The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 798 – 807

Speech-based Emotion Recognition Methodologies

S. M. M. Naidu, Vedant Shinde, Varun Kulkarni, Akanksha Wadekar, Y. A. Chavan

International Institute of Information Technology (I^{2}IT), Pune, India

mohans@isquareit.edu.in

Abstract

Speech is used by humans to communicate with one another. Speech not only carries the content but also context of the information. There have been advancements in the field of identification and classification of speech. Also, the emergence of consumer applications that are speech-centric has also added to the rigor for perfecting human-machine interaction. This study aims to review the research in this field focusing on the papers from 2010 onwards.

Keywords— Speech-based Emotion Recognition, Neural Networks, Utterance-level Features, Speech Corpus

I. INTRODUCTION

Speech is one of the primary ways through which humans communicate along with other factors like posture, facial expressions, behavior, etc. Traditionally, machines have interacted with humans through text. Recently advancements in consumer applications have facilitated communication with machines using speech. The next step in this trend is to make the interactions between computer and machine interfaces personalized and seamless. Applications of speech-based emotion recognition range from detecting severity of calls in emergency call-centers to automatic safety protocols to detect 'lazy' emotions. Speaker's emotions can be determined using a multitude of factors like contents of speech, physical attributes of the speaker like age and gender and characteristics of the sound itself. This makes determining emotion difficult. Extensive research has been done in increasing the accuracy of the classification models. This paper aims to list the research carried out in this domain so far.

The paper is organized as follows. Section II is further divided into two sections which outline different conventional and non-conventional methodology-based research papers. In section III, discussion of the gatherings from reviewing these papers. Finally, this paper is concluded in section IV.

II. REVIEW

Selecting a suitable database is crucial in determining performance of a speech recognition system (Ayadi et al. 2011). Important factors to be considered while selecting a database include the number of samples, type of database (natural/simulated), number of people or actors, language, etc. Anger, fear, sensory pleasure, sadness, excitement, pleasure, amusement, satisfaction, contentment, pride, shame, guilt, disgust, contempt, embarrassment and relief are basic emotions (Eckman 1992). Determination of emotion using speech is more difficult than using, say, facial expression. Emotions, like anger and happy, are highly misclassified (Ghosh et al. 2016). This is because anger has a high fundamental frequency. Also, it depends on factors like whether the data is real or simulated. Differences in language and gender could also lend to misclassification of emotions. Some databases are developed using voice actors, others are developed using data from call centers, TV and radio programs. Selecting proper features for speech-based emotion recognition is paramount in increasing the classification accuracy.

A. Features and databases used in Non-ANN based papers

In 2003, Schuller et al. created a database consisting of speech samples in English and German languages. The corpus was collected in a soundproof room using two methods to collect samples from five individuals. The larger test set, consisting of four speakers, involved acting out emotions, while the other method collected spontaneous emotions for comparison to recognition results. To model each emotion, a single-state Gaussian Mixture Model (GMM) Hidden Markov Model (HMM) was used, with up to four mixtures of Gaussian distributions approximating the Probability Distribution Function of each feature. Continuous HMMs (CHMMs) were formed using Baurn Welsh Reestimation, and four Gaussian mixtures were used to estimate the probability density function. Overall, seven models were created, each corresponding to an emotional state. The model with the maximum likelihood was selected to represent the recognized emotion.

In 2015, Wang et al. discussed three emotional speech databases: the German Emotional Corpus (EMODB), the Chinese Emotional Database (CASIA), and the Chinese Elderly Emotional Speech Database (EESDB). EMODB contains 10 sentences that cover seven classes of emotion commonly used in everyday communication, including anger, fear, happiness, sadness, disgust, boredom, and neutrality. CASIA consists of 9,600 wave files featuring different emotional states, including happiness, anger, sadness, surprise, fear, and neutrality, with 4 actors. EESDB contains recordings from 11 Chinese elderly males and females over the age of 60, and includes seven classes of emotions, including neutrality. The authors extracted the first 20 Harmonic Coefficients and calculated the minimum, maximum, mean, median, and standard deviation of Fourier Parameters Feature Vector for each database. They also extracted Mel-frequency cepstral coefficients (MFCC) features to compare with Fourier Parameters features. To eliminate variability due to different speakers and recordings, while maintaining the ability to effectively discriminate between emotions, they included maximum, minimum, median, and standard deviation as MFCC features.

