Improving speech emotion recognition with data expression aware multi-task learning

Pooja Kumawat, Aurobinda Routray
Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India
pk28@iitkgp.ac.in,aroutray@ee.iitkgp.ac.in

***Abstract*—Majority of speech emotion recognition (SER) systems are developed using databases with simulating speech performed by professional actors. Whereas, concerning real-world deployment, the SER inputs are mostly spontaneous utterances. Several SER researchers reported that the performance of SER models developed using acted emotional data degrades for spontaneous inputs. In this work, we improve the SER performance under the elicitation-based data expression mismatch scenarios by utilizing multi-task learning (MTL) with data expression recognition as the auxiliary task. We use the ECAPA-TDNN architecture with MFCCs and wav2vec 2.0 pre-trained embeddings as features. We conduct this study on the IEMOCAP and BAUM-1 databases. The proposed MTL-based method achieves state-of-the-art performance on the SER task. Further, we conduct an emotion-specific analysis and show that the data expression knowledge mostly helps to classify the highly aroused emotions.**

***Index Terms*—Speech emotion recognition, Data expression mismatch, Acted and spontaneous emotions, Self-supervised features, Multi-task learning.**

I. Introduction

Speech emotion recognition (SER) is an important task of affective computing, that aims to enable the machine to automatically recognize the human emotions from only audio information. The trend of verbal human-to-computer interaction (HCI) is gradually increasing with the evolution of smart applications. With this, the relevance of effective SER systems is increasing. SER is utilized in several real-world applications such as for voice assistants, conversation monitoring, and analysis in call centers or interview scenarios [1].

A major challenge for developing effective SER systems is to generalize for domain variations, majorly caused by mismatches in speakers, languages, background noise, and microphones. Many notable research works attempted to address this issue of domain-sensitive SER performance using cross-speaker, cross-language, cross-dataset, cross-modality frameworks [2]–[4]. However, another important aspect of SER domain variability lies in the form of the emotion elicitation process. From the existing SER literature, we observe that the majority of the widely used SER databases contain acted emotions [5], [6]. The data collection protocols of these databases make professional actors perform different emotions on some predefined scripts. Due to the trained manner of expressing the emotions and the standard articulations, such data contains well-defined distinctions among the emotions [7]. However, considering real-world usability, the inputs to the SER models trained with acted data are more likely to contain spontaneous (spont.) utterances. The perceivable emotions under such situations may not be very well-differentiated [8]. For example, for any specific emotions, the pitch contours may vary significantly for acted data, but for spontaneous, it may remain relatively flat [9]. Different studies reported that the performance of SER models developed using acted emotional data degrades for spontaneous inputs, which is not desirable for practical SER applications [8]–[11]. Hence, addressing the impact of data expressions, such as acted or spontaneous, due to various emotion eliciting approaches is important.

In the SER literature, few studies have attempted to address these data expression mismatch issues. In [12], Li et al. investigated various transfer learning techniques using a feed-forward neural network and progressive neural network for leveraging acted speech data to improve emotion recognition of spontaneous speech. In [13], Feng et al. used a few-shot learning approach that transfers emotion-specific knowledge from acted to spontaneous speech. In [14], Li et al. investigated transfer learning between utterances with fixed scripted lexical contents and utterances recorded spontaneously. They used domain adversarial training with softlabel loss. However, while trying to adapt for target data, the domain adaptation based methods have been reported to degrade performance on utterances coming from domains similar to the source-data.

In this work, concerning the above mentioned factors, we follow a different approach for the assessment and improvement of SER generalization against data expression mismatch conditions. Instead of focusing only on the spontaneous target domain, we aim to improve the overall SER system performance by utilizing elicitation-dependent emotional cues. As a preliminary experiment, we first train three independent SER systems with acted, spontaneous, and combined training data and evaluate them on acted and spontaneous test utterances separately. Apart from the earlier reported [8], [9] performance mismatch caused by data expression-mismatch, we further observe that compared to the combined model, SER systems can further improve the performance if it individually focuses on elicitation-specific emotion discriminating cues. This observation motivates us to induce data expression discriminating knowledge during the SER training phase so that utterances of each kind of data expression can focus on their own emotion discriminating cues. Therefore, we develop *Multi-Task Learning* (MTL) based SER systems with elicitation-dependent data expression recognition as additional tasks.

