Grenze International Journal of Engineering and Technology, June Issue

Malayalam Kannur Dialect Speech Emotion Recognition using Machine Learning

Amitha I C[1], Amal Surendran[2], Likhith N V[3], Adithya P P[4] and Aswin K[5]
[1]Assistant professor, St. Thomas College of Engineeringand Technology, Mattannur Kerala, India
Email*:* hod@stthomaskannur.ac.in
[2-5]St. Thomas College of Engineeringand Technology, Mattannur Kerala, India
Email*:* reamal2016@gmail.com, likhithnvmzh@gmail.com, adithyaprakash222@gmail.com, ashwin.jagadeeshan@gmail.com

***Abstract*—Speech is the most natural form of communication among various modes such as verbal and non verbal communications. Speech include information about languages, Speaker, messages and emotions. Emotions make speech more expressive and effective. There are so many language based speech emotion recognition systems are available. And there are several Malayalam speech emotion recognition systems. But there is no speech emotion recognition system in kannur dialect, which is used in northern district of Kerala. So we implementing speech emotion recognition system in Malayalam of Kannur dialect. Currently there is no data set for our project. So we decide to create a dataset of kannur dialect. Here we use Mel- frequency spectral coefficient (MFCC), Chroma shift, Spectral centroid, Spectral bandwidth, rolloff, Zero crossing rate and mel spectrogram as the walls of the spectral features for extraction. The support vector machine (SVM) and Random forest algorithmare decided to use in our project for classification. The proposed system will provide a Novel speech emotion recognition suitable for malayalam kannur dialect.**

***Index Terms*—Speech, Random forest, Emotion recogniton,Malayalam Kannur dialect**

I. INTRODUCTION

Speech is generally considered the easiest, fastest and most natural form of communication between humans, however this is not true for machines. Current machine cannot understand the emotions from local languages. Communication is necessary because machines cannot grasp human emotions. The need for more human-interactive applications has made emotion recognition from speech using machine learning algorithms a hot topic in recent years. Due to the availbility of datasets for these languages, most European and Asian languages—including German, English, Spanish, Arabic and Dutch are used in the implementation of emotion recogni- tion systems. Due the limited number of malayalam dataset, the proposed system investigates emotion recognition using spoken Malayalam Kannur dialect, which is used in northern region of Kerala. The dataset, which was constructed from freely accessible YouTube films and WhatsApp chats. The proposed system is mainly meant for recogniton of an endto end system which is capable of recognising the humanemotions such as angry, happy, sad and neutral for MalayalamKannur dialect speech. Numerous spectrum component were recovered, including the mel-frequency cepstral coefficient (MFCC), Chroma shift, Spectral centroid, Spectral bandwidth, rolloff, Zero crossing rate, spectral contrast, mel spectrogram and subsequently support vector machine (SVM) and Randomforest classification techniques are used for classifying theemotions, These two classification techniques are compared.

Grenze ID: 01.GIJET.9.2.380

II. Problem Definition

Our thoughts and behaviours are significantly influenced by our emotions. The emotions we experience every day caninfluence the decisions we make about our lives. In orderto improve the effectiveness of human-machine interaction, it is essential for machines to be capable of perceive human emotions. One of the most natural methods we as humans can communicate is through speech. So we are create a Malayalamvoice emotion recognition system that has a variety of uses.

III. Existing System

Speech emotion recognition systems are implemented inmany languages like Arabic, English, Dutch, Spanish. various spectral features like mfcc, melspectrogram are used for feature extraction. Most commonly recognised emotions are anger, happiness and sadness. SVM (support vector machine), RNN (recurrent neural network) and ANN (artificial neural network) were used for classifying emotions from audiodatasets.

