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

Applying TDNN Architectures for Analyzing Duration Dependencies on Speech Emotion Recognition

Pooja Kumawat, Aurobinda Routray

Indian Institute of Technology Kharagpur, India
pk28@iitkgp.ac.in, aroutray@ee.iitkgp.ac.in

Abstract

We have analyzed the Time Delay Neural Network (TDNN) based architectures for speech emotion classification. TDNN models efficiently capture the temporal information and provide an utterance level prediction. Emotions are dynamic in nature and require temporal context for reliable prediction. In our work, we have applied the TDNN based x-vector and emphasized channel attention, propagation & aggregation based TDNN (ECAPA-TDNN) architectures for speech emotion identification with RAVDESS, Emo-DB, and IEMOCAP databases. The results show that the TDNN architectures are very efficient for predicting emotion classes and ECAPA-TDNN outperforms the TDNN based x-vector architecture. Next, we investigated the performance of ECAPA-TDNN with various training chunk durations and test utterance durations. We have identified that in spite of very promising emotion recognition performance the TDNN models have a strong training chunk duration-based bias. Earlier research work revealed that individual emotion class accuracy depends largely on the test utterance duration. Most of these studies were based on frame level emotions predictions. However, utterance level based emotion recognition is relatively less explored. The results show that even with the TDNN models, the accuracy of the different emotion classes is dependent on the utterance duration.

Index Terms: Speech emotion recognition, emotion dynamics, TDNN, x-vector.

1. Introduction

Emotion recognition based on speech plays an important role in Human-Computer Interaction (HCI). In the past few decades, interest in automatic Speech Emotion Recognition (SER) has grown rapidly in the field of modern HCI systems. Automatic SER helps virtual assistants and smart speakers to understand their users better. Many attempts have been made to implement the automatic emotion recognition system using audio and visual cues. Most of the speech based emotion recognition systems uses common features such as Mel Frequency Cepstral Coefficients (MFCCs), pitch, intensity, formant frequency, etc. to represent the different emotions. Human emotions are dynamic in nature that changes continuously and sequentially. Moreover, different amounts of information are required over time to perceive the emotional expressions of humans. This varies as a function of emotion classes. Human perception studies have also found that different durations are required to correctly recognize emotions [1]. For example, angry emotion requires a short utterance duration whereas happy emotion requires a relatively large utterance duration for correct classification. It emphasizes the need for optimizing the time window length for analyzing the individual emotion classes in the existing SER systems.

Sarma et al. [2] experimented with different DNN architectures to model the long temporal context for recognizing emotional cues in longer speech utterances/sentences. TDNN with unidirectional LSTM (TDNN-LSTM) setup and time restricted attention mechanisms were also explored in their work which enables the DNN to be more attentive to emotionally sensitive portions of the speech. It was observed that TDNN-LSTM-attention set up trained with fixed length examples with longer chunk of 1 second (s) duration outperforms all other setups. In [3], a new framework was proposed to characterize utterances based on interpretable measures of affective dynamics. In [4], Huang et al. investigated the correlation between emotion and emotion dynamics using joint modeling and probabilistic fusion approaches with a Kalman filter. It is suggested that a speech segment longer than 0.25s carries sufficient information for detecting the emotion present in it [5]. Tzirakis et al. [6] used the CNN-LSTM architecture to learn features and temporal context directly from raw audiovisual signals. Satt et al. [7] developed a system to detect emotions at the segment level. The system was based on an end-to-end deep neural network, applied directly to raw spectrograms. The performance of the developed system was comparable to the state-of-the-art model in SER. In [8], Mower et al. presented a novel hierarchical static-dynamic emotion classification framework. The results were presented across four different sentence lengths. The performance gain between static modeling, dynamic modeling, and static-dynamic modeling was higher for the longer sentences. They have considered only a single database in their work. It has been suggested to apply this framework to additional data collected in alternative settings. In [9], Kim et al. proposed a data-driven framework to explore timing and duration's effect (temporal patterns) of individual emotion classes. Their work is based on window averaging of audio-visual cues to find the influential regions over time for utterance-level emotion inference.

