INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speech Emotion Recognition via Multi-Level Cross-Modal Distillation

Ruichen Li$^+$*, Jinming Zhao*$^+$*, Qin Jin**

School of Information, Renmin University of China, China
{ruichen, zhaojinming, qjin}@ruc.edu.cn

Abstract

Speech emotion recognition faces the problem that most of the existing speech corpora are limited in scale and diversity due to the high annotation cost and label ambiguity. In this work, we explore the task of learning robust speech emotion representations based on large unlabeled speech data. Under a simple assumption that the internal emotional states across different modalities are similar, we propose a method called Multi-level Cross-modal Emotion Distillation (MCED), which trains the speech emotion model without any labeled speech emotion data by transferring emotion knowledge from a pretrained text emotion model. Extensive experiments on two benchmark datasets, IEMOCAP and MELD, show that our proposed MCED can help learn effective speech emotion representations which generalize well on downstream speech emotion recognition tasks.

Index Terms:speech emotion recognition, cross-modal transfer, pretraining

1. Introduction

Emotions play the important role in speech communications. Speech emotion recognition has attracted more and more attention from researchers. It has a wide range of applications, such as assisting mental health analysis [1], improving natural human-machine interaction [2] etc. However, the research and application of speech emotion recognition suffer greatly from data scarcity [3]. Current state-of-the-art speech emotion recognition benchmarks have great limitations in terms of naturalness, speaker diversity, annotation quality, and data scale. Previous works have proposed various methods to improve the generalization ability of speech emotion recognition models, including data augmentation [4, 5], semi-supervised learning [6, 7], etc. However, the essential obstacle to the robustness and generalization ability caused by data scarcity remains unsolved.

With the rapid development in the field of computer vision and neutral language processing, the performance of facial emotion recognition and text sentiment recognition has been greatly improved. Therefore, cross-modal transferring-based methods, which transfer expression from one domain (such as visual and text) to another domain (such as speech) through cross-modal distillation, have provided another possibility to solve the problem. Cross-modal distillation aims to transfer supervision and knowledge between different modalities. It normally adopts a teacher-student learning mechanism, where the teacher model is reliably trained on one modality with sufficient supervised data, and then the student model on another modality without supervised data is trained by transferring from the teacher model. The distillation methods usually involve the traditional response-level knowledge distillation [8, 9, 10], which uses the logits as the supervision, and feature-level distillation [11, 12], which encourages a student network to mimic the intermediate representations of a teacher network. Won et al. [13] propose to learn an end-to-end spoken language understanding (student) model using supervision from a BERT language model, which transfers the response-level knowledge from text modality to the audio modality. Gupta et al. [14] propose to learn an image model on depth or flow modality via distilling feature-level knowledge from the RGB modality. For the speech emotion recognition task, Albanie et al. [15] propose a method to train a speech emotion recognition model via response-level distillation from a pre-trained facial emotion recognition model given visual-audio pairs without labeled speech data. However, most above previous works simply apply the coarse-grained supervision for cross-modal distillation, such as only one distillation strategy or coarse-grained feature-level distillation, which can hardly make full use of the information from the teacher model.

In this work, we proposed a Multi-level Cross-modal Emotion Distillation (MCED) method to make better use of the knowledge from the teacher model. In MCED, in addition to using the traditional response-level knowledge distillation, we propose a novel feature-level fine-grained cross-modal transfer method that applies on multiple intermediate layers of the teacher's hidden representations for incremental knowledge extraction. We collect a large dataset with unlabelled audio-text pairs as a bridge to transfer the knowledge from the teacher model to better train a robust speech emotion recognition student model. Different from the visual-audio cross-modal distillation [15], we consider using text-audio cross-modal transfer, which can be more reasonable and efficient as follows: 1) the text modality naturally has a stronger connection with the audio modality as it contains the intention of what the speaker wants to express. 2) movement of lips and facial muscles when speaking may interfere with facial expression recognition, which may result in inaccurate emotion prediction [16, 17, 18]. 3) text modality naturally expresses utterance-level emotion, while facial emotion recognition model is based on frame-level, and how to get the utterance-level emotion through the frame-level is also a difficult problem. Therefore, in this work, we propose an approach to train a speech emotion recognition model via text-audio cross-modal distillation.

