Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.mosuljournals.com

The Intelligent Recognition of Speech Emotions: Survey Study

Ali Abdulwahhab Yehya[1],[*] Fawziya M. Ramo [2]

[1,2]Department of Computer Sciences, College of Computer Sciences and Mathematics, University of Mosul, Mosul, IRAQ*
Email: ali_alsaffar@.uomosul.edu.iq1, fawziyaramo@uomosul.edu.iq[2]

Article information

Article history:
Received:28/2/2023
Accepted :15/5/2023
Available online:

Abstract

Speech emotion recognition (SER) is a challenging task in the field of artificial intelligence and machine learning. Over the years, researchers have proposed various approaches to recognize emotions from speech signals. This article will analyze and discuss some of the previous works on machine learning and deep learning in SER. This survey study focuses on the importance of the human voice in determining emotional and psychological states. Various methods were used to successfully classify emotions such as anger, sadness, happiness, fear, disgust, neutral, and surprise. Reviewed in this study was conducted in sequential stages including pre-processing treatment, feature selection, classification, and evaluation of results. different data sets were also reviewed for international languages such as English, Hindi, German, Urdu, Tamil, French, and Arabic. The study primarily focused on artificial intelligence and machine learning algorithms due to their flexibility and ease of understanding with distinct results.

Keywords:
Machine learning, Deep learning, SER, MFCC, Feature Extraction, CNN.

Correspondence:
Author: Ali Abdulwahhab Yehya
Email: ali_alsaffar@.uomosul.edu.iq

I. INTRODUCTION

Speech emotion identification detects human emotions from speech signals. In recent years, this technique has gained recognition as a key link in the human-computer interaction system. Speech emotion recognition is used in medical diagnostics[1], sleepy driving detection [2], and student emotional state monitoring during e-learning[3].Emotion identification from audio data is becoming an increasingly popular area of study in the field of computer science. Emotion detection systems have been modeled using voice data from a number of languages[4].There is active research into the possibility of using deep learning to decipher emotional states from audio recordings. Here, convolutional neural networks (CNNs) could as well be the norm[5].Our emotional state is a complex psychological condition that controls every facet of our lives. Better communication may be achieved by increasing the focus on exploring and understanding people's emotions and reactions [6]. Two different ways of depicting emotions exist: the discrete emotional approach, in which emotions are categorized into discrete labels such as happiness, boredom, anger, surprise, etc., and the dimensional emotional approach, in which emotions are depicted along a spectrum of dimensions such as arousal (how strong the feeling is), valence (whether the feeling is positive or negative), and power (describe the strength or weakness of the person)[7].Feature extraction and feature classification are the two main aspects of speech emotion recognition (SER). In the field of Feature extraction, researchers have derived several features such as source-based excitation features, prosodic features, vocal traction factors, and other hybrid features[8]. The second process is feature classification, which may be done with linear or nonlinear classifiers. Bayesian networks (BN), the maximum likelihood principle (MLP), and support vector machines (SVM) are the most popular linear classifiers used for emotion identification. The voice signal is typically viewed as dynamic. Therefore, non-linear classifiers are thought to be useful for SER . For SER, various non-linear classifiers are offered, such as the Gaussian mixture model (GMM) and the hidden Markov model (HMM). These are commonly used to classify data at the most fundamental feature level. Emotional expression in speech

may be effectively recognized using energy-based characteristics like linear predictor coefficients (LPC), mel energy-spectrum dynamic coefficients (MEDC), mel-frequency cepstrum coefficients (MFCC), and perceptual linear predictor coefficients (PLP). Emotion recognition also makes use of other classifiers, such as K-nearest neighbors (KNN), principal component analysis (PCA), and decision trees.[9]. As an illustration of where and how these studies are applied, consider the following frameworks [10] :

1) realm of education, a course system for distant learning may identify when its users are bored and adjust the presentation mode or difficulty level, as well as provide psychological rewards or concessions.
2)driver's ability to function safely behind the wheel and his or her emotional condition are often related inside. As a result, these technologies may be employed to enhance the driving experience and boost efficiency.
3) Detecting strong emotions like dread and anxiety, they may be employed as security systems in public places.
4) To better serve customers, contact centers may incorporate automated emotion recognition technology with interactive voice response systems to enhance communication.
5)In terms of health, autistic persons may benefit from using portable gadgets to better comprehend their own emotions and, in turn, modify their social behavior.

2. Dataset

Recognition systems are properly trained using a strong database, otherwise their performance and reliability will suffer. Figure(1) represent component of SER system.

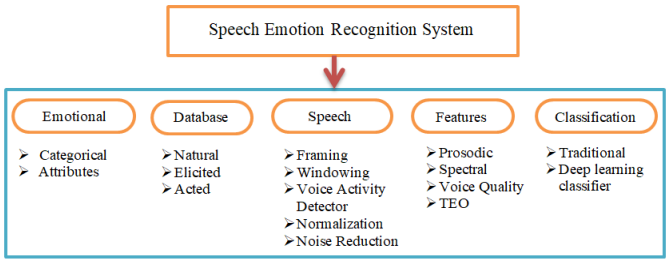

Fig. 1: component of SER system

It is thus crucial to have appropriate words in the database to train the emotion detection system and then assess its performance. Databases can be divided into three types: those containing acted emotions, those containing natural, spontaneous emotions, and those containing induced emotions[9] .

