Classification of causes of speech recognition errors using attention-based bidirectional long short-term memory and modulation spectrum

Jennifer Santoso*, Takeshi Yamada*, and Shoji Makino*
* University of Tsukuba, Japan
E-mail: s1820748@s.tsukuba.ac.jp

***Abstract*—In this paper, we address the problem of classifying four common utterance characteristics related to the utterance speed, which cause speech recognition errors. We previously proposed bidirectional long short-term memory (BLSTM) as a classifier and the modulation spectrum as an acoustic feature. However, the performance of it is still insufficient, since BLSTM classified the utterance characteristics from the overall utterance, while most of the recognition errors resulted from utterance characteristics occur in only a small part of utterance. In this paper, we propose an approach to enhance classifier by using attention mechanism (attention-based BLSTM). Attention-based BLSTM enables the classifier to weight each frame according to its importance instead of directly measuring overall information from the speech. Furthermore, we investigate the correspondence of utterance characteristics to different modulation spectrum block lengths. To evaluate the performance of the proposed method, we conducted a classification experiment on Japanese conversational speeches with four different utterance characteristics: 'fast', 'slow', 'filler', and 'stutter'. As a result, the proposed method improved the F-score by 0.033–0.129 compared with the previously proposed method using BLSTM. This result confirms the effectiveness of attention-based BLSTM in classifying cause of errors based on utterance characteristics.**

I. Introduction

Along with the spread of portable devices, speech recognition systems have become more prevalent in recent years. Nowadays, speech recognition systems cater to various needs, from transcribing short commands in commercial voice assistants to dictating long-duration utterances such as in lecture transcribers. Although the performance of speech recognition systems is improving, there are still cases when speech recognition systems fail to properly recognize utterances from users. Several studies have attempted to detect speech recognition errors by using confidence measures [1] and to ask the user to clarify the specific words that might be incorrectly recognized in a spoken dialogue system [2]. However, users do not know what kind of utterances or which part of the utterances caused the error. Despite users' unawareness of the causes of the error, the system requests the user to repeat the utterance without giving any informative feedback. The lack of informative feedback decreases the usability of speech recognition systems and has discouraged users to continue using them, having deemed them as user-unfriendly. Therefore, improving the usability of speech recognition systems is important to improve users' next utterance and to encourage users to use the systems. This can be achieved by specifying the causes of error and presenting them in a way that is easy for users to understand.

A number of studies have attempted to provide feedback regarding specific causes of errors. One study suggested informing users of the utterance volume required for speech recognition in noisy environments [3]. In this study, an appropriate utterance volume was predicted from a noisy input signal, and the resulting volume was then notified to the user. The method used in the study reduced the potential recognition errors caused by a noisy signal.

In another study, utterance characteristics were estimated from the speech data. Pronunciation error detection on the utterance with scoring [4] or diagnostic feedback [5] was proposed to improve the usability of speech recognition-based computer-assisted language learning (CALL) systems. Although the scoring and feedback improved the pronunciation of the user, the user cannot improve the pronunciation immediately and the feedback tends to be complicated. Utterance characteristics such as utterance speed were also estimated from speech data [6]. The causes of errors occurring in daily-use speech recognition systems, such as 'fast', 'slow', 'filler', and 'stutter', were classified using the modulation spectrum (MS) as the acoustic feature and bidirectional long short-term memory (BLSTM) as the classifier. However, the performance of the proposed model is still insufficient as BLSTM classified the utterance characteristics from the overall utterance. In fact, most of the utterances that failed to be recognized correctly resulted from utterance characteristics occurring in only a small part of utterance.

To address this problem, we propose a classification method using attention-based BLSTM [7], which can automatically focus on the important part of an utterance with certain utterance characteristics as the cause of error. To evaluate the performance of the proposed method, we conduct a classification experiment on Japanese conversational speeches with four different utterance characteristics used in the previous study [6].