The main contributions of this work are: 1) We propose a novel Multi-level Cross-modal Emotion Distillation (MCED) approach to learn the robust speech emotion representations via transferring from the text modality. 2) We collect a large movie data with audio-text pairs for the cross-modal distillation, which will be available at https://github.com/AIM3-RUC/MCED. 3) Our proposed model achieves state-of-the-art performance on both IEMOCAP and MELD datasets.

2. Method

We aim to learn useful emotional speech representations without any labeled speech emotion data, which can benefit downstream low-resource speech emotion recognition tasks. We propose a model with Multi-level Cross-modal Emotion Distilla-

+ Equal Contribution
* Corresponding Author

http://dx.doi.org/10.21437/Interspeech.2021-785

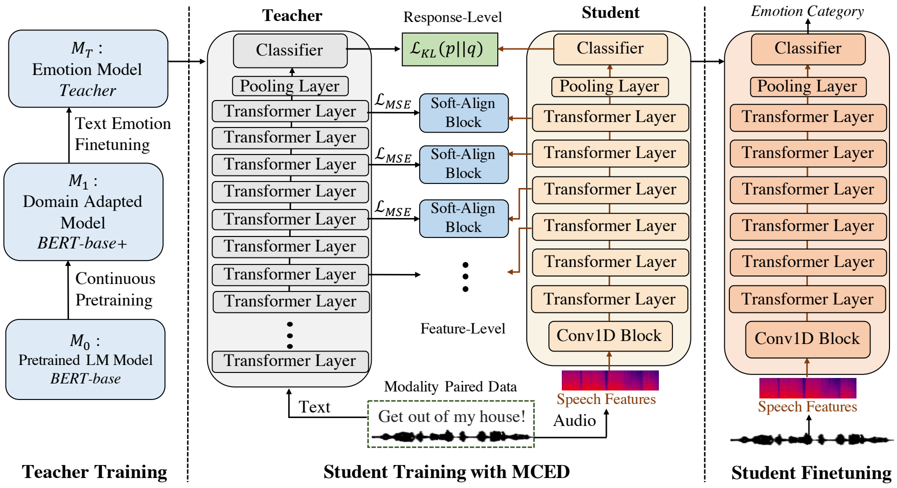

Figure 1: *Illustration of the proposed pipeline, which consists of three phases: 1) training a strong teacher model through a three-stage training strategy 2) training the student model via multi-level cross-modal emotion distillation (MCED) with the pretrained teacher model fixed on large unlabeled audio-text pairs; 3) finetuning the student model on the benchmark datasets for evaluation.*

tion (MCED) based on the teacher-student learning mechanism as shown in Figure 1. A teacher model on the **text** modality is trained through three-stage strategy on a large labeled text emotion recognition corpus first. Then the student model on the **speech** modality is trained through transferring knowledge from the teacher model on a large audio-text pairs dataset without any labels. Finally, the pretrained student model is finetuned on the downstream speech emotion recognition benchmark datasets to evaluate the performance of speech emotion recognition.

2.1. Teacher Model

The teacher model is responsible for providing supervision to train the student speech emotion model. Therefore, building a strong teacher model is critical. We propose a three-stage training strategy: 1) adopt a well-pretrained language model M_0; 2) keep pretraining M_0 on a large domain-related text corpus to build M_1; 3) fine-tune M_1 on a large supervised emotion recognition corpus to build the teacher text emotion model M_T. In this work, we first adopt the pretrained "BERT-base" model[1]. Motivated by the continuous pretraining method [19], we then continue pretraining the BERT-base model on the OpenSubtitle Dataset [20][2] with the pretaining task of masked language modeling to get the domain adapted model "BERT-base+". Finally, we finetune the "BERT-base+" model on a large labeled text emotion recognition corpus (Section 3.1.1) to build the **teacher model**. Once the teacher model is trained, we can get the logits z_i^t from the teacher model for a text sentence L_i, $z_i^t = f(L_i; \theta_t)$, where θ_t are parameters of the teacher model.

