WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Deep Learning Based Speech Emotion Detection for Music Recommendation

Eswar Benarjee Naidu Polamarishetty [a]*, *G. Ramana Reddy* [b], *A. Vinay Viswanath* [c], *A. Malaya Maruthi* [d]

[a] *Department of Computer Science and Engineering, GMR Institute of Technology, Rajam,Andhra Pradesh, India.*

ABSTRACT

Speech emotion recognition is a technology that extracts emotional features from speech signals by a computer. It contrasts and analyses the characteristic parameters and the emotional change acquired. Positive feelings (for example, satisfaction) are emphatically connected with efficiency. Conversely, gloomy feelings (for example, dissatisfaction) are negatively connected with efficiency and positively connected with unfortunate activities and mischievous activities in the work environment. Subsequently, maybe, programmed proposal frameworks can recommend activities to dispense with undesired gloomy feelings in the working environment. These frameworks could support their activities in light of the dependability of feeling detectors. Our profound learning model endeavors to foresee the six essential feelings in the Crema dataset. Furthermore, it utilized no relevant information, but it accomplished hearty results. Conventional one-layer CNN and two-layer CNN can't catch optimal highlights from sound records. To overcome this, data augmentation is performed prior to extricating the features. Feelings are ordered using deep learning techniques, and it is integrated with the Spotify API to prescribe music by recommendation system based on the inclination.

Keywords: Emotion detection, Data Augmentation, Convolutional Neural Network (CNN), Recommendation System, Learning model.

1. Introduction

Speech emotion recognition can detect using many deep learning techniques. By using deep learning techniques, we recognize many emotions like happy, angry, sadness, neutral, disgust, fear. A new model for speech emotion recognition by using sequence selections and extraction via non-linear RBFN based method to find similarity level in clustering was proposed. In this model, a key segment is selected from the complete cluster which is nearer to the centroid of the clusters and represents all the segments. After developing the SER model, it was tested on various standard datasets like IEMOCAP, EMO-DB, and RAVDESS. The performance metrics to measure the accuracy of the model with these three datasets, confusion matrix is used.

A new model for speech emotion recognition by using sequence selections and extraction via non-linear RBFN based method to find similarity level in clustering was proposed. In this model, a key segment is selected from the complete cluster which is nearer to the centroid of the clusters and represents all the segments. After developing the SER model, it was tested on various standard datasets like IEMOCAP, EMO-DB, and RAVDESS. The performance metrics to measure the accuracy of the model with these three datasets, confusion matrix is used.

There are many features extraction and feature selection techniques used for above temporal features. The features like MFCC. The basic target of a feature selection (FS) model is to choose optimal set of features which can reduce the computational cost and storage requirement, as well as enhance the classification accuracy of the problem in hand. After future selection we use various deep learning techniques like Deep convolution neural network. In the end ,different performance metrics are used to evaluate the proposed model.

2. Related Work

[1]. In this paper the author discusses about speech emotion recognition using artificial neural network on Mfcc features. He uses ANN over CNN algorithm because CNN requires image datasets or converting audio files into images dataset to perform optimal capacity. But here ANN takes audio files as input emotion recognition. ANN model takes less time for training on the given dataset .So it is not time-consuming process and we access fast data. He uses MFCC features extraction technique is used to convert the conventional frequency to Mel scale frequency because it gives better results. The proposed model gives accuracy of 88.72% on RAVDESS dataset and 86.80 on SAVEE dataset.

[2]. In this journal, the author proposed a model name Weighted TrBaidu Algorithm. Using the MELD Dataset (Multimodal Emotion Lines Dataset), it consists of Audio, Video. Only audio files were taken for this model. After conversion of Audio to Spectrogram, Features have been passed to Baidu's first CNN and BN layers. And the output of these were passed to Baidu's second CNN and BN layers. And then the output was Activated and was passed to Bi LSTM layer, and fully connected layer. Finally, the Emotion State Probability Distribution was calculated.

[3]. The main aim of this paper is to compare Multitask and Single Task Models using MSP-IMPROV dataset. The Dataset was passed to a DNN architecture, which consisted of independent layers, Shared layers and Input layer. The output of this layer was passed to Multi-Layer Perceptron layers, which collects the features and classifies the output using SoftMax layer. Along with Multi-Layer Perceptron Layers, the output of DNN was also passed to Single Task Layer and the output of this layer is predicted by only one score.