2.1 Type of Speech in Variance Dataset

Identifying whether or not the speech is Acted or Natural is one of the obstacles to developing a dataset of emotional speech. Having an actor perform each sample makes it easy to generate a balanced dataset with all the desired emotions, but there is also a chance that the feeling won't be genuine or won't be accurately represented. It is uncommon to be able to successfully replicate emotions via speech, and often the actor must think back on earlier experiences in order to feel the appropriate mood [11].

2.1.1 SPONTANEOUS(Natural) SPEECH

The database that is most often referenced is spontaneous speech since it displays the most genuine and natural expressions of emotion. In this scenario, covert techniques are used to capture the speaker's precise emotional states. This gives the subject the freedom to reply normally while being unaware that his responses are being observed.[12] However, the development of the technology that is employed for the data gathering of emotions has never been a simple endeavor. Samples of them are gathered using a variety of methods. These may be obtained via many mediums, such as television talk programs, interviews, and other venues of a similar kind[11].

2.1.2 ACTED SPEECH

The creation of the database for acted speech does not the database for natural speech, which makes it simpler to manage. According to Cao et al. performed speech is simply adapting to the sort of emotion .number of performed speech compilations may also include recordings with experienced actors and performers [13].

2.1.3 ELICITED SPEECH

Elicited speech is the method used in eloquent discourse to evoke a desired feeling [14]. It is possible to manipulate a subject's emotional response by placing them in a certain scenario. A recording is then made of the speaker's remarks. However, the introduced feelings tend to be somewhat subdued. The induction approach, on the other hand, allows for some manipulation of the stimuli. Emotion recognition makes use of several different datasets. Table 1 provides a summary of some of the most notable datasets [11].

Table 1. Examples of speech corpora Types

Name	Lang.	Emotion Type	Emotions	Corpus Type
RAVDEES [15]	English	Discrete	neutral, calm, happy, sad, angry, fearful, surprised, and disgusted	Acted
CREMA-D [16]	English	Discrete	neutral, happiness, anger ,disgust, fear, and sadness.	Acted
SAVEE [17]	English	Discrete	anger, disgust, fear, happiness, sadness, surprise, and neutrality	Acted
TESS [17]	English	Discrete	anger, disgust,fear, happiness, pleasant surprise, sadness, and neutral	Acted
Serbian database of acted emotions [18]	Serbian	Discrete	Neutral, anger, happiness, sadness, fear	Acted
IEMOC [19]	English	Discrete and continuous	Anger, happiness, sadness, neutrality, valence, activation and dominance	Induced
SmartKom [20]	German	Discrete and continuous	joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise, neutral, unidentifiable	Induced
TUM AVIC [21]	English	continuous	Disinterest, indifference, neutrality, interest, curiosity	Natural
CEMO [22]	French	discrete	Fear, anger, sadness, neutral,relief	Natural
RECOLA [23]	French	continuous	Social behaviors (agreement,dominance, engagement, performance, rapport)	Acted
KSUEmotions[24]	Arabic	discrete	Neutral, happiness, sadness, surprise, anger	Acted
ANAD [4]	Arabic	discrete	Happy, angry, surprised	Acted

3.ML and DL Techniques in SER

Compared to more conventional approaches, Deep Learning methods for SER have several advantages, such as the ability to detect complex structure and features without the need for manual feature extraction and tuning, a tendency toward extraction of low-level features from the given raw data, and the ability to deal with un-labeled data[25]. Study on speech emotion recognition aims to increase the recognition rate and precision. The use of feature set helps to improve the recognition rate of the framework. Table 2 provides details and comparison of different classifiers machine learning and deep learning techniques in SER .

Table 2. Comparison of ML and DL in SER

REF.	Dataset	year	Features extraction	ML or DL Classifier	Acc (%)
[26]	Berlin Database	2011	MFCC, Prosodic features	GMM, HMM, MLS&	68.57%
				Hierarchical model	71.75%
[27]	Berlin Database	2017	thirteen (MFCC) & thirteen acceleration components	CNN& LSTM	80%
[28]	Berlin Database	2019	Fusion method	SVM	72.52%
[29]	Berlin Database	2015	Fourier Parameters, MFCC SVM	SVM	88.88%
[30]	RAVDESS	2016	Wavelet transformations &prosodic coefficients	SVM	60.1%
[31]	RAVDESS	2021	MFCC	DCNN and a BLSTM	82.7%,
[32]	RAVDESS	2019	MFCCs, spectral, and centroids	Bagged ensemble of SVM	75.69%
[33]	Urdu	2021	MFCC	SVM	95.25%
[34]	Urdu	2021	pitch and MFCC	Randomforest	78.75%
[35]	Tamel Language	2021	MFCC, MFCC delta	LSTM and BiLSTM	84%
[36]	Hindi emotional speech	2016	jitter and shimmer features	ANN model	83.3%
[37]	TESS	2019	MFCC	SVM	96%
[38]	Arabic KSU Emotions	2019	MFCC	CNN-BLSTM-DNN	87.2 %

3-1 Analyzing and discussing previous works

Speech emotion recognition (SER) is a challenging task that involves identifying the emotional state of a speaker based on their speech. Over the years, researchers have developed various algorithms and feature extraction techniques to improve the accuracy of SER systems. In this article, we will discuss the results obtained from SER experiments, including accuracy, algorithms, feature types, and the best methods used.

