IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 6, December, 2021, pp. 1861-1873
Journal homepage: http://iieta.org/journals/ts

A Perspective Study on Speech Emotion Recognition: Databases, Features and Classification Models

Kogila Raghu[1,2*], Manchala Sadanandam[1]

[1] Department of CSE, Kakatiya University, Warangal 506001, India
[2] Department of CSE, Vardhaman College of Engineering, Hyderabad 501218, India

Corresponding Author Email: raghu_kogila@vardhaman.org

https://doi.org/10.18280/ts.380631

Received: 29 October 2021
Accepted: 3 December 2021

Keywords:
ASR, HCI, SER, Telugu emotional speech, acoustic, SVM, MLP, CNN

ABSTRACT

Automatic Speech Recognition (ASR) is a popular research area with many variations in human behaviour functionalities and interactions. Human beings want speech for communication and Conversations. When the conversation is going on, the information or message of the speech utterances is transferred. It also consists of message which includes speaker's traits like emotion, his or her physiological characteristics and environmental statistics. There is a tremendous number of signals or records that are complex and encoded, but these can be decoded quickly because of human intelligence. Many academics in the domain of Human Computer Interaction (HCI) are working to automate speech generation and the extraction of speech attributes and meaning. For example, ASR can regulate the usage of voice command and maintain dictation discipline while also recognizing and verifying the speech of the speaker. As a result of accent and nativity traits, the speaker's emotional state can be discerned from the speech. In this Paper, we discussed Speech Production System of Human, Research Problems in Speech Processing, SER system Motivation, Challenges and Objectives of Speech Emotion Recognition, so far the work done on Telugu Speech Emotion Databases and their role thoroughly explained. In this Paper, our own Created Database i.e., (DETL) Database for Emotions in Telugu Language and the software Audacity for creating that database is discussed clearly.

1. INTRODUCTION

There are a variety of uses for speech processing, and it's becoming more popular in specific fields. Providing a natural and pleasant form of communication and interaction with a non-private computer, speaker focus will become an increasingly important technological advancement. It is possible to eliminate typing, free up your hands and walk away from your devices, all while ensuring security, by using only your voice. Because of this, Speech Recognition and its alloyed Technologies, as well as their primary goal of developing a system that mimics or mimics human behavior, particularly the ability to speak naturally and respond appropriately to spoken language, have been a leading factor in the development of speech processing [1]. The acoustic points of speech have been established to differ from person to person for a long time and have been used in speaker recognition for a long time. Reproducing each individual's anatomy (e.g., a person's throat, size and shape of mouth) and learned behavioral patterns is done through these acoustic shapes (e.g., pitch, style of talking).

1.1 Basics of speech production, speech reorganization

Humans' ability to express themselves verbally is an essential part of their capacity for social interaction. It's made by altering the vocal tract's interest rate over time. When the lungs are empty, the voice chords begin to vibrate. The glottis vibrates, which causes air to flow into the supra-glottal area, which is above the glottis, in quasi-periodic pulses. Together, the supra glottal region is comprised of the pharynx, nasal cavities, and oral cavities. The quasi-periodic pulses generated as air moves through the oral cavity and throat modify numerous articulators (such as lips, teeth, tongue, etc.). Reduced velum also alters the airflow through the nasopharynx and nostrils after it leaves the nose. Changing the pulses and air flow in the gadget causes sound waves to radiate from the nose and mouth. Figure 1 depicts the principal components of the Human Speech Production System.

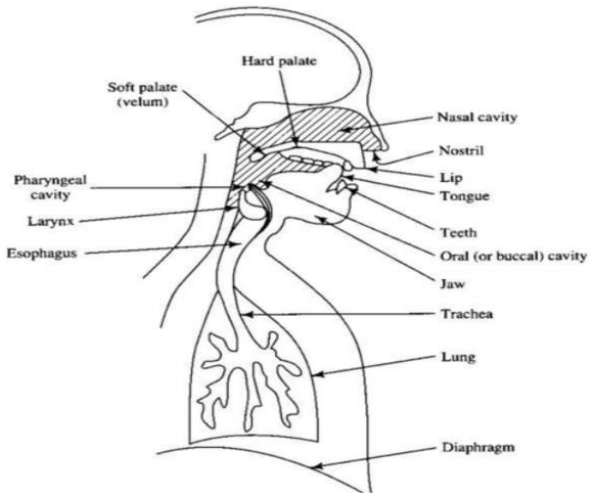

Figure 1. Speech production system of humans

The generation of a voiced sound is possible when the excitation source is the vibrating vocal folds. Unvoiced sounds can be heard when the air is forced to flow through an extremely narrow space. Random movement of air particles compared to quasi-periodic flow generates turbulence. Vocal tract noises that lack a distinct voice might have a variety of locations in the vocal tract where they originate, depending on the sound's articulation point and pitch. The size of the vocal tract is dealt and subject to the noise generator only. Based on the vocal tract portions in front and backside of constriction, noise will be altered. In some frequency ranges, the noise strength amplitude suddenly decreases. Fricative sounds can be detected by listening to these. The rapid release of the 'stop' consonants results in a loud noise blast. The 'stop' consonant is modified by the vocal tract filter mostly based on four articulation points. Because of the articulator's flexibility, a wide range of speech alterations are possible. Using a basic microphone, one can obtain a wide range of speech characteristics, including the complete frequency spectrum. Using a basic microphone is preferable because of the signal's high perceptual exceptionality and intelligibility.

1.2 Recent advances and developments in speech emotion recognition (SER)

In the early 1950s, investigators tried to work on acoustic-phonetics in the field of speech processing. In 1952, a researcher tried to find digits that were too far away for a normal single speaker to hear. For the first Automatic Speech Recognition (ASR) systems, the full speech was segmented and speech recognition algorithms were devised. These segmented utterances aided in the recognition of different people's voices. When electronics advanced in the 1970s, the pace of speaker recognition development accelerated. As early as 1979, Sakoe and Chiba [2] suggested a method for recognizing related phrases. In 1981, Myers and Rabiner [3] introduced a novel technique, Dynamic Time Warping (DTW), which is similar to the Dynamic Programming algorithm. While it may have seemed complicated at first, that machine was also efficient and versatile. The Hidden Markov Model (HMM) [4-6] was employed for ASR lookups in the early 1980s. Research in this area is still a work in progress, and a great deal of progress is needed in the future. Since its introduction in 1996, Speech Emotion Recognition (SER) has found numerous uses in our day-to-day lives (2009). Human-computer interaction (HCI) and brain-computer interaction (BCI) are becoming increasingly automated thanks to recent trends, advancements, and improvements in modern information technology. Emotion is a powerful indicator of a person's mental state; therefore, SER is utilized in a broad range of applications, including medical diagnosis, call centers, banking, self-driving automobiles, and silent communication. It is difficult to do current study since there is no standard database for a specific region's spoken language. SER development is influenced by the many aspects of speakers and languages. The SER system should be able to handle any situation, no matter how challenging it may be.

1.3 Accent recognition

As a result of linguistic and ethnic distinctions between speakers of a given language, accents are formed. For native and non-native speakers, there is a noticeable difference in the acoustic region spanned by phonemes. Accent recognition relies heavily on factors such as intonation, length, rhythm, and the time taken to release the voice. As a result of the recognition of accents, the overall performance of speech processing systems will increase. Accent Recognition technology enhances Human-Computer Interaction (HCI) After recognizing the user's accent in IVRS systems, the computer can converse with the user in their native language. With regard to accent, there are various degrees of accuracy in speech. The vocal tract produces a range of sonic devices at the segmental level, most of which are depending on the accent of the speakers.

The vocal tract's spectral envelope is determined by the MFCC, or Mel-Frequency Cepstral Coefficients. Vocal folds of the open/closed section duration contain accent-specific data. A voice signal contains a variety of data, such as the strength characteristics, pitch dynamics, duration, and dialect. In the sub-segmental level, glottal pulses with time periods in proportion to the open phases as well as closed phases of the vocal folds may be available. It is possible to identify every language's accent using features from the segmental and supra-segmental levels. For the extraction of segment-level features, a speech segment of standard length (20-30 ms) is employed. These are also referred to as spectral facets because they are found in the speech segment's frequency spectrum.

