Interspeech 2022
18-22 September 2022, Incheon, Korea

Performance Improvement of Speech Emotion Recognition by Neutral Speech Detection Using Autoencoder and Intermediate Representation

Jennifer Santoso[1], *Takeshi Yamada*[1], *Kenkichi Ishizuka*[2], *Taiichi Hashimoto*[2], *Shoji Makino*[1,3]

[1]University of Tsukuba, Japan
[2]RevComm, Inc., Japan
[3]Waseda University, Japan

j.santoso@mmlab.cs.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp, ishizuka@revcomm.co.jp, taiichi.hashimoto@revcomm.co.jp, s.makino@waseda.jp

Abstract

In recent years, classification-based speech emotion recognition (SER) methods have achieved high overall performance. However, these methods tend to have lower performance for neutral speeches, which account for a large proportion in most practical situations. To solve the problem and improve the SER performance, we propose a neutral speech detector (NSD) based on the anomaly detection approach, which uses an autoencoder, the intermediate layer output of a pretrained SER classifier and only neutral data for training. The intermediate layer output of a pretrained SER classifier enables the reconstruction of both acoustic and text features, which are optimized for SER tasks. We then propose the combination of the SER classifier and the NSD used as a screening mechanism for correcting the class probability of the incorrectly recognized neutral speeches. Results of our experiment using the IEMOCAP dataset indicate that the NSD can reconstruct both the acoustic and textual features, achieving a satisfactory performance for use as a reliable screening method. Furthermore, we evaluated the performance of our proposed screening mechanism, and our experiments show significant improvement of 12.9% in the F-score of the neutral class to 80.3%, and 8.4% in the class-average weighted accuracy to 84.5% compared with state-of-the-art SER classifiers.

Index Terms: speech emotion recognition, neutral speech detection, autoencoder, intermediate representation, screening mechanism

1. Introduction

Speech emotion recognition (SER) is the task of recognizing a speaker's emotional state from his/her speeches. SER has become an increasingly important task with the rise of voice-based technologies such as virtual voice assistants, call-center conversation analysis, and dialog systems. Based on deep learning-based methods, the state-of-the-art SER methods [1–7] have achieved high overall performance in classifying emotions. However, these methods tend to have lower performance for neutral speeches than for emotional speeches. Particularly in business conversations, neutral speeches account for a large proportion, so there is an urgent need to solve this problem. In general, increasing the amount of training data can be effective in improving classification performance. However, it is challenging to balance the number of neutral and emotional speeches because of the high cost of collecting emotional speeches. Another way to increase the training data is to use a data augmentation method for SER [8]. However, the main focus was not to improve the performance of neutral speeches but on the underrepresented emotions.

In several practical settings, such as business conversation analysis, most conversations do not contain emotions or are considered neutral. Emotional speeches might indicate potential trouble or unanticipated events in conversations. On the basis of this idea, we focus on the anomaly detection approach that uses only neutral speeches as training data. The anomaly detection model learns the representation of the normal data used in training and identifies anomalous data, that is the data deviating from normal behavior. Here, normal data can be regarded as neutral speeches and anomalous data as emotional speeches. The anomaly detection approach has been investigated in the acoustics domain, namely anomalous sound detection in machines [9]. In this approach, raw spectrograms are used as input and an autoencoder as a reconstructor. Here, the reconstructor is trained to minimize the reconstruction error of normal machine sounds. The normal machine sounds therefore will be successfully reconstructed, whereas the anomalous ones will not be reconstructed well. The anomalies in faulty machines were successfully detected from sounds using this approach.

There are several challenges to applying the anomaly detection approach to speech data in an SER domain. First, the reconstruction of speech is difficult because of the high dimensionality of a spectrogram and the variability of speech length. Second, it is difficult to deal with textual information, which is often utilized as input features in SER, to provide additional hints for possible anomalies in the emotion domain. These two problems can be solved by representing a raw spectrogram and textual information with fixed-length low-dimensional vectors. One way to realize it is to utilize the intermediate layer representation of the SER classifier as the input for anomaly detection, which has been studied in the field of image anomaly detection [10] and proven to be effective in improving the anomaly detection performance.

