Speech Emotion Recognition with Multi-task Learning

Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, Kenneth Church

Baidu Research, USA

{xingyucai,jiahongyuan,renjiezheng,lianghuang,kennethchurch}@baidu.com

Abstract

Speech emotion recognition (SER) classifies speech into emotion categories such as: *Happy, Angry, Sad* and *Neutral*. Recently, deep learning has been applied to the SER task. This paper proposes a multi-task learning (MTL) framework to simultaneously perform speech-to-text recognition and emotion classification, with an end-to-end deep neural model based on wav2vec-2.0. Experiments on the IEMOCAP benchmark show that the proposed method achieves the state-of-the-art performance on the SER task. In addition, an ablation study establishes the effectiveness of the proposed MTL framework.

Index Terms: speech emotion recognition, multi-task learning

1. Introduction

Emotions such as *Happy, Angry, Sad* and *Neutral*, play an important role in human communication process. Emotion has been described as an "implicit channel" that is transmitted in addition to the explicit messages [1]. Participants in a conversation can communicate more effectively if they can recognize each others' emotion states. Although it may not be very hard for humans to perceive others' emotions, it remains a challenging task for computers. Considerable efforts have been devoted into emotion recognition (ER) in the human-computer interaction field since decades ago.

People express emotions in many ways including body language, facial expressions, choice of words, tone of voice and more. Therefore, a variety of ER systems based on different types of input signals are proposed in the past, e.g. face emotion recognition [2]. Emotions are even correlated with human's electrochemical characteristics such as EEG signals [3], suggesting electrochemical probes can be used to capture emotions. In this paper, we focus on the speech emotion recognition (SER) task that takes audio speech as input, and outputs emotion classes such as: *Happy, Angry, Sad, Neutral*.

SER systems typically consist of several major cascading components: feature extraction, feature selection and classification [4]. Many systems make use of spectral features, as well as explicit representations of prosodic features, voice quality features and teager energy operator based features [5]. These approaches require strong domain knowledge and a deep understanding of speech. In recent years, end-to-end systems have tended to outperform those traditional systems based on carefully engineered features. In particular, end-to-end deep neural models learn to extract features implicitly, via trainable blocks such as convolutional layers. Thanks to the much larger model capacity (significantly larger number of parameters) and the development of efficient learning algorithms, the deep neural models have become the dominant and preferred systems for the SER task [6].

In this paper, we propose an end-to-end deep neural SER model that is trained using multi-task learning (MTL). The main contributions of this paper are:

- We build an end-to-end model that achieves the state-of-the-art SER results on the standard IEMOCAP [7] dataset.
- We leverage the pretrained wav2vec-2.0 as the backbone for speech feature extraction, and fine-tune on SER data through two tasks: SER (emotion classification) and ASR (speech recognition).
- Ablation study verifies the effectiveness of the MTL approach, and discusses how the ASR affects the SER.
- The speech transcription could be obtained as a byproduct.

The rest of the paper is organized as follows: Section 2 reviews recent related work on SER, MTL and the pretrained wav2vec-2.0. Next, in Section 3, we describe the proposed model, as well as the training and inference processes. Empirical results and ablation studies are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Related Work

2.1. Speech Emotion Recognition

Speech emotion recognition detects the speakers' emotion state from their speech signals. It is often treated as a classification task. Typically, each input utterance is assigned a single class label, which is a predefined emotion category, e.g. *Happy*.

There is a considerable literature on SER. Much of this work makes use of steps such as preprocessing, feature extraction and classification [5, 8]. In early work [4], it was common to extract features such as pitch, energy, formants, mel-band energies, and mel-frequency cepstral coefficients (MFCCs) as the base features, as well as utterance level features such as speaking rate. The next step is to feed those features as input into machine learning classifiers, e.g. SVMs, LDA, QDA and HMMs. SVMs and HMMs performed relatively well, in terms of classification accuracy. Ensembling methods were often found to be more effective [9, 10, 11].

