Library Progress International
Vol.44 No. 3, Jul-Dec 2024: P. 11487-11496

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

A Systematic Literature Review On Speech Emotion Recognition Approaches Using Different Methodologies

[1]Dr. Kamatchi Sundravadivelu,[2]Mr. S. Muthukumar,[3]Ms. A. Sirin Vifakga
[4]Aditi Chaudhary,[5]Dr. Shabnam Gulati

1Assistant Professor, Department of Computer Science, Madurai Kamaraj University, Madurai-625 021, Tamil Nadu, Indiasvadivelu2021@gmail.com
[2]Head and Assistant Professor, Department of Computer Science, PVP College, K. Singarakottai, Dindugul, Tamil Nadu, India. kumar.muthu31@gmail.com

3PG Student, Department of Computer Science, Madurai Kamaraj University, Madurai-625 021, Tamil Nadu, India sirinvifakgaa@gmail.com

[4] Mittal School of Business, Lovely Professional University, Phagwara, Punjab, India
aditichaudhary208@gmail.com

5Mittal School of Business, Lovely Professional University, Phagwara, Punjab, India

How to cite this article: Kamatchi Sundravadivelu,S. Muthukumar, A. Sirin Vifakga, Aditi Chaudhary, Shabnam Gulati (2024) A systematic literature review on speech emotion recognition approaches using different methodologies. *Library Progress International,* 44(3), 11487-11496.

ABSTRACT
This paper explores Speech Emotion Recognition (SER), crucial for applications like education, customer service, and healthcare. SER aims to predict emotions conveyed through speech, utilizing various frameworks accurately. The paper evaluates existing methods based on feature extraction and classification, highlighting their strengths, limitations, and overall accuracy. By analyzing these parameters, the paper provides insights into the effectiveness of current SER approaches. This outline is an important asset for scientists and specialists looking to comprehend and further develop feeling acknowledgment frameworks in true settings.

Keywords: SER-SVM-Mel Frequency Cepstral Coefficients (MFCC)-(RNN)-Hidden Markov Model (HMM)-Emotional Database, etc.,

1. Introduction:

Speech is a vital technique for human correspondence, inciting researchers to examine its actual limit with respect to fruitful human-PC participation. Talk Feeling Affirmation (SER) has gained unquestionable quality in various fields like clinical assessment, client backing, and preparing, working with nuanced correspondences. Seeing sentiments from talk presents hardships, primarily in picking ideal components that exactly perceive different up close and personal states. Experts have proposed a couple of frameworks using various capacities, for instance, prosodic (e.g., pitch, energy) and supernatural components (e.g., MFCC, LPCC). Request procedures like Bayesian classifiers, Hidden away Markov Models, and Cerebrum Associations are then used to recognize sentiments like fulfillment, shock, and sharpness. This outline reviews different strategies in feature extraction and portrayal, wanting to overhaul the accuracy and significance of SER systems in veritable circumstances.

2. Related Work:

The paper [1] explores various Speech Emotion Recognition (SER) methods, focusing on feature extraction and classification techniques to improve detection accuracy, with an emphasis on cross-corpus adaptation strategies like local domain adaptation (L-DA) and weighted maximum mean discrepancy to address variability in emotional speech data. Supplementing this, [2] examines both customary AI strategies, like Innocent Bayes and K-Closest Neighbor, and high level profound learning procedures, including Vanilla Profound Brain Organizations, Convolutional Brain Organizations (CNNs), and Long-Transient Memory organizations (LSTMs), investigating the effect of Mel-Recurrence Cepstral Coefficients (MFCCs) on model execution. Moreover, [4] assesses nineteen self-regulated discourse portrayals close by traditional acoustic highlights, showing the viability of models like WavLM Enormous and UniSpeech-SAT Huge in SER assignments while addressing difficulties connected with little and unequal datasets. Further, [7] audits ongoing headways in highlight extraction, featuring the job of acoustic elements, for example, MFCC, Zero-Crossing Rate (ZCR), Teager Energy Administrator (TEO), and Music to-Clamor Proportion (HNR), alongside cutting edge dimensionality decrease strategies like auto-encoders to further develop characterization precision. Section [8] analyzes different SER frameworks, showing that joining MFCC and balance unearthly elements with classifiers like Repetitive Brain Organizations (RNNs), Backing Vector Machines (SVMs), and Various Straight Relapse (MLR) can accomplish high precision, with RNNs arriving at up to 94% on the Spanish information base. Besides, [11] shows that consolidating prosodic and ghostly elements like energy, pitch, Direct Prescient Cepstral Coefficients (LPCC), and LPCMCC can essentially improve characterization exactness, accomplishing 82.5% with SVM. [12] explores emotion recognition in Arabic speech, noting that SVM outperforms other methods with 77.14% accuracy. Expanding on these discoveries, [15] accentuates the designing parts of feeling acknowledgment, zeroing in on worldly component coordination strategies like momentary measurements, ghostly minutes, and autoregressive models for ordering discourse signals and featuring the significance of multiresolution examination. Finally, [18] presents a cross-corpus SER framework leveraging L-DA and weighted maximum mean discrepancy to address challenges in emotional speech data, improving generalization by evaluating category-grained discrepancies between source and target domains, thereby enhancing performance compared to global adaptive and non-adaptive methods.

