2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2019)
Honolulu, Hawaii, USA, March 4-7, 2019

Categorizing Error Causes Related to Utterance Characteristics in Speech Recognition

Jennifer Santoso, Takeshi Yamada and Shoji Makino

University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305–8577 Japan
Phone/FAX:+81-29-853-2111
E-mail: s1820748@s.tsukuba.ac.jp

Abstract

Speech recognition systems are now widely used in our daily lives. However, sometimes speech recognition systems fail to recognize utterances. In many cases, users do not know what causes the failure, while the system asks users to repeat the utterance. When this situation continues, users consider that speech recognition systems are not user-friendly. The usability of speech recognition systems can be improved by specifying causes of error and presenting them in a way that users can easily understand, allowing them to improve the utterance.

In this study, our aim is to categorize causes of error related to utterance characteristics occurring in daily-use speech recognition systems and present the feedback to users. Here, we focus on causes of error related to the utterance speed, such as 'fast', 'slow', 'filler', and 'stuttered', since they are easy for users to correct and frequently occur in natural speech. We propose a categorization method with bidirectional long short-term memory (BLSTM) as the categorization model. In this paper, we compare the Mel filter bank with that of the modulation spectrum as feature extraction methods.

We perform an experiment in which it is decided whether a cause of error is present in a given utterance. The results indicate that our method using the modulation spectrum can reduce the number of false detections of causes of error related to the utterance speed, compared with the method using the Mel filter bank.

1. Introduction

Speech recognition systems are now widely used in our daily lives. Nowadays, speech recognition systems cater for various needs from transcribing short commands in commercial voice assistants to dictating long-duration utterances in systems such as lecture transcribers.

Unfortunately, sometimes speech recognition systems fail to recognize utterances. In many cases, users do not know what causes the failure, while the system asks users to repeat the utterance. When this situation continues, users consider that speech recognition systems are not user-friendly. To improve the usability of speech recognition systems, causes of error that may occur in speech recognition should be specified and presented in a way that users can understand.

Several approaches have been proposed as means of providing feedback regarding the causes. One approach suggests informing the users of the volume required for speech recognition in noisy environments [1]. In this approach, an appropriate utterance volume is predicted from an input noisy signal, and the resulting volume is then notified to the users. This approach enables the reduction of the potential recognition errors caused by a noisy signal.

Another approach is to estimate utterance characteristics from speech data [2]. Metrics for measuring utterance characteristics known as impression assessment indexes (i.e. 'activeness', 'easy-to-listen', 'smoothness', 'fluency') [3], which are thought to correlate with causes of speech recognition error, were evaluated. The estimates were then evaluated for speech data that were incorrectly recognized. The result showed that the estimator model used in their study, bidirectional long short-term memory (BLSTM), performs well in this task. Also, utterance characteristics can be estimated simply by using spectrogram-based features.

There are two points worth noting from the previous approaches. First, although impression assessment indexes correlate with causes of error, they cannot be used directly as the causes of error. Second, although the previous work focused on spectrogram-based features, there might be other features that are more closely related to the causes of error. It might be possible to categorize the causes of error more precisely by incorporating these features than the generally used spectrogram-based features.

In this study, we aim to categorize the causes of error occurring in daily-use speech recognition systems. In this study, causes of error are directly used as the categorization target, and acoustic feature extraction methods that contain information related to the utterance speed, such as 'fast', 'slow', 'filler', 'stuttered', are investigated. We propose a method of determining causes of error related to the utterance speed. It uses the Mel filter bank (MFB) and the modulation spectrum

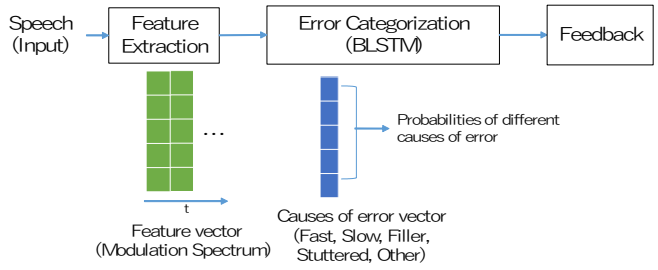

Figure 1: Proposed System Flow

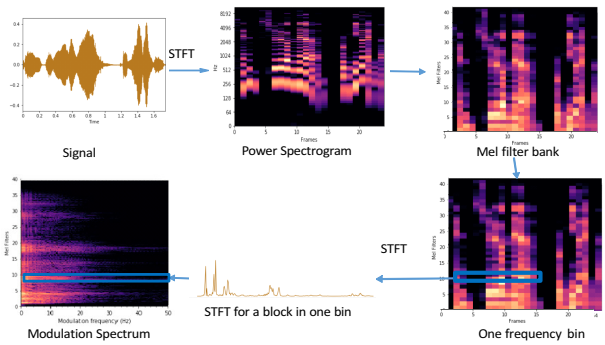

Figure 2: Modulation Spectrogram Process

(MS) as feature extraction methods, and BLSTM as the categorization model.

2. Proposed Method

In this section, the overall flow of the system and the details of the categorization task for the causes of error is explained. There are three points to be discussed in detail: the causes of error, the acoustic feature extraction methods used in the experiment, and the error categorization model.