Thanks to the advances of deep learning, neural based models dominate the recent trends in SER studies. The authors in [12] evaluated on CNN and LSTM architecture, and found a concatenation of 3 convolutional layers plus a bi-LSTM layer, yields the best results. This is largely due to the feature learning ability from the convolutional layer, and sequence modeling ability from the RNN model. In [13], a much larger backbone convolutional network, ResNet-101, is adopted to provide stronger feature extraction. More recently, attention mechanism started to play a very important role in NLP domain, and extends to the speech and vision areas. In [14], the authors proposed a model that consists of an attention sliding recurrent neural network (ASRNN). The authors from [15] combine both encoded linguistic and acoustic features, and build a multi-head self-attention model to study the influence of both features on the SER task. In [16], two models, CNN plus attention, and bi-LSTM plus attention, are evaluated and compared from several aspects. A comprehensive survey regarding the recent deep neural models for SER is given in [6].

2.2. Multi-task Learning

Multi-task learning simultaneously optimize multiple objectives in different tasks, using a shared backbone model. The advantages come from auxiliary information and cross regularization from different tasks (implicitly, task A could be the regularizer for task B's objective). At the same time, the challenges arise from the joint optimization [17].

MTL is widely adopted in different deep neural models in various areas. For example, in the computer vision domain, a recent work [18] brought a MTL model that operates on 12 different datasets at the same time, and achieved state-of-the-art results on 11 of them. In speech recognition, [19, 20] place a Connectionist Temporal Classification (CTC) [21] mapping layer beside an attention-based decoder, on top of a shared attention encoder, to perform end-to-end speech recognition (ASR task). The text-to-speech (TTS) model FastSpeech-2 [22] explicitly and jointly learns the mel-spectrogram as well as prosody: pitch, duration and energy. The well-known deep language model BERT [23] employs two tasks in the pretraining phase: masked token prediction, and next sentence prediction. For the SER task, [24] classifies input utterances by both gender and emotion at the same time; this approach is shown to outperform models that predict emotions only.

In addition, MTL is often closely related to other techniques such as transfer-learning [25] and continuous learning [26]. For example, the deep language model ERNIE-2.0 [27] combines continuous learning in MTL framework and achieves state-of-the-art results on GLUE tasks.

2.3. Wav2Vec-2.0: Pretraining with Fine-Tuning for Speech

Combinations of pretraining and fine-tuning have shown to be a very effective learning scheme. It becomes extremely popular for natural language applications. The pretraining phase typically trains a model such as BERT, in an unsupervised manner. A large dataset, e.g. Wikipedia and BookCorpus, is necessary in this phase, to let the model learn meaningful representations of the text. Once pretraining is done, the model could be fine-tuned for specific downstreaming tasks using a relatively small amount of supervised training data with labels. This combination could outperform models trained only on the task-specific data, because of the knowledge gained in the pretrainig phase. Another advantage is that, the model tends not to overfit the task-specific data, because the pretrainig phase behaves like a regularizer that brings a lot of prior information. Due to these benefits, this approach of combining pretraining and fine-tuning, is moving from natural language processing domain [23, 28, 27] to the speech domain [29].

A very recent pretrained model, wav2vec-2.0 [30], learns representations of speech by pretraining on large quantities of audio data, using a similar unsupervised approach to what adopted by BERT. It tries to recover the randomly masked portion of the encoded audio feature. After pretraining, wav2vec-2.0 has been fine-tuned on Librispeech [31] with impressive performance in terms of word-error-rate (WER). In this paper, we also start with a pretrained wav2vec-2.0 model, but fine-tune it for a different downstream task, speech emotion recognition (SER), using two different task-specific heads.

3. Proposed Method

In this section, we will present the details of our proposed model, and the multi-task training scheme.