II. Methodology

In this section, the overall flow of the method and the details of the classification task to determine the causes of error are explained. Three points are discussed in detail: the causes

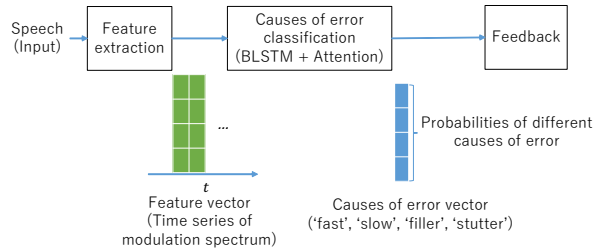


Fig. 1. Proposed method flow

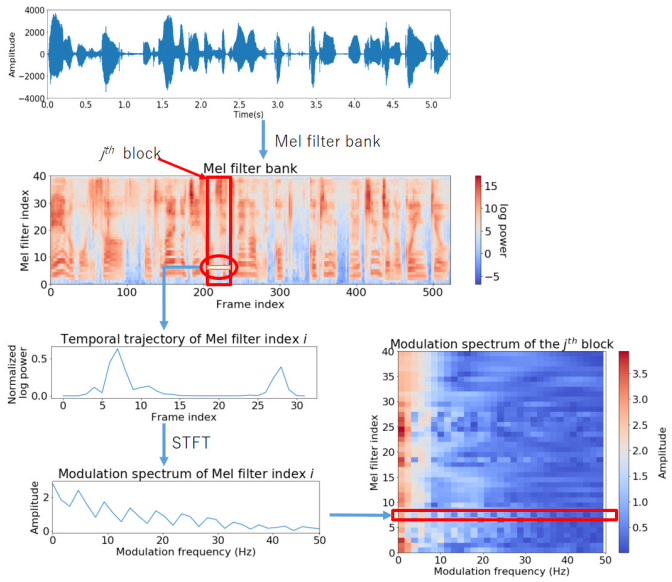


Fig. 2. Calculation process to obtain the modulation spectrum

of error, the acoustic feature extraction method used in the experiment, and the classifier.

A. Overview

The proposed method is shown in Fig. 1. First, the speech is inputted, then features of the input speech are extracted, providing a time series of feature vectors. The feature vectors are used as the input for the classifier. The result of the classification is expressed as probabilities of different causes of error which are processed to provide user-friendly feedback.

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, fillers, and stutter. As utterance characteristics affect the recognition error rate [8][9], they are the main focus of this paper.

To determine the utterance characteristics causing the error, characteristics satisfying two conditions are considered [6]: those that are easy for users to improve in the next utterance and those that occur frequently in natural speech data. Therefore, the selected causes of error related to the utterance speed are 'fast', 'slow', 'filler', and 'stutter' utterances [6].

B. 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 acoustic features. Spectrogram-based features are mostly represented by time versus frequency signals. As spectrogram-based features contain phoneme information, they are capable of visualizing phonemes. However, the information needed in this study is related to speed, not the spectral envelope. Therefore, it is important for the acoustic feature to be able to show information on changes in time series. This kind of acoustic feature is found in the MS [10][11].

In this study, we use the MS as the acoustic feature. MS shows the irregularities in syllable changes clearly, and is independent of the speech content (phoneme information). As the MS has been successfully applied to speech-based depression classification [12] and speech emotion recognition tasks [13], we also use the MS to extract features related to the utterance speed.

The MS is defined as the spectral representation of a temporal trajectory of a feature. In this study, the MS is represented by the acoustic frequency versus modulation frequency [11] of a speech signal. It provides information about energy modulation frequencies in the carriers of a signal and its dynamic characteristics, such as syllable changes. The MS is also known to be related to speech rhythm [14].

Fig. 2 illustrates the process of calculation used to obtain the MS. First, we compute the power spectrogram by applying a short-term Fourier transform (STFT) to the speech signal. The power spectrogram is then applied to Mel filters, producing the MFB. We then proceed to divide the frames into blocks of consecutive frames. For each block, we apply normalization for each Mel filter index as follows:

$$y_{ik} = log\left(\frac{exp(x_{ik})}{\sum_{k=1}^{t} exp(x_{ik})}\right), \quad (1)$$

$$y_i = [y_{i1}, y_{i2}, ..., y_{it}, 0, 0, ...0], \quad (2)$$

where x denotes log power, i denotes the Mel filter index, k denotes a frame index from the block, t is the number of frames in the block, and y denotes the normalized log power of the block. This normalization reduces the phoneme dependence. The normalized block is then zero-padded up to $Q \geq t$ points to increase the resolution of the modulation frequency. Finally, for each Mel filter index in the normalized block, we apply a Q-point STFT and combine the results to form the MS.