[4]. In this paper the author discusses about “Clustering –based BILSTM model”. He used BILSTM model for speech emotion Recognition. He uses speech signals as input. He uses “IEMOCAP dataset, which is commonly used for speech emotion recognition. The dataset consists of 10 actors’ audio-visual data including video, audio, motion of faces, speech. Apart from IEMOCAP dataset he also uses EMO-DB and RAVDESS dataset and it contain 535 utterances recorded by 10 actors. RAVDESS dataset contain 8 emotions having 24 actors, both male and female. The emotions like happy, sad, anger, calm, neutral, disgust, surprise recorded by different male and female. Accuracies over IEMOCAP, EMO-DB, and RAVDESS dataset 72.25%, 85.57%, and 77.02%. The proposed system proved the robustness and significance for SER to correctly recognize the emotional state of the speaker using spectrograms of speech signals.

[5]. In this paper, the author has studied about two different features techniques for emotion detection. One of them is Empirical Features i.e., F0, energy and voice probability. The other is spectrogram based statistical features. The author has used a Kernel Extreme Learning Machine(KELM) which can learn more from features. Emo-DB and IEMOCAP datasets are used. This paper concludes that fusion framework will outperforms the state of art methods. The accuracy for KELM is 87.49%.It consists of future gaps A stricter requirement for selecting the Auditory-based features would give more accuracy. A better feature extraction should be found as sound is dynamic.

[6]. In this paper the author discusses a “Speech emotion recognition using recurrent neural network with directional self –attention". He uses BILSTM-DSA model for speech emotion recognition. He uses IEMOCAP dataset and EMO-DB datasets are used for this algorithm. Especially in emotion recognizing of anger and happiness ,BILSTM-DSA achieve the highest recognition accuracies. He uses BLSTM-DSA because the speech signal features are decoded by bi-directional LSTM forward and backward respectively, then encode them with directional self-attention mechanism. The proposed method gives more correlation on EMO-DB dataset when compared to IEMOCAP dataset. The proposed algorithm can not only discover the correlation of signals in sentences through self-attention mechanism, but also improve the diversity of information through direction mechanism.

[7]. In this paper the author discusses about “Speech Emotion recognition Using Various Deep Learning Techniques”. Signal Pre-Processing, Feature Extraction and Classification are used. For Development of emotional speech systems, Speech Data bases are classified into three types they are Simulated Database, Induced Database and the Natural Database. The most commonly used databases are Berlin EmoDB stands for Emotional Speech Database, DES stands for Danish Emotional Speech Database that contains the recorded voices of 10 performers. Deep Belief Networks((DBN), , Convolutional Neural Networks (CNN),Auto Encoder (AE) are few Deep Learning techniques that improves the Overall Performance of the proposed system.

[8]. In this paper the author discusses about “Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition”. He proposed a parallelized convolutional recurrent neural network (PCRN) with spectral res for acquiring better emotional features in Speech Signals. To calculate the performance of the proposed model .The author tested the model on four common data sets they are CASIA, EMO-DB, ABC, SAVEE. The Author uses LOSO Cross-Validation Strategy for the evaluation in this experiment. Different from other works on speech emotion recognition, the author uses 3-D log Mel-spectrograms and frame-level features. He extracts 3-D log Mel-spectrograms and frame-level features from the original speech signal as inputs to our PCRN model to prevent the loss of emotional information in time-frequency domain learning. The recognition rate is at least 9.75% and 8.89% higher than that of all comparative experiments.

[9]. In this paper the author discusses about metaheuristic algorithms. Author proposed modified Cuckoo Search and NSGA-II based feature selection .The author uses two data sets in this paper they are, USC Interactive IEMOCAP stands for Emotional Dyadic Motion Capture, the EMO-DB stands for Berlin Emotional Speech. The author evaluated the Performances of NSGA-II and Cuckoo Search metaheuristics search algorithms for feature selection in speech emotion recognition and 87.66% accuracy obtained.