Inspired by the results of the study on the anomalous sound detection and tackling the classification task problems, we aim to improve the classification performance of SER through the anomaly detection approach combined with existing SER methods. First, we propose a neutral speech detector (NSD) that uses an autoencoder and the output of an intermediate layer in a pretrained SER classifier. The NSD reconstructs the output of the intermediate layer extracted by the feature extraction part of the pretrained SER classifier instead of reconstructing the raw speech spectrogram. The pretrained SER classifier extracts both the acoustic and text features and produces the fixed-length vector representation containing richer and optimized information for SER tasks. Then, we propose the combination of the SER classifier and the NSD used as a screening mechanism. Here, the NSD is employed to screen neutral speeches. The remaining speeches except for ones detected as neutral are classified according to the class probability of the SER classifier. The screening mechanism corresponds to the correction of the class probability of the speeches detected in the neutral class, which are incorrectly recognized in the SER classifier. We investigate the performance of the proposed method by comparing it with the state-of-the-art SER methods.

10.21437/Interspeech.2022-584

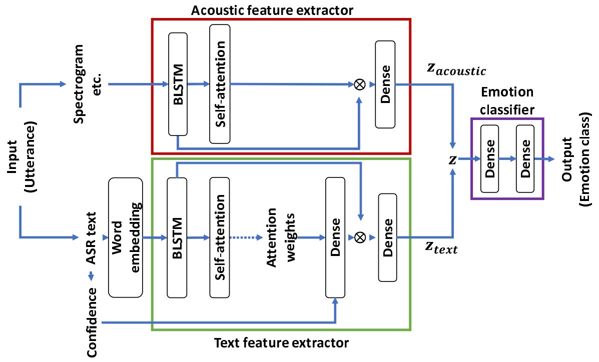

Figure 1: *Architecture of base SER*

2. Overview of base SER classifier

The SER classifier illustrated in Fig. 1 is based on our previous work [5]. The SER classifier consists of the acoustic feature extractor, the text feature extractor, and emotion classifier. The SER classifier receives input of speech and its automatic speech recognition (ASR) result, which are then fed to the acoustic and text feature extractor, respectively. The extracted feature representations of acoustic and text are then concatenated and fed to the emotion classifier to obtain the output of emotion class probability.

The input acoustic feature in the pretrained SER classifier consists of Mel-frequency cepstrum coefficients (MFCCs), constant Q-transform (CQT), and fundamental frequency (F0). These features are then fed to the bidirectional LSTM (BLSTM) [11] network to obtain $\mathbf{e}_i$, which is defined as

$$\mathbf{e}_i = \mathbf{g}_i \oplus \mathbf{h}_i, \tag{1}$$

where $\mathbf{g}_i$, $\mathbf{h}_i$, and $\oplus$ represent the forward hidden states of BLSTM, backward hidden states of BLSTM, and concatenation, respectively. $\mathbf{e}_i$ is then weighed for its importance by the self-attention mechanism [12] defined as

$$\alpha_i = softmax(\mathbf{w}_i tanh(\mathbf{W}\mathbf{e}_i^T)). \tag{2}$$

Here, α_i is the attention weight for each frame, and $\mathbf{w}_i$ and $\mathbf{W}$ are trainable parameters. Therefore, the weighted sum $\mathbf{v}$ from BLSTM and attention weights is defined as

$$\mathbf{v} = \sum_{i=1}^{T} \alpha_i \mathbf{e}_i. \tag{3}$$

After the weighted sum $\mathbf{v}$ is calculated, it is fed to a single fully connected layer to obtain a fixed-length intermediate layer representation, $\mathbf{z}_{acoustic}$, of acoustic features.

The text feature extraction in the pretrained SER classifier uses the ASR text from the same speech data as the input. The ASR result is first encoded by BERT word embedding [13]. The resulting features are then fed to the text feature extractor, with the flow similar to that of the acoustic feature extractor with the addition of the self-attention correction mechanism explained in our previous paper [5]. Here, we obtain the fixed-length intermediate layer representation $\mathbf{z}_{text}$ of text features.

The emotion classifier receives the output of the intermediate layer $\mathbf{z} = \mathbf{z}_{acoustic} \oplus \mathbf{z}_{text}$, which is then fed to layers of a dense network, providing emotion class probabilities. The emotion class is selected from the highest emotion class probability.