3-1-1 The accuracy : The accuracy of SER systems varies depending on several factors such as dataset size, feature extraction techniques, and classification algorithms. In general, the accuracy of SER systems ranges from 60% to 90%. However, some studies have reported higher accuracies of up to 97%.

3-1-2 Algorithms: Several classification algorithms have been used in SER systems such as support vector machines (SVM), k-nearest neighbors (KNN), decision trees (DT), and artificial neural networks (ANN). SVM is one of the most commonly used algorithms in SER due to its high accuracy and robustness. KNN is also popular due to its simplicity and ease of implementation [32].

3-1-3 Features Types: Feature extraction is a crucial step in SER that involves extracting relevant information from speech signals. Various feature types have been used in SER such as prosodic features, spectral features, and cepstral features. Prosodic features include pitch, duration, and energy variations in speech signals. Spectral features include frequency-based information such as formants and harmonics. Cepstral features are derived from spectral information using techniques such as discrete cosine transform (DCT) or mel_frequency cepstral coefficients (MFCC)[39] .

3-1-4 Best Methods Used: The best method for SER depends on several factors such as dataset size, feature type, and classification algorithm. However, some studies have reported high accuracies using specific methods. For example, a study by Lindgren et al. (2019) reported an accuracy of 96% using SVM with MFCC features for emotion recognition in Mandarin Chinese speech [40] .

3-2 inferential comparisons between ML and DL

In the field of speech emotion recognition, machine learning (ML) and deep learning (DL) are two popular approaches used for modeling and analyzing emotional speech signals. While both ML and DL have their own strengths and weaknesses, there are some key differences between them that can impact their performance in this domain.

One of the main advantages of DL over ML is its ability to automatically learn complex features from raw data, which can be particularly useful in speech emotion recognition where emotional cues may be subtle and difficult to extract. DL models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been shown to achieve state-of-the-art performance on various speech emotion recognition tasks [25].

In terms of feature extraction, DL has shown to be more effective than ML. DL models can automatically learn features from raw data without the need for manual feature engineering. This is particularly useful in speech emotion recognition since emotions are often expressed through subtle changes in tone, pitch, and other acoustic features that may be difficult to extract manually. Some popular DL models for speech emotion recognition include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks[32] .

For feature selection, ML algorithms such as Recursive Feature Elimination (RFE) can be used to select the most relevant features for classification. DL algorithms can also perform feature selection through techniques such as Dropout and L1 regularization. [41] .

In terms of time, DL models generally require more training time than ML models due to their complexity. However, once trained, DL models can make predictions much faster than ML models since they can process large amounts of data in parallel [42] .

However, DL models also require a large amount of labeled data for training, which can be a challenge in some cases. In contrast, ML models such as support vector machines (SVMs) and decision trees can often achieve good performance with smaller amounts of labeled data. ML models also tend to be more interpretable than DL models, which can be important for understanding the underlying mechanisms behind emotional speech recognition[42].

Another important consideration is the computational resources required for training and inference. DL models typically require more computational power than ML models due to their larger number of parameters and more complex architectures. This can make them less practical for deployment on resource-constrained devices or in real-time applications [43] .

For classification, both ML and DL algorithms have been shown to perform well in speech emotion recognition tasks. ML algorithms such as SVMs and Random Forests can achieve high accuracy with relatively small datasets. DL algorithms such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have also shown promising results in speech emotion recognition tasks[13].

Overall, both ML and DL have their own strengths and weaknesses in the context of speech emotion recognition. The choice between these approaches will depend on factors such as the availability of labeled data, computational resources, interpretability requirements, and desired level of performance[44].

4. Processing of Speech

Manipulating signals to modify their key properties or extract crucial information from them, such as the recoded audio signals, which include the target speaker's speech with background noise and the voices of non-target speakers. The following procedures make up speech processing[11] :

4-1: Preprocessing

After data is gathered, the first step is preprocessing. A classifier in a SER system would be trained using the gathered data. Some of these preprocessing steps are used for feature extraction, while others handle feature normalization to ensure that little differences in speech recordings do not negatively impact the recognition outcome[45].

4-2: Framing

The subsequent process is known as signal framing. It is the method of dividing up continuous speech signals into segments of a known length in order to get around some of the problems encountered with SER systems. As a consequence of the non-stationary nature of the signals, the speaker's emotions are prone to fluctuate during the course of a speech. Despite this, the speech stays the same for a short amount of time, say 20–30 ms. The semi_fixed and local characteristics of an area may be estimated with the use of the speech signal[46]. By consciously covering 30_40% of these segments, we may also preserve the link as well as the data that exists between the frames. The use of processing techniques like the discrete Fourier transform (DFT) for feature extraction, which may be controlled by persistent voice inputs, is one example. As a consequence of this, classifiers such as artificial neural networks (ANNs) work best with frames of a constant size while keeping emotion data in speech[11] .