[10]. In This paper the sequential GMM-DNN classifier has been contrasted with support vector machines (SVMs) and multilayer perceptron (MLP) classifier. Proposed GMM-DNN model gives better results than the greatly used SVM and MLP based classifiers. Accuracies on various models like GMM-DNN:83.97%,SVM:80.33%,MLP: 69.78%.Their GMM-DNN model yields the results similar to those obtained by human judges in a subjective assessment context. The used classifiers work for some features only for other features the performance is poor. The main limitation of their work is that the collected Emirati speech dataset is acted and unspontaneous due to the difficulty of capturing spontaneous emotions from human beings. To improve the accrued limitations in future and capturing spontaneous emotions from human beings.

[11]. In this paper, INCA model(iterative neighborhood component analysis) is used. By using the spectrograms of speech signals, which are utilized as the inputs of the convolutional neural networks (CNNs) to learn and extract the spatial features .The spectrum and the spectrogram of the speech signal and extracted the high-level discriminative features by using the 2D and 1D-CNN architectures. In this SAVEE dataset is used. Now we use INCA algorithm for best feature selection removed the redundancy from the fused futures. By using above strategy, we get highest recognition rate is 85%.

[12]. In this paper the author Novel Approach model is used. Here he used Deep Feature Extraction from Pre-Trained CNN Models in which Data Augmentation, Feature Selection with Relief, Classifications are also used for prediction. Here in this paper, he mentioned the approach,

Features used and classifiers, Training-Testing Data Splitting, Dataset and Accuracy.

[13]. In this paper the author uses a Deep and Acoustic feature on a novel approach model. Here he used Deep Feature Extraction from Pre-Trained CNN Models in which Data Augmentation, Feature Selection with Relief, Classifications are also used for prediction. Here in this paper, he mentioned the approach, Features used and classifiers, Training-Testing Data Splitting, Dataset and Accuracy.In this paper, the author introduce in this paper, the author only uses the utterances that are the affective improvisational scenarios spontaneously and don't take account with the performance of theatrical scripts. The author selects four emotions like Angry, Sad, Happy and Neutral, in IEMOCAP corpus and then adapt all the utterances in Berlin EMO-DB database from seven emotion(Angry, Excited, Frustrated, Happy, Neutral, Sad, and Surprise). In this paper, the author also performs the cross-corpus experiment between two databases of IEMOCAP and EMODB. The author adopts the IEMOCAP to treat as the training set and the test the model in the EMODB database. The author proposed that experimental results shows that the designed ARDNN network does not only recognize the speech emotion effectively, but also has better generalization ability and the ARDNN networks have a certain advantage over ACRNN networks in overall performance.

[14]. In this paper, the author proposed a joint deep learning model that combines 3D convolutions and attention based sliding recurrent neutral networks (ASRNNS) as the back-ends of the SER system. In the proposed system, the author describes the front-end is used to generate temporal modulation, and the attention-based back-end is used to identify the emotional state in natural speech. To show the benefit of the proposed model, the author evaluates it on the IEMOCAP and MSP-IMPROV datasets by comparing the various models with the proposed model. The author said that proposed model can achieve better results compared with the traditional model on both the datasets and can effectively obtain the emotional information.

[15]. In this paper, the author developed an SER method. It consists of low computational complexity. In this paper one-dimensional local pattern is used which ternary pattern. It is applied on the raw dataset.by using NCA approach useful features extracted. The author compares with the various datasets finally author concludes that the obtained accuracies are low.

[16]. In this paper the author discusses about clustering-based GA-optimized feature set. He proposed a feature optimization approach that uses a clustering-based genetic algorithm . So, to improve the emotional recognition performance in terms of accuracy he proposed the clustering-based GA for feature reduction. They are used Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for outlier detection at the feature optimization level, further he used Principal component analysis(PCA). the results are improved from 84.09% to 87.76% when only PCA is used. For EMO-DB, RAVDESS, and SAVEE datasets the recognition rates are 89.65%, 82.5%, and 77.74% .