The supra-segmental characteristics have a duration of more than 100 milliseconds and are referred to as prosodic features. It takes fewer than three milliseconds for sub-segmental features to enter the speech phase. According to the literature study, a few research have attempted to identify dialects in Japanese and English using an automated method. However, dialect evaluation in Indian languages is still in its infancy. Literature survey results show that relatively little effort has been put into the use of speech features extracted from dialects. As a result of this motivation, the current suggested effort focuses on extracting Telugu speech and dialect characteristics. Telugu has a wide variety of regional accents, especially in Andhra Pradesh and Telangana, the two states that make up the Telugu speaking region.

1.4 Emotion recognition

Additionally, these speech samples help to identify the speaker and additional examination helps to evaluate the individual's behavioral state of mind by analyzing their speech signals. The said speech generates a speech signal that includes both the physical expression of the words and the emotional feelings they convey. To a greater or lesser extent, all people share the same basic emotional reaction to a certain expression. Each type of verbal emotion has a corresponding prosody. Many elements influence this prosody, such as the community's living conditions, the language's rules, and the location of the individual culture. As a general rule, emotion recognition is a means of analyzing the behavior of an individual based on the content of a speech sample they have uttered. It can also be viewed as a science of understanding the minds of individual**s.**

The organization of the paper is as follows. Section 2 explained SER: Review which contains Speech Emotion Data bases and its importance, Feature Extraction Techniques and Classification Models. After that in section 3 Research Problem Motivation given. In section 4, Research Problem Objectives discussed and at the last, in section 5 Conclusions are given.

2. SPEECH EMOTION RECOGNITION (SER): REVIEW

When confronted with a certain event, each person displays a distinct emotional response. One of the most difficult tasks is determining an individual's emotional state of mind when they respond to a real-world circumstance. The ability to discern a person's emotional state helps foster social and interpersonal interactions. An individual's emotional state of mind can be determined by analyzing their spoken language and its symbolic representation. Figure 2 depicts the general block diagram of the SER System.

Speech Emotion Recognition (SER) system contains mainly three phases. They are Speech Emotion Database, Feature Extraction Phase and Classification Model Phase. These are explained in the following Sections.

2.1 Speech emotion databases and its importance

The first stage in developing speech-processing algorithms is to create a database tailored to the application. For both training and testing purposes, this database could be utilized. In the early stages of the project, acoustic-phonetic research was the primary focus and the research strategy was laid out. Text-dependent and text-free speech databases were created in the second phase of the project. In addition to identifying and verifying speakers, speech databases can also be used to study topics such as emotional expression, gender, age [7] and regional and social dialects. Various Speech Emotion datasets are shown in Table 1. In terms of size, diverse emotions, and construction method, these Speech Emotion Corpus datasets are all unique.

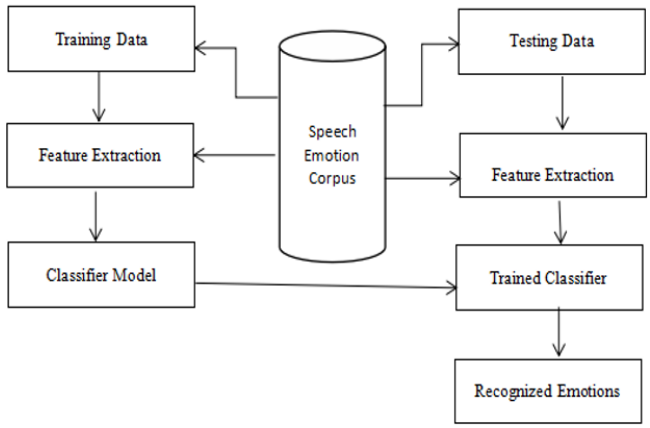

Figure 2. General block diagram of SER

Table 1. Various speech emotion databases and their description

S.No	Database	Language	Type	Size	Emotions
1	Dutch Emotional Database (1984) [8]	DutchS	Simulated	8speakers, 4 phrases (4females, 4 males)	Anger, neutral, contempt, disgust, surprise, fear, interest, joy.
2	Danish Emotional Speech (1997) [9]	Danish	Simulated	(2 Speakers) neutral recordings, extra recordings (3speakers),4 speakers in total (2F,2M).	Anger, happiness, sadness, surprise, neutral
3	SUSAS (1999) [10]	English	Simulated	32 Speakers (13M, 19F) 16000 Utterances	Happy,Anger,Sad,Fear and Neutral
4	EMO-DB (1999) [11]	German	Simulated	10 Speakers 800 Utterances	happy, angry, anxious, fearful, bored and disgust
5	Emotional Speech DB (1998) [12]	English	Semi-Natural	5 Speakers 23*5 Speakers	Anger, fear, happy, sad, surprised, neutral
6	Emotional Speech (1999) [13]	Spanish	Simulated	Single Actor	Anger, disgust, happiness, and sadness, 2000 phones per Emotion.
7	Emotional Speech (2000) [14]	English	Simulated	40 Speakers * 5 Sentences	Anger, fear, happiness, sadness, neutral
8	Emotional Speech (2000) [15]	English	Simulated	22 Patients and 19 Healthy Persons	Depression and Neutral
9	Spanish Emotional DB (2001) [16]	Spanish	Simulated	2 Actors	Anger, fear, disgust, sadness, neutral,joy,neutral
10	Emotional Speech (2001) [17]	English	Simulated	8 Actors	Anger,Joy,Sad, Neutral,Surprise and Fear
11	Emotional Speech (2001) [18]	Chinese	Simulated	Native TV Actors 721 Utterances	Anger, happiness, neutral and sad
12	RUSLANA (2002) [19]	Russia	Simulated	61 Actors 610 Utterances	Anger, neutral, happiness, sadness, fear, surprise
13	Speech Emotion DB (2002) [20]	Chinese	Simulated	9 Native Speakers 288 Sentences per Emotions	Anger, Sad, Fear, Joy and Neutral
14	Emotion Speech DB (2003) [21]	Japanese	Simulated	6 Speakers 4800 Utterances	Anger, joy, sorrow, normal
15	Speech Emotion DB (2003) [22]	English	Simulated	Single Actor	Anger,Fear,Happy,Sad, Neutral
16	Emotion Speech DB (2003) [23]	Japanese	Simulated	2 Native Speakers	Anger,Joy and Sadness

