Speech Emotion Recognition

Yash Dixit[1], Sakshi Chauhan[2], Suraj Yadav[3], Shivam Singh[4], Surabhi[5]

[1,2,3,4]Student, Department of Computer Science and Engineering, GNIOT, Greater Noida, UP, India
[1,2,3,4]Teacher, Department of Computer Science and Engineering, GNIOT, Greater Noida, UP, India

Abstract:
Speech emotion recognition is a vital area of research with applications starting from human-computer interaction to mental health monitoring. This paper will provide a comprehensive survey of the techniques, methods, applications, and challenges in speech emotion recognition. It begins by checking the significance of recognizing emotions from speech and its diverse applications across various fields. The paper then shows the method employed for emotion speech recognition, encompassing traditional machine learning techniques, such as support vector machines and Gaussian mixture models, as well as contemporary approaches, including deep learning and multimodal fusion. Moreover, it examines benchmark datasets commonly used for training and evaluation purposes in emotion speech recognition research. Speech Emotion Recognition (SER) has a wide range of applications and there has been a lot of research going on in this fascinating area in recent years. However, the entertainment sector suffers from a lack of study in this research. Many of use Neural Network (NN) and Long Short-Term Memory (LSTM) architectures to categorize the emotions in audio recordings captured by actors expressing various emotions.

Moreover, our survey explores real-world applications of emotion speech recognition like virtual assistants, health sector, market sector, education, Mental health diagnosis. The paper discusses the challenge associated with emotion speech recognition, including the variability of emotional expressions, cultural influences, humor and privacy concerns. In future it will help others to deal with many Noisy dataset and various cultural effects of emotion.

Keywords: Emotion speech recognition, RNN, LSTM, sequential model, machine learning, deep learning, multimodal fusion,benchmark datasets, applications, challenges, future directions.

2. INTRODUCTION

Speech is the main means of transmitting information. It contains a wide variety of information, and it is important to understand and express rich emotional information through the emotions it contains and visualize it in response to objects, scenes or events. However, it can also lead us to do opinion mining and the sentiment of public can be find out using the audio data only. There is not much difference between textual analysis and audio analysis, the only difference the source of data. Whether we are getting data in the form of text or audio format. Similar to sentiment, emotions can be analyzed computationally. However, the goal of emotion analysis is to recognize the emotion, rather than sentiment, which makes it a more difficult task as differences between some emotion classes are more subtle than those between positive and negative. The goal of this survey is to provide an overview of recent methods of emotion and sentiment analysis as applied to a text. The survey is directed at researchers looking for an introduction to the existing research in the field of sentiment and emotion analysis of a (primarily, literary) text. We do not cover applications of emotion analysis in the areas of digital humanities that are not focused on text. Neither do we provide an in-depth overview of all possible applications of emotion analysis. In differentiating between various emotions which particular speech features are more useful is not clear. Because of the existence of the different sentences, speakers, speakingstyles, speaking rates variability was introduced, because of which speech features get directly affected and was not easy to detect. The same utterance may show different emotions. Each emotion may correspond to the different portions of the spoken utterance. Therefore, it is very difficult to differentiate these portions of utterance. Another problem is that Emotion is depending on the speaker and his or her culture and environment. As the culture and environment gets change the speaking style also gets change, which is another challenge in front of the speech emotion recognition system. There may be twoor more types of emotions, long term emotion and transient one, so it is not clear which type of emotion the recognizer will detect.

