APSIPA Transactions on Signal and Information Processing, 2025, 14, e17

Original Paper

Speech Emotion Recognition Using Sequences of Fine-grained Emotion Labels with Phoneme Class Attributes

Ryotaro Nagase[1*], Takahiro Fukumori[2] and Yoichi Yamashita[2]

[1] *Graduate School of Information Science and Engineering, Ritsumeikan University, Japan*
[2] *College of Information Science and Engineering, Ritsumeikan University, Japan*

ABSTRACT

Recently, much research has been actively conducted on speech emotion recognition (SER) using deep learning, which predicts emotions conveyed by speech. Our study focused on a method of recognizing emotions at each frame level. One challenge with this approach is that emotion label sequences, which are used for training the frame-based SER, do not sufficiently account for phonemic characteristics. To overcome this limitation, we propose a new frame-based SER methods using fine-grained emotion label sequences that considers phoneme class attributes, such as vowels, voiced consonants, unvoiced consonants, and other symbols. As a result, we found that the proposed methods improve the utterance- and frame-level performance compared with conventional methods.

Keywords: Speech emotion recognition, deep learning, emotion label sequence, phoneme class attribute

*Corresponding author: Ryotaro Nagase, rnagase@fc.ritsumei.ac.jp. This work was supported by JST SPRING, Grant Number JPMJSP2101.

Received 24 October 2024; revised 18 March 2025; accepted 11 June 2025
ISSN 2048-7703; DOI 10.1561/116.20240077

1 Introduction

Speech emotion recognition (SER) is the process of predicting the emotion conveyed by speech. This technique can be applied to call center automation [3], mental health analysis [11], and e-learning systems [19]. In particular, it primarily deals with two types of emotion [36]. One is categorical emotion, which is the class of emotions, including happiness and sadness [28]. The other is dimensional emotion, including valence and arousal, which is expressed as a score on an axis [30]. In this study, we investigate how to classify utterances into categorical emotions.

Researchers have proposed many methods for improving the performance of SER. Among them, deep learning-based methods have achieved significant improvements. For instance, Satt *et al.* proposed a method of training convolutional neural networks (CNNs) and bidirectional long short-term memory (BLSTM) by extracting features robust to background noise from utterances divided into intervals of 3 [31]. Li *et al.* also proposed a method of training networks combining CNNs, BLSTM, and self-attention by the multitask learning of emotion and gender classification [20]. In other methods, various input features, model structures, and training strategies are used [6, 8, 26]. In recent years, methods using pretrained self-supervised learning (SSL) models representing speech information have also been proposed. For instance, Pepino *et al.* proposed a method of learning BLSTM using the embedded representation obtained from pretrained SSL models [27]. They showed that their method might be more effective for SER than methods using low-level descriptors and spectrograms. Cai *et al.* also proposed a multitask learning method for automatic speech recognition (ASR) and SER using pretrained SSL models, which improved the performance of SER [5]. In many other methods, pretrained SSL models, which are highly effective for improving SER performance [34, 23, 33, 40, 38], are used. The models trained by these methods are utterance-based SER ones, which estimate one emotion category for the entire utterance.

One limitation of these methods is that they cannot estimate emotional states that change in utterances. For example, emotional expressions in utterances may change, such as from neutral to happy. Previous studies have indicated the relationship between changes in acoustic features, such as pitch and power, and those in emotions [32, 2]. Since acoustic features continuously shift throughout utterances, it is considered that emotions also change dynamically in a similar manner. To train a model that can recognize fine-grained emotional expressions that change in utterances, it is necessary to use correct emotions in smaller units than utterances. Therefore, in some previous reports, methods have been proposed to train a frame-based SER model using emotion label sequences that consider different emotional expressions for each frame. This method is one of the time-continuous SER approaches, which can predict emotional states for each frame. A similar approach is to

estimate dimensional emotions for each speech sample from the waveform [16]. The difference between these approaches is the length of the labeled section. Although the estimation interval of the frame-level is longer than that of the speech sample, this approach makes it easier to train the model because the emotion labels become simpler. For instance, Fayek *et al.* proposed training a frame-based SER model that considers silent frames within an utterance [10]. In this method, sequences of emotion labels, including emotional and silent labels, are used in training the model. Han *et al.* also proposed training a connectionist temporal classification (CTC) model of SER reflecting the feature of voiced phonemes [13]. In their method, they use the emotion label sequences constructed under the condition that voiced phonemes indicate emotional states and other symbols indicate non-emotional states.

A problem with conventional methods using frame-based SER is that they do not consider the effects of acoustic differences such as vowels and consonants on emotion representation. In the conventional method [13], emotion label sequences are defined by assuming the emotion state in the case of voiced phonemes and the non-emotional state in other cases. Therefore, the acoustic differences in vowels, voiced and unvoiced consonants, as well as those in each phoneme between emotions, are not considered. In addition, the possibility that unvoiced phonemes and other symbols represent emotion states is not considered. Many studies have been conducted on the relationship between phonemes and emotions. For instance, Lee et. al. compared emotion recognizers for each phoneme class and investigated the effects of different phonemes on the accuracy of SER [18]. They found that the accuracy of recognizing emotions and the tendency to recognize them differed between phoneme classes. Aryani et. al. analyzed the frequency of phonemes for each emotion [9]. Their analysis suggested that the frequencies of phoneme classes differed depending on the type of emotion, such as tender, aggressive, positive, negative, and others. Yenigalla et. al. improved the utterance-based SER method by using phoneme embeddings and spectrograms [39]. Their results showed that the information of phoneme level symbols, such as pronunciations and their intervals, might be useful for utternace-based SER. Therefore, by reflecting the differences in phoneme classes such as vowels, voiced consonants, unvoiced consonants, and other symbols in an emotion label sequence, we considered that it may be possible to train a frame-based SER model considering the fine-grained differences among emotions.