2.1 Overview

The proposed system is shown in Fig. 1. First, the speech is inputted to the system. Features of the input speech are extracted, providing a time segment of feature vectors. The feature vectors are used as the input for the error categorization model. The result of the categorization is expressed as probabilities of different causes of error, which are processed to provide user-friendly feedback.

2.2 Causes of speech recognition error

Causes of error in speech recognition are factors that might cause failure in speech recognition. There are three typical causes of error in speech recognition. The first is environmental conditions, which are interference from outside such as noise, echo, reflection, and reverberation. The second is system factors such as unknown words, which are not listed in the dictionary. The third is utterance characteristics such as utterance speed, utterance volume, pronunciation, filler, and stutter. As utterance characteristics affect the recognition error rate [4], they are the main focus of this study.

To determine the utterance characteristics as the cause of error in this study, characteristics satisfying two conditions are considered: those that are easy for users to improve in the next utterance and have high occurrences in natural speech data. Therefore, the selected causes of error related to the utterance speed are ‘fast’, ‘slow’, ‘filler’, and ‘stuttered’ utterances.

2.3 Acoustic Feature Extraction

Acoustic feature extraction is an essential part of retrieving information from audio-based data. In speech recognition systems, spectrogram-based features such as the power spectrogram, Mel-frequency cepstrum coefficient (MFCC), and Mel filter bank (MFB) are commonly used as conventional feature extraction methods. Spectrogram-based features are mostly represented by time versus frequency signals. As spectrogram-based features contain phoneme information, they are capable of visualizing phonemes.

In this study, we use the modulation spectrum (MS) for the feature extraction method. This is because the MS shows the irregularities in syllable changes more clearly, and is independent of the speech content (phoneme information). As the MS has been successfully applied to a voice emotion recognition task [5], we attempt to apply the MS to extract features related to the utterance speed.

The MS is a representation of a filtered spectrum focusing on the temporal structure [6]. The MS also provides information on dynamic characteristics in a signal, such as syllable changes, and is related to the speech rhythm [7]. In this study, the computation of the MS is conducted in three steps. First, the power spectrogram is computed by applying a short-term Fourier transform (STFT) to the speech signal. The power spectrogram is then applied to Mel filters, producing the MFB. Finally, to obtain the MS, an STFT is applied again to each filter in the filter bank, which are then reconstructed to form the MS. The process is shown in Fig. 2.

2.4 Error Categorization Model

In the proposed method, BLSTM [8] is used as the error categorization model. As a variant of the recurrent neural network (RNN) classifier model, BLSTM consists of two long short-term memory (LSTM) networks that move forward and backward while storing time-series information. As speech data can be easily represented by time series and rely on the data in previous frames, BLSTM is suitable for handling audio-related tasks and has been successfully applied to

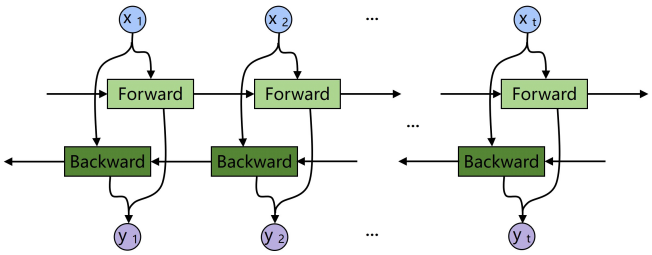

Figure 3: BLSTM Architecture

Table 1: Speech Data Specifications

Database	PASD	UUDB
Speakers	8 male 2 female	2 male 8 female
Sampling Rate	16 kHz	16 kHz
Bit Rate	16 bits	16 bits
Length	1−10 s	1−10 s
Dataset ID	kyo0121, kyo0221 kyo0321, osa0910 osa0918, uec0001 uec0002, uec0003 uec0004	C001, C024 C032, C052 C064

sound event classification [9]. BLSTM is illustrated in Fig. 3.

3. Experiment

In this paper, we consider a task that decides whether each cause of error is present in a given utterance. In this experiment, BLSTM consists of one hidden layer and one output layer with two units representing the probability of the presence and absence of each cause of error in any given utterance.

3.1 Experimental Data

This study incorporates data from two Japanese conversational speech databases, namely Priority Areas Spoken Dialogue Simulated Spoken Dialogue (PASD) [10] and Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) [11]. The data are selected among several datasets in PASD and UUDB.

In this study, we collect all incorrectly recognized speech data as our training data. To obtain the training data, speech data are inputted to Julius Speech Recognizer [12], and the recognition result is compared directly with the correct sentence. If the results are different, the speech data are then added and manually labeled according to the causes of errors (‘fast’, ‘slow’, ‘filler’, ‘stuttered’). In total, this experiment uses 1053 Japanese speech files from 10 male and 10 female speakers. Details are shown in Table 1.