2.2. Student Model

The student model (the speech emotion recognition model) aims to learn the emotional speech representations and predict the emotion categories given speech signals. Let A denotes the acoustic feature sequence of an audio, we first use the Conv1D block [21] to capture the local emotional representations of acoustic features. The segment-level local representations from the output of Conv1D block are fed into a transformer encoder network [22] to learn temporal-aware features. A max-pooling layer is used to get the emotion-salient utterance-level representations based on the output of the last layer of the transformer encoder network. Finally, we use several fully connected layers as a classifier to predict the probability distribution of emotion classes. The emotion distribution of the student model can be formulated as $z_i^s = f(A_i; \theta_s)$, where θ_s and z_i^s are the parameters of the student model and the output of its logits layer for the i^{th} sample, respectively.

2.3. Multi-level Cross-modal Emotion Distillation

The proposed multi-level cross-modal emotion distillation (MCED) method involves transferring the response-level and feature-level knowledge from the teacher model to the student model. Following the protocol of classical knowledge distillation [8, 23, 24], we transfer the response-level knowledge from the teacher to the student model through minimizing the distance between the "soft-target" produced by the teacher and the student models. To be specific, we use KullbackLeibler divergence [25] to measure the distance of the "soft-target" emotional probability distribution. The objective function can be formulated as:

$$\begin{aligned} p_i^t &= softmax(z_i^t/temp) \\ p_i^s &= softmax(z_i^s/temp) \\ \mathcal{L}_{RL} &= -\frac{1}{N_{data}} \sum_i KL(p_i^t || p_i^s) \end{aligned} \tag{1}$$

where z_i^t and z_i^s are the logits output from the teacher and student model given i-th audio-text pair. N_{data} denotes the total number of audio-text pairs for training. p_i^t and p_i^s denote the "soft-targets" of the i-th sample produced by the teacher and the student model respectively. $temp$ refers to the temperature hyper-parameter described in [8], which could soften the predicted distribution and better train the student model. By optimizing the response-level distillation loss $\mathcal{L}_{RL}$, higher-level emotional semantic information could be transferred from the teacher to the student model.

Moreover, we propose a feature-level knowledge transfer strategy, so that the student can learn more knowledge from the hidden representations of multiple intermediate layers in

[1] https://github.com/google-research/bert

[2] we use 10 million movie subtitle sentences of the "en-zh" subset.

Table 1: *Data Statistics of Teacher Text Emotion Dataset*

Emotion	Neutral	Happiness	Surprise	Sadness	Anger
Train	10193	3285	2908	2111	3175
Val	1270	419	389	287	481
Test	2889	838	792	564	921

the teacher model. However, the word feature sequence from the text and the acoustic feature sequence produced by the student's CNN block are not aligned. Inspired by [26, 27], we apply attention mechanism to learn the audio-to-text alignment via a soft-align network block, which consist three weight matrices: $W_q \in \mathbb{R}^{d_t \times d_t}$, $W_k \in \mathbb{R}^{d_s \times d_t}$ and $W_v \in \mathbb{R}^{d_s \times d_t}$. W_q performs linear transformation of the query, which is applied as the teacher's hidden states of each word. W_k and W_v are perform linear transformation of the key and value, which transforms the student hidden states to the same dimension as the teacher's. To be specific, we denote the set of intermediate layers in the teacher model for distillation as I^t, and the corresponding transformer layers of the student network as I^s, where $|I^t| = |I^s|$ (eg:$I^t = \{6, 8, 10, 12\}$ and $I^s = \{3, 4, 5, 6\}$). Given the j-th intermediate layer pairs, let us denote $m = I^t_j$ and $n = I^s_j$ for aligning the hidden states of the m-th layer of the teacher model and n-th layer of the student model, and h^t_m and h^s_n for the hidden states of the corresponding layers respectively, where $h^t_m \in \mathbb{R}^{T^t \times d_t}$, $h^s_n \in \mathbb{R}^{T^s \times d_s}$, and $T^{(\cdot)}$ and $d_{(\cdot)}$ represent the sequence length and feature dimension respectively. The calculation of the soft-align block can be formulated as:

