluent* speech than when detecting *fluent* speech.

[3] More specifically, within each of the total 201 speech data, seven-rank rating of fluency-disfluency is annotated to its constituent part 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 part and a "neutral" rated part, or both a "disfluent" rated part and a "neutral" rated part. In those cases, we remove those "neutral" rated parts and only keep "fluent" rated or "disfluent" rated parts.

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

[5] In both of the binary classifications of fluent speech detection and disfluent speech detection, we used the three types of target optimization functions: i.e., recall, precision, and f-measure of the target class to be detected.

(a) Binary classification of fluent vs. neutral-disfluent classes

rank	optimizing recall		optimizing f-measure	
1	SilR	75.4	SpR	43.3
2	SpR	74.3	Ps/Mr	39.5
3	MrWF/Mr	67.4	MrWF/Mr	39.3
4	WF/Mr	65.4	SilR	37.9
5	MrFP/Mr	65.3	WF/Mr	35.8
6	FP/Mr	62.4	MrFP/Mr	35.1
7	Ps/Mr	53.8	FP/Mr	34.1

(b) Binary classification of fluent-neutral vs. disfluent classes

rank	optimizing recall		optimizing f-measure	
1	SpR	84.2	MrWF/Mr	50.3
2	MrWF/Mr	69.3	SpR	50.1
3	SilR	68.7	SilR	48.3
4	Ps/Mr	66.1	WF/Mr	47.5
5	WF/Mr	58.9	Ps/Mr	46.6
6	MrFP/Mr	43.1	FP/Mr	32.2
7	FP/Mr	34.6	MrFP/Mr	28.6

Table 5: Results of running SVM with a single prosodic / disfluency-based feature by optimization through grid search (individual features are ranked in descending order of optimized recall/f-measure (%))

Table 6 further examines the experimental results of feature combinations selected from a list of exhaustive combinations of all the seven features. For each of feature combinations, again, the kernel functions and hyperparameters were optimized through grid search by maximizing recall or f-measure. We compared feature combinations that satisfy one of the following requirements:

(1) achieving highest recall/f-measure,

(2) achieving highest recall/f-measure with feature combinations other than four disfluency-based features,

(3) achieving highest recall/f-measure with feature combinations other than two filled pauses features,

(4) achieving highest recall/f-measure with feature combinations other than two word fragments features.

Overall, it is obvious that, in this feature evaluation by SVM, we achieved higher recall/f-measure when detecting *disfluent* speech compared to detecting *fluent* speech. This result indicates that, when detecting fluent/disfluent speech, both disfluency-based and prosodic features are generally inevitable. Furthermore, it is clear from the results shown in Table 6 (b) that the differences in recall/f-measure values in Table 6 (b) are much greater compared to the differences shown in Table 6 (a).

5. Integrating Disfluency-based and Prosodic Features with Acoustics through LSTM

Next, we employed an LSTM based framework in order to integrate the disfluency-based and prosodic features with time sequential acoustic features.

(a) Binary classification of fluent vs. neutral-disfluent classes (recall/precision/f-measure of detecting fluent class)

features		(macro ave.)
baseline (minority)		100 / 22.8 / 37.1
optimizing recall	SpR + SilR + WF/Mr + MrFP/Mr (highest recall)	**82.2** / 35.7 / 48.1
	SilR + MrFP/Mr (highest recall w/o word fragments)	81.5 / 28.9 / 42.1
	SpR + MrWF/Mr (highest recall w/o filled pauses)	80.5 / 33.6 / 45.8
	SpR + Ps/Mr + SilR (highest recall w/o disfluency-based features)	71.2 / 32.7 / 43.2
optimizing f-measure	SpR + Ps/Mr + SilR + FP/Mr + MrWF/Mr (highest f-measure)	67.1 / 39.1 / **48.7**
	SpR + Ps/Mr + SilR + FP/Mr (highest f-measure w/o word fragments)	69.7 / 37.6 / 48.0
	SpR + Ps/Mr + SilR + WF/Mr (highest f-measure w/o filled pauses)	67.0 / 38.5 / 47.5
	SpR + SilR (highest f-measure w/o disfluency-based features)	65.1 / 34.4 / 44.0

