University Of AL-Qadisiyah

Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)

A Survey for Emotion Recognition Based on Speech Signal

Fatima A. Hameed [a]***,Dr. Loay E. George*** [b] *

[a] *Informatics Institute for Postgraduate Studies/ Iraqi Commission for Computers & Informatics (IIPS/ICCI),Baghdad, Iraq. Ms202010636@iips.icci.edu.com*

[b] *University of Information Technology And Communication , Baghdad, Iraq , Loayedwar57@uoitc.edu.iq*

ARTICLE INFO

Article history:

Received: 02 /03/2022

Rrevised form: 28 /03/2022

Accepted : 14 /04/2022

Available online: 20 /04/2022

Keywords:

Speech Analysis, Emotional speech databases, Feature extraction, classification, Audio Speech Characteristics

ABSTRACT

There are many characteristics of human beings, such as fingerprints, DNA, and retinal pigmentation that are essential. A person's voice is unique to each individual. Humans use speech to communicate their thoughts and feelings. The process of determining one's mental state involves expressing one's basic emotions in words. A person's emotions play a significant part in his or her daily existence. In order to convey one's thoughts and feelings to others, it is essential to use this method. Emotions can be discerned from speech information because humans have a built-in ability to do so. The selection of an emotion recognition body (speech database), the identification of numerous variables connected to speech, and the selection of a suitable classification model are the main hurdles for emotion recognition. To identify emotions, an emotion identification system analyzes auditory structural elements of speech. The analysis is based on multiple research papers and includes an in-depth examination of the methodology and data sets. The study discovered that emotion detection is accomplished using four distinct methodologies: physiological signal recognition, facial expression recognition, a variety of speech signals, and text semantics in both objective and subjective databases such as JAFFE, CK+, Berlin emotional database, and SAVEE. In general, these techniques enable the identification of seven basic emotions. To determine the emotion, the audio expression for eight emotions (happy, angry, sad, depressed, bored, anxious, afraid, and apprehensive), all published research maintain an average level of accuracy. The major goal of this survey is to compare and contrast numerous previous survey methodologies, which are backed up by empirical evidence. This study covered signal collection processing, feature extraction, and signal classification, as well as the pros and downsides of each approach. It also goes over a number of strategies that may need to be tweaked at each step of speech emotion recognition.

MSC.41A25; 41A35; 41A3

https://doi.org/10.29304/jqcm.2022.14.1.905

1. Introduction

Speech signal is the quickest and most natural way for humans to communicate. This characteristic has prompted academics to consider voice as a quick and efficient way of human-machine interaction [1]. In contrast to humans,

*Corresponding author: Fatima A. Hameed

Email addresses: *Ms202010636@iips.icci.edu.com*

Communicated by : Dr. Alaa Taima Abd Akadhem machines have a hard difficulty noticing emotions. Consequently, an emotion detection system aims to improve communication between humans and machines by exploiting emotion-related knowledge[2].

Uniqueness and individuality are inherent in every human being. One of the things that makes us uniquely human is our ability to hear. Emotion recognition and categorization on verbal signals can be done using text, speech, facial expressions, and gestures [3]. For the purpose of identifying human emotions in speech, a variety of classifiers have been employed, including the Hidden Markov Model (HMM)[4], the Neural Network (NN)[5], the Maximum Likelihood Bayes Classifier (MLBC), the Gaussian Mixture Model (GMM)[6], the Kernel Deterioration and K-Nearest Neighbors approach (KNN), the Support Vector Machine (SVM)[7] and There are numerous applications in which emotion plays a significant role, including call centers for detecting angry customers [8][9][10][11], entertainment electronics for capturing emotional user input, automatic speech recognition (ASR) for resolving linguistic ambiguities [12][13], and text-to-speech systems for generating emotionally more natural speech [9][10].

The extraction of features or distinct qualities from speech can be used to identify an emotion, and the system will need to be trained on a large number of speech databases to be accurate. The steps in developing an emotion identification system include selecting or implementing an emotional speech corpus, extracting emotion-specific characteristics from those speeches, and lastly, using a classification model to classify emotions.

Section headings should be left justified, bold, with the first letter capitalized and numbered consecutively, starting with the Introduction. Sub-section headings should be in capital and lower-case italic letters, numbered 1.1, 1.2, etc., and left justified, with second and subsequent lines indented. All headings should have a minimum of three text lines after them before a page or column break. Ensure the text area is not blank except for the last page.

