Speech Emotion Recognition Techniques: A Review

Manoj
Student of J. C. Bose University of Science and Technology, Faridabad (2020-2022)

Abstract-With the recent advancements of the technology and the growing research areas like machine learning (ML), audio processing and speech processing, the emotional states will be inevitable part of the human-computer interaction. There are more and more studies that are working on providing the computers with abilities like recognizing, interpretation and simulation of emotional states. A general SER detection system based on ML consists of four main steps. The various technique of machine learning and deep learning techniques are reviewed in this paper. The deep learning techniques include CNN, LSTM. The machine learning techniques like support vector machine, KNN, Random forest etc. It is analysed that deep learning techniques give high performance as compared to machine learning techniques in terms of accuracy.

Keywords-Speech Classification, Emotion, Deep Learning, Machine Learning

INTRODUCTION

Human abilities to perceive, adapt, and learn about the environment are generally the three key elements of the characterization of what constitutes intelligent behaviour. Several studies over the past few decades have been signifying that this definition of intelligent behaviour has missed a very crucial element known as emotional intelligence. Emotional intelligence is a person's ability to sense, convey, self-regulate, distinguish, and control the emotional condition of other people [1]. In psychology, the definition of emotional state involves a complex condition leading to mental and physical changes and affect the behaviour and thinking process of an individual. Emotions are cornerstones to folks, influencing perspicacity and exercises of daily living like communication, learning, and decision-making. Speech, face expressions, body signs and other silent gestures are different ways to express emotions. Speech emotion classification is the analysis of verbal behaviour as an indicator of affect, with an emphasis on gestural features of speech. Its is based on the theory that the voice consists of a group of quantitatively measurable parameters reflecting the affective state of a person in currently expression. Automated emotional speech classification is an integral constituent of mounting research domains such as robotic, automobile sector, the entertainment sector, the marketing sector, and different parallel sectors [2].

Automated speech emotion classification may contribute in making speech recognition process more accurate. It appears that automated emotional speech classification will soon replace man-machine interaction. An emotion classification framework is intended to distinguish the emotional state being experienced by the speaker. The emphasis is generally on how something [3] is expressed, not what is expressed. In addition to approaches focusing only on analyzing the speaker's voice, a variety of methods can be used to identify emotional states. Some approaches include analysis of voice and spoken words while others focus only on facial expressions. Some analyze the responses of various emotional states in the human brain. In addition, some composite solutions use a mixture of the stated approaches. Generally speaking, the pipeline of human emotions analysis consists of two methods. The first method bifurcates emotions into two classes of discrete and distinct recognition. In the second method, emotional states are represented in 2D or 3D space, where parameters such as emotional distance, activity level, domination level, and pleasure level are noticed [4].

1.1 Speech Emotion Classification (SEC) System based on Machine learning

The current technological developments and evolving research hotspots such as machine learning (ML), audio processing and speech processing have made emotional conditions an integral element of man-machine communication. There are ever more studies relating to provisioning computers with multiple capabilities such as recognition, understanding and simulation of emotional states. A typical ML based

SEC system possess four crucial steps. The first step is to collect verbal samples [5]. Next feature vector is created by deriving features. The next step is intended to determine the most relevant features for distinguishing each emotional state. These features feed machine learning classifiers for classification. Figure 1 illustrates the process of speech emotion classification.

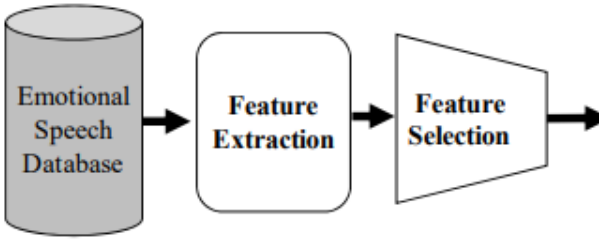

Figure 1: Emotional Speech Classification

The whole process of ML based Speech emotion detection is described below in detail:

i. Emotional speech database: Some widely used speech emotion databases in literature include German, English, Japanese, Spanish, Chinese, Russian, Dutch, etc. A key feature of the speech emotion database is the kind of emotional states expressed in verbal communication: whether they are fake or they are derived from the situations of daily life. The benefit of having mock speech is that the investigator has full control over the emotion he or she expresses and full control over the voice quality [6]. The shortcoming, however, is that the naturalness and spontaneity of speech degrades. At other side, the non-simulated emotion database includes a speech derived from real-time conditions such as call-centres, interviews, meetings, movies, short videos and similar situations where genuineness and naturalness are maintained. The drawback of these databases is that they do not have complete control over the expressed emotions. Also, there may be a problem of low sound quality.