[17]. In this paper the author discusses about A multimodal hierarchical approach to speech emotion recognition from audio and text. He uses a hierarchical DNN based approach to emotion recognition from speech and text in both unimodal and multimodal systems. The audio-based unimodal system proposes using a combination of 33 features, which include spectral, and voice quality-based audio features. Further, for the multimodal system, both the above-mentioned audio features and additional textual features are used. He uses future extraction techniques for both audio and text features for improving the accuracy. The proposed hierarchical unimodal model offered average unweighted recall scores of 81.2% and 81.7% on the RAVDESS and SAVEE datasets, respectively, while the proposed multimodal system offered an AUR of 74.5% on IEMOCAP.

[18]. In this paper the author works on Autonomous vehicle. The author proposes an SER-enhanced model. They convert the vehicle speech information data into spectrograms and input them into an Alex Net network model to obtain the high-level features of the vehicle speech acoustic model. They convert the vehicle speech information data into text information. The input representation is the Bidirectional Encoder to obtain high level features. As the foundation for an ITS, the in-vehicle communication network makes use of advanced wireless communication technology to realize vehicle-to-vehicle and vehicle-to-road communication and the organic integration of traffic participants, vehicles and their environment to improve the safety and efficiency of the transportation system The accuracy is improved.

[19]. The proposed model constructs a balanced music video emotion dataset including diversity of territory, language, culture and musical instruments. The author tested dataset over four unimodal and four multimodal convolutional neural networks (CNN) of music and video, separately fine-tuned each pre-trained unimodal CNN and test the performance on unseen data. The predictive model by integration of all multimodal structure achieves 88.56% in accuracy, 0.88 in f1-score, and 0.987 in area under the curve (AUC) score. In this paper, RECOLA dataset is used. In future author try to fine-tune the audio and video networks separately or train 1D CNN for audio with the constructed small emotion dataset.

[20]. In this paper, a probability and integrated learning (PIL) based classification algorithm is proposed for solving high-level human emotion recognition problems. This paper also presented three new analyses methods based on classification probability including the emotional sensitivity, emotional decision preference and emotional tube. vel human emotion recognition problems. The authors study expects that the proposed method could be used in the affective computing for video, and may play a reference role in artificial emotion established for robot with a natural and humanized way. There is significant scope in extending the present study in context of methodology and more complex applications. There are several problems and unanswered questions which are worth to be explored and discussed in the future.

[21]. The paper implementation of Convolutional Neural Network for emotion detection and there by playing a song. In this paper, the dataset given is in order to obtain minimal processing, multilayer perceptron's are implemented by CNNs. CNNs are observed to have less processing. Filters used in CNNs are advanced.This proposed model can be extended to driver detection to find if the driver is sleepy according to his voice modulation.

[22]. In this paper, A deep learning based hierarchical approach is proposed for both unimodal and multimodal SER systems in this work. The

audio based unimodal system proposes using a combination of 33 features which include prosody spectral, voice quality-based audio. Along with above features, additional textual features are also used. Dataset used in this paper are IEMOCAP, SAVEE and RAVDESS. The accuracies are 81.7%, 81.2% and 74.5% for SAVEE, RAVDESS and IEMOCAP.

[23]. In this paper, the proposed modal is a Convolutional Neural Network built around VGGNet and a novel post processing technique. In this paper, the Datasets used are Soundtracks, Bi-Modal and MER_Taffe. In this paper, the author stated that it is difficult to obtain a consistent feature set that works across the classifier and datasets. To get rid of feature design, deep learning-based approach is followed.

[24]. In this study, the author developed a framework integrating three distinct classifiers. They are DNN, CNN and RNN. The framework was used for categorical recognition including angry, happy, neutral and sad. In this paper, the dataset used is SAVEE and IEMO-CAP. Frame-level low level descriptors (LLDs), Segment-level met spectrograms, and utterance level outputs of high-level statistical functions on LLDs are passed to RNN, CNN and DNN respectively.

[25]. The Author proposed an approach for music emotion recognition based on convolutional long short-term memory deep neural network (CLDNN) architecture. The dataset used for this model training was Turkish emotional music database. The features obtained by feeding convolutional neural network (CNN) layers are utilized with log-Mel filter bank energies and Mel frequency cepstral coefficients (MFCCs) in addition to standard acoustic features. This model obtained an accuracy of 99.19% is obtained using the proposed system with 10-fold cross-validation.

