# Review of Performance Factors of Emotional Speaker Recognition System: Features, Feature Extraction Approaches and Databases

[1]Satish Kumar Das, [2]Uttpal Bhattacharjee, [3]Amit Kumar Mandal

•

Department of Computer Science & Engineering
Rajiv Gandhi University, Arunachal Pradesh, India
[1]satish.das@rgu.ac.in
[2]utpal.bhattacharjee@rgu.ac.in
[3]amitkumar.tu@gmail.com

**Abstract**

*Emotion is a conscious mental reaction accompanied by physiological and behavior changes in human body.In speaker authentication system, emotional state of the speaker plays a vital role. Recently, the field of speaker recognition in emotional context attracts more and more attention of many research focuses. However, to implement more realistic and intelligent emotional speaker recognition system it is interesting to study this system under real life conditions. Speech emotion recognition is a system in which speech signals are processed to classify the embedded emotions. In recent past, speaker emotion recognition has gained a lot of attention from different researchers as it has many applications. In this regards, study of prior works is useful for further research in the field of speaker verification in emotional context. So, performance and reliability of Emotional Speaker Recognition System depend on the proper selection of features to characterize different emotional states, feature extraction approaches and databases. In this paper we briefly discuss about different features, feature extraction approaches and emotion recognition and speaker verification databases.*

**Keywords:** speech emotion recognition, features, feature extraction approaches, databases

## I. Introduction:

The human voice is used to expresses the emotions. The speech has potential to be an essential mode of interaction with the computer. Recognition of emotion from human speech is an active research area in signal processing. This is due to the enormous possibilities in the field of human-machine interaction [1].Speech emotion recognition is using in different fields from medical scienceto business, from entertainment to interactive systems.

A speech signal is a logical arrangement of sounds from which our brain acquires the information and knowledge. But how a machine interprets the human speech signals and gains information and knowledge from is in the heart of speech recognition system. The basic goal of speech recognition is to offer an interaction between a person and a machine.

In this paper we are trying to present comprehensive literature review of various features, feature extraction techniques and datasets for emotion recognition and speaker verification. In our previous paper [2], we tried to present comprehensive literature review of various classifiers for speech emotion recognition. The paper is not to compare and evaluate the works by different

authors. Moreover, we are not surveying an exhaustive list of all research works. Also we are not providing justification for choosing particular features, extraction methods and datasets for specific tasks as they can be combined differently which may yield different results.

## II. Features of Speech Emotion Recognition:

To efficiently characterize different emotions, a SER system needs to extract suitable features. Performance of classification is affected by the proper extraction and selection of features. There are different speech features for SER, but we can say which one is best. Speech features can be classified as continuous, spectral, qualitative and Teager energy operator based features. Continuous features are pitch, energy, formants; qualitative features are voice qualities such as hoarse, tense, breathy; spectral features are LPC, MFCC, LFPC and TEO-based features are TEO-FM-Var, TEO-Auto-Env, TEO-CB-Auto-Env.

### I. Continuous features:

Continuous features can be prosody features and acoustics features. According to many researchers [3, 4, 5] prosody features such as pitch and energy represents the larger portion of the emotional content of utterance. Several studies [6, 7, 8] show that the arousal state of the speaker can affect the overall energy and energy distribution across frequency spectrum. Continuous speech features such as vocal cues [9] have been used by many researchers in SER. Different researchers [5, 10, 11] grouped the acoustic features into pitch-related features, energy-related features, timing features, formants features articulation features.

Some basic global features are fundamental frequency, energy, duration, formants, etc. Different researchers [5, 9, 10, 11, 12, 13, 14, 15, 16] have studied the relationships between these speech features and the basic emotions and shown that prosodic features are good indicators of emotions. But for some researchers [17, 18], prosodic features do not have much impact on emotions.

### II. Qualitative features:

Voice quality is strongly related to emotional content of utterance [5, 19, 20]. Many researchers studied the auditory aspects of emotions and tried to define a relation [5, 10, 12]. Voice quality has the subjective impression from the contribution of different phonetic variables [10]. Cowie et al. [5] grouped the acoustic correlates which are related to the voice quality into the voice level, voice pitch, phrase, phoneme, word and feature boundaries and temporal structures.

But role of voice quality of subject in determining emotions has some confusion and problem. First, the term tense, harsh, and breathy can have different interpretations for different researchers based on their understanding [20]. According to Scherer [22], anger, joy, and fear are associated with tense voice while sadness is associated with lax voice. But for Murray [10], both anger and happiness are associated with breathy voice while resonant voice is associated with sadness. Second, it is the difficulty to automatically decide those voice qualities directly from the speech signal.

### III. Spectral features:

Spectral features have significant role in SER. Spectral signals convey the speech signal's frequency content and provide complementary information to be used in prosodic features [23]. They are selected as short term representation for speech signal with longer temporal information [24]. Emotional content of a speech signal can affect on the distribution of the spectral energy

across the spectrum of speech [25]. Utterances with happiness imply the high energy at high level of frequency; on the other hand utterances with the sadness imply low energy at the same level of frequency [9, 26].

However, there are limitations of short-term spectral features for speech recognition [27]. Conventional spectral features such as MFCCs convey only the short-term spectral properties of the speech signal, not the temporal behavior information. To overcome the limitation, Wu et al. [23] proposed long-term modulation spectral features (MSFs) using both acoustic and temporal modulation frequency components.