In SER literature, MTL has been explored with different additional tasks, such as identification of gender [15], corpus domain recognition [4]. However, the exploration of elicitation-based data expression mismatch is little explored. Mangalam et al. [16] used hierarchical classifier and spontaneity detection based MTL tasks in SER. Zhang et al. [17] considered acted and spontaneous utterances from two different databases of different languages and showed that merging and applying MTL on them helps improving generalization. However, by doing so, we suspect that database mismatches, such as differences in language, gender, and age distributions of speakers, recording environment, and equipment, can aid the MTL performance with an overfitted discrimination between acted and spontaneous classes [18]. Hence, it is not clear whether the elicitation-aware knowledge is effectively incorporated in the SER system. To mitigate this issue, we utilize the acted and spontaneous utterances within the same database.

The contribution of our work is the following: (1.) We develop elicitation-based data expression aware SER systems under a more practical consideration, where both acted and spontaneous data are collected from the same database. (2.) We validate the effectiveness of the proposed within-corpora MTL-based SER independently on two different databases and languages. (IEMOCAP-English database and BAUM-1-Turkish database). We then compare our system by developing an inter-corpora MTL system, where acted and spontaneous utterances are collected from different databases (languages). (3.) We conduct an extensive emotion-specific study revealing recognition of what kinds of emotions benefit the most by incorporating elicitation-awareness using the MTL framework.

The rest of this work is organized as follows: Section II provides the database descriptions. In Section III, the methodologies are discussed. Experimental setups and results are presented in Section IV followed by the conclusions in Section V.

II. Database description

We have used two different emotional speech databases in this work. These are the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and the BAUM-1 database. In both databases, audio recordings are collected in two different scenarios: scripted play (acted) and spontaneous dialogs.

IEMOCAP database: IEMOCAP [19] database is a large and widely used multimodal conversational English database in SER research. It consists of five dyadic sessions performed by ten speakers, each with a scripted (acted) and improvised (spontaneous) interaction. This database includes speech, facial recordings, and text transcriptions for anger, happy, neutral, sad, fear, disgust, frustration, excitement, and surprised emotions. Based on prior research [20], we only use the audio data from anger, excitement, happy, neutral, and sad emotion classes and merge the excitement utterances with the happy emotion class to deal with the data imbalance. The final dataset has 5,531 utterances (angry 1,103, happy 1,636, neutral 1,708, sad 1,084). Out of 5,531 utterances, 2588 utterances are scripted, and 2,934 are improvised. The database is sampled at 16 kHz, with a mean utterance duration of 4.52s.

BAUM-1 database: BAUM-1 [21] database is an emotional audio-visual face database of affective and mental states in the Turkish language. This database includes 199 acted audio utterances in BAUM-1a data and 501 spontaneous audio data in BAUM-1s data representing six (anger, boredom, disgust, fear, happiness, and sadness) emotional states. The audio samples in this database are collected from 31 subjects which are recorded at 44.1 kHz sampling rate. The mean duration of the audio files in this database is 4.10s.

III. Methodology

A. Data preprocessing and feature extraction

We have downsampled the BAUM-1 database to 16 kHz sampling frequency before the feature extraction step. To obtain reliable results, we conducted 5-fold cross-validation (CV) approach for both databases and reported the average results [22]. For the IEMOCAP database, we perform leave-one-session-out 5-fold CV. At each fold, four sessions are used for training and validation sets, and the remaining one session is used for testing set [23]. For the BAUM-1 database, we use an 80:10:10 ratio split training, validation, and testing approach for each fold of 5-fold CV.

As a baseline feature, we extract the most widely used 40-dimensional MFCCs speech features with a hamming window of 20ms and a hop size of 10ms. Following the effectiveness of self-supervised learning in different speech processing applications [24], we primarily focus on using wav2vec 2.0 pre-trained model ("wav2vec2-large-960h") [25] as a feature extractor in our work. wav2vec 2.0 optimizes a contrastive predictive coding (CPC) based loss function during pre-training with 960 hours of unlabeled Librispeech data. It consists of three stages: (1.) local encoder: convolutional blocks that extract embeddings from raw waveform as latent representation, (2.) contextualized encoder: consisting of 24 transformer blocks with 16 attention heads, (3.) quantization module: for discretizing latent representation of local encoder.