C. Classifier

In the previously proposed method, BLSTM [15] was used as the classifier. The BLSTM architecture is illustrated in Fig. 3. As a variant of the recurrent neural network (RNN) [16], the BLSTM consists of two long short-term memory (LSTM) [17] networks that move forward and backward while

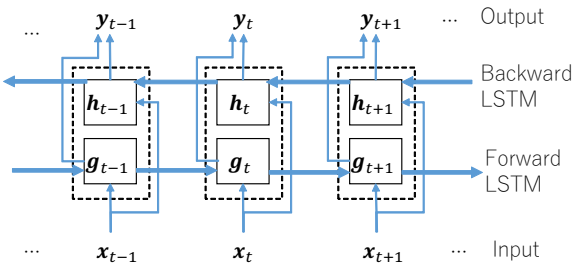


Fig. 3. BLSTM architecture

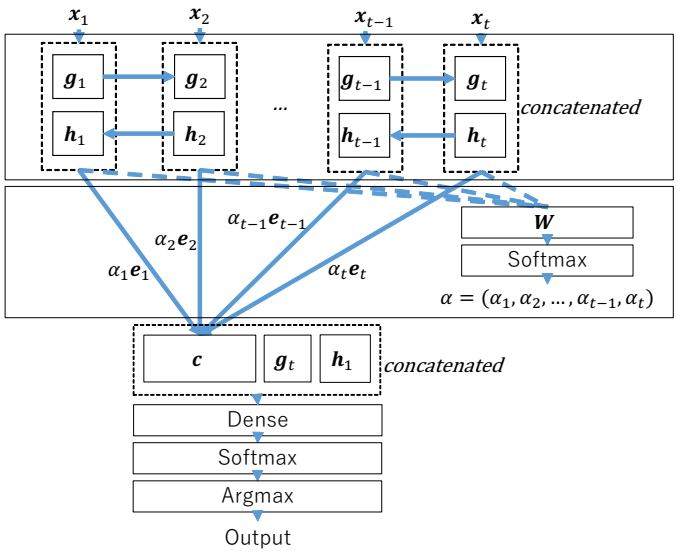


Fig. 4. Attention-based BLSTM in the proposed method

storing time-series information. As speech data can be easily represented by time series and depend on the data in previous frames, BLSTM is suitable for handling audio-related tasks and has been successfully applied to sound event classification [18].

In our method, we classify the causes of errors from utterances that failed to be recognized correctly using the system. However, in most cases, errors are not caused by the whole utterance, but by only a small part of the utterance that contains a specific characteristic. To address this problem, we propose using the attention mechanism [7] to enable the model to focus on specific parts of an utterance by weighting each parts to assist the decision making process. This mechanism has been successfully applied to classification tasks in various fields, such as psychological stress detection [19], speech emotion recognition [20][21], and acoustic event tagging [22].

We adopt part of the attention mechanism in the classification task[23], as illustrated in Fig. 4. The attention mechanism can be represented by

$$\mathbf{e}_i = \mathbf{g}_i \oplus \mathbf{h}_i, \tag{3}$$

$$\mathbf{u}_i = tanh(\mathbf{W}\mathbf{e}_i + \mathbf{b}), \tag{4}$$

$$\boldsymbol{\alpha}_i = \frac{exp(\mathbf{u}_i^T \mathbf{u}_i)}{\sum_{i=1}^{t} exp(\mathbf{u}_i^T \mathbf{u}_i)}, \text{ and} \tag{5}$$

$$\mathbf{c} = \sum_{i=1}^{t} \boldsymbol{\alpha}_i \mathbf{e}_i. \tag{6}$$

In these equations, $\mathbf{g}_i$ and $\mathbf{h}_i$ represent the forward and backward hidden states of BLSTM, respectively, which are concatenated to $\mathbf{e}_i$, as shown in Eq. (3). $\mathbf{e}_i$ is fed to the attention layer to determine the attention weight $\boldsymbol{\alpha}_i$ of each frame, which is determined by Eqs. (4) and (5). The output of attention mechanism is the weighted sum of $\mathbf{e}_i$, represented by vector $\mathbf{c}$, as shown in Eq. (6).

