INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining

Yuan Gao, Chenhui Chu, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{gao,kawahara}@sap.ist.i.kyoto-u.ac.jp chu@nlp.ist.i.kyoto-u.ac.jp

Abstract

This paper addresses effective pretraining of automatic speech recognition (ASR) and gender recognition to improve wav2vec 2.0 embedding for speech emotion recognition (SER). Specifically, we propose a two-stage finetuning method, which first pretrains the self-supervised learning (SSL) model with ASR to learn the linguistic information and address the gradient conflict problem of conventional multi-task learning. Experimental results on the IEMOCAP dataset show that ASR pretraining can significantly outperform the simple MTL with ASR, and thus demonstrate the effectiveness of the two-stage finetuning method. We also investigate how to combine gender recognition with ASR pretraining to derive more effective embedding for SER. As the upper layers of the SSL model are focused on ASR, incorporating skip-connection can effectively embed the gender information. Compared with the single-task learning baseline, our method achieves a UA of 76.10% with an absolute improvement of 3.97%.

Index Terms: speech emotion recognition, human-computer interaction, wav2vec 2.0, multi-task learning

1. Introduction

Understanding the emotional state from speech is crucial for conversational robots and intelligent devices to generate empathetic responses to the user [1]. Therefore, speech emotion recognition (SER) has become an active research area to promote the development of natural and friendly HCI. Traditional SER systems used hand-crafted features, designed by experts on acoustic and speech signal processing, as the input of statistical machine learning models such as support vector machines [2, 3]. With the development of deep learning, researchers attempt to learn emotional representations from speech using deep neural networks such as convolutional neural network (CNN) [4, 5] and long short-term memory network (LSTM) [6, 7].

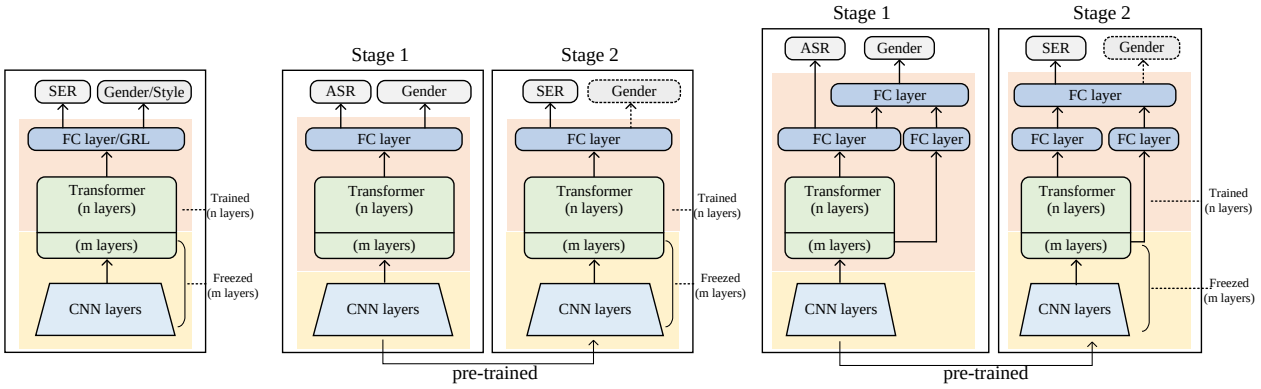
Although deep neural networks can provide promising performance in speech signal processing, data sparsity limits the performance of the supervised SER models [8, 9]. Since collecting emotional speech with human-annotated labels is time-consuming, the existing public emotional datasets are insufficient to train a robust supervised learning model. One potential solution for this problem is pretraining a model with self-supervised learning (SSL) using a large amount of unlabeled speech data and finetuning with the contextual and acoustic information for SER. In recent years, SSL models like wav2vec 2.0 [10] have been widely used for many speech processing tasks [11, 12]. This model is first pretrained by unlabeled speech data, and then finetuned by the target datasets for many downstream tasks, such as low-resource automatic speech recognition (ASR) [13] and speech translation [14]. Compared with conventional supervised learning models, it can achieve promising performance with only a small amount of the target data. Prior works have incorporated SSL models to extract a robust feature representation for SER [15, 16, 17]. In [16], Pepino et al. proposed a transfer learning method and used the SSL embedding using a simple network for SER. They compared it with traditional CNN and LSTM for SER, and showed that the pretrained model can significantly improve the recognition performance on two emotional speech datasets.

