Unsupervised Speech Representation and Multi-task Learning for Speech Emotion Recognition

G001 (s1407182, s1902530, s1968246)

Abstract

Our project focuses on using unsupervised neural speech representations and multi-task learning to improve performance on task of Speech Emotion Recognition (SER). SER requires to recognize categorical emotion from individual utterances in speech. We have two research questions to investigate. The first one is to explore how unsupervised neural speech representations can benefit SER when they are used in feature-based manner. The second research question is to investigate the effectiveness of gender classification as an auxiliary task for multi-task learning (MTL) for SER, when use Mel-based features. In order to answer these research questions, we train our LSTM model both with and without unsupervised representation and gender classification. The experiments are ran on IEMOCAP, which is a commonly used dataset for SER. Our experiments indicate that the chosen unsupervised representations, MPC, does not benefit SER when it is used in feature-based manner. The experiments also show that gender classification also does not improve SER when MPC features are used.

1. Introduction

Unsupervised neural representation is a long-standing research area for the Natural Language Processing (NLP) community which can directly benefit downstream tasks by pre-training on large unlabelled corpus (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2019). Recently, BERT achieved SOTA performance on eleven NLP tasks (Devlin et al., 2019). However, unsupervised representation is a relatively new idea for the Speech community. We are particularly interested in Contrastive Predictive Coding (CPC) (van den Oord et al., 2018; Schneider et al., 2019), Autoregressive Predictive Coding (APC) (Chung et al., 2019) and Masked Predictive Coding (MPC) (Liu et al., 2019) which leverages comparatively large corpus to train speech representation in an unsupervised manner. CPC uses convolutional neural network (CNN) for unsupervised pre-training, APC uses recurrent neural network (RNN) and MPC uses BERT-like Transformer encoder (Vaswani et al., 2017) to reconstruct masked out input features. The core idea behind these unsupervised representation learning is to predict future or masked out features to train a network, that can produce representations which are directly used on downstream tasks such as phoneme classification and sentiment classification (Liu et al., 2019). These unsupervised representations are trained on Librispeech (Panayotov et al., 2015), a dataset for Automatic Speech Recognition. It contains a total of 960 hours of clean or noisy speech, compared to only about 7 hours data from IEMOCAP we are using for our experiments.

The goal of our experiments are to find out whether unsupervised representations and gender multi-task classification improve the task of SER. This could benefits the Speech community in terms of two aspects. Firstly, unsupervised representations are trained from large unlabelled corpus, meaning that we could have used these to improve the performance on downstream tasks with relatively small dataset. Secondly, gender multi-task classification uses gender label, which requires relatively less human annotation effort. IEMOCAP offers ground-truth gender labels.

Our first research task is to investigate how well MPC unsupervised representation can be generalized to our IEMOCAP SER task. Our baseline will be using traditional Mel-frequency cepstral coefficients (MFCC) as input feature. We will be putting raw log Mel-Spectrogram into pre-trained network to generate MPC embeddings[1]. Then we will be using their output as input into our model, and compare with our baseline. This means that we are only changing our input, leaving any other network structure unchanged. To distinguish our work from existing work, the original paper of MPC (Liu et al., 2019) we refer to have done comparison between APC and MPC on the MOSEI dataset on the task of sentiment classification which is a different task from SER (Bagher Zadeh et al., 2018). We expect to see MPC can improve our MFCC baseline in our SER task.

Multi-task learning (MTL) (Caruana, 1997) refers to using different tasks to improve each other. Particularly, gender classification has been used to improve SER in the setting of using raw wave input (Li et al., 2019). They argued that classify gender explicitly can benefit SER, as sad woman voice can sound like neutral man voice and happy man voice can sound like neutral woman voice. They also argued that this may probably because they were using raw wave input features which is not as good as MFCC for gender classification (Bisio et al., 2013), which means training explicitly on gender classification might not help SER when

[1] https://github.com/andi611/Mockingjay-Speech-Representation

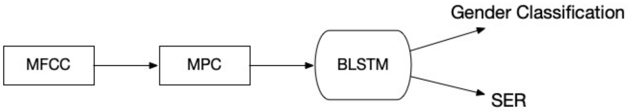

Figure 1. Overall network architecture.