We improve the classification performance of BLSTM using the attention mechanism to direct the focus of the model on the important part that decides whether and where utterance characteristics occur. The classification process begins with feeding the acoustic features to the BLSTM model. The resulting hidden states of the BLSTM from both the forward and backward layers are then fed to the attention layer, resulting in the weighted sum vector $\mathbf{c}$. This vector is then concatenated with the final state of backward BLSTM $\mathbf{h}_1$ and final state of forward BLSTM $\mathbf{g}_t$, and fed to the dense layer for final classification. The prediction is the probability of the utterance characteristics occurring. We choose the highest probability to obtain the final result.

III. Experiment

The aim of the experiment is to assess the effectiveness of the attention mechanism in enhancing the classification.

TABLE I
SPEECH DATA SPECIFICATIONS

Database	PASD	UUDB
Speakers	8 male 2 female	2 male 8 female
Sampling rate	16 kHz	16 kHz
Quantization	16 bits	16 bits
Length	1−10 s	1−10 s
Dataset ID	kyo0121, kyo0221, kyo0321 osa0910, osa0918, uec0001 uec0002, uec0003, uec0004	C001, C002, C021, C022 C023, C024, C031, C032 C033, C051, C052, C053 C061, C062, C063, C064

A. Speech data

Table I shows the speech data specifications used in this study. This experiment is conducted using data from Japanese conversational datasets: Priority Area Speaking Dialogue (PASD) [24] and Utsunomiya University Database (UUDB) [25]. As we focus on incorrectly recognized speeches and their causes of error, all of the conversational data are inputted to the Julius speech recognizer [26]. The incorrectly recognized speeches are carefully examined for the occurrence of any utterance characteristics: 'filler' and 'stutter' information is available from the correct reference sentences; therefore, their occurrence can be extracted from the sentences. However, 'fast' and 'slow' occurrences are based on the utterance speed and cannot be extracted directly, so all speeches are manually labeled. In total, there are 1400 speech files of length ranging from 1–10 s compiled from 10 male and 10 female speakers.

TABLE II
MEL FILTER BANK SPECIFICATIONS

Sampling rate	16 kHz
FFT sample points	512
Mel filters	40
Frame length	25 ms
Frame shift length	10 ms

TABLE III
MODULATION SPECTRUM SPECIFICATIONS

Base feature	40-dimensional MFB		
Block length (frames)	80 ms (=8)	320 ms (=32)	640 ms (=64)
FFT size	64		
Dimension	$40 \times (64/2) = 1280$		

B. Feature extraction

Tables II and III show the acoustic feature extraction of the MFB and MS respectively. For the feature extraction, the base feature for the MS is extracted from a 40-dimensional MFB taken with a frame length of 25 ms and a frame shift length of 10 ms. The MS is then taken for a constant block length (8, 32, or 64 frames) for each of the 40 Mel filter indexes, each shifted every one frame. For each of the blocks, we apply zero padding to apply a 64-point FFT. In total, the MS for the input has a dimension of 1280 ($= 40 \times 32$).

C. Classifier

In this study, we compare the proposed method employing attention-based BLSTM with our previously proposed method employing BLSTM (baseline) as the classifier. For each method, the classifier is trained separately for each class. The number of hidden layers is taken to be 1 or 2 with hidden unit of 64 or 128; the dropout used in the experiment is set from 0.0 to 0.4; chosen optimizer is Adam with a learning rate 0.001, 0.0005, 0.0001, or 0.00005. The training in all models is conducted up to 50 epochs. To ensure validity, five-fold cross-validation is conducted on each class. Details of the classifier are given in Table IV. The model is evaluated using the F-score averaged over five folds, defined as follows:

$$F\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

where

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}, \text{ and} \tag{8}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}. \tag{9}$$

D. Experiment results and discussion

Evaluation results for the previous and proposed models, along with the set block size, are shown in Table V. From the results, BLSTM enhanced with the attention mechanism outperforms the previously proposed method in terms of the F-score for all classes by 0.033–0.129. Also, we can see that the F-score for every utterance characteristics depends on the block length. For instance, 'slow' and 'filler' are classified best with a block length of 64, whereas 'fast' is classified equally well with block lengths of 8 and 64. On the other hand, 'stutter' is classified best with a block length of 32.