B. Classifier architecture description

In this work, we use the ECAPA-TDNN [26] architecture to develop the SER systems. In our previous study [27], we have shown that this architecture outperformed the conventional x-vector based TDNN architecture. ECAPA-TDNN introduces 1-dimensional Squeeze-Excitation Residual block (SE-Res2Block), which models the channel interdependencies and enhances the multi-scale local contexts. We have used 512 channels in the convolutional frame layers and 192 nodes in the final fully connected layer. The outputs from the shallower level SE-Res2Block are concatenated and then fed to the next Conv1D+ReLU layer. Instead of statistical pooling of the x-vector, ECAPA-TDNN applies attention mechanism across the channels and captures global utterance characteristics in the pooling layer to obtain an utterance-level feature vector, i.e., embeddings. The embeddings learned from this model carry more robust target-dependent information due to the *Multi-layer Feature Aggregation* (MFA). We use Adam as the optimizer with a learning rate of 0.001 and Additive Angular

Margin (AAM)-softmax as loss function. The batch size is set to 64, and the number of epochs is limited to 50.

IV. Experiments & Results

The performance of the SER systems are evaluated using both *Weighted Accuracy* (WA) and *Unweighted Accuracy* (UWA) as the test sets for both the databases are imbalanced between different emotion categories.

A. Preliminary experiment: assessment of performance degradation due to data expression mismatch

In the preliminary experiment, out of the five sessions in the IEMOCAP database, we keep the first four sessions for system training and the fifth one for evaluation. From the training set, we separate the acted and spontaneous utterances and create three different training sets: (i) acted training set, (ii) spontaneous training set, (iii) combined training set. Similarly, the test set is also segregated into acted and spontaneous parts. In Table I, we present the number of utterances for each emotion in the acted and spontaneous training and test sets. The training set is split randomly in training and validation parts by $80:20$ ratio and individual SER systems are trained using the MFCC feature and ECAPA-TDNN architecture.

TABLE I: Number of acted and spontaneous utterances in the preliminary experiment for each emotion class in the IEMOCAP training and testing sets.

Emotion	**Training**			**Testing**		
	Total	Acted	Spont.	Total	Acted	Spont.
Anger	933	675	258	170	139	31
Happy	1194	527	667	442	162	280
Neutral	1324	512	812	384	97	287
Sad	839	364	475	245	112	133

The results of the preliminary experiment are presented in Table II. From this Table, we observe the performance degradation due to the data expression mismatch in train-test pairs. Compared to the combined trained model, the accuracy of the acted trained model improves by 5.50% on the acted test data, and the accuracy of the spontaneous trained model improves by 4.15% on the spontaneous test data. The acted and spontaneous trained models learn their respective expression-specific emotion discriminating cues, and that results in better SER performances compared to the combined trained models. From the results, we hypothesize that if we incorporate the data expression awareness into our combined SER model, the overall SER performance can be improved.

TABLE II: SER classification accuracy (in $\%$) of the preliminary experiment with the IEMOCAP database.

Feature	Testing	Training		
		Combined	Acted	Spont.
MFCC	Acted	50.57	**56.07**	47.70
	Spont.	54.70	50.33	**58.85**

B. Data expression recognizing multi-task learning (MTL) in ECAPA-TDNN based SER model

To incorporate the data expression specific knowledge in our system, we apply the MTL approach in the ECAPA-TDNN architecture by hard-parameter sharing. After the channel attentive pooling layer in ECAPA-TDNN, we insert another parallel classifier branch consisting of a fully-connected layer followed by a softmax layer. We choose the binary classification of elicitation-based data expression recognition as the additional task. MTL enables the ECAPA-TDNN model to jointly learn both the tasks by optimizing a combined loss function corresponding to the two tasks, $\mathcal{L}_{emo}$ (loss of emotion prediction) and $\mathcal{L}_{exp}$ (loss of expression prediction). The total loss to be optimized is expressed as:

$$\mathcal{L} = (1-\alpha) * \mathcal{L}_{emo} + \alpha * \mathcal{L}_{exp} \quad (1)$$

α decides the relative weights of both tasks. We experiment with different choices of α and set it empirically to 0.1 with a linear increase of 0.05 in each epoch. The working principle of applying MTL in the ECAPA-TDNN classifier is shown in Fig. 1. The MTL-incorporated ECAPA-TDNN is trained with the similar training protocols as stated in Section III-B.