Human emotion expressions are closely related to linguistic and acoustic information in speech [18]. As the size and the shape of the vocal tract is much different between male and female [19], including gender identification and speaker-related tasks can enhance the feature extraction process for SER. In [20], Li et al. investigated the effect of gender information on SER and showed that learning gender information can largely improve the performance of SER. More recently, Sharma et al. incorporated gender, language, and energy-related information using the multi-task learning (MTL) method to achieve rich transcription and improve the performance of SER on 25 datasets [21].

Moreover, it is widely known that using the linguistic information can benefit SER systems [18]. However, ASR of emotional speech is difficult compared with reading speech. Therefore, instead of using the output transcription, Feng et al. [22] proposed an end-to-end model using the hidden output of the ASR model, and Li et al. [23] proposed hierarchical co-attention to fuse the hidden output of wav2vec 2.0 for SER. Their experimental results demonstrate that combining the latent representation of ASR model can largely improve the performance of supervised learning-based SER model. Since the SSL models can effectively learn both the contextual information and emotional information from speech, Cai et al. [24] designed MTL of ASR and SER based on the wav2vec 2.0 model. They showed that joint training with ASR based on the SSL model led to better performance for SER.

Conventionally, researchers used MTL to include auxiliary tasks to promote the feature extraction of the main task, e.g. ASR for SER. However, their training objectives are much different, and thus the gradients of the different tasks often conflict if they point away from one another, i.e., have a negative cosine similarity [25]. Moreover, the optimization of ASR is more complicated, and the network tends to be focused on the ASR in MTL models. To address this problem, in this paper, we introduce a novel two-stage finetuning method for SSL-based SER. We first train the SSL model with auxiliary tasks including ASR and then finetune it for SER in the second stage to avoid the conflicting gradient problem. Furthermore, we also investigate the effective combination of gender recognition. Comparative ex-

10.21437/Interspeech.2023-756



(a) MTL and adversarial learning in single-stage finetuning

(b) ASR and gender pre-training in two-stage finetuning

(c) ASR and gender pre-training with skip-connection in two-stage finetuning

Figure 1: *Network structure of wav2vec 2.0-based models. In two-stage finetuning methods, we finetune all the Transformer layers in stage 1 and n Transformer layers (pink part) in stage 2.*

periments of this two-stage finetuning and MTL are conducted to demonstrate the effectiveness of our proposed approach.

2. Proposed Methods

Wav2vec 2.0 is a self-supervised representation learning framework consisting of 1) convolutional layers to extract a deep representation from raw audio, and 2) Transformer layers to learn the contextual information from the output of the CNN layers. This model is pretrained by a large amount of unlabeled data, then we can finetune it using the target data, and the output of the last Transformer layer can be used for SER. The proposed methods in this study are described in this Section.

2.1. MTL and adversarial learning for SER

As described in Section 1, linguistic and other attribute information can benefit emotional feature extraction. In this work, we incorporate ASR, gender, and emotional expression style (spontaneous vs. acted) as the auxiliary task in the MTL model. The overall loss function L of using all three auxiliary tasks can be defined as:

$$L = \alpha L_{emotion} + \frac{(1-\alpha)}{3}(L_{ASR} + L_{gender} + L_{style}) \quad (1)$$

where $L_{emotion}$, L_{ASR}, L_{gender}, L_{style} are the loss functions of emotion, ASR, gender, and style recognition tasks, and α is a weight parameter.

In this work, we also try adversarial learning and compare it with MTL. Instead of learning the corresponding information of the subtask, the adversarial learning model introduces a gradient reversal layer (GRL) to multiply a certain negative constant by the gradient of the subtask classification:

$$L = \gamma L_{emotion} + (1-\gamma)G(L_{gender/style}) \quad (2)$$

Here we use G to represent the GRL. This adversarial learning is applied to gender and style in this study. The MTL and adversarial learning applied to wav2vec 2.0-based SER is illustrated in Figure 1 (a).

2.2. Two-stage finetuning of wav2vec 2.0

MTL has been widely used for many related tasks including SER with ASR. However, the training objective and complexity of these tasks are much different. As a result, the gradient of different subtasks can conflict with each other, which may influence the recognition performance of the main task. Thus, we propose two-stage finetuning of the wav2vec 2.0 model. ASR and other auxiliary tasks are pretrained in the first stage of the proposed method so that the model can take advantage of the emotional-related information and address the gradient conflict problem. For the SSL model which is pretrained by unlabeled audio data, it is possible to finetune progressively with a different objective in each finetuning stage before the target task.