3. Feature Extraction:

Highlight extraction is the fundamental phase of discourse feeling acknowledgment. Elements can be utilized to perceive the contrast between a few close to home states. The discourse highlights are parted into two gatherings incorporates prosodic and ghastly elements. Prosodic elements are energy, pitch and phantom highlights are MFCC, LPCC, and MEDC.

3.1 Prosodic Feature:

Prosodic features are essential in conveying emotions through speech. They include **pitch**, which indicates how high or low the voice sounds; **energy**, reflecting loudness and intensity; **duration**, which measures the length of speech segments; and **speech rate**, or how quickly someone speaks. Variations in these features can reveal emotional states, such as excitement or sadness, beyond the literal meaning of words. For instance, a rising pitch might suggest a question, while increased energy could signal enthusiasm or anger. Analyzing these features helps in understanding the speaker's emotional context more deeply.

3.2 Spectral Feature:

Spectral features analyse the frequency content of speech, capturing its texture and timbre. MFCC (Mel-Recurrence Cepstral Coefficients) address the power range of discourse, reflecting phonetic and speaker-explicit qualities. **LPCC (Linear Predictive Coding Coefficients)** model the vocal tract's shape, aiding in distinguishing phonetic sounds. **MEDC (Mel-Energy Cepstral Coefficients)** extend MFCC by incorporating energy variations across frequency bands. These features are vital for distinguishing emotional states by examining the detailed frequency patterns in speech, enhancing the recognition of different emotional expressions beyond simple prosody.

3.2.1 MFCC:

MFCCs are crucial in speech recognition for their human-perception-based representation of voice signals, involving mapping Fourier-transformed spectra onto the Mel-frequency scale and computing DCT of Mel log energies [8]. Typically, the first 12 DCT coefficients are used in classification, computed from 60-dimensional feature vectors of speech frames sampled at 16 kHz. MFCC values are extracted from audio signals, comprising 39 features including 12 MFCC parameters, log-energy, 13 delta, and 13 acceleration coefficients. Frames are 25ms with a 10ms rate using a Hamming function, adjusted based on wave file lengths. Extra prosodic elements like F0 recurrence, voicing likelihood, and clamor forms, adding up to 35 highlights, are additionally extricated utilizing the Open Grin tool compartment [14].

3.2.2 MEDC:

MEDC (Mel Energy Range Dynamic Coefficients) is a component extraction strategy likened to MFCC, used in discourse handling. It includes preprocessing, outlining, windowing, FFT, Mel channel bank, and recurrence wrapping. Dissimilar to MFCC, MEDC works out the logarithmic mean of energies post Mel Channel bank as opposed to applying a logarithm straightforwardly. Additionally, MEDC computes the first and second differences of these logarithmic mean energies. This differential computation captures temporal changes in energy distribution across frequency bands, enhancing sensitivity to dynamic features in speech signals for tasks like emotion recognition and speaker verification.