In 2017, Ghai et al. utilized the Berlin database, which contains 535 emotional utterances from 10 German speakers. The total length of the utterances is 1487 seconds, with an average length of 2.77 seconds, and each file consists of 16-bit PCM and mono channels. The audio files are divided into frames, and feature extraction is performed. The audio signal is sampled at 16000 Hz, and the frame duration is selected as 0.025 seconds. The majority of the audio signal has a frequency of 8 kHz, although ideally it would be 16 kHz. The authors extracted features for each frame, including Mel-frequency cepstral coefficients (MFCC) to represent the logarithmic perception of loudness and pitch, the Mel scale to give a unit of pitch such that equal distances in pitch sounds equally distant to the listener, and energy to represent the intensity of the speech, which is calculated by the sum of the square of the amplitude of each frame. These features are fed to a Support Vector Machine (SVM) as a classification model. To improve classification accuracy, additional copies of the SVM were fitted on the same dataset, with the weights of incorrectly classified samples adjusted to make successive classifiers focus on more complex cases. Gradient boosting was used for this purpose. The authors also used Random Decision Forest classifiers, which initialize multiple decision trees during training and output the mean/mode of the class of individual trees to prevent overfitting of the training data. The Random Decision Forest classifier gave the highest accuracy of 81.05% for the Berlin database.

In their study, (Deshmukh et al. 2019) utilized the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) as their audio dataset to analyze three emotions - anger, happiness, and sadness. The dataset was divided into training and testing sets. Preprocessing of the audio signals was carried out as the first step, with the continuous time audio signal being converted to a discrete time signal for better analysis through sampling. The low variation parts of the audio were removed as they did not contribute to emotion determination, and a high-pass filter was applied to emphasize the high-frequency components, which represented the rapidly changing signal. Parts of the audio that consisted of absolute silence were removed, as they did not contribute to emotion determination. The input signal was divided into small constituent frames of a specific time interval to achieve a stationary form, with a frame duration of 20-30 ms being ideal for speech processing. Windowing techniques such as rectangular, Hamming, Blackman, etc., were applied for analyzing the signal. After preprocessing, feature extraction was carried out by extracting the energy from each frame, and then obtaining a short-term energy plot of the signal, which indicated the high-frequency frames using larger peaks. After that, MFCC feature vector extraction was performed for each frame. MFCC represents the short-term power spectrum envelope, which determines various sound characteristics. Since MFCC is not extractable in the time domain, it is converted to the frequency domain using FFT, and the power spectral density is calculated. Mel filter banks were initialized and applied on frames to get 26 output values per frame, which relate the perceived sound frequency to its actual frequency. The logarithm of frame energies is performed to normalize the variation in loudness. Additional extraction, such as pitch classification, was performed. A linear SVM classification algorithm was used, which was divided into two classes. Therefore, models were built for each emotion versus the rest. The classifier formed a model for each of the mentioned emotion classes earlier, which was tested using a testing dataset.

Reference & Database	Emotional Labels	Features	Techniques & Results
Iliev et al. (2010) & Independent Database using RODE NT2	Happy, Angry, Sad , Neutral	Glottal Symmetry and MFCC	Optimum Path Forest Classifier & OPF Classifier performed better than SVM Classifier
Yeh and Chi (2010) & Berlin Emotional Speech Database(EMO DB)	Anger, Happiness, Fear, Disgust, Boredom, Sadness , Neutral	Spectral and Prosodic Features	SVM Classifier & Spectro temporal modulation features are more robust to additive white and babble noises than MFCC's combined with Prosodic Features
Espinosa et al. (2010) & German EMO DB	Arousal, Valence and Dominance	Voice Quality, Spectral and Prosody	SVM and Pace Regression Classifiers & In Activation, Spectral, Voice Quality, Energy Contour and Pitch Contour are the most important feature groups.
Bitouk et al. (2010) & English Emotional Speech Database from Linguistics Data Consortium (LDC), Berlin Database	Anger, Fear, Disgust, Happy, Neutral, Sadness	MFCC computed over Unstressed Vowels, Stressed Vowels are Consonants in utterance.	Class Level Spectral Feature Analysis, SVM Classifiers with Radial Basis Kernels & Combined Accuracy: 1. LDC Dataset: Anger-71.9%, Fear-58.6%, Disgust-48.4%, Happy- 59.2%, Sadness-59.4%, Neutral-79.8% 2. Berlin Database:

			Anger- 89.0% , Fear- 82.8%, Disgust- 88.2% , Happy- 65.5% , Sadness- 92.9%, Neutral- 89.4%
Sun and Moore (2011) & SEMAINE Corpus	Activation, Expectation, Power, Valence	Glottal Waveform Parameters, TEO	Glottal Waveform Parameters, and Teager Energy Parameter & Average Accuracy rate: 1.Activation:64.3% 2. Expectation: 62.3% 3.Power:61.2% 4.Valence:58.6%
Rozgic et al. (2012) & USC-IEMOCAP Database	Anger, Happiness, Sadness, Neutral	Spectral, Prosodic and Lexical	SVM, Gaussian Mixture Model & Emotion Recognition Accuracy of 65.7% is delivered by the fusion of Lexical and Acoustic Features.
Jeon et al. (2013) & Chinese ACC, English AMA, German Emo DB, ENGLISH IEMOCAP Databases	Angry, Happy, Neutral and Sad	Spectral and Prosody	WEKA implemented SVM, 4-way classification discrimination, Polynomial Kernel of Order 3 & Automatic Classification, in general, shows better performance on Within Corpus than Cross Corpus. It also has degradation on Cross Corpus than Human Perception.
Gangamohan et al. (2013) & IIT-H Telugu Emotion Database, Berlin EMO DB Database	Angry, Happy , Neutral , Sad	Excitation Source	KL Distance Values & Accuracy: 1.IIT-H Telugu Emotion Database:76% 2.Berlin EMO DB Database: 69%
Gangamohan et al. (2014) & Berlin Emo-DB, IIIT-H Telugu Emotion Database	Neutral, Angry , Happy	Excitation Source Signal, Strength Of Excitation, Feature Band Energy	Zero Frequency Filtering Method, Spectral Band Magnitude Energies & Accuracy 1. IIT-H Telugu Emotion Database:75% 2. Berlin EMO-DB:68%

Table 1. Literature survey of Features, databases, techniques and results of ANN based papers

B. Features and databases used in ANN based papers

4 public databases were used by (Huang et al 2014): surrey audio-visual expressed emotion (SAVEE database), Berlin emotional database (Emo-DB), Danish emotional speech database (DES) and mandarin emotional speech database (MES). Spectrogram was provided as input to the convolutional neural network. Using unsupervised feature learning the system obtains a long feature vector. This is fed to semi-supervised CNN which produces affect-salient features and nuisance features the affect-salient features are provided as input to a linear SVM.

Berlin database was used by (Lalitha et al 2015).7 emotions were classified- anxiety, disgust, happy, boredom, neutral, sadness, angry. Features that were extracted were teager energy, LPCC,mel-energy spectral dynamic coefficients (MEDC),ZCR, shimmer, spectral roll-off, spectral centroid , HNR, short time energy, entropy of energy and pitch this features were feed to SVM for classification. Accuracy of 85.71% was achieved.

(Ghosh et al. 2016) have used representational learning for Speech Emotion Recognition (SER). Glottal volume velocity spectrogram is used along with MFCC for feature extraction. They aimed to investigate whether spectrograms are an effective representation of audio for SER. Auto encoders were used to learn the lower dimensional representation of input data. tanh was used as an activation function. Recursive Neural Networks (RNNs) are effective for learning time series data and temporal correlation but vanishing gradient problem is encountered. To overcome this problem, bidirectional Long short-term memory (LSTM) is used along with RNN. Majority voting scheme is used for predicting emotions label for each time step(frame). Softmax layer is used for final classification. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database was used.4 emotion labels are classified- neutral, angry, sad, happy. Glottal source waveform filters out the varying factors such as speakers, identity, and phonetic information. IAIF algorithms were used to calculate glottal flow. Happy and angry classes were highly misclassified. Accuracy of 51.64% was achieved using a representational learning-based model.

(Lim et al. 2016) report that conventional methods use low-level features and train the machine accordingly. Selecting and optimising these features is difficult and time consuming. Deep neural architectures overcome this problem by naturally progressing from low-level to high-level structures. 2D representation of audio is analysed. Most commonly, time-frequency analysis. STFT is used to get this 2D representation. This 2D representation is fed to CNNs and Long Short-term Memory(LTSM) architectures. Each level in the deep learning models increases the level of abstraction. Results of each audio frame are classified using a sum of probabilities. CNN is ideal for analysing multi-dimensional data. Hence, CNN is suitable for audio applications. A Berlin database was used.