17	Emotional Speech DB (2004) [24]	Greek	Simulated	1 Female 20 short sentences, 25 long sentences, 12 passages, 10 single words of fluent speech	Anger, fear, joy, sadness, neutral
18	Emotional Speech DB (2004) [25]	Swedish	Simulated	One Single Speaker	Happiness and Neutral
19	Emotional Speech DB (2004) [26]	Italian	Simulated	Single Speaker	Anger,Disgust,Joy,Sad, Fear,Disgust
20	Emotion Speech DB (2005) [27]	Korean	Simulated	10 Speakers 5400 files	Angry, joyful, sad, neutral
21	Emotion Speech DB (2005) [28]	Berlin	Simulated	10 Actors 800 Speech files	Anger,Joy,Fear,Boredom, Disgust,Neutral
22	Speech Emotion DB (2006) [29]	German	Simulated	51 School Children (21F + 30M)	Different Emotions
23	Emotion DB (2006) [30]	Swedish	Simulated	7619 utterances	Emphatic, negative, neutral
24	IEMOCAP (2008) [31]	English	Simulated	12 hours	happiness, sadness, anger and frustration
25	Speech Emotion DB (2008) [32]	German	Natural	104 Native Speakers	Two basic emotions
26	IIT-KGP-SESC (2010) [33]	Telugu	Simulated	10 Speakers 12000 Utterances	Anger, fear, happy, disgust, compassion, sarcastic, surprise, neutral
27	TESS (2020) [34]	English	Simulated	2800 Utterances	Anger, fear, happiness, disgust, surprise, neutral, sadness
28	IIT-KGP-SEHSC (2011) [35]	Hindi	Simulated	10 Speakers 12000 Utterances	Anger, fear, happy, disgust, neutral, sarcastic, compassion, surprise
29	SAVEE (2011) [36]	English	Simulated	4 Actors 480 Utterances	Anger, Sad, Happy, Fear, Neutral, Disgust and surprise
30	Keio University Japanese Emotional Speech Database (Keio-ESD) (2009) [37]	Japan	Simulated	940 utterances, 71 speaker (male)	Anger, disgusting, downgrading, happiness, gentle, funny, worried, relief, shameful, etc.(47emotions)
31	RECOLA (2013) [38]	French	Natural	46 speakers (19 males, 27 females) 7 hour speech Utterances	Five social behaviours (engagement, agreement, dominance, rapport performance)
32	Chinese Natural Emotional Audio-Visual Database (CHEAVD) (2017) [39]	Mandarin	Semi-Natural	238 speakers 140 minutes emotional speech From TVs, movies.	Anger, disgust, happiness, anxious, neutral, surprise, sadness and worried
33	Mixed Two Language Speech Emotion DB (2014) [40] (PDREC)	Latin-American, Japanese	Natural	57 Speakers (30-Latin American and 17 Japanese)	Negative,Positive
34	Persian Drama Radio Emotional Corpus (2014) [41]	Persian	Simulated	33 Speakers (15M,18F) 748 Utterances	Anger,Happy,Sad,Fear, Disgust,Boredom,Neutral
35	EMOVO (2014) [42]	Italian	Simulated	6 Professional Actors 14 Sentences and 588 recordings	Anger,Sad,Joy,Fear, Disgust,Surprise,Neutral
36	Multi language Speech Emotion DB (2016) [43]	German (Berlin DB), English (eNTERFACE)	Simulated	10 Speakers (494 files) and 42 speakers (1170 files)	Anger,Happy,Sad, Surprise,Boredom, Disgust
37	Multi Lingual Speech Emotion DB (2018) [44]	Tamil, Malayalam, English	Simulated	10 Speakers	Happy,Sad,Anger
38	RAVDESS (2019) [45]	English	Simulated	24 Speakers 1440 Utterances	Calm, happy, sad, angry, fearful, surprise, and disgust
39	SHEMO (2000) [46]	Persian	Semi-Natural	87 Speakers 3000 Utterances	Anger, fear, happiness, sadness and surprise
40	IIITH-TEMD (2007) [47]	Telugu	Semi-Natural	38 Speakers 5317 Utterances	Anger, Happy, Sad, Surprise, Neutral, Sarcastic, etc.

A platform and improved recognition accuracy are not provided by traditional Telegu databases. Most of the Indian languages' speech research is in its infancy. As a result, a database specifically for Indian languages should be built in the lab [5]. It is possible to discern emotions in Telugu speech using a variety of databases, as shown in Table 2.

Table 2. Overview of various Telugu speech emotion databases

Sl.No	References	Database Type	Description	Purpose and Approach	Emotions
1	Koolagudi et al. (2009) [48]	Simulated	This Data Base created and collected from All India Radio (AIR), Vijayawada. It contains 10 Professional Artists (5male & 5female) -Total number of Utterances are 12,000	Recognition Design, acquisition, post processing and evaluation of IITKGP SESC database	Anger, disgust, fear, happy, compassion, neutral, sarcastic, surprise
2	Kadiri et al. (2015) [49]	Telugu Semi natural and simulated	This Database is created by IIT-H. It is collected from Students (2 female & 5 male). Total number of Utterances are 200.	Recognition Excitation source feature analysis for speech emotion recognition	Anger, happy, neutral, sad
3	Vijai Bhaskar and Ramamohana Rao (2013) [50]	Natural simulated	Database is created with 9 persons, each person has 4 emotional speech signals	Telugu speech emotion classification system using K-NN Classifier.	normal, happy, sad and angry
4	Prasad Reddy et al. 2010 [51]	Natural simulated	Emotional speech database, Feature extraction, HMM model and Recognized emotion output	The Emotional speech database is from Telugu. Rural Dialects of Andhra Pradesh.	anger, surprise, happiness, sadness and neutral
5	Gangamohan et al. (2013) [5]	Natural SImulated Excitation source	Excitation source KL distance values taken as consideration	The emotion recognition system for the 4-class	anger, happy, neutral and sad
6	Pravena and Govind et al. (2013) [52]	Natural SImulated Excitation source	GMM model was considered	IITKGP-SESC Telugu speech emotion databases for classification of three emotions	anger, happy and sad
7	Ram and Ponnusamy (2018) [53]	Telugu Semi natural and simulated	Support vector machine was considered	Telugu_DB	Anger, neutral, happy and sadness
8	Rambabu et al. (2020) [54]	Semi Natural	This Database is created by IIIT-H, contains 38 native speakers, 19 professionals and 19 Non-Professionals. Overall 5317 speech files are there.	TEMD created for Speech Processing like Speaker, emotion recognition, etc.	Happy, Sad, Surprise, Anger, Neutral

2.2 DETL (Database for emotions in Telugu language)

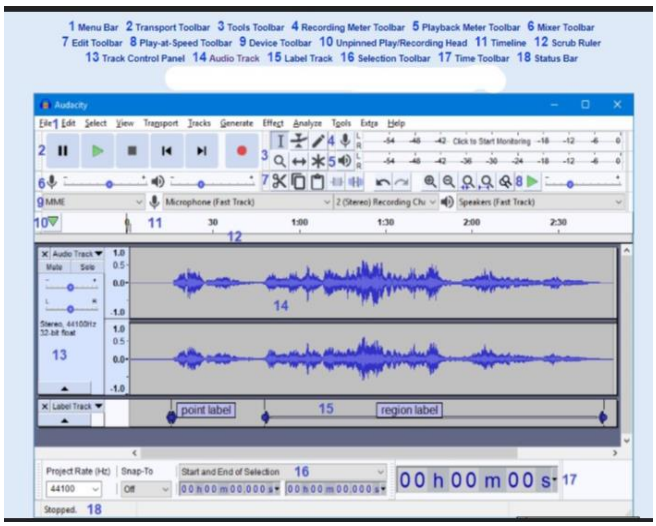

Figure 3. Audacity software opening window

In order to improve the database size and verification of system, we create a new database in addition to existing database to improve research work in Emotion recognition of Speech. It is a Simulated Emotion of Telugu Language Speech, which is designed and developed by Audacity Software. This Database is collected from various sources like movies, news readings and natural utterances spoken from native Telugu speakers. The Speakers are aged between 25 years to 55 years and in mixed of male and female. The Audacity Software opening window as shown in Figure 3.

For the collection of DETL, first recording has to be done with record option which is in red colour. Select the clipping from the computer and play it or in natural way by speaker utterances we can record, then click on the Record option and stop it. After recording, the file has to be saved in the computer, just click on the File menu in that Export option is there select it and save it in the prescribed format (.wav) for simplifying the process. The recording may be different length of Speech Signal, from this every Speech file recorded could be cut down into below 10 sec. Because short duration Speech Utterances only used for doing Speech Processing tasks. In this every Speech Sample is recorded with a sample rate of 44100Hz for good quality. Then Pre-Processing of the recorded file has to be takes place. Choose the recorded audio and click on the "Noise Removal" button in Figure 4 to begin the first step of noise removal. In the next step, choose all of the noise you wanted to remove and click the "Clear Noise" button.