We propose new frame-based SER methods considering fine-grained acoustic differences. In this paper, we introduce phoneme class attributes to emotion label sequences and compare three sets of phoneme class attributes that are suitable for the frame-based SER methods. These methods are expected to consider phoneme-dependent acoustic diversity during the training of frame-based SER models. One of the key points of this study is that, unlike previous research [24], we evaluate the effectiveness of the proposed methods using mul-

tiple pretrained models. In addition, another key aspect of this study is the evaluation of the accuracy of frame-based SER models using utterances with changes in emotions. This is the first time in the field of SER that the performance of frame-level SER based on CTC has been evaluated using these utterances. The results show that the frame-based SER models can recognize emotions at the frame level and that the proposed methods are effective for improving the performance of frame-based SER models. This paper proceeds as follows. In Section 2, we describe the conventional frame-based SER methods, which considers only voiced phonemes. In Section 3, we present the proposed methods for frame-based SER, which consider various phoneme class attributes. In Section 4, we explain the setup for the experiment and show the results. Finally, in Section 5, we present our conclusions and future work.

2 SER Using Emotion Label Sequences

A frame-based SER model is trained using emotion label sequences. In a previous study [13], an emotion label sequence is constructed from an utterance-level emotion label, in which the emotional state for voiced phonemes and the non-emotional state for silent intervals and unvoiced phonemes are assumed. The outline of training the frame-based SER model and predicting the emotion is shown in Figure 1. Examples of an emotion label sequence and CTC paths of the conventional method are highlighted in red, while those of the proposed method are highlighted in blue in Figure 1. We manually gather the transcription of utterances. In this paper, we will explain the conventional and proposed methods using the transcription from IEMOCAP (the interactive emotional dyadic motion capture) dataset as an example.

Before the training phase, we prepare the emotion label sequence. Figure 2 shows the method of converting the emotion label sequence. The section with highlighted in red represents the conventional sequence, whereas that with highlighted in blue represents the proposed sequence.

We convert the transcript corresponding to the input utterance into phonemes on the basis of the CMU Pronouncing Dictionary.[1] Among these phonemes, vowels and voiced consonants are converted into emotion sequence labels, whereas the rest are removed. The emotion label sequence consists of as many emotion labels as the number of voiced phonemes. For example, the utterance "YES, YES. [LAUGHTER]" (/jɛs, jɛs. [LAUGHTER]/) in the emotion of 'happiness' (H) contains four voiced phonemes: two voiced consonants (/j/) and two vowels (/ɛ/). Hence, this transcript is converted into {H, H, H, H}. Other phonemes are considered non-emotional states and are not included in the emotion label sequence.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

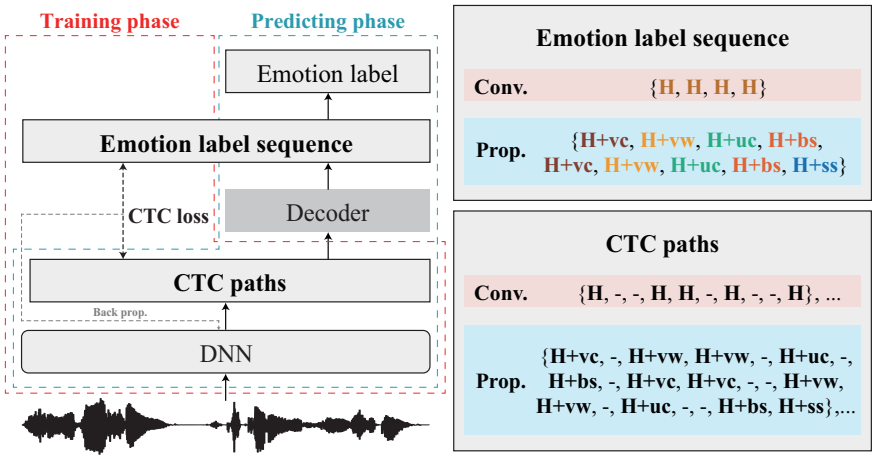


Figure 1: Outline of training the frame-based SER model and predicting the emotion. (In the right panels, the red block is the ground truth of conventional method, and the blue block is the ground truth of proposed method. 'H' is happiness, "bs" is basic symbol,"vw" is vowel, "vc" is voiced consonant, "uc" is unvoiced consonant, and "ss" is special symbol. A combination of emotion label and phoneme class attribute, represented by notations such as "H+vw," indicates one token of the emotion label sequences with phoneme class attribute.)

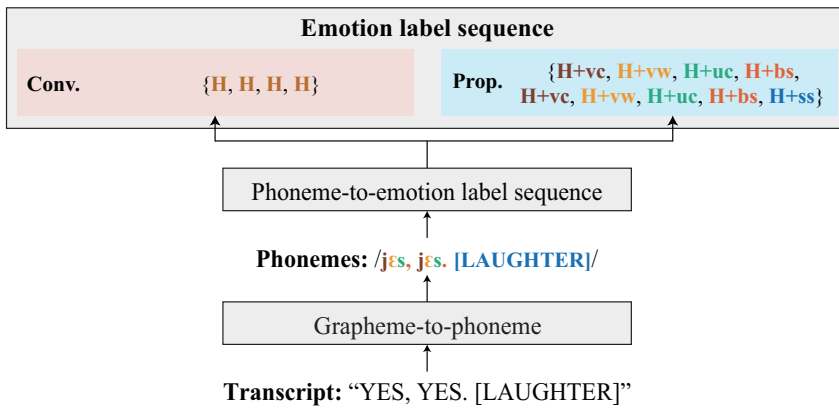


Figure 2: How to convert the transcript into the emotion label sequence. (The red and blue blocks have the same meaning as in Figure 1. Similarly, each symbol, such as 'H,' "vw," and "H+vw," has the same meaning as in Figure 1.).