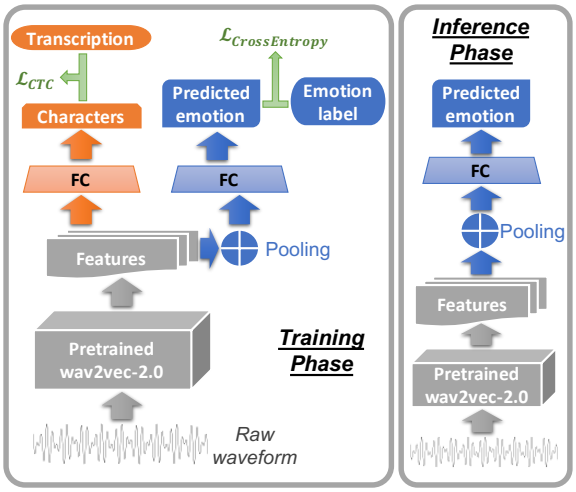

Figure 1: *The proposed training model takes a single input (raw waveform) and produces two outputs (predicted characters and predicted emotions). There are two losses: $\mathcal{L}_{CE}$ and $\mathcal{L}_{CTC}$ which measure losses between predictions and gold labels. The inference phase discards the orange path in the diagram.*

3.1. Model Architecture

We propose an end-to-end model that inputs speech as raw waveforms and outputs predicted emotions. Figures 1 shows the training and inference processes. There are three colors in the figure. The gray color represents the shared backbone pretrained wav2vec-2.0 model (bottom part). Orange (upper left part in the training phase diagram) and blue (upper right in training, upper part in inference) are used to separate the two tasks. The blue task is SER: SER inputs speech and outputs emotion labels. The orange task is ASR: ASR inputs speech and outputs text. The blue task is the main task of interest for this paper. The orange task plays a supporting role during training, but is not used at inference time. The ablation study in Section 4.5 will show that this combination of orange and blue tasks improves performance on the blue task, even though the orange part is not used at inference.

We denote the pretrained wav2vec-2.0 model as $f_\theta(\cdot)$. It is the gray box in the bottom. Let the input waveform be $x \in \mathcal{R}^L$, whose length is L (a total of L samples), we get wav2vec-2.0's last hidden layer's output as the feature z (the gray component labeled as Features), i.e. $z \in \mathcal{R}^{L\times d} = f_\theta(x)$, where d is the hidden dimension, typically 768, and θ represents the parameters in f. Both the orange and blue parts take z as their input.

For the orange (ASR) path, we use a vocabulary that consists of $V = 32$ characters, within which there are 26 letters in English plus a few punctuation characters. As shown in Figure 1, after obtaining the feature z, we apply a fully-connected layer (the orange FC block in the left), denoted as g_ϕ that maps $z \in \mathcal{R}^{L\times d}$ to logits $y \in \mathcal{R}^{L\times V}$. At the end of the orange path, we obtain predictions for characters in terms of logits: $y = g_\phi\left(f_\theta(x)\right)$.

The blue path in Figure 1 starts with a pooling layer that sums over the sample length L. This transforms a sequence (length L) of vectors, denoted as $z \in \mathcal{R}^{L\times d}$, into a single vector $\hat{z} \in \mathcal{R}^d$. Suppose we have C emotion categories. We apply another fully-connected layer h_ϕ (the blue FC block) that maps $\hat{z}$ to the logits $c \in \mathcal{R}^C$. At the end of the blue path, we obtain predictions for emotion classes in terms of logits:

$c = h_\phi(\sum_{i=0}^{L-1} z_i)$. Note that ϕ represents the parameters associated with the two fully-connected layers g and h.