IV. Related Works

Many researches has been conducted to study about speech emotion recogniton. Speech recogniton systems areimplemented in many ways. Different types of algorithmsand audio features are analysed by this study

A. Hybrid LSTM-Transformer Model for Emotion RecognitionFrom Speech Audio Files

This paper proposes a hybrid LSTM-Transformer model that extracts Mel Frequency Cepstral Coefficient (MFCC) features from spectrograms of speech samples to classify emotions, achieving improved recognition success rates compared to existing models. The model uses MFCC to extract speech features and achieve significant recognition improvement com- pared to existing models in published works. Speech emotion recognition is crucial in daily human activities, and speech is an essential channel for studying human emotions. The proposed model shows significant improvement in the RAVDESS, Emo-DB, and language-independent dataset,

B. End-to-End Speech Emotion Recognition With GenderInformation

In this study, introduces an end-to-end approach for emotion recognition that incorporates speaker gender information, using deep learning algorithms to extract emotional information from raw speech data without the need for explicit modeling of emotional characteristics. This study proposes a novel approach for speech emotion recognition that incorporates speaker gender information without relying on speech acoustic features. The proposed algorithm uses deep learning algorithms to automatically extract important information fromraw speech signals and classify emotions. The algorithm comprises a residual convolutional neural network and a gender information block. The results show that the proposed method achieves accuracy improvements compared to existing state- of-the-art speech emotion recognition algorithms with differentstructures, even when the gender information is removed. The study provides further validation of the use of raw speech signals to capture emotional information without artificial intervention.

C. Facial Recognition System to Detect Student Emotions andCheating in Distance Learning

The purpose of this study is to develop a system that uses computer vision and deep learning algorithms to support lecturers in monitoring and managing students during online learning sessions and E-exams. The system employs software methods, computer vision algorithms, and deep learning algorithms to achieve its objective. The study found that the system achieved high accuracy for student identification in real-time, student follow-up during the online session, and cheating detection.

The system uses a pre-processing architecture representation. The student identification system verifies a student'sidentity based on the algorithms used. It compares the existingface in the database with the face taken by the camera. If the face is recognized, the name of the face appears as it is saved in the database. The result of the performance and accuracy for the deep metric learning algorithm and the ResNet-34 algorithm in this system is about 99.38 percent.

The system also demonstrated accuracy and high performance in detecting abnormal behaviors of students during e-exams. The gaze tracking model achieved an accuracy of up to 96.95 percent, and the facial movement tracking model achieved up to 96.24 percent. Additionally, the face detection and object detection models achieved an accuracy of around 60 percent. The system was able to detect happy behavior with an accuracy of 45.27 percent, and fear with an accuracy of 30.20 percent.

The study found that the system can effectively help teachers in online exams to identify expected cheating cases. The system accurately tracked the movement of the student's head, face, and expressions of fear emotion during exams. These findings were consistent with earlier research that studied mechanisms that students use to cheat in online exams.

D. Two-Way Feature Extraction for Speech Emotion Recognition using Deep Learning

This paper presents an Two-way feature extraction for speech emotion recognition using deep learning. Two-way feature extraction for speech emotion recognition involves using a combination of PCA and deep neural network (DNN) techniques to extract potential features from speech data. Addi-tionally, a pretrained VGG-16 model is trained on the features obtained from the proposed two-way feature extraction model. The approach works directly on the audio dataset and involves the use of various numerical features such as MFCC, Log Mel-Spectrogram, Chroma, Spectral centroid, and Spectral roll- off. After feature extraction, normalization and dimensionality reduction techniques are applied. In the first approach, a DNN model with 13 layers, including 7 dense and 6 dropout layers, is used for the network architecture. In the second approach, log mel-spectrogram images from the input audio dataset are extracted and saved to particular emotion classes. The VGG-16 model with 16 layers is then used for the model architecture. The proposed two-way feature extraction model for speech emotion recognition using deep learning combines these two approaches and is trained on multimodal speech data.

E. Spoken Language Identification System Using Convolutional Recurrent Neural Network

The article proposes a spoken language identification system that utilizes a hybrid convolutional Recurrent Neural Network (CRNN) for identifying seven languages, including Arabic. The system compares Gammatone Cepstral Coefficient(GTCC), Mel Frequency Cepstral Coefficient (MFCC), and a combination of both features at the feature extraction stage. The speech signals are represented as frames and used as input for the CRNN architecture, which combines the advantages of both CNN and RNN. The system performs better with combined GTCC and MFCC features than GTCC or MFCC alone, and is effective in identifying Arabic and Russian and discriminating among similar languages. The proposed system has potential for developing Arabic-related automatic speech recognition (ASR) systems and can be extended to other languages and evaluated with other speech corpora. The output of the system can be used for various purposes, including multi-lingual automatic translation.

