INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech Emotion Recognition by Estimating Emotional Label Sequences with Phoneme Class Attribute

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

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

is0368sh@ed.ritsumei.ac.jp, fukumori@fc.ritsumei.ac.jp, yyama@is.ritsumei.ac.jp

Abstract

In recent years, much research has been into speech emotion recognition (SER) using deep learning to predict emotions conveyed by speech. We studied the method that detected the emotion for the whole utterance using the frame-based SER, which estimates emotions in each frame rather than in a whole utterance. One of the problems with this method is that the emotional label sequence, which is used in training the frame-based SER, does not sufficiently consider phonemic characteristics. To solve this problem, we propose new methods of recognizing the emotion for the whole utterance using frame-based SER that considers the phoneme class attribute such as vowels, voiced consonants, unvoiced consonants, and other symbols in training. As a result, we found that the proposed methods significantly improve the performance of the result for the whole utterance compared to conventional methods.

Index Terms: speech emotion recognition, deep learning, emotional label sequence, phoneme class attribute

1. Introduction

Speech emotion recognition (SER) is a technique of predicting the emotion conveyed by speech. This technique can be applied to call center automation [1], mental health analysis [2], and e-learning systems [3]. There are two primary representation schemes of emotion. One is categorical emotion, which is the class of emotions, such as happiness and anger. The other is dimensional emotion, which is expressed as a score on the axis, such as arousal or valence. In this study, we examined speech classification into categorical emotion.

Researchers have proposed many methods using deep learning to improve the performance of SER. 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 3s [4]. Li et al. also proposed a method of training networks combining CNNs, BLSTM, and self-attention by multitask learning of emotion and gender classification [5]. Other methods use various input features, model structures, and training strategies [6, 7, 8]. In recent years, methods have also been proposed using pretrained self-supervised learning (SSL) models representing speech information. For instance, Pepino et al. proposed a method to learn BLSTM using the embedded representation obtained from pretrained SSL models [9]. They showed that it might be more effective for SER than 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 [10]. Many other methods use pretrained SSL models, which are highly effective for improving SER's performance [11, 12, 13, 14, 15]. The above methods train a model to predict emotion per utterance using one emotional label. We call this model utterance-based SER in this paper. Note that the prediction result per utterance obtained from the model is used in the evaluation.

However, because the expression of emotion often changes in an utterance, the emotional label should be given for each frame rather than an entire utterance. Therefore, in some previous reports, methods have been proposed to train a model using emotional label sequences that considered different emotional expressions for each frame. We call this model frame-based SER in this paper. For instance, Fayek et al. proposed training the frame-based SER that considered silent frames within an utterance. This method uses emotional label sequences, including emotional and silent labels in training the model [16]. Han et al. also proposed training a connectionist temporal classification (CTC) model of SER reflecting the feature of voiced phonemes. The method uses emotional label sequences constructed under the condition that voiced phonemes indicate emotional states and other symbols indicate non-emotional states [17]. Note that these methods aggregate the output of each frame and evaluate the prediction result per utterance. The challenge of the conventional method for frame-based SER is how to describe the emotional label of the frames in training the model. In the conventional method, it was assumed that only voiced phonemes express emotional states. However, the acoustic differences between vowels and voiced consonants, or those of each phoneme between emotions, were not assumed. Moreover, the possibility that unvoiced phonemes and other symbols represent emotional states was not considered. Other studies [18, 19, 20] suggested that emotions are expressed differently for each phoneme and symbol. Thus, we consider that reflecting phoneme class differences, such as vowels, voiced consonants, unvoiced phonemes, and other symbols in the emotional label sequence make it possible to train the model of SER more effectively.

We propose new methods of predicting the emotion for the whole utterance using the frame-based SER. This emotion is called the utterance-level emotion in this paper. The proposed methods include new training schemes of frame-based SER that is introduced the phoneme class attribute to the emotional label sequence to consider the phoneme-dependent acoustic diversity of the input frame. This research contributes to realizing frame-based SER that recognizes differences in the attribute of phoneme class and outperforms the performance of conventional methods. This paper proceeds as follows. In Section 2, we describe the conventional method. In Section 3, we present the proposed methods for frame-based SER. 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.