3.2. Training and Inference

The training phase is supervised. That is, the training process has access to gold labels for both the orange and blue paths. For each utterance in the training set, the orange path has access to the gold transcription, t, and the blue path has access to the ground-truth emotion label, l. At the end of both paths, we apply a softmax operator on both y and c to convert them to probability vectors (a probability vector $v \in \mathcal{P}^d$ s.t. $\sum_{i=1}^{d} v_i = 1$ and $v_i \geq 0$). For the character encoding (orange task), we compute the Connectionist Temporal Classification (CTC) [21] loss against the encoding of the given gold transcription. CTC is the standard technique to map the input signal to the output target, when they don't have the same length and no alignment information is provided. Here the speech signal length L is typically significantly longer than the text transcript length, because multiple frames correspond to a single phoneme. CTC could be used as a loss function and we can backpropogate the gradient efficiently. Details could be found in [21]. Thus, we have the CTC loss as:

$$\mathcal{L}_{\text{CTC}} = \text{CTC}(\hat{y}, t), \text{ where } \hat{y} = \text{softmax}(y) \in \mathcal{P}^{L \times V} \quad (1)$$

Meanwhile, we compute the cross-entropy between the predicted probability distribution and the true emotion label. Cross-entropy loss is widely utilized for classification tasks.

$$\mathcal{L}_{\text{CE}} = \text{CrossEntropy}(\hat{c}, l), \text{ where } \hat{c} = \text{softmax}(c) \in \mathcal{P}^{C} \quad (2)$$

We introduce a hyper-parameter α to combine two losses into a single one. α controls the relative importance of the CTC loss. The optimal choice of α can be found by grid search. Finally, we optimize the following objective w.r.t θ and ϕ:

$$\min_{\theta,\phi} \mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{CTC}} \quad (3)$$

At inference time, we replace softmax with the argmax operator, and select the most probable emotion class label as output. To obtain predictions for transcriptions, we need to further add a CTC decoder to convert y to the most probable text $\hat{t}$.

Note that gold transcriptions, t, are used to fine-tune the network during training only. There is no need for t at inference time. In other words, our model predicts $P[\text{emotion} \mid \text{waveform}]$, rather than $P[\text{emotion} \mid (\text{waveform}, \text{transcription})]$. In addition, there is no need for explicit language models under the proposed approach. This is the key difference between our model and other multi-modal models, e.g. [32], that use gold transcription or other source in the inference time. Therefore, our approach is uni-modal during inference.

4. Experiments

4.1. IEMOCAP Benchmark Data and Metrics

Following much of the literature on SER such as the references in Table 2, we use the IEMOCAP [7]. This dataset contains approximately 12 hours of speech from 10 speakers. Each utterance is labeled with an emotion. The literature in Table 2 selects $N_{\text{total}} = 5531$ utterances that are labeled with one of five categories: *Happy, Angry, Neutral, Sad, Exited.* This set is mapped down to four categories by merging *Exited* and *Happy* into a single category. We perform 10-folds cross validation, following [33]. For each fold, we leave one speaker out as the test set and use the remaining nine speakers for training. The final weighted accuracy (WA) is computed as

$$\text{acc} = \frac{1}{N_{\text{total}}} \sum_{k=1}^{10} N_{\text{correct}}^{(k)} \quad (4)$$

where $N_{\text{correct}}^{(k)}$ is the number of correct emotion predictions on the test set in k-th fold. Much of this literature also reports unweighted accuracy (UA), the average accuracy of different classes. Nevertheless, we list whichever is higher between UA and WA, for those baselines in Table 3.

4.2. Hyper-parameters

We conduct experiments using the pretrained model: wav2vec-2.0-base. The wav2vec-2.0-base has 12 transformer blocks and 7 convolutional blocks (each has 512 channels). Our implementation is based on the Huggingface transformers repository [34]. We take the pretrained model as the shared backbone. Note that as shown in Equation (3), parameters for the backbone, θ, as well as parameters for both task-specific heads, ϕ, are optimized jointly during fine-tuning.