3. Proposed method

3.1. Overview

The process flow of the proposed method is illustrated in Fig. 2. The proposed method consists of the feature extractor, the NSD, and the screening mechanism part. The feature extractor is taken from the pretrained SER classifier explained in Sect. 2, following the steps until the output of the intermediate layer representation $\mathbf{z}$, which is then fed to the autoencoder-based NSD. The NSD works by reconstructing $\mathbf{z}$, resulting in the reconstructed feature vector $\hat{\mathbf{z}}$ and having the reconstruction error calculated as the anomaly score. When the anomaly score exceeds the decision threshold value, the input speech is classified as emotional (anomalous). Otherwise, it is classified as neutral (normal). Finally, the screening mechanism part decides the emotion class by correcting the neutral class probability based on the anomaly score.

3.2. NSD

The NSD of our proposed method consists of a deep autoencoder [14], which is a deep-learning architecture primarily used to represent higher-dimensional data, typically for efficient dimensionality reduction. In the proposed method, the autoencoder, which consists of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ in the form of two neural networks, is used to learn the representation of neutral speech through the output of the intermediate layer $\mathbf{z}$ of the SER classifier. The most attractive feature of our NSD is that it can deal with not only acoustic information but also textual information as the target of reconstruction. $\mathbf{z}$ is transformed into a compact bottleneck representation $\mathbf{v}$ with the encoder $\mathcal{E}$, whereas the decoder $\mathcal{D}$ maps back the bottleneck representation into the reconstructed intermediate layer representation $\hat{\mathbf{z}}$. The process is defined as

$$\mathbf{v} = \mathcal{E}(\mathbf{z}|\theta_E), \tag{4}$$

$$\hat{\mathbf{z}} = \mathcal{D}(\mathbf{v}|\theta_D), \tag{5}$$

where θ_E and θ_D represent the parameter set of an encoder and a decoder respectively. The reconstruction error of the autoencoder, hereby defined as the anomaly score, is computed as the mean square error (MSE)

$$r = \sum_{i=1}^{dim} ||\mathbf{z}_i - \hat{\mathbf{z}}_i||^2, \tag{6}$$

where dim is the dimension of $\mathbf{z}$. As the autoencoder is trained using only neutral speeches, $\mathbf{z}$ here represents the intermediate layer representation of neutral speeches.

We investigated the anomaly scores of the neutral speeches in the training data by the reconstruction experiment. As a result, it was found that the distribution of the anomaly scores is asymmetric. So the neutral/emotional decision is conducted using a decision threshold obtained from the value applied to the percentile point function of the Gamma distribution [15] of the anomaly scores in the training data.

3.3. Screening mechanism

We introduce a screening mechanism to combine the results of SER and the NSD to improve the SER performance further. In the screening mechanism, the NSD is utilized as the main decider for the final class decision, where speeches detected as neutral are automatically regarded as neutral in the SER result. In the following equation, we will assume $p_1, p_2, ..., p_k, ..., p_C$ as the SER class probability, C as the number of emotion classes, p_k as the neutral probability, and r as the reconstruction error. We compare two screening mechanisms in this study.

Weak screening Speeches detected as neutral by the NSD are regarded as neutral in the final SER class decision. On the

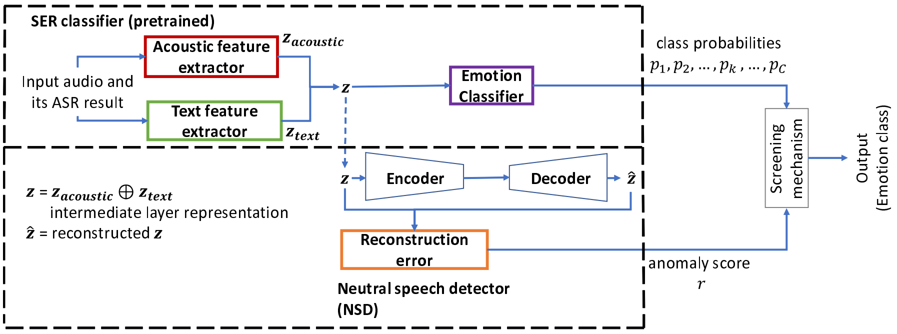
SER classifier (pretrained)
Acoustic feature extractor
$z_{acoustic}$
Input audio and its ASR result
Text feature extractor
z_{text}
z
Emotion Classifier
class probabilities
$p_1, p_2, \dots, p_k, \dots, p_C$
Screening mechanism
Output (Emotion class)
$z = z_{acoustic} \oplus z_{text}$
intermediate layer representation
$\hat{z}$ = reconstructed z
z
Encoder
Decoder
$\hat{z}$
Reconstruction error
anomaly score
r
Neutral speech detector (NSD)