[26]. In this the proposed model deals with the singing voice and speech but finally prove that songs are more emotional using acoustic features. The dataset used is Ryerson Audio-Visual Database of Emotional Speech (RAVDESS). They evaluate different features sets, feature types, and classifiers on both song and speech emotion recognition. In this paper, three feature sets: GeMAPS, py Audio Analysis, and Librosa, two feature types, low-level descriptors and high-level statistical functions are used. The classified used in this paper are multilayer perceptron, LSTM, GRU, and Convolution Neural Networks.

[27]. The proposed model deals with the personalized music playlist generation. The mood is statistically inferred from various data sources primarily: audio, image, text, and sensors. In this paper, Machine learning model Random Forest classification algorithm and eMusic's classifiers are used. eMusic, a new way of personalizing songs playlist by using machine learning technique Currently, this model was performed on a small dataset and limited number of features.

[28]. In this the proposed model presents the evaluation of SER based on the features extracted from MFCC and GFCC to study the emotions from different versions of audio signals. In this paper, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is used. The sound signals are segmented by extracting and parametrizing each frequency calls using MFCC, GFCC, and combined features (M-GFCC) in the feature extraction stage, the proposed model proposes a Deep Convolutional-Recurrent Neural Network (Deep C-RNN) approach to classify the effectiveness of learning emotion variations in the classification stage. Accuracy-rate more than 80% and comparatively less loss during the training and testing process.

[29]. In this proposed model pre-processing step in which speech segments with similar formant characteristics are clustered together and labeled as the same phoneme. The phoneme occurrence rates in emotional utterances are then used as the input features for classifiers. In this papers, six databases (EmoDB, RAVDESS, IEMOCAP, SHEMO, DEMOS and MSP-Improv) for evaluation. Accuracy obtained as follows 62% on the IEMOCAP database and 71% on EmoDB.

[30]. In this paper, Semi-supervised ladder networks are used. supervised loss and auxiliary unsupervised cost function are trained. The addition of the unsupervised auxiliary task provides powerful discriminative representations of the input features, the accuracies obtained are 57.0%, 58.5% and 59.7% respectively. The results give better performance for "angry", "happy" and "sad". The ladder network achieves its best performance on "angry", "happy" and "neutral", less on sad Thus, the enhanced network structure is beneficial to the performance improvement of "angry" and "happy". In the future, Deeper semi-supervised learning and other network structures like recurrent neural networks (RNNs) will be explored.

3. Conclusion

Speech Emotion Recognition is a difficult task that entails extracting features from speech, plotting the wave plot and spectrogram, and classifying it into its respective emotion. Wave plots and spectrograms are representations of speech where augmentation is done to remove noise from the data. In this process, a multi-layer perceptron model for speech emotion recognition was proposed based on extracted features as input data. The model achieved 89.75% accuracy for the Ravdess dataset.Music is recommended based on the emotion received from the multi-layered perceptron. Playlists are suggested to users using the Spotify API, where users can choose the playlist from the listed playlists. For future scope, we will try Feature extraction using pretrained models like Xception. Apart from that we will also try Deep Convolution Neural network and Bidirectional long short-term memory model which works on the data enhancement and datasets balancing. we will also try Detection using Extreme Learning Models which gives high performance and faster response time compare with other methods. These extreme models have several advantages. Those are easy to use, learning speed is more, high performance, it also gives correctness for most of the nonlinear kernel functions and also for activation functions. By implementing all those we will better our models in all aspects. In This way we improve our model in future.

References

[1]. Dolka, H., VM, A. X., & Juliet, S. (2021, May). Speech emotion recognition using ann on mfcc features. In *2021 3rd International Conference*

on Signal Processing and Communication (ICPSC) (pp. 431-435). IEEE.

[2]. Oliveira, J., & Praça, I. (2021). On the usage of pre-trained speech recognition deep layers to detect emotions. *IEEE Access*, *9*, 9699-9705.

[3]. Atmaja, B. T., Sasou, A., & Akagi, M. (2022). Speech Emotion and Naturalness Recognitions with Multitask and Single-task Learnings. *IEEE Access*.

[4]. Sajjad, M., & Kwon, S. (2020). Clustering based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, *8*, 79861-79875.