During the training of the model using this emotion label sequence, we can utilize various network structures. For example, there are models that combine BLSTM and various attention mechanisms [41] and models that combine parallel CNNs, the squeeze-and-excitation network (SENet), and the dilated

residential network (DRN) [42]. The model for frame-based SER is trained on the basis of CTC, which is a framework for estimating paths (CTC paths) containing blank symbols (-) and repeated symbols [12]. Note that a blank symbol represents the non-emotional state. This method can estimate output sequences even when the output length is smaller than the input length. During CTC-based training, given an input $\mathbf{x} = [x_0, \ldots, x_{\mathrm{T}}]$ of length T, we maximize the probability of obtaining an emotion label sequence $\mathbf{y} = [y_0, \ldots, y_{\mathrm{L}}]$ of length $\mathrm{L}(\leq \mathrm{T})$. The probability $p(\mathbf{y}|\mathbf{x})$ is given in Equation 1, and the CTC loss function $\mathcal{L}_{\mathrm{ctc}}$ is given in Equation 2. Let x_t be the input of time t, π_t the emotion label of time t, π the CTC path for the emotion label sequence, and $\Phi(y)$ the set of π. Moreover, let U be the set of training data.

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\pi \in \Phi(\mathbf{y})} \prod_{t=1}^{\mathrm{T}} p(\pi_t|x_t) \tag{1}$$

$$\mathcal{L}_{\mathrm{ctc}} = -\sum_{(\mathbf{x},\mathbf{y}) \in U} \log p(\mathbf{y}|\mathbf{x}) \tag{2}$$

The number of classes estimated in each frame of the CTC path is the number of emotions + 1 (blank symbol). In the prediction phase, an emotion label sequence is obtained by gathering estimated symbols within the CTC path by deleting blank symbols and merging repeated characters. Eventually, the CTC path is regarded as a result of the frame-level prediction, whereas the emotion with the highest frequency among the emotion labels in a sequence is regarded as a result of the utterance-level prediction.

3 SER Using Emotion Label Sequences with Phoneme Class Attributes

We propose methods of training the frame-based SER model considering the phoneme class attributes. In this study, we define five phoneme classes: basic symbols (bs), vowels (vw), voiced consonants (vc), unvoiced consonants (uc), and special symbols (ss). Basic symbols are mainly punctuation marks, such as '?', '!', and '.'. Special symbols are unique information for each dataset, such as '[BREATHING]'. Lee *et al.* have discussed that vowels have an important role in SER [18]. Aryani *et al.* have also suggested that voiced or unvoiced consonants may express various emotions [9]. As for other symbols, previous studies [40, 39] have shown that inputting embedded representations of phonemes, including silent and special symbols, into models improves performance. Therefore, we propose the methods that consider the attributes of vowels, voiced and unvoiced consonants, and other symbols, and we expect to improve the performance of the frame-based SER models. Any symbol not falling under the above attributes is considered a non-emotional state. The

blue highlight in Figure 1 shows examples of the emotion label sequence with phoneme class attributes and CTC paths of the proposed method, whereas that in Figure 2 shows the proposed sequence. A significant difference from the conventional method is that the emotion label sequence explicitly considers various types of attribute information of phonemes. For example, the utterance "YES, YES. [LAUGHTER]" (/jɛs, jɛs. [LAUGHTER]/) in the emotion of 'happiness' (H) has two base symbols (/,/, /./), two vowels (/ɛ/), two voiced consonants (/j/), two unvoiced consonants (/s/), and one special symbol (/[LAUGHTER]/). Thus, this transcript is converted into {H+vc, H+vw, H+uc, H+bs, H+vc, H+vw, H+uc, H+bs, H+ss}. Note that 'emotion label+phoneme class attribute' is an emotion label with the phoneme class attribute. The number of classes estimated in each frame of the CTC path is the number of emotions × the number of attributes + 1 (blank symbol). During the training phase, the DNN model is trained using emotion label sequences with the phoneme class attribute. During the prediction phase, the emotion label sequence is obtained from the estimated CTC path. The CTC path is considered the frame-level result, and the emotion with a high frequency of appearance is considered the utterance-level result.

4 Experimental Setup

In this section, we describe the experimental setup. First, we compared the utterance-level accuracies of the models trained by conventional and proposed methods for emotion label sequence recognition. We also compared the results with those reported in previous studies [13, 41, 42]. In addition, to investigate whether the proposed method improved the frame-level performance, we compared the results using the combined data of different emotion utterances. The dataset, models, and metrics used in this experiment are as follows.

4.1 Dataset

We utilized the IEMOCAP database, which includes English emotional speech [4]. It is composed of acted or improvised utterances during dialogues. This dataset consists of five sessions, and each session has a dialogue between a man and a woman. This dataset consists of 10,039 utterances, 5,255 of which are acted and 4,784 are improvised dialogues. The utterances are categorized into ten emotion labels: neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and others. Each utterance is assigned to one of these emotion labels on the basis of a majority vote by annotators.

In the utterance-level evaluation, we exclusively utilized improvised dialogues to avoid semantic interference [41, 42]. We developed and evaluated the models to predict four emotion categories: anger, happiness, sadness, and

neutral. In addition, we included the excitement utterances within the happiness data. Consequently, the total number of utterances for each emotion label amounted to 2,943, distributed as follows: anger (289), happiness (947), sadness (608), and neutral (1,099). Table 1 presents the number of utterances per session and per emotion in this dataset. The average duration of the utterances utilized in the experiment was approximately 4.5 s. During the training phase, utterances that were 15 s or shorter were only used, as in related works [20, 1].

Table 1: Number of utterances per emotion and per session.