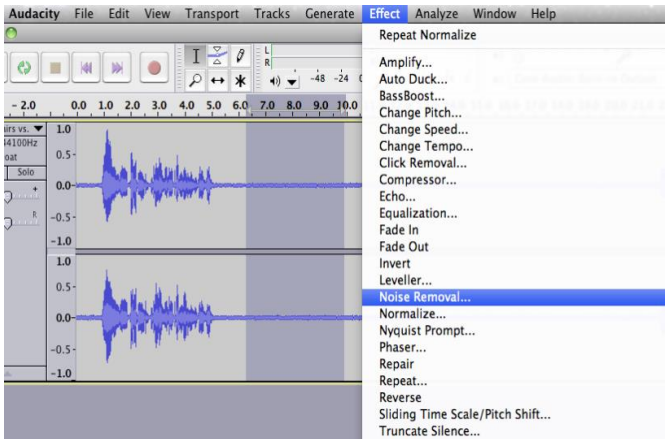

Figure 4. Noise removal in audacity software

Table 3. Validation of DETL database

Speech Files(.wav)		Validation (Given to Professionals) Total Professionals	Professionals Correctly given
Happy	100	10	8
Sad	100	10	10
Anger	100	10	7
Surprise	100	10	9
Neutral	100	10	7
Total		**500**	

In this Database, there are five emotions mainly Happy, Sad, Anger, Surprise and Neutral. Here every Emotion contains 100 Speech files. Overall 100 Speech Utterances * 5 Emotions, total 500 speech files (.wav) with different emotions created. Each Speech file name is manually typed and labelling, because to design the system with easy identification and Emotion Recognition of Speech. After this DETL Database is given to the Professionals for the validation of correct emotions as shown in Table 3.

2.3 Feature extraction

Features are very important in any pattern recognition problem. Feature is a major prominent characteristics data that reflects the whole data. Features do a vital role in Speech Emotion Recognition Systems (SER), a good feature subset will give better accuracy. In the Speech Processing, features along with the statistical values like min, max, mean, standard deviation of speech utterances are taken consideration. Mainly three categories of Features are used in Speech Processing. They are: i) Prosodic Features ii) Spectral Features and iii) Hybrid Features.

2.3.1 Prosodic features

These are also called as acoustic features or Excitation Features. Prosodic features are Human beings perception and carrying paralinguistic information, generally these are in supra-segmental in nature. The basic Prosodic Features are:

(a) Fundamental Frequency (F0): It is the Frequency of vocal tract vibrations. Vocal tract folds generate oscillations.

F0 = Avg (no.of oscillations / Sec)

(b) Pitch: It is the lowness or highness of voice tone. It is very much related to F0 in human being perception.

$$\text{Pitch Period}(m) = \arg_m \max R_m(\tau),\ m=1, 2, M \qquad (1)$$

Pitch is calculated from Eq. (1) as

$$F0 = f_s / \textit{pitch period}\ (\tau) \qquad (2)$$

where, f_s represents the frequency sampling.

(c) Zero-Crossing Rate (ZCR): ZCR is the rate at which Signal variations it's sign. While a signal marks the zero axes range of times, it specifies the occurrence of higher frequency contents during this signal, i.e., the signal oscillate earlier.

Generally, emotions of speech signal are frequency dependent, hence, the ZCR will give the stated emotional information. The short-time average ZCR are often stated as:

$$z(m) = \sum_{n=-\infty}^{\infty} \left\|\text{sgn}(s(n) - \text{sgn}(s(n-1))\right\| w(m-n) \qquad (3)$$

$$\text{sgn}(s(m)) = \begin{cases} 1 & if\ \ s(m) \geq 0 \\ -1 & if\ \ s(m) \prec 0 \end{cases} \qquad (4)$$

where, s(m) is that the windowed speech signal.

(d) Short-Time Energy (STE): The energy of speech signal offers amplitude variations and unveil the loudness of signal. As human beings, emotions are completely various arousal levels, the signal owns totally different levels of energy and amplitude in an emotional speech utterance. For an indication s(n), STE is specified by

$$\text{EST}(m) = \sum_{n=-\infty}^{\infty} \left\|s(n)w(m-n)\right\|^2 \qquad (5)$$

where, w(m) is that the analyzing window.

Nicholson et al. [46] used pitch and LPCC features in their work to recognize the emotions and the proposed SER got average accuracy results.

Rao et al. [47] proposed a work with LP residual energy as features in the Speech Emotion Recognition process. Yegnanarayana et al. also worked in the environment of multi speaker and used LP residual Features in SER system.

Busso et al. [31] presented a work with Fundamental Frequency (F0) contour with its statistical values mean,max,min to identify the emotions in speech. By their experimentation it is observed that sentence level features surpass from voice region.

Koolagudi et al. [55] worked on IIT-KGP-SESC emotional database to recognize the 6 categories of emotions. The work used epoch parameters of LP residuals and Zero Frequency Filtered Speech Signal.

Sun and Moore et al. [56] presented a work with parameters of glottal wave forms to identify the binary class emotions from four class.

Rambabu et al. [54] reported a work on Emotion Recognition by using Excitation Features like F0, Strength of Excitation (SoE). The work used KL distance and IIITH-Telugu emotion Database to recognize four class emotions.

2.3.2 Spectral features

Human Speech sound is filtered by vocal tract shape. Speech can be determined by this shape. Generally Vocal tract characteristics are represented in the frequency domain, Fourier Transformation can change Speech signal from time domain into frequency domain. Spectral Features are generally

LPCC, MFCC, ΔMFCC, ΔΔMFCC, LFPC, etc.

(a) LPCC (Linear predictive cepstral coefficients): These features also contain speaker's vocal tract characteristics used to differentiate emotions in Speech. LPCC is defined from LPC; here Linear Prediction Coefficients are chosen to minimize the residual error.

(b) MFCC (Mel frequency cepstral coefficients): MFCC (Mel Frequency Cepstral Coefficient) features are used in recent research problems of Speech Processing, because it mostly resembles the human auditory system. In MFCC firstly we take input speech signal, then apply the steps which are Pre-emphasis, Framing, Windowing, FFT, Filter Bank and Frequency Wrapping and finally apply Logarithmic to the spectrum. Mel-Frequency calculated by using the formulae:

$$Mel(f)=2595*Log10(1+f/700) \quad (6)$$

where, f - frequency in Hz.

(c) Delta1 MFCC (or) Δ MFCC: It is the first order derivative of the MFCC with respect to Time domain.

(d) Delta2 MFCC (or) Δ Δ MFCC: It is the second order derivative of the MFCC with respect to Time domain.

MFCC Spectral Features [57] are also used to find the facial expression.

Lee et al. [58] reported a work with only MFCC features based Emotion Recognition system, the work used phoneme level process used for recognizing emotions.

Sato and Obuchi [59] worked with segmented MFCC features to identify the emotions from speech. In their work, each frame is labeled with clustering of MFCC multi templated. The work compared with prosody, but their new approach gave better results.

Nwe et al. [60] presented a work on Speech Emotion Recognition with LFPC, MFCC and LPCC of text independent emotion classification for mandarin and Burmese languages.

Koolagudi et al. [48] proposed a work to recognize the emotions, the features used in SER is MFCC only and IIT-KGP-SESC semi natural Telugu language database used.

2.3.3 Hybrid features

In this combination of Features are used, like prosodic features combination (pitch+Energy+ZCR), spectral features combination (LPCC+MFCC or LPCC+MFCC+ΔMFCC+ΔΔMFCC) and prosodic features + spectral features combinations (pitch+Energy+ZCR+MFCC+ΔMFCC), etc. to improve the accuracy or minimize the error rate of the System.

2.4 Classification models

Most of the researchers considered Speech Emotion Recognition (SER) system as a Machine Learning (ML) problem, because ML are data driven solutions. In the SER, generally Speech Signals are in the discrete time signals form that contains much information and used for further process in SER model. The primary and foremost SER objective is to find emotions or patterns from shortest duration of speech utterances.

Classification is a two-step process. i) Training Phase and ii) Testing Phase. In the Training Phase, Input data is trained based on the Classification Algorithm model, after that in the Testing Phase generated reference model from Training Phase is tested with testing samples to classify the emotions. Various Classification Models used in SER is shown in Figure 5.