10.21437/Interspeech.2023-1163

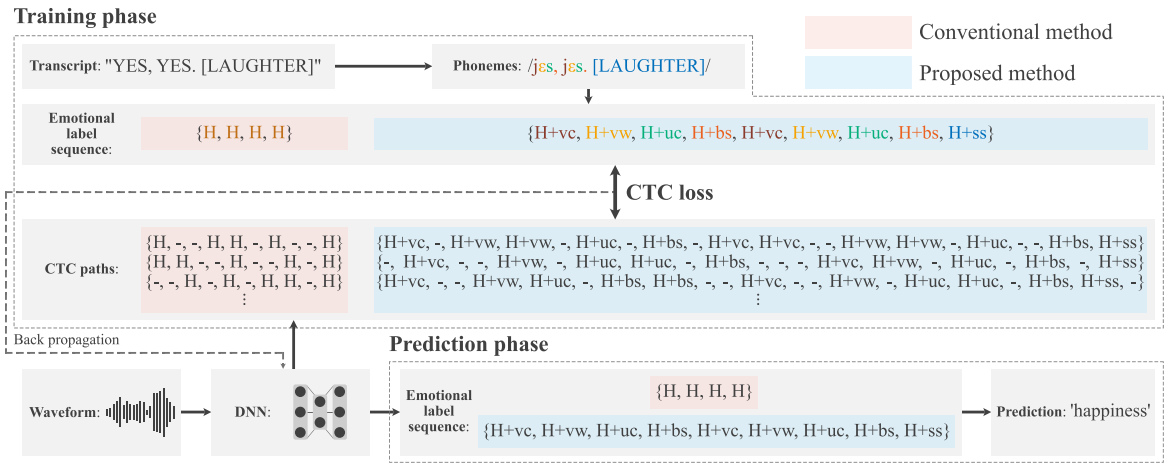


Figure 1: *Illustration of training the frame-base SER by estimating emotional label sequences*

2. SER by estimating emotional label sequences

Frame-based SER is trained by estimating emotional label sequences. In a previous study [17], an emotional label sequence is constructed from the utterance-level emotional 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 and predicting the emotion is shown in Figure 1. In the training phase, the transcript corresponding to the input utterance is converted into phonemes, and the numbers of vowels and voiced consonants are counted the basis of the CMU Pronouncing Dictionary[1]. The emotional label sequence consists of as many emotional labels as the number of voiced phonemes. For example, the utterance "YES, YES. [LAUGHTER]" (/jɛs, jɛs. [LAUGHTER]/) in the emotion of 'happiness' (H) has four voiced phonemes, including /j/ and /ɛ/. Thus, this transcript is converted into {H, H, H, H}. The methods use various network structures for estimating emotional label sequences. For example, there are models that combines BLSTM and various attention mechanisms [21] and models that combines parallel CNNs, squeeze-and-excitation network (SENet), and dilated residential network (DRN) [22]. The model for SER is trained basis of CTC. CTC is a framework for estimating paths (CTC paths) containing blank symbols (-) and repeated symbols [23]. 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. In CTC-based learning, given an input $\mathbf{x} = [x_0, \ldots, x_{\mathrm{T}}]$ of length T, we maximize the probability of obtaining an emotional 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 be the emotional label of time t, π be the CTC path for the emotional label sequence, and $\Phi(y)$ be the set of π. Moreover, let U be the set of training data.

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

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

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

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 emotional label sequence is obtained by gathering estimated symbols within the CTC path by deleting blank symbols and merging repeated characters. The final predicted utterance-level result is the most frequent emotion in the sequence.