2. Literature review of Speech Emotion Recognition Technologies

A lot of studies have gone into using speech statistics to determine emotions in the previous few years. Emotion identification is a difficult problem, especially when it is done through the use of voice signals. There have been several key types of a study presented in this sector, with the primary hurdles being the selection of a speech database, the identification of distinct speech aspects, and the appropriate selection of the classification approach [7].Previous techniques and algorithms are discussed in this section. A variety of methods for distinguishing emotions are investigated and contrasted:

Lee et al. [16] (2011), introduced a computational hierarchy for the detection of emotions was presented. As a result of the subsequent layers of binary classification, Is (), was assigned to one of the corresponding emotion classes. In order to minimize the spread of errors, the various levels of a tree are designed to fulfill the categorization assignment in the simplest manner possible. The AIBO and USC IEMOCAP datasets were utilized to evaluate the categorization system. The absolute result improves accuracy by 72.44% - 89.58% when compared to the baseline SVM. A hierarchical approach to categorizing emotional speech in various datasets has been shown successfully as a result.

Albornoz et al [17] (2011), investigated a novel spectral component for the purpose of classifying groups and determining emotions. In this work, emotions are classified using auditory features and a novel hierarchical classifier. Various classifiers, including HMM, GMM, and MLP, were evaluated against a variety of configurations and input data in order to develop a novel hierarchical technique for emotion categorization. The proposed method is unique in two ways: first, it selects the best performing features, and second, it employs the class-wise classification performance of total features, which is identical to the classifier's performance. When decuple cross-validation is used, the hierarchical approach outperforms the best standard classifier on the Berlin dataset. For instance, the normal HMM approach performed at 68.57 %, whereas the hierarchical model performed at 71.75 %.

Cao et al. [18] (2015), have indicated that in order to solve the binary classification challenge, a ranking SVM algorithm was used to synthesize information on emotion recognition. Each utterer's data is treated as a distinct query in this technique, which then combines all ranker forecasts to apply multi-class prediction. One of the benefits of ranking SVM is that the training and testing operations are conducted in a speaker-independent way, allowing it to obtain speaker-specific data. It also considers the fact that each speaker may communicate a variety of emotions when determining the dominant emotion. Ranking approaches outperform standard SVM in two publicly available datasets of performed emotional speech, Berlin and LDC. Emotional utterances, such as neutral, forceful emotional utterances, were easier to detect with ranking-based SVM than with traditional SVM algorithms. 44.4 % of the unweighted average (UA) was accurate or correct.

Shegokar and P. Sircar [19] (2016), proposed a speech-based emotion detection technique that uses continuous wavelet transformations and prosodic coefficients to pick features for feature selection. A variety of SVMs are employed in the system described here to serve as a classification model. The trial's best recognition rate came in at 60.1%, according to the results.

Basu et al. [20] (2017), recommended using the Mel Frequency Cepstral Coefficients (MFCC) and thirteen acceleration components like features, as well as a Convolution Neural Network (CNN) with Long Short Term Memory (LSTM) as the classification technique, for speech-based emotion recognition. Around 80 % of the time, the answers were correct. This technique can yield better results when fed with a larger database. Using MFCC features and SVM as a classification model, M. S. Likitha et al.[14] Achieved a similar level of accuracy.

Han and Wang [22] (2017), developed a technique for speech-based emotion recognition utilizing SVM and Gaussian Kernel Nonlinear Proximal SVM. The features of prosody and quality of speech are first extracted from the preprocessed voice signal. SVM and Proximal SVM were used as classification models for emotion recognition, with an average recognition rate of 80.75 percent for SVM and 86.75 percent for Proximal SVM. Proximal SVM is three times faster than SVM in recognizing emotions, according to these data. Features that are more efficient are needed to achieve the best results.

Zhang et al. [23] (2018). Have employed CNN, 3D-CNN, and the Deep Belief Network to construct an audio-visual emotion identification system (DBN). Mel-spectrogram of the speech is submitted to CNN. 3D-CNNs are utilized to extract emotional content from video footage. The outputs of audio and visual networks are combined using a deep DBN model in order to merge aural and visual information. eNTERFACE'05's audio path findings were 66.17% for the RML database and 78.08% for the RML database's audio path. The eNTERFACE'05 database's visual path findings for RML were 68% and 54%, respectively. 80.36 And 85.97 %, respectively, were the results of data fusion.

