STUDY OF SPEECH EMOTION RECOGNITION USING BLSTM WITH ATTENTION

Dalia Sherman
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel
dalia.sherman@live.biu.ac.il

Gershon Hazan
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel
hazanshl@gmail.com

Sharon Gannot
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel
sharon.gannot@biu.ac.il

***Abstract*—We present a study of a neural network-based method for speech emotion recognition that uses audio-only features. In the studied scheme, the acoustic features are extracted from the audio utterances and fed to a neural network that consists of convolutional neural networks (CNN) layers, bidirectional long short-term memory (BLSTM) combined with an attention mechanism layer, and a fully-connected layer. To illustrate and analyze the classification capabilities of the network, we used the t-distributed stochastic neighbor embedding (t-SNE) method. We evaluate our model using Ryerson audio-visual dataset of emotional speech and song (RAVDESS) and interactive emotional dyadic motion capture (IEMOCAP) datasets achieving weighted accuracy (WA) of 80% and 66%, respectively.**

***Index Terms*—Speech Emotion Recognition, Deep Neural Network, Attention Mechanism**

I. Introduction

Speech emotion recognition (SER) is a well-studied problem with multiple applications, such as human-robot interaction (HRI) for robots with social capabilities interacting with humans in populated environments; chatbots in call centers; commerce recommendation systems, and many more. Detecting the emotional state of the interacting person is, therefore, a crucial component in deciding the preferred way to proceed with the interaction. Our research is part of the "Socially Pertinent Robots in Gerontological Healthcare" (SPRING) project. SPRING's mission is to create a social robot that will be deployed in an elderly-care hospital. The robot will interact with the patients, and based on their emotional state, it will have to decide whether to continue engaging in a conversation or leave. Our SER method, which will be later integrated with the visual and text modalities, will be responsible for this task.

Previous works in speech emotion recognition used CNN, deep neural networks (DNN), recurrent neural networks (RNN), and long short-term memory (LSTM) architectures. A combination of CNN and RNN layers demonstrated improved results as compared with conventional classification methods [1], [2], [3]. A network that combines CNN and LSTM can take advantage of the strengths of both blocks, namely the ability of the CNN layers to capture higher-level representations of the data [4], and of the LSTM layer to infer long-term dependencies [5]. In [6], normalizing the CNN features and feeding them to a BLSTM layer further improved the recognition accuracy. In [7], using blocks of CNN+LSTM was proven effective.

An attention mechanism is designed to utilize the most relevant parts of the input sequence by giving weights to the input vectors, with the highest weights given to the most important input vectors [8]. Thanks to the ability to focus on the relevant parts of the input and the use of context information, this mechanism is suitable for sequence modeling. It can therefore reduce misclassification of the audio utterances and improve the SER measures [4], [9], [10], [11].

Feature selection is another essential component in properly designing SER systems. Log-mel filterbank energies and log-mel spectrogram were used as features in [4] and [10], [12], respectively. In [13], an area attention-based CNN was adopted, and the log-mel spectrogram was used as an input feature. This method obtained good results on the interactive emotional dyadic motion capture (IEMOCAP) dataset. Using log-mels filterbank energies and mel frequency cepstral coefficients (MFCC) as features, providing good performance on the eNTERFACE'05 corpus [3]. In [14], MFCCs were extracted from the signals and then classified with k-NN algorithm [15]. Several works used a combination of multiple features at the frontend. In [16], five different features were used, namely MFCC, mel-scaled spectrogram, chromagram, spectral contrast feature and tonnetz representation. In [17] root mean square (RMS) energy, spectrum centroid, MFCC, and zero-crossing rate (ZCR) were used as features and provided accuracy rates that outperform other classifications methods, including [16].