3. SER by estimating emotional label sequences with phoneme class attribute

We propose methods of training the frame-based SER considering the phoneme class attribute. 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 include punctuation marks, such as '.', '!', '?'. Special symbols are unique information for each dataset, such as '[BREATHING]'. Lee et al. have discussed that vowels are necessary for SER performance [18]. Aryani et al. have also suggested that voiced or unvoiced consonants may express various emotions [19]. As for other symbols, previous studies [14, 20] have shown that inputting embedded representations of phonemes, including silent and special symbols, into models improves performance. Therefore, the proposed methods of estimating emotional label sequences considers the attributes of vowels, voiced and unvoiced consonants, and other symbols. Any symbol not falling under the above attributes is considered a non-emotion state. In Figure 1, the proposed methods construct an emotional label sequence by combining the phoneme class attribute and emotional labels and uses it for training the model by CTC. A significant difference from the conventional method is that the emotional 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, two voiced consonants, two unvoiced consonants, and one special symbol. 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 'emotional label+phoneme class attribute' is an emotional 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). In the training phase, the DNN model is trained using emotional label sequences with the phoneme class attribute. In the prediction phase, the emotional label sequence is obtained

Table 1: *Correspondence 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]

Table 2: *Numbers of classes in different methods*

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

from the estimated CTC path, and the emotion with the high frequency of appearance is the prediction result.

4. Experiments

4.1. Dataset

We used the interactive emotional dyadic motion capture database (IEMOCAP), which contains emotional speech in English [24]. The IEMOCAP database provides an audiovisual emotional dataset that contains speech and facial and hand movements during dialogue. It consists of five sessions, each containing a dialogue between a man and a woman. This dataset contains 10,039 utterances, 5,255 of which are acted utterances and 4,784 of which are improvised dialogues. The utterances are assigned to ten emotional labels: neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and others. In this experiment, we used only improvised dialogues, and we developed and evaluated the model of predicting four emotion categories: anger, happiness, sadness, and neutral. We also added the utterances of excitement to those of happiness. Therefore, the number of utterances for each emotion totaled 2,943 (Anger, 289; Happiness, 947; Sadness, 608; Neutral, 1,099). The average length of utterances used for the experiment was about 4.5s. In the training phase, utterances of 15s or less were only used.

Phoneme sequences necessary for constructing emotional label sequences were obtained using a grapheme-to-phoneme (g2p) conversion toolkit[2]. Table 1 shows the correspondence between the phoneme class attribute and symbols. The numbers of bs, vw, vc, uc, and ss in the dataset were 37,304, 40,474, 37,388, 21,791, and 133, respectively. Moreover, the number of voiced phonemes is 77,862. In this experiment, to investigate the effect of the proposed methods, we experimentally evaluated with three methods, each considering different phoneme class attributes. Table 2 shows the phoneme class attributes considered in the conventional and the proposed methods and the number of estimated classes in each frame of the CTC path.

The evaluation method was ten fold cross-validation without speaker overlap. We split the dataset into eight speakers for the training data, one speaker for the validation data, and one speaker for the testing data in each fold.

4.2. Models and metrics

We use the conventional frame-based SER as comparative methods [17, 21, 22]. We also set up another comparative method based on wav2vec2.0 trained by the conventional method. Wav2vec2.0 is one of the self-supervised representation learning frameworks [25]. It trains the representation of speech by introducing quantization and Transformer layers into conventional wav2vec. We used the pretrained model of wav2vec2.0[3] provided by Hugging Face [26]. This model was pretrained in speech representation learning and fine-tuned in ASR with Libri-Light and Librispeech, which are English speech datasets. The model used for the comparative method was constructed by combining one fully connected (FC) layer with wav2vec2.0 consisting of seven CNN layers and 24 Transformer layers. In the training phase, the model parameters of the CNN layers were fixed, and the Transformer layers and the FC layer were fine-tuned. The proposed methods also used the same model based on wav2vec2.0. The inputs to the model were speech waveforms, and the outputs were the CTC paths of emotional label sequences or emotional label sequences with the phoneme class attribute. The number of epochs was 50, the batch size was 8, the learning rate was 0.0001, and the optimization method was RAdam [27]. In the training phase, we added gradient clipping with a threshold of 5.0. The model calculated CTC loss.