Session	Ang.	Hap.	Sad.	Neu.	Total
1	62	132	104	223	521
2	22	191	100	217	530
3	90	149	190	198	627
4	84	195	81	174	534
5	31	280	133	287	731
Total	289	947	608	1099	2943

In the frame-level evaluation, we utilized the same training data used for the utterance-level evaluation, and we created and used the evaluation data in which emotion labels change within utterances. The procedure for creating the evaluation data is shown in Figure 3. First, we selected one utterance

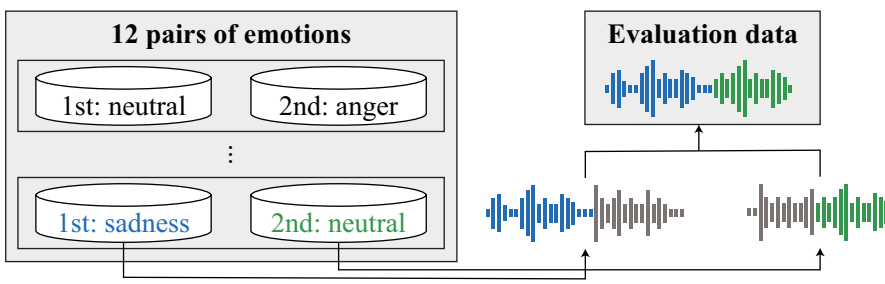


Figure 3: Outline of creating the evaluation data (blue: sadness, green: neutral).

from each of two different emotion categories. Both were longer than the average utterance length of the dataset. Because there were four emotions in this study, the number of permutation pairs was 12. Next, we calculated the word alignment for each utterance by ctc-segmentation [17] and split the utterances at the midpoint on the basis of the total number of words. This process ensured that the utterances were not cut in the middle of a word.

Finally, we combined the first half of the utterance from the first emotion with the second half of the utterance from the second emotion in each pair, and used 600 combined utterances in each fold.

We produced the phonemes for generating emotion label sequences with a grapheme-to-phoneme (g2p) conversion toolkit.[2] The mapping between the phoneme class attribute and the symbols is shown Table 2. The dataset contained the following tokens: 37,304 instances of bs, 40,474 instances of vw, 37,388 instances of vc, 21,791 instances of uc, and 133 instances of ss. Furthermore, the total number of voiced phonemes was 77,862. In this experiment, we evaluated the proposed methods using three different sets of phoneme class attributes to investigate their effectiveness. Table 3 illustrates the phoneme class attributes considered in the conventional and proposed methods, along with the number of predicted classes per frame in the CTC path. Conv. defined the voiced phoneme that combines vowels and voiced consonants [13]. Prop. I distinguished between vowels and voiced consonants, Prop. II extended Prop. I with the distinction of unvoiced consonants, and Prop. III further extended it with the incorporation of the distinction of basic symbols. We compared them and attempted to clarify phoneme class attributes to be considered during the training of the frame-based SER model. Special symbols were distinguished in all methods because they depended on the dataset. Note that we trained the model five times with different random seeds and took the average of the results.

Table 2: Mapping between the phoneme class attribute and symbols.

Phoneme class	Symbols
Basic symbols (bs)	!, ?, ', ,, -, ., >
Vowels (vw)	AA, AE, AH, AO, AW, AY, EH, ER, RY, IH, IY, UH, UW, OW, OY
Voiced consonants (vc)	B, D, DH, G, L, M, N, NG, JH, R, V, W, Y, Z, ZH
Unvoiced consonants (uc)	CH, F, HH, K, P, S, SH, T, TH
Special symbols (ss)	[LAUGHTER], [LIPSMACK], [GARBAGE], [BREATHING]

We evaluated the utterance-level and frame-level results for each method. The evaluation scheme was ten folds cross-validation with no speaker overlap. In each fold, the dataset was divided into eight speakers for the training data, one speaker for the validation data, and one speaker for the test data. During the utterance-level evaluation, we utilized the test data in each fold. During

[2] https://github.com/Kyubyong/g2p

Table 3: Numbers of classes in each method.

	Phoneme class	# of estimated classes
Conv.	voiced	5 (4 emos. + 1 blk.)
Prop. I	vw, vc, ss	13 (4 emos. × 3 atts. + 1 blk.)
Prop. II	vw, vc, uc, ss	17 (4 emos. × 4 atts. + 1 blk.)
Prop. III	bs, vw, vc, uc, ss	21 (4 emos. × 5 atts. + 1 blk.)

the frame-level evaluation, we combined utterances with different emotions to artificially create test data with changes in emotions and used them in each fold.

4.2 Models

For the model of frame-based SER, we utilized a pretrained model, wav2vec2.0 and HuBERT. Wav2vec2.0 is a self-supervised representation learning framework [1]. In this framework, the model is trained by contrastive learning, which estimates the quantized speech representation of masked sequences with that of unmasked sequences as negative samples. Hubert is also a self-supervised representation learning method [14]. In this framework, the model is trained by predicting frame-level classes from speech. These classes are defined on the basis of the results of clustering the frames using acoustic features. We utilized the pretrained models wav2vec2.0[3] and HuBERT[4] provided by Hugging Face [37]. These were fine-tuned models of automatic speech recognition, which were pretrained in speech representation learning with Libri-Light [15] and Librispeech [25]. Their architecture consisted of seven CNN layers and 24 Transformer layers. We trained the model that combines one linear layer with either wav2vec2.0 or HuBERT. During the training phase, we fixed the model parameters of the CNN layers and fine-tuned that of the Transformer layers and the FC layer. The inputs of the models were waveforms of speech, and the outputs were emotion sequences represented by CTC paths. We set the number of epochs to 50, the batch size to 8, and the learning rate to 0.0001 using RAdam [21] as the optimization method. During the training phase, we applied gradient clipping with a threshold of 5.0. The model parameters were updated by minimizing the CTC loss.

4.3 Metrics

The utterance-level-evaluated metrics were the weighted accuracy (WA) and unweighted accuracy (UA). WA was the overall accuracy, and UA was the

[3] facebook/wav2vec2-large-960h-lv60
[4] facebook/hubert-large-ls960-ft

average recall of each emotion. If no emotion was recognized, we classified the input utterance as 'neutral.' Higher WAs and UAs indicate higher model performance. The frame-level-evaluated metric was the emotion match rate (EMR). This score indicates the percentage of frames correctly predicted out of all frames that were predicted to be in an emotional state. If the predicted token was a blank symbol, we considered it a non-emotional state. Higher EMRs indicate the higher frame-level recognition performance of the model. We compared the WA, UA, and EMR of the proposed methods with those of the conventional methods. Additionally, we performed tests to examine the significant differences between the conventional and proposed methods. For WA and UA, we set the null hypothesis that there was no significant difference in results among the methods when the frequencies of positive and negative differences in the number of recognition errors were equal. For EMR, we established the null hypothesis that there was no significant difference in results among the methods when the frequencies of positive and negative differences in the EMR produced per speech were equal. On the basis of these hypotheses, we performed a two-sided sign test.