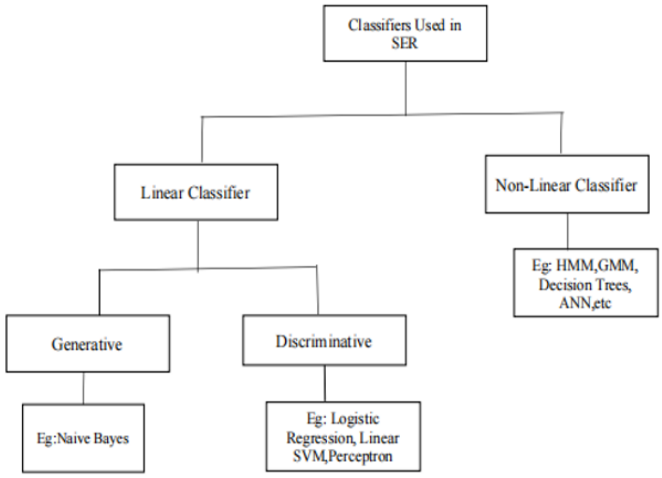

Figure 5. General classifiers used in SER

2.4.1 Hidden Markov Model (HMM)

It is a very popular stochastic model and successful in Speech processing, because HMM is very suitable to sequences data. General Structure of HMM model is shown in Figure 6. HMM contains mainly two steps a) Transition Model, which is used to change state transition and b) Observation Model, which is decides the given hidden sequence pattern. It depends on Markov chain property in which the future possible event completely depends on the current possible event, but not on the previous event.

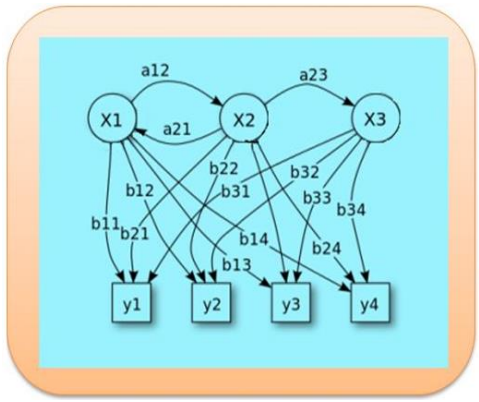

Figure 6. General structure of HMM model

HMM as an Evaluation Problem, which is used in testing phase of any pattern recognition application. It estimates the probability of observation sequence (O) against given Hidden Markov Model.

HMM as a learning problem, used in training phase of Pattern recognition including Speech Processing model and Speech Emotion Recognition system. It gives a reference model by using given an HMM model λ and sequence of inputs.

In SER Problems, generally feature vectors are extracted from Telugu speech utterances of different emotions and HMM is design with the Learning Problem mentioned above. In testing phase of the system, those feature vectors are extracted from unknown utterances of Telugu speech and evaluated against HMM of emotions of observation sequence is calculated using evaluation problem of HMM.

Observations:

- It uses either left to right or fully connected topology network system.

- It is very hard to decide the number of states that are optimal in SER.
- HMM breaks every speech utterance into a small frame sequence of phonemes clearly.

2.4.2 Gaussian mixture model (GMM)

It is one of the major salient models used by many researchers in the domain of ASR systems. It is also a stochastic model and one of the special mixture models and mainly uses the concept of Probability Distribution Estimation (PDE). EM Algorithm is used to train GMM model.

E(Expectation) Step: Used to find the probability that each point d_i belongs, distributions C1, C2, ..., Cs.

M(Maximization) Step: By considering the results of E-step update the μ, σ, π values.

In SER, extracted features from Telugu speech utterances of different emotions and GMM model is trained. In testing phase of the system, those extracted feature vectors from unknown utterances of Telugu speech and evaluated against GMM of emotions. Using the EM algorithm maximum probability is calculated.

Observations:
- GMM can't be used to temporal structure data.
- Checking whether the number of Gaussian elements that are optimal is a big challenge.

2.4.3 Vector quantization (VQ)

VQ model maps n-dimensional vector space to a finite set Code Book CB = {C1, C2, C3, ..., CN}. Generally, Code Book consists of N number of code vectors. The key to VQ is the good code book. The most common method used to generate code book is the Linde-BuzoGray (LBG) algorithm. Speech Emotion Recognition system using VQ is considered as a Classification Problem, can be done in two Phases a) Training and b) Testing.

Training Phase: In this Phase, Features training is done with feature vectors of the input speech samples and performs code book training with the help of the LBG vector quantization (VQ) algorithm.

Testing Phase: In this Phase, Feature matching or testing is done with a matching score. Generally, it is a Similarity measure of the extracted features from the unknown Speech Utterance and the stored Utterances in the Code book. The unknown utterance is determined by the minimum matching value in the Code book database.

Observations:
- VQ can't be good if there are a greater number of Speakers and for Multi-Language Database.
- Difficult to design good VQ code book from large number of Speakers and different ages.

2.4.4 Support vector machines (SVM)

Speech Utterances data points or features can be plotted on n-dimensional space. To recognize the emotions (support vectors) used linear separable plane or hyper plane. Mathematical equation for linear mode is

$$f(x,w)=w^{T}.x+b \quad (7)$$

For finding the multi class problems in SVM, kernel functions may be used, i.e., linear kernel, RBF kernel or polynomial kernel.

Observations:
- Choosing kernel functions is a complicated task to identify the best features.
- To avoid over fitting, taking complete data is not always recommended.

2.4.5 Deep learning methods

Now a days many researchers are working on Deep Learning techniques in ASR, because Human beings perception is same as Deep Learning hence every domain is replaced with this automation. Deep Learning Techniques mainly includes Multi-Layer Perceptron (MLP), Recurrent Neural Networks (RNN) and Convolution Neural Network (CNN). The DNN is more popular because it is able to extract even complex features from large amounts of data even from hours of speech data both Linear as well as non-linear because of the hidden layer architecture as shown in Figure 7.

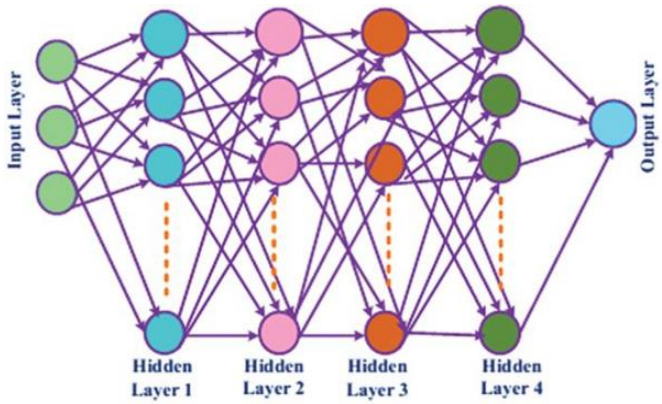

Figure 7. Basic DNN structure

Observations:
- SER implementation can be done using Keras and Tensor flows is a complex task.
- Many Tensors summed up with back propagation and all may consume lot of memory space and time also, hence GPU is required.
- Features sets are increased when using Deep Learning methods.

Nogueiras et al. [16] presented a work using prosodic features like pitch, energy and their contour values with HMM model to recognize the emotions and the system got 70% recognition rate.

Schuller et al. [61] used two methods HMM and GMM with a comparison study. The work used prosodic features pitch, energy and contour values. In this GMM used global values that derived from basic prosody features and HMM used low level features. The recognition rate of the system is 86% and 79% respectively for GMM and HMM.

Nwe et al. [60] used HMM to recognize the emotions form the Speech utterances. The system used MFCC, LPCC and LFPC. The accuracy is calculated for above three features and got good results for LFPC features other than two.

Neiberg et al. [30] presented a work using GMM model with MFCC and MFCC-low features to recognize the emotions. In this GMM model is trained by using EM (Expectation Maximization) technique. It is noticed from the experimentation that combination features gave better results.