4-3: Windowing

The window function is applied to the frame when the framing of a voice signal has been completed successfully. Leakages arise in the course of the Fast Fourier Transform (FFT) of information due to discontinuities at the edges of the signals; these leakages are therefore decreased by the windowing function. In general, the Hamming window is considered one of the types of windowing function, and it is defined in Eq. (1) [47].

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{m-1}\right) \quad \ldots (1)$$

Where:

$w(n)$ is the frame, M is the window size and $0 \le n \le M - 1$

4-4: Detection of Vocal Activity

Unvoiced speech, voiced speech, and silence are the three parts of utterance. Voiced speech is generated if vocal cords are used to create sound [8]. However, if the vocal cords are not moving, the speaker is not voicing their words. Because of its periodic behavior, spoken speech can be isolated and identified. One possible application of a voice activity detector is to identify vocalized and unvocalized speech, as well as periods of quiet, in an audio signal [11].

4-5: Normalization

It's a technique for setting the volume to a uniform level . For normalization, the maximum value of the signal is obtained, and then the whole signal sequence is divided by the calculated maximum to estimate that every sentence has a similar level of volume. Z-normalization is generally used for normalization and is calculated as shown in Eq.(2) :

$$z = \left(\frac{x - \mu}{\alpha}\right) \quad \ldots (2)$$

where μ is the average and α represents the standard deviation of the specified vocal signal [48].

4-6: Decreasing Noise

Every speech signal contains environmental noises as well as the sounds being communicated. The presence of noise in the speech signal will severely degrade its accuracy. There are a number of noise reduction techniques available for use in this context, including minimum mean square error (MMSE) and log-spectral amplitude MMSE [49].

5. Feature Extraction

There are many parameters in the voice signal that can be used to infer the speaker's mood. The use of appropriate features is a major stumbling block in the field of emotion identification. Energy, pitch, formant, and spectral properties, including Linear prediction cepstral coefficients (LPCC) , mel-frequency cepstrum coefficients (MFCC), and modulation, have all been found to be useful in recent studies[50].The field of emotion detection has seen a proliferation of feature types in recent years, and mel-frequency cepstral coefficients (MFCCs) are just one of them. linear predictive cepstral (LPC) and perceptual (LPP) coefficients of prosody and perception (PLPs). Replacement features extracted from a long-term spectro-temporal model motivated by the human ear have been the subject of recent research[51]. As can be seen in Fig. 3, El Ayadi et al. divided certain features of speech into four distinct groups. Features can be broken down into several distinct types, such as those that are continuous, qualitative, spectral, or based on the Teager Energy Operator (TEO). Since speech signals are non-stationary and most traditional features are based on short-term analysis, it is possible to achieve high

performance by selecting them; nonetheless, constraints have encouraged researchers to design their own feature sets[52]. In order to extract features, the speech waveform must be converted to a parametric representation at a low enough data rate to be suitable for further processing and analysis.

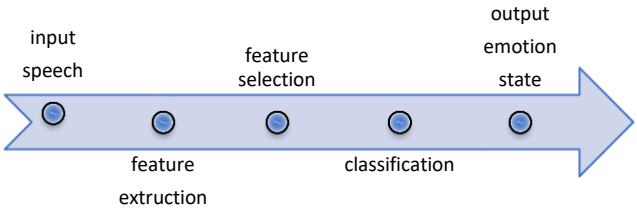

Fig. 2 : Stages of speech emotion recognition system

5-1 : Categories of Speech Features

To a large extent, SER may be defined by its linguistic traits. Every well-crafted combination of elements that adequately characterizes every emotion improves the recognition rate. In SER frameworks, several characteristics have been used. However, there is no agreed-upon set of characteristics that permits precise and detailed categorization[53]. All investigations that have been conducted so far have been experimental. We understand that speech is a continuous transmission that may be of variable duration and that carries both information and emotion. As a result, we may choose between extracting global features and extracting local features, depending on the circumstances. Global features, also known as supra-segmental or long-term features, describe the all-encompassing statistics, including maximum and minimum values, standard deviation, and mean .In contrast, local features, also known as segmental or short-term features, are used to reflect temporal dynamics and provide an imperfect representation of a stable state [54]. The certainty that emotional components are not consistently appropriated across all places of the speech signal is the source of the relevance of these fixed features. Prosodic features, spectral features, voice quality features, and Teager Energy Operator (TEO)-based characteristics are the four categories through which the global and local features of SER frameworks are investigated [55].

5 -1-1: Prosodic Features

Prosodic features are energy, pitch. Prosodic characteristics, also known as para-linguistic features, are those aspects of speech that are not represented by individual phonemes, such as rhythm and intonation. Long-lasting and prosodic characteristics are disassembled from larger parts. It is via these characteristics that the distinctive traits of emotional material may be transmitted for use in speech emotion identification. Typical prosodic traits are dependent on factors including energy, duration, and fundamental frequency [53] .

5 -1-2: Spectral Features

spectral features are MFCC, LPCC, and MEDC[55].When a human being makes a sound, their vocal tract modifies it. The quality of one's voice is determined by how their vocal tract is shaped. Having the vocal tract's form accurately represented allows for a more accurate description of the sound produced and the delivery mechanism. The frequency domain accurately portrays the characteristics of the vocal tract [56]. Spectral characteristics are obtained using the Fourier transform, which inverts the signal from the time domain to the frequency domain. Fig. 3 show Categories of speech features.