5 Results

5.1 Utterance-level Evaluation

Table 4 shows WA and UA, the utterance-level results for each method. A comparison between conventional and proposed methods shows that all the proposed methods outperform the conventional method in improving the WA and UA. Comparing proposed method I with the conventional method, WA improved by 1.9% and UA by 1.6% with wav2vec2.0+FC, and WA improved by 2.5% and UA by 2.8% with HuBERT+FC. By distinguishing between vowels and voiced consonants within voiced phonemes, the model could recognize fine-grained emotions attributed to the differences in phoneme-level acoustic features. Comparing proposed method II with the conventional method, WA improved by 2.2% and UA by 1.8% with wav2vec2.0+FC, and WA improved by 2.6% and UA by 2.9% with HuBERT+FC. With unvoiced consonants added to the distinguished phoneme class attributes, the model might have recognized emotions related to phonemes characterized by variations in breath speed and pitch rise during speech. Comparing proposed method III with the conventional method, WA improved by 1.4% and UA by 1.2% with wav2vec2.0+FC, and WA improved by 3.1% and UA by 3.3% with HuBERT+FC. It might be possible for the model to recognize emotions related to subtle changes in breath speed by considering basic symbols.

Comparing the three proposed methods, both method II with wav2vec2.0 +FC and method III with HuBERT +FC showed the highest performance.

Table 4: Utterance-level results of each method.

Model	wav2vec2.0+FC		HuBERT+FC	
	WA (%)	UA (%)	WA (%)	UA (%)
Conv.	73.3	72.7	71.3	69.4
Prop. I	75.2*	74.3*	73.8*	72.2*
Prop. II	**75.5***	**74.5***	73.9*	72.3*
Prop. III	74.7*	73.9*	**74.4***	**72.7***

* $p < 0.05$ for significant difference compared with Conv.

The observed performance differences are considered due to variations in the pretraining methods of the SSL models used. The model parameters of wav2vec2.0 were trained using contrastive learning, whereas those of HuBERT were trained with frame-wise clustering results. Therefore, when we utilized HuBERT pretrained with fine-grained frame-level classification, proposed method III, which considers most phoneme class attributes, was considered to have achieved the highest WA and UA among the proposed methods. On the other hand, when we utilized wav2vec2.0 pretrained by contrastive learning between masked and unmasked frames, the effect of considering the basic symbols was not clear. These results are considered due to whether the self-teaching labels clearly included information about the basic symbols during pretraining. From the above results, we found that the most effective proposed method varies depending on the SSL models used.

Comparing the results among the SSL models used for training, WA and UA when were higher using wav2vec2.0+FC than those when using HuBERT+FC. Because wav2vec2.0 was trained on the differences between similar and dissimilar frames using contrastive learning and enhanced the ability to identify differences between frame-level information, it is considered suitable for the proposed methods.

Table 5 shows the WA and UA of the proposed methods as well as those in previous studies. The performance of the proposed methods is comparable to or higher than that in previous studies. In particular, proposed method II with wav2vec2.0+FC improves WA by 2.4% and UA by 8.2% compared with a conventional method in a previous study [42]. This result shows that a fine-tuned SSL model effectively improves the utterance-level performance of a frame-based SER model.

Table 5: Comparison between the utterance-level accuracies of proposed methods and those of conventional methods.

Conventional methods	WA (%)	UA (%)
BLSTM [13]	64.2	65.7
BLSTM+Component Attention [41]	69.0	67.0
PCNSE+SADRN [42]	73.1	66.3
Proposed methods		
HuBERT+FC (prop. III) [Ours]	74.4	72.7
wav2vec2.0+FC (prop. II) [Ours]	**75.5**	**74.5**

5.2 Frame-level Evaluation

Table 6 shows the EMRs for the conventional and proposed methods.

Table 6: Frame-level results of each method.

Model	Match Rate (%)	
	wav2vec2.0+FC	HuBERT+FC
Conv.	46.6	44.1
Prop. I	47.6*	45.1*
Prop. II	48.8*	46.9*
Prop. III	**49.0***	**47.0***

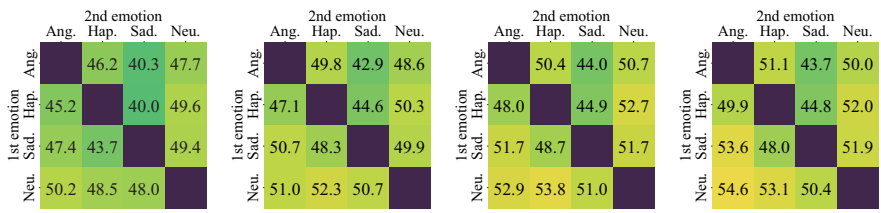
* $p < 0.05$ for significant difference compared with Conv.

There was a significant difference between the EMR of each proposed method and that of the conventional method at $p < 0.05$. A comparison between conventional and proposed methods showed that the performance of all proposed methods was higher than that of the conventional method. This result indicated that incorporating phoneme class attributes into emotion label sequences effectively improved the frame-level performance of the frame-based SER model.

Comparing the three proposed methods, we found that Prop. III was the most effective in improving the EMR. The EMR of proposed method III was improved by 2.4% with wav2vec2.0+FC and by 2.9% with HuBERT+FC. These results suggest that a more detailed consideration of phoneme class attributes might improve the frame-level performance.

