*Grenze International Journal of Engineering and Technology, Jan Issue*

# Sentiment Analysis in Voice, Speech, and Audio: A Comprehensive Review of Feature Extraction and Detection Techniques

M.M. Bapat[1], C.H. Patil[2] and S.M.Mali[3]
[1-3]School of Computer Science, Dr. Vishwanath Karad MIT World Peace University,
Pune, Maharashtra, India 411 038
Email: mmbapat@mitacsc.ac.in, chpatil.mca@gmail.com, shankarmali007@gmail.com
[1]MAEERS MIT Arts, Commerce & Science College, Alandi (D), Pune, Maharashtra, India

***Abstract*—With the increasing use of speech and audio in modern communication, sentiment analysis using these modalities has become an important area of research. This article offers a thorough discussion of the most recent developments in speech emotion recognition. It covers a wide range of topics, including emotional speech corpora, various speech features, and models for emotion recognition from speech. In addition, a large number of representative voice datasets have been evaluated in terms of the languages used, the number of speakers included, the range of emotions covered, and the aim of the collection. This study presents a comprehensive overview of sentiment analysis from three distinct vantage points, namely databases, feature extraction techniques, and methods of detection. This is accomplished by conducting an in-depth analysis of previously published works. This review aims to provide researchers with a comprehensive understanding of the current state of sentiment analysis in voice, speech, and audio by synthesizing the research that has been done on this topic and providing a summary of that research.**

***Index Terms*— Sentiment detection, Speech emotion recognition, Speech feature extraction, Deep learning, Machine learning.**

## I. INTRODUCTION

Paralinguistic and non-linguistic information is also used by humans to communicate with one another in addition to language [1]. Included in paralinguistic data are style, attitude, and intention. Sentiment and emotion are examples of non-linguistic information. Sentiment can be thought of as a division of emotion. It is comparable or identical to valence in terms of emotions with dimensions. Yet, sentiment and categorical emotion has been the subject of distinct investigations in the past. Speech sentiment analysis usually follows the algorithm shown in Figure 1.

Speech is a sophisticated communication that carries data about the speaker, the message, the language, the emotion, and more. The majority of currently available speech systems process studio-recorded, neutral speech well, but they perform poorly when processing emotional speech. This is because it is challenging to model and characterize the emotions that are expressed in speech. Speech becomes more natural when emotions are present. Non-verbal communication in a conversation conveys significant information, such as the speaker's goal. The emotion recognition from the speech follows the process shown in Figure 2.

*Grenze ID: 01.GIJET.10.1.554_1*

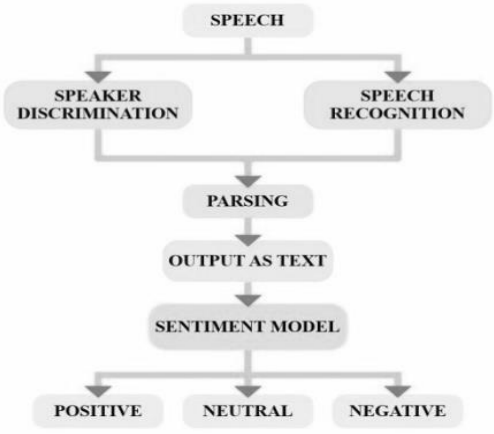


Figure 1: Speech Sentiment Analysis

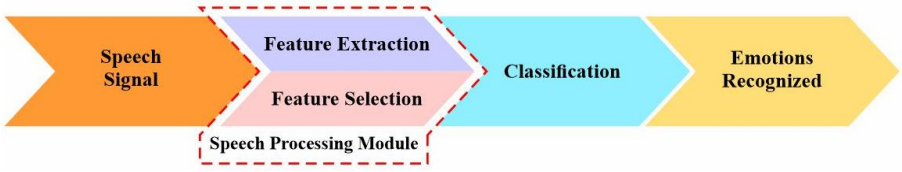


Figure 2: Speech Emotion Recognition System

In addition to text-based communication, the way the words are delivered also carries important non-linguistic information. By combining the right emotions, the same textual message could be sent with varied meanings. Depending on how it is said, spoken text can be interpreted in several various ways. For instance, in English, the phrase "OKAY" might be used to convey appreciation, skepticism, agreement, apathy, or an assertion. To clarify the semantics of a given utterance, one must grasp the spoken language as well as the written text. Yet, speech systems must be able to process both the message and non-linguistic information, such as emotions [2]. This paper is limited to the analysis of speech-related emotions. Understanding emotions that are available from the speech and synthesizing desirable emotions in a speech following the targeted message are the two fundamental goals of emotional speech processing. From a computer's point of view, classifying or separating different emotions might be considered as understanding speech emotions. You could think of the synthesis of emotions as putting emotion-specific knowledge into speech synthesis.

## II. Literature Review

Early studies on speech emotion identification focused on recognizing acoustic parameters like frequency, utterance intensity, and bandwidth. Teager energy operated-based features, formants, Mel frequency cepstral coefficients (MFCC), log frequency power coefficients (LFPC), linear prediction cepstral coefficients (LPCC), and other features provide the basis for further research.


Figure 3: Plutchik's emotion wheel

The emotions expressed in different mental conditions are well explained by Plutchik [3] in a work that presents the classification of the emotions by employing a dimensional approach, in the form of a wheel, as shown in Figure 3.