Kwon et al. [62] worked on prosodic and spectral features like pitch, energy, MFCC with Gaussian Kernal Function SVM(GSVM), Linear SVM and HMM on their database. It is

observed that the system accuracy is good for GSVM and HMM.

Pan et al. [63] reported a work on SER with LPCC, MFCC and MEDC using SVM classifier on Berlin Emo-DB and their own Chinese Database. From their experimentation it is observed that the system accuracy is better for MFCC and MEDC combination Features.

Qayyum et al. [64] presented a work on SAVEE Database with MFCC, MFCC+MS using SVM and CNN classifiers. The experimentation results showed that CNN gives better 83% accuracy and SVM got 76% recognition rate.

Koduru et al. [65] proposed a work to recognize the emotions of RAVDESS Database using Hybrid combination features MFCC, Pitch, Energy, ZCR and DWT using LDA, SVM and Decision Tree. The accuracy of the system is 65%, 70% and 85% respectively.

Parikh et al. [66] worked on Deep CNN to recognize the emotions of RAVDESS database using hybrid combination features MFCC, Spectral Contrast Features. The recognition rate of the SER is 71%.

In Table 4, Various Classification models used in Speech Emotion Recognition with References, in Table 5, Combination of Classification models in SER described and shown below.

Table 4. Various classification models used in SER with references

Classifiers	References
HMM	Bitouk et al. [67], Fernandez and Picard [68], Zhou et al. [69], Schuller et al. [61], Lee et al. [58]
GMM	Breazeal et al. [70], Slaney and McRoberts [71], Mubarak et al. [72], Jeon et al. [73], Lugger and Yang [74]
K-NN	Wang and Guan [75], Dellaert et al. [76], Pao et al. [77], Yu et al. [25], Petrushin [78]
SVM	Kwon et al. [62], Lee et al. [58], Schuller et al. [61], Sun and Moore [56], Hu et al. [79]
DNN	Chen et al. [80], Mirsamadi et al. [81], Neumann and Vu [82], Parthasarathy and Tashev [83], Koduru et al. [65], Parikh et al. [66]

Table 5. Combination of classification models in SER

Reference	Classifier Used	Accuracy Rate
Yu et al. [18]	SVM, ANN	71%,42%
El Ayadi et al. [84]	ANN, HMM	55%,71%
He et al. [85]	GMM, KNN	77%,77%
Li et al. [86]	DNN-HMM, RBM	74.28%
Chernykh and Prikhodko [87]	LSTM with loss function	71.08%
Zhao et al. [88]	2D+Conv Layer+LSTM	76.64%
Fahad et al. [89]	DNN-HMM	65.93%

2.5 Challenges in SER

- Speech Emotion Recognition is an uncertain problem, why because it is subjective, behaviour and expression vary from person to person.
- SER is always challenging, because there is a huge gap between low level features and Subjective features of Speech.
- No standard database is available in market.
- Cross-Linguistic Data recognition rate is low.
- There is a high accuracy rate between low-arousal and high-arousal emotions.
- Using shortest duration, Performance of the system is improved.
- Even increase the database size, the performance may be degraded.

3. RESEARCH PROBLEM MOTIVATION

Digital electronics and computer systems are advancing rapidly, making it easier to create and apply speech processing algorithms for improved Human-Computer Interaction (HCI). Deep learning and machine learning are now being integrated into this region. Any language's statistics can be extracted via the field of speech recognition. Digitization, the process of breaking down a spoken utterance into individual words or groups of words, is how this is accomplished. Researchers in the domain of recognition and identification consider speech to be a reliable and essential biometric trait. Compared to other existing structures like scanning a unique fingerprint or a retina of the any human being, the facts acquisition method for speech is easy and straightforward.

Even if humans aren't participating, a high-fantastic cell phone or a phone network can be used if the recording system is of high quality. Dialect/accent refers to a community's unique method of speaking/pronouncing a language. In the states of Andhra Pradesh and Telangana, there are primarily three distinct accents to be found. When the state is divided, the speaker's focus shifts to a specific area, and this can be detected by the usage of the speaker's accent or dialect. Automatic Speech Recognition (ASR) systems could benefit from a better understanding of how speech sounds, words, and sentences are structured. Identifying the speaker's place of birth is essential for locating him inside the state. Similarly, after locating the speaker to a certain area, the algorithm can identify the voice more precisely. The voice and speaker identification systems face a number of obstacles, which inspired the current work. The dialect, accent, and socioeconomic status of a person's upbringing all play a role in how he or she speaks. To model speaker-independent structures that can accept input from any language variation, it is necessary to account for changes in the characters used in each language. Emotion, dialect or accent can be determined mechanically from a sample speech supplied in this work. In the absence of a modern database for Telugu talks [9-12], laboratory settings were used to increase the training set and also the testing set. In addition to making, it easier to understand the speech's substance, this will simplify the process of processing speech. Emotions are also taken into account when determining the speech and speaker in this study, based on speaker and speech recognition. The essence of Telugu speech is seen as crucial because of all these difficulties in the suggested job.

4. RESEARCH PROPOSAL OBJECTIVES

Emotions are employed in this study to identify and classify Telugu speech samples and the speakers in them. The following is a list of the goals of this research work.

- To recognize Telugu speech samples based on their emotional content, create algorithms to do so.

- A database of Telugu speech of emotions like happy, sad, angry, and neutral for training and testing purposes.
- Determining emotions from Speech Utterances of Short duration may be discerned in brief statements.
- Techniques for designing high-recognition systems.
- Use the Text-independent data to identify the speaker's psychological qualities.

5. CONCLUSIONS

Speech production and basic concepts are the focus of this study. Research in the field of speech processing is also discussed in terms of its significance, difficulties, and sources of inspiration. Concepts about accent and emotion identification are thoroughly covered in this article. For the time being, we'll focus on the work done so far in the Telugu emotional speech database. The use of Audacity software to create the DETL (Database for Emotions in Telugu Language) is described in detail. Speech processing has also been included in this list. There is also a discussion of the goals and the numerous tasks involved in achieving these goals.

REFERENCES

[1] Lawrence, R., Juang, B.H. (1993) Fundamentals of Speech Recognition. Pearson Edition.

[2] Sakoe, H., Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1): 43-49. https://doi.org/10.1109/TASSP.1978.1163055

[3] Myers, C., Rabiner, L. (1981). A level building dynamic time warping algorithm for connected word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(2): 284-297. https://doi.org/10.1109/TASSP.1981.1163527

[4] de Veth, J., Gallopyn, G., Bourlard, H. (1993). Limited parameter HMMs for connected digit speaker verification over telephone channels. IEEE Trans. Acoust., Speech, Signal Processing, pp. 247-250.

[5] Gangamohan, P., Kadiri, S.R., Yegnanarayana, B. (2013). Analysis of emotional speech at subsegmental level. INTERSPEECH, 2013: 1916-1920.

[6] Forsyth, M. (1995). Discriminating observation probability (DOP) HMM for speaker verification. Speech Communication, 17(1-2): 117-129. https://doi.org/10.1016/0167-6393(95)00020-O

[7] Houari, H., Guerti, M. (2020). Study the influence of gender and age in recognition of emotions from Algerian dialect speech. Traitement du Signal, 37(3): 413-423. https://doi.org/10.18280/ts.370308

[8] Bezooijen, R.A.M.G. (1984). Characteristics and Recognizability of Vocal Expressions of Emotion. Foris Publications. https://doi.org/10.1515/9783110850390

[9] Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P. (1997). Design, recording and verification of a Danish emotional speech database. Fifth European Conference on Speech Communication and Technology, Rhodes, Greece, pp. 1695-1698.

[10] Hansen, J.H.L. (1999). Philadelphia: Linguistic Data Consortium. SUSAS LDC99S78. Web Download. https://doi.org/10.35111/x4at-ff87

[11] Paeschke, A., Kienast, M., Sendlmeir, W.F. (1999). F_0-contours in emotional speech. Proceedings of the ICPHs in San Francisco, pp. 929-932.