Caihua [24] (2019).SVM-based machine learning was found to be useful for analyzing voice consumer sentiment. He proposed a multimodal speech emotion recognition system based on SVMs. The tests show that the SVM method has evolved significantly by applying this SVM approach to the common database classification problem. Finally, he utilizes a technique for recognizing and interpreting emotional expressions through successful communication.

Pane et al. [25] (2019), developed an approach that combines lateralization of emotion and ensemble learning. Time-domain, frequency-domain, and wavelet characteristics of EEG data were recovered using four alternative channel sequences and combinations. The DEAP datasets were then categorized using a random forest algorithm with a classification accuracy of 75.6%.

Liu et al. [26] (2020), introduced a novel multimodal music emotion grouping method was created using music audio quality and text for music lyrics. In the case of audio, it is recommended to use an LSTM network for classification, as the classification effect is greatly enhanced when compared to other machine learning approaches. Bert is proposed as a means of expressing the emotions expressed in songs in a way that effectively addresses long-term reliance. In terms of multimodal fusion, the lyrics refer to LSFM. The emotion dictionary is used to alter the emotional classification of lyrics. The neural network is constructed utilizing stage fusion in conjunction with linear weighted decision-making to maximize efficiency and precision.

Husam [32] (2021), The goal of this project is to create and test a new feature extraction approach that can extract features to recognize various moods. The newly suggested extracted statistical characteristics were used to build and execute a unimodal speech, real-time, gender, and speaker independent speech emotion recognition (SER) framework. The technique employed in extracting the statistical characteristic utilized numerous degrees of standard deviation (SD) on either side of the mean rather than two SDs on either side of the mean, as all previous studies did. The degrees of deviation on either side of the mean are used to investigate the feature distribution variation around the mean in this study (0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.5 and 4). The Ryerson Audio-Visual Database of Emotional Speech and Song dataset (RAVDESS), which contains eight emotions, the Berlin dataset (Emo-DB), which contains seven emotions, and the Surrey Audio-Visual Expressed Emotion dataset (SAVEE), which contains seven emotions, were employed in this study. The classification accuracy attained in this study was near flawless when compared to state-of-the-art unimodal SER techniques, with 86.1 %, 96.3 %, and 91.7 % for the RAVDESS, Emo-DB, and SAVEE datasets, respectively.

Table (1): Comparison between different approaches of Speech Emotion Recognitions.

Ref.	Database	Emotion Type	Features	Classifier Type	Accuracy
Lee et al., (2011) [16]	AIBO&USC IEMOCAP Data set	Angry, Emphatic Neutral, Positive Rest	Energy, pitch, ZCR, Root mean square, MFF	SVM	72.44%-89.58%
Albornoz et al., (2011) [17]	Berlin Database	Joy, Anger, Fear Disgust, Boredom Neutral and Sadness	MFFC, Prosodic features, and (MLS) mean of log-spectrum	GMM, HMM, MLS& Hierarchical model	Hierarchical model 71.75% HMM 68.57%
Cao et al. (2015) [18].	Berlin &LDC Database	Disgust, Happy Anger, Fear, Sad and Neutral	Prosodic & Spectral features	SVM	44.4%
Shegokar & P. Sircar (2016) [19]	RAVDESS Data set	Angry, Clam, Sad, Neutral, Happy, Fear, Surprised, and Disgust	Wavelet transformations &prosodic coefficients	SVMs	60.1%
Basuetal, (2017) [20]	Berlin (EmoDB)	Angry, Sad, Neutral, Happy, Fear, Boredom, and Disgust	thirteen (MFCC) & thirteen acceleration components	CNN& LSTM	80%
Han and J. Wang (2017) [22]	Speech database in Chinese language.	Angry, Joy, Sadness, and Surprise	------------------	SVM & Proximal SVM	80.75 % with SVM 86.75 % with Proximal SVM
Zhang et al. (2018) [23]	RML database, eNTERFACE05 database, BAUM-1s database	Anger, Disgust, Fear, Joy, Sadness, and Surprise.	Mel-spectrogram	CNN, 3D-CNN, and Deep Belief Network (DBN)	66.17 % &78.08 % for audio path 68.9% and 54.35 % for visual path 80.36 % & 85.97 %for fusion two datasets