$$\hat{h}^s_n = softmax(\frac{h^t_m W_q \sigma(h^s_n W_k)^\top}{\sqrt{d_t}})\sigma(h^s_n W_v) \tag{2}$$

where σ denotes the GELU function, $\hat{h}^s_n$ denotes the soft aligned acoustic sequences, which has the same shape with the teacher's word hidden states h^t_m. We use mean square loss (MSE) to measure the feature-level distance. The loss function can be formulated as:

$$\mathcal{L}_{FL} = \frac{1}{N_{data}} \sum_i \sum_j^{|I^t|} \lambda(j) \left\| h^t_{i,m} - \hat{h}^s_{i,n} \right\|_2^2 \tag{3}$$

where λ denotes the hyper-parameters of the loss weight for each intermediate layer pairs. Fine-grained knowledge could be transferred by optimizing the word aligned feature-level objective function $\mathcal{L}_{FL}$. We jointly optimize the objective function to train the student model, with the fixed parameters of the teacher model, which can be formulates as:

$$\theta_s = \arg\min_{\theta_s}(\mathcal{L}_{RL} + \mathcal{L}_{FL}) \tag{4}$$

3. Experiments

3.1. Datasets

3.1.1. Teacher Text Emotion Datasets

To build a strong and robust teacher model, we combine two text emotion recognition corpora XED [28] and EmotionLines [29]. The XED dataset contains multilingual emotion annotated movie subtitles and we only use the utterances in English. The EmotionLines dataset contains utterances with emotion annotation in dialogue and we only use the Friends part of EmotionLines. Moreover, in order to combine these two datasets and build a class-balanced training set, we choose the 5 most common emotion categories (neutral, happiness, surprise, sadness, and anger) to form a training set for the teacher model, which we name as "Teacher Text Emotion Dataset". The data distribution is shown in Table 1.

Table 2: *Data statistics of the collected EmoATMovie dataset. The emotion category of each utterance is predicted by the teacher model as mentioned in Section 2.1*

Emotion	Neutral	Happiness	Surprise	Sadness	Anger
Train	69562	42509	35651	32992	40191

Table 3: *Data Statistics of MELD and IEMOCAP*

Emotion	MELD Train	Dev	Test	IEMOCAP
Neutral	4710	470	1256	1708
Happy	1743	163	402	1636
Anger	1109	153	345	1103
Sadness	683	111	208	1084
Surprise	1205	150	281	–
Disgust	271	22	68	–
Fear	268	40	50	–

3.1.2. EmoATMovie Dataset

Training a robust student model via transferring knowledge from the teacher model requires modality paired data. We collect 351 movies and TV shows and extract the soundtrack and corresponding subtitles of each spoken utterance. We filter out the utterance whose duration is shorter than 1.5sec or the transcript with only one word. In order to ensure the utterances are emotion salient, we feed each utterance text into the pretrained teacher model to predict its emotion category. We then select the emotion salient utterances based on the prediction probability. In the end, we collect a dataset called "EmoATMovie", which contains about 221k utterances with paired audio and text transcript. The detailed information of EmoATMovie is shown in Table 2.

3.1.3. Benchmark Datasets

The Multimodal EmotionLines Dataset (MELD) dataset [30] and Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [31] are two popular emotion recognition benchmark datasets, which are used to evaluate our student model. MELD contains more than 1300 dialogues and 13000 utterances from the Friends TV series with 304 speakers in total. We follow the standard data split [30], which contains 7 emotion categories. Note that there's no overlap between the valid/test set of MELD and the training set of Teacher Text Dataset in text modality. IEMOCAP dataset contains recorded videos from 10 speakers in 5 dyadic conversation sessions. Each session contains dialogues between 2 speakers. We follow the data split setting as in [26, 6] to form the four-class (angry, happy, neutral and sad) emotion recognition task. The statistics of the two datasets are shown in Table 3.