[12] Li, Y., Zhao, Y. (1998). Recognizing emotions in speech using short-term and long-term features. Fifth International Conference on Spoken Language Processing, Sydney, Australia, pp. 2255-2258.

[13] Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Enriquez, E., Pardo, J.M. (1999). Analysis and modelling of emotional speech in Spanish. Proc. of ICPhS, 2: 957-960.

[14] McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: A rough benchmark. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, pp. 207-212.

[15] Ambrus, D.C. (2000). Collecting and recording of an emotional speech database technical Report. University of Maribor.

[16] Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J.B. (2001). Speech emotion recognition using hidden Markov models. Seventh European Conference on Speech Communication and Technology.

[17] Alpert, M., Pouget, E.R., Silva, R.R. (2001). Reflections of depression in acoustic measures of the patient's speech. Journal of Affective Disorders, 66(1): 59-69. https://doi.org/10.1016/S0165-0327(00)00335-9

[18] Yu, F., Chang, E., Xu, Y.Q., Shum, H.Y. (2001). Emotion detection from speech to enrich multimedia content. Pacific-Rim Conference on Multimedia, Beijing, China, pp. 550-557. https://doi.org/10.1007/3-540-45453-5_71

[19] Makarova, V., Petrushin, V.A. (2002). RUSLANA: A database of Russian emotional utterances. Seventh International Conference on Spoken Language Processing, Denver, Colorado, USA.

[20] Yuan, J., Shen, L., Chen, F. (2002). The acoustic realization of anger, fear, joy and sadness in Chinese. Seventh International Conference on Spoken Language Processing, Denver, Colorado, USA, pp. 2025-2028.

[21] Pierre-Yves, O. (2003). The production and recognition of emotions in speech: Features and algorithms. International Journal of Human-Computer Studies, 59(1-2): 157-183. https://doi.org/10.1016/S1071-5819(02)00141-6

[22] Cowie, R., Cornelius, R.R. (2003). Describing the emotional states that are expressed in speech. Speech Communication, 40(1-2): 5-32. https://doi.org/10.1016/S0167-6393(02)00071-7

[23] Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. Speech Communication, 40(1-2): 161-187. https://doi.org/10.1016/S0167-6393(02)00081-X

[24] Fakotakis, N. (2004). Corpus design, recording and phonetic analysis of Greek emotional database. International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, pp. 1391-1394.

[25] Nordstrand, M., Svanfeldt, G., Granström, B., House, D. (2004). Measurements of articulatory variation in expressive speech for a set of Swedish vowels. Speech Communication, 44(1-4): 187-196. https://doi.org/10.1016/j.specom.2004.09.003

[26] Caldognetto, E.M., Cosi, P., Drioli, C., Tisato, G., Cavicchio, F. (2004). Modifications of phonetic labial

targets in emotive speech: effects of the co-production of speech and emotions. Speech Communication, 44(1-4): 173-185. https://doi.org/10.1016/j.specom.2004.10.012

[27] Kim, E.H., Hyun, K.H., Kwak, Y.K. (2005). Robust emotion recognition feature, frequency range of meaningful signal. ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, Nashville, TN, USA, pp. 667-671. https://doi.org/10.1109/ROMAN.2005.1513856

[28] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. (2005). A database of German emotional speech. In INTERSPEECH.

[29] Batliner, A., Biersack, S., Steidl, S. (2006). The prosody of pet robot directed speech: Evidence from Children. Proc. of Speech Prosody 2006, Dresden, pp. 1-4.

[30] Neiberg, D., Elenius, K., Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. International Conference on Speech and Language Processing (ICSLP), Pittsburgh, USA, pp. 809-812.

[31] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Lang Resources & Evaluation, 42: 335. https://doi.org/10.1007/s10579-008-9076-6

[32] Grimm, M., Kroschel, K., Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, pp. 865-868. https://doi.org/10.1109/ICME.2008.4607572

[33] Chauhan, A., Koolagudi, S.G., Kafley, S., Rao, K.S. (2010). Emotion recognition using LP residual. 2010 IEEE Students Technology Symposium (TechSym), Kharagpur, India, pp. 255-261. https://doi.org/10.1109/TECHSYM.2010.5469162

[34] Pichora-Fuller, M.K., Dupuis, K. (2020). Toronto emotional speech set (TESS). Scholars Portal Dataverse. https://doi.org/10.5683/SP2/E8H2MF

[35] Rao, K.S., Koolagudi, S.G. (2011). Identification of Hindi dialects and emotions using spectral and prosodic features of speech. IJSCI: International Journal of Systemics, Cybernetics and Informatics, 9(4): 24-33.

[36] Jackson, P., Haq, S. (2014). Surrey audio-visual expressed emotion (SAVEE) database. University of Surrey: Guildford, UK.

[37] Moriyama, T., Mori, S., Ozawa, S. (2009). A synthesis method of emotional speech using subspace constraints in prosody. Journal of Information Processing, 50(3): 1181-1191.

[38] Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, pp. 1-8. https://doi.org/10.1109/FG.2013.6553805

[39] Li, Y., Tao, J., Chao, L., Bao, W., Liu, Y. (2017). CHEAVD: A Chinese natural emotional audio–visual database. Journal of Ambient Intelligence and Humanized Computing, 8(6): 913-924. https://doi.org/10.1007/s12652-016-0406-z

[40] Quiros-Ramirez, M.A., Polikovsky, S., Kameda, Y., Onisawa, T. (2014). A spontaneous cross-cultural emotion database: Latin-America vs. Japan. KEER2014. Proceedings of the 5th Kanesi Engineering and Emotion Research; International Conference, Linköping, Sweden, pp. 1127-1134.

[41] Esmaileyan, Z., Marvi, H. (2014). A database for automatic Persian speech emotion recognition: Collection, processing and evaluation. IJE Transactions A: Bascis, 27(1): 79-90.

[42] Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., Di Natale, C. (2014). Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowledge-Based Systems, 63: 68-81. https://doi.org/10.1016/j.knosys.2014.03.019

[43] Song, P., Ou, S., Zheng, W., Jin, Y., Zhao, L. (2016). Speech emotion recognition using transfer non-negative matrix factorization. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 5180-5184. https://doi.org/10.1109/ICASSP.2016.7472665

[44] Livingstone, S.R., Russo, F.A. (2018) 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, 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

[45] Nezami, O.M., Lou, P.J., Karami, M. (2019). ShEMO: A large-scale validated database for Persian speech emotion detection. Language Resources and Evaluation, 53(1): 1-16. https://doi.org/10.1007/s10579-018-9427-x

[46] Nicholson, J., Takahashi, K., Nakatsu, R. (2000). Emotion recognition in speech using neural networks. Neural Computing & Applications, 9(4): 290-296. https://doi.org/10.1007/s005210070006

[47] Rao, K.S., Prasanna, S.R.M., Yegnanarayana, B. (2007). Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. IEEE Signal Processing Letters, 14: 762-765. https://doi.org/10.1109/LSP.2007.896454

[48] Koolagudi, S.G., Maity, S., Kumar, V.A., Chakrabarti, S., Rao, K.S. (2009). IITKGP-SESC: Speech database for emotion analysis. International Conference on Contemporary Computing, Noida, India, pp. 485-492. https://doi.org/10.1007/978-3-642-03547-0_46

[49] Kadiri, S.R., Gangamohan, P., Gangashetty, S.V., Yegnanarayana, B. (2015). Analysis of excitation source features of speech for emotion recognition. In Sixteenth Annual Conference of the International Speech Communication Association, pp. 1324-1328.

[50] Vijai Bhaskar, P., Ramamohana Rao, S. (2013). Emotional Telugu speech signals classification based on k-NN classifier. International Journal of Research in Engineering and Technology, 2(1): 85-93.