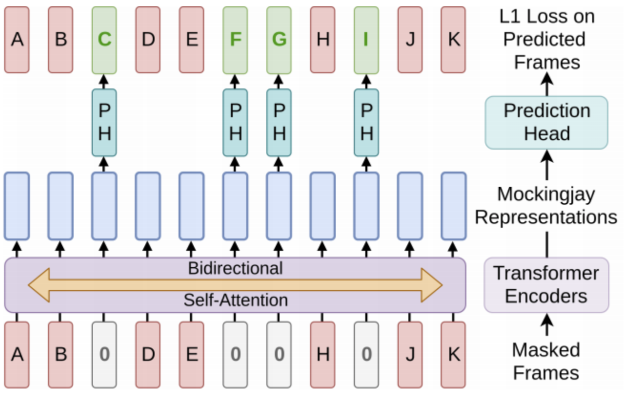

Figure 2. MPC architecture.

input is MFCC feature.

Our second research task is to investigate whether gender classification will help SER as an auxiliary task under the setting of MTL, particularly when input is Mel-based features. We investigate this on MPC features. The difference between our research and the literature (Li et al., 2019) is that they have used gender classification on raw wave input, while we will be using gender classification as MTL on our MPC input and MPC receives log Mel-spectrogram as input. We expect to see gender classification MTL will not improve accuracy on MPC features, APC and MPC on SER, as it is Mel-based features.

Overall, we inspect whether MPC features will improve the SER task on IEMOCAP. We will also be validating whether gender classification MTL improves SER when using MPC input features. We found that both unsupervised MPC features and gender classification do not improve the task of SER.

2. Data set, pre-processing and task

2.1. Data set and pre-processing

	CPC	APC	MPC
Input	raw wave form	Mel	Mel
Backbone	CNN	RNN	Transformer
Bi-directional	×	×	√
Loss	Discriminative	Predictive	Masked

Table 1. Comparison between CPC, APC and MPC.

Interactive Emotional Dyadic Motion Capture (IEMOCAP) is a multi-modal emotional classification dataset which contains approximately 12 hours of speech, with corresponding text transcripts, video and face motion capture (Busso et al., 2008). It was collected from 10 actors, five male and five female. Each sentences have their corresponding categorical labels, dimensional labels and self-evaluation labels. Categorical labels are annotated by 6 annotators. The original IEMOCAP has 10 categorical labels which are imbalanced.

We discard text transcripts, videos, face motion capture, dimensional labels and self-evaluation labels. We follow previous work to discard all the sentences which do not fall into the category of *ANGRY*, *SAD*, *HAPPY*, *EXCITED* and *NEUTRAL*, then we merge *EXCITED* into *HAPPY* (Li et al., 2019). Instead of using soft labels (Kim & Kim, 2018), we simply use majority vote among six annotators as categorical ground truth and whenever there is a tie, we discard the sentence. As a result, we obtained 5,531 sentences, which constitutes approximately 7 hours of speech in total. There are 1,103 angry utterances, 1,636 happy utterances, 1,708 neutral utterances and 1,084 sad utterances. The original dataset does not contain gender label, but the gender information can be easily extracted from audio file name which is marked as 'M' or 'F'. We randomly shuffle the dataset into 70/20/10 as the training/validation/testing set with random seed 123456.

We will be using both weighted accuracy (WA), unweighted accuracy (UA) (Neumann & Vu, 2019) to evaluate our unsupervised representation and gender classification MTL. Accuracy is a common indicator to evaluate this dataset. WA is the accuracy for the entire dataset and UA is the average accuracy for each emotion category.

We will be using librosa to extract MFCC and log Mel-spectrogram features from raw audio file (McFee et al., 2020). For log Mel-spectrogram, we set the window size to be 25ms and stride to be 10ms, and we extract 160-dimensional features for each window (Liu et al., 2019). For MFCC, we use the same window size and stride and we extract 40-dimensional features (Tripathi et al., 2019). We will not be using cross-validation (Li et al., 2019) due to the limitation of computational resource.

2.2. Task Formulation

The task of SER is formulated as follows. Each raw wave audio file a contains a single utterance from one actor, which is 4 seconds in average from IEMOCAP. a is pre-processed into MFCC-based speech feature or processed by APC, CPC or MPC. The pre-processed feature is denoted as i. Given i, the models are required to output one of the $\{$*ANGRY*, *SAD*, *HAPPY*, *NEUTRAL*$\}$ as the final output.