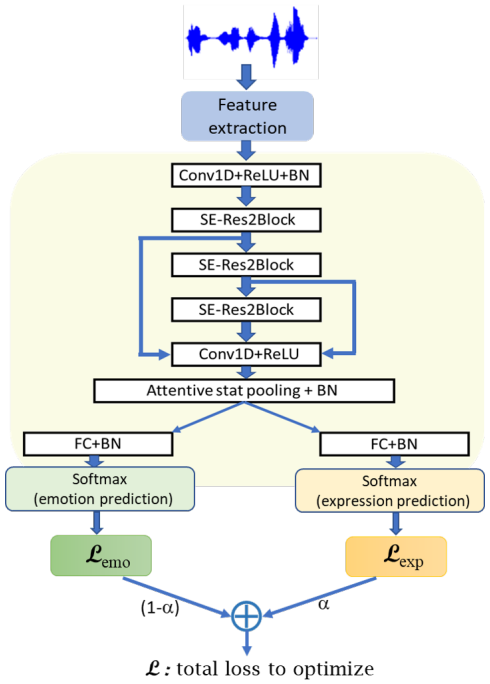

Fig. 1: Working principle of the ECAPA-TDNN architecture with the data expression recognition based MTL extension.

TABLE III: Performance of IEMOCAP and BAUM-1 databases for two different features in terms of WA(%) and UWA(%).

Feature	Method	IEMOCAP		BAUM-1	
		WA	UWA	WA	UWA
MFCC	w/o. MTL	54.25	53.78	50.24	44.12
	w. MTL	55.43	55.98	53.08	46.89
wav2vec 2.0	w/o. MTL	67.71	68.80	50.82	48.51
	w. MTL	72.79	72.82	59.02	53.92

For both the databases, we train the individual MTL-extended ECAPA-TDNN models corresponding to each fold and present the averaged SER performances in Table III.

This Table also includes the SER performances without the MTL for comparative analysis. Table III shows that for the MFCC feature, the MTL experiments improve the SER WA by 1.18% and 2.84% for the IEMOCAP and BAUM-1 databases, respectively. For wav2vec 2.0 feature, the WA improvements due to MTL is 5.08% and 8.20%, for IEMOCAP and BAUM-1 database, respectively. The performance improvements across multiple databases and features prove the effectiveness of our hypothesis that inducing data expression awareness while learning emotion discrimination is useful. The relative performance improvements due to MTL inclusion are more for the wav2vec 2.0 features compared to MFCC features. The self-supervised pre-training of the wav2vec 2.0 architecture enables the extracted features to carry more generalized underlying audio representations [25]. It leads to better capturing of the data expression distinctive information for the wav2vec 2.0 extracted features. On IEMOCAP database, incorporating our MTL method in SER with wav2vec 2.0 feature achieves 72.82% UWA, which is comparable to the previous works as shown in Table IV. For BAUM-1 database, due to different data organization and evaluation protocols, we are unable to compare the results with the existing literature. However we observe significant improvements in the performance of the BAUM-1 database by introducing MTL.

TABLE IV: Performance comparison for IEMOCAP database.

Method	Modalities	CV	UWA (%)
Peng et al. [28]	Audio	5-fold	62.60
Liu et al. [23]	Audio	5-fold	70.78
Sajjad et al. [29]	Audio	5-fold	72.25
Ours (w/o. MTL)	Audio	5-fold	**68.80**
Ours (w. MTL)	Audio	5-fold	**72.82**

C. Quantitative method explaining how the interior modulation of MTL improves SER performance

To quantitatively explain the interior modulation of the MTL framework, using wav2vec 2.0 feature, we compute the pattern separability for the emotion recognition branch's embedding for the two classifiers with and without MTL. From these SER models, embeddings are extracted for the evaluation set of IEMOCAP and BAUM-1 database. For N emotions, to compute pattern separability, we measure the average of ${}^{N}C_{2}$ pair-wise Fisher discriminant ratios of the embeddings.

We also measure pattern separability for the embeddings of two elicitation-based data expression categories. As shown in Fig. 2, with MTL, the emotion embeddings show higher emotion-class separability and greater robustness to data expression mismatch due to reduced expression-class separability. This quantitative experiment shows the effectiveness of the developed MTL-based SER system.