ii. Feature extraction: The speech signal consists of plenty of metrics that indicate emotion related features. An important point in emotion classification is the selection of appropriate characteristics. Numerous regular features (e.g., energy, pitch, formant), and some spectrum features (e.g., linear prediction coefficient (LPC), mel-frequency cepstrum coefficient (MFCC) [7], and modulation spectral features) are extracted for speech emotion classification. Some of these features are described below:

- Mel-frequency cepstrum coefficient (MFCC): MFCC is the most common representative of the spectral quality of speech signals. These are ideal for voice signals recognition because it takes into account human insight sensitivity in terms of frequencies. Fourier transform and the energy spectrum per frame are estimated and mapped to the Mel-frequency extent. Discrete cosine transformation (DCT) of the Mel log energy is estimated, and the first 12 DCT coefficients includes MFCC values that are used for classification.
- Modulation spectral features (MSFs): MSFs are extracted from an auditory-induced long-lasting spectro-temporal illustration. These features are derived from simulating spectro-temporal (ST) processing carried out in the human speech system and treats regular acoustic frequency combined with the modulation frequency. The first step to obtain ST illustration is to disintegrate the speech signal using an audio filterbank [8]. The Hilbert envelope of the crucial-band output is then computed to produce the modulation signal. Next, frequency analysis is performed by applying a modulation filterbank to the Hilbert envelope. The spectral content of the modulation signals is called modulation spectra, and the obtained features are named MSFs.

iii. Feature selection: Feature selection in ML aims to decrease the number of features describing a dataset in order to improve the performance of a learning algorithm on a specified task. The purpose would be to maximize the classification accuracy in a given task for a specified learning algorithm. It reduces the number of features through collateral effects to adopt the final classifier framework. The purpose of feature selection (FS) is to select a subset of the relevant features from the original features based on some relevance assessment principle, which generally improves the detection accuracy [9]. This can significantly shorten the operating period of the learning algorithm. Recursive feature elimination (RFE) is an efficient feature selection technique. It selects either the ideal or worst performing feature using a model (e.g., linear regression or SVM) and then extracts the feature. These estimators allocate

weights to features (for example, the coefficients of a linear model), thus the RFE recursively considers a progressively decreasing feature set to select features. The estimator is initially trained on a preliminary feature set, and then the predictive strength of every feature is quantified. Then, the redundant features are separated from the available feature set. This process is repeated over the sorted set recursively until the required number of features are selected [10].

iv. Classification Methods: The extracted, standardized and selected features are used to build a feature vector database. Each data template in the database represents an example, that is, a feature vector, which is intended for classification. Since an appropriate emotion state is labeled for each instance, the use of a supervised classification algorithm is appropriate [11]. Several machine learning algorithms have been used for discrete emotion classification. Speech emotions can be classified using various machine learning algorithms. These algorithms are intended to learn from training templates and classify new perceptions based on this learning. In fact, there is no definite answer to the selection of learning algorithm; every method has its benefits and shortcomings. The commonly used machine learning algorithms for speech emotion detection are explained below:

a. Decision Tree Algorithm: The implementation of this algorithm relies on the concept of game theory. DT algorithm is implemented by selecting dividing features with the maximum information gain by means of equation (1) [12], as the probability of incidence of a feature is contingent on the volume of information that can be attached to the feature. could. Assume D and $H(D)$ represent data in a specific dataset, and c is the corresponding class, then

$$Gain(D,S) = H(D) - \sum_{i=1}^{S} p(D_i)H(D_i) \quad (1)$$

The information gain of a feature is quantified using the idea of entropy by gauging the scale of arbitrariness in a dataset, as depicted in Equation (2). Entropy tends to zero when the data refers to one dataset without uncertainty, as demonstrated in Equation (2) [13].

$$Entropy: H(p_1, p_2, \dots, p_s) = \sum_{i=1}^{S} (p_i[log(1 / p_i)]) \quad (2)$$

One of the main advantages of the DT classifier is that it continuously divides the specified dataset into subsets of all constituents, where the final subsets pertain same class.