(b) Binary classification of fluent-neutral vs. disfluent classes (recall/precision/f-measure of detecting disfluent class)

features		(macro ave.)
baseline (minority)		100 / 11.7 / 20.9
optimizing recall	6 features (w/o MrFP/Mr) (highest recall)	**93.0** / 32.1 / 47.4
	SilR + MrWF/Mr (highest recall w/o filled pauses)	90.4 / 34.0 / 49.3
	SpR + FP/MR (highest recall w/o word fragments)	84.7 / 34.9 / 48.8
	SpR + SilR (highest recall w/o disfluency-based features)	84.5 / 34.2 / 47.7
optimizing f-measure	SpR + SilR + WF/Mr (highest f-measure = highest f-measure w/o filled pauses)	73.8 / 50.9 / **58.4**
	SpR + Ps/Mr (highest f-measure w/o disfluency-based features)	67.8 / 42.2 / 50.6
	SpR + FP/Mr (highest f-measure w/o word fragments)	77.6 / 36.4 / 48.9

Table 6: Experimental results of feature combination for prosodic and disfluency-based features by SVM (%)

5.1. LSTM Framework

The employed LSTM architecture is shown in Figure 1, which is implemented through the TensorFlow framework. Its input vector consists of the time sequential acoustic features and the static disfluency-based / prosodic seven features. The time sequential acoustic features employed in this paper are fundamental ones, namely, root mean square (RMS) frame energy, zero-crossing-rate (ZCR) from the time signal, fundamental frequency (F0), and voicing probability by autocorrelation function. They are selected from the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009) feature set, which are available through the OpenSMILE[6] toolkit. Other than those four fundamental features, we exclude mel-frequency cepstral coefficients (MFCC) from the INTERSPEECH 2009 Emotion Challenge feature set for the sake of simplicity.

[6] https://www.audeering.com/opensmile/

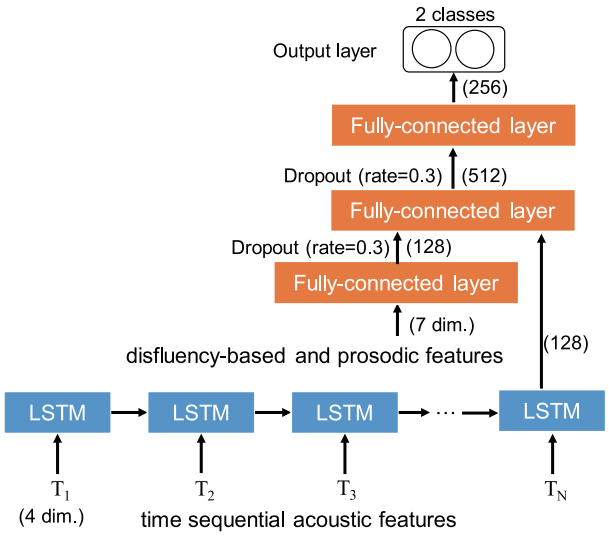


Figure 1: Neural network architecture for the LSTM framework

mini-batch size	32
num. of epochs	300
activation for the hidden layers	ReLU / Leaky ReLU
dropout	0.3
batch normalization	Yes
loss func.	binary cross entropy
optimizer	Adam
initial learning rate	0.00001 / 0.000001

Table 7: Training conditions of the LSTM model

The employed LSTM architecture is a neural network composed of four hidden layers:

(1) an LSTM layer for the time sequential acoustic features,
(2) one fully-connected layer for disfluency-based and prosodic features,
(3) two fully-connected layers after concatenating the two sorts of features (1) and (2).

Its output layer has two class nodes. The training condition of the LSTM is described in Table 7. 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 training and the development datasets, where the model with the number of epochs which minimizes the loss against the development dataset is evaluated against the test dataset. The hyper parameters of Table 7 are selected also by consulting the performance against each development dataset in five fold cross validation.

5.2. Experiments

In the evaluation of the LSTM framework of this paper, we compare models *with* and *without* time sequential acoustic features as presented in Table 8. Their training conditions and disfluency-based and prosodic features are heuristically selected through the evaluation against the development datasets. For the models *without* time sequential acoustic features, we compare the following seven models:

(a) seven features,
(b) disfluency-based features (four features),
(c) filled pauses and prosodic features (five features),
(d) word fragments and prosodic features (five features),
(e) filled pauses features (two features),
(f) word fragments features (two features), and
(g) prosodic features (three features).