3.2. Implementation Details

For audio features, we extract 130 dimensional low-level descriptors using OpenSmile Toolkit [32] with the configuration of "IS16_ComparE". For text input, we tokenize the sentence into a sequence of WordPieces [33]. The teacher model is built by finetuning BERT-base+ on the Teacher Text Emotion Dataset for max 6 epochs and the BERT-base+ is obtained by continuous pretraining BERT-base for 20 epochs on the OpenSubti-

Table 4: *Performance comparison of teacher model candidates on Teacher Text Emotion Dataset*

Models	WA	WF1
Finetune BERT-base	0.686	0.685
Finetune BERT-base+ (Our Teacher)	**0.696**	**0.694**

Table 5: *Speech emotion recognition performance comparison on IEMOCAP and MELD datasets.*

	IEMOCAP		MELD	
	Approaches	UAR	Approaches	WF1
1	ARE [34]	0.580	ICON [35]	0.373
2	LSTM+ATT [36]	0.588	BC-LSTM [37]	0.364
3	CNN+LSTM [38]	0.602	DialoueRNN [39]	0.340
4	Zhao et.al [40]	0.601	ConGCN [41]	0.422
5	SSMM [6]	0.585	SSMM [6]	0.402
6	Student−MCED	0.589	Student−MCED	0.372
7	Student+MCED	**0.614**	Student+MCED	**0.446**

tle subset. The student model consists of 4 Conv1D layers as Conv1D block, 6 transformer layers (4 heads and 512 hidden units), and 2 fully-connected layers as the classifier. For the training parameters of the cross-modal distillation, we set $I_t = \{6, 8, 10, 12\}$, $I_s = \{3, 4, 5, 6\}$ and $\lambda = \{0.1, 0.1, 0.2, 0.3\}$. The student model is trained for at most 50 epochs on EmoAT-Movie with an initial learning rate of 1e-4, which decays linearly during training. When finetuning the pertained student model, we use different classifiers on the benchmark datasets because of different emotion categories.

We use the 10-fold speaker-independent cross-validation for evaluation. We run each model three times to alleviate the influences of random parameter initialization on both IEMOCAP and MELD. We use two evaluation metrics on IEMOCAP: weighted accuracy (WA) and unweighted average recall (UAR). Due to emotion category imbalance on MELD, we use the weighted F1 (WF1) as the evaluation metric.

3.3. Experiment Results

3.3.1. Teacher Model

A strong teacher model is critical for training the student model. We compare the performance of different teacher model candidates in Table. 4. As mentioned in Section 2.1, our teacher model "Finetune BERT-base+" is built through the three-stage training strategy and "Finetune BERT-base" is built through the first M_0 and the third M_T stage. The results show that the extra pretraining stage (M_1 in Figure 1) can further improve the teacher model.

3.3.2. Student Model

In order to evaluate our proposed MCED method, we conduct downstream speech emotion recognition experiments on the two benchmark datasets. We compare our student model (row 7) to other state-of-the-art models (row 1-5) in Table. 5. Our student model outperforms other SOTA models on both benchmark datasets. We also compare the performance of the student model without MCED based pretraining (row 6), which means that the model has the same structure as our student model but is trained from scratch on the benchmark training data. The performance drop of "Student−MCED" compared to "Student+MCED" demonstrates that pretraining the student model on large unlabeled data with our proposed MCED is effective. Our pretrained student model can be easily applied to various downstream speech emotion recognition tasks.