Caihua (2019) [24]	Berlin Emotional DB	Happiness, Anger, Fear, Anxiety, Boredom, Disgust, and Normal	Fusion method	SVM	72.52%
Pane et al. (2019) [25]	DEAP datasets	Happy, Relaxed, Angry, and Sad	frequency domain features, Time-domain features, wavelet features of EEG signals	random forest	75.6%
Liu et al. (2020) [26]	777 songs (Music Mood) the data set	Exuberance, Anxious, Contentment, and Depression	Bert model, LSFM	LSTM	5.77% Improvement
Husam. (2021)[32]	RAVDESS, Emo-DB, and SAVEE datasets	Anger ,Neutral ,sad ,happy, disgust , fear,suprise	Zero Crossing (ZC), entropy, energy, the deviation of the ZC, the deviation of the energy, the Haar, the Fourier function, the MATLAB fitness function, the loudness function, the pitch function, the MFCC function according to time and frequency, the Gamma tone Cepstrum Coefficient (GTCC) according to time and frequency, and the harmonic ratio.	Cubic Spline Interpolation , ANN	86.1 %, 96.3 %, and 91.7 %

3. Speech Database

Several speech databases are used in this survey to validate the proposed methods for recognizing speech emotion. In published studies, two types of datasets are used: (1) Public datasets and (2) private datasets; Berlin [16] and AIBO are the most often used databases. Burkhardt et al.[16] The scene was filmed in German. The Department of Technical Acoustics at the Technical University of Berlin was the site of the recording. Five male and five female German actors contributed to the dataset by reading one of the selected statements. The emotions anger, indifference, fear, disgust, happiness, and sadness have all been documented.[17][18][19][20]. Another emotional database called AIBO [21] was created in real time by connecting and playing with fifty-one children using Sony's robot AIBO, which was controlled by a human operator to extract the children's spoken language. Positive, neutral, furious, restful, and emphatic are the five emotions collected in AIBO. [22]. Several more investigations made use of private datasets.

4. Feature Extraction

In the preceding section of this research, we addressed the literature review and comparative evaluation of several approaches and algorithms for voice emotion recognition systems. Emotion-relevant features can only be selected via the feature extraction step in our voice recognition system, which is why it is so important to our study. The speech emotion recognition challenge is complex by the extraction and accurate selection of distinguishing elements from speech emotion data. Previous research has used two types of features: statistical and time-domain features. The set of statistical characteristics enables for separation across feature categories because the suggested statistics are so diverse. With the exception of the first time series difference, none of them has a particularly high weighted relative frequency score. The statistical feature strength of the signal is very strong. This isn't to say that it's a highly regarded feature.in

[22][17][19][14][23]. Although time-domain features aren't generally used in speech emotion recognition, there are a number of techniques for discovering time-series qualities that differ between emotional states; in[18] [24] [20][25][26], they used Time-domain features.

5. Classification Techniques

A classification system is a way of classifying each utterance based on the information retrieved from it. Emotions can be recognized using a variety of clarefssifiers. Different classification techniques are used to generate proper classifiers for modelling emotional states. For Speech emotion recognition systems, many classification methods were used such as: support vector machine (SVM) in [22][18][24][14][20] . Under the conditions when there is a limited of training data. For machine learning algorithms, SVM is a more efficient and faster calculation technique. It's commonly used for pattern recognition and classification problems [27]. Whereas In [17], Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Mean of Log-Spectrum(MLS) were used. In [19], Convolution Neural Network (CNN) with Long Short Term Memory (LSTM) was used. In [23], Convolution Neural Network (CNN), 3D-CNN, and Deep Belief Network(DBN) were used. In [25], a random forest was used. In [26], LSTM was used. In[28], Support Vector Machine (SVM) and Gaussian Kernel Nonlinear Proximal SVM(SVM &Proximal SVM) were used.

6. Performance Analyses

While state-of-the-art SER solutions are solely based on data-driven machine learning techniques, selecting an appropriate speech database is a logical first step in the development of such systems. When choosing a suitable dataset, several factors must be considered, including the degree of naturalness of the emotions, the size of the database, and the quantity of emotions accessible. In Table 1, the most widely used databases of emotional speech are listed[29]. There are many different methods for recognizing feelings through speech and many classification methods Table 2 presents the common classifiers that are used in recognizing emotions by using speech[30].

Table 3. Comparison of databases of emotional speech.