Table 1 lists a number of settings of hyper-parameters that we used in the fine-tuning experiments. It was necessary to reduce the learning rate when $\alpha = 0$ (performing emotion classification only). Otherwise, fine-tuning can easily diverge after a few epochs. The best choice we found is to set the learning rate to be 10^{-5} when $\alpha = 0$. For $\alpha > 0$, the best learning rate is 5×10^{-5}. For each run, we use 2 GPUs (TitanX), with a batch size of 1, accumulated 4 forward pass per backward pass, resulting in accumulated batch size of 8.

Table 1: *The hyper-parameters used for fine-tuning.*

sample frequency	16k Hz
training epochs	100
optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$)
α	0, 0.001, 0.01, 0.1
learning rate	10^{-5} if $\alpha = 0$, 5×10^{-5} otherwise
warm-up ratio	0.1 , linear warm-up
batch size	8 (accumulated batch size)

4.3. Baseline methods

We compare with a few most recent baselines. We also provide a brief summary of the baseline methods, along with the associated publications and years, listed in Table 2. These baselines use the same settings as ours.

Table 2: *Baseline methods from the literature*

Method	Description	Year
Wu et al. [35]	capsule network	2019
Sajjad et al. [13]	ResNet-101 + bi-LSTM	2020
Lu et al. [33]	pretrained ASR + bi-LSTM + attention	2020
Liu et al. [36]	local + global representation learning	2020
Wang et al. [37]	Dual-Sequence LSTM	2020
Pappagari et al. [38]	ResNet based x-vector model	2020
Peng et al. [14]	3D convolution + ASRNN	2020

Note that some other literatures are not included due to their different settings. For example, [39] uses a set of six emotion categories rather than four; [24] takes advantage of aux-

iliary gender information; and [16] does not use the same cross-validation settings as other baselines. Because of these mismatches, we limited the comparisons in this paper to the methods listed in Table 2.

4.4. Evaluation Results

In this section, we report the acc of our proposed method, against all the baselines. We also list the cross-validation settings. The results are in Table 3.

Table 3: *Speech emotion recognition (SER) results.*

method	cross-validation	acc
Wu et al. [35]	10-fold	72.73%
Sajjad et al. [13]	5-fold	72.25%
Lu et al. [33]	10-fold	72.6%
Liu et al. [36]	5-fold	70.78%
Wang et al. [37]	5-fold	73.3%
Pappagari et al. [38]	5-fold	70.30%
Peng et al. [14]	5-fold	62.6%
ours	10-fold	**78.15%**

Clearly, our proposed model significantly improves the state-of-the-art performance in the SER task on IEMOCAP benchmark dataset. We outperform the best baseline by a large margin (about 5% improvement).

4.5. Ablation Study

In this subsection, we study whether the multi-task learning could help the SER task. Recall that the hyper-parameter α controls the strength of CTC loss. Therefore, we vary α from 0 to 0.1, and obtain the different acc. Note that when $\alpha = 0$, the model is degraded to single-task training, using only the emotion labels. As α increases, the CTC loss becomes more and more important. Table 4 reports acc for different values of α. To show the impact of CTC loss on the speech-to-text part, we also list the average word-error-rate (wer) obtained during the experiments. It is computed as

$$\text{wer} = \frac{1}{10}\sum_{k=1}^{10} \text{wer}^{(k)} \tag{5}$$

where $\text{wer}^{(k)}$ is the word-error-rate on the test set in k-th fold.

Table 4: *The impact of the CTC loss. $\alpha = 0$ corresponds to single task (emotion classification only), which leads to poor acc. Both acc and wer improve with larger α.*

	acc	wer
$\alpha = 0$	71.66%	0.9981
$\alpha = 0.001$	73.97%	0.2233
$\alpha = 0.01$	76.34%	0.2007
$\alpha = 0.1$	78.15%	0.1929

Table 4 reports accuracy (acc) and word error rates (wer) for four choices of α. $\alpha = 0$ corresponds to training on just a single task (the blue SER task), which yields a poor acc that is worse than the baselines. Larger values of α improve both acc and wer. This verifies that the performance improvement comes from multi-task learning scheme.