V. PROPOSED SYSTEM

The proposed system is focus to identify various emotions from Malayalam Kannur dialect speech. Random forest algorithm is used to classify emotions from speech data. The human voice is given as the input. The proposed system accept the audio input from user, then the input is converted into wave format and it campare the input data with already trained audio dataset in the model. And display the emotion and an emoji corresponding to the predicted emotion and also displays accuracy. Python librosa library function is used for audio analysis and feature extraction. Happy, Anger, sad and neutral were identified.

A. Architecture

The architecture diagram of the system is shown below:

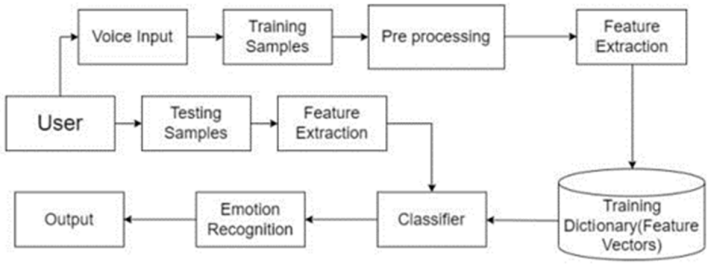


Fig. 1. Architecture

It consists of eight main components, which include input, pre-processing, region detection, feature extraction, feature database matching, classification, and output. The database block consists of various gesture images.

1)Input:- As soon as we login to the web page by providing a username and password, the user can input the audio. The audio is in the wave format.
2)Training Samples:-Training data is the data used to train a machine learning model.
Training data is the initial step to create models for a dataset. It's a set of data samples used to fit the parameters of a machine learning model.
3)Testing samples:- Test data provides a final, real-world check of an unseen data to confirm that the machine learning algorithm was trained effectively.
4) Pre-processing:- In pre-processing step, data cleaning is done for classification. The noise, background music, and silence were removed.
5) Feature extraction:-The features such as energy, pitch, energy rate are extracted using spectral features like MFCC, mel-spectrogram, spectral contrast, and then the frequency-based features were obtained by converting the time-based signals.
6) Feature vector:-A feature vector is an ordered list of numerical properties of observed phenomena. It represents input features to a machine learning model that makes a prediction.
7) Classifier:-In classification the feature extracted from training and testing phase are compared, then the emotion is recognised. Here random forest algorithm is used as classification algorithm.

B. Kannur Malayalam speech emotion recognition dataset (KMSER Dataset)

In this work we have created a Malayalam Kannur dialect dataset from YouTube videos taken from popular YouTube channels, Audio clips taken from WhatsApp chats and self- trained audio files. A group of audios was viewed and checked to select the scenes that include clear emotional references to test the ability of different machine learning algorithms to detect emotions. After that, the audio files were extracted from videos, which were divided into smaller audio clips and converted into wave-format. the duration of the sound clip varry from 2 sec to 9 sec depending on the completion of the project. divided into smaller audio clips and converted into wave-format.

C. Random forest

Random forest algorithm is used for classification and also for regression.Random forest or Random decision forest is a method that operates by constructing multiple decision trees.The final decision is obtained by majority votes of the decision tree. In random forest algorithm there is no over fittings, multiple trees reduce the risk of overfitting. It's training time is less. It has high accuracy and runs efficiently in large database. For large data it produces highly accurate predictions.

D.Emotion Detection

The audio is extracted, and converted in the wave format. There is no pre-trained dataset available for the classification of emotions. Here a new dataset is created which consists of 130 audios that includes emotions like happy, sad, angry, neutral for training and 30 percent of audios that includes all other audios for testing. The audios in the dataset are classified into four categories of emotions consisting of anger, happy, sad, neutral of humans. These mp3 format audios in the data sets are converted into wave format using librosa library and are saved in a data list and the corresponding labels are stored in the label list.
By conducting a preprocessing step, the audio aims to improvedata to suppress unwilling distortions or enhance some audio features in the audio which is important for further processing,where unwanted noice are removed.
Preprocessing is done to enhance audio properties and to extract valuable features from the audios such as Chroma stft, Spectral centroid, Spectral bandwidth,Rolloff, Zero crossing rate, and 20 Mfcc features. Basic preprocessing techniquesare used here for removing noice of the audios.