3. Methodology

The whole network architecture is illustrated in Figure 1. MFCC and Mel-spectrogram (MSPC) are the traditional features used in SER (Tripathi et al., 2019) is extracted by using librosa . Our baseline skips the MPC module and feeds MFCC directly into the bi-directional LSTM (BLSTM). When those unsupervised trained representations are used, MSPC is firstly fed into the MPC network, then the output from them are fed into BLSTM to be used. Since we are doing a classification task here, the final hidden states of the LSTMs are extracted and fed into two non-weight sharing output layers. One is for binary gender classification and the other one is for multi-class emotion classification.

3.1. MFCC Spectrogram

MFCC(Mel-frequency cepstral coefficient) are coefficients that represent short-term power spectrum of audio data. The coefficient is based on a linear cosine transformation of the log power spectrum of mel scale frequency, which is:

$$Mel(f) = 2595log(1 + \frac{f}{700}) \tag{1}$$

where f is the frequency. This function obtains MSPC, which is further fed into DCT (discrete cosine transform) to obitan mFCC. In order to approximate the response of human auditory system more closely, the frequency bands in the MFCC are equally spaced on the mel scale frequency. MFCC is a type of frequency warping which has a better representation of sound.

During our experiments, When mapping the spectrum mentioned above into mel scale, these sample points are divided into collections called frame. Every frame has a length of 25ms which is the size of the window, and we sample a frame every10ms. The overlap of the 'slide' operation is for the purpose of avoiding the significant change of two neighbour frames. During our baseline experiments, we set the dimensionality of MFCC to 40, such that for every second of raw audio, the output will be a vector with 40×100 dimensionality.

MSPC (Mel-frequency Spectrogram) is can be seen as MFCC without DCT. MFCC has a compressible representation with only a fixed number of coefficients, while MSPC keeps more information of the original data. In the MPC model, we use MSPC as the upstream transformation.

3.2. CPC, APC and MPC

As in Table 1, CPC uses raw wave form as the input, but not MFCC. This is why we argue that gender classification MTL can be useful when CPC is used. It uses CNN and tries to discriminate future wave signal from negative sampled signal. APC and MPC both use MFCC. While APC uses RNN to predict future framed MFCC feature, it is not bi-directional, meaning no future frame can be used together with past frame to predict current frame. MPC, however, benefited from the masked MFCC loss trained on Transformer encoder backbone, can use bi-directional information to reconstruct a masked out frame, meaning both past and future frames can be used to predict one masked frame, thus MPC is deeply bi-direction. This also means that, ideally, MPC should works better than APC and APC should works better than CPC.

Figure 2 presents the overall network architecture for MPC. Some of the input is masked as 0 and the network tries to reconstruct them. We are mainly interested in the Mockingjay representations, which should be extracted and fed into the BLSTM part of our model. Due to computation resource limitation, we will not be training our own CPC, APC and MPC models and we will be using their trained models which are publicly available.

3.3. BLSTM

LSTM is a commonly used RNN architecture (Hochreiter & Schmidhuber, 1997) and it is not the main focus for our study, hence, we only describe it briefly here. Since we have a classification task, we obtain two of the hidden representations from last hidden states of the BLSTM and concatenate them:

$$h = [h_f; h_b] \tag{2}$$

where h_f is from the forward LSTM and h_b is from the backward LSTM. h is fed forward to gender classification and SER.

3.4. Gender Classification and SER

The representation h comes from the BLSTM part of our model are fed into both gender classification layer and SER layer. Both of the layer have non-shared parameters:

$$\begin{aligned} &softmax(W_g h) \\ &softmax(W_s h) \end{aligned} \tag{3}$$

where W_g are the trainable parameters for gender classification and W_s are the trainable parameters for SER. The softmax layer for gender classification produces a two-way classification results correspond to *MALE* and *FEMALE*. The softmax for SER produces a four-way classification results correspond to *ANGRY*, *HAPPY*, *SAD* and *NEUTRAL*.

The final loss is thus:

$$L = \alpha L_s + \beta L_g \tag{4}$$

where L_s is the loss for SER and L_g is the loss for gender classification. These two losses are backpropagated simultaneously for each training samples. α and β are hyperparameters.

4. Experiments

All of our experiments are ran for 300 epochs. Test performance is reported on the model with best validation accuracy.