*A. Databases*

The databases of the audio that are used for emotion recognition are comprised of different types of emotions as can be seen from Figure 4. "The Toronto Emotional Speech Set (TESS)" [4] is a dataset made up of 2,800 brief audio clips pronounced by two women who are, respectively, 20 and 60 years old. For 200 different sentences, they each mimicked seven sentiments. A team of 56 students blindly labeled the dataset.

Busso et al. [5] proposed the "Interactive Emotion Dyadic Motion Capture (IEMOCAP)" dataset, which is frequently used in speech sentiment analysis applications, was first proposed in 2008. Professional actors gather it through many hours of talk and dialogue in various scenes. It lasts up to 12 hours and has five segments. The dataset contains nine different emotions in total. Three or more factors analyze each conversation, and when higher than half of the factors reach a consensus, the dialogue is given the associated label. The conversation is labeled "Other" when the raters cannot agree. This method of gathering data is more natural. Different types of emotions considered for the sentiment analysis from speech or audio are shown in Figure 4.

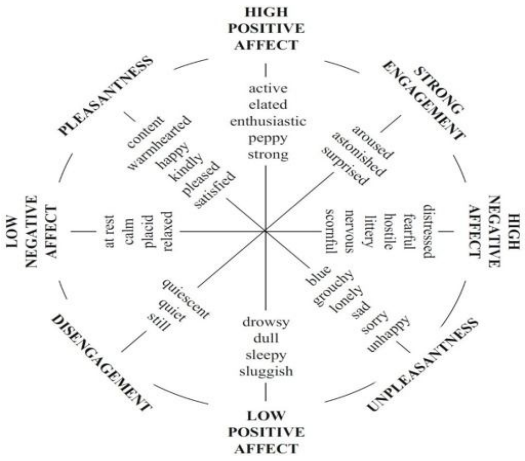


Figure 4: A two-dimensional structure of emotional space

The Danish emotional database, "DES (Danish Emotional Speech)" [6, 7] is supported by the European Union. It accommodates speech recordings from four professional actors, two of whom are male and two of whom are female. The two separated words "yes" and "no," nine brief phrases – four of which are questions – and two paragraphs make up the recordings. The five emotional states were all imitated in each of the spoken words. One of the most popular databases in this sector has been used by scholars. Its uses include evaluating SVM methods on their own or associating with hidden Markov models.

The "Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" dataset [8] consists of audio and visual English recordings of 24 individuals (12 male and female). 7,356 files are present in RAVDESS (total size: 24.8 GB). In this instance, only the speech samples are utilized, and the sentences are said with eight various emotional expressions. The difficulty level of sentiment analysis from speech or audio depends on the database, as can be seen in Figure 5.

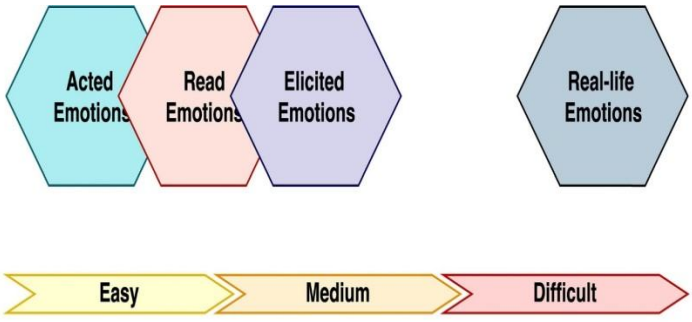


Figure 5: Emotion recognition databases and their difficulty level

The study by Ramakrishnan [9] highlights the normalization of signals, which is a pre-processing stage that is carried out before the extraction of the features, in addition to the databases, features, and classifiers in the SER systems. Additionally, he offers SER system application areas that are not SER technologies but have an impact on them in terms of needs and design. In a recent yet succinct assessment, Basu et al. [10] highlight articles involving databases, noise reduction methods for pre-processing of signals, features, and classifiers, including the latest developments like convolutional and recurrent neural networks.
Spanish professional actors, one male, and one female, are recorded for INTER1SP [11]. It has 184 utterances, including single words, numbers, and complete phrases. 6,040 samples totaling approximately 4 hours of data from each speaker are included. It was developed as part of a European effort. Both traditional machine learning and deep learning methods use it as a dataset. The 535 German audio files in the "EMODB" dataset [12] are broken down into seven different emotion categories. Ten performers—five males and five females—simulated the various emotions while uttering ten German expressions—five brief and five lengthy—that could be used in normal conversation and could be understood in the context of all employed emotions. The dataset was assessed using a perceptual test that looked at emotions' naturalness and capacity to be recognized. More than 60 people deemed the utterances to be natural. The data was phonetically labeled with unique markers depending on the quality of voice, phonatory and articulatory settings, and articulatory characteristics. The wide variety of databases considered for this review article is mentioned in Table I.