In terms of precision and recall calculation, a trained LSTM predicts confidence $conf$ by the softmax function for given inputs and requires a confidence threshold c for deterministic prediction. Inputs whose predicted confidence $conf$ are above the threshold are positive predictions, constituting set $S(conf \geq C)$. Supposing that the set R is the reference set, precision and recall is formally defined as below.

$$\text{Recall}(conf \geq c) = \frac{|R \cap S(conf \geq c)|}{|R|}$$

$$\text{Precision}(conf \geq c) = \frac{|R \cap S(conf \geq c)|}{|S(conf \geq c)|}$$

Figure 2 $\sim$ Figure 4 present evaluation results by plotting recall-precision curves for the models to be compared. Each curve represents all the precision/recall pairs as confidence threshold c varies from 1 to 0. Figure 2 (a), Figure 3 (a), and Figure 4 (a) presents the evaluation results of binary classification of fluent vs. neutral-disfluent classes, where the fluent class is regarded as positive and its recall-precision curves are plotted. Figure 2 (b), Figure 3 (b), and Figure 4 (b), on the other hand, presents the evaluation results of binary classification of fluent-neutral vs. disfluent classes, where the disfluent class is regarded as positive and its recall-precision curves are plotted.

It is obvious by comparing the evaluation results of Figure 2 (a) and Figure 2 (b) that, when detecting *fluent* speech (Figure 2 (a)), time sequential acoustic features contribute to improving the models *without* time sequential acoustic features, while they do not contribute when detecting *disfluent* speech (Figure 2 (b)). Furthermore, when detecting *disfluent* speech (Figure 2 (b)), it is quite interesting to note that, the model *without* time sequential acoustic features performs best even without word fragments features, but only with filled pauses and prosodic features.

Next, in the comparison of disfluency-based features, word fragments + prosodic features, and filled pauses + prosodic features, all of which are *without* time sequential acoustic features (Figure 3), it is also quite interesting to note that the models without prosodic features perform worst. This finding indicates that the prosodic features contribute more to detecting fluent/disfluent speech than the disfluency-based features. Finally, in the comparison of filled pauses features, word fragments features, and prosodic features, all of which are *without* time sequential acoustic features (Figure 4), those small number of features alone do not perform well in detecting fluent/disfluent speech.

From those evaluation results, we conclude the following: (1) when detecting *fluent* speech, the best performing model is required that all of disfluency-based, prosodic, and time

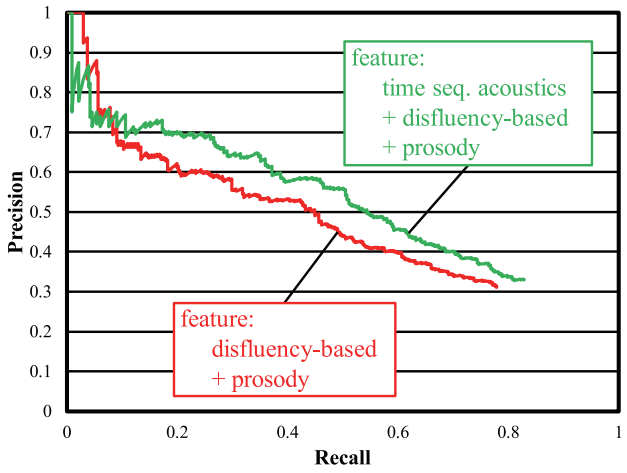
(a) Binary classification of fluent vs. neutral-disfluent classes

	training conditions	disfluency-based / prosodic features
w/ time seq. acoustic features	initial learning rate: 0.000001, activation for the hidden layers: Leaky ReLU	7 features
w/o time seq. acoustic features		7 models (7 combinations of filled pauses / word fragments / prosody features)