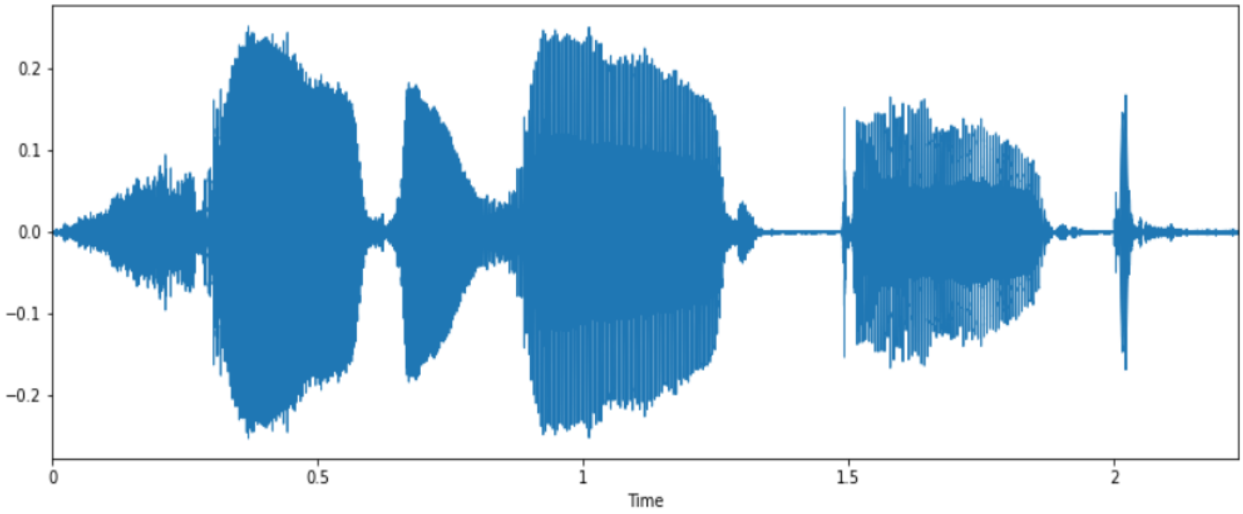


Fig 1. Showing amplitude variation in a voice note

3. LITERATURE SURVEY

This section will provide overview of all the methods and techniques which were used in recent times and in past**.** In this section, we review the most recent studies related to current work. In 2018, Swain et al. revised studies between 2000 and 2017 in the SER programs according to the three retained methods, input domain, and separators. Animportant phase of data research and feature releases; however, only traditional machine learning methods such as CNN, KNN, SVM are considered a distinguishing tool, and the authors feelremorse for the neural networks and deep learning methods. A year later, Khalil et al. reviewed comprehensible approaches to theSER using in-depth reading. Many in-depth, updated learning methods include deep neuralnetwork (DNN), convolution neural network (CNN), repetitive neural network (RNN), and auto encoder, spoken and some of their issuesand strengths in the study. In 2020, Bas et al. published a brief review of the importance of data sets and features of speech recognition, audio removal; finally, they analyzed the importance of differentiated approaches involving SVM and HMM. The power of the study was to identify a number of factors related to the recognition of speech emotions; however, its weakness is the leak of modern research methods and is briefly mentioned in the interaction of repetitive neural networks as an in-depth learning method. In our model, the data set we needed to train was the RAVDEES data set. In this training process, we were able to achieve 93% accuracy.

4. METHODOLOGY

Speech emotion recognition basically works on finding out feature from a given audio file. Audio files can be also converted to textual format and then it would extract features on various parameters. In this research paper researchers will get to know about Librosa library which is used to visualize and extract the audio. Speech emotion recognition technology utilizes various techniques and tools to analyze and interpret emotional cues in spoken language. Some of the key technologies and methods used for speech emotion recognition include:

1. Feature Extraction: Techniques such as Mel-frequency cepstral coefficients (MFCC), pitch, energy, and formant analysis are used to extract acoustic features from speech signals that are indicative of emotional content.2. Machine Learning and Pattern Recognition: Supervised learning algorithms, including Support Vector Machines (SVM), Random Forest, and deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are employed to classify and recognize emotional states based on extracted features.3. Natural Language Processing (NLP): NLP techniques are used to process and analyze the linguistic content of speech, including sentiment analysis and language modeling, to complement acoustic features in determining emotional states.4. Emotion Databases and Corpora: Databases of emotional speech samples and corpora annotated with emotional labels are used for training and testing emotion recognition models, enabling the development of robust and accurate systems.5. Signal Processing and Audio Analysis*: Signal processing methods, such as time-frequency analysis, spectrogram analysis, and audio feature extraction, are applied to capture emotional characteristics from speech signals. 6. Emotion Recognition APIs and Frameworks*: Various software libraries, APIs, and frameworks, such as openSMILE