The major contribution of this work is summarised as follows:

- We have explored the emphasized channel attention, propagation, and aggregation based TDNN (ECAPA-TDNN) [10] system for emotion recognition task and compared the performance with the widely used x-vector [11] TDNN architecture. This model was found to outperform the x-vector model for speaker verification tasks and was not applied in emotion recognition earlier.
- We have trained the TDNN models with several fixed lengths of training segment duration. During evaluation, we have shown that the TDNN based models learn a duration dependent bias and performance gradually degrades if the test utterance duration deviates from the training segment duration.

http://dx.doi.org/10.21437/Interspeech.2021-2168

- Further, with the TDNN based models we have conducted an evaluation with various test utterance duration and analyzed the variations of prediction performance of the individual emotion classes.

2. Database description

We have used three different widely used emotional speech databases for implementing the emotion recognition system. These three databases are: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) database, Berlin Database of Emotional Speech (Emo-DB) database, and Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. RAVDESS and Emo-DB databases are the simulated databases whereas the IEMOCAP database is a semi-natural.

2.1. Meta information analysis of the used databases

2.1.1. RAVDESS database

The RAVDESS [12] database consists of surprised, disgusted, happy, sad, angry, fearful, calm, and neutral emotions. This database includes a total of 1440 speech audio clips performed by 24 actors, 60 sentences for each actor. Each emotion is performed in two different intensities. The database was recorded at a 48 kHz sampling rate and the average length for utterances is 3.7s.

2.1.2. Emo-DB database

Emo-DB [13] database comprises of disgust, sad, happy, angry, fear, boredom, and neutral emotions. This database contains a total of 535 utterances performed by 10 professional speakers (five male and five female) recorded at a 48 kHz sampling rate. The average length for utterances in this database is 2.7s.

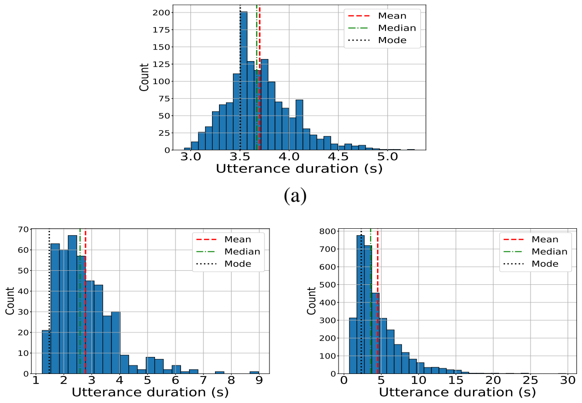
2.1.3. IEMOCAP database

IEMOCAP [14] database is more diverse and larger than other emotional databases. The database is recorded in five sessions of conversations performed by ten speakers. Each session contains utterances from two speakers (one male and one female). The database includes audio, audio+video, and transcriptions for angry, happy, sadness, disgust, fear, excitement, surprise, frustration, and other emotions. In our work we used the audio database of four emotion classes angry, happy, neutral, and sad, where samples from excitement and happy classes are merged into the single happy class to deal with the data imbalance problem [15], [16], [17]. The final dataset contains 5,531 utterances (1,103 angry, 1,636 happy, 1,084 sad, 1,708 neutral) grouped into 5 sessions. The database is sampled at 16kHz sampling frequency and the average length for utterances in this database is 4.52s.

We have discarded some emotion classes that are not present in more than one database. For example, we have discarded the calm emotion class from RAVDESS database and boredom emotion class from Emo-DB database.

2.2. Duration analysis of the used databases

To choose the training chunk size and the offset effectively, in Figure 1 we have plotted the histograms of the utterance durations for each of the three used databases. Each database has speech utterances of varying duration. While training the utterance level classifier, the training chunk size should be selected accordingly based on the utterance durations. We also see that most of the audios contain some small portion of the initial silence portion before the speech. So, a small offset duration can be discarded from the utterances by choosing the proper offset value.



(a)

(b)

(c)

Figure 1: *Histogram of utterance duration for the audio files in (a.) RAVDESS , (b.) Emo-DB, (c.) IEMOCAP database*

3. Methodology

3.1. Data preprocessing and feature extractions

The RAVDESS and Emo-DB databases are downsampled to 16 kHz sampling frequency before extracting the features. We have split the RAVDESS and Emo-DB databases into 80:10:10 ratio for training, validation, and testing purposes. For the IEMOCAP database, we have used the leave one session out cross-validation approach [18]. We used 10% of IEMOCAP training data as a hold-out validation set for hyper-parameter tuning. The widely used 40-dimensional MFCCs audio features are used to implement the speech emotion recognition system. These features are computed using the hamming window of size 20ms with a constant frameshift of 10ms.