In our work, we only use the audio modality. Other SER methods use multi-modal data to design systems that analyze both visual and acoustic modalities [18], [19], [20]. Such systems are beyond the scope of this contribution. We are proposing an SER model, which is a variant of the system proposed in [4]. We are using four convolutional layers and a Bahdanau attention mechanism [8], while in [4], two convolutional layers and a convolutional attention mechanism were implemented. In [4], only log-mels were used as features, while in our work we analyze several combinations of features. Our implementation can classify audio utterances accurately using IEMOCAP dataset that is more realistic as compared

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

with the commonly used datasets, such as eNTERFACE'05 that is used in e.g. [3], [4]. We aim at the implementation of a simple yet effective model compared to other, more complex models [6], [7], [17], [21]. We are also studying and comparing the contribution of the various components of the studied system to the overall performance by carrying out an ablation study. Moreover, we are using the t-SNE visualization tool [22] to analyze the classification capabilities of the studied scheme across different datasets.

II. Studied Approach

A. Features

In our evaluation, we used both Ryerson audio-visual dataset of emotional speech and song (RAVDESS) [23] and interactive emotional dyadic motion capture (IEMOCAP) [24] datasets. The sampling rate of the signals in the IEMOCAP dataset is 16 kHz, and in RAVDESS, it is 48 kHz. We, therefore, resampled the latter to 16 kHz as well. While processing, we used a fixed utterance length of 5 sec, and either truncated or zero-padded the signals accordingly. We used a short-time Fourier transform (STFT) with frame-size 256, hop-size 128, and fast Fourier transform (FFT) resolution of 512.

After trying different combinations of features, the best results for the RAVDESS dataset were obtained by using the mel-spectrogram feature [12] with 128 mel bands. For the IEMOCAP dataset, after reviewing previous works [5], [13], [16], [17] we experimented with different combinations of the features. For each combination, we used all the features listed in Table I,[1] while excluding only one chosen feature, as described in Table II. We conclude that using a concatenation of all features provides the best results.

B. Architecture

Our starting point is the SER model presented in [4] with several modifications adopted from [7]. After experimenting with different structures and parameters, we made several architectural changes to the model that resulted in improved results, as compared with the original structure, when applied to the relevant datasets.

We used a convolutional long- short-term deep neural network (CLDNN) model that comprises three parts. The first part of the scheme consists of 4 layers of a Conv2D network[2] with a kernel size of 3×3 and stride of 1. The first two layers comprise 64 filters and the following two layers comprise 128 filters each. The layers combine BatchNormalization, 'relu' activation, MaxPooling2D with a size of 4×4, a stride of 4×4, and a dropout rate of 0.2. The two dimensions of the convolutional layers correspond to the time and frequency axes, respectively. 2D convolutions may demonstrate deep dependencies, which go beyond the decorrelating effect along the frequency axis of the MFCCs features. The second part of the scheme is a BLSTM+attention layer. The BLSTM is implemented with a size of 256 and outputs the hidden states, the hidden forward state, and the hidden backward state that are fed into an attention layer.[1] We implemented the attention mechanism using the architecture proposed by Bahdanau et al. [8]. We use the Bahdanau attention mechanism rather than the simpler convolutional attention mechanism that was used in [4], as it is reported to be more suitable for the task of classifying sequential data, such as speech utterances. This was implemented by first concatenating the hidden forward state and the hidden backward state and denoting the concatenated state as the last state. In addition, we implemented a fully-connected (FC) layer with size 256 and 'tanh' activation function, that is fed by the hidden states, and subsequently another FC layer with size 1 and a linear activation function. After flattening and normalizing, the attention weights, which determine the amount of attention given to each hidden state, are obtained. Then, the hidden states are multiplied by the attention weights, to obtain a context vector, which is the weighted sum of the hidden states. Finally, we concatenated the result with the last state, as described above, to obtain the output of the BLSTM+attention block. The third and last part of the system is an FC layer with a softmax activation function, indicating the probabilities of each emotion. The entire architecture of the proposed SER is depicted in Fig. 1.