3.2.3 LPC:

Linear Predictive Coding (LPC) is a strategy utilized in discourse handling to display the otherworldly envelope of discourse signals. It gauges coefficients for a direct expectation channel that limits expectation blunder by approximating discourse as a straight blend of past examples. LPC utilizes the source-channel model, where the vocal plot move capability is addressed as an all-shaft channel [17]. Widely applied in speech analysis, synthesis, and compression, LPC effectively captures vocal tract characteristics, enhancing both recognition and coding efficiency across various speech-related applications.

3.2.4 LPCC:

Linear Prediction Cepstral Coefficients (LPCCs), got from Linear Prediction Coefficients (LPCs) [13], provide a reliable method for feature extraction in speech emotion recognition [13]. LPCCs capture individual channel characteristics essential for recognizing varying emotional tones [13]. The computation of LPCCs is efficient, effectively describing vowels but less capable with consonants and noise resilience [12]. The extraction cycle incorporates pre-complement, framing, windowing, Quick Fourier Change (FFT), Mel scale channel bank, log energy computation, and Discrete Cosine Change (DCT) [13]. Typically, 12-order LPCC coefficients are extracted per speech frame, resulting in 48-dimensional feature vectors used for statistical analysis [11]. LPCCs are widely utilized in emotion recognition systems for their effectiveness in distinguishing emotional states in speech [13].

3.2.5 ZCR:

Zero Crossing Rate (ZCR) ascertains how frequently a sound sign changes its sign inside short edges. By dividing the signal into frames, ZCR quantifies the rate of sign changes normalized by frame length. This feature is crucial in speech and audio processing, providing insights into signal dynamics and temporal characteristics. In Speech Emotion Recognition (SER), ZCR helps identify emotional cues based on signal variability. Its application spans from basic feature extraction to advanced machine learning models, enhancing understanding and classification of emotional states in speech.

3.3 Advanced Techniques:

Advanced techniques in Speech Emotion Recognition (SER) enhance the accuracy and effectiveness of detecting emotions from speech. These techniques include methods for reducing background noise to improve speech clarity, approaches for adapting models to work effectively across different datasets, and innovative learning methods that extract meaningful features from speech without needing extensive labeled data. Additionally, sophisticated methods for selecting and combining the most relevant features help refine emotion detection. Together, these techniques contribute to more precise and reliable recognition of emotional states in various speech applications.

3.3.1 Wiener filters:

The Wiener filter is a sign handling procedure utilized for sound decrease. It works in the recurrence area by limiting the mean square mistake between the first sign and the assessed signal polluted by commotion. It accomplishes this by weakening commotion parts in view of their phantom qualities, making it compelling in upgrading signals debased by added substance clamor.

3.3.2 L-DA:

Local Domain Adaptation (L-DA) is a technique intended to upgrade cross-corpus discourse feeling acknowledgment (SER). It centers around adjusting the source and target areas at the feeling class level instead of universally. By utilizing a **weighted Maximum Mean Discrepancy (MMD)**, L-DA measures the distance between domains with a focus on specific emotion categories, ensuring more precise and relevant adaptation. This targeted approach improves the model's ability to generalize across different corpora, resulting in significantly better performance in cross-corpus SER compared to traditional global adaptation methods [18].