The architecture of the two-stage finetuning is shown in Figure 1 (b). In the first stage, we finetune the pretrained wav2vec 2.0 model to embed the linguistic and gender information. Note that we finetune all Transformer layers in the first stage to enhance the performance of the ASR task. Specifically, we feed the output of the last Transformer layer to a fully-connected layer (FC layer) with connectionist temporal classification (CTC) for ASR and cross entropy (CE) loss function for gender classification. The purpose of the first finetuning stage is to embed the attribute information to the wav2vec 2.0 model, and thus the SER task is not included in this stage. In the second stage, the pretrained model after the first finetuning stage is trained for SER. We freeze the parameter of the CNN feature encoder and the early m layers of the Transformer (yellow part in Figure 1) and finetune the top n Transformer layers. By this freezing of the early layers, the emotion recognition task can benefit from the information learned in the first stage.

2.3. Skip-connection for gender recognition

When we finetune ASR and gender recognition jointly using a SSL model, it is expected that the higher layers of the Transformer are focused on the linguistic information while gender information is learned with low-level features. Therefore, we incorporate a skip-connection to directly use the feature encoding of the gender information of early m Transformer layers, as shown in Figure 1 (c). In the first finetuning stage, we concatenate the output of m_{th} and the last Transformer layer to enhance the feature encoding of gender recognition. Then in the second stage, we freeze the lower m Transformer layers, to keep the gender information learned in the first stage, and train the top n layers of the Transformer for the SER task. For the two-stage

finetuning methods (Figure 1 (b) and (c)), we also compare the performance of single-task learning (STL) and MTL with gender recognition (denoted as dotted lines in Figure 1) in the second stage.

3. Experimental Setup

3.1. Database

In this study, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [26] to evaluate the proposed methods. This dataset contains emotional data of approximately 12 hours. Ten American English speakers were engaged in recording five speaker-independent dyadic sessions, and each session is performed by two speakers (one male and one female) with either a series of scripts or improvisational scenarios. For each speech utterance, three annotators assigned the emotional categorical labels. We adopt the common practice of merging "happy" and "excited" into one emotion class titled "happy" [8], resulting in 5,531 utterances with four emotion classes: happy (1,636), sad (1,084), angry (1,103), and neutral (1,708).

3.2. Implementations

We implemented the proposed model using PyTorch and the Huggingface Transformers repository [27]. Following previous studies [16, 23, 28, 29, 30, 31], we choose the wav2vec-2.0-base model, which was pretrained with LibriSpeech of 960h [32] without transcriptions. This model consists of a CNN feature extractor to map the raw audio to a latent representation and 12 layers of Transformer to learn the contextual information. For arbitrary length input of the SSL model, we pad all the sentences of each batch. During training, we froze the CNN feature extractor and finetuned the Transformer layers for 100 epochs. For the output of the last Transformer layer, we use mean pooling to reduce the dimension of the feature, and one FC layer with dropout is used for SER. The learning rate was 10e-5, and the mini-batch size was 16. In experiments with MTL and adversarial learning, the weight parameters of auxiliary tasks α and γ are 0.1. We use the unweighted accuracy (UA) and unweighted accuracy (WA) for the emotion classification task and the word error rate (WER) for ASR as the evaluation measure, respectively. In this work, we follow the common practice [8, 23] and report the performance of 5-fold speaker-independent cross-validation on the IEMOCAP dataset.

4. Experimental Results and Analysis

4.1. Finetuning wav2vec 2.0 for SER

We first evaluate the performance of wav2vec 2.0 in SER by comparing it with the conventional CNN-LSTM model [33, 34]. A previous study on wav2vec 2.0 [35] shows that partially finetuned SSL model (trained only Transformer layers and frozen CNN feature encoder) can generate comparable performance to the case finetuning all layers. To extend this research, we also make a preliminary study of finetuning different numbers of the Transformer layers for SER.

As shown in Table 1, the wav2vec 2.0 model significantly outperforms the supervised CNN-LSTM method by absolute 6.09% and 4.57% on UA and WA, even without finetuning. Moreover, finetuning eight Transformer layers of wav2vec-2.0-base can achieve a comparable result to the case finetuning all layers. Therefore, in the following experiments, we set m to 4 and n to 8.