In Figure 2, we further study how α affect the training phase. We plot the acc and wer against training epochs, using one run. From the figure, we can see that the bigger α,

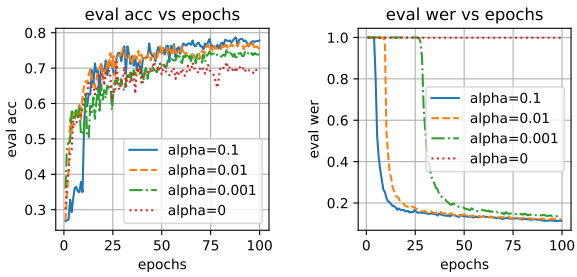

Figure 2: *The acc and wer against training epochs, for different α. Note that when $\alpha = 0.1$ (strongest CTC loss), the acc converges slower than others in the beginning (the blue curve in left plot). However, after the wer converges (at around epoch 12), the acc climbs quickly and outperforms others in the end.*

the earlier that the model reaches low word-error-rate (the blue curve in the right plot first reaches lower wer). However, before reaching a satisfactory wer, acc grows slower (the blue curve in the left plot is below other curves when epoch < 15). Interestingly, after the model reaches a good wer, the performance on acc climbs quickly for larger α, and they soon outperform smaller α cases (the blue curve catches other curves quickly and leads in the later epochs). This indicates that the ASR task is very important, and a better ASR trained model (lower wer) yields better SER (higher acc).

5. Conclusions

In this paper, we propose a simple end-to-end model for speech emotion recognition (SER). The model leverage a pretrained speech model, wav2vec-2.0, as the feature extractor backbone. We add a CTC head and a classification head on top of the backbone, to simultaneously obtain predictions for transcriptions and emotion categories. Comprehensive experiments were conducted using the IEMOCAP benchmark dataset. The proposed model improved the state-of-the-art results by a large margin (about 5% improvement). The ablation study established the effectiveness of the proposed multi-task learning (MTL) approach. It shows training on the combination of the ASR plus SER tasks, performs better than training on a single task.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] D. Pandit, "A comprehensive survey of different phases for involuntary system for face emotion recognition," in *Recent Trends in Image Processing and Pattern Recognition: Third International Conference, RTIP2R 2020, Aurangabad, India, January 3–4, 2020, Revised Selected Papers, Part I.* Springer Nature, 2021, p. 169.

[3] J. Yan, S. Chen, and S. Deng, "A eeg-based emotion recognition model with rhythm and time characteristics," *Brain informatics*, vol. 6, no. 1, pp. 1–8, 2019.

[4] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[5] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, sup-

porting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[6] E. Lieskovska, M. Jakubec, and R. Jarina, "Speech emotion recognition overview and experimental results," in *2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA)*. IEEE, 2020, pp. 388–393.

[7] 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.

[8] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[9] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[10] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2010.

[11] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

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

[13] M. Sajjad, S. Kwon *et al.*, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE Access*, vol. 8, pp. 79 861–79 875, 2020.

[14] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16 560–16 572, 2020.

[15] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7189–7193.

[16] M. A. Jalal, R. Milner, and T. Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition," *Proc. Interspeech 2020*, pp. 4113–4117, 2020.

[17] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[18] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 437–10 446.

[19] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[20] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[22] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[25] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[26] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[27] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8968–8975.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[29] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "fairseq s2t: Fast speech-to-text modeling with fairseq," *arXiv preprint arXiv:2010.05171*, 2020.

[30] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[31] 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.

[32] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," *arXiv preprint arXiv:1912.02610*, 2019.

[33] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end asr models," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7149–7153.

[34] 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[35] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu *et al.*, "Speech emotion recognition using capsule networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6695–6699.

[36] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7174–7178.

[37] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6474–6478.

[38] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[39] J. Sebastian, P. Pierucci *et al.*, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts." in *Interspeech*, 2019, pp. 51–55.