TABLE I. MOST CITED SER DATABASES. #EM = TOTAL NUMBER OF EMOTIONS, AV = DATA OF AUDIO-VISUAL, NAT. = NATURAL V/S ACTORS RECORDING, #MS = TOTAL MALE SUBJECTS #FS = TOTAL FEMALE SUBJECTS, #UT = NUMBER OF WORDS, SENT = NUMBER OF SENTENCES

| Database | #em | language | AV | Nat. | #ms | #fs | #ut | #sent |
|---|---|---|---|---|---|---|---|---|
| TESS [**Error! Bookmark not defined.**] | 7 | English | n | actors | - | 2 | 200 | 2800 |
| IEMOCAP [**Error! Bookmark not defined.**] | 10 | English | y | actors | 5 | 5 | - | - |
| DES [**Error! Bookmark not defined.**, **Error! Bookmark not defined.**] | 5 | Danish | n | actors | 2 | 2 | 13 | |
| RAVDESS [**Error! Bookmark not defined.**] | 8 | English | y | actors | 12 | 12 | 2 | 1440 |
| INTER1SP [**Error! Bookmark not defined.**] | | Spanish | n | actors | 1 | 1 | 184 | 6040 |
| EMODB [**Error! Bookmark not defined.**] | 7 | German | n | actors | 5 | 5 | 10 | 800 |
| MSP-PODCAST [**Error! Bookmark not defined.**] | 8 | English | n | natural | - | - | - | 18000 |

"MSP-PODCAST" [13] is a strategy to efficiently develop a big, realistic emotional database with overall emotional content. It depends on currently available spontaneous recordings gathered from audio-sharing services issued under public licenses. It contains more than 18,000 natural emotional sentences, with a length of over 27 hrs, from numerous persons with audio chunks ranging from 2.75 sec. to 11 sec. Then, a set of roughly 300 assessors tagged the samples with emotional qualities ("arousal", "valence", "dominance") and an enlarged list of category emotions. Only the labeling with emotional qualities is balanced. It has been utilized to evaluate a listener-dependent strategy to employ architectures of deep learning to SER.

### *B. Methods of feature extraction*

To classify emotions, Gao et al. [14] employ a linear kernel SVM with sequential minimal optimization. Pitch, intensity, "MFCC (Mel Frequency Cepstral Coefficients)", "LSP (Line Spectral Pairs)", and "ZCR (Zero Crossing Rate)" are calculated on the audio files as local characteristics of windows between 20 and 100 ms. To get values for the duration and overlap of windows, they also employ a depth-first search. The statistics of all speech local features that were extracted from each utterance then smoothed, normalized, and used as global features are then obtained. A sample of the extraction of MFCC features from audio signals is shown in Figure 6. Since it is generally known that the audio part processing in the voice sentiment analysis task is an activity that goes with a certain sequence. For the model to properly interpret the audio signal's sentiment aspects, the sequence of the audio signal is crucial. The Long Short-Term Memory (LSTM) network is regarded by the majority of researchers as the ideal model to handle sequential modeling tasks since the advent of deep learning.

Zhang et al. [15] suggested a unique heterogeneous parallel convolutional Bi-LSTM model while taking into account the acoustic model and the efficiency of feature extraction. The model understands spatiotemporal feature information using heterogeneous parallelism, permitting more effective feature extraction.
"Convolutional neural networks (CNNs)" and "LSTM" networks are combined in the method for "context-aware" emotionally relevant feature extraction that Trigeorgis et al. [16] proposed. On this kind of job, deep neural networks typically perform well after being trained on basic acoustic data, however, continuous recognition of dimensional emotion from audio is not been extensively considered. It is handled as a sequential

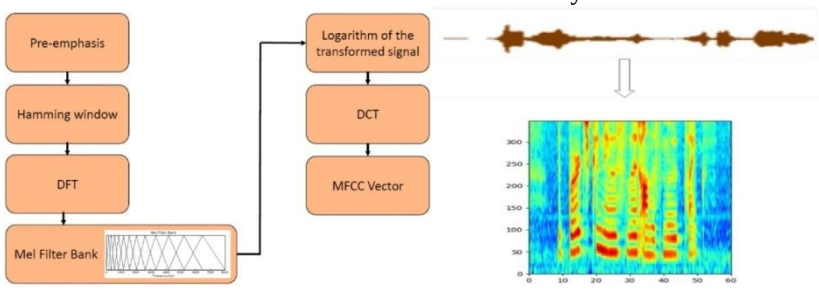


Figure 6: (a) MFCC extraction process (b) Example of extracting MFCC features from audio signals

regression issue, where the objective is to minimize the average deviation while maximizing the correlation between successive regression outputs and continuous-valued emotion contours [17]. Figure 7 depicts an example of a deep learning architecture that combines CNNs and LSTMs to process the input MFCC image and identify the emotions.

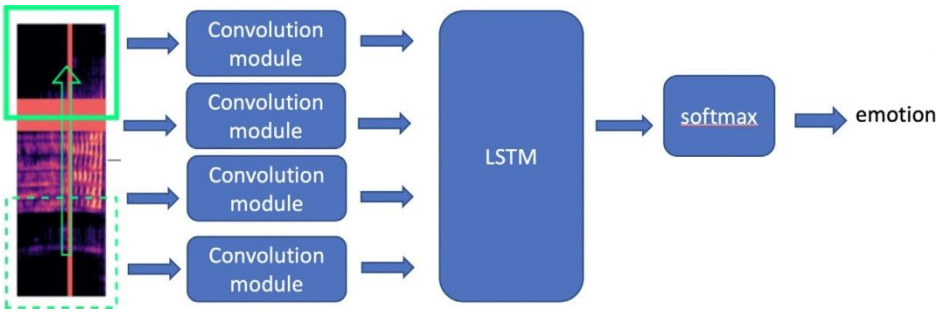


Figure 7: A deep learning architecture for SER based on MFCC input as a sequence of images