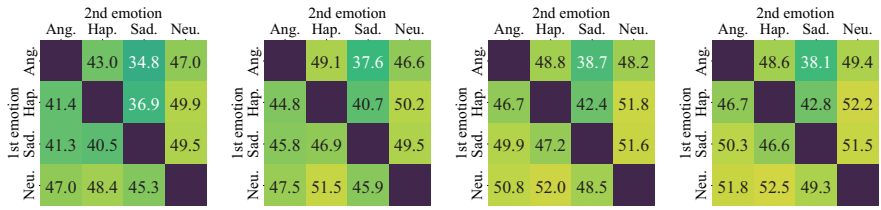
Comparing the results of each SSL model used for training, the EMR of wav2vec2.0+FC was higher than that of HuBERT+FC. As stated previously, because wav2vec2.0 was trained to distinguish frame-level differences, it was considered suitable for frame-based SER.

Figures 4 and 5 show the average EMR for each emotion pair in the evaluation data of the frame-based SER model using wav2vec2.0+FC and HuBERT+FC, respectively. Each axis label represents the ground truth of emotions, with the row labels indicating emotions for the first half of the evaluation data and the column labels indicating emotions for the second half of the evaluation data. From the results shown in the figures, the proposed methods outperform the conventional method in all permutations of emotions. In particular, for the evaluation data including "neutral," the results indicate that increasing the number of phoneme class attributes generally leads to the improvement in EMR. On the other hand, for the evaluation data of permutations where the first emotion is "anger" and the second emotion is "sadness," the EMR is consistently low regardless of the method or model used. It may be easier to estimate emotions in the permutations including "neutral" than in those of distinct emotion classes, such as anger, happiness, and sadness. In addition, the EMRs when the first emotion is anger and the second emotion is sadness are higher than those when the order was reversed. This shows that the difficulty of recognition may change depending on the order in which emotions are expressed. Overall, the proposed methods improve the EMR compared with the conventional method, indicating their effectiveness in improving the performance of the frame-based SER models.



(a) Conv. (b) Prop. I (c) Prop. II (d) Prop. III

Figure 4: Average EMR for each emotion pair in the evaluation data (wav2vec2.0+FC; Blue means lower values, and yellow means higher values).



(a) Conv. (b) Prop. I (c) Prop. II (d) Prop. III

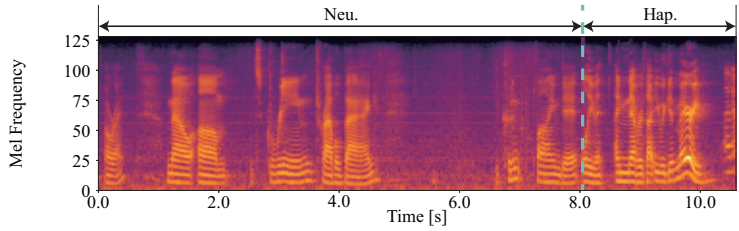
Figure 5: Average EMR for each emotion pair in the evaluation data (Hubert+FC; Blue means lower values, and yellow means higher values).

Figures 6 illustrate examples of the output from models trained by each method for wav2vec2+FC, which had high frame-level evaluation results overall. For the evaluation, we used the utterance which emotion changes from "neutral" to "happiness." Figure 6a illustrate the mel spectrogram of the evaluation utterance, and Figures 6b through 6e illustrate the log-likelihood of the model trained by each method. In addition, the blue dotted line in Figure 6a and the orange dotted line in Figures 6b through 6e illustrate the emotion changing point in the correct emotion label sequence. Therefore, the emotion before the dotted line is "neutral" and the emotion after the dotted line is "happiness."

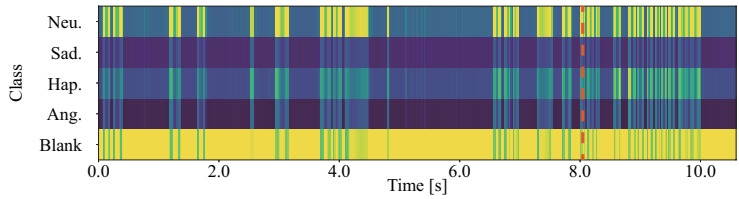
Comparing Figures 6b through 6e shows that in each case, the model predicted the emotion label sequence in the spoken interval and blank symbols in the unspoken interval. These results suggested that models might have been trained to predict emotions by focusing on spoken interval. Also, as the number of phoneme class attributes increases, the class with the highest log-likelihood changed before and after the dotted line. In particular, Figure 6e illustrates that the emotions class with high log-likelihood changed from "neutral" to "happiness" on both sides of the dotted line. Therefore, it can be said that the fine-grained and accurate emotion recognition became more possible than the conventional method by considering the acoustic difference of each phoneme.

6 Conclusion

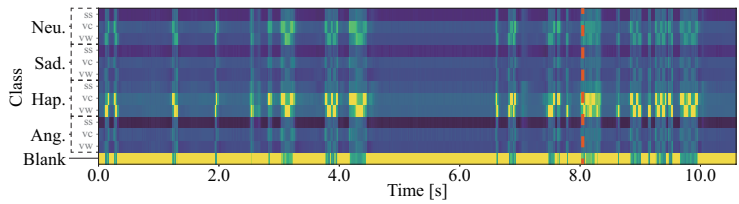
In this study, we proposed the training methods for frame-based SER models using emotion label sequences with phoneme class attributes, which were not considered in previous studies, and compared them with conventional methods. As a result, we confirmed that the proposed methods improve the accuracy of frame-based SER at both the utterance and frame levels. Proposed method II could improve WA by 2.4% and UA by 8.2% compared with the conventional method used in a previous study. In addition, proposed method III with wav2vec2.0+FC could improve EMR by 2.4% compared with the conventional method. We also found that considering vowels, voiced consonants, unvoiced consonants, and special symbols enabled models to recognize the fine-grained emotion changes within the utterance. The frame-based SER model trained by the proposed method is effective for recognizing emotions that change during utterances. The following are the tasks to be addressed in the future. First, we will experiment with other datasets such as MSP-Potcast [22] and BIIC-Podcast [35] to confirm the effectiveness of the proposed methods. Since it has been suggested that the performance of g2p affects the performance of automatic speech recognition and speech synthesis [29, 7], we also need to investigate the effect of g2p in frame-level SER. Furthermore, we will investi-