(Trigeorgis et al. 2016) segmented raw waveforms into 6 sec long sequences. These 6 sec sequences were further divided into 150 smaller subsequences. ReCOLA dataset, which is a french dataset, was used. Raw waveforms(16kHz) were fed to the convolutional layer. Output of the convolutional layer was input to recurrent LSTM. The convolution layers replace the requirement of hard-engineering features which were used till now. The researched method performs significantly better in comparison to traditionally designed features on the ReCOLA database.

(Badshah et al. 2017) directly fed spectrograms to the classification model. Spectrogram is a 2D time frequency representation of an audio signal. Intensity of audio signal at any given point can be determined by the amplitude as well as colour at that point. This spectrogram is obtained only by applying FFT to speech signals. CNN was used for feature extraction and the softmax layer for classification of emotions. Extraction of distinct features from spectrograms is harder than in images. Berlin speech emotion dataset was used. 7 emotions: anger, boredom, disgust, fear, sad, happy and neutral were classified. Multiple spectrograms were created per audio files. Prediction after analysing individual spectrograms was used to update the belief values of that audio file for all emotions. Two experiments were performed: traditional learning and transfer learning. Traditional learning performed better than transfer learning.

(Zhao et al. 2018) used the IEMOCAP database. CNN was used to extract features and its output was tested against various models like Fully Convolutional Network, Attention-Long Short-term Memory model and the proposed attention- Bidirectional LSTM-FCM models.

(Wu et al. 2019) states 2 major challenges in SER are extraction of high-level features and construction of utterance-level features. Traditional SER required hard-engineering of low-level features, like pitch. Recently, Spectrogram is directly fed to CNNs to extract relevant features. But CNNs lose spatial and global information. This is because spectrograms of frames (chunks) of an audio file are classified individually. Both local and global information is paramount in determining emotion. Further, pooling is used to further classify utterance-level features from the output of CNN. This gives results by discriminating between regions of spectrogram to only use relevant information. But spatial information is lost when pooling is used. Positional info of intensity is lost in time-frequency representation. Capsule networks have been proposed to overcome this. Spatial information, which CNNs fail to capture, are captured by Capsule networks. Group of neurons which maintain activity vectors form a capsule. Magnitude of the vector represents the probability of activity and the direction of the vector represents the detailed Instantiation Parameter like rotation & translation info. A routing algorithm is used to connect similar activity vectors at lower levels to activate corresponding layers at upper levels. Temporal information viz. Is vital in SER has not been taken into consideration by CapsNets previously. In this paper, Input is sliced into windows. These windows are provided to the CapsNet iteratively. Aggregate output of 1st CapsNet is provided to 2nd CapsNet. The output of 2nd CapsNet gives utterance-level features. In this way, temporal information is retained. IEMOCAP database was used. 4 emotions -> Neutral, angry, happy, sad.

The feature extraction and model of emotion recognition are main in SER. In Nischay Parikh et al. (2021), attention-based mechanism was used to calculate the relevance weights signal in time domain and its feature and then by differentiating the weight by another time domain in the signal and then by choosing the greater weight from the time signal than it contributes more on salient signal so that the main signal would not be lost. So, further for data processing the RAVDESS dataset was quadrupled (5760 audio-samples) by using the method AWGN(additive gaussian white noise) for segmenting the features into spectrograms.

Kadiri et al. (2015)	IIIT-H Telugu Emotion Database, Berlin Speech Emotion Database	Anger, Happy, Neutral and Sad	Excitation Source Signal	Hierarchical Binary Decision Tree	Accuracy: 1. SVM Classifier: 79.5%
Lotfian & Busso (2015)	SEMAINE Database	Arousal , Valence	WPCC	Polynomial Classifier	2.1%-2.8% Improvement in Performance in classifying low versus high levels of arousal and valence.