b. K-Nearest Neighbour (KNN): This algorithm relies on distance measures between classes. It tries to search k features in the training data that appear nearest to the test instance. Next, it assigns the new model the most frequent label in these instances. Whenever a classification is performed, it firstly measures its distance to each feature available in the dataset and considers k nearest instances only [14].

c. Support Vector Machine (SVM): It is one of the most popular supervised ML (Machine Learning) algorithms in which data is divided into two or more classes to identify the optimal linear or non-linear boundary. First, the kernel function is chosen because this function is used to separate the data [15]. The extended kernel exists in the form of a linear function or a Gaussian function. After training a set of models, the settings are placed to choose the remaining parameters of the SVM (Support Vector Machine) algorithm for the model with low error. This algorithm is vulnerable to non-discriminatory magnitudes. Thus, a dimensionality reduction methos is applied to the input variables for simplifying the training and obtaining a better simplification to LR (Linear Regression) [16]. The main drawback of this algorithm is the increasing memory cost with the increase in the amount of data that needs processing. This algorithm can efficiently recognize non-linear and sparse input data. SVM (Support Vector Machine) is an important algorithm and many studies implement it to achieve the maximum efficiency.

LITERATURE REVIEW

2.1 Emotional Speech Classification using Deep Learning

B. Puterka, et.al (2019) presented a CNN (Convolutional Neural Network) algorithm which focused on analyzing the window time length to recognize the speech emotion [17]. The spectrograms, which a Hamming window of different length had

computed, considered to test the system. Two Conv (convolutional) layers and one FCL (fully-connect layer) was utilized to train the presented algorithm. The major aim was at determining the finest resolution of spectrograms concerning time or frequency while classifying the speech signal into 7 emotional categories. The outcomes indicated that the presented algorithm was effective to pre-process the speech signal in time domain while recognizing the speech emotion.

Y. Dong, et.al (2020) introduced an ASESM (affect-salient event sequences modelling) technique on the basis of CTC (connectionist temporal classification) [18]. A label sequence of sentence was employed as a chain of ASE (affect-salient event) states. This technique concentrated on modelling a CTC based CNN (convolutional neural network) for assigning a label of ASE to the emotional segments and Null to the non-emotional segments in automatic way. Thereafter, the introduced technique helped in avoiding the reaction delay compensation based on events as the target and mitigating the impact of noises. The experimental results of RECOLA dataset revealed that the introduced technique was effective for classifying the emotional speech

Y. -Y. Lin, et.al (2020) suggested an approach to recognize the emotional voice for assisting in understanding the demands of the people for offering them treatment in advance [19]. EMD (Empirical mode decomposition) model was put forward for enhancing the efficiency to detect and recognize the emotional sounds. Furthermore, the issue related to overfitting was alleviated using an ensemble CNN (Convolutional Neural Network). The experimental outcomes validated that the suggested approach offered the accuracy up to 91.6% while classifying the emotion speech in comparison with other methods.

A. Muppidi, et.al (2021) projected a QCNN (quaternion convolutional neural network) based SER (speech emotion recognition) system for encoding the Mel-spectrogram attributes of speech signals in an RGB quaternion domain [20]. This system was computed on three datasets: RAVDESS having 8 classes, IEMOCAP of 4 classes and Berlin EMO-DB consisted of 7 classes. The accuracy of the projected system was computed 77.87% on initial dataset, 70.46% on second, and 88.78% on last dataset. The experimental outcomes confirmed that the projected system was capable of encoding the internal dependencies as compared to other algorithm

2.1 Table

Author	Year	Technique Used	Findings	Limitations
B. Puterka, et.al	2019	CNN (Convolutional Neural Network)	The outcomes indicated that the presented algorithm was effective to pre-process the speech signal in time domain while recognizing the speech emotion.	The presented algorithm had generated poor outcomes with50% overlap of consecutive frames.
Y. Dong, et.al	2020	ASESM (affect-salient event sequences modelling) technique and CTC based CNN (convolutional neural network)	The experimental results of RECOLA dataset revealed that the introduced technique was effective for classifying the emotional speech.	Due to the constricted potential of decoding of this technique, the efficacy of arousal dimension was found lower.
Y. -Y. Lin, et.al	2020	EMD (Empirical mode decomposition) model and ensemble CNN (Convolutional Neural Network)	The experimental outcomes validated that the suggested approach offered the accuracy up to 91.6% while classifying the emotion speech in comparison with other methods.	This approach had not contained more emotional voice samples representing more varieties of emotional kinds due to which the services of recognizing the emotional speech worked ineffectively.
A. Muppidi, et.al	2021	QCNN (quaternion convolutional neural network) based SER	The accuracy of the projected system was computed 77.87% on initial dataset, 70.46% on	This system was proved unsuitable to discover other attributes of waveforms namely MFCC