## IV. TEO-based features:

Teager Energy Operator is a non-liner operator, which represents the frequency and the changes of the signal amplitude which instantaneous [28]. It was initially proposed for nonlinear speech signals modeling, but later on used widely in the audio signal processing.TEO was introduced by Teager [29] who believed that speech is a function of non-linear flow of air in vocal system [30]. The fundamental frequency changes under stressful conditions. TEO of multi-frequency signals represents the individual frequency components and relationships between these components [31]. So, to detect stress in speech signal, features based on TEO can be used. Caims et al. [32] used the Teager energy of the pitch contour to classify clear, angry, loud, neutral and Lombard effects of speech. Zhou [31] proposed some other TEO-based features, such TEO-FM-Var, TEO-Auto-Env and TEO-CB-Auto-Env for detecting stress speech versus neutral speech. These features were also used to classify the stressed speech into angry, loud, and Lombard.

Proper feature selection in SER system mainly depends on the considered classification task. From review of this section, we may conclude that, continuous features like the fundamental frequency and the pitch are more preferable for high-arousal versus low-arousal emotions classification. While TEO-based features should be used for detecting stress in speech and the spectral features are the more promising for representation of speech. Features are somehow related to each other and so these features can be combined to get better classification results.

# III. Feature Extraction

Feature extraction directly implies the accuracy and the performance of the system. Feature extraction is performed to represent a speech signal by a set of predetermined components of the speech signal [33]. Infeature extraction, the speech signal is transformed to a form which is more logical but concise and reliable than the original signal. It is a process of changing the speech waveform to a parametric form for subsequent processing and analysis where data rate is relatively lesser. Some commonly used approaches are:

- Mel Frequency Cepstral Coefficients (MFCC)
- Linear Prediction Coefficients (LPC)
- Linear Prediction Cepstral Coefficients (LPCC)
- Line Spectral Frequencies (LSF)
- Discrete Wavelet Transform (DWT)

## I. Mel Frequency Cepstral Coefficients (MFCC):

MFCC is a static feature extraction approach. MFCC is replication of the human's ear system. MFCC is considered as frequency domain feature and designed to represents signal as the short-term power spectrum. It is based on discrete cosine transform of logarithm power spectrum on nonlinear Mel frequency scale [34]. The formula to calculate the mels of frequency is [35,36]:

$$Mel(f) = 2595 * \log_{10}(1 + F/100) \quad (1)$$

Where f is the frequency of signal in Hz. The MFCCis calculated using the following equation [35, 37]:

$$\hat{C}_n = \sum_{n=1}^{k} (\log \hat{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right] \quad (2)$$

where $\hat{C}_n$ is the final MFCC coefficients, $\hat{C}_n$is the output of filterbank and k is the number of mel cepstrum coefficients.

Continuous Speech →Pre-emphasis → Frame Blocking → Windowing → Fast Fourier Transform → Mel-Scale Filter Bank → Log → Discrete Cosine Transform → MFCC

**Figure 1: Steps in MFCC approach**

MFCC can effectively be used the low frequency region compared to the high frequency region. So, it can be used to compute low frequency formants and for the vocal tract resonances description [33]. MFCC features are used to distinguish between information of speech and non-speech signals [38]. Features with lower order MFCC can contains phonetic information and with higher order contains non-speech information. In case of stable and consistent source characteristics, MFCC is perfect representation for sound [36]. Information with frequencies at a maximum of 5 kHz, which covers most energy of human generated sounds [39] can be captured by it. But in the presence of background noise [40, 41], MFCC features do not work accurately and might not be well suited for generalization [36].

## II. Linear Prediction Coefficient (LPC):

LPC is a static feature extraction method that imitates the human vocal tract [42]. LPC is used to represent the spectral part of the speech in compressed manner, using linear predictive model. LPC is applied to get the filter coefficients which are equivalent to the vocal tract by minimizing the mean square error [43]. The linear predictive model is given [44, 45] as:

$$\hat{S}(n) = \sum_{k=1}^{p} a_k s(n-k) \quad (3)$$

Where $\hat{s}$is the predicted sample, p is the predictor coefficients, s is the speech sample. Then the prediction error is given as [16, 25]:

$$e(n) = s(n) - \hat{s}(n) \quad (4)$$

LPC is derived by the following equation:

$$a_m = \log\left[\frac{1-k_m}{1+k_m}\right] \quad (5)$$

Here $a_m$ is the linear prediction coefficient, and $k_m$ is the reflection coefficient.

Speech signal →Frame blocking → Windowing → Auto Correlation Analysis → LPC Analysis → LPC

**Figure 2: Steps in LPC approach**

LPC is used for speech reconstruction, coding and synthesis and known for its computation speed and accuracy. LPC helps to encode high quality speech signal at low bit rate [46, 47, 48]. Linear predictive analysis has the ability to efficiently select the vocal tract information from speech [42]. The steady and consistent behaviors are excellently represented by LPC [36]. LPC is very accurate in estimating speech parameters and highly sensitive to quantization noise [49].

## III. Linear Prediction Cepstral Coefficients (LPCC):

LPCCs are the coefficients of the Fourier transformation of the log magnitude spectrum [50] of LPC. LPCCs are LPC coefficients presented in cepstral domain and used to represent spectral envelope [49].

LPCC is calculated using the following equation:

$$C_m = a_m + \sum_{k=1}^{m-1} \left[\frac{k}{m}\right] c_k \, a_{m-k} \quad (6)$$

where $C_m$ is the cepstral coefficient and $a_m$ is the linear prediction coefficient.

Speech signal →Frame blocking → Windowing → Auto Correlation Analusis → LPC Analysis → LPC Parameter Conversion → LPCC

**Figure 3: Steps in LPCC approach**

LPCC has the ability to perfectly represent the speech waveforms and the speech characteristics with limited features [51].