Database	Language	Num. of Subjects	Num. of Utterances	Discrete Labels	Modality
EmoDB[17]	German	5 Female/5 Male	500	A, B, D, F, H, N, S	A
eNTERFACE'05[23]	English	42	5 utt./emotion	A, D, F, H, N, S, Sr	A, V
IEMOCAP[16]	English	5 Female/5 Male	10,039	A, D, E, F, Fr , H, N, s, Sr	A, V, T, MCF
RAVDESS[19]	English	12 Female/12 Male	104	A, D, F, H, N, S, Sr	A, V
SAVEE[15]	English	4 M	480	A, D, F, H, S, Sr , N	A, V

AIBO[16]	German	30 Female/21 Male (children)	18 216	A, B, Em, He, I, J, M, N, O, R, Sr	A

Meaning of acronyms are as follows: Discrete labels: A—anger, B—boredom, C—contempt, D—disgust, E—excitement, Em—emphatic, F—fear, H—happiness, He—helplessness, I—irritation, J—joy, M—motherese, N—neutral, O—other, R—reprimanding, S—sadness, Sr—surprise; Dim. Labels: dimensional labels (arousal, valence, dominance); Modality: A—audio, V—video, T—text, MCF—motion capture of face[31]

Table 2: Different Classifiers in SER

Traditional based classifiers	
Classifier	Description
Hidden Markov Model (HMM)[17]	In this model, the current state depends on previous state. This model uses little data to recognize speech emotions from contextual information. Its strong side is in natural databases.
Singular Vector Machine (SVM)[16]	It achieves better in case of small databases and high dimension features in which these two features are common in SER. It also cares about both testing and training data.
Gaussian Mixture Model (GMM)[17]	It can work well when it is combined with discriminate classifiers like SVM because it can generate and learn the hidden features of speech emotion.
Deep learning based classifiers	
CNN[23]	It has ability to decrease the signal processing, automatic learning of discriminative and global emotional features.
Artificial Neural Networks (ANN)[15]	This classifier achieves good results for nonlinear emotional features. Its latency is very short to predict the features, hence it is effective for applications that are sensitive to time.
LSTM[26]	It has the ability to process the long contextual information and the long variance input utterance features.

7. Conclusions

In this paper, we present a survey of emotion recognition and review and analyze several methods based on speech and emotion recognition systems. The different databases used in this field are presented with their method of construction. Important applications of emotion recognition are also studied: namely, emotion-based emotion recognition. Speech, face-based emotion recognition, and auditory and visual emotion recognition. Much of the research covered in this study has presented different sets of characteristics. We have provided a comprehensive comparison of the various feature extraction approaches available using machine learning techniques. Speech signal processing, feature extraction, feature selection, and classification methods have been done in the future. We will focus on improving the accuracy of the system classification by applying algorithms to create a more complete and user-friendly discrimination system. A classification method should be created to detect verbal emotions in a way that does not threaten the accuracy of the system. . Through more efficient deep learning approaches, such as developing a more powerful emotion recognition model.

References

[1] M. J. Gangeh, P. Fewzee, A. Ghodsi, M. S. Kamel, and F. Karray, "Multiview supervised dictionary learning in speech emotion recognition," IEEE Trans. Audio, Speech Lang. Process., 2014, doi: 10.1109/TASLP.2014.2319157.

[2] I. Chiriacescu, "Automatic Emotion Analysis Based on Speech," Delft University, 2010.

[3] R. D. Shah, A. C. Suthar, and M. E. Student, "Speech Emotion Recognition Based on SVM Using MATLAB", Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ., 2007.

[4] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition based on SVM/HMMs fusion system", 2008, doi: 10.1109/ICALIP.2008.4590144.

[5] R. P. Gadhe, R. R. Deshmukh, and V. B. Waghmare, "KNN based emotion recognition system for isolated Marathi speech," Int. J. Comput. Sci. Eng., Vol. 4, No. 04, Pp. 173–177, 2015.

[6] I. J. Tashev, Z. Q. Wang, and K. Godin, "Speech emotion recognition based on Gaussian Mixture Models and Deep Neural Networks," 2017, doi: 10.1109/ITA.2017.8023477.

[7] Gang, H., Jiandong, L., & Donghua, L. (2004, May). Study of modulation recognition based on HOCs and SVM. In 2004 IEEE 59th Vehicular Technology Conference. VTC 2004-Spring (IEEE Cat. No. 04CH37514) (Vol. 2, pp. 898-902). IEEE.