An intriguing solution to the features extraction challenge is offered by Kadiri et al. [18] By recording the differences between the emotional and the neutral, non-emotional speech, they take a different technique than that often used in the state-of-the-art. According to their report, their hierarchical binary decision tree-based approach is on par with or superior to the current prosody and spectral features.
Frame-level low-level descriptors (LLD) and high-level statistical functions (HSF) were integrated by Vlasenko et al. [19] for effective speech emotion identification. The findings highlighted feature integration at various feature extraction stages. There is, however, no research examining the balance between the employment of LLD, HSF, and hybrid characteristics. It was not suggested how to speed up the extraction of functions, which calls for computing LLD. A feature extraction algorithm was proposed by Koduru et al. [20]. To assess the speech signal, this algorithm primarily extracts the acoustic characteristics, like discrete wavelet transform, pitch, over-zero rate, "MFCC", and energy.
Deep learning techniques are now receiving more and more interest in audio research. Using a hybrid of pitch, zero crossing, formants, MFCC, and its statistical characteristics, Dahake et al. [21] retrieved the features. After comparing it to autocorrelation and AMDF, the cepstral algorithm determines the pitch. MFCC, intensity, spectral flux (SF), beat histogram, and the spectral centroid are audio variables that are crucial for sentiment analysis. The openSMILE software must be capable of successfully extracting the audio elements related to vocal pitch and intensity. This software is used to extract audio parameters such as the pause time, voice quality, beat histogram, pitch, and spectral centroid (SC). It is also possible to extract statistical functions such as amplitude mean, skewness, and standard deviation. The well-known openSMILE program, which is used to extract audio features, will extract the features with a sliding window at a rate of 30 Hz. The identification of attitudes based on audio is crucial for human-computer interaction, claim the researchers (Tajadura-Jiménez and

Västfjäll [22]; Vogt et al. [23]). Local and global characteristics are the two categories of speech-related features that can be retrieved. By dividing the audio modality into various overlapped and non-overlapped segments, the audio modality can be simply examined. Within each section, the signal is thought to be static.
For recognizing emotions, Jiang et al. [24] suggest using parallelized convolutional recurrent neural networks. They start by identifying traits in each speech that are remembered through the use of long short-term memory. They also compute the Mel spectrograms and extract visual attributes using a CNN. To determine the emotion from the utterance, these two collections of high-level features are recognized, normalized, and introduced into a softmax classifier.
The research team Tamulevicius et al. [25] has used a variety of emotional speech-based databases that span different languages to perform emotion recognition from voice data. According to the investigation conducted by the researchers Khalil et al. [26], it is simple and effective to determine the sentiment by taking the emotions out of the speech. There are some speech-based sentiment analysis database options available.
The majority of current studies primarily focus on extracting speech features, although the models' precision and prediction rates still need to be increased. In a study by Pang et al. [27], a new framework is presented to enhance the extraction followed by the fusion of speech sentiment feature information. To improve the model's accuracy, the framework uses a model of hierarchical conformer and a model of attention-based GRU. Learning groups of local and global features are the two fundamental components of the method. To more accurately obtain the spatiotemporal properties and the long- and short-term dependencies of the speech signal, this work combines convolution and a transformer. This method takes into account the need to create suitable learning groups for local and global features. To improve the extraction of local, long, and short-term feature information, convolution and transformer are combined. The attention mechanism then performs the feature fusion to access the weights of feature information once the AUGRU model has extracted the global features. The proposed system in this study achieves good results, with IEMOCAP and REVDESS achieving recognition accuracy rates of 80% and 81%, respectively.

TABLE II. DIFFERENT FEATURE EXTRACTION TOOLS USED FOR SPEECH EMOTION RECOGNITION

| Method | Description | Reference |
|---|---|---|
| MFCC | Speech and music discrimination. Higher-order MFCCs have a higher proportion of music-specific information, while fewer MFCCs have a higher proportion of speech-specific information. | Gao et al. [**Error! Bookmark not defined.**], Koduru et al. [**Error! Bookmark not defined.**], Dahake et al. [**Error! Bookmark not defined.**] |
| MFCC with CNN-LSTM | MFCC is used for the extraction of images from an audio signal which will act as an input to CNN-LSTM | Trigeorgis et al. [**Error! Bookmark not defined.**] |
| LLD | LLDs are hand-crafted acoustic features from short parts that are combined with RNN | Vlasenko [**Error! Bookmark not defined.**] |
| HSF | HSFs, which integrate DNN with specific statistics based on LLDs over all frames of an utterance | Vlasenko [**Error! Bookmark not defined.**] |

### *C. Methods of detecting sentiments*

The two major methods that can be employed for sentiment analysis from speech or audio are deep learning and machine learning methods. The flow mechanism of deep learning and machine learning is shown in Figure 8.