4.1. Baseline

Our baseline experiments are ran on a simple bi-directional LSTM network. LSTM network has been used as a standard architecture for the task of SER (Tripathi et al., 2019; Li et al., 2019). The baseline uses 40-dimensional MFCC features.

We use default Adam optimizer setting in Pytorch, which uses 0.001 as learning rate, 0.9 as β_1, 0.999 as β_2, 1e-8 as epsilon, and 0 weight decay. We keep the hidden size to be 256 and alter the number of layers from {1, 2, 3}. For experiments with multiple layers, we use dropout with $p = 0.5$.

# of layers	best val WA acc	best val UA acc
1	**0.602**	**0.595**
2	0.585	0.592
3	0.581	0.576

Table 2. Best WA and UA validation accuracy for baseline MFCC features.

The best WA and UA validation accuracy built by the MFCC baseline are 60.2% and 59.5% respectively. We establish this as our baseline. This is close to the other arts which have employed BLSTM (Lee & Tashev, 2015), which have achieved 62.8% on WA and 63.9% on UA.

4.2. Experiment A

To improve the baseline, we follow the feature-based methods employed in the original MPC paper (Liu et al., 2019) to extract 768-dimensional features from the last layer of the MPC model. The input to the MPC model is 160-dimensional log Mel-Spectrogram features. We replace the MFCC input features with extracted MPC input features.

Firstly, we found that a too big learning rate would cause NaN value in the loss. A too small learning rate would cause slow learning. We found that 5e-4 is a good value for learning rate for the MPC last layer feature. Before fine-tuning on the learning rate, we use 5e-4 as the learning rate to find whether adding multiple layers and dropout to the LSTM network would improve performance for MPC feature. We keep all the hyperparameter for the Adam optimizer as default which has 0.9 as β_1, 0.999 as β_2, 1e-8 as epsilon, and 0 weight decay.

As in Table 3, we found that the best validation accuracy is achieved on a single layer BLSTM with no dropout.

Then, we further tune the learning rate with one-layer BLSTM that does not have any dropout.

As in Table 4, using a too large learning rate would cause NaN value in the loss which the model will not be updated anymore in the training iterations and stuck at 20% accuracy. Using a too small learning rate such as 5e-4 would cause a slow learning. Thus, we adpot 5e-4 as the default learning rate in our later experiments.

# of layers	dropout	best val WA acc	best val UA acc
1	0	**0.600**	**0.590**
2	0	0.554	0.562
2	0.1	0.572	0.568
2	0.3	0.538	0.552
2	0.5	0.598	0.567
2	0.7	0.545	0.542
3	0	0.552	0.540
3	0.1	0.575	0.582
3	0.3	0.535	0.569
3	0.5	0.545	0.567
3	0.7	0.542	0.553

Table 3. Best WA and UA validation accuracy for MPC features on LSTM with different number of layers and dropout probability.

learning rate	best val WA acc	best val UA acc
1e-2	NaN	NaN
5e-3	0.547	0.552
1e-3	0.553	0.561
5e-4	**0.600**	0.590
1e-4	0.423	0.435

Table 4. Best WA and UA validation accuracy for baseline MPC features with different learning rates.

Hidden size	best val WA acc	best val UA acc
128	0.554	0.542
256	**0.600**	**0.590**
512	0.562	0.573
1024	0.572	0.582

Table 5. Best WA and UA validation accuracy for baseline MPC features with different hidden sizes.

Layers NO.	best val WA acc	best val UA acc
1	0.458	0.451
2	0.503	0.523
3	0.483	0.476
4	0.482	0.465
5	0.511	0.486
6	0.482	0.460
7	0.479	0.490
8	0.440	0.450
9	0.481	0.500
10	0.512	0.523
11	0.543	0.553
12	**0.600**	**0.590**

Table 6. Best WA and UA validation accuracy of MPC features from different layers in the original MPC model.