		(speech emotion recognition) system	second, and 88.78% on last dataset. The experimental outcomes confirmed that the projected system was capable of encode internal dependencies as compared to other algorithms.	(Mel-frequency cepstral coefficients), chromograms, etc.

2.2 Emotional Speech Classification using Machine Learning

Y. Sun, et.al (2018) established a distance similarity algorithm and a KMC (K-means clustering) algorithm to evaluate the emotional speech samples and extract the neutral as well as 3 negative emotions such as anger, fear and sadness from EMO-DB and CASIA database [21]. These samples were considered to split the grades of 3 emotions. Meanwhile, SVM (support vector machine) was adopted to determine the split outcomes in the experimentation. The experimental outcomes demonstrated the smoothness of the established algorithm for recognizing all types of emotions.

J. Heredia, et.al (2022) developed an adaptive and flexible framework for recognizing a speech emotion which was applicable on multiple sources and modalities of information and managing diverse levels of data quality and missing data [22]. An analysis was performed on every modality for aggregating the partial outcomes. The prior technique known as EmbraceNet+ was put together with the developed framework. Various tests that integrated dissimilar modalities were carried out on the developed framework for quantifying the developed framework during classifying emotions with regard to 4 classes namely happiness, neutral, sadness, and anger. The outcomes exhibited the adaptability of the developed framework as it offered optimal outcomes for classifying the emotions against the classic techniques.

J. Pribil, et.al (2019) formulated a GMM (Gaussian Mixture Model) based system to detect the speech artefacts automatically [23]. This system was useful for detecting one or more emotions in synthetic speech. The continual GMM algorithm of classifying the emotional states was employed in 2-D (two dimensional) space of valence and arousal in the whole sentence. The major focus was on measuring the final change in the evaluated emotions. The formulated system functioned better to attain effective results in experimentation. Moreover, this system affected the number of mixtures, emotional classes, and kinds of speech attributes on the evaluated emotional shift.

S. Yan, et.al (2020) devised Wechat program of a system utilized to recognize the speech emotion on the basis of RF (random forest) algorithm. The fundamental goal of this system was to pre-process the gathered speech signals for alleviating the noise [24]. Thereafter, these signals were employed to generate sixteen acoustic attributes. Twelve statistical functions were exploited to the original acoustic attributes for acquiring the emotional attributes of speech. Two algorithms: RF (Random Forest) and SVM (Support Vector Machine) were implemented to classify the speech emotions on BSE (Berlin Speech Emotion) dataset. The results depicted that the primary algorithm offered 83% accuracy and the later offered 89% accuracy. The RF algorithm was effectual for constructing a system to recognize the speech emotion.

2.2 Table

Author	Year	Technique Used	Findings	Limitations
Y. Sun, et.al	2018	A distance similarity algorithm and a KMC (K-means clustering) algorithm	The experimental outcomes demonstrated the smoothness of the established algorithm for recognizing all types of emotions.	This approach had not performed well on the professional voice library audio which the ORI recorded.

J. Heredia, et.al	2022	An adaptive and flexible framework	The outcomes exhibited the adaptability of the developed framework as it offered optimal outcomes for classifying the emotions against the classic techniques.	This framework consumed much time to execute in real time scenarios and on datasets containing numerous modalities.
J. Pribil, et.al	2019	GMM (Gaussian Mixture Model) based system	The formulated system functioned better to attain effective results in experimentation. Moreover, this system affected the number of mixtures, emotional classes, and kinds of speech attributes on the evaluated emotional shift.	The formulated was inapplicable for localizing the artefact position within the sentence.
S. Yan, et.al	2020	RF (Random Forest) and SVM (Support Vector Machine)	The results depicted that the primary algorithm offered 83% accuracy and the later offered 89% accuracy. The RF algorithm was effectual for constructing a system to recognize the speech emotion.	The format of the gathered speech signal in devised program was MP3 whose processing was not possible in open SMILE software.