E. Emotion Recognition

By performing feature matching we find related featuresin two similar audios. The brute-force approach is used for feature matching which takes the descriptor of one feature in the first audio and compares it with descriptors of all features in the second audio using some distance calculations. Then the closest one is returned in a resulting pair. By performing classification, the process of categorizing and labeling groups of pixels or vectors within an image is done based on specific characteristics.

F. Training

Creating a csv file with header chroma stft, spectral cen- troid, spectral bandwidth, rolloff, zero crossing rate,20 mfcc featuters and ouput label. Feature extraction of each audio in the data set using librosa Finding mean value

calculation of chroma stft, spectralcentroid, spectral bandwidth, rolloff, zero crossing rate,20 mfcc featuters. writing extracted features to csv file.

G. Softwares Used

- Flask:- Flask is a very important web application frame-work it is written in python.It can use for developing webapplication, creating web APIs and in machine learning application to create and to end projects.
- Pandas:-Pandas is a python module that makes data science easy and effective. Pandas is a python module that makes data science easy and effective.It perform High performance data analysis.It is used for working with large datasets. It support files with different format.In pandas data is represented in tabular ways.
- Scikit-learn :- It is a well-documented and well-loved Python machine learning library. The library is main-tained and reliable, offering a vast collection of machinelearning algorithms for you to incorporate into your projects. If you haven't tried scikit-learn.
- NumPy:- NumPy stands for numerical python.It is afundamental python package for scientific computation or programming. It contain fundamental tools and tech- niques to solve mathematical problems.Numby library functions is used to perform mathematical and logical operations on array and linear algebra and random number genaration.
- Librosa:- Librosa is an important package for music and audio analysis.If you need to do any kind of analysis there is no way around librosa.This will give as one dimensional NumPy array so we can work with this and also with the sample rate then it has built in function to easily plot the data.It can used for any kind of feature analysis.

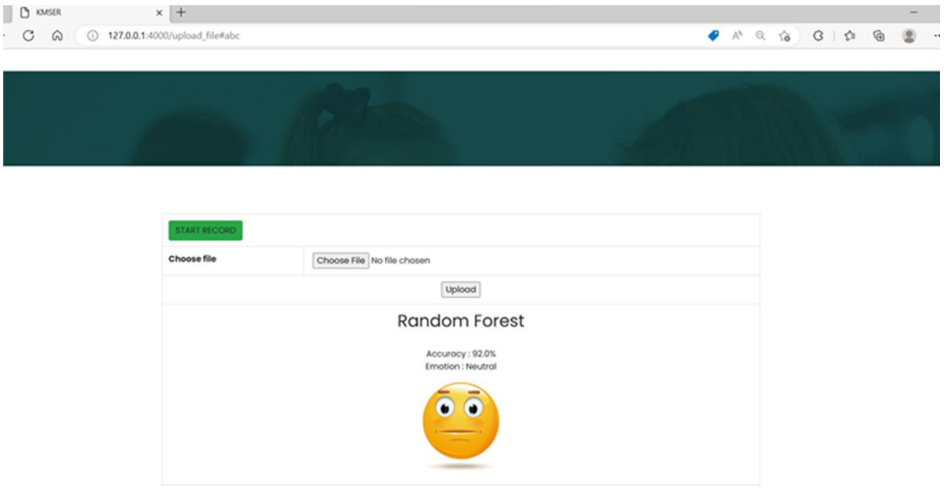
VI. RESULT

This section will compare and analyze the results of the proposed random forest model and SVM. Since there is no predefined dataset, a new dataset is created which is then trained and tested. There are 130 audios in the dataset, where 35 audio clips for happy, 35 audio clips for angry 25 audio clips for sad, 35 audio clips for neutral for training and some audio clips that includes emotions audio for testing. The system successfully classified the emotions like angry, happy, sad and neutral.