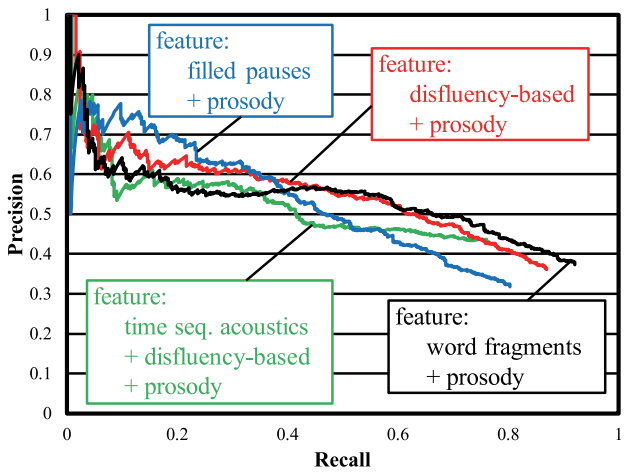
(b) Binary classification of fluent-neutral vs. disfluent classes

	training conditions	disfluency-based / prosodic features
w/ time seq. acoustic features	initial learning rate: 0.00001, activation for the hidden layers: ReLU	SpR + Ps/Mr + SilR + MrWF/Mr
w/o time seq. acoustic features	initial learning rate: 0.000001, activation for the hidden layers: Leaky ReLU	7 models (7 combinations of filled pauses / word fragments / prosody features)

Table 8: Comparison of w/ and w/o time seq. acoustic features: training conditions and disfluency-based / prosodic features

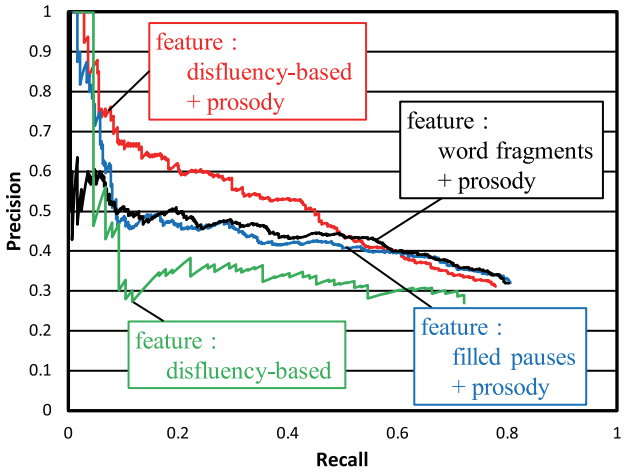


(a) Binary classification of fluent vs. neutral-disfluent classes

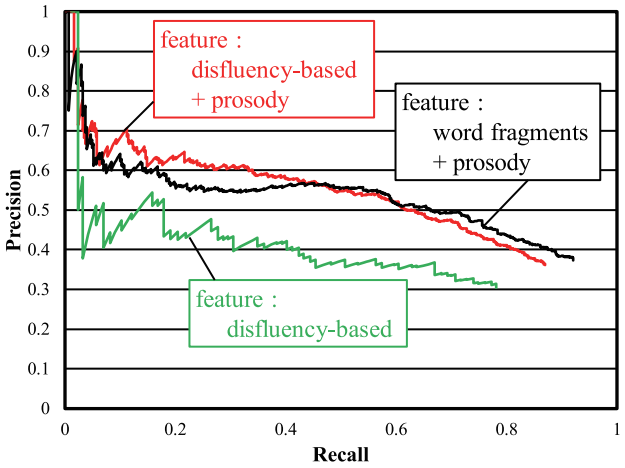


(b) Binary classification of fluent-neutral vs. disfluent classes

Figure 2: Experimental results of the LSTM framework (1): Comparison of w/ and w/o time seq. acoustic features



(a) Binary classification of fluent vs. neutral-disfluent classes



(b) Binary classification of fluent-neutral vs. disfluent classes

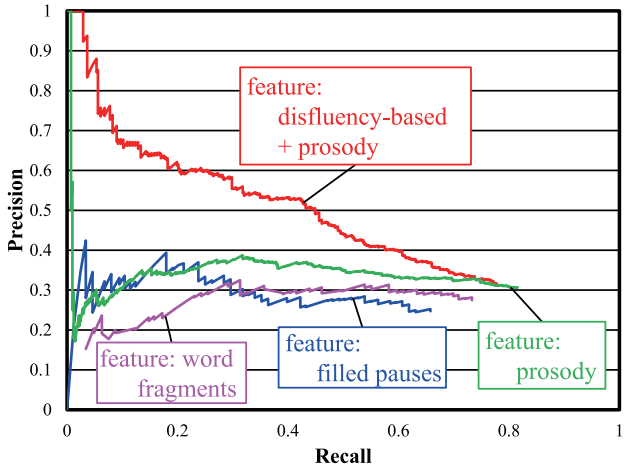
Figure 3: Experimental results of the LSTM framework (2): Comparison of disfluency-based / word fragments + prosody (/ filled pauses + prosody) features