We evaluated the model performance for estimating emotional label sequences. The metric used in this evaluation was the token error rate (TER) in emotional label sequences. TER was calculated for tokens with phoneme classes and emotions and tokens with only emotions. TER is given in Equation 3. Let S be the number of substitutions, D be the number of deletions, I be the number of insertions, and N be the length of the correct phoneme class sequence.

$$\mathrm{TER} = \frac{S + D + I}{N} \tag{3}$$

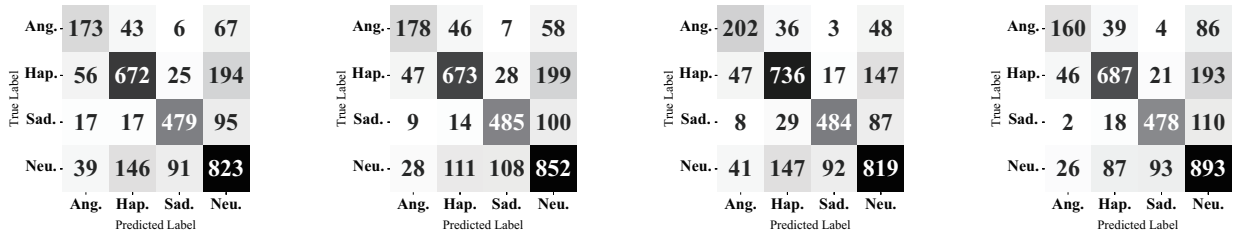
The lower TER is, the more correctly the model estimates each token. We also evaluated the model performance for recognizing emotions per utterance. The metrics used in this evaluation were the weighted accuracy (WA) and the unweighted accuracy (UA). When nothing was estimated, we regarded the emotion of input utterance as 'neutral.' The higher WA and UA are, the more correctly the model predicts emotions. We compared the TER, WA, and UA of the proposed methods with those of the conventional methods.

4.3. Results

Table 3 shows TER of emotional label sequences with the phoneme class attribute in each method. Note that all tokens contain phoneme classes and emotions, and emotional tokens contain emotions only. A comparison of all methods shows that the proposed methods show lower TERs of all tokens or only emotional tokens than the conventional method. In particular, proposed method II improved with TER of all tokens by 11.2% and TER of only emotional tokens by 14.5% over the conventional method. These results indicate that the proposed methods can recognize acoustic differences of the phoneme class at-

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

[3] facebook/wav2vec2-large-960h-lv60-self



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

Figure 2: *Confusion matrices in each method used fine-turned wav2vec2.0*

Table 3: *TER of each method used fine-turned wav2vec2.0*

	TER (%)	
	All tokens	Emotional tokens
Conv.	39.7	39.7
Prop. I	29.7	27.1
Prop. II	**28.5**	**25.2**
Prop. III	31.3	27.0

Table 4: *WA and UA of each method*

Conventional methods	WA (%)	UA (%)
BLSTM [17]	64.2	65.7
BLSTM+Component Attention [21]	69.0	67.0
PCNSE+SADRN [22]	73.1	66.3
Wav2vec2.0+FC (conv.)	73.4	70.3
Proposed methods		
Wav2vec2.0+FC (prop. I) [Ours]	74.6	73.0
Wav2vec2.0+FC (prop. II) [Ours]	**76.1**	**75.7**
Wav2vec2.0+FC (prop. III) [Ours]	75.6	72.2

tribute or the emotion in a sequence rather than the conventional method.

Table 4 shows WA and UA for each method. The performance with wav2vec2.0 and FC layer is comparable to or better than that in previous studies. This model improved UA more than the system in previous studies by about 3.3%. This result shows that fine-tuned wav2vec2.0 effectively improves the performance of frame-based SER.

Comparing the conventional methods and the proposed methods shows that all the proposals are more effective in improving the WA and UA than the conventional methods. Proposed method I improves WA by 1.2% and UA by 2.7% compared with the conventional method using fine-tuned wav2vec2.0. By distinguishing voiced phonemes into vowels and voiced consonants, the model in proposed method I could recognize the emotions corresponding to phonemes with the acoustic differences in detail. Proposed method II improves WA by 2.7% and UA by 5.4% compared with the conventional method using fine-tuned wav2vec2.0. In particular, proposed method II improves WA by 3.0% and UA by 9.4% compared with conventional methods in a previous study [22]. By considering unvoiced consonants, the model may have recognized emotions corresponding to phonemes in which the speed of breath and the rise of tone in speech differ. Proposed method III improves WA by 2.2% and UA by 1.9% compared with the conventional method using fine-tuned wav2vec2.0. By comparing all the proposed methods, we found that the performance improvement of proposed method III is slight. The improvement of UA in proposed method III was one-third of that in proposed method II. This result shows that considering basic symbols may adversely affect the performance of SER using emotional label sequences.