Figure 2: *Proposed method flow*

other hand, the class decision for speeches not detected as neutral will defer back to the initial SER class probability. This is described as

$$p_k = \begin{cases} 1, & r \leq T, \\ p_k, & r > T \end{cases} \tag{7}$$

where T is the decision threshold.

Strong screening This is similar to the weak screening in terms of speeches detected as neutral. However, speeches not detected as neutral are regarded as any class other than the neutral class. In this case, the SER class decision takes the neutral class probability out of the equation and takes the remaining class with the highest probability as the result. The mechanism can be described as

$$p_k = \begin{cases} 1, & r \leq T, \\ 0, & r > T. \end{cases} \tag{8}$$

4. Experiments

4.1. Overview

The experiment aims to examine the effectiveness of the proposed method in improving SER performance. First, we evaluate the performance of the NSD by reconstructing $\mathbf{z}_{acoustic}$, $\mathbf{z}_{text}$, and $\mathbf{z}$, then comparing the F-score of neutral class with the result of the pretrained SER classifier. Then, we evaluate the overall performance of our proposed method of using the NSD as a screening mechanism by comparing with the state-of-the-art SER methods.

4.2. Dataset

In this study, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [16], one of the benchmark datasets for emotion recognition, to evaluate the effectiveness of the proposed method. The IEMOCAP dataset consists of scripted and improvised emotional speeches divided into five sessions, each containing one male and one female speaker. There are ten speakers (five males and five females) in the IEMOCAP dataset.

For the pretrained SER classifier, we used the data from four classes (happy, sad, neutral, and angry). To make it similar to previous works, we included the utterances labeled as excited to the utterances labeled as happy. The experiments were performed in five fold cross-validation. The training set comprises four sessions, and the test set comprises the remaining one session to ensure speaker independence. The F-score reported are based on the combined results from all five folds, not from averaging the F-score in each fold. The details of the dataset are shown in Table 1.

For the NSD, we used the same five fold cross-validation setting with the pretrained SER classifier. However, because we aim to train the neutral data representation, the training set contains only the neutral speeches from each of the four sessions.

Table 1: *Dataset specifications*

Dataset	IEMOCAP	
Speakers	5 males and 5 females	
Utterance length	1−19 s	
# of utterances	Happy	1689
	Sad	1084
	Neutral	1708
	Angry	1103

On the other hand, the test set of the NSD uses the same dataset as that for the SER but with the labels being neutral and the rest of the classes being emotional.

4.3. Input features

The features inputted to the pretrained SER classifier were divided into two parts for acoustic feature extraction and textual feature extraction [5]. For the acoustic feature extraction, we extracted a 33-dimensional feature consisting of 20-dimensional Mel-frequency cepstral coefficients (MFCCs), 12-dimensional constant Q-transform (CQT), and one-dimensional fundamental frequency (F0). All of the acoustic features are extracted using Librosa [17]. For the textual features, first, we conducted ASR on the input speeches using a recognizer pretrained with the Librispeech [18] dataset and Kaldi speech recognition toolkit [19]. Librispeech consists of approximately 1000 hours of speech sampled at 16 kHz. Next, we encoded the ASR texts using pretrained BERT [13], which was trained from lower-case English texts. The pretrained BERT consists of 12-layer and 110M parameters, resulting in 768-dimensional textual features.

4.4. SER classifier and NSD specifications

The pretrained SER consists of a feature extractor (acoustic feature extractor and textual feature extractor) and the emotion classifier. The feature extractor used BLSTM with 128 units and a self-attention mechanism with 128 units for the acoustic feature extractor and the additional confidence measure-based correction mechanism for the text feature extractor. The resulting intermediate layer representation $\mathbf{z}$ from the SER is a 256-dimensional vector, consisting of a 128-dimensional vector from each of the acoustic and text features.