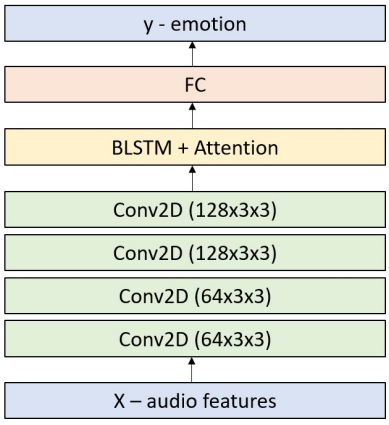


Fig. 1: Architecture of the network.

C. Learning Strategy

The model was trained with 80% of the data as a train set, and 20% of the data as a validation set. The partitioning was done randomly, so the same actors appear in both train and validation sets. We tried different values and eventually chose to use the Adam optimizer with a learning rate of 0.001, a categorical cross-entropy loss function, a batch size of 16, and set the number of epochs to 200. To avoid over-fitting, we added an early stop strategy with patience of 30. Therefore the number of epochs eventually was around 100. The overall number of parameters was about 1.180M for the RAVDESS dataset and about 1.4420M for the IEMOCAP dataset. Our code is available on github.[3]

[1] https://riccardo-cantini.netlify.app/post/speech_emotion_detection/
[2] https://www.kaggle.com/code/bhavikjain/cnn-lstm/notebook
[3] github.com/Dalia-Sher/Speech-Emotion-Recognition-using-BLSTM-with-Attention

TABLE I: Proposed features for the IEMOCAP dataset

Feature	Parameter Value	Description
MFCC	sr=16000, hop_length=512	Spectrogram in mel-scale
MFCC Derivative	width=9, mode='interp', order=1, axis=-1	Local estimates of the MFCC derivative
Spectral Centroid	sr=16000, hop_length=5120	The frequency of the center of mass of the spectrum
Spectral Contrast	sr=16000, hop_length=512	Ratio of the average power in the upper and lower quadrants
Spectral Bandwidth	sr=16000, hop_length=512	3dB Bandwidth
Spectral-roll off	sr=16000, hop_length=512	Threshold frequency below which a specified percentage of the total spectral energy lies
ZCR	hop_length=512	Zero-crossing rate of the time-domain signal
RMS	hop_length=512	The root-mean-square value of the signal

TABLE II: Feature analysis for the IEMOCAP dataset.

Excluded Feature	WA%	**Excluded** Feature	WA%
None	**66**	Spectral Bandwidth	64
RMS	64	MFCC Derivative	63
Spectral Centroid	62	Spectral-roll off	61
ZCR	61	Spectral Contrast	59
MFCC	57		

III. DATA AND EXPERIMENTS

Data: We evaluated the algorithm using two datasets, RAVDESS and IEMOCAP. RAVDESS [23] is a publicly available audio-visual dataset. In this work, we only use the audio modality. The dataset comprises 24 actors, evenly distributed between male and female speakers, each uttering 60 English sentences. Hence, there are 1,440 utterances in total, expressing 8 different emotions: 'sad', 'happy', 'angry', 'calm', 'fearful', 'surprised', 'neutral', and 'disgust'. All utterances are transcribed in advance. Consequently, emotions are more artificially expressed as compared with spontaneous conversation. Another drawback of the dataset is the small number of utterances.

The second dataset is the IEMOCAP [24]. This dataset comprises approximately 12 hours of audio-visual data, including video, speech, motion capture of the face, and text transcriptions. It is, therefore, suitable for multi-modal tasks, but in this work, we only use the audio modality. The dataset consists of conversations between two people that are either improvised or played according to a predetermined transcript that was chosen to evoke different emotions. The dataset consists of 10 actors, evenly distributed between male and female speakers. The utterances are classified into 9 different emotions: 'neutral', 'happiness', 'sadness', 'anger', 'surprise', 'fear', 'disgust', 'frustration', and 'excited'.

Both datasets were used without augmentation since according to our experience, it did not provide a significant improvement. We are aware of several works that were able to report significant improvements due to augmentation [4], [13], [16].