5-1-3 : Qualitative Features

voice quality refers to the characteristics of a person's voice that are not related to the words they are speaking. These features include pitch, volume, tone, and rhythm. One aspect of voice quality is the presence of certain qualitative features such as harshness, tension, and breathiness [57].

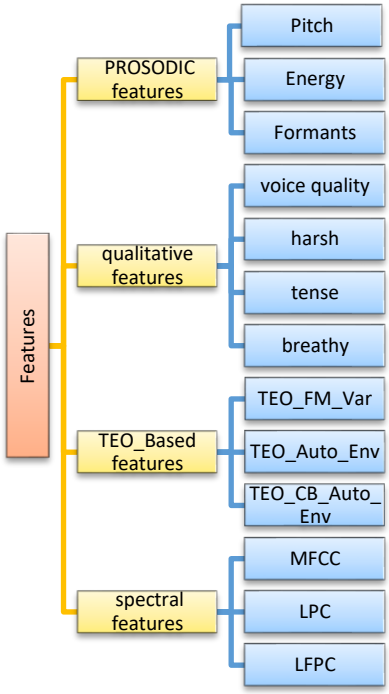

Fig. 3 : Categories of Speech Features

5-1-4: Teager Energy Operator Based Features(TEO)

Teager and Kaiser [58], [59] first proposed the Teager energy operator (TEO). TEO was based on the discovery that the ear is the primary sense used for detecting energy. A shift in fundamental frequency and crucial bands is thought to occur under pressure due to a shift in the distribution of harmonics. When someone is in a tense situation, the pressure in their muscles changes, which alters the airflow they use to create their voice. Kaiser documented Teager's operator, which quantifies the power of a speaker's words through the nonlinear process of Eq. (3).

$$w[x(n)] = x^2(n) - x(n-1)x(n+1) \quad \ldots(3)$$

where x (n) is the recorded speech signal and W is the TEO. Three novel TEO-based features are presented in [60], the crucial band-based TEO auto-correlation envelope area, the TEO-decomposed frequency modulation variation, and the normalized TEO auto-correlation envelope area.

5-2: Features Extraction Techniques

Some of the best-known techniques for extracting speech features are Subband based cepstral (SBC), Mel Energy Spectrum Dynamic Coefficients (MEDC) [61],[55],Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP). Given their widespread use and proven effectiveness, these strategies are widely trusted and accepted[62].

5-2-1: Mel-frequency Cepstrum Coefficients (MFCC)

The properties of the human ear's hearing serve as the foundation for MFCC. This device models the human auditory system with the help of a nonlinear frequency unit models the human auditory system with the help of a nonlinear frequency unit. The mel frequency scale is the aspect of speech that is utilized the most often because of its ease of computation, high capacity to distinguish between sounds, and anti-noise properties, among other benefits [63]**.** Fig. 4 shows MFCC extraction steps.

5-2-2 :Linear Prediction Cepstral Coefficients (LPCC)

The performance of the recognition system is enhanced by using a feature extraction method. First three LPCC stages are identical to the first three MFCC phases. In terms of efficiency, LPCC excels above MFCC. When compared to MFCC, LPCC has a higher noise sensitivity[64].

5-2-3:Mel Energy Spectrum DynamicCoefficients(MEDC)

The extraction of features using MEDC (Mel Energy Spectrum Dynamic Coefficients) is analogous to that using MFCC. Preprocessing, framing, windowing, the FFT-Mel filter bank, and frequency wrapping are all components of the MEDC [65] . The main difference between MFCC and MEDC is that the former uses a logarithmic mean of energies following the Mel filter bank, while the latter uses a logarithmic wraparound of frequencies. MEDC can also calculate the first and second differences for this characteristic [55].

5-2-4: Subband Based Cepstral (SBC)

Features are extracted from speech using a set of parameters. As with MFCC, SEBC has its roots in the same mathematical construction. However, wavelet packet transform is used instead of Fourier transform in the calculation of filter bank energies. As opposed to the MFCC, the SBC's parameters are more robust in the face of background noise. Consequently, SBC improves recognition performance and accuracy[61].

5-2-5: Wavelet Based Features (WBF)

It is challenging to do an analysis on a speech signal since it is a non-stationary signal that has abrupt transitions, drifts, and trends. Wavelets may be used to produce a time-frequency representation of the signals being sent. The discrete wavelet transform is the most effective method for applications involving the identification of the emotional state of the speaker [63].