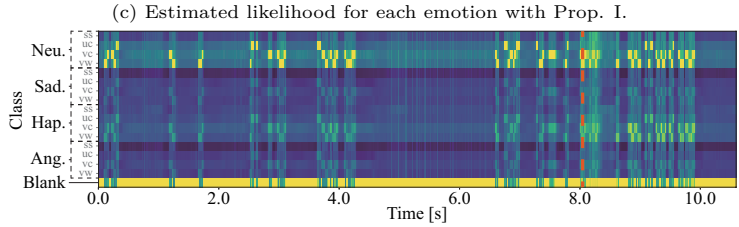
(a) Mel spectrogram.



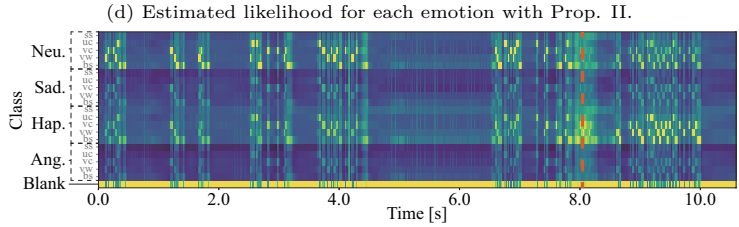
(b) Estimated likelihood for each emotion with Conv.



(c) Estimated likelihood for each emotion with Prop. I.



(d) Estimated likelihood for each emotion with Prop. II.



(e) Estimated likelihood for each emotion with Prop. III.

Figure 6: Examples of predictions from models trained by conventional and proposed methods. (The blue dotted line and orange dotted line illustrate the emotion changing points. In Figure 6a, black means lower values, and yellow means higher values. In Figure 6b through 6e, blue means lower values, and yellow means higher values.)

gate how the duration of each phoneme class attribute and other units, such as syllables, affects frame-level SER. Finally, we will conduct a subjective evaluation of the frame-level SER to determine whether the emotion sequence and waveform are properly aligned.

Biographies

Ryotaro Nagase received his B.E. and M.E. and Ph.D. degrees from Ritsumeikan University in 2020, 2022 and 2025, respectively. From 2022 to 2025, he was a Ritsumeikan Advanced Research Academy (RARA) Student fellowship. He is currently a specially appointed assistant professor at Ritsumeikan University. His current research interests include speech emotion recognition. He is a member of IEEE and ASJ.

Takahiro Fukumori received the B.E., M.E., and Ph.D. degrees from Ritsumeikan University in 2010, 2012, and 2015, respectively. From 2012 to 2015, he was a JSPS research fellowship for young scientists (DC1). From 2015 to 2020, he was an assistant professor with Ritsumeikan University. From 2020 to 2025, he was a lecturer with Ritsumeikan University. His research interests include speech recognition and speech enhancement. He is a member of IEEE, ASJ, IEICE, and IPSJ.

Yoichi Yamashita received his B.E., M.E. and Ph.D. degrees from Osaka University in 1982, 1984 and 1993, respectively. He has worked for the Institute of Scientific and Industrial Research of Osaka University as a Technical Official, a Research Associate, and an Assistant Professor from 1984 to 1997. He has worked for Ritsumeikan University as an Associate Professor and a Professor in the College of Science and Engineering and the College of Information Science and Engineering from 1997 to 2025. He continuously works for Ritsumeikan University as a fixed-term Professor. His research interests include speech recognition, speech synthesis, acoustic signal processing, and spoken document processing. He is a member of IEICE, ASJ, IPSJ, ISCA and IEEE.

References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", in *Proc. NIPS'20 – 34th International Conference on Neural Information Processing Systems*, No. 1044, Vancouver, BC, Canada, December 2020, 12449–60.

[2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.", *Journal of personality and social psychology*, 70(3), 1996, 614–36, DOI: 10.1037/0022-3514.70.3.614.

[3] M. Bojani, V. Deli, and A. Karpov, "Call Redistribution for a Call Center Based on Speech Emotion Recognition", *Applied Sciences*, 10(13), 2020, ISSN: 2076-3417, DOI: 10.3390/app10134653.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database", *Language Resources and Evaluation*, 42(4), 2008, 335–59, DOI: 10.1007/s10579-008-9076-6.

[5] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning", in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czech, September 2021, 4508–12, DOI: 10.21437/Interspeech.2021-1852.

[6] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition", *IEEE Signal Processing Letters*, 25(10), 2018, 1440–4, DOI: 10.1109/LSP.2018.2860246.

[7] S. Cheng, P. Zhu, J. Liu, and Z. Wang, "A Survey of Grapheme-to-Phoneme Conversion Methods", *Applied Sciences*, 14(24), 2024, ISSN: 2076-3417, DOI: 10.3390/app142411790.

[8] Y. Chiba, T. Nose, and A. Ito, "Multi-Stream Attention-Based BLSTM with Feature Segmentation for Speech Emotion Recognition", in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, October 2020, 3301–5, DOI: 10.21437/Interspeech.2020-1199.

[9] "Extracting salient sublexical units from written texts: "Emophon," a corpus-based approach to phonological iconicity", *Frontiers in Psychology*, 4, 2013, DOI: 10.3389/fpsyg.2013.00654.

[10] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition", *Neural Networks*, 92, 2017, 60–8, DOI: 10.1016/j.neunet.2017.02.013.

[11] Y. Gao, Z. Pan, H. Wang, and G. Chen, "Alexa, My Love: Analyzing Reviews of Amazon Echo", in *Proc. SmartWorld/SCALCOM /UIC/ATC/CBDCom/IOP/SCI – 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, Guangzhou, China, October 2018, 372–80, DOI: 10.1109/SmartWorld.2018.00094.