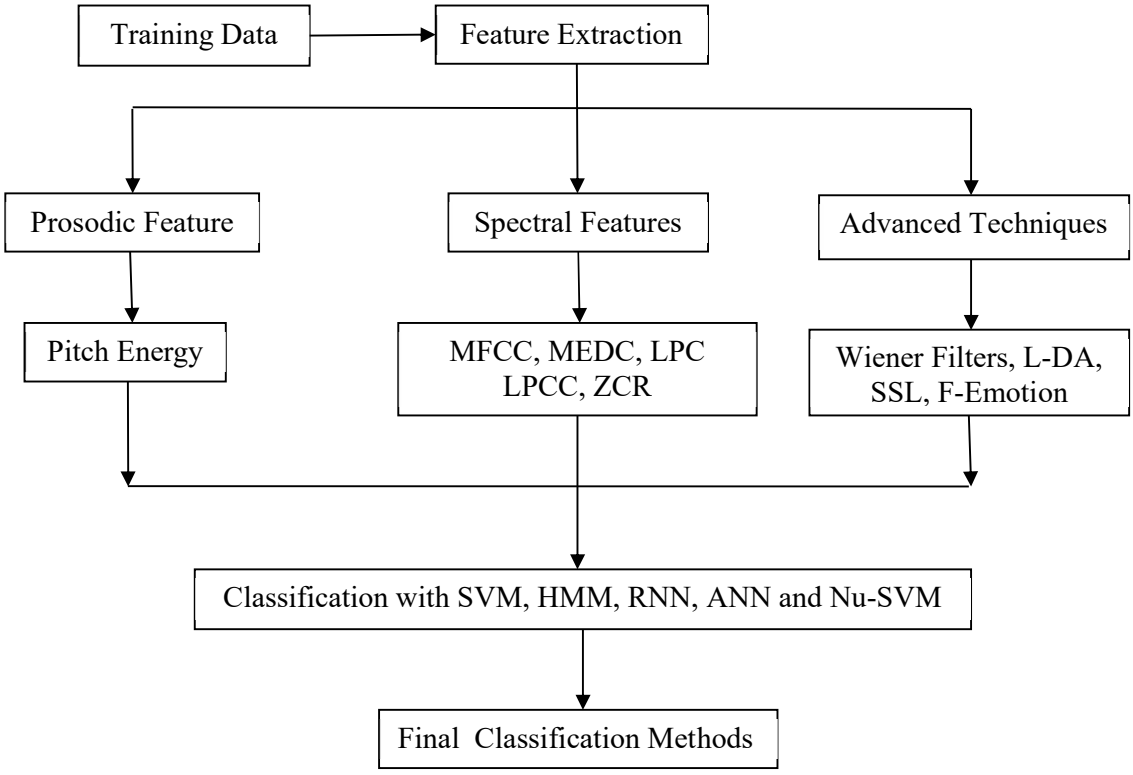
3.3.3 SSL:

Evaluating self-supervised learning (SSL) algorithms such as WavLM Large, UniSpeech-SAT Large, and HuBERT Large for speech emotion recognition (SER) reveals their effectiveness compared to traditional acoustic features. A consistent classifier is applied across five SER datasets to maintain fairness. The evaluation includes calculating effect sizes to measure performance differences among representations and using confusion matrices to analyze predictions across different emotion categories and dataset conditions. This approach benchmarks SSL-based representations against classical methods, providing insights into their strengths and limitations in SER tasks [4].

3.3.4 F-Emotion:

The **F-Emotion algorithm** improves Speech Emotion Recognition (SER) by selecting optimal features through clustering analysis. It first extracts features from speech data, calculates the F-Emotion value to evaluate their effectiveness, and identifies the best feature combinations. A parallel deep learning model is then trained on these features to recognize different emotions individually. The outputs from these models are combined using a decision fusion mechanism with weighted results to enhance accuracy. Tested on RAVDESS and EMO-DB datasets, this approach achieved recognition accuracies of 82.3% and 88.8%, respectively, demonstrating its effectiveness in emotion recognition [9].

Figure 1: Architecture of Speech Emotion Recognition



4. Classification:

4.1 SVM:

Support Vector Machines (SVM) use bit capabilities to change input highlight vectors into higher-layered spaces for ideal class division. They find an ideal hyperplane by expanding separation utilizing support vectors. In discourse feeling acknowledgment, SVMs are adjusted for multi-class arrangement utilizing Choice Coordinated Non-cyclic Diagrams (DDAG), permitting characterization among N classes in light of pairwise qualifications. Normal portion capabilities incorporate direct, polynomial, and Outspread Premise Capability (RBF), empowering viable order in both straight and non-direct component spaces [17].

4.2 HMM:

Hidden Markov Models (HMMs) are statistical tools for systems with hidden states and observable outputs, used to analyze speech for emotion recognition [13]. In this specific situation, Well model feelings as covered up states and detectable elements, for example, pitch and tone got from discourse. They learn transition probabilities between emotions and the likelihood of observing features given an emotion, inferring the most probable emotional state over time. To further develop feeling acknowledgment, calculations incorporate element values utilizing techniques like momentary measurements, ghostly minutes, autoregressive models, and wavelet disintegration[13]. These techniques fuse feature sets at both feature and log-likelihood levels to enhance the accuracy of identifying emotional states.