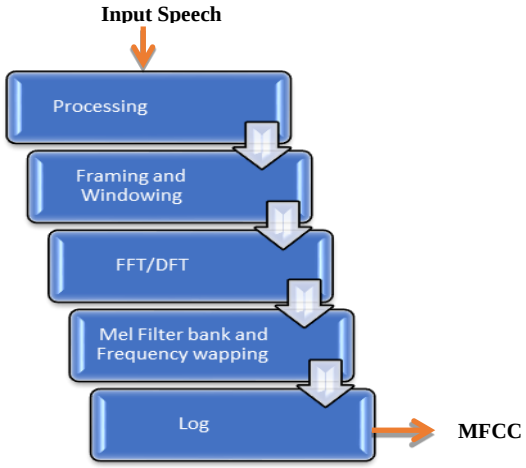

Fig. 4: steps involved in MFCC feature extraction

6. Feature Selection

Learning times for SER systems may be sped up with the use of feature selection techniques, which help narrow down the sometimes overwhelming quantity of characteristics [66].There are many characteristics and feelings associated with speech, and because it is difficult to determine for sure which set of characteristics need to be represented, there is a need for the application of feature selection approaches[67].This must be done to protect the classifiers from the perils of high dimensionality, extended training times, and overfitting, all of which have a significant impact on the accuracy of predictions [11]. Even if they were, not all of these traits would be useful for identifying emotions because different feelings may affect various aspects of a person's voice. In order to improve emotional recognition accuracy while decreasing training time, feature selection techniques are used [68].Forward Feature Selection (FFS) and Backward Feature Selection (BFS) [69], Sequential Floating Forward Selection (SFFS) [70], wrapper approach with forward selection [71], Principal Component Analysis (PCA), or Linear Discriminate Analysis (LDA) [72], [73] are some of the feature selection methods that are utilized in the SER [74] Forward Feature Selection (FFS) and Backward Feature Selection (BFS) [69] as well as the Elimination of Recursive Features (RFE)[75]. The ability to extract speech characteristics that effectively characterize the emotional content of the speech while at the same time being

independent of the speaker or the word content is an essential component of the SER[76].

7.Classification Approaches

The classification stage in speech emotion recognition involves using ML or DL algorithms to analyze the extracted features from the speech signal and classify it into one or more emotional categories such as happy, sad, angry, or neutral , elc . This stage is crucial for accurately identifying the emotional state of the speaker and providing appropriate responses or actions[77].ML (Machine Learning) and DL (Deep Learning) are two popular approaches used in speech emotion recognition,ML classification approaches involve the use of algorithms that learn from data to classify emotions. These algorithms include Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and Random Forests. SVM is a popular algorithm used in speech emotion recognition due to its ability to handle high-dimensional data and its robustness against noise[78].DL classification approaches, on the other hand, involve the use of neural networks that learn from large amounts of data to classify emotions. These neural networks include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. CNNs are commonly used in speech emotion recognition due to their ability to extract features from raw audio signals[11].

Conclusion

In this research paper, all the methods and techniques mentioned for detecting emotion in human speech and everything that is presented and explained is a survey of previous research papers. The system for detecting speech emotions consists as an initial stage, which is the initial processing of the audio signal and the process of collecting the types of data sets used, then the stage of extracting features from this data using various techniques. From this step, many features will be produced for the purpose of training, which will enter the feature selection stage. Before training, which will help reduce training time and give optimal results. Finally, the traditional classification algorithms and deep learning algorithms were reviewed, which previous experiments have proven to be very promising and give excellent prediction results. Previous experiments have also shown that using convolutional neural networks and spectral features such as MFCC is a commonly used option and gives good results.

References

[1] D. Gupta, P. Bansal, and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition," *Speech Lang. Process. Human-Machine Commun. Proc. CSI 2015*, pp. 195–207, 2018.

[2] H. Cai, Y. Lin, and R. R. Mourant, "Study on Driver Emotion in Driver-Vehicle-Environment Systems Using Multiple Networked Driving Simulators," no. January, 2007.

[3] S. O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and Detection of Cognitive Load and Frustration in Drivers ' Speech," pp. 8–11, 2009.

[4] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Enhancement of an Arabic speech emotion recognition system," *Int. J. Appl. Eng. Res.*, vol. 13, no. 5, pp. 2380–2389, 2018.

[5] Mohamed, Omar, and Salah A. Aly. "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset." arXiv preprint arXiv:2110.04425 (2021).

[6] S. R. Krothapalli and S. G. Koolagudi, *Emotion recognition using speech features*. Springer, New York, 2013.

[7] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[9] A. D. Dileep and C. C. Sekhar, "GMM-Based Intermediate Matching Kernel for Classification of Varying Length Patterns of Long Duration Speech Using Support Vector Machines," vol. 25, no. 8, pp. 1421–1432, 2014.

[10] H. Aouani and Y. Ben Ayed, "ScienceDirect ScienceDirect Speech Emotion Recognition with deep learning Emotion Recognition with learning Ben deep," *Procedia Comput. Sci.*, vol. 176, pp. 251–260, 2020, doi: 10.1016/j.procs.2020.08.027.

[11] T. M. Wani *et al.*, "A Comprehensive Review of Speech Emotion Recognition Systems," pp. 47795–47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

[12] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 186–202, 2015.

[13] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking : Studies on acted and spontaneous speech ‹," *Comput. Speech Lang.*, pp. 1–17, 2014, doi: 10.1016/j.csl.2014.01.003.

[14] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *2017 International conference on inventive communication and computational technologies (ICICCT)*, 2017, pp. 109–114.

[15] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS One*, vol. 13, no. 5, p. e0196391, 2018.

[16] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014.

[17] P. Jackson and Sju. Haq, "Surrey audio-visual expressed emotion (savee) database," *Univ. Surrey Guildford, UK*, 2014.

[18] S. T. Jovicic, Z. Kasic, M. Dordevic, and M. Rajkovic, "Serbian emotional speech database: design, processing and evaluation," in *9th Conference Speech and Computer*, 2004.