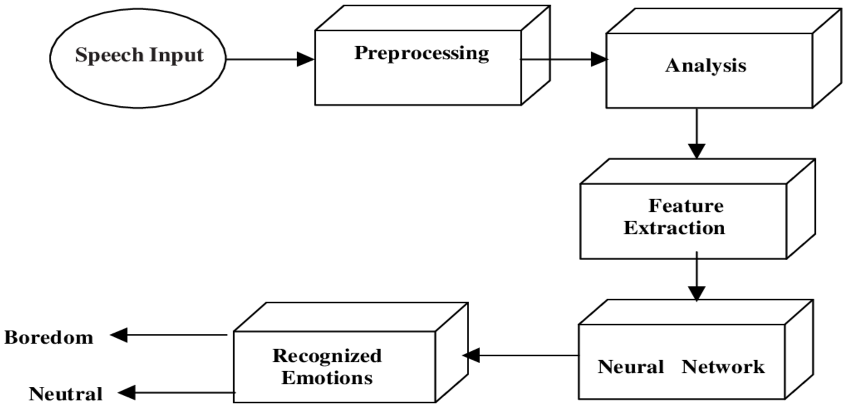
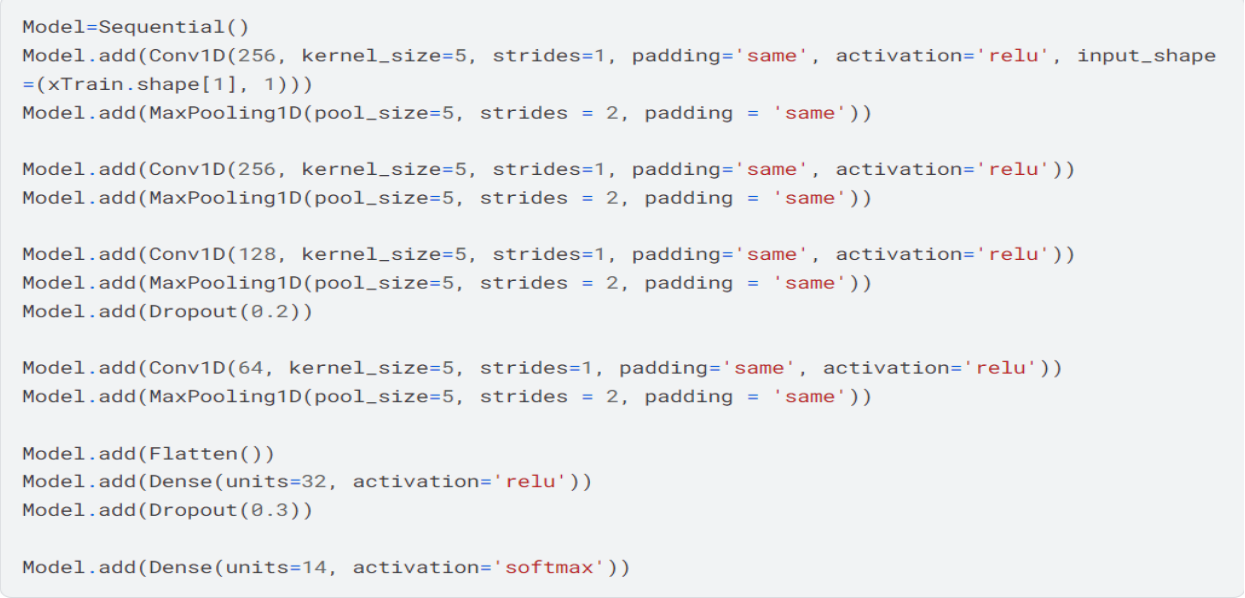


Fig 2. Showing process of Speech emotion recognition.