TABLE IV
CLASSIFIER SPECIFICATIONS

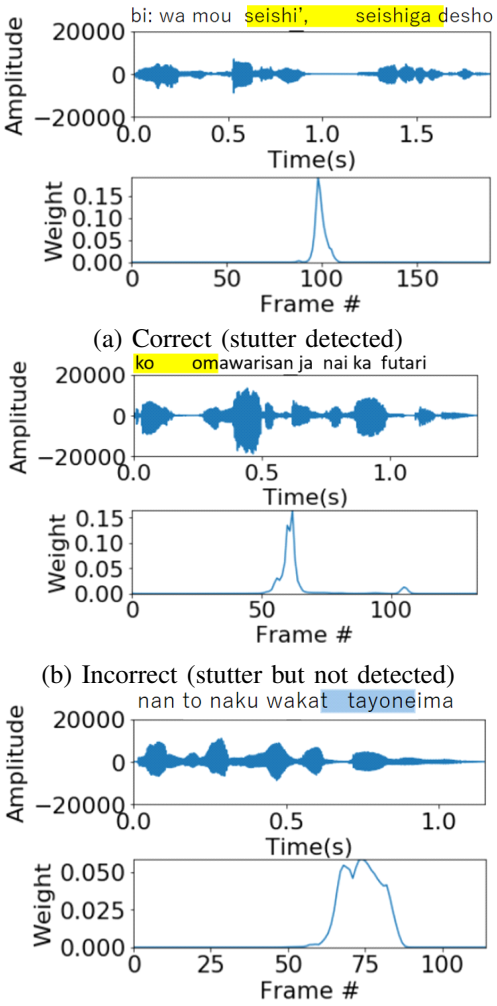
Classifier	Attention-based BLSTM
Layers and units	1280–(64–64)–256–2 1280–(128–128)–512–2
Optimizer	Adam
Learning rate	0.001, 0.0005, 0.0001, 0.00005
Dropout	0.0–0.4
Loss function	Softmax cross entropy
Epoch	1, 2, ..., 50
Cross-validation	5-fold speaker open test (training : test = 8 male+8 female : 2 male+2 female)

TABLE V
EXPERIMENT RESULT: F-SCORE

	BLSTM	Attention-based BLSTM		
Block length	32	8	32	64
Fast	0.577	**0.610**	0.587	**0.610**
Slow	0.576	0.611	0.616	**0.705**
Filler	0.641	0.666	0.682	**0.686**
Stutter	0.650	0.621	**0.691**	0.649

The process carried out by attention mechanism is also visualized in Fig. 5. For ease of visualization, we only show the attention mechanism for 'stutter'. Here, we extract the attention weights from the best trained model and evaluate the speech data, with each part of the speech representing the presence or absence of the utterance characteristics. The figures show the utterance transcription aligned with the waveform and the weight of each part of the speech. The parts highlighted in yellow indicate the occurrence of 'stutter', and the parts highlighted in light blue indicate the normal part of the utterance incorrectly detected to have a high probability of 'stutter'. Parts with a higher weight indicate greater importance in the decision of the classification. In the case of 'stutter', the sudden stops and a continuing similar pattern are important in confirming the presence of stutter, as demonstrated by the correct attention result in Fig. 5(a). Some repeated phonemes that are not sudden stops are weighted as not important by the attention mechanism in the case of 'stutter'. Also, some parts with a normal utterance speed have less weight as they are not closely related to utterance characteristics.

However, the model with the attention mechanism did not perform well with several types of speech data. For instance, some utterance characteristics failed to be recognized because the utterance speed was closer to normal in most parts, as demonstrated by the example of 'stutter' but not detected as 'stutter' in Fig. 5(b). It is likely that the output of the attention mechanism also depends on the output of BLSTM, which is also dependent on the block length of the MS. In the final example in Fig. 5(c), the utterance transcript highlighted in light blue is detected as 'stutter', possibly due to the assimilated sound in Japanese, indicated by double consonant phoneme, which resemble a 'stutter' pattern.