Performance Evaluation: The proposed scheme was evaluated using weighted accuracy (WA) metric and confusion matrix. The weighted accuracy (WA) metric, is calculated as the multiplication of the number of cases correctly predicted for each class by the relative occurrence of the class in the dataset, and then summing over all available classes. The Weighted Accuracy (WA), rather than the simple accuracy, which does not take into account the population of each class, is used as our main evaluation metric, as it is commonly used in many references in the field such as [2], [13]. The confusion matrix compares the actual target values with the predicted values and presents the percentage of the correct and incorrect predictions for each class. The confusion matrix sheds light on the distribution of the errors over different emotions rather than just their Weighted Average (WA), allows us to identify the specific errors between the different emotions, and adds insights when analyzing it alongside the t-SNE plot, which in turn adds information on the classification capabilities of the proposed method and on the difference between the datasets. Note that, depending on the application, some errors might have a strong impact on the applicability of the system, while others are more tolerable.

Results for the RAVDESS dataset: Since the emotions 'calm' and 'neutral' are very similar, and since the number of utterances in 'neutral' is only half of the number of utterances in the other emotion classes, we decided to combine both emotions under the label 'neutral'. In total, the network classified the data into seven different emotions and obtained a WA of 80%. The results are also presented as a confusion matrix as depicted in Fig. 2. It is important to note that the classes 'happy' and 'sad' have significantly lower accuracy than the other classes.

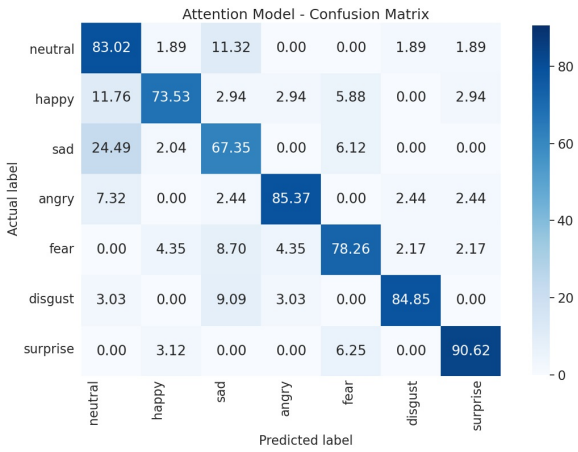


Fig. 2: Confusion matrix of the results on RAVDESS dataset.

Results for the IEMOCAP dataset: In many cases reported in the literature [2], [6], [7], [13], [17], [21], [25], only

the emotions ‘neutral’, ‘happiness + excited’, ‘sadness’, and ‘anger’ are used while training and evaluating the performance of a SER method on the IEMOCAP dataset since these classes are balanced in the number of their utterances. The emotions ‘happiness’ and ‘excited’ have a certain degree of similarity and there are too few utterances of ‘happiness’. Therefore these emotions are combined together to create the ‘happy’ label with an approximately similar number of utterances as the other three emotions used for evaluation. In total, the network classified the data into four different emotions and obtained a WA of 66%. The results are presented as a confusion matrix as depicted in Fig. 3. It is important to note that the class ‘happy’ has lower accuracy than the other classes. Similar effects were reported in [2], [5], [6], [7] and it may indicate that the ‘happy’ emotion is more difficult to characterize.

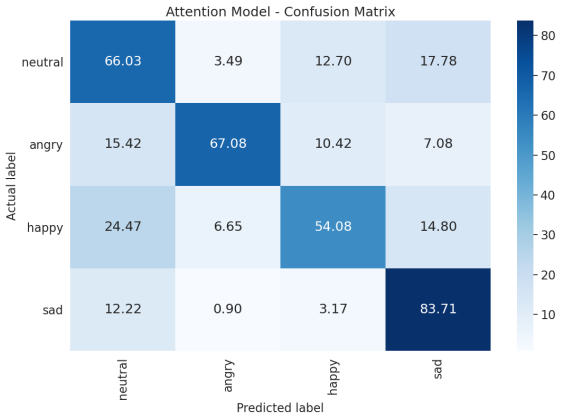


Fig. 3: Confusion matrix of results on IEMOCAP dataset.