A sequential model in the context of neural networks refers to a linear stack of layers, where each layer has exactly one input tensor and one output tensor. It's the most common type of model used in deep learning, especially for tasks like image recognition, natural language processing, and sequence prediction. Here's a simplified breakdown of a sequential model:1. Initialization: When creating a sequential model, you start by initializing an instance of the sequential model class provided by the deep learning framework you're using, such as TensorFlow or PyTorch.2. *Layer Stacking*: Layers are added to the sequential model one by one, creating a linear stack of layers. These layers can include various types, such as dense (fully connected), convolutional, recurrent, activation, and normalization layers.3. *Input Shape*: The first layer added to the sequential model should specify the input shape, which defines the shape of the input data that the model will expect. Subsequent layers automatically infer their input shapes based on the output shapes of the previous layers.4. *Model Compilation*: Once the layers are added, the model is compiled by specifying the loss function, optimizer, and metrics to be used during training.5. *Training*: The model is trained using a training dataset, where the input data is fed through the layers, and the model's weights are updated through backpropagation to minimize the defined loss function**.**



```
Model=Sequential()
Model.add(Conv1D(256, kernel_size=5, strides=1, padding='same', activation='relu', input_shape
=(xTrain.shape[1], 1)))
Model.add(MaxPooling1D(pool_size=5, strides = 2, padding = 'same'))

Model.add(Conv1D(256, kernel_size=5, strides=1, padding='same', activation='relu'))
Model.add(MaxPooling1D(pool_size=5, strides = 2, padding = 'same'))

Model.add(Conv1D(128, kernel_size=5, strides=1, padding='same', activation='relu'))
Model.add(MaxPooling1D(pool_size=5, strides = 2, padding = 'same'))
Model.add(Dropout(0.2))

Model.add(Conv1D(64, kernel_size=5, strides=1, padding='same', activation='relu'))
Model.add(MaxPooling1D(pool_size=5, strides = 2, padding = 'same'))

Model.add(Flatten())
Model.add(Dense(units=32, activation='relu'))
Model.add(Dropout(0.3))

Model.add(Dense(units=14, activation='softmax'))
```

Fig 3. Sequential model creation

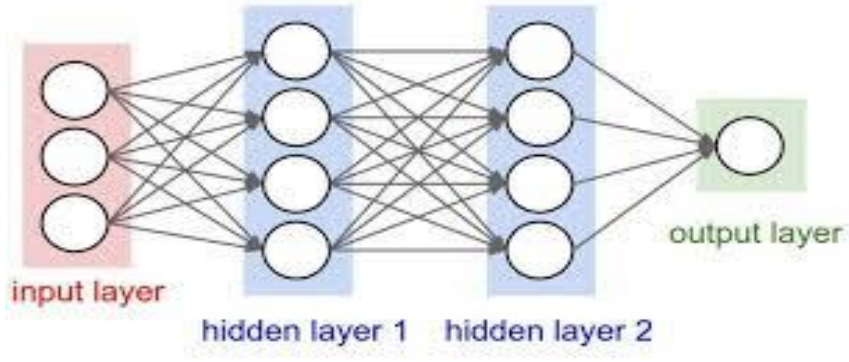
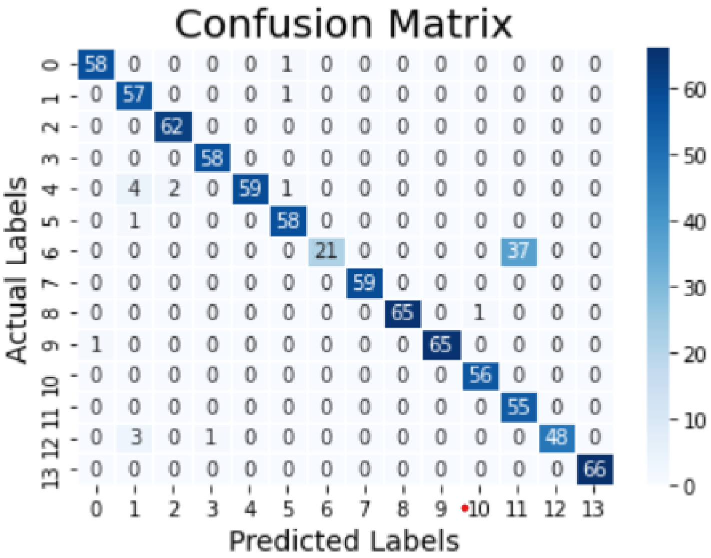