[5]. Guo, L., Wang, L., Dang, J., Liu, Z., & Guan, H. (2019). Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. *IEEE Access*, *7*, 75798-75809.

[6]. Li, D., Liu, J., Yang, Z., Sun, L., & Wang, Z. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, *173*, 114683.

[7]. Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, *7*, 117327-117345.

[8]. Jiang, P., Fu, H., Tao, H., Lei, P., & Zhao, L. (2019). Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access*, *7*, 90368-90377.

[9]. Yildirim, S., Kaya, Y., & Kılıç, F. (2021). A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. *Applied Acoustics*, *173*, 107721.

[10]. Shahin, I., Nassif, A. B., & Hamsa, S. (2019). Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE access*, *7*, 26777-26787.

[11]. Kwon, S. (2021). Optimal feature selection-based speech emotion recognition using two-stream deep convolutional neural network. *International Journal of Intelligent Systems*, *36*(9), 5116-5135.

[12]. Er, M. B. (2020). A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access*, *8*, 221640-221653.

[13]. Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE access*, *7*, 125868-125881.

[14]. Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2020). Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access*, *8*, 16560-16572.

[15]. Sönmez, Y. Ü., & Varol, A. (2020). A speech emotion recognition model based on multi-level local binary and local ternary patterns. *IEEE Access*, *8*, 190784-190796.

[16]. Kanwal, S., & Asghar, S. (2021). Speech emotion recognition using clustering based GA-optimized feature set. *IEEE Access*, *9*, 125830-125842.

[17]. Singh, P., Srivastava, R., Rana, K. P. S., & Kumar, V. (2021). A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, *229*, 107316.

[18]. Tan, L., Yu, K., Lin, L., Cheng, X., Srivastava, G., Lin, J. C. W., & Wei, W. (2021). Speech Emotion Recognition Enhanced Traffic Efficiency Solution for Autonomous Vehicles in a 5G-Enabled Space–Air–Ground Integrated Intelligent Transportation System. *IEEE Transactions on Intelligent Transportation Systems*, *23*(3), 2830-2842.

[19]. Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, *80*(2), 2887-2905.

[20]. Jiang, D., Wu, K., Chen, D., Tu, G., Zhou, T., Garg, A., & Gao, L. (2020). A probability and integrated learning-based classification algorithm for high-level human emotion recognition problems. *Measurement*, *150*, 107049.

[21]. Deebika, S., & Indira, K. A. (2019, March). A machine learning based music player by detecting emotions. In *2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (Vol. 1, pp. 196-200). IEEE.

[22]. Singh, P., Srivastava, R., Rana, K. P. S., & Kumar, V. (2021). A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, *229*, 107316.

[23]. Sarkar, R., Choudhury, S., Dutta, S., Roy, A., & Saha, S. K. (2020). Recognition of emotion in music based on deep convolutional neural network. *Multimedia Tools and Applications*, *79*(1), 765-783.

[24]. Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Communication*, *120*, 11-19.

[25]. Hizlisoy, S., Yildirim, S., & Tufekci, Z. (2021). Music emotion recognition using convolutional long short term memory deep neural networks. *Engineering Science and Technology, an International Journal*, *24*(3), 760-767.

[26]. Atmaja, B. T., & Akagi, M. (2020, November). On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers. In *2020 IEEE REGION 10 CONFERENCE (TENCON)* (pp. 968-972). IEEE.

[27]. Sarda, P., Halasawade, S., Padmawar, A., & Aghav, J. (2019). Emousic: emotion and activity-based music player using machine learning. In *Advances in Computer Communication and Computational Sciences* (pp. 179-188). Springer, Singapore.

[28]. Kumaran, U., Radha Rammohan, S., Nagarajan, S. M., & Prathik, A. (2021). Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. *International Journal of Speech Technology*, *24*(2), 303-314.

[29]. Liu, Z. T., Rehman, A., Wu, M., Cao, W. H., & Hao, M. (2021). Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. *Information Sciences*, *563*, 309-325.

[30]. Tao, J. H., Huang, J., Li, Y., Lian, Z., & Niu, M. Y. (2019). Semi-supervised ladder networks for speech emotion recognition. *International Journal of Automation and Computing*, *16*(4), 437-448.