2.3 Emotional Speech Classification using General Techniques

O. Kwon, et.al (2019) investigated an E2E-TTS (end to end-text to speech) system called controlled weight-based method to create emotion embedding vectors based on the trained GSTs (global style tokens) [25]. First of all, all the samples were distributed according to the kind of every emotion. After that, the centroid of the distribution was employed as an emotion-specific weighting value with the objective of controlling the emotion of synthesized speech. In the end, the investigated system performed more effectively as compared to other techniques concerning perceptual quality and accuracy while classifying the speech emotion.

M. Manjutha, et.al (2019) constructed a new technique to enhance the accuracy for classifying the speech emotion [26]. This technique employed PSO (Particle Swarm Optimization) and SFO (Synergistic Fibroblast Optimization) algorithms for selecting the optimal features from the traditional methods. The optimized cepstral attributes, extracted from the pre-processed Tamil speech data, were helped in discriminating diverse kinds of speech signals such as normal, moderate and sever stutter. For this, ML (machine learning) algorithms namely SVM (Support Vector Machine) and NB (Naive Bayes) were implemented. The SFO algorithm with NB yielded an accuracy around 96.08% in comparison with other algorithms.

P. Powroznik, et.al (2021) intended a DTW (Discrete Wavelet Transform) for classifying the speech emotion [27]. The technique to process the scalograms was put forward for extracting the input data for NLP (natural language processing) models so that the emotional state was recognized. The databases consisted of recordings of emotional speech was applied in the experimentation of the intended model. The results validated that the intended algorithm offered efficacy of 94%. Moreover, the fuzzy classification algorithms led to enhance the time and accuracy to classify the speech emotion.

2.3 Table

Author	Year	Technique Used	Findings	Limitations
O. Kwon, et.al	2019	Controlled weight-based method	In the end, the investigated system performed more effectively as compared to other techniques concerning perceptual	This system was not suitable to produce effective or manifold categories of emotional speech in a single sentence.

			quality and accuracy while classifying the speech emotion.	
M. Manjutha, et.al	2019	PSO (Particle Swarm Optimization) and SFO (Synergistic Fibroblast Optimization) algorithms	The SFO algorithm with NB yielded an accuracy around 96.08% in comparison with other algorithms.	This technique was incapable of classifying in depth stuttering syllable repetition, etc., for Tamil dataset.
P. Powroznik, et.al	2021	DTW (Discrete Wavelet Transform)	The results validated that the intended algorithm offered efficacy of 94%. Moreover, the fuzzy classification algorithms led to enhance the time and accuracy to classify the speech emotion.	The accuracy was not exceeded from 94% as the learning data and the number of samples of emotional speech recordings were selected at random.

CONCLUSION

There are several emotional speech databases that are extensively used in the literature: German, English, Japanese, Spanish, Chinese, Russian, Dutch etc. One main characteristic of an emotional speech database is the type of the emotions expressed in the speech: whether they are simulated or they are extracted from real life situations. The advantage of having a simulated speech is that the researcher has a complete control over the emotion that it is expressed and complete control over the quality of the audio. The schemes of deep learning and machine learning for the speech emotion are reviewed. The deep learning techniques give maximum accuracy for the speech emotion classification. In future novel deep learning model will be designed which gave maximum accuracy for the speech emotion recognition.

REFERENCES

[1] J. Wang, Y. Chin, B. Chen, C. Lin and C. Wu, "Speech Emotion Verification Using Emotion Variance Modeling and Discriminant Scale-Frequency Maps," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 10, pp. 1552-1562, Oct. 2015

[2] R. V. Darekar and A. P. Dhande, "Improving emotion detection with speech by enhanced approach," 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), 2016, pp. 364-369

[3] Z. Qing, W. Zhong and W. Peng, "Research on Speech Emotion Recognition Technology Based on Machine Learning," 2020 7th International Conference on Information Science and Control Engineering (ICISCE), 2020, pp. 1220-1223,

[4] T. Ramakrishna and G. Krishna, "Significance of Accurate Vowel Region Detection for Speech based Emotion Recognition," 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), 2021, pp. 345-349

[5] O. A. Mohammad and M. Elhadef, "Arabic Speech Emotion Recognition Method Based on LPC And PPSD," 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2021, pp. 31-36

[6] H. Nishizaki and K. Watase, "Emotion classification of spontaneous speech using spoken term detection," 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), 2017, pp. 1-5