Fig 4. Showing Sequential model

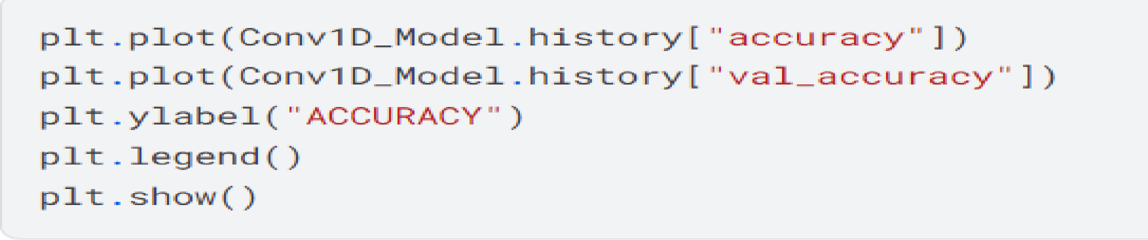
Recurrent Neural Networks (RNNs): RNN has Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, which help for sentiment analysis due to their ability to capture sequential dependencies in textual data, making them suitable for processing language and contextual information related to emotions. When using a Recurrent Neural Network (RNN) for emotion analysis, particularly in the context of textual data such as written expressions of emotion, the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures are commonly employed due to their ability to capture sequential dependencies and long-term contextual information. Here's a simplified overview of how an RNN, specifically an LSTM-based model, can be used for emotion analysis:1. Input Representation: The input text data, such as sentences or phrases expressing emotions, is typically encoded using techniques like word embeddings (e.g., Word2Vec, GloVe) to represent words as dense vectors that capture semantic relationships.2. Sequential Processing: The LSTM model processes the input text sequentially, taking into account the order of words and their contextual dependencies within the text. 3. Long Short-Term Memory Cells: Within the LSTM architecture, the memory cells maintain and update information over the sequence, allowing the model to retain important contextual information related to emotions over longer spans of text. 4. Gates: LSTM units have gates (input gate, forget gate, and output gate) that regulate the flow of information, enabling the model to selectively remember or forget information based on the context of the input text. 5. Training: The LSTM model is trained using labeled emotion data, where the model learns to associate specific patterns in the input text with corresponding emotional categories or sentiment labels. In order to check the accuracy and outcome of our model we are using confusion matrix which shows the actual parameter and the predicted labels. Not only it gives us the brief idea about the model efficiency but also shows the various points which model considers.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	58	0	0	0	0	1	0	0	0	0	0	0	0	0
1	0	57	0	0	0	1	0	0	0	0	0	0	0	0
2	0	0	62	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	58	0	0	0	0	0	0	0	0	0	0
4	0	4	2	0	59	1	0	0	0	0	0	0	0	0
5	0	1	0	0	0	58	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	21	0	0	0	0	37	0	0
7	0	0	0	0	0	0	0	59	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	65	0	1	0	0	0
9	1	0	0	0	0	0	0	0	0	65	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	56	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	55	0	0
12	0	3	0	1	0	0	0	0	0	0	0	0	48	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	66

Fig 5. Showing Confusion Matrix for our testing model

In the speech emotion recognition system after calculation of the features, the best features are providedto the classifier. A classifier recognizes the emotion in the speaker's speech utterance. Various types of classifiers have been proposed for the task of speech emotion recognition. Gaussian Mixtures Model (GMM), K-nearest neighbors (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), etc. are the classifiersused in the speech emotion recognition system. Each classifier has some advantages and limitations over the others. Only when the global features are extracted fromFeature Extraction Feature Selection Classifier Recognized emotion Speech input the training utterances, Gaussian Mixture Model is more suitable forspeech emotion recognition. All the training and testingequations are based on the supposition that all vectors are independent therefore GMM cannot form temporal structure of the training data. For the best features a maximum accuracy of 78.77% could be achieved usingGMM. In speaker independent recognition typical performance obtained of 75%, and that of 89.12% for speaker dependent recognition using GMM. Other classifier that is used for the emotion classification is anartificial neural network (ANN), which is used due to itsability to find nonlinear boundaries separating the emotional states. Out of the many types, feed forward neural network is used most frequently in speech emotion recognition. Multilayer perceptron layer neural networks are relatively common in speech emotion recognition as it is easy for implementation andit has well defined training algorithm.



```
plt.plot(Conv1D_Model.history["accuracy"])
plt.plot(Conv1D_Model.history["val_accuracy"])
plt.ylabel("ACCURACY")
plt.legend()
plt.show()
```