To prevent overfitting problem and for better generalization, we have augmented the training data using Additive White Gaussian Noise (AWGN). The AWGN is parameterized by the Signal-to-Noise Ratio (SNR). We choose random SNR values in the range of 15dB to 30dB.

3.2. Model architecture description

In this work, we have explored two TDNN based architectures for the emotion recognition task. Our first model is the x-vector ($TDNN_1$) architecture [11]. This model incorporates five TDNN layers, followed by a statistical pooling layer and two fully connected linear layers. Our second model is the ECAPA-TDNN ($TDNN_2$) architecture [10] which is built upon the x-vector architecture. This model uses a 1-dimensional Squeeze-Excitation Residual block (SE-Res2Block) in between the convolutional layers. We have used 512 channels in the convolutional frame layers and 192 nodes in the final fully connected layer. In the x-vector architecture, the statistical pooling is done only on the outputs from the preceding TDNN layer. However, in ECAPA-TDNN, the outputs from the shallower level SE-Res2Block are concatenated and fed to the next Conv1D+ReLU layer followed by the pooling operation. Due to this *Multi-layer Feature Aggregation* (MFA), the learned embeddings can carry more robust target dependent information.

Table 1: *Layer wise description of ECAPA-TDNN architecture*

Layer	Layer type	Input shape	Output shape
1	Conv1D+ReLU+BN	40×T	512×T
2	SE-Res2Block	512×T	512×T
3	SE-Res2Block	512×T	512×T
4	SE-Res2Block	512×T	512×T
5	Conv1D+ReLU	1536×T	1536×T
6	Attentive stat pooling+BN	1536×T	3072×1
7	FC+BN	3072×1	192×1
8	AAM-Softmax	192×1	N×1

Finally, in the pooling layer, channel dependent attention is used to capture the global context during the pooling operation. The global context helps to utilize the important pattern information that is not present in the currently processed frame. The detailed description of $TDNN_2$ model is given in Table 1. N and T correspond to the total number of emotion classes and temporal dimension of intermediate feature maps respectively.

For implementing both the $TDNN_1$ and $TDNN_2$ models, 40-dimensional MFCC features are used as input. The initial silence portions are removed from audio clips using offset value. Adam is used as the optimizer with a learning rate of 0.001. Additive Angular Margin (AAM)-softmax [19] is used as loss function. The margin is empirically set to 0.4 and the pre-scale used is 30. We have utilized both the TDNN models in an end-to-end approach.

4. Results

The performance of $TDNN_1$ and $TDNN_2$ architectures is evaluated by using both *Weighted Accuracy* (WA) and *Unweighted Accuracy* (UWA) [20]. For WA calculation, the class-specific recalls are weighted by the prior probabilities of the respective emotion classes and UWA is the unweighted average of the class-specific recalls [21]. Further, the performance of $TDNN_2$ architecture is also analyzed for different training chunks and test utterances size in terms of overall classification accuracy. Each emotion class performance is also evaluated for $TDNN_2$ architecture for different test utterance durations in terms of True Positive Rate (TPR).

4.1. Performance comparison of the TDNN models

At first, we have trained the two TDNN architectures using the three emotional speech databases. During training, we used the training chunk size of 3s with an initial offset of 0.5s for RAVDESS database. For Emo-DB database, the chunk size of 2s with 0.1s offset and for IEMOCAP database, the chunk size of 4s with an offset of 0.1s is used for training. As discussed in section 2.2, we can set the appropriate training chunk length and test utterance duration from Figure 1. The overall emotion recognition performance is then compared between the two trained architectures $TDNN_1$ and $TDNN_2$ for each of the three databases. The evaluation results are shown in Table. 2 in terms of WA and UWA. From the results, it is evident that both the TDNN models are learning the emotion class information efficiently. But the ECAPA-TDNN ($TDNN_2$) model outperforms the x-vector ($TDNN_1$) model for all three databases. For example, in IEMOCAP, we achieved WA and UWA of 61.69% and 58.67% respectively for the $TDNN_2$ model which is comparable with state-of-the-art results for the four emotion classes with leave one session out cross-validation approach [18]. It is difficult to strictly compare the results because there is no predefined train/validation/test split followed in the IEMOCAP database.