Ablation Study: To test whether the additional number of layers compared to [4] provides the observed improvement, we carried out a series of experiments in which we used the full structure of the blocks but with a different number of CNN layers in each experiment. It can be verified from Table III that using four CNN layers indeed yields the best performance on both IEMOCAP and RAVDESS datasets.

To understand the contribution of each of the blocks in the proposed architecture to the overall performance on both IEMOCAP and RAVDESS datasets, we applied an ablation study, as depicted in Table IV. It is clearly observed that each of the blocks indeed contributes to the overall performance, with the full scheme achieving the best results. In addition, we compared our work with two recent works that were applied to the RAVDESS and IEMOCAP datasets as well. In [16] one-dimensional CNN layers are applied after a feature extraction stage. The best-reported results are 71% on the RAVDESS dataset and 64% on the IEMOCAP dataset, both inferior to our best results, which are 80% and 66%, respectively. In [7] the raw signals are used to classify utterances into emotions, in an end-to-end manner. The results reported are 80% on the RAVDESS dataset (similar to our results) and 75% on the IEMOCAP (better than our results). In this paper,

TABLE III: RAVDESS and IEMOCAP: # CNN layers.

Architecture	Dataset	WA%
Full structure with 2 Conv	IEMOCAP	60
Full structure with 3 Conv	IEMOCAP	62
Full structure with 4 Conv	IEMOCAP	**66**
Full structure with 5 Conv	IEMOCAP	64
Full structure with 2 Conv	RAVDESS	62
Full structure with 3 Conv	RAVDESS	68
Full structure with 4 Conv	RAVDESS	**80**
Full structure with 5 Conv	RAVDESS	74

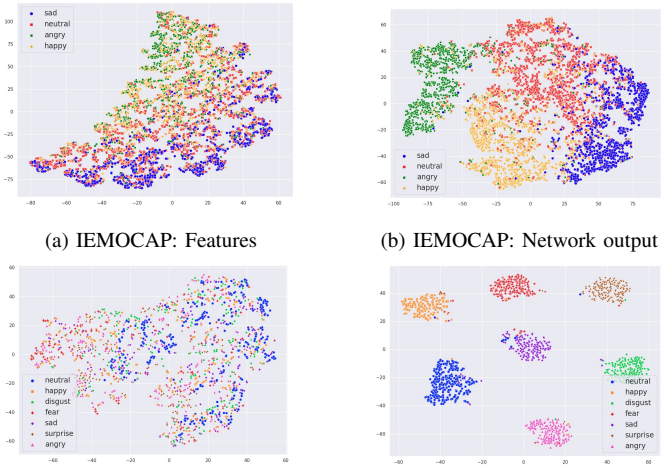
TABLE IV: RAVDESS and IEMOCAP: Various architectures.

Architecture	Dataset	WA%
BLSTM	IEMOCAP	54
BLSTM + Att.	IEMOCAP	62
Conv + BLSTM	IEMOCAP	61
Conv + BLSTM + Att. (Proposed)	IEMOCAP	66
Issa et al., 2020 [16]	IEMOCAP	64
Mustaqeem and Kwon, 2020 [7]	IEMOCAP	**75**
BLSTM	RAVDESS	66
BLSTM + Att.	RAVDESS	73
Conv + BLSTM	RAVDESS	74
Conv + BLSTM + Att. (Proposed)	RAVDESS	**80**
Issa et al., 2020 [16]	RAVDESS	71
Mustaqeem and Kwon, 2020 [7]	RAVDESS	**80**

four layers of ConvLSTM were used. The ConvLSTM is a heavier variant of the LSTM, used in our paper, and will therefore impose higher computational resources (such as memory and processing power), due to the added complexity of the convolutional layers.