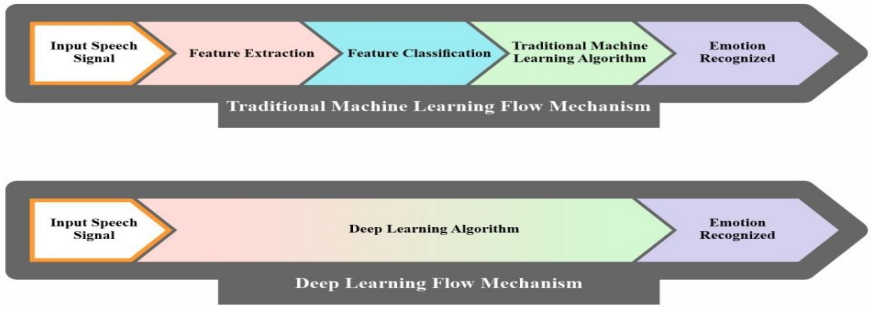


Figure 8: Machine Learning Flow v/s Deep Learning Flow

*D. Deep Learning Approach*

Most of the current academic work has employed models of deep learning for tasks of voice sentiment analysis, such as “CNNs”, “RNNs”, “LSTMs”, and “DNNs”. Indeed, deep learning models can keep the original low-level properties while learning deeper features from voice signals than typical machine learning models. “Deep Stochastic Convolutional Neural Network (DSCNN)” architecture was presented by Mustaqeem and Kwon [28]. The design employs a shared network methodology to identify distinguishable characteristics in the voice signal's spectral map. To enable better implementation, the architecture has been improved. At the time of the experiments on audio speeches with an emotion dataset, higher recognition rates are produced by combining MFCC (Mel-frequency cepstral coefficient), discriminant analysis, and NSL (Neural Structure Learning) than by using other conventional approaches like “MFCC-DBN (Deep Belief Network)”, “MFCC-CNN (Convolutional Neural Network)”, and “MFCC-RNN (Recurrent Neural Network)”. The strategy outlined above, put forth by Uddin et al. [29], can be used in intelligent settings like homes or clinics to deliver effective healthcare. NSL can be tested on edge devices with constrained datasets gathered from edge sensors because it is quick and simple to build.

Slimi et al. [30] put forth the utilization of a DL network to be employed with small-sized databases. The technique does not employ typical data enhancement techniques. Firstly, the spectrogram connected to each recording is created by utilizing traditional procedures, and scaled while keeping the aspect ratio. It has been claimed that the greater the final image dimensions, the greater the accuracy, albeit it also entails increased computing complications, and increasing the amount of neural network characteristics. The neural network comprises just one unseen layer and greater accuracy has been reported than the state of the art.

Standard features, temporal feature integration strategies, and 1D and 2D “DCNN” architectures were thoroughly compared by Vrysis et al. [31]. The developed convolutional algorithms outperformed the conventional feature-based methods in terms of performance. The best “2D DCNN” architecture outperformed “1D DCNN” in accuracy while using a similar amount of parameters. In addition, “1D DCNN” executed four times more slowly.

Many scholars have developed speech sentiment analysis frameworks and tools over the past few years. In this part, some latest unique methods and improvements to these methods are given. In a study, Lu Zhiyun et al. [32] recommended using pre-trained ASR model parameters to resolve the downstream investigation of speech sentiments. End-to-end ASR characteristics are demonstrated to incorporate acoustic and textual speech data and to deliver proof-of-concept outcomes. They used RNN as a self-aware sentiment classifier that also enables a clear display utilizing weights to comprehend model predictions. They conducted their analysis using an “IEMOCAP” dataset and a brand-new, sizable “SWBD” – sentiment dataset. This method obtains an accuracy of 70.10% on “SWBD” – sentiment and raises the state-of-the-art accuracy of “IEMOCAP” from 66.6% to 71.7%.

An innovative voice sentiment recognition system was put out by Chernykh and Prikhodko [33]. The “Deep Recurrent Neural Network (DSRNN)” and a connectionist time-based classification approach are used in the algorithm. A bunch of acoustic features calculated by small speech intervals is used to train the network. The system can forecast the emotional sequence of sentences and take into consideration the non-emotional components that are included in emotionally charged words and phrases.

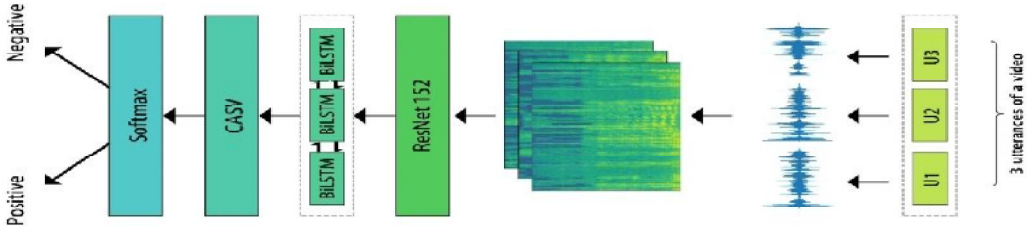


Figure 9: Overview of ResNet152 CNN Model [50]

The Conversational Transform Network (CT Net), developed by Lian et al. [34], simulates intra-modal and cross-modal interactions between multimodal features using a transformer-based design. Multi-headed attention networks aid in the formation of speaker- and context-sensitive dependency models and also the improved temporal information capture in discourse. The system suggested in this paper uses a model conformer, which is a fusion of a “CNN” with a transformer and generally performs great in learning local features. The “CNN” branch employs the “ResNet” structure and the transformer branch uses the “ViT” structure in the conformer, which is a parallel two-body network structure. A ResNet152 CNN model can be seen in Figure 9.

*E. Machine Learning Approach*

The machine learning approach that can be used for sentiment analysis using speech or audio has wide classification as shown in Figure 10.

For voice emotion recognition, Hajarolasvadi and Demirel [35] presented a "3D CNN" model. The utterances were processed as overlapping frames, and for each frame, 88-dimensional characteristics and a spectrogram were retrieved. K-means clustering was used with the retrieved features to pick the k-most discriminant frames for use in the 3D spectrogram representation. This method makes it possible to record both temporal and spectral data. The proposed architecture was able to outperform a "2D CNN" model that had been trained before being applied to the SER challenge.