Table 2: *Performance of $TDNN_1$ and $TDNN_2$ model for three different databases in terms of WA (%) and UWA (%)*

Model	Metric	RAVDESS	Emo-DB	IEMOCAP
$TDNN_1$	WA	82.30	74.58	51.25
	UWA	81.42	70.05	50.71
$TDNN_2$	WA	97.33	77.08	61.69
	UWA	97.50	76.90	58.67

4.2. Analyzing training chunk dependent bias using $TDNN_2$ architecture

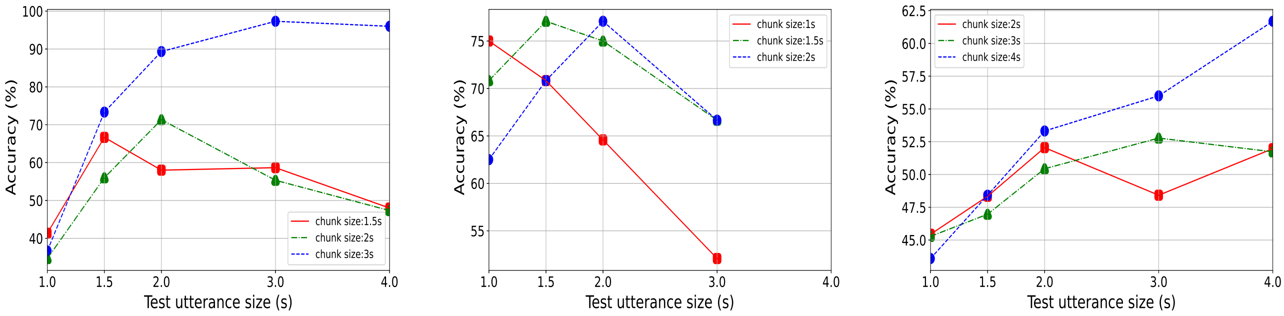
As the ECAPA-TDNN ($TDNN_2$) outperforms the x-vector ($TDNN_1$) model for all the three databases, we have conducted the further analysis using $TDNN_2$ architecture. In this set of experiments, we have trained the model with different initial offset and training chunk lengths for each of the three databases. During evaluation, the test utterances are presented in different duration using appropriate clipping. The experiments are carried out for each of the three databases. The results are presented in Table 3 in the terms of overall classification accuracy to evaluate the performance of the $TDNN_2$ model for different training chunks and test utterance duration. 4s test utterance duration is not used for Emo-DB database as the majority of utterances are less than 4s.

Table 3: *Performance of $TDNN_2$ model for different training chunks and test utterance size in overall accuracy (%)*

Database	Chunk	Offset	Test utterance size (s)				
	(s)	(s)	1	1.5	2	3	4
RAVDESS	1.5	0.5	41.33	**66.66**	58.00	58.66	48.00
	2	0.5	34.66	56.00	**71.33**	55.33	47.33
	3	0.5	36.66	73.33	89.33	**97.33**	96.00
Emo-DB	1	0.1	**75.00**	70.83	64.58	52.08	-
	1.5	0.1	70.83	**77.08**	75.00	66.66	-
	2	0.1	62.50	70.83	**77.08**	66.66	-
IEMOCAP	2	0.1	45.44	48.34	**52.05**	48.42	51.97
	3	0.1	45.28	46.97	50.44	**52.78**	51.73
	4	0.1	43.59	48.42	53.32	56.00	**61.69**

The TDNN architectures embed the variable utterance duration into fixed dimensional representation in the pooling layers. Then the fully connected layers along with the softmax provide utterance level predictions. When we apply the $TDNN_2$ model for the emotion recognition performance, we observe that for most of the cases the overall evaluation accuracy is maximum when the test utterances are of the same duration as of the training chunk size.

If the test utterance duration is decreased or increased from this training chunk duration, the accuracy value decreases gradually. This training duration bias of the $TDNN_2$ model is visually represented using Figure 2. For real-world applications, where the test utterances will be of varying duration, such biases are not desired.



(a) (b) (c)

Figure 2: *Accuracy for TDNN$_2$ model evaluated for (a.) RAVDESS, (b.) Emo-DB, (c.) IEMOCAP database with different training chunk size and testing utterance size*

Table 4: *Individual emotion class accuracy for TDNN$_2$ architecture with different training chunks and testing utterance duration*