4.3 KNN:

The k-Nearest Neighbor (KNN) calculation is a basic, non-parametric, and natural AI technique utilized for

grouping and relapse undertakings. It works by distinguishing the 'k' nearest information focuses to a given test and making expectations in light of the larger part class (for characterization) or the typical incentive (for relapse) of these neighbors. KNN does not require a training phase and relies solely on the distance metric (like Euclidean distance) to measure similarity between data points[12]. Its simplicity and effectiveness make it popular for various applications, including pattern recognition and emotion classification.

4.4 WMMD:

Weighted Maximum Mean Discrepancy (WMMD) In light of Feeling Class is utilized to upgrade cross-corpus discourse feeling acknowledgment. This technique tends to disparities among source and target spaces by zeroing in on feeling explicit arrangements. In contrast to conventional greatest mean error (MMD), which adjusts generally circulations, WMMD allocates loads to various feeling classifications. This approach thinks about the class earlier dissemination in the two spaces, prompting more exact arrangement and further developed move effectiveness, subsequently helping acknowledgment precision [18].

4.5 GMM:

The Gaussian Mixture Model (GMM) is a probabilistic model used to address a circulation of pieces of information. It expects that information is created from a combination of a few Gaussian conveyances, each portrayed by its mean, covariance, and weight. GMMs are prepared utilizing the Assumption Amplification (EM) calculation, which repeats between allotting information focuses to Gaussian parts in view of likelihood and refreshing the boundaries of the parts to augment the probability of the noticed information. GMMs are successful for displaying complex appropriations and are generally utilized in feeling acknowledgment frameworks to characterize information by contrasting new examples against the learned Gaussian disseminations [13].

4.6 RNN:

Recurrent Neural Networks (RNNs) are effective for learning time series data and performing classification tasks, particularly for emotion recognition by modeling temporal dependencies in speech data [8][13]. RNNs comprise of neurons with associations that permit data to continue across time steps, catching worldly elements fundamental for grasping feelings [13]. They share similar boundaries (U, V, and W) across all time advances, and the secret state at each time step (st) is figured utilizing these boundaries and the info (xt) from the past step [8]. To address the evaporating angle issue in longer arrangements, long short-term memory (LSTM) RNNs was presented, using memory cells to hold data and catch long-range conditions [8]. RNNs are prepared utilizing backpropagation through time (BPTT), which updates loads in light of mistake slopes over the succession [13].

4.7 ANN:

The Artificial Neural Network (ANN) utilized has seven layers: an information layer (number of MFCCs), trailed by thick layers of 255, 128, and 64 neurons individually, each with a 0.3 dropout rate. The actuation capability utilized in these layers is ReLU. After the third thick layer, there is a straighten layer, trailed by a fourth thick layer of 64 neurons [17]. The result layer, comprising of 8 classes, utilizes the Softmax actuation capability. The model was prepared for 2000 ages with a clump size of 16, utilizing RMSProp enhancer with a learning pace of 0.001, ρ (rho) of 0.9, ϵ (epsilon) of 1e-07, and no rot rate. Execution measurements incorporate preparation and testing exactness, as well as preparing and testing misfortune.

4.8 MLR:

Multiple Linear Regression (MLR) predicts a result variable in light of numerous information factors by fitting a direct condition. It estimates coefficients for each input, modeling relationships between variables to make predictions, assuming a linear relationship between inputs and output.In speech emotion recognition using Multiple Linear Regression (MLR), acoustic features (like pitch, intensity) are input variables. MLR fits a linear equation to predict emotional states (e.g., happy, sad) based on these features. It estimates coefficients for each feature to model relationships between acoustic cues and emotions, providing a quantitative approach to classify

emotions from speech data.

4.9 RBF:

The Radial Basis Function (RBF) calculation is a bit strategy utilized in AI for characterization and relapse. It transforms data into a higher-dimensional space using radial basis functions, enabling linear separation in this new space for complex patterns**.** In speech emotion recognition, the Radial Basis Function (RBF) algorithm processes audio features like pitch and tone. It maps these features into a higher-dimensional space using radial basis functions, allowing the algorithm to classify emotions (e.g., happy, sad) more effectively by finding linear separations in this transformed space.