The trained model is used to predict the class labels of the dataset, and returns the predicted classes as a NumPy array. The model's predictions on the test dataset are computed using model.predict classes() and model's performance is evaluated on a test dataset by computing its confusion matrix using the ,confusion matrix() function from scikit-learn. The load model() function is used to load a pre-trained machine learning model from a file, and the predict() function is used tomake predictions on the input dataset using the loaded model. Nearly 80% of dataset is used for training stage and 20% of dataset is for testing stage.

The random forest learns to make more accurate predictions and assigns higher probabilities to the correct class for each input data point by using categorical cross-entropy.

We compare two classifiers random forest and SVM , Randomforest shows highest accuracy.



I. Fig. 2. Upload

VII. Conclusion

Speech Emotion recognition refers to the ability of a computer to understand human emotion. The information or data gathered in this way may be used to execute further commands. In our proposed system, Random forest classification of audio data set is used. The user inputs an audio into the model, model converts the user inputed mp3 audio into wave format, which compares it to the audio already in the data set and the corresponding output is displayed with corresponding emojies.

In the future, this system can be by adding more emotions intothe dataset for detecting human emotions. Future work could focus on expanding the dataset to include a more diverse range of speakers and emotional expressions. Further, different new techniques can be used in order to improve the performanceof the model.

In conclusion, the Malayalam speech emotion recognition project aimed to develop a system capable of accurately iden- tifying and categorizing emotions expressed in spoken Malay- alam language. Through the implementation of advanced ma- chine learning algorithms and techniques, the project success- fully achieved its objectives and produced promising results.

References

[1] T.-W. Sun: End-to-End Speech Emotion Recognition With Gender Information. ttps://creativecommons.org/licenses/by/4.0/. Digital Object Identifier 10.1109/ACCESS.2020.3017462.

[2] F. Andayani et al.: Hybrid LSTM-Transformer Model for Emo- tion Recognition From Speech Audio Files. Digital Object Identifier 10.1109/ACCESS.2022.3163856.; creativecommons.org/licenses/by-nc- nd/4.0 .

[3] Ozdamli, F.; Aljarrah, A.; Karagozlu, D.; Ababneh, M. Facial Recognition System to Detect Student Emotions and Cheating in Distance Learning. Sustainability 2022, 14, 13230. https://doi.org/ 10.3390/su142013230.

[4] Alashban, A.A.;Qamhan, M.A.; Meftah, A.H.; Alotaibi, Y.A. Spoken Language Identification System Using Convolutional Recurrent Neural Network. Appl. Sci. 2022, 12, 9181. https://doi.org/10.3390/app12189181.

[5] Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.A.; Alhadlaq, A.; Lee, H.-N. Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. Sensors 2022, 22,2378. https://doi.org/10.3390/ s22062378.

[6] Fika Hastarita Rachman;Riyanarto Sarno;Chastine Fatichah; (2020). Music Emotion Detection using Weighted of Audio and Lyric Fea- tures . 2020 6th Information Technology International Seminar (ITIS), (),doi:10.1109/itis50118.2020.9321046.

[7] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Ci- chocki. 2018. “Emotionmeter: A multimodal framework for recognizing human emotions.” IEEE transactions on cybernetics, 99, Pp. 1–13.

[8] JOUR Juan, Wan 2020/12/23 1 11 Gesture recognition and information recommendation based on machine learning and virtual reality in distance education 40 10.3233/JIFS-189572 Journal of Intelligent Fuzzy Systems.

[9] M. R. Prasetya, A. Harjoko, and C. Supriyanto, “Speech emotion recognition of Indonesian movie audio tracks based on MFCC and SVM,” in Proc.Int. Conf. Contemp. Comput. Informat. (IC3I), Dec. 2019, pp. 22–25.

[10] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, “Speech based human emotion recognition using MFCC,” in Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET), Mar. 2017, pp. 2257–2260.

[11] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emo- tion recognition: Features, classification schemes, and databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.