(a) Correct (stutter detected)

(b) Incorrect (stutter but not detected)

(c) Incorrect(not stutter but detected)

Fig. 5. Attention mechanism visualization of 'stutter' class

IV. Conclusion

In this paper, we proposed a method that uses attention mechanism to enhance BLSTM in classifying the causes of speech recognition error based on the utterance speed. Attention-based BLSTM enables the model to focus on specific parts that have the greatest importance in influencing the causes of errors. We conducted an experiment using Japanese conversational data from PASD and UUDB, most of which were incorrectly recognized by the Julius speech recognizer and labeled according to the utterance characteristics occurring in the speech data. Attention-based BLSTM improves the F-score from that obtained with the previously proposed method using BLSTM by 0.033–0.129 for all classes.

V. Acknowledgement

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant No. 17K00224.

References

[1] P.S. Huang, K. Kumar, C. Liu, Y. Gong and L. Deng. "Predicting speech recognition confidence using deep learning with word identity and score features," Proc. ICASSP 2013, pp. 7413–7417, 2013.

[2] E. Pincus, S. Stoyanchev and J. Hirschberg, "Exploring features for localized detection of speech recognition errors," Proc. SIGDIAL 2013, pp. 132–136, 2013.

[3] 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.

[4] R. Srikanth and L. B. J. Salsman, "Automatic pronunciation evaluation and mispronunciation detection using CMUSphinx," Proc. CICLing 2012, pp. 61–68, 2012

[5] R. Ai "Automatic pronunciation error detection and feedback generation for CALL applications," Proc. LCT 2015, pp. 175–186, 2015

[6] J. Santoso, T. Yamada and S. Makino, "Categorizing error causes related to utterance characteristics in speech recognition," Proc. NCSP'19, pp. 514–517, 2019.

[7] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," ICLR 2015, arXiv preprint arXiv, 1409.0473, 2014.

[8] M. A. Siegler and R. M. Stem, "On the effects of speech rate in large vocabulary speech recognition systems," Proc. ICASSP 1995, Vol. 1, pp. 612–615 1995.

[9] 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, Vol. 52, No. 3, pp. 181–200, 2010.

[10] H. Hermansky, "Should recognizers have ears?" Speech Communication, Vol. 25, No. 1–3, pp. 3-27, 1998.

[11] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," EURASIP Journal on Applied Signal Processing, No. 7, pp. 668-675, 2003.

[12] E. Bozkurt, O. Toledo-Ronen, A. Sorin and R. Hoory "Exploring modulation spectrum features for speech-based depression level classification," Proc. Interspeech 2014, pp. 1243-1247, 2014.

[13] 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, Vol. 39, No. 3, pp. 234–242, 2018.

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

[15] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," Proc. ICANN 2005, Vol. 2, pp. 799–804, 2005.

[16] C. L. Giles, G. M. Kuhn, and R. J. Williams, "Dynamic recurrent neural networks: theory and applications," IEEE Transactions on Neural Networks, Vol. 5, No. 2, pp. 153–156, 1994.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, Vol. 8, No. 9, pp. 1735–1780, 1997.

[18] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Roux and K. Takeda, "Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection," DCASE2016 Challenge, Tech. Rep., 2016.

[19] G. I. Winata, O. P. Kampman and P. Fung, "Attention-based LSTM for psychological stress detection from spoken language using distant supervision," Proc. ICASSP 2018, pp. 6204-6208, 2018.

[20] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," Proc. ICASSP 2017, pp. 2227–2231, 2017.

[21] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y.Tu "Attention based fully convolutional network for speech emotion recognition." Proc. APSIPA ASC 2018, pp. 1771–1775, 2018

[22] Y. Xu, Q. Kong, Q. Huang, W. Wang and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," Proc. Interspeech 2017, pp. 3083-3087, 2017.

[23] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao and B. Xu. "Attention-based bidirectional long short-term memory networks for relation classification." Proc. ACL 2016, Vol. 2, pp. 207–212. 2016.

[24] Priority Areas "Spoken Dialogue" Simulated Spoken Dialogue, http://research.nii.ac.jp/src/en/PASD.html, Accessed: 2019-06-12.

[25] Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies, http://research.nii.ac.jp/src/en/UUDB.html, Accessed: 2019-06-12.

[26] Open-Source Large Vocabulary CSR Engine Julius, http://julius.osdn.jp/en_index.php, Accessed: 2019-06-12.