D. Criticality of using acted and spontaneous speech recorded under the same condition

A key contribution in this paper compared to [17] is the choice to use datasets that have both acted/spontaneous speech recorded under the same condition. To demonstrate the criticality of this change, we simulate the condition of selecting individual types of data expressions from different

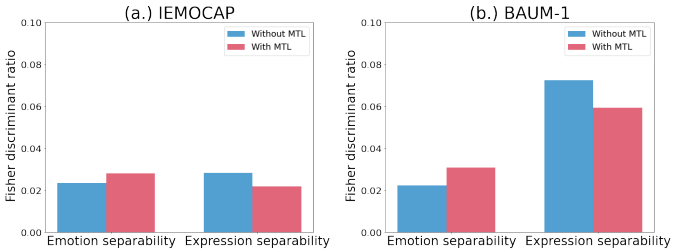

Fig. 2: Pattern separability of emotion and data expression embeddings with and without applying MTL.

datasets. We have applied the MTL on two cross-dataset settings, (i) by selecting acted and spontaneous speech utterances from IEMOCAP and BAUM-1 databases, respectively (Cross-dataset1), and (ii) spontaneous and acted speech utterances from IEMOCAP and BAUM-1 databases, respectively (Cross-dataset2). We consider the common emotions (angry, happy, sad) between the two datasets for the cross-dataset settings. We then present the data expression recognition accuracy (ERA) results for same-dataset and cross-dataset settings in Table V. For cross-dataset settings, we suspect that the additional task does not purely incorporate elicitation-aware information in the SER performance. This is because other dataset mismatch factors, such as language, speaker, and recording environment, can also contribute as discriminating cues in the additional task, which is reflected by the prominently higher ERA (nearly 100%) for the cross-dataset setting as compared to the same-dataset setting.

TABLE V: Performance of the same-dataset and cross-dataset MTL frameworks for the IEMOCAP and BAUM-1 databases.

Corpora setting	Same-dataset				Cross-dataset			
Method	IEMOCAP		BAUM-1		Cross-dataset1		Cross-dataset2	
	WA	UWA	WA	UWA	WA	UWA	WA	UWA
w/o. MTL	67.71	68.80	50.82	48.51	80.49	80.74	80.15	80.67
w. MTL	72.79	72.82	59.02	53.92	83.23	82.80	82.33	81.12
ERA (%)	88.72		87.50		98.30		98.15	

E. Individual emotion class specific analysis

In Table VI, we present the UWA of each emotion class for both the databases. Here, our key focus is the relative performance improvements for the individual emotion classes due to the MTL incorporation. From the results, we observe that for IEMOCAP and BAUM-1 databases with MFCC feature, the highest performance improvement of 7.66% and 6.67% is achieved for the angry emotion, respectively. Similarly, with wav2vec 2.0 feature, the highest performance improvement of 11.94% and 14.29% is achieved for the angry emotion for IEMOCAP and BAUM-1 database, respectively. We further observe that apart from angry, the happy emotion also gains noticeable performance improvements due to the MTL training. Whereas, we find that there is lesser variation in the SER performance for the neutral emotions due to the inclusion of MTL. These observations reveal a trend that the prediction performance of the highly aroused emotion classes attains the maximum benefit due to data expression-aware MTL training for both databases.

TABLE VI: Emotion-specific SER performance analysis for the inclusion of data expression recognizing MTL approach.

Database	Feature	Model	True Positive Rate (TPR) (%)							UWA (%)
			Angry	Happy	Neutral	Sad	Disgust	Fear	Boredom	
IEMOCAP	MFCC	w/o. MTL	59.41	45.57	55.75	54.40	-	-	-	53.78
		w. MTL	67.07	47.26	53.90	55.69	-	-	-	55.98
	wav2vec 2.0	w/o. MTL	68.65	59.78	75.45	71.33	-	-	-	68.80
		w. MTL	80.59	64.67	72.69	73.33	-	-	-	72.82
BAUM-1	MFCC	w/o. MTL	47.30	27.77	-	28.33	61.04	35.10	65.19	44.12
		w. MTL	53.97	25.15	-	36.22	64.35	36.28	65.37	46.89
	wav2vec 2.0	w/o. MTL	78.57	71.43	-	56.86	29.43	33.33	21.43	48.51
		w. MTL	92.86	74.51	-	56.50	29.41	41.67	28.57	53.92