Kunxia et al. (2015)	German Emotional Corpus(EMODB), Chinese Emotional Database(CASIA) , Chinese Elderly Emotional Speech Database(EESDB)	Happiness, Sadness, Angry, Neutral ,Fear, Disgust	MFCC, LPCC, Fourier Parameter	SVM Classifier, Feature Normalization	Improvement in Speaker Independent Emotion Recognition using proposed FP Features: 1. German Database: 16.2 Points 2. CASIA Database: 6.8 Points 3. EESDB Database: 16.6 Points
Lalitha et al. (2015)	Berlin Emo Database	Anxiety, Disgust, Happy, Boredom, Neutral , Sadness, Anger	DWT, LPCC, MFCC, MEDC	Multi SVM Classifier,	Accuracy of 85.71% was achieved
Ruhul et al. (2016)	Berlin Emo Database, Danish Emotional Database, INTERFACE05, LDC Emotional Speech and Transcripts, Interactive Emotional Dynamic Motion Capture	Anger, Disgust, Fear , Happiness, Joy , Sadness	Spectrogram	SVM Classifier, Deep Boltzmann Machine(DBM) , Recursive Neural Network(RvNN), Deep Belief Network(DBN)	Multimodality has potential in Improving SER
Mohan et al. (2017)	Berlin Database	Anger, Boredom, Disgust, Anxiety, Happiness, Sadness, Neutral	MFCC , Energy as a Feature	Random Decision Forest Classifier, SVM Classifier, Gradient Boosting	Maximum Accuracy of 81.05% obtained by using Random Decision Forest Classifier.

Deshmukh et al. (2019)	RAVDEES Database	Anger, Happiness, Sadness	MFCC	Preprocessing, Sampling, SVM Linear Classification	Accuracy of all the 3 Features: 80%
Rahul et al. (2019)	Berlin Emotional Database, Danish Emotional Database, INTERFACE 05 Database, LDC Emotional Speech and Transcripts, Interactive Emotion Dyadic Emotion Capture	Anger, Disgust, Fear , Happiness, Joy, Sadness	MFCC	Deep Boltzmann Machine(DBM), Recursive Neural Network(RvN N), Deep Belief Network(DBN)	Gaussian Mixture Model(GMM) is compared with DNN-HMM.
Xu et al. (2021)	RAVDESS Database	Surprised, neutral, calm, happy, sad, angry, fearful, disgust	MFCC, CNN and attention mechanism for weight calculation purpose	AWGN (additive Gaussian white noise) to quadruple the dataset (5760 audio-samples)	Accuracy was found to be 80.46%
Nishchay et al. (2021)	RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)	Calm, Happiness, Fearfulness and Disgust	Mfcc, chroma, mel spectrogram	MLP (Multi-Layer Perceptron) Classifier	The accuracy of SER by using real-time recognition was found to be 78.65%

Table 2. Literature survey of Features, databases, techniques and results of Non-ANN based papers

III. DISCUSSION

As can be observed from the trends in the tabular representations of ANN and non-ANN based speech emotion recognition techniques, it is clear that ANN-based models didn't emerge till the later half of this decade. Instead of hard-engineering features of the audio samples, which is cumbersome and ineffective if not done correctly, a 2D representation of the audio signal was fed directly to a neural network. RNNs, along with LSTMs, proved to be ideal for classification of emotions as they could hold not only the local but also the global, spatial and temporal information of the audio signal. SER is still in it's infancy as it is computationally expensive to classify and there is a lot of ambiguity in determining emotions like 'happy' and 'anger'. Also, the accuracy of classification ranges anywhere from 50% - 80% which is not effective for critical applications.

IV. RESULTS AND CONCLUSION

In this review paper, an array of research papers from 2010 onwards were studied and catalogued into 2 categories: ANN-based and Non-ANN-Based. Parameters studied for these papers included features, databases, classifiers and results. Major Classification models used in most of the papers included SVM, HMM, CNN, RNN, LTSM, Capsule networks or some combination of these. A 2-D representation of the audio signal in time-frequency domain was used in almost all the research work. We have found out the accuracy of 57.71% by four emotions extraction- fear, calm, happy, disgust.