[12] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and S. Rajan, “Emotion recognition from audio signals using support vector machine,” in Proc. IEEE Recent Adv. Intell. Comput. Syst. (RAICS), Dec. 2015, pp. 139–144.

[13] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, “Acoustical properties of speech as indicators of depression and suicidal risk,” IEEE Trans. Biomed. Eng., vol. 47, no. 7, pp. 829–837, Jul. 2000.

[14] L. Abdel-Hamid, “Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features,” Speech Commun., vol. 122, pp. 19–30, Sep. 2020.

[15] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, “Emotion recognition in Arabic speech,” Analog Integr. Circuits Signal Process., vol. 96, no. 2, pp. 337–351, Aug. 2018.

[16] M. B. Mustafa, M. A. M. Yusoof, Z. M. Don, and M. Malekzadeh, “Speech emotion recognition research: An analysis of research focus,” Int. J. Speech Technol., vol. 21, no. 1, pp. 137–156, Mar. 2018.

[17] (2021). Arabic Population. Accessed: Aug. 3, 2021. [Online]. Available: https://worldpopulationreview.com/country-rankings/arab-countries.

[18] Y. Hifny and A. Ali, “Efficient Arabic emotion recognition using deep neural networks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 6710–6714.

[19] I. Hadjadji, L. Falek, L. Demri, and H. Teffahi, “Emotion recognition in Arabic speech,” in Proc. Int. Conf. Adv. Electr. Eng. (ICAEE), Nov. 2019, pp. 1–4.

[20] A. Meftah, Y. Alotaibi, and S.-A. Selouani, “Designing, building, and ana- lyzing an Arabic speech emotional corpus,” in Proc. Workshop Free/Open- Source Arabic Corpora Corpora Process. Tools Workshop Programme, 2014, p. 22.

[21] Telfaz11. Youtube Channel. Accessed: Mar. 25, 2021. [Online]. Avail- able: https://www.youtube.com/user/telfaz11.

[22] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “Librosa: Audio and music signal

analysis in Python," in Proc. 14th Python Sci. Conf., 2015, pp. 18–25.
[23] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in Proc. IEEE Int. Conf. Multi- media Expo, Aug. 2002, pp. 113–116.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011.
[25] J. Sidorova, "Speech emotion recognition," Master Thesis, Univ. Pompeu Fabra, Barcelona, Tech. Rep., 2007, doi: 10.13140/RG.2.1.3498.0724
[26] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "A review on speech emotion recognition: Case of pedagogical inter- action in classroom," in Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP), Fez, Morocco, May 2017, pp. 1–7, doi: 10.1109/AT- SIP.2017.8075575.
[27] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," Frontiers Comput. Sci., vol. 2, p. 14, May 2020, doi: 10.3389/fcomp.2020.00014.
[28] K. H. Lee, H. Kyun Choi, B. T. Jang, and D. H. Kim, "A study on speech emotion recognition using a deep neural network," in Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC), Jeju Island, South Korea, Oct. 2019, pp. 1162–1165, doi: 10.1109/ICTC46691.2019.8939830.
[29] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Calgary, AB, Canada, Apr. 2018, pp. 5089–5093, doi: 10.1109/ICASSP.2018.8462677.
[30] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 27, no. 11, pp. 1675–1685, Jul. 2019, doi: 10.1109/TASLP.2019.2925934.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, arXiv:1706.03762.
[32] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," 2019, arXiv:1912.02610.
[33] S. Lee, D. K. Han, and H. Ko, "Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition," Sensors, vol. 20, no. 22, p. 6688, Nov. 2020, doi: 10.3390/s20226688.
[34] B. McFee et al., "Librosa/librosa," Zenodo, 2020, doi: 10.5281/zenodo. 591533.
[35] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.
[36] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," IEEE Trans. Affect. Comput., vol. 6, no. 2, pp. 97–108, Apr. 2015.
[37] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., May 2004, pp. 577–580.
[38] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 5688–5691.
[39] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2227–2231.
[40] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," IEEE/ACM Trans. Audio, Speech, Lang., Process., vol. 26, no. 1, pp. 31–43, Jan. 2018.
[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.
[42] K. Wang, N. An, B. Nan Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 69–75, Jan. 2015.