IV. Visualization

In order to analyze and visualize the separation capabilities of the proposed scheme on the two databases, we used the t-SNE visualization method, which is a nonlinear method of dimensionality reduction from a high-dimension to a two- or three-dimension representation that can be graphically visualized. The method embeds close data points in the high-dimension to representations with small inter-distance, while remote points are mapped to representations with large inter-distance [22]. The t-SNE method was proven as a very powerful tool in accurately demonstrating the quality of the clustering in SER applications [21], [25]. In Fig. 4 (a) and (b), we apply the t-SNE method to the features extracted from the IEMOCAP database and to the network output and mark each emotion with a different color and shape to visualize the quality of the clustering process. It is easily observed that the application of the network improves the classification quality, but that a significant amount of overlap between classes is still encountered. Specifically, it is clearly depicted that ‘happy’ and ‘neutral’ substantially overlap, thus explaining the poor results obtained for the former. We applied a similar procedure to the RAVDESS dataset, as depicted in Fig. 4 (c) and (d). The significant clustering improvements from the feature level to the output level are evident. We can also observe that the clusters ‘happy’ and ‘sad’ are contaminated by other emotions, in accordance with the lower WA results for these emotions as depicted in Fig. 2. Further inspection of the t-SNE



(a) IEMOCAP: Features (b) IEMOCAP: Network output

(c) RAVDESS: Features (d) RAVDESS: Network output

Fig. 4: t-SNE visualization.

visualization on both datasets, clearly demonstrates the large performance gap between RAVDESS and IEMOCAP. This gap can be attributed to the construction of the dataset. While RAVDESS is using actors in a fully-transcribed scenario, IEMOCAP also comprises spontaneous conversations.

V. Conclusions

In this paper, we carried out a study of an SER architecture, which is a modification of the method presented in [4], and demonstrated that a network comprising CNN and BLSTM with attention blocks provides good recognition results. The main architectural modifications are the number and structure of the convolutional layers and, more importantly, the implementation of the attention scheme as proposed by Bahdanau et al. [8], which is widely used in many audio processing applications. The proposed architecture is still simple and the number of parameters is kept relatively small. Another important modification is related to the input features. We analyzed various combinations of features and selected the most appropriate combination for each of the examined datasets. We applied the SER to two popular datasets, namely IEMOCAP and RAVDESS. The results obtained for the former are inferior to the results obtained for the latter, most probably because the former comprises more realistic data. Still, the obtained scores, especially for the RAVDESS dataset, are high. Furthermore, we carried out an ablation study demonstrating the contributions of all blocks to the overall performance. Finally, using the t-SNE visualization tool, we demonstrated the enhanced classification capabilities due to the application of the proposed SER scheme. The results also shed some light on the performance difference between the examined datasets.

References

[1] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *IEEE Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, 2016.

[2] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Interspeech*, 2018, pp. 3683–3687.

[3] C.-W. Huang and S. Narayanan, "Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition," *arXiv preprint arXiv:1706.02901*, 2017.

[4] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *IEEE international conference on multimedia and expo (ICME)*, 2017, pp. 583–588.

[5] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

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

[7] Mustaqeem and S. Kwon, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network," *Mathematics*, vol. 8, no. 12, pp. 2133–2151, 2020.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[9] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125 868–125 881, 2019.

[10] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[11] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, "Attention-Enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition," in *Interspeech*, 2019, pp. 206–210.

[12] M. Seo and M. Kim, "Fusing visual attention CNN and bag of visual words for cross-corpus speech emotion recognition," *Sensors*, vol. 20, no. 19, p. 5559, 2020.

[13] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6319–6323.

[14] S. Demircan and H. Kahramanlı, "Feature extraction from speech data for emotion recognition," *Journal of Advances in Computer Networks*, vol. 2, no. 1, pp. 28–30, 2014.

[15] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "K-NN model-based approach in classification," in *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, 2003, pp. 986–996.

[16] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[17] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221 640–221 653, 2020.

[18] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.

[19] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, "Multimodal end-to-end sparse model for emotion recognition," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5305–5316.

[20] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Transactions on Affective Computing*, 2019.

[21] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE Access*, vol. 9, pp. 51 231–51 241, 2021.

[22] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[23] 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.

[24] 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.

[25] X. Wang, M. Wang, W. Qi, W. Su, X. Wang, and H. Zhou, "A novel end-to-end speech emotion recognition network with stacked transformer layers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6289–6293.