[19] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.

[20] F. Schiel, S. Steininger, and U. Türk, "The SmartKom Multimodal Corpus at BAS.," in *LREC*, 2002.

[21] B. Schuller *et al.*, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1760–1774, 2009.

[22] L. Vidrascu and L. Devillers, "Real-life emotions in naturalistic data recorded in a medical call center," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006)). Genoa, Italy*, 2006, pp. 20–24.

[23] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and*

gesture recognition (FG), 2013, pp. 1–8.
[24] A. H. Meftah, M. A. Qamhan, Y. Seddiq, Y. A. Alotaibi, and S. A. Selouani, "King Saud University emotions corpus: construction, analysis, evaluation, and comparison," *IEEE Access*, vol. 9, pp. 54201–54219, 2021.
[25] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition using Deep Learning Techniques : A Review," *IEEE Access*, vol. PP, p. 1, 2019, doi: 10.1109/ACCESS.2019.2936124.
[26] 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.
[27] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, pp. 333–336.
[28] C. Caihua, "Research on multi-modal mandarin speech emotion recognition based on SVM," in *2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2019, pp. 173–176.
[29] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, 2015.
[30] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *2016 10th International conference on signal processing and communication systems (ICSPCS)*, 2016, pp. 1–8.
[31] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep CNN and BLSTM networks," *IEEE Access*, vol. 10, pp. 564–578, 2021.
[32] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Syst.*, vol. 184, p. 104886, 2019.
[33] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Appl. Acoust.*, vol. 179, p. 108046, 2021.
[34] M. Farhad, H. Ismail, S. Harous, M. M. Masud, and A. Beg, "Analysis of emotion recognition from cross-lingual speech: Arabic, English, and Urdu," in *2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 2021, pp. 42–47.
[35] B. Fernandes and K. Mannepalli, "Speech Emotion Recognition Using Deep Learning LSTM for Tamil Language.," *Pertanika J. Sci. Technol.*, vol. 29, no. 3, 2021.
[36] A. Jacob, "Speech emotion recognition based on minimal voice quality features," in *2016 International conference on communication and signal processing (ICCSP)*, 2016, pp. 886–890.
[37] A. Lindgren and G. Lind, "Language Classification Using Neural Networks." 2019.
[38] Y. Hifny and A. Ali, "Efficient arabic emotion recognition using deep neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6710–6714.
[39] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimed. Tools Appl.*, pp. 1–68, 2021.
[40] W. Zhang *et al.*, "Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services," *Softw. Pract. Exp.*, vol. 47, no. 8, pp. 1127–1138, 2017.
[41] S. Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Appl. Acoust.*, vol. 173, p. 107721, 2021.
[42] A. Biswas, S. Chakraborty, A. N. M. Y. Rifat, N. F. Chowdhury, and J. Uddin, "Comparative analysis of dimension reduction techniques over classification algorithms for speech emotion recognition," in *Emerging Technologies in Computing: Third EAI International Conference, iCETiC 2020, London, UK, August 19–20, 2020, Proceedings 3*, 2020, pp. 170–184.
[43] J. Chauhan, J. Rajasegaran, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "Performance characterization of deep learning models for breathing-based authentication on resource-constrained devices," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 4, pp. 1–24, 2018.
[44] L. Deng and D. Yu, "Deep learning: methods and applications," *Found. trends® signal Process.*, vol. 7, no. 3–4, pp. 197–387, 2014.
[45] B. M. Nema and A. A. Abdul-kareem, "Preprocessing Signal for Speech Emotion Recognition," vol. 28, no. 3, 2017.
[46] T. Özseven, "Evaluation of the Effect of Frame Size on Speech Emotion Recognition," *2018 2nd Int. Symp. Multidiscip. Stud. Innov. Technol.*, pp. 1–4, 2018.
[47] H. Beigi and H. Beigi, *Speaker recognition*. Springer, 2011.
[48] B. M. Nema and A. A. Abdul-kareem, "Preprocessing Signal for Speech Emotion Recognition," no. November, 2017, doi: 10.23851/mjs.v28i3.48.
[49] J. Pohjalainen, F. Fabien Ringeval, Z. Zhang, and B. Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," in *Proceedings of the 24th ACM international Conference on Multimedia*, 2016, pp. 670–674.
[50] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech Emotion Recognition : Methods and Cases Study," vol. 2, no. Icaart, pp. 175–182, 2018, doi: 10.5220/0006611601750182.
[51] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011.
[52] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 2007, vol. 4, pp. IV–957.
[53] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *Int. J. Speech Technol.*, vol. 21, pp. 93–120, 2018.
[54] Y. Gao, B. Li, N. Wang, and T. Zhu, "Speech emotion recognition using local and global features," in *Brain Informatics: International Conference, BI 2017, Beijing, China, November 16-18, 2017, Proceedings*, 2017, pp. 3–13.
[55] Y. Pan, P. Shen, and L. Shen, "Feature extraction and selection in speech emotion recognition," in *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2005), Como, Italy*, 2005.
[56] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, and D. Mürbe, "Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall," *Biomech. Model. Mechanobiol.*, vol. 14, pp. 719–733, 2015.
[57] I. R. Titze and K. V. Abbott, *Vocology: The science and practice of voice habilitation*. National Center for Voice and Speech, 2012.
[58] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech Prod. speech Model.*, pp. 241–261, 1990.
[59] J. F. Kaiser, "Some useful properties of Teager's energy operators," in *1993 IEEE international conference on acoustics, speech, and signal processing*, 1993, vol. 3, pp. 149–152.
[60] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. speech audio Process.*, vol. 9, no. 3, pp. 201–216, 2001.
[61] K. V. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using MFCC and wavelet features," *Proc. 2013 3rd IEEE Int. Adv. Comput. Conf. IACC 2013*, pp. 842–847, 2013, doi: 10.1109/IAdCC.2013.6514336.
[62] S. A. Alim and N. K. A. Rashid, *Some commonly used speech feature extraction algorithms*. IntechOpen London, UK:, 2018.
[63] M. Babu, "INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS ISSN 2320-7345 EXTRACTING MFCC AND GTCC FEATURES FOR