[7] S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in IEEE Access, vol. 8, pp. 60382-60391, 2020

[8] R. Lotfian and C. Busso, "Lexical Dependent Emotion Detection Using Synthetic Speech Reference," in IEEE Access, vol. 7, pp. 22071-22085, 2019

[9] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018

[10] J. Oliveira and I. Praça, "On the Usage of Pre-Trained Speech Recognition Deep Layers to Detect Emotions," in IEEE Access, vol. 9, pp. 9699-9705, 2021

[11] Z. Zhao et al., "Exploring Deep Spectrum Representations via Attention-Based Recurrent and

Convolutional Neural Networks for Speech Emotion Recognition," in IEEE Access, vol. 7, pp. 97515-97525, 2019
[12] M. T. Teye, Y. M. Missah, E. Ahene and T. Frimpong, "Evaluation of Conversational Agents: Understanding Culture, Context and Environment in Emotion Detection," in IEEE Access, vol. 10, pp. 24976-24984, 2022
[13] M. S. Hossain and G. Muhammad, "Emotion-Aware Connected Healthcare Big Data Towards 5G," in IEEE Internet of Things Journal, vol. 5, no. 4, pp. 2399-2406, Aug. 2018
[14] J. Zhang, L. Xing and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition", Computers & Industrial Engineering, vol. 12, no. 4, pp. 1187-1194, 10 March 2022
[15] H. Lai, H. Chen and S. Wu, "Different Contextual Window Sizes Based RNNs for Multimodal Emotion Detection in Interactive Conversations," in IEEE Access, vol. 8, pp. 119516-119526, 2020
[16] R. Thirumuru, K. Gurugubelli and A. K. Vuppala, "Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition", Digital Signal Processing, vol. 1, no. 8, pp. 76454-76460, 29 Oct. 2021
[17] B. Puterka, J. Kacur and J. Pavlovicova, "Windowing for Speech Emotion Recognition," 2019 International Symposium ELMAR, 2019, pp. 147-150
[18] Y. Dong and X. Yang, "Affect-Salient Event Sequences Modelling for Continuous Speech Emotion Recognition Using Connectionist Temporal Classification," 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), 2020, pp. 773-778
[19] Y. -Y. Lin, J. -Y. Yang, C. -Y. Kuo, C. -Y. Huang, C. -Y. Hsu and C. -C. Liu, "Use Empirical Mode Decomposition and Ensemble Deep Learning to Improve the Performance of Emotional Voice Recognition," 2020 IEEE 2nd International Workshop on System Biology and Biomedical Systems (SBBS), 2020, pp. 1-4
[20] A. Muppidi and M. Radfar, "Speech Emotion Recognition Using Quaternion Convolutional Neural Networks," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6309-6313
[21] Y. Sun, X. Zhang, J. Ma and Z. Zhang, "Classification of Negative Emotion Speech Intensity Based on Similarity Algorithm," 2018 IEEE International Conference on Information Communication and Signal Processing (ICICSP), 2018, pp. 94-97
[22] J. Heredia et al., "Adaptive Multimodal Emotion Detection Architecture for Social Robots," in IEEE Access, vol. 10, pp. 20727-20744, 2022
[23] J. Pribil, A. Pribilova and J. Matousek, "Artefact Determination by GMM-Based Continuous Detection of Emotional Changes in Synthetic Speech," 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), 2019, pp. 45-48
[24] S. Yan, L. Ye, S. Han, T. Han, Y. Li and E. Alasaarela, "Speech Interactive Emotion Recognition System Based on Random Forest," 2020 International Wireless Communications and Mobile Computing (IWCMC), 2020, pp. 1458-1462
[25] O. Kwon, I. Jang, C. Ahn and H. -G. Kang, "An Effective Style Token Weight Control Technique for End-to-End Emotional Speech Synthesis," in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1383-1387, Sept. 2019
[26] M. Manjutha, P. Subashini, M. Krishnaveni and V. Narmadha, "An Optimized Cepstral Feature Selection method for Dysfluencies Classification using Tamil Speech Dataset," 2019 IEEE International Smart Cities Conference (ISC2), 2019, pp. 671-677
[27] P. Powroznik, P. Wojcicki and S. W. Przylucki, "Scalogram as a Representation of Emotional Speech," in IEEE Access, vol. 9, pp. 154044-154057, 2021