[12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", in *Proc. ICML'06 – 23rd International*

Conference on Machine Learning, Pittsburgh, Pennsylvania, USA, June 2006, 369–76, DOI: 10.1145/1143844.1143891.

[13] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards Temporal Modelling of Categorical Speech Emotion Recognition", in *Proc. INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, September 2018, 932–6, DOI: 10.21437/Interspeech.2018-1858.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29, 2021, 3451–60, DOI: 10.1109/TASLP.2021.3122291.

[15] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-Light: A Benchmark for ASR with Limited or No Supervision", in *Proc. ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, May 2020, 7669–73, DOI: 10.1109/ICASSP40776.2020.9052942.

[16] S. Khorram, M. G. McInnis, and E. M. Provost, "Jointly Aligning and Predicting Continuous Emotion Annotations", *IEEE Transactions on Affective Computing*, 12(4), 2021, 1069–83, DOI: 10.1109/TAFFC.2019.2917047.

[17] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition", in *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 79, 2020, Proceedings*, St. Petersburg, Russia, October 2020, 267–78, DOI: 10.1007/978-3-030-60276-5_27.

[18] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes", in *Proc. INTERSPEECH 2004 – ICSLP, 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, October 2004, 889–92, DOI: 10.21437/Interspeech.2004-322.

[19] W. Li, Y. Zhang, and Y. Fu, "Speech Emotion Recognition in E-learning System Based on Affective Computing", in *Proc. ICNC 2007 – Third International Conference on Natural Computation*, Vol. 5, Haikou, China, August 2007, 809–13, DOI: 10.1109/ICNC.2007.677.

[20] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning", in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, September 2019, 2803–7, DOI: 10.21437/Interspeech.2019-2594.

[21] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the Variance of the Adaptive Learning Rate and Beyond", in *Proc. ICLR – 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, April 2020.

[22] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings", *IEEE Transactions on Affective Computing*, 10(4), October 2019, 471–83, DOI: 10.1109/TAFFC.2017.2736999.

[23] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech Emotion Recognition Using Self-Supervised Features", in *Proc. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, May 2022, 6922–6, DOI: 10.1109/ICASSP43922.2022.9747870.

[24] R. Nagase, T. Fukumori, and Y. Yamashita, "Speech Emotion Recognition by Estimating Emotional Label Sequences with Phoneme Class Attribute", in *Proc. INTERSPEECH 2023 – 24th Annual Conference of the International Speech Communication Association*, Dublin, Ireland, August 2023, 4533–7, DOI: 10.21437/Interspeech.2023-1163.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books", in *Proc. ICASSP 2015 – 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, QLD, Australia, April 2015, 5206–10, DOI: 10.1109/ICASSP.2015.7178964.

[26] J. Parry, E. DeMattos, A. Klementiev, A. Ind, D. Morse-Kopp, G. Clarke, and D. Palaz, "Speech Emotion Recognition in the Wild using Multi-task and Adversarial Learning", in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, September 2022, 1158–62, DOI: 10.21437/Interspeech.2022-10581.

[27] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings", in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czech, September 2021, 3400–4, DOI: 10.21437/Interspeech.2021-703.

[28] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice", *American scientist*, 89(4), 2001, 344–50, DOI: 10.1511/2001.28.344.

[29] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks", in *Proc. ICASSP2015 – 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Queensland, Australia, April 2015, 4225–9, DOI: 10.1109/ICASSP.2015.7178767.

[30] J. A. Russell, "A circumplex model of affect.", *Journal of personality and social psychology*, 39(6), 1980, 1161–78, DOI: 10.1037/h0077714.

[31] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms", in *Proc. INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, August 2017, 1089–93, DOI: 10.21437/Interspeech.2017-200.

[32] K. R. Scherer, "Vocal affect expression: a review and a model for future research.", *Psychological bulletin*, 99(2), 1986, 143–65, DOI: 10.1037/0033-2909.99.2.143.

[33] M. Sharma, "Multi-Lingual Multi-Task Speech Emotion Recognition Using wav2vec 2.0", in *Proc. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, May 2022, 6907–11, DOI: 10.1109/ICASSP43922.2022.9747417.

[34] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition", in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, October 2020, 3755–9, DOI: 10.21437/Interspeech.2020-1212.

[35] S. G. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. N. Salman, C. Busso, and C.-C. Lee, "An Intelligent Infrastructure Toward Large Scale Naturalistic Affective Speech Corpora Collection", in *Proc. ACII2023 – 11th International Conference on Affective Computing and Intelligent Interaction*, Cambridge, Massachusetts, USA, September 2023, 1–8, DOI: 10.1109/ACII59096.2023.10388175.

[36] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common", *Frontiers in Psychology*, 4, 2013, DOI: 10.3389/fpsyg.2013.00292.

[37] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-Art Natural Language Processing", in *Proc. EMNLP – 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, October 2020, 38–45, DOI: 10.18653/v1/2020.emnlp-demos.6.

[38] Y. Wu, Z. Zhang, P. Peng, Y. Zhao, and B. Qin, "Leveraging Multi-Modal Interactions among the Intermediate Representations of Deep Transformers for Emotion Recognition", in *Proc. MuSe'22 – 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, Lisboa, Portugal, October 2022, 101–9, DOI: 10.1145/3551876.3554813.

[39] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding", in *Proc. INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, September 2018, 3688–92, DOI: 10.21437/Interspeech.2018-1811.

[40] Z. Zhao, Y. Wang, and Y. Wang, "Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition", in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, September 2022, 4725–9, DOI: 10.21437/Interspeech.2022-10230.

[41] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, "Attention-Enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition", in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, September 2019, 206–10, DOI: 10.21437/Interspeech.2019-1649.

[42] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B. W. Schuller, "Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition", *Neural Networks*, 141, 2021, 52–60, ISSN: 0893-6080, DOI: 10.1016/j.neunet.2021.03.013.