sequential acoustic features are incorporated into the model and contribute to detecting clues to *fluent* speech, (2) when detecting *disfluent* speech, on the other hand, prosodic features are inevitable, and filled pauses features contribute most to improving the recall-precision curve even without word fragments and time sequential acoustic features.

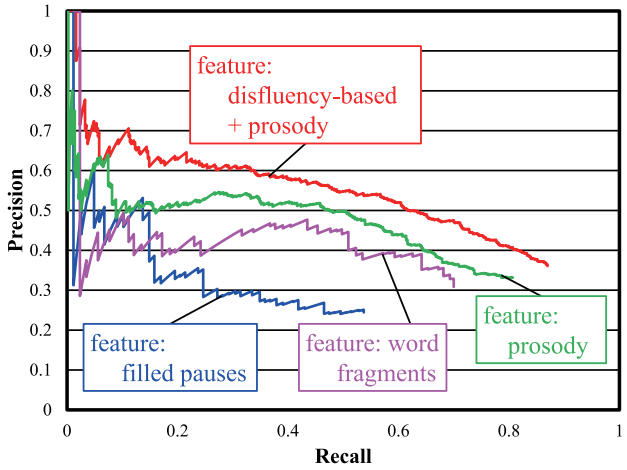
In addition to the two conclusions above, we further evaluate the models *with* time sequential acoustic features, where either difluency-based or prosodic features are discarded. When detecting *fluent* speech, we confirmed that their recall-precision curves perform worse than when all the seven features are incorporated. This result indicates that both difluency-based and prosodic features are inevitable even when incorporated together with time sequential acoustic features, which coincides with the conclusion (1) above. When detecting *disfluent* speech, on the other hand, we confirmed that the model *with* time sequential acoustic features and prosody only performs relatively close to the model when all the seven features are incorporated. This result again coincides with the conclusion (2) above, supporting that prosodic features are inevitable, while time sequential acoustic features also contribute to improving the recall-precision curve even without disfluency-based features.

6. Concluding Remark

This paper has demonstrated that the disfluency-based features derived from word fragments and filled pauses were effective relative to evaluating fluent/disfluent speech, especially when combined with prosodic features. The experimental evaluation results of those integrated diverse features further indicated that time sequential acoustic features contribute to improving the model with disfluency-based and prosodic features when detecting *fluent* speech, but not when detecting *disfluent* speech. Furthermore, when detecting *disfluent* speech, the model *without* time sequential acoustic features performs best even without word fragments features, but only with filled pauses and prosodic features. In the future, we plan to employ a number of automatic techniques to detect various disfluencies, such as filled pauses and word fragments within speech and text (Zayats and Ostendorf, 2019; Yulia et al., 2013; Fujimura et al., 2018; Ferguson et al., 2015; Dutrey et al., 2014; Wang et al., 2017; Zayats et al., 2016;



(a) Binary classification of fluent vs. neutral-disfluent classes



(b) Binary classification of fluent-neutral vs. disfluent classes

Figure 4: Experimental results of the LSTM framework (3): Comparison of filled pauses / word fragments / prosody features

Dong et al., 2019; Maekawa and Mori, 2015). Here, acoustic features should help detecting disfluencies and are expected to contribute to the proposed framework based on disfluecy-based features. In addition, further extension to the automatic evaluation of impressions of public speaking (Maekawa, 2014) other than fluency/disfluency could be the focus of future work.

7. Bibliographical References

Deshmukh, O. D., Kandhway, K., Verma, A., and Audhkhasi, K. (2009). Automatic evaluation of spoken English fluency. In *Proc. 34th ICASSP*, pages 4829–4832.