4.10 Nu-SVM:

Support Vector Machines (SVM) is highly effective for classification tasks across various domains, including Speech Emotion Recognition (SER). In 2000, Schölkopf et al. introduced the Nu-SVM, which uses the ν parameter instead of the standard SVM's C parameter to regulate the number of support vectors. This parameter, defined as a fractional value between 0 and 1, balances the number of support vectors and the margin, ensuring efficient handling of both linear and non-linear input data [1].When applied to speech emotion recognition, the Nu-SVM algorithm leverages the ν parameter to manage learning errors and support vectors effectively, thereby enhancing the classification accuracy of emotional states in speech data.

5. Comparison:

Topic	Algorithm	Key and Contextual Details	Dataset	Accuracy
Modeling Speech Emotion Recognition via ImageBind Representations	Nu-support vector machine(Nu-svm) Radial basis function(RBF)kernel	Introduces ImageBind representations and Nu-SVM with RBF kernel for SER.	IEMOCAP	80.58%
Speech emotion recognition using support vector machine	Support Vector Machines (SVM) with a linear kernel.	Demonstrates effectiveness of SVM in handling speech emotion features, explores different kernel configurations for optimal performance.	Berlin Japan Thai emotion databases.	89.80% 93.57% 98.00%
Automatic Speech Emotion Recognition Using Machine Learning	mulitivariate linear regression(MLR) support vector machine(SVM) Recurrent neural network(RNN)	Explores different machine learning approaches and their applications in automatic SER.	Berlin spanish emotion databases	83% 94%
Speech Emotion Recognition with Deep Learning	Gaussian Mixture Models (GMM) Hidden Markov Models (HMM)	Focuses on deep learning models specifically designed for speech emotion recognition tasks.	German and English Databases	86%
Multimodal Speech emotion recognition using Audio and text	Recurrent neural networks (RNNs)	A method combines audio and text for better speech emotion classification than previous approaches.	IEMOCAP	68.8% to 71.8%

6. Discussion:

The study of Speech Emotion Recognition (SER) reveals the strengths and limitations of various algorithms employed in this field. Algorithms such as Nu-SVM with RBF kernel and traditional SVM with linear kernels have shown impressive performance, achieving high accuracy across diverse datasets like IEMOCAP and Berlin. These models are successful in dealing with various element spaces and can oversee both straight and non-direct connections in feeling arrangement. Recurrent Neural Networks (RNNs) and their high level variations, as Long Short-Term Memory (LSTM) organizations, succeed in demonstrating worldly conditions, which is urgent for catching the elements of profound articulation after some time. However, issues such as data variability, language

differences, and cultural nuances present ongoing challenges. Techniques like Weighted Maximum Mean Discrepancy (WMMD) improve cross-corpus recognition by addressing domain discrepancies and enhancing feature alignment. Future research should focus on integrating multimodal data—combining audio, text, and visual inputs—to improve SER accuracy and robustness. Expanding datasets to include a wider range of emotional states and cultural contexts will be essential for creating more generalized and effective SER systems, paving the way for practical applications in real-time emotion detection and human-computer interaction.

7. Conclusion:

Research in Speech Emotion Recognition (SER) has explored a variety of algorithms including Nu-SVM with RBF kernel, SVM with linear kernel, MLR, SVM, RNN, GMM, and HMM, achieving accuracies ranging from 68.8% to 98.00% on datasets like IEMOCAP, Berlin, Japanese, Thai, German, and English databases. Deep learning models such as RNNs, GMM, and HMM show promise in capturing intricate speech patterns for emotion classification. Future research aims to advance SER by refining feature representation with additional modalities such as text and physiological signals. Improving model performance through advanced deep learning architectures and expanding datasets to encompass diverse emotional expressions and cultural contexts are crucial for better generalization. Real-time SER applications are promising for integration into interactive systems, while exploring multimodal fusion techniques—combining audio, text, and visual data—aims to further enhance emotion recognition capabilities. These advancements seek to propel SER systems for practical applications requiring nuanced emotion understanding and response in human-computer interaction and related domains.