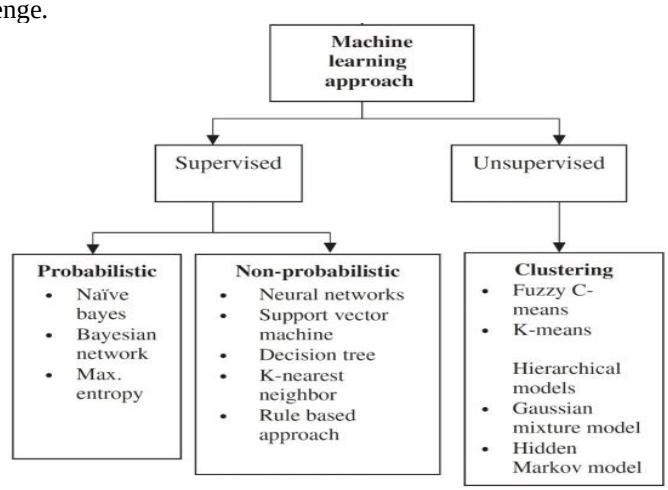


Figure 10: Classification of machine learning approach

To identify emotions, Milton et al. [36] use SVM classifiers with linear and RBF kernels and MFCC features. By suggesting the use of a 3-stage SVM, they advance the state of the art. Speech emotions are categorized by Sinith et al. [37] using SVM and a combination of MFCC, pitch, and energy-based characteristics. A speech-based emotion detection system is described by Zhu-Zhou et al. [38] that uses MFCCs as features while simulating various real-world noise and reverberation settings.

In a study by Ghriss et al. [39], an effort was made to incorporate sentiment information into the job of voice emotion recognition. In that study, the authors trained a dataset using the loss of automatic speech recognition (ASR) and loss functions of cross-entropy sentiment to accommodate the sentiment-aware technique to enhance speech emotion identification performance. The model was subsequently improved to predict valence, arousal, and dominance using the "Concordance Correlation Coefficient Loss (CCCL)" function. It has been demonstrated that CCCL produces higher results than other loss functions [40]. The findings highlighted the relationship between sentiment analysis and prediction of valence by showing an enhancement in valence prediction in the identification of emotions in speech.

Source features are speech attributes that are obtained from excitation source signals. Speech is used as the excitation source signal after vocal tract (VT) features have been suppressed. This is accomplished by first estimating the VT information from the speech signal using filter coefficients (LPCs), and then differentiating it using an inverse filter formulation. The emerged signal, known as the linear prediction residual, mainly tells us where the excitation came from. Features generated from LP residual are indicated as the source of excitation, sub-segmental, or just source characteristics in research by Makhoul [41]. To better understand glottal pulse features, glottis open and closure phases, and excitation intensity, sub-segmental analysis of speech signals is used. The excitation source features can be used to estimate the elements of the glottal activity that are unique to the emotions.

In a work by Liu et al. [42], the number of retrieved features is reduced using "LDA (Linear Discriminant Analysis)" and "PCA (Principal Component Analysis)" methods, and the classification stage is then completed using a genetic algorithm. The accuracy of the speaker-dependent outcomes is 71.05% for the "FAU Aibo" dataset, 76.40% for the "SAVEE" dataset, and 90.28% for the "CASIA" dataset, whereas the speaker-independent results are 64.60% ("FAU Aibo"), 38.55% ("CASIA"), and 44.18% ("SAVEE") (SI).

Speech emotion and sentiment recognition for interactive dialogue systems were carried out by Bertero et al. [43]. The "TED-LIUM" release 2 corpora was annotated by the authors for emotion recognition, and Twitter and

Movie Review corpora were used for sentiment analysis. These authors used word2vec for sentiment analysis and raw speech for emotion recognition. The authors utilized two different models for each task, even though they recommended doing both sentiment analysis and emotion recognition (using different datasets). The usage of "SVM" and "CNN" techniques for emotion recognition was assessed by the authors. The authors compared the "CNN" and "LIWC (Linguistic Inquiry and Word Count)" approach for sentiment analysis.

Same to the investigation of sentiments in text, the analysis of sentiments in speech is critical to pinpoint sentiments by utilizing certain acoustic parameters including pitch, bandwidth, speaking rate, intensity, and other factors. It is favored for voice sentiment recognition because some prosodic elements better convey emotions. According to a study by Koolagudi et al. [44], acoustic parameters are dependent on personality characteristics that influence sentiment polarity.

The speaker-dependent approach and the speaker-independent approach are two categories into which existing techniques for sentiment identification from speech can be divided. As demonstrated in [45], the speaker-dependent strategy produced significantly better performance than the speaker-independent technique. More accuracy was attained by combining the "Mel frequency cepstral coefficients (MFCC)" with the "Gaussian mixture model (GMM)" as a classifier. The speaker-dependent strategy, however, is impractical in many applications. The best outcome for the speaker-independent application thus far is just 81% [46], where features are chosen using the sequential floating forward selection technique] (SFFS).