Dong, Q., Wang, F., Yang, Z., Chen, W., Xu, S., and Xu, B. (2019). Adapting translation models for transcript disfluency detection. In *Proc. 33rd AAAI*, pages 6351–6358.

Dutrey, C., Clavel, C., Rosset, S., Vasilescu, I., and Adda-Decker, M. (2014). A CRF-based approach to automatic disfluency detection in a French call-centre corpus. In *Proc. 15th Interspeech*, pages 2897–2901.

Ferguson, J., Durrett, G., and Klein, D. (2015). Disfluency detection with a semi-Markov model and prosodic features. In *Proc. NAACL-HLT*, pages 257–262.

Fontan, L., Le Coz, M., and Detey, S. (2018). Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with Japanese learners of French. In *Proc. 19th Interspeech*, pages 2544–2548.

Fujimura, H., Nagao, M., and Masuko, T. (2018). Simultaneous speech recognition and acoustic event detection using an LSTM-CTC acoustic model and a WFST decoder. In *Proc. 43th ICASSP*, pages 5834–5838.

Kjellgren, F. and Nordstrom, J. (2016). Convolutional neural networks for semantic classification of fluent speech phone calls. In *Proc. 6th SLTC*.

Kobayashi, K., Somiya, M., Nishizaki, H., and Sekiguchi, Y. (2008). Is a speech recognizer useful for characteristic analysis of classroom lecture speech? In *Proc. 9th Interspeech*, pages 1341–1344.

Maekawa, K. and Mori, H. (2015). Voice-quality analysis of Japanese filled pauses: A preliminary report. In *Proc. DiSS*.

Maekawa, K. (2014). Prosodic speaking styles extracted from the X-JToBI annotation of the corpus of spontaneous Japanese. *Journal of the Phonetic Society of Japan*, 18(1):70–82. (in Japanese).

Mahesha, P. and Vinod, D. S. (2012). An approach for classification of dysfluent and fluent speech using k-NN and SVM. *IJCSEA*, 2(6):23–32.

Mao, S., Wu, Z., Jiang, J., Liu, P., and Soong, F. K. (2019). NN-based ordinal regression for assessing fluency of ESL speech. In *Proc. 44th ICASSP*, pages 7420–7424.

Nishizaki, H., Somiya, M., Kobayashi, K., and Sekiguchi, Y. (2007). The effect of filled pauses in a lecture speech on impressive evaluation of listeners. In *Proc. 8th Interspeech*, pages 2673–2676.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rasipuram, S., Rao, P., and Jayagopi, D. (2016). Automatic prediction of fluency in interface-based interviews. In *Proc. INDICON*, pages 1–6.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. 10th Interspeech*, pages 312–315.

Tohyama, H. and Matsubara, S. (2006). Influence of pause length on listeners' impressions in simultaneous interpretation. In *Proc. Interspeech*, pages 893–896.

van Dalen, R. C., Knil, K. M., and Gales, M. J. F. (2015). Automatically grading learners' English using a gaussian process. In *Proc. SLaTE*, pages 7–12.

Wang, S., Che, W., Zhang, Y., Zhang, M., and Liu, T. (2017). Transition-based disfluency detection using LSTMs. In *Proc. EMNLP*, pages 2785–2794.

Yulia, T., Zaid, S., and Florian, M. (2013). Identification and modeling of word fragments in spontaneous speech. In *Proc. 38th ICASSP*, pages 7624–7628.

Zayats, V. and Ostendorf, M. (2019). Giving attention to the unexpected: Using prosody innovations in disfluency detection. In *Proc. NAACL-HLT*, pages 86–95.

Zayats, V., Ostendorf, M., and Hajishirzi, H. (2016). Disfluency detection using a bidirectional LSTM. In *Proc. 17th Interspeech*, pages 2523–2527.

8. Language Resource References

Kagomiya, T., Yamasumi, K., Maki, Y., and Maekawa, K. (2007). Development and analysis of a psychological evaluating database of public speaking. *Japanese Journal of Language in Society*, 9(2):65–76. (in Japanese).

Maekawa, K. (2003). Corpus of spontaneous Japanese: Its design and evaluation. In *Proc. SSPR*, pages 7–12.