Table 6: *Ablation study of MCED. RL and FL refer to response-level and feature-level distillation respectively. FL-klayer means using the last k layers for feature-level distillation.*

Dataset	IEMOCAP		MELD	
Models	UAR	F1	WA	WF1
RL	0.599	0.573	0.496	0.438
FL-4layers	0.593	0.570	0.481	0.423
RL+FL-1layer	0.608	0.588	0.499	0.434
RL+FL-2layers	0.603	0.591	**0.501**	0.438
RL+FL-3layers	0.611	0.592	0.482	0.442
RL+FL-4layers	**0.614**	**0.601**	0.497	**0.446**

Table 7: *Ablation study on the amount of finetuning data. "1/2" means we only use half of the training data in the corresponding benchmark dataset when finetuning.*

Dataset	Ratio	IEMOCAP		MELD	
Settings		UAR	F1	WA	WF1
Student+MCED	1/2	0.590	0.571	0.470	0.424
	1/3	0.577	0.545	0.479	0.394
	1/5	0.541	0.525	0.468	0.377
Student−MCED	full	0.589	0.569	0.458	0.372

3.4. Ablation Study

Different level distillation: We conduct experiments to ablate the contributions of different components of our proposed Multi-level Cross-modal Emotion Distillation method as shown in Table 6.The table shows that the model trained with multi-level distillation, including both response-level and feature-level, outperforms the one with only single level distillation, and the performance keeps increasing as we use more layers to perform the soft-aligned feature-level distillation, which demonstrates the effectiveness of our proposed MCED method.

Amount of Finetuning Data: In order to validate the generalization ability of our pretrained student model, we conduct experiments on benchmark datasets using different amounts of training data in the finetuning process as shown in Table 7, where $1/n$ means we only use $1/n$ of the training set to finetune our pretrained student model. Our student model finetuned with only 1/5 of the training data on MELD and 1/2 of the training data on IEMOCAP can outperform the "Student−MCED" trained on the full training set, which demonstrates that our pretrained student model with MCED method can generalize well even with limited labelled data.

4. Conclusion

In this paper, we propose a novel multi-level cross-modal emotion distillation (MCED) method to train a student speech emotion model without any labeled speech data through transferring from a teacher text emotion model. The pretrained student model can benefit downstream emotion recognition tasks, especially in low-resource supervised training conditions. Extensive experiments on two public benchmark datasets demonstrate the effectiveness and robustness of our proposed model.

5. Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2020AAA0108600), National Natural Science Foundation of China (No. 62072462 & 61772535), Beijing Natural Science Foundation (No. 4192028).

6. References

[1] Fabien Ringeval and Björn Schuller and Michel Valstar and others, “AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition,” in *ACM MM*. Seoul, Korea: ACM, October 2018.

[2] N. Fragopanagos and J. G. Taylor, “Emotion recognition in human-computer interaction.” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2002.

[3] B. Schuller, S. Steidl, A. Batliner *et al.*, “Paralinguistics in speech and languagestate-of-the-art and the challenge,” *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[4] C. Etienne, G. Fidanza, A. Petrovskii *et al.*, “Cnn+ lstm architecture for speech emotion recognition with data augmentation,” *arXiv preprint arXiv:1802.05630*, 2018.

[5] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos *et al.*, “Data augmentation using gans for speech emotion recognition.” in *Interspeech*, 2019, pp. 171–175.

[6] J. Liang, R. Li, and Q. Jin, “Semi-supervised multi-modal emotion recognition with cross-modal distribution matching,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2852–2861.

[7] S. Parthasarathy and C. Busso, “Semi-supervised speech emotion recognition with ladder networks,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2697–2709, 2020.

[8] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[9] T. Afouras, J. S. Chung, and A. Zisserman, “Asr is all you need: Cross-modal distillation for lip reading,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2143–2147.

[10] V. Sanh, L. Debut, J. Chaumond *et al.*, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.

[11] A. Romero, N. Ballas, S. E. Kahou *et al.*, “Fitnets: Hints for thin deep nets,” *arXiv preprint arXiv:1412.6550*, 2014.

[12] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, “A comprehensive overhaul of feature distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1921–1930.

[13] W. I. Cho, D. Kwak, J. W. Yoon *et al.*, “Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation,” in *Proc. Interspeech*, 2020, pp. 896–900.

[14] S. Gupta, J. Hoffman *et al.*, “Cross modal distillation for supervision transfer,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.