Table 1: *Comparison of CNN-RNN and pretrained SSL model using single-stage finetuning. Num. layer denotes the number of finetuned Transformer layers. When Num. layer is 0, the wav2vec 2.0 model is not finetuned, and only the FC layer are trained for SER.*

Model	Num. layer	UA	WA
CNN-LSTM	-	55.47	55.81
Wav2vec 2.0	0	61.88	60.24
	2	63.94	62.70
	4	70.51	69.38
	6	70.24	68.69
	8	**72.30**	71.52
	10	72.05	71.19
	12	72.13	**71.55**

Table 2: *Comparsion of MTL and adversarial learning using single-stage finetuning.*

Method	Task	UA	WA
STL	SER	72.13	71.55
MTL	SER+ASR	73.15	72.70
	SER+gender	72.94	72.47
	SER+style	72.67	72.29
	SER+gender+style	72.39	71.33
	SER+ASR+gender	**74.16**	**73.85**
	SER+ASR+style	73.55	72.84
	SER+ASR+gender+style	73.56	73.08
Adversarial learning	SER+gender	72.75	71.94
	SER+style	72.33	71.28

4.2. Comparing MTL and adversarial learning in single-stage finetuning

Then, we conduct MTL for ASR, gender, and style recognition. Furthermore, we also compare the adversarial learning to explore whether eliminating the gender and style information can benefit SER. In this Section, we apply single-stage finetuning to all experiments. As shown in Table 2, compared with the STL baseline, incorporating ASR with SER using MTL improved the performance of SER. This result confirms that learning contextual information can benefit the SSL model in learning emotional information [24, 23]. Using the adversarial learning method achieved a similar UA but slightly degraded WA to the baseline STL model. Taking advantage of gender and style information by MTL can help the model to learn more discriminative features and enhance the SER task. In the single-stage finetuning methods, combining ASR and gender recognition can provide the best performance with a UA of 74.16%. These results demonstrate that learning both contextual and gender information together can benefit SER. However, we conclude that style and gender information shows no synergy in the SER task since jointly training the style and gender recognition leads to decreased performance compared with using only the gender recognition.

Table 3: *Results of two-stage finetuning methods. When the subtask is only used in the first finetuning stage, we denote $*$ to indicate the UA_gender or WER of the first stage.*

Model	Task		UA	WA	UA_gender	WER
	1st stage	2nd stage				
ASR Pretraining	ASR	SER	74.29	73.96	-	23.73*
	ASR	SER+ASR	74.61	73.55	-	23.36
	ASR	SER+gender	75.15	74.13	98.28	23.73*
Gender Pretraining	Gender	SER	64.30	61.74	97.54*	-
ASR & Gender Pretraining	ASR+gender	SER	74.54	73.17	97.13*	23.94*
	ASR+gender	SER+ASR	74.29	73.05	97.13*	22.52
	ASR+gender	SER+gender	74.31	72.80	97.28	23.94*
ASR & Gender (skip-connected) Pretraining	ASR+gender	SER	**76.10**	74.94	98.57*	22.80*
	ASR+gender	SER+ASR	76.06	75.01	98.57*	**22.19**
	ASR+gender	SER+gender	76.02	**75.17**	**98.92**	22.80*

Table 4: *Comparasion with previous works using wav2vec 2.0 embedding in the same setting.*

Approaches	Year	UA	WA
Pepino et al. [16]	2021	67.20	-
Xia et al [28]	2021	66.90	65.40
Keesing et al. [29]	2021	65.60	-
Zou et al. [30]	2022	69.80	71.05
Li et al.[23]	2022	-	63.40
Yue et al. [31]	2022	70.82	-
Ours	2023	**76.10**	**74.94**

4.3. Results of two-stage finetuning methods

Based on the results in the previous subsection, we use ASR and gender recognition for the proposed methods.

From Table 3, pretraining the model with ASR in the first stage consistently outperforms the single-stage finetuning results reported in Table 2. Among them, the best performance is achieved using ASR in the first stage, and SER and gender in the second stage. Compared with the single-stage MTL with the same auxiliary tasks, learning the contextual information and emotional information in different finetuning stages can effectively address the gradient conflict problem of MTL and achieved a 1.14% improvement in UA. However, learning only gender information in the first stage results in poor SER performance (64.30% on UA and 61.74% on WA, which is lower than the STL baseline of 72.13% and 71.55% on UA and WA, respectively). This indicates that ASR pretraining is crucial for SER in two-stage finetuning. Furthermore, although gender recognition was effective in Table 2, directly combining ASR with gender recognition in the first stage does not improve the SER performance. We hypothesize that since the training loss of the gender recognition task converges much easier than ASR, when it is jointly trained with ASR, the higher layers of the Transformer are mainly focused on ASR. Thus, pretraining of ASR and gender in the first finetuning stage does not have synergy and benefit SER. Therefore, we introduced a skip-connection for gender recognition so that the learned information in the early Transformer layers can benefit SER.