Table 2: Mel filter bank Specifications

Sampling Rate	16 kHz
FFT Sample Points	2048
Mel Filters	40
Frame Length	25 ms
Frame Shift Length	10 ms

Table 3: Modulation Spectrum Specifications

Base Feature	40-dimensional MFB
Block length	320 ms
FFT Size (# frames)	32 (320 ms / 10 ms)
Dimension	40 * (32 / 2) = 640

3.2 Feature Extraction

The experiment involves the comparison of the proposed method using MFB with that using the MS as the feature extraction method. In this experiment, a 40-dimensional MFB is incorporated for feature extraction. The training using the MFB is compared with that using the MS derived from the MFB, which has 640 dimensions. The specifications of the MFB and MS can be seen in Tables 2 and 3, respectively.

3.3 Training Model and Evaluation

The training is carried out separately for each class, with each training model comprising of two classes of output, 1 or 0, representing the existence of each cause of error. Each training model consists of one hidden layer, tested with 16, 32, and 64 units in the hidden layer. Among these three models, the model that yields the best result is selected. The same training models are tested for both the MFB and MS. To ensure the validity of the training, five fold cross-validation is conducted for each class. The specifications of BLSTM for each class model are shown in Table 4. The models are evaluated using evaluation scales (precision, recall, and F-score) for each categorized potential cause of error. The precision, recall, and F-score are defined in Eqs. (1), (2), and (3), taking the average of each evaluation metric.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (1)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (2)$$

$$Fscore = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

3.4 Results and Discussion

Tables 5 and 6 show the evaluation results for the presence of different causes of error using the MFB and MS, respectively. For all causes of error except ‘stuttered’, the F-scores were 0.06−0.21 higher for the MS than for the MFB.

Table 4: Training Model Specifications

Classifier	BLSTM
Hidden Layers	1 (units tested: 16, 32, 64)
Optimizer	Adam
Learning Rate	0.001
Dropout	0.0
Loss Function	Softmax Cross Entropy
Epoch	50
Cross-validation	5-fold (Speaker Open Test)

Table 5: Results of BLSTM with MFB

	Precision	Recall	F-score
Fast	0.485	0.444	0.450
Slow	0.486	0.189	0.269
Filler	0.566	0.331	0.397
Stuttered	0.505	0.340	0.381

The higher F-scores were due to the high recall values, which means that false detection of the causes of error for the MS was greatly reduced. Among the causes of error, the F-score of 'filler' was the highest at 0.583, while the F-score of 'stuttered' was the lowest at 0.379. One of the possible reasons for the slightly lower F-score of 'stuttered' may have been the similarity of the modulation spectrum patterns of 'filler' and 'stuttered' speech.

4. Conclusion and Future Work

In this study, a method of categorizing causes of speech recognition errors related to the utterance speed was proposed. The results indicate that our method using the modulation spectrum can reduce the number of false detections for all causes of error related to the utterance speed, compared with the method using the Mel filter bank. Future work includes the enhancement of BLSTM as a categorization model and experiments with other neural network architectures as the categorization model. Also, to increase the categorization performance of the system, combining the use of multiple feature extraction methods while setting their optimum parameters should also be considered.

Acknowledgment

This research was supported by KAKENHI (17K00224).

References

[1] T. Goto, T. Yamada and S. Makino, "Novel speech recognition interface based on notification of utterance volume required in changing noisy environment", Proc. NCSP'18, pp. 192–195, 2018.

Table 6: Results of BLSTM with MS

	Precision	Recall	F-score
Fast	0.489	0.561	0.517
Slow	0.511	0.469	0.479
Filler	0.539	0.664	0.583
Stuttered	0.454	0.409	0.379

[2] T. Goto, T. Yamada and S. Makino, "Impression rating score estimation for cause of error notification in speech recognition", Proc. ASJ 2019 Spring Meeting, pp. 117–120, 2018.

[3] K. Yamazumi, T. Kagomiya, Y. Maki and K. Maekawa, "Impression rating scale of lecture speech", Journal of Acoustical Society of Japan, 61(6), pp. 303–311, 2005.

[4] S. Goldwater, D. Jurafsky and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical and disfluency factors that increase speech recognition error rates", Speech Communication, 52(3), pp. 181–200, 2010.

[5] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Contributions of the temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech", Acoustical Science and Technology, 39(3), pp. 234–242, 2018.

[6] H. Hermansky, "Should recognizers have ears?", Speech Communications, 25(1–3), pp. 3–27, 1998

[7] T. Kinnunen, K. Lee and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification", Odyssey, p. 30, 2008.

[8] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", Neural Networks, 18(5), pp. 602–610, 2005.

[9] T. Matsuyoshi, T. Komatsu, R. Kondo, T. Yamada and S. Makino, "Weakly labeled learning using BLSTM-CTC for sound event detection", Proc. APSIPA 2018, Paper ID 1478, pp. 1918–1923, 2018.

[10] Priority Areas Spoken Dialogue Simulated Spoken Dialogue, http://research.nii.ac.jp/src/en/PASD.html, Accessed: 2018-12-10.

[11] Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies, http://research.nii.ac.jp/src/en/UUDB.html, Accessed: 2018-12-10.

[12] Open-Source Large Vocabulary CSR Engine Julius, http://julius.osdn.jp/en_index.php, Accessed: 2018-12-10.