EMOTION RECOGNITION FROM AUDIO SPEECH SIGNALS," vol. 2, no. 8, pp. 46–63, 2014.
[64] M. Rana and S. Miglani, "Performance analysis of MFCC and LPCC techniques in automatic speech recognition," *Int. J. Eng. Comput. Sci.*, vol. 3, no. 08, 2014.
[65] Y. D. Chavhan, B. S. Yelure, and K. N. Tayade, "Speech emotion recognition using RBF kernel of LIBSVM," in *2015 2nd international conference on electronics and communication systems (ICECS)*, 2015, pp. 1132–1135.
[66] [66] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, pp. 155–177, 2015.
[67] T. Özseven, "The acoustic cue of fear: investigation of acoustic parameters of speech containing fear," *Arch. Acoust.*, vol. 43, no. 2, pp. 245–251, 2018.
[68] T. Özseven, "A novel feature selection method for speech emotion recognition," *Appl. Acoust.*, vol. 146, pp. 320–326, 2019.
[69] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and Y.-H. Chang, "Emotion recognition and evaluation of Mandarin speech using weighted D-KNN classification," in *Proceedings of the 17th conference on computational linguistics and speech processing*, 2005, pp. 203–212.
[70] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," in *2006 14th European Signal Processing Conference*, 2006, pp. 1–5.
[71] J. Sidorova, "Speech emotion recognition with TGI+. 2 classifier," in *Proceedings of the Student Research Workshop at EACL 2009*, 2009, pp. 54–60.
[72] S. Haq, P. J. B. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*, 2008.
[73] T. Özseven and B. E. Özseven, "A Content Analysis of the Research Approaches in Music Genre Recognition," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2022, pp. 1–13.
[74] D. Gharavian, M. Sheikhan, A. Nazerieh, and S. Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network," *Neural Comput. Appl.*, vol. 21, pp. 2115–2126, 2012.
[75] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning." IntechOpen, 2019.
[76] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process.*, vol. 22, no. 6, pp. 1154–1160, 2012.
[77] S. Supraja, S. Tatinati, K. Hartman, and A. W. H. Khong, "Automatically linking digital signal processing assessment questions to key engineering learning outcomes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6996–7000.
[78] *K. R. Bagadi and C. M. R. Sivappagari, "An evolutionary optimization method for selecting features for speech emotion recognition," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 21, no. 1, pp. 159–167, 2023*

التعرف الذكي على عواطف الكلام: دراسة استقصائية

علي عبد الوهاب يحيى — **فوزية محمود رمو**
كلية علوم الحاسبات والرياضيات — **كلية علوم الحاسبات والرياضيات**
ali_alsaffar@.uomosul.edu.iq — fawziyaramo@uomosul.edu.iq

تاريخ القبول: 2023/5/15 — **تاريخ الاستلام: 2023/2/28**

الملخص

يعد التعرف على عاطفة الكلام *(SER)* مهمة صعبة في مجال الذكاء الاصطناعي والتعلم الآلي , على مر السنين ، اقترح الباحثون أساليب مختلفة للتعرف على المشاعر من إشارات الكلام. ستحلل هذه المقالة وتناقش بعض الأعمال السابقة حول التعلم الآلي والتعلم العميق في ***SER.*** تركز هذه الدراسة الاستقصائية على أهمية صوت الإنسان في تحديد الحالات العاطفية والنفسية. تم استخدام طرق مختلفة لتصنيف المشاعر بنجاح مثل الغضب والحزن والسعادة والخوف والاشمئزاز والحياد والمفاجأة. تمت المراجعة في هذه الدراسة على مراحل متسلسلة بما في ذلك المعالجة المسبقة واختيار الميزات والتصنيف وتقييم النتائج. كما تمت مراجعة مجموعات البيانات المختلفة للغات الدولية مثل الإنجليزية ، والهندية ، والألمانية ، والأردية ، والتاميلية ، والفرنسية ، والعربية. ركزت الدراسة بشكل أساسي على الذكاء الاصطناعي وخوارزميات التعلم الآلي نظرًا لمرونتها وسهولة فهمها بنتائج مميزة.

الكلمات المفتاحية : تعلم الالة , التعلم العميق , تمييز عواطف الكلام , معاملات درجة النغم, استخلاص المعلومات , الشبكة العصبية الالتفافي