We report the results of using the skip-connection for gender recognition in the last rows of Table 3. Compared with the standard pretraining of ASR and gender, it achieved an absolute improvement of 1.56% and 1.77% on UA and WA, respectively. This result supports the effect of the skip-connection for gender recognition to be combined with ASR pretraining. It is also confirmed in Table 3 that finetuning ASR or gender recognition in the second stage does not help SER, but it results in the best performance in ASR and gender recognition.

4.4. Comparasion with previous studies

Finally, we compare the performance of the proposed method with the results reported in other recent studies using the same SSL model and the same setting. The compared publications and years are also provided in Table 4.

The proposed two-stage finetuning method outperforms other recent studies by more than 5.28% and 3.89% on UA and WA, respectively. This confirms that incorporating ASR and gender pretraining in the two-stage finetuning can enhance SER based on the SSL model.

5. Conclusion and Future Work

In this work, we first confirmed that learning the linguistic, gender, and style information using MTL can benefit the SER system. Then, we proposed a two-stage finetuning method, and showed that pretraining the SSL model using ASR is more effective than the simple MTL by addressing the gradient conflict problem. Moreover, we presented that skip-connected gender recognition effectively improved SER performance. Finally, the comparison with previous works using wav2vec 2.0 showed that our proposed approach outperformed state-of-the-art results. In the future, we plan to apply the proposed two-stage finetuning method to multi-lingual SER tasks.

6. Acknowledgment

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123, Grant-in-Aid for Scientific Research KAKENHI (JP19H05691, 20H00602 and 23H03454), and JST Moonshot R&D (JPMJPS2011).

7. References

[1] B. Luo, R. Y. Lau, C. Li, and Y.-W. Si, "A critical review of state-of-the-art chatbot designs and applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 1, p. e1434, 2022.

[2] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *2005 international conference on machine learning and cybernetics*, vol. 8. IEEE, 2005, pp. 4898–4901.

[3] A. K. Samantaray, K. Mahapatra, B. Kabi, and A. Routray, "A novel approach of speech emotion recognition with prosody, quality and derived features using svm classifier for a class of northeastern languages," in *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*. IEEE, 2015, pp. 372–377.

[4] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[5] N. Hajarolasvadi and H. Demirel, "3D-CNN-based speech emotion recognition using k-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, 2019.

[6] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*. IEEE, 2019, pp. 40–44.

[7] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[8] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[9] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6319–6323.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] A. Rouhe, A. Virkkunen, J. Leinonen, M. Kurimo *et al.*, "Low resource comparison of attention-based and hybrid asr exploiting wav2vec 2.0," in *Interspeech*. International Speech Communication Association, 2022, pp. 3543–3547.

[12] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.

[13] L. Borgholt, T. M. Tax, J. D. Havtorn, L. Maale, and C. Igel, "On scaling contrastive representations for low-resource speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3885–3889.

[14] C. Wang, A. Wu, J. Pino, A. Baevski, M. Auli, and A. Conneau, "Large-scale self-and semi-supervised learning for speech translation," *arXiv preprint arXiv:2104.06678*, 2021.

[15] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition," *arXiv preprint arXiv:2110.06309*, 2021.

[16] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[17] Y. Li, Y. Mohamied, P. Bell, and C. Lai, "Exploration of a self-supervised speech model: A study on emotional corpora," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 868–875.

[18] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2004, pp. I–577.

[19] A. P. Simpson, "Dynamic consequences of differences in male and female vocal tract dimensions," *The journal of the Acoustical society of America*, vol. 109, no. 5, pp. 2153–2164, 2001.

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

[21] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.

[22] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model." in *INTERSPEECH*, 2020, pp. 501–505.

[23] Y. Li, P. Bell, and C. Lai, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.

[24] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning." in *Interspeech*, vol. 2021, 2021, pp. 4508–4512.

[25] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[26] 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.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[28] Y. Xia, L.-W. Chen, A. Rudnicky, and R. M. Stern, "Temporal Context in Speech Emotion Recognition," in *Proc. Interspeech 2021*, 2021, pp. 3370–3374.

[29] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech." in *Interspeech*, 2021, pp. 3415–3419.

[30] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7367–7371.

[31] P. Yue, L. Qu, S. Zheng, and T. Li, "Multi-task learning for speech emotion and emotion intensity recognition," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1232–1237.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[33] D. Tang, J. Zeng, and M. Li, "An end-to-end deep learning framework for speech emotion recognition of atypical individuals." in *Interspeech*, 2018, pp. 162–166.

[34] S. K. Pandey, H. S. Shekhawat, and S. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2019, pp. 1–6.

[35] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.