V. Conclusions

In the literature, researchers have reported that the performance of the SER systems, trained with acted data, degrades when spontaneous utterances are used for evaluation. In this work, we first analyze this performance degradation due to the data expression mismatch. Our analysis also shows that the overall SER performance can potentially improve by separately focusing on elicitation-based data expression cues. Based on this observation, we extend the SER classifier using multi-task learning with data expression recognition as the auxiliary task. The results show that across multiple databases, the MTL approach prominently improves the overall SER performance by achieving results that are comparable with the state-of-the-art. Finally, we present an emotion specific analysis and reveal a trend that the highly aroused emotions benefit the most in SER performance due to the MTL inclusion. Further investigation into this can be a promising future research direction. This study addresses a practical challenge in SER research and enhances the efficient applicability of SER systems for real-world scenarios.

References

[1] R. Cowie, E. Douglas-Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Context-dependent domain adversarial neural network for multimodal emotion recognition," in *Proc. of INTERSPEECH*, 2020, pp. 394–398.

[3] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *Proc. of ICASSP*, 2015, pp. 5058–5062.

[4] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," in *Proc. of INTERSPEECH*, 2017, pp. 1113–1117.

[5] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[6] F. Burkhardt, A. Paeschke, M. Rolfes *et al.*, "A database of german emotional speech," in *European Conference on Speech Communication and Technology*, 2005.

[7] R. Jürgens, A. Grass, M. Drolet, and J. Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *Journal of nonverbal behavior*, vol. 39, no. 3, pp. 195–214, 2015.

[8] S. Zhang, X. Tao, Y. Chuang, and X. Zhao, "Learning deep multimodal affective features for spontaneous speech emotion recognition," *Speech Communication*, vol. 127, pp. 73–81, 2021.

[9] N. Audibert, V. Aubergé, and A. Rilliard, "Prosodic correlates of acted vs. spontaneous discrimination of expressive speech: a pilot study," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[10] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 474–477.

[11] F. Chenchah and Z. Lachiri, "Speech emotion recognition in acted and spontaneous context," *Procedia Computer Science*, vol. 39, pp. 139–145, 2014.

[12] Q. Li and T. Chaspari, "Exploring transfer learning between scripted and spontaneous speech for emotion recognition," in *International Conference on Multimodal Interaction*, 2019, pp. 435–439.

[13] K. Feng and T. Chaspari, "Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation," *IEEE Transactions on Affective Computing*, 2021.

[14] H. Li, Y. Kim, C.-H. Kuo, and S. S. Narayanan, "Acted vs. Improvised: Domain adaptation for elicitation approaches in audio-visual emotion recognition," in *Proc. of INTERSPEECH*, 2021, pp. 3395–3399.

[15] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *Proc. of ICASSP*, 2020, pp. 7179–7183.

[16] K. Mangalam and T. Guha, "Learning spontaneity to improve emotion recognition in speech," in *Proc. of INTERSPEECH*, 2018, pp. 946–950.

[17] H. Zhang, M. Mimura, T. Kawahara, and K. Ishizuka, "Selective multi-task learning for speech emotion recognition using corpora of different styles," in *Proc. of ICASSP*, 2022, pp. 7707–7711.

[18] B. L. Sturm, "A simple method to determine if a music information retrieval system is a "horse"," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1636–1644, 2014.

[19] C. Busso, M. Bulut, C.-C. Lee *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Journal of Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. of INTERSPEECH*, vol. 2016, pp. 3603–3607.

[21] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "Baum-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2016.

[22] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *Proc. of ICASSP*, 2021, pp. 6269–6273.

[23] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *Proc. of ICASSP*, 2020, pp. 7174–7178.

[24] Y. Li, P. Bell, and C. Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in *Proc. of ICASSP*, 2022, pp. 7362–7366.

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

[26] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. of INTERSPEECH*, vol. 2020, pp. 1–5.

[27] P. Kumawat and A. Routray, "Applying tdnn architectures for analyzing duration dependencies on speech emotion recognition," in *Proc. of INTERSPEECH*, 2021, pp. 3410–3414.

[28] Z. Peng, X. Li *et al.*, "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.

[29] 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.