8. Reference:

[1] Adil Chakhtouna,Sara Sekkate, Abdellah Adi," Modeling Speech Emotion Recognition via ImageBind Representations", International Symposium on Green Technologies and Applications (ISGTA'2023) ,https://doi.org/10.1016/j.procs.2024.05.050

[2] Andry Chowandaa,Irene Anindaputri Iswantob, Esther Widhi Andangsarib,"Exploring deep learning algorithm to model emotions recognition from speech",Procedia Computer Science, 2023, p. 706-713, https://doi.org/10.1016/j.procs.2022.12.187,

[3] Ashish B. Ingale, D. S. Chaudhari,"Speech Emotion Recognition",International Journal of Soft Computing and Engineering (IJSCE)",ISSN: 2231-2307, Volume-2 Issue-1, March 2012

[4]Bagus Tris Atmaja ,Akira Saso,"Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition",IEEE Access,DOI:10.1109/ACCESS.2022.3225198.

[5] Bjorn Schuller, Gerhard Rigoll, and Manfred Lang,"Hidden markov model-based speech emotion recognition",2003 IEEE International Conference on Acoustics, Speech, & Signal Processing, April 6-10, 2003,

[6] Chen Caihua,"Reserch on Multi-modal Mandarin Speech Emotion Recognition Based on SVM",2019 IEEE International Conference on Power, Intelligent Computing and Systems(ICPICS),DOI: 10.1109/ICPICS47731.2019.8942545

[7] Hadhami Aouani,Yassine Ben Ayed,"Speech Emotion Recognition with Deep Learning",24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems,https://doi.org/10.1016/j.procs.2020.08.027

[8] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, MohamedAli Mahjoub, Catherine Cléder,"Automatic Speech Emotion Recognition Using Machine Learning",4 Jun 2022,HAL open science,DOI: 10.5772/intechopen.84856

[9]Li-Min Zhang, Giap Weng Ng ,Yu-Beng Leau, Hao Yan ,"A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering",IEEE Transactions on Affective Computing,DOI: 10.1109/ACCESS.2023.3294274

[10] Misaki Sakurai, Tetsuo Kosaka, "Emotion Recognition COmbining Acoustic and Linguistic Features Based

on Speech Recognition Results", 2021 IEEE 10th Global Conference on Consumer Electronics(GCCE), DOI: 10.1109/GCCE53005.2021.9621810

[11] Peipei Shen,Zhou Changjun,Xiong Chen,"Automatic Speech Emotion Recognition Using Support Vector Machine" ,Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology,DOI: 10.1109/EMEIT.2011.6023178.

[12] Reem Hamed Aljuhani, Areej Alshutayri , and shahd alahdal,"Arabic Speech Emotion Recognition From Saudi Dialect Corpus",IEEE Access, vol. 9, pp. 127081-127085,DOI: 10.1109/ACCESS.2021.3110992 .

[13] Saikat Basu∗, Jaybrata Chakraborty, Arnab Bag†and Md. Aftabuddin,"A Review on Emotion Recognition using Speech",2017 International Conference on Inventive Communication and Computational Technologies (ICICCT),DOI: 10.1109/ICICCT.2017.7975169

[14] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung,"Multimodal Speech emotion recognition using Audio and text",2018 IEEE Spoken Language Technology Workshop (SLT),DOI: 10.1109/SLT.2018.8639583

[15] Stavros Ntalampiras and Nikos Fakotakis,"Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition" ,IEEE Transactions on Affective Computing (Volume: 3, Issue: 1, Jan.-March 2012),DOI: 10.1109/T-AFFC.2011.31.

[16] Taiba Majid Wani , Teddy Surya Gunawan ,Syed Asif Ahmad Qadri , Mira Kartiwi ,and Eliathamby Ambikairajah , "A Comprehensive Review of Speech Emotion Recognition Systems", DOI: 10.1109/ACCESS.2021.3068045.

[17] Thapanee Seehapoch,Sartra Wongthanavasu,"Speech emotion recognition using support vector machine",2013 5th International Conference on Knowledge and Smart Technology (KST),DOI: 10.1109/KST.2013.6512793

[18]Zhao Huijuan, Ye Ning, Wang Ruchuan,"Improved Cross-Corpus Speech Emotion Recognition Using Deep Local Domain Adaptation"*Chinese Journal of Electronics*, Vol. 32, No. 3, May 2023,DOI**:** 10.23919/cje.2021.00.196