As in Table 5, we found that the best hidden size is 256. Increasing the hidden size does not bring increase in validation accuracy. Due to time limitation, we tuned exhaustively and we leave further tuning as further experiments. We also

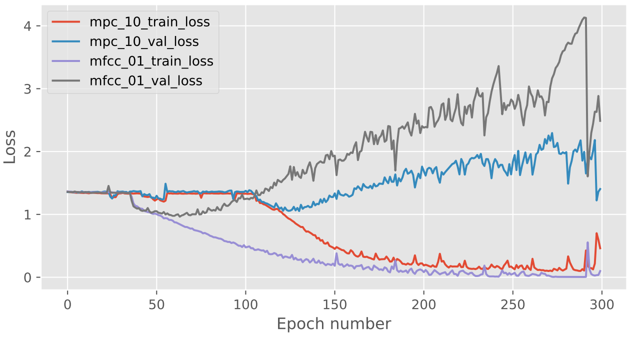

Figure 3. The loss graph for the baseline and the MPC features with the best hyper-parameter setting we have found.

have not run K-fold validation. As our dataset only have 5,531 examples, the reported results could have high variance. We leave this as future work.

Using MPC features from the final MPC model does not improve SER. Also, as in Figure 3, using MFCC hits the best validation performance quicker than MPC.

4.3. Experiment B

As in previous experiments, we have found that using the representations from the final layer of the MPC model does not improve over MFCC baseline. We thus extract the features from the previous layers to see whether using them would get better accuracy than using the final layers. We keep all the hyper-parameters as the same and we use the features from the previous 1st to 11th layers to compare with the final 12th layer in the MPC model.

As in Table 6, we have found that using features from other layers is worse than using the final layers. This is probably because the final layers contains most valuable information for the task of SER.

4.4. Experiment C

Combining all the layers from the MPC model could have improved performance. The implementation is straightforward, however, due to time limitation, we leave this as future experiments.

Fine-tuning on the original MPC model could also improve the performance. However, the parameters of this model is too large to be hold in a V100 GPU. So we have tried a trick called mixed precision training using apex developed by Nvidia, hoping to reduce the demand of VRAM and speed up the fine-tuning. However, the average time for one epoch of fine-tuning would costs around 1 hour. Besides, MPC employs self-attention, so the largest batch size we could have set is 1. After training for 15 epochs, the training accuracy is still below 50%. So we decide to leave it as future experiments. This could be potentially sped up by dividing the original audio into small pieces so that we could have used a larger batch size to decrease overheads of switching between GPU and CPU. With multiple GPUs, we could also use model parallelism to fine-tune without precision loss, or data parallelism to speed up the fine-tuning.

Beta	best val WA acc	best val UA acc
1e-3	0.550	0.561
5e-3	0.584	0.582
1e-2	0.588	0.590
5e-2	0.567	0.550
1e-1	0.543	0.561
5e-1	0.550	0.562
0	**0.600**	**0.590**
1e-0	0.456	0.435
5e-0	0.350	0.367
1e1	0.456	0.480
5e1	0.430	0.412
1e2	0.394	0.402
5e2	0.374	0.365
1e3	0.398	0.412
5e3	0.345	0.360

Table 7. Best WA and UA validation accuracy of MPC features from the final layers with different β values.

Beta	best val WA acc	best val UA acc
1e-3	0.588	0.573
5e-3	0.583	0.596
1e-2	0.592	0.581
5e-2	0.583	0.593
1e-1	0.562	0.545
5e-1	0.576	0.549
0	**0.602**	**0.595**
1e-0	0.562	0.571
5e-0	0.552	0.542
1e1	0.531	0.521
5e1	0.535	0.543
1e2	0.525	0.530
5e2	0.512	0.502
1e3	0.452	0.443
5e3	0.431	0.452

Table 8. Best WA and UA validation accuracy of MFCC features with different β values.

4.5. Experiment D

To answer our second research question and find about whether gender classification works, we run gender classification and SER simultaneously on one sample and backward the loss together. The main goal here is to improve the performance of SER. Since we have achieved a reasonable performance on SER in our previous experiments with learning rate 5e-4, we do not want to tune this more and thus we keep it as 5e-4 and tune the coefficient β for the gender classification loss. We keep the coefficient for the SER loss α as 1.

As in Table 7, it is found that using gender MTL does not improve SER on MPC features. When β gets larger, the SER accuracy continually decreases. As in Table 8, similar phenomenon is the observed on MFCC features.

This is probably because that gender classification is an easy task for both MPC and MFCC features. It is observed that it can achieve over 99% validation accuracy in the above experiments.