Moreover, Jing et al. [47] look for the ideal qualities for rendering audio files. It is suggested a novel sort of prominence-related feature that, along with conventional acoustic parameters, can be used to categorize seven different types of common emotional states. In speaker-dependent and speaker-independent studies using classifiers from the "Support Vector Machines (SVM)", "k closest neighbors (kNN)", "Naive Bayes (NB)", and "Partial Least Squares-Discriminatory Analysis" families, the prominence features are confirmed. According to the findings, utilizing combination characteristics increases average identification rates by 6.2% in speaker-independent studies and by 6.6% in speaker-dependent experiments as compared to utilizing simply acoustic features.

Using a "1D CNN" for the raw audio recording, a "2D CNN" for the Mel-spectrogram created out of the original waveform, and a "3D CNN" for temporal-spatial dynamic, Zhang et al. [48] present a template of "multi-CNN" to identify emotion in utterances. An average-pooling layer integrates the outputs to create the categorization outcome. The tests use databases in the Mandarin language and only slightly enhance other cutting-edge solutions.

The fragmentation of the audio signals into acoustically homogenous portions is a step in the pre-processing of speech signals. These areas are then divided into speech and non-speech zones, and they aid in speaker identification. After utilizing specific techniques to denoise the voice data, the background regions can also be isolated. To determine whether the words were spoken by the same speaker, it is important to identify the speaker. It also addresses identifying gender and socioeconomic factors. A speaker identification system that separates the audio stream into homogeneous sections depending on the speaker's identity has been created by the researchers Sinha et al. [49].

## III. CONCLUSION

The field of sentiment analysis has recently emerged as a major area of research, prompting extensive investigation and a plethora of studies. This paper provides a comprehensive overview of the most recent advancements in speech and audio sentiment analysis. The review has highlighted the difficulties posed by emotional speech corpora as well as the diverse techniques and models for emotion recognition from speech. In addition, a large number of representative voice datasets have been evaluated in terms of their language, speaker count, emotional range, and collection purpose. The review has presented sentiment analysis from three unique perspectives: databases, feature extraction techniques, and detection methods. This review aims to provide researchers with a comprehensive understanding of the current state of sentiment analysis in voice, speech, and audio by conducting an in-depth analysis of previously published works. Even though the field of speech emotion recognition is maturing in terms of the relevant signal features and the widespread use of deep learning architectures, additional validation and analysis are required to determine their sensitivity to data sampling and overfitting. The joint exploitation of available corpora can help alleviate the data deficiency and lead to the development of more general emotion recognizers capable of identifying emotions across the diverse characteristics of different databases. In addition, the rapid development of deep learning architectures and features is anticipated to lead to significant improvements in sentiment analysis performance shortly, given the

importance of emotional user interaction. Overall, this paper is a valuable resource for researchers interested in speech and audio sentiment analysis.

## REFERENCES


[1] H. Fujisaki, "Prosody, Information, and Modeling with Emphasis on Tonal Features of Speech," In Proceedings of the Workshop on Spoken Language Processing, 9–11 January 2003, Mumbai, India.

[2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, Vol. 15, pp. 99–117, 2012.

[3] R. Plutchik, "Emotion: a psychoevolutionary synthesis," Harper and Row, New York, 1980.

[4] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto emotional speech set," Canadian Acoustics, Vol. 39, No. 3, pp. 182–183, 2011

[5] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, Vol. 42, No. 4, pp. 335–359, 2008.

[6] I. S. Engberg and A. V. Hansen, "Documentation of the Danish emotional speech database," Internal AAU report, Center for Person Kommunilation, Denmark, 1996.

[7] S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in Proceedings of 5th European Conference on Speech Communication and Technology, pp. 1695–1698, 1997.

[8] 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, e0196391, 2018

[9] S. Ramakrishnan, "Recognition of Emotion from Speech: A Review," Speech Enhancement, Modeling and Recognition-Algorithms and Applications, Intech Open Publishers, London, UK, March 2012

[10] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," In: International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 109–114, 2017.

[11] V. Mapelli, "Inter1sp: Spanish emotional speech synthesis database," European Language Resources Association, 2011

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," Interspeech, Lisbon, Portugal, 4–8 Sep. 2005

[13] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing, Vol. 10, No. 4, pp. 471–483, 2019.

[14] Y. Gao, B. Li, N. Wang, and T. Zhu, "Speech emotion recognition using local and global features," International Conference on Brain Informatics, pp. 3–13, 2017.

[15] H. Zhang, H. Huang, and H. Han, "A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition," Applied Sciences, Vol. 11, No. 21, 9897, Oct. 2021.

[16] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," In: Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 5200–5204, 20–25 March, 2016.

[17] F. Weninger, F. Ringeval, E. Marchi, B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI, New York, USA, pp. 2196–2202, 9–15 July 2016.

[18] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, P. Alku, and B. Yegnanarayana, "Excitation features of speech for emotion recognition using neutral speech as reference," Circuits, Systems, and Signal Processing, Vol. 39, pp. 4459–4481, 2020.

[19] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," In: Proceedings of Interspeech 2007, pp. 2249–2252, 2007

[20] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," International Journal of Speech Technology, Vol. 23, pp. 45–55, 2020

[21] P. P. Dahake K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and support vector machine," In: International conference on automatic control and dynamic optimization techniques (ICACDOT), pp. 1080–1084, 2016.

[22] A. Tajadura-Jiménez and D. Västfjäll, "Auditory-induced emotion: A neglected channel for communication in human–computer interaction," In Affect and emotion in human–computer interaction, Berlin Heidelberg: Springer, Vol. 4868, pp. 63–74, 2008.