The NSD is an autoencoder consisting of nine layers with units (256–128–64–32–16–32–64–128–256). The optimizer is set to Adam [20] with a learning rate of 0.00001 and dropout to 0.2. For the anomaly score calculation, we use the Gamma distribution of the reconstruction error of neutral speeches. The decision threshold is taken from the distributions' percentile point function with a value of 0.8, which yields the best performance among the tested percentile values. We evaluate the results for the NSD using F-score for neutral and the results for the SER using the average unweighted accuracy (UA), average weighted

Table 2: *Comparison of reconstructed feature*

Reconstructed feature	Neutral F-score (%)
Base SER method	67.4
Text	61.1
Acoustics	76.0
Acoustics + Text	**80.3**

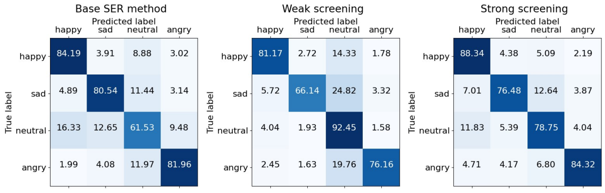

Figure 3: *Confusion matrices (in %) for base SER method, weak screening mechanism and strong screening mechanism*

accuracy (WA), and F-score of each emotion class. The pretrained SER model used as the feature extractor was taken from the model that yields the highest WA of the test data out of 100 epochs. Meanwhile, the results for the NSD were taken from the highest neutral F-score of the test data out of 100 epochs.

4.5. Results

Table 2 shows the F-score of our proposed method's neutral class in reconstructing the different features. Results of our experiment show that in all the different features reconstructed, the proposed method outperforms the SER method in terms of the neutral F-score. The base SER method (pretrained SER classifier) obtained a neutral F-score of 67.4%. The performance in reconstructing only the textual feature representation and the acoustic feature representation yields neutral F-scores of 61.1% and 76.0%, respectively. On the other hand, the reconstruction of both the acoustic and text features achieves a neutral F-score of 81.0%, which shows significant improvement from the base SER method and the reconstruction of a single feature. One possible explanation is that the intermediate layer output from the base SER method is produced by considering both the acoustic and text features in the training phase. Therefore, it is necessary for the NSD to use the representation from both acoustic and text features to achieve the best reconstruction. From the results, the NSD can be expected to have sufficient reliability as an input to the screening mechanism.

Table 3 shows UA and WA of our proposed method and the state-of-the-art SER classifiers with acoustic and text features as the input. Overall, our proposed method using the NSD for strong screening mechanism achieved UA and WA of 82.7% and 83.2% respectively. On the other hand, the use of NSD for the weak screening mechanism achieved UA and WA of 81.0% and 84.5% respectively. These results indicate the significant improvement of our method compared with the base SER method, achieving UA and WA of 75.9% and 76.1%, respectively. Table 4 shows the F-score of each emotion class of our proposed method and F-score reported in the state-of-the-art SER classifiers. The F-score of neutral is improved from 67.4% to 78.6% and 80.3% in the weak and strong screening mechanisms, respectively. The screening mechanism results indicate that the SER performance and the F-score of neutral speeches can be increased simply just by prioritizing the NSD screening result, where neutral speeches are automatically regarded as neutral. In results of the weak screening mechanism, most of the emotional classes show some performance increase because the speeches incorrectly recognized as emotional classes were corrected to neutral. However, the strong screening mechanism further improves the performance for neutral, angry, and sad speeches by 1.7%–2.8% from those of the weak screening mechanism.

Table 3: *SER performance comparison (UA, WA) with state-of-the-art methods. The symbol '–' means that the value does not described in the paper.*

Method	UA (%)	WA (%)
Neumann and Vu [2]	–	56.1
Feng et al. [3]	69.7	68.6
Siriwardhana et al. [4]	75.5	–
Base SER method [5]	75.9	76.1
Wang et al. [6]	77.1	76.8
Priyasad et al. [7]	79.2	80.5
(Proposed) Weak screening	81.0	**84.5**
(Proposed) Strong screening	**82.7**	83.2

Table 4: *SER performance comparison (F-score) with state-of-the-art methods. The symbol '–' means that the value does not described in the paper.*