[51] Prasad Reddy, P.V.G.D., Prasad, A., Srinivas, Y., Brahmaiah, P. (2010). Gender based emotion recognition system for Telugu rural dialects using hidden Markov models. Journal of Computing, 2(6): 94-98.

[52] Pravena, D., Govind, D. (2017). Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. International Journal of Speech Technology, 20(4): 787-797. https://doi.org/10.1007/s10772-017-9445-x

[53] Ram, C.S., Ponnusamy, R. (2018). Toward design and enhancement of emotion recognition system through speech signals of autism spectrum disorder children for

Tamil language using multi-support vector machine. Proceedings of International Conference on Computational Intelligence and Data Engineering, pp. 145-158. https://doi.org/10.1007/978-981-10-6319-0_13

[54] Rambabu, B., Botsa, K.K., Paidi, G., Gangashetty, S.V. (2020). IIIT-H TEMD semi-natural emotional speech database from professional actors and non-actors. Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, pp. 1538-1545.

[55] Koolagudi, S.G., Rao, K.S. (2012). Emotion recognition from speech using source, system, and prosodic features. International Journal of Speech Technology, 15(2): 265-289. https://doi.org/10.1007/s10772-012-9139-3

[56] Sun, R., Moore, E. (2011). Investigating glottal parameters and teager energy operators in emotion recognition. International Conference on Affective Computing and Intelligent Interaction, Memphis, TN, USA, pp. 425-434. https://doi.org/10.1007/978-3-642-24571-8_54

[57] Demircan, S., Örnek, H.K. (2020). Comparison of the effects of Mel coefficients and spectrogram images via deep learning in emotion classification. Traitement du Signal, 37(1): 51-57. https://doi.org/10.18280/ts.370107

[58] Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S, Narayanan, S. (2004). Emotion recognition based on phoneme classes. Eighth International Conference on Spoken Language Processing.

[59] Sato, N., Obuchi, Y. (2007). Emotion recognition using Mel-frequency cepstral coefficients. Information and Media Technologies, 2(3): 835-848. https://doi.org/10.11185/imt.2.835

[60] Nwe, T.L., Foo, S.W., De Silva, L.C. (2003). Speech emotion recognition using hidden Markov models. Speech Communication, 41(4): 603-623. https://doi.org/10.1016/S0167-6393(03)00099-2

[61] Schuller, B., Rigoll, G., Lang, M. (2003). Hidden Markov model-based speech emotion recognition. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03), Hong Kong, China. https://doi.org/10.1109/ICASSP.2003.1202279

[62] Kwon, O.W., Chan, K., Hao, J., Lee, T.W. (2003). Emotion recognition by speech signals. Eighth European Conference on Speech Communication and Technology, pp. 125-128.

[63] Pan, Y., Shen, P., Shen, L. (2021). Speech emotion recognition using support vector machine. International Journal of Smart Home, 6(2):101-108.

[64] Qayyum, A.B.A., Arefeen, A., Shahnaz, C. (2019). Convolutional neural network (CNN) based speech-emotion recognition. 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), Dhaka, Bangladesh, pp. 122-125. https://doi.org/10.1109/SPICSCON48833.2019.9065172

[65] Koduru, A., Valiveti, H.B., Budati, A.K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, 23(1): 45-55. https://doi.org/10.1007/s10772-020-09672-4

[66] Parikh, N., Mistry, K., Bhavsar, Y., Hakimi, A., Magare, A. (2021). Real time speech emotion recognition using machine learning. International Research Journal of Engineering and Technology (IRJET), 8(6): 2786-2792.

[67] Bitouk, D., Verma, R., Nenkova, A. (2010). Class-level spectral features for emotion recognition. Speech Communication, 52(7-8): 613-625. https://doi.org/10.1016/j.specom.2010.02.010

[68] Fernandez, R., Picard, R.W. (2003). Modeling drivers' speech under stress. Speech Communication, 40(1-2): 145-159. https://doi.org/10.1016/S0167-6393(02)00080-8

[69] Zhou, G., Hansen, J.H., Kaiser, J.F. (2001). Nonlinear feature based classification of speech under stress. IEEE Transactions on Speech and Audio Processing, 9(3): 201-216. https://doi.org/10.1109/89.905995

[70] Breazeal, C., Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. Autonomous Robots, 12(1): 83-104. https://doi.org/10.1023/A:1013215010749

[71] Slaney, M., McRoberts, G. (2003). Babyears: A recognition system for affective vocalizations. Speech Communication, 39: 367-384. https://doi.org/10.1016/S0167-6393(02)00049-3

[72] Mubarak, O.M., Ambikairajah, E., Epps, J. (2005). Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources. In: Proceedings of the 8th International Symposium on Signal Processing and Its Applications, 2: 619-622. https://doi.org/10.1109/ISSPA.2005.1581014

[73] Jeon, J.H., Xia, R., Liu, Y. (2011). Sentence level emotion recognition based on decisions from subsentence segments. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, pp. 4940-4943. https://doi.org/10.1109/ICASSP.2011.5947464

[74] Lugger, M., Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, USA, pp. IV-17-IV-20. https://doi.org/10.1109/ICASSP.2007.367152

[75] Wang, Y., Guan, L. (2004). An investigation of speech-based human emotion recognition. IEEE 6th Workshop on Multimedia Signal Processing, Siena, Italy, pp. 15-18. https://doi.org/10.1109/MMSP.2004.1436403

[76] Dellaert, F., Polzin, T., Waibel, A. (1996). Recognizing emotion in speech. Proceedings of the 4th International Conference on Spoken Language, 3: 1970-1973. https://doi.org/10.1109/ICSLP.1996.608022

[77] Pao, T.L., Chen, Y.T., Yeh, J.H., Liao, W.Y. (2005). Combining acoustic features for improved emotion recognition in mandarin speech. International Conference on Affective Computing and Intelligent Interaction, Beijing, China, pp. 279-285. https://doi.org/10.1007/11573548_36

[78] Petrushin, V.A. (2000). Emotion recognition in speech signal: Experimental study, development, and application. In Sixth International Conference on Spoken Language Processing, Beijing, China.

[79] Hu, H., Xu, M., Wu, W. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP'07, vol. Honolulu, HI, USA, pp. IV-413-IV-416.

https://doi.org/10.1109/ICASSP.2007.366937
[80] Chen, M., He, X., Yang, J., Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, 25(10): 1440-1444. https://doi.org/10.1109/LSP.2018.2860246
[81] Mirsamadi, S., Barsoum, E., Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, pp. 2227-2231. https://doi.org/10.1109/ICASSP.2017.7952552
[82] Neumann, M., Vu, N.T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. arXiv preprint arXiv:1706.00612.
[83] Parthasarathy, S., Tashev, I. (2018). Convolutional neural network techniques for speech emotion recognition. 2018 16th international workshop on acoustic signal enhancement (IWAENC), Tokyo, Japan, pp. 121-125. https://doi.org/10.1109/IWAENC.2018.8521333
[84] El Ayadi, M.M.H., Kamel, M.S., Karray, F. (2007). Speech emotion recognition using Gaussian mixture vector autoregressive models. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007, pp. IV-957-IV-960. https://doi.org/10.1109/ICASSP.2007.367230
[85] He, L., Lech, M., Maddage, N.C., Allen, N.B. (2011) Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. Biomedical Signal Processing and Control, 6(2): 139-146. https://doi.org/10.1016/j.bspc.2010.11.001
[86] Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E. (2013). Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) based speech emotion recognition. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 312-317. https://doi.org/10.1109/ACII.2013.58
[87] Chernykh, V., Prikhodko, P. (2017). Emotion recognition from speech with recurrent neural network. https://arxiv.org/abs/1701.08071.
[88] Zhao, J., Mao, X., Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47: 312-323. https://doi.org/10.1016/j.bspc.2018.08.035
[89] Fahad, M.S., Deepak, A., Pradhan, G., Yadav, J. (2021). DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features. Circuits, Systems, and Signal Processing, 40: 466-489. https://doi.org/10.1007/s00034-020-01486-8