[23] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation," In Affect and emotion in human–computer interaction, Berlin Heidelberg: Springer, Vol. 4868, pp. 75–91, 2008.

[24] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," IEEE Access, Vol. 7, pp. 90368–90377, 2019.

[25] G. Tamulevicius, G. Korvel, A. B. Yayak, P. Treigys, J., Bernataviciené, and B. Kostek, "A study of cross-linguistic speech emotion recognition based on 2D feature spaces," Electronics, Vol. 9, No. 10, 1725, 2020

[26] 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, 7, pp. 117327–117345, 2019.
[27] P. Zhao, F. Liu, and X. Zhuang, “Speech Sentiment Analysis Using Hierarchical Conformer Networks,” Applied Sciences, Vol. 12, No. 16, 8076, Aug. 2022.
[28] Mustaqeem and S. Kwon, “A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition,” Sensors, Vol. 20, No. 1, 183, Dec. 2019
[29] M. Z. Uddin and E. G. Nilsson, “Emotion recognition using speech and neural structured learning to facilitate edge intelligence,” Engineering Applications of Artificial Intelligence, Vol. 94, 103775, 2020
[30] A. Slimi, M. Hamroun, M. Zrigui, and H., Nicolas, “Emotion recognition from speech using spectrograms and shallow neural networks,” In: ACM Int. Conf. Advances in Mobile Computing and Multimedia, pp. 298–301, 2020.
[31] L. Vrysis, N. Tsipas, I. Thoidis, and C. Dimoulas, “1D/2D Deep CNNs vs. Temporal Feature Integration for General Audio Classification,” Journal of Audio Engineering Society, Vol. 68, pp. 66–77, 2020.
[32] Z. Lu, L. Cao, Y. Zhang, C. C. Chiu, and J. Fan, “Speech sentiment analysis via pre-trained features from end-to-end ASR models” In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7149–7153, 2020.
[33] V. Chernykh and P. Prikhodko, “Emotion recognition from speech with recurrent neural networks,” arXiv:1701.08071, 2017.
[34] Z. Lian, B. Liu, and J. Tao, “CTNet: Conversational transformer network for emotion recognition,” IEEE/ACM Transaction Audio Speech Language Processing, Vol. 29, pp. 985–1000, 2021.
[35] N. Hajarolasvadi and H. Demirel, “3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms,” Entropy, Vol. 21, 479, 2019.
[36] A. Milton, S. S. Roy, and S. T. Selvi, “SVM scheme for speech emotion recognition using MFCC feature,” International Journal of Computer Applications, Vol. 69, No. 9, pp. 34–39, 2013.
[37] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and S. Rajan, “Emotion recognition from audio signals using support vector machine,” In: 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 139–144, 2015.
[38] F. Zhu-Zhou, R. Gil-Pita, J. Garcia-Gomez, and M. Rosa-Zurera, “Robust multiscenario speech-based emotion recognition system,” Sensors, Vol. 22, No. 6, 2022.
[39] A. Ghriss, B. Yang, V. Rozgic, E. Shriberg, and C. Wang, “Sentiment-Aware Automatic Speech Recognition Pre-Training for Enhanced Speech Emotion Recognition,” In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7347–7351, Singapore, 23–27 May 2022.
[40] B. T. Atmaja and M. Akagi, “Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition” Journal of Physics Conference Series, Vol. 1896, 012004, 2021.
[41] T. J. Makhoul, “Linear prediction: A tutorial review,” in Proceedings of the IEEE, Vol. 63, No. 4, pp. 561-580, April 1975.
[42] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao, J. P. Xu, and G. Z. Tan, “Speech emotion recognition based on feature selection and extreme learning machine decision tree,” Neurocomputing, Vol. 273, pp. 271–280, 2018.
[43] D. Bertero, F. B. Siddique, C. S. Wu, Y. Wan, R. Ho, Y. Chan, and P. Fung, “Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems,” In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, pp. 1042–1047, 1–5 November 2016.
[44] S. G. Koolagudi, N. Kumar, and K. S. Rao, “Speech emotion recognition using segmental level prosodic analysis,” In 2011 international conference on devices and communications (ICDeCom), pp. 1–5, 2011.
[45] E. Navas, I. Hernaez, and I. Luengo, “An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS,” IEEE Transactions on Audio Speech Language Processing, Vol. 14, No. 4, pp. 1117–1127, 2006.
[46] H. Atassi and A. Esposito, “A speaker independent approach to the classification of emotional vocal expressions,” In: 20th IEEE international conference on tools with artificial intelligence, pp. 147–152, Dayton, Ohio, USA, 3–5 November 2008.
[47] S. Jing, X. Mao, and L. Chen, “Prominence features: Effective emotional features for speech emotion recognition,” Digital Signal Processing: A Review Journal, Vol. 72, pp. 216–231, 2018.
[48] S. Zhang, X. Tao, Y. Chuang, and X. Zhao, “Learning deep multimodal affective features for spontaneous speech emotion recognition,” Speech Communication, Vol. 127, pp. 73–81, 2021.
[49] R. Sinha, S. Tranter, M. Gales, and P. Woodland, “The Cambridge University March 2005 speaker diarisation system,” 2005
[50] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, “Tensor Fusion Network for Multimodal Sentiment Analysis,” In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Stroudsburg, PA, USA, pp. 1103–1114, September 2017.