Figure 2 shows the confusion matrices in the methods using fine-tuned wav2vec2.0. In the comparison between the conventional method and proposed method I, the number of correct answers for each emotion increased slightly in proposed method I. In particular, the number of data in 'neutral' increased by 29. This result indicates that distinguishing between voiced consonants and vowels may improve the performance of recognizing utterances as 'neutral'. In the comparison between the conventional method and proposed method II, the number of correct answers other than 'neutral' increased significantly. In addition, the number of correct answers for 'happiness' increased by 64, and the number of mistakes in recognizing 'happiness' as 'neutral' decreased by 47. For example, a utterance containing five unvoiced consonants in 'happiness', e.g., "OH, THEN YOU HAD A WHOLE WEEKEND OF CAMPING?" was mistakenly recognized as 'neutral' by the conventional method and correctly recognized by the proposed method II. This result shows that proposed method II considered both voiced and unvoiced phonemes can reduce mistakes in recognizing 'happiness' as 'neutral'. In the comparison between the conventional method and proposed method III, the number of correct answers for 'neutral' increased by 70, whereas those for other emotions remained almost unchanged. This result indicates the possibility that using basic symbols increases the number of correctly classified utterances as 'neutral.'

5. Conclusion

In this study, we investigated SER using emotional label sequences. We proposed methods for training models using emotional label sequences with the phoneme class attribute, which was not considered in previous studies, and compared them with conventional methods. The results showed that the proposed methods outperformed all conventional methods in terms of phoneme class or emotional TER, WA, and UA. In particular, proposed method II could improve WA by 3.0% and UA by 9.4% compared with the latest conventional methods. We also found that considering vowels, voiced consonants, unvoiced consonants, and special symbols significantly reduced the number of mistakes in recognizing 'happiness' as 'neutral' and enabled models to recognize the changes within the utterance. In the future, we will perform experiments using other pretrained models and datasets in different languages. Moreover, because we evaluated the proposed methods using the data assigned with utterance-level emotional labels in this experiment, we will verify the effectiveness of our proposed methods using data in which emotions change in utterances.

6. References

[1] M. Bojanić, V. Delić, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Applied Sciences*, vol. 10, no. 13, 2020.

[2] 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, Oct. 2018, pp. 372–380.

[3] 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, Aug. 2007, pp. 809–813.

[4] 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, Aug. 2017, pp. 1089–1093.

[5] 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, Sep. 2019, pp. 2803–2807.

[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*, vol. 25, no. 10, pp. 1440–1444, 2018.

[7] 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, Oct. 2020, pp. 3301–3305.

[8] 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, Sep. 2022, pp. 1158–1162.

[9] 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, Sep. 2021, pp. 3400–3404.

[10] 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, Sep. 2021, pp. 4508–4512.

[11] 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, Oct. 2020, pp. 3755–3759.

[12] 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, pp. 6922–6926.

[13] 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 (ICASSP)*, Singapore, Singapore, May. 2022, pp. 6907–6911.

[14] 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, Sep. 2022, pp. 4725–4729.

[15] 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, Oct. 2022, pp. 101–109.

[16] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[17] 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, Sep. 2018, pp. 932–936.

[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, Oct. 2004, pp. 889–892.

[19] A. Aryani, M. Conrad, and A. Jacobs, "Extracting salient sublexical units from written texts: "Emophon," a corpus-based approach to phonological iconicity," *Frontiers in Psychology*, vol. 4, 2013.

[20] 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, Sep. 2018, pp. 3688–3692.

[21] 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, Sep. 2019, pp. 206–210.

[22] 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*, vol. 141, pp. 52–60, 2021.

[23] 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, Jun. 2006, pp. 369–376.

[24] 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*, vol. 42, no. 4, pp. 335–359, 2008.

[25] 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*, Vancouver, BC, Canada, Dec. 2020.

[26] 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, Oct. 2020, pp. 38–45.

[27] 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, Apr. 2020.