Database	Chunk	Offset	Test Utterance	True Positive Rate (TPR) (%)							UWA
	(s)	(s)	(s)	Angry	Happy	Sad	Neutral	Surprised	Disgust	Fear	(%)
RAVDESS	1.5	0.5	1.5	55.00	55.00	70.00	40.00	55.00	85.00	80.00	65.00
	2	0.5	2	85.00	45.00	60.00	70.00	80.00	90.00	75.00	71.25
	3	0.5	3	95.00	100	90.00	100	100	100	100	97.50
Emo-DB	1	0.1	1	61.54	62.50	100	100	-	60.00	71.43	75.91
	1.5	0.1	1.5	69.23	87.50	100	87.50	-	60.00	57.14	76.90
	2	0.1	2	69.23	75.00	85.71	100	-	60.00	71.43	76.90
IEMOCAP	2	0.1	2	68.82	36.20	52.65	62.50	-	-	-	55.04
	3	0.1	3	64.12	40.05	41.63	69.53	-	-	-	53.83
	4	0.1	4	48.53	66.47	50.49	69.19	-	-	-	58.67

4.3. Individual emotion class accuracy with different testing utterance duration

In the literature, several research works have shown that the different emotion classes need different temporal contexts for reliable prediction [22], [23]. Mainly frame level classifiers have been explored to analyze the temporal length related dependencies for the individual emotion classes. In this work, we have computed the classification accuracy for each emotion class when the test utterance duration is varied. But instead of frame level classifiers, we have conducted this experiment with the utterance level TDNN$_2$ architecture. The experiment is conducted for each of the three databases to evaluate the accuracy for individual emotion classes. The results are presented in Table 4 for each emotion class. Based on the analysis done in Figure 2, to achieve the best performance, the evaluations of the varying test utterance durations are carried out with the models trained on the same training chunk lengths. From the results, it is shown that even with the utterance level TDNN classifiers, the prediction performance of certain emotional classes is more accurate only if the test utterance duration lies in a certain range. For example, the prediction performance of the sad emotion present in IEMOCAP data degrades if the test utterance duration is increased from 2s. Whereas for the angry emotion, prediction performance improves gradually with shorter test utterances. For the Emo-DB database, the emotion recognition performance for disgust does not vary with the test utterance duration. For neutral emotion class, the prediction performance does not vary significantly for Emo-DB and IEMOCAP databases. Whereas, for happy emotion in all the three databases, the predictions become significantly accurate if longer test utterances are presented to the TDNN$_2$ model.

5. Conclusions

In this paper, we have explored the TDNN based architectures for the speech emotion recognition task with three emotional speech databases used in the literature. Our first architecture is the x-vector architecture which is widely being utilized in speaker and language recognition tasks. Further, we have implemented the ECAPA-TDNN architecture for the emotion recognition task. To the best of our knowledge, this architecture was not explored in the emotion recognition field extensively. We have shown that with various modifications in terms of skip connects, multi-layer feature aggregation, channel attentive pooling, the ECAPA-TDNN outperforms the x-vector architecture for all three databases. Even though the overall accuracy is promising, we have analyzed two potential issues with the TDNN models in emotion recognition. The first issue lies with the pre-dominant bias based on the training chunk duration of the models. The second issue is that the individual emotion class accuracy is still very much dependent on the test utterance duration even for the utterance level TDNN models. In future work, we would like to mitigate the above mentioned two issues to make TDNN based SER system more efficient.

6. References

[1] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS One*, vol. 6, no. 11, p. e27256, 2011.

[2] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using dnns." in *INTERSPEECH*, 2018, pp. 3097–3101.

[3] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *IEEE ICASSP*. IEEE, 2013, pp. 3677–3681.

[4] Z. Huang and J. Epps, "An investigation of emotion dynamics and kalman filtering for speech-based emotion prediction." in *INTERSPEECH*. Stockholm, 2017, pp. 3301–3305.

[5] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *IEEE ICASSP*. IEEE, 2013, pp. 3682–3686.

[6] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

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

[8] E. Mower and S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," in *IEEE ICASSP*. IEEE, 2011, pp. 2372–2375.

[9] Y. Kim and E. M. Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 92–99.

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

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*. IEEE, 2018, pp. 5329–5333.

[12] 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.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *European Conference on Speech Communication and Technology*, 2005.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

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

[16] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *IEEE ICASSP*. IEEE, 2015, pp. 4749–4753.

[17] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*. IEEE, 2012, pp. 1–4.

[18] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *IEEE ICASSP*. IEEE, 2019, pp. 7390–7394.

[19] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[20] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: Hsf-dnn, ms-cnn and lld-rnn," *Speech Communication*, vol. 120, pp. 11–19, 2020.

[21] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[22] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 552–558.

[23] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, 2019.