REFERENCES

1. Badshah, A.M., Ahmad, J., Rahim, N., & Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. International Conference on Platform Technology and Service (PlatCon), 2017, 1-5. DOI: 10.1109/PlatCon.2017.7883728.
2. Xu, A. D., & Zhou R. (2021). Speech emotion recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. Journal of Physics: Conference Series, 1861(1): 012064.
3. Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 2003), II-1. DOI: 10.1109/ICASSP.2003.1202279.
4. Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. Speech Communication, 52(7–8): 613–625.
5. Eckman, P. (1992). An argument for basic emotions. Cognition and Emotion. 6: 169–200.
6. Ayadi, M., Kamel, M. S., & Karray, F., (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 1(44): 572–587.
7. Espinosa, H. P., Garcia, J. O., & Pineda, L. V. (2010). Features selection for primitives' estimation on emotional speech. In: ICASSP, Florence, Italy, 5138–5141.
8. Deshmukh, G., Gaonkar, A., Golwalkar, G., & Kulkarni, S. (2019). Speech based emotion recognition using machine learning. In: 3rd International Conference on Computing Methodologies and Communication (ICCMC 2019), 812-817. DOI: 10.1109/ICCMC.2019.8819858.
9. Trigeorgis, G. et al. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), 5200-5204. DOI: 10.1109/ICASSP.2016.7472669.
10. Gangamohan, P., Kadiri, S. R., & Yegnanarayana, B. (2013). Analysis of emotional speech at sub segmental level. In: Interspeech, Lyon, France, 1916–1920.
11. Gangamohan, P., Kadiri, S. R., Gangashetty, S. V., & Yegnanarayana, B. (2014). Excitation source features for discrimination of anger and happy emotions. In: INTERSPEECH, Singapore, 1253–1257.
12. Ghosh, S., Eugene L., Louis-Philippe, M., & Stefan, S. (2016). Representation learning for speech emotion recognition. In: Interspeech, 3603–3607.
13. Iliev, A. I., Scordilis, M. S., Papa, J. P., & Falco, A. X. (2010). Spoken emotion recognition through optimum-path forest classification using glottal features. Computer Speech and Language, 24(3):445–460.

14. Jeon, J. H., Le, D., Xia, R., & Liu, Y. (2013). A preliminary study of cross-lingual emotion recognition from speech: Automatic classification versus human perception. In: Interspeech, Layon, France, 2837–2840.
15. Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L. (2015). Speech emotion recognition using Fourier parameters. IEEE Transactions on Affective Computing. 6(1): 69-75. DOI: 10.1109/TAFFC.2015.2392101.
16. Kadiri, S. R., Gangamohan, P., Gangashetty, S. V., & Yegnanarayana, B. (2015). Analysis of excitation source features of speech for emotion recognition. In: Interspeech 2015, Dresden, 1324–1328.
17. Ruhul, A. K., Edward, J., Mohammad, I. B., Tariqullah, J., Mohammad, H. Z., & Thamer, A. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE Access 7 (2019): 117327–117345.
18. Lotfian, R., & Busso, C. (2015). Emotion recognition using synthetic speech as neutral reference. In: IEEE International conference on ICASSP-2015, 4759–4763.
19. Ghai, M., Lal, S., Duggal, S., & Manik, S. (2017). Emotion recognition on speech signals using machine learning. In: International Conference on Big Data Analytics and Computational Intelligence (ICBDAC 2017), 34-39. DOI: 10.1109/ICBDACI.2017.8070805.
20. Rozgic, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Vembu, A. N., & Prasad, R. (2012). Emotion recognition using acoustic and lexical features. In: INTERSPEECH, Portland, USA, 2012.
21. Lalitha, S., Mudupu, A., Nandyala, B. V., & Munagala, R. (2015). Speech emotion recognition using DWT. In: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC 2015), 1-4. DOI: 10.1109/ICCIC.2015.7435630.
22. Sun, R., & Moore, E. (2011). Investigating glottal parameters and teager energy operators in emotion recognition. In: Affective Computing and Intelligent Interaction, 425–434.
23. Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2016), 1-4. DOI: 10.1109/APSIPA.2016.7820699.
24. X. Wu et al. (2019). Speech emotion recognition using capsule networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), 6695-6699. DOI: 10.1109/ICASSP.2019.8683163.
25. Yeh, L., & Chi, T. (2010). Spectro-temporal modulations for robust speech emotion recognition. In: INTERSPEECH, Chiba, Japan, 2010, 789–792.
26. Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM international conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 801–804. DOI: https://doi.org/10.1145/2647868.2654984.
27. Zhao, Z., Zhao, Y., Bao, Z., Wang, H., Zhang, Z., & Li., C. (2018). Deep Spectrum Feature Representations for Speech Emotion Recognition. In: Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data (ASMMC-MMAC'18). Association for Computing Machinery, New York, NY, USA, 27–33. DOI: https://doi.org/10.1145/3267935.3267948.