[8] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, 2011, Vol. 2, Pp. 621–625.

[9] C. S. Ooi, K. P. Seng, L. M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," Expert Syst. Appl., 2014, doi: 10.1016/j.eswa.2014.03.026.

[10] A. Milton, S. Sharmy Roy, and S. Tamil Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," Int. J. Comput. Appl., 2013, DOI: 10.5120/11872-7667.

[11] S. S. Agrawal, N. Prakash, and A. Jain, "Transformation of Emotion based on Acoustic Features of Intonation Patterns for Hindi Speech," IJCSNS Int. J. Comput. Sci. Netw. Secur., 2010.

[12] S. Bahuguna and Y. P. Raiwani, “Study of Speaker’s Emotion Identification for Hindi Speech,” Int. J. Comput. Sci. Eng., Vol. 5, No. 7, Pp. 596, 2013.

[13] B. Panda, D. Padhi, K. Dash, and S. Mohanty, "Use of SVM classifier & MFCC in speech emotion

recognition system," Int. J. Adv. Res. Comput. Sci. Softw. Eng., Vol. 2, No. 3, Pp. 225–230, 2012.

[14] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "IITKGP-SEHSC: Hindi speech corpus for emotion analysis," in 2011 International conference on devices and communications (ICDeCom), 2011, Pp. 1–5.

[15] Wu, C., Huang, C., & Chen, H. (2018). Text-independent speech emotion recognition using frequency adaptive features. Multimedia Tools and Applications, 77(18), 24353-24363.

[16] C. C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Commun., 2011, DOI: 10.1016/j.specom.2011.06.004.

[17] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Comput. Speech Lang., Vol. 25, No. 3, Pp. 556–570, 2011.

[18] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., 2015, DOI: 10.1016/j.csl.2014.01.003.

[19] Shegokar, P., & Sircar, P. (2016, December). Continuous wavelet transform based speech emotion recognition. In 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1-8). IEEE.

[20] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture", 2018, DOI: 10.1109/CESYS.2017.8321292.

[21] Likitha, M. S., Gupta, S. R. R., Hasitha, K., & Raju, A. U. (2017, March). Speech based human emotion recognition using MFCC. In 2017 international conference on wireless communications, signal processing and networking (WiSPNET) (pp. 2257-2260). IEEE.

[22] Han, Z., & Wang, J. (2017, October). Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine. In 2017 Chinese Automation Congress (CAC) (pp. 2513-2516). IEEE.

[23] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition", IEEE Trans. Circuits Syst. Video Technol., 2018, DOI: 10.1109/TCSVT.2017.2719043.

[24] Caihua, C. (2019, July). Research on multi-modal mandarin speech emotion recognition based on SVM. In 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 173-176). IEEE.

[25] E. S. Pane, A. D. Wibawa, and M. H. Purnomo, "Improving the accuracy of EEG emotion recognition by combining valence lateralization and ensemble learning with tuning parameters", Cogn. Process., 2019, DOI: 10.1007/s10339-019-00924-z.

[26] Liu, G., & Tan, Z. (2020, June). Research on Multi-modal Music Emotion Classification Based on Audio and Lyirc. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 1, pp. 2331-2335). IEEE.

[27] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005, September). A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

[28] Bozkurt, E., Erzin, E., Erdem, Ç. E., & Erdem, A. T. (2010, April). Interspeech 2009 emotion recognition challenge evaluation. In 2010 IEEE 18th Signal Processing and Communications Applications Conference (pp. 216-219). IEEE.

[29] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone

applications based on gmm supervectors and support vector machines", in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, Pp. 1605–1608.[1]M. J. Gangeh, P. Fewzee, A. Ghodsi, M. S. Kamel, and F. Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE Trans. Audio, Speech Lang. Process.*, 2014, doi: 10.1109/TASLP.2014.2319157.

[30] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey," in *2021 International Wireless Communications and Mobile Computing (IWCMC)*, 2021, pp. 681–687.

[31] F. Ringeval *et al.*, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th international workshop on audio/visual emotion challenge*, 2015, pp. 3–8.

[32] Abdulmohsin, H. A. (2021). A new proposed statistical feature extraction method in speech emotion recognition. *Computers & Electrical Engineering*, *93*, 107172.