[15] S. Albanie, A. Nagrani, A. Vedaldi *et al.*, “Emotion recognition in speech using cross-modal transfer in the wild,” in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 292–301.

[16] S. Mariooryad and C. Busso, “Factorizing speaker, lexical and emotional variabilities observed in facial expressions,” in *ICCV*. IEEE, 2012, pp. 2605–2608.

[17] C. Busso and S. S. Narayanan, “Interrelation between speech and facial gestures in emotional utterances: a single subject study,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007.

[18] S. Mariooryad and C. Busso, “Facial expression recognition in the presence of speech using blind lexical compensation,” *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 346–359, 2015.

[19] S. Gururangan, A. Marasović, S. Swayamdipta *et al.*, “Dont stop pretraining: Adapt language models to domains and tasks,” in *ACL*, 2020, pp. 8342–8360.

[20] P. Lison and J. Tiedemann, “Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 923–929.

[21] M. Steinschneider, K. V. Nourski, and Y. I. Fishman, “Representation of speech in human auditory cortex: is it special?” *Hearing research*, vol. 305, pp. 57–73, 2013.

[22] A. Vaswani, N. Shazeer, N. Parmar *et al.*, “Attention is all you need,” in *NIPS*, 2017.

[23] L. J. Ba and R. Caruana, “Do deep nets really need to be deep?” in *NIPS Systems-Volume 2*, 2014, pp. 2654–2662.

[24] J. Li, R. Zhao, J.-T. Huang *et al.*, “Learning small-size dnn with output-distribution-based criteria,” in *INTERSPEECH*, 2014.

[25] S. Kullback and R. A. Leibler, “On Information and Sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79 – 86, 1951.

[26] H. Xu, H. Zhang, K. Han *et al.*, “Learning alignment for multimodal emotion recognition from speech,” *Proc. Interspeech 2019*, pp. 3569–3573, 2019.

[27] Y.-H. H. Tsai, S. Bai, P. P. Liang *et al.*, “Multimodal transformer for unaligned multimodal language sequences,” in *ACL. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[28] E. Öhman, M. Pàmies, K. Kajava *et al.*, “Xed: A multilingual dataset for sentiment analysis and emotion detection,” in *ACL*, 2020, pp. 6542–6552.

[29] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo *et al.*, “Emotionlines: An emotion corpus of multi-party conversations,” *arXiv preprint arXiv:1802.08379*, 2018.

[30] S. Poria, D. Hazarika, N. Majumder *et al.*, “Meld: A multimodal multi-party dataset for emotion recognition in conversations,” in *ACL*, 2019, pp. 527–536.

[31] C. Busso, M. Bulut, C.-C. Lee *et al.*, “Iemocap: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[32] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *ACM Multimedia*, 2010, pp. 1459–1462.

[33] J. Devlin, M.-W. Chang, K. Lee *et al.*, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[34] S. Yoon, S. Byun, and K. Jung, “Multimodal speech emotion recognition using audio and text,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[35] D. Hazarika, S. Poria, R. Mihalcea *et al.*, “Icon: interactive conversational memory network for multimodal emotion detection,” in *EMNLP*, 2018, pp. 2594–2604.

[36] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *ICASSP*. IEEE, 2017, pp. 2227–2231.

[37] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, “Context-dependent sentiment analysis in user-generated videos,” in *ACL (volume 1: Long papers)*, 2017, pp. 873–883.

[38] S. Latif, R. Rana, S. Khalifa *et al.*, “Direct modelling of speech emotion from raw speech,” *arXiv preprint arXiv:1904.03833*, 2019.

[39] N. Majumder, S. Poria, D. Hazarika *et al.*, “Dialoguernn: An attentive rnn for emotion detection in conversations,” in *AAAI*, vol. 33, 2019, pp. 6818–6825.

[40] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, “Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition,” in *Proc. Interspeech*, 2018, pp. 272–276.

[41] D. Zhang, L. Wu, C. Sun *et al.*, “Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations,” in *IJCAI*, 2019, pp. 10–16.