5. Related Work

5.1. Unsupervised Representation Learning

Unsupervised representation learning has been widely applied to NLP community. The central idea is to use nearby words for word prediction. This is based on distributional hypothesis (Harris, 1954) which argues that nearby words should have similar meaning. Based on such hypothesis, unsupervised algorithms use unlabelled corpus that is particularly large to project words into dense vectors that could be used directly on downstream tasks. Word2Vec (Mikolov et al., 2013) has used Continous-bag-of-words (CBOW) and Skip-Gram (SG) to predict whether words should be within the same context window. This produces word embeddings that could be used on downstream tasks. GloVe (Pennington et al., 2014) uses the difference between nearby words as the loss, which is weighted by their global occurrences, and it was empirically proved to have on par performance with Word2Vec (Naili et al., 2017). ELMo (Peters et al., 2018) uses LSTM to contextualize the word embeddings so it produces different embeddings for each word, under different context. GPT (Radford, 2018) uses Transformer encoder to replace RNN structure in ELMo. BERT (Devlin et al., 2019) produces deep bi-direction contextualized words. Although BERT also uses Transformer encoder as its backbone, it has a different masked word loss to reconstruct some masked out words, which improves GPT by bi-directional contextualization. GPT-2 (Radford et al., 2019) uses larger dataset and much more parameters than GPT and BERT. It zero-shots state-of-the-art performance on 7 tasks without any fine-tuning, while BERT and GPT should be fine-tuned to be adopted to downstream tasks.

However, unsupervised representation learning is a comparatively new idea to the Speech Language Processing community. One can refer CPC as the speech version of Word2Vec which uses Convolutional Neural Network (CNN) to produce representations from raw wave. One can refer APC as the speech version of ELMo which uses Recurrent Neural Network (RNN) to produce representations from log Mel-spectrogram features. One can refer MPC as the speech version of BERT which uses masked loss to produce representations from log Mel-spectrogram features. Our work aims at finding our whether those representations can benefit the task of SER on IEMOCAP.

5.2. Multi-task Learning

Multi-task learning uses multiple tasks to improve each other. Firstly, it provides implicit regularisation as we try to fit our model parameters to different tasks, thus decrease the probability of overfitting. Secondly, it could augment tasks with low-resource with tasks with high-resource, which boosts performance for the low-resource task. Thirdly, it gives 'hints' to the tasks. For example, for Speech Emotion Recognition, training explicitly on gender classification (Li et al., 2019) gives performance boost. Without gender classification, a sad woman voices could sound like neutral man voice. However, this was used for raw-wave input (Li et al., 2019), and our work aims at investigating whether this is effective when MFCC is used as input feature.

5.3. Speech Emotion Recognition

Speech emotion recognition (SER) aims at recognizing utterance emotion from given audio clips. IEMOCAP is a commonly used dataset for SER. The state-of-the-art method has a weighted accuracy of 81.6% and unweighted accuracy of 82.8% on it (Li et al., 2019). It receives a raw-wave form input, which is passed to a CNN for feature extraction. The extracted features are passed into bi-LSTMs. The hidden states of the LSTMs are extracted and fed into a self-attention layer. The attended features are fed into two softmax layer, one for SER and one for gender classification.

5.4. Speech Sentiment Analysis

Speech Sentiment Analysis (SSA) is a different task from SER. In SER, models are required to output the emotion of the speakers as for example, *HAPPY* and *SAD*. In SSA, models are required to classify the attitude of the speakers as either *POSITIVE*, *NEUTRAL* or *NEGATIVE*. MPC (Liu et al., 2019) has been used instead of log Mel-spectrogram features as input into a singler layer RNN, which has been empirically proved to improve performance. They extracted the 768-dimensional hidden representation as features from the last layer of MPC and feed that into the RNN. They also reported two alternative ways which also improved performance. On one hand, they extracted features from all the layers and feed them into a feed-forward layer before feeding them into the RNN. On the other hand, they fine-tune the whole MPC model with downstream tasks.

In their report, Phoneme classification and Speaker Identification were also reported to be benefited from MPC, however, they did not report on SER.

6. Conclusions and Future Work

Our experiments have shown that using MPC features do not increase SER accuracy. Using features from previous layers in the MPC model also do not improve accuracy as well. Furthermore, using gender classification does not improve SER accuracy as well. Unfortunately, it was unable to run fine-tuning experiments on the fine-tuning experiments or using features from all the layers in the MPC model, due to GPU resource limitation. We thus leave this as future work.

References

Bagher Zadeh, AmirAli, Liang, Paul Pu, Poria, Soujanya, Cambria, Erik, and Morency, Louis-Philippe. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208.

Bisio, I., Delfino, A., Lavagetto, F., Marchese, M., and Sciarrone, A. Gender-driven emotion recognition through speech signals for ambient intelligence applications. *IEEE Transactions on Emerging Topics in Computing*, 1(2):244–257, Dec 2013. ISSN 2376-4562. doi: 10.1109/TETC.2013.2274797.

Busso, Carlos, Bulut, Murtaza, Lee, Chi-Chun, Kazemzadeh, Abe, Mower Provost, Emily, Kim, Samuel, Chang, Jeannette, Lee, Sungbok, and Narayanan, Shrikanth. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008. doi: 10.1007/s10579-008-9076-6.

Caruana, Rich. Multitask learning. *Machine Learning*, 28 (1):41–75, Jul 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734.

Chung, Yu-An, Hsu, Wei-Ning, Tang, Hao, and Glass, James. An Unsupervised Autoregressive Model for Speech Representation Learning. *arXiv e-prints*, art. arXiv:1904.03240, Apr 2019.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Harris, Zellig S. Distributional structure. *<i>WORD</i>*, 10(2-3): 146–162, 1954. doi: 10.1080/00437956.1954.11659520.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

Kim, Y. and Kim, J. Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5104–5108, April 2018. doi: 10.1109/ICASSP.2018.8462011.

Lee, Jinkyu and Tashev, Ivan. High-level feature representation using recurrent neural network for speech emotion recognition. 09 2015.

Li, Yuanchao, Zhao, Tianyu, and Kawahara, Tatsuya. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In *Proc. Interspeech 2019*, pp. 2803–2807, 2019. doi: 10.21437/Interspeech.2019-2594.

Liu, Andy T., Yang, Shu-wen, Chi, Po-Han, Hsu, Po-chun, and Lee, Hung-yi. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. *arXiv e-prints*, art. arXiv:1910.12638, Oct 2019.

McFee, Brian, Lostanlen, Vincent, McVicar, Matt, Metsai, Alexandros, Balke, Stefan, Thomé, Carl, Raffel, Colin, Malek, Ayoub, Lee, Dana, Zalkow, Frank, Lee, Kyungyun, Nieto, Oriol, Mason, Jack, Ellis, Dan, Yamamoto, Ryuichi, Seyfarth, Scott, Battenberg, Eric, , , Bittner, Rachel, Choi, Keunwoo, Moore, Josh, Wei, Ziyao, Hidaka, Shunsuke, nullmightybofo, Friesch, Pius, Stöter, Fabian-Robert, Hereñú, Darío, Kim, Taewoon, Vollrath, Matt, and Weiss, Adam. librosa/librosa: 0.7.2, January 2020.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

Naili, Marwa, Habacha, Anja, and Ben Ghezala, Henda. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, 112:340–349, 12 2017. doi: 10.1016/j.procs.2017.08.009.

Neumann, M. and Vu, N. T. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7390–7394, May 2019. doi: 10.1109/ICASSP.2019.8682541.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, April 2015. doi: 10.1109/ICASSP.2015.7178964.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

Peters, Matthew, Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, and Zettlemoyer, Luke. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

Radford, Alec. Improving language understanding by generative pre-training. 2018.

Radford, Alec, Wu, Jeff, Child, Rewon, Luan, David, Amodei, Dario, and Sutskever, Ilya. Language models are unsupervised multitask learners. 2019.

Schneider, Steffen, Baevski, Alexei, Collobert, Ronan, and Auli, Michael. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pp. 3465–3469, 2019. doi: 10.21437/Interspeech.2019-1873.

Tripathi, Suraj, Kumar, Abhay, Ramesh, Abhiram, Singh, Chirag, and Yenigalla, Promod. Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions. *arXiv e-prints*, art. arXiv:1906.05681, Jun 2019.

van den Oord, Aaron, Li, Yazhe, and Vinyals, Oriol. Representation Learning with Contrastive Predictive Coding. *arXiv e-prints*, art. arXiv:1807.03748, Jul 2018.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Polosukhin, Illia. Attention Is All You Need. *arXiv e-prints*, art. arXiv:1706.03762, Jun 2017.