Method	F-score (%)			
	Happy	Sad	Neutral	Angry
Neumann and Vu [2]	58.2	51.9	52.8	66.5
Feng et al. [3]	69.1	70.5	61.0	77.3
Siriwardhana et al. [4]	77.1	78.4	64.7	81.9
Base SER method [5]	81.5	76.2	67.4	80.4
Wang et al. [6]	–	–	–	–
Priyasad et al. [7]	–	–	–	–
(Proposed) Weak screening	**85.2**	75.6	78.6	82.5
(Proposed) Strong screening	85.0	**78.0**	**80.3**	**85.3**

The confusion matrices from the base SER method, the proposed method with the weak screening mechanism, and the proposed method with the strong screening mechanism are shown in Fig. 3. The strong screening mechanism improves the neutral classification performance from the base SER method by using only the neutral NSD detection result. As a result, it can be observed that the strong screening mechanism tends to improve the performance of both neutral and emotional classes in a well-balanced manner. On the other hand, the weak screening method drastically improves the neutral classification performance by using the NSD detection result and the SER class decision. However, in the weak screening mechanism, it can be observed that there is a tendency to incorrectly classify the emotional speeches as neutral.

5. Conclusions

To improve the performance of SER particularly for neutral speeches, we proposed an NSD that uses an autoencoder and the output of the intermediate layer from the pretrained SER classifier. It can deal with not only the acoustics information but also the textual information as the target of reconstruction. We then proposed a screening mechanism to screen the neutral speeches ahead and correct the class probability of the SER result. Experimental results confirmed that the NSD has a sufficient reliability as an input to the screening mechanism, and the screening mechanism achieved show significant improvement of 12.9% in the F-score of the neutral class to 80.3%, and 8.4% in the class-average weighted accuracy to 84.5% compared with the state-of-the-art SER classifiers.

6. Acknowledgements

Part of this work was supported by JST SPRING, Grant Number JPMJSP2124.

7. References

[1] Y. Li, T. Zhao, T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," Proc. Interspeech, pp. 2803–2807, 2019.

[2] M. Neumann, N.T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in Proc. Interspeech, pp. 1263–1267, 2017.

[3] H. Feng, S. Ueno, T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word ASR," Proc. Interspeech pp. 501–505, 2020.

[4] S. Siriwardhana, A. Reis, R. Weerasekera, S. Nanayakkara, "Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition, "Proc. Interspeech, pp. 3755–3759, 2020.

[5] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, T. Hiramura,"Speech emotion recognition based on attention weight correction using word-level confidence measure," Proc. Interspeech, pp. 1947–1951, 2021.

[6] Y. Wang, G. Shen, Y. Xu, J. Li, Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition," Proc. Interspeech, pp. 4518-4522, 2021.

[7] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, C. Fookes, "Attention driven fusion for multi-modal emotion recognition," Proc. ICASSP, pp. 3227–3231, 2020.

[8] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, S. Narayanan, "Data augmentation using GANs for speech emotion recognition," Proc. Interspeech, pp. 171–175, 2019,

[9] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma," IEEE/ACM Transactions on Audio Speech and Language Processing, pp. 212–224, 2019.

[10] V. Abdelzad, K. Czarnecki, R. Salay, T. Denounden, S. Vernekar, and B. Phan. "Detecting out-of-distribution inputs in deep neural networks using an early layer output." arXiv preprint arXiv:1910.10307, 2019.

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

[12] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, "A structured self-attentive sentence embedding," Proc. ICLR, 2017.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL, pp. 4171-4186, 2019.

[14] D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning internal representations by error propagation," Parallel Distributed Processing. Vol 1: Foundations, chapter 8, pp. 318-–362. MIT Press, Cambridge, 1986.

[15] K. O. Bowman, L. R. Shenton, "Gamma Distribution," International Encyclopedia of Statistical Science, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 573–575. doi: 10.1007/978-3-642-04898-2_269.

[16] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture dataset," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[17] B. McFee, C. Raffel, D. Liang,D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, "librosa: Audio and music signal analysis in python," Proc. Python in Science Conference, vol. 8, pp. 18-25, 2015.

[18] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "Librispeech: an ASR corpus based on public domain speech books," in Proc. ICASSP, pp. 5206–5210, 2015.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovský, "The Kaldi speech recognition toolkit," Proc. ASRU, pp. 1-4, 2011.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.