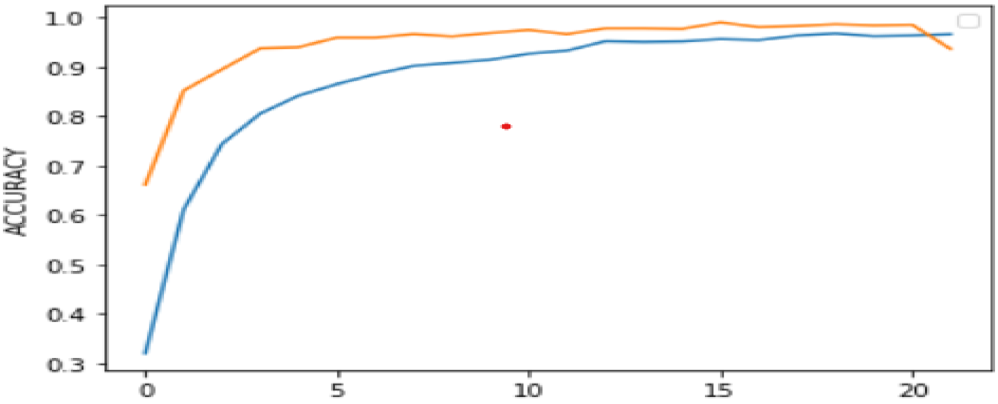


Fig 6. Showing Accuracy graph

5. CONCLUSION

In conclusion to this speech emotion recognition, we would say that it’s an unveiling and evolving fields of computer science and medical Science. If it would achieve and accuracy around 98 percent then would help a lot of medical patients and as well as doctors to monitor their patient mental health and stability. These field would also let law and enforcement to check for false statement passed by any one in the court and would punish the guilty. The impact of audio emotion recognition extends to a wide array of applications, from enhancing user experiences in human-machine interactions to providing valuable insights in healthcare, customer service, and entertainment settings. Ongoing research and innovation in audio emotion recognition are expected to yield more accurate, robust, and context-aware systems, enabling deeper insights into human emotions and fostering empathetic and personalized interactions in the digital realm. As the field of audio emotion recognition continues to evolve, the potential for creating more empathetic and responsive technologies that understand and adapt to human emotions is both exciting and promising.

6. REFRENCES

1. Schuller, B., Steidl, S., Batliner, A., Hantke, S., & Vinciarelli, A. (2018). The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-assessed Affect, Crying & Heart Beats. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.
2. Lee, C. M., Narayanan, S. S. (2019). Recent advances in speech emotion recognition. IEEE Signal Processing Letters, 26(1), 116-130.
3. Eyben, F., Weninger, F., & Schuller, B. (2019). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 16-20). IEEE.
4. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Schwenker, F.

(2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6885-6889). IEEE.

5. Khorram, S., Peck, E. M., Sundararajan, A., & Narayanan, S. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 623-630). IEEE.
6. Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2018). Recognizing Affect in Speech. In S. S. Narayanan (Ed.), The Oxford Handbook of Affective Computing (pp. 267-282). Oxford University Press.
7. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Wöllmer, M. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.
8. Deng, J., Zhang, Z., Marchi, E., & Schuller, B. (2013). In search of the acoustical correlates for speech emotion: A survey of the 20th century databases. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 105-110). IEEE.
9. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32-80.
10. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Schwenker, F. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 815-819). IEEE.
11. Schuller, B., Steidl, S., Batliner, A., Hantke, S., & Vinciarelli, A. (2020). The INTERSPEECH 2020 Computational Paralinguistics Challenge: COVID-19 Cough, Breathing & Speech. In 2020 28th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.
12. Han, K. J., Zhang, Y., Schuller, B., & Wu, J. (2021). Emotion recognition in the wild with deep transfer learning and data augmentation. IEEE Transactions on Affective Computing, 12(2), 342-354.
13. Schuller, B., Steidl, S., Batliner, A., Hantke, S., & Vinciarelli, A. (2021). The INTERSPEECH 2021 Computational Paralinguistics Challenge: Atypical & Self-assessed Affect, Crying & Heart Beats. In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.
14. Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2022). Speech emotion recognition with self-attention mechanism. IEEE Access, 10, 13491-13500.
15. Schuller, B., Steidl, S., Batliner, A., Hantke, S., & Vinciarelli, A. (2023). The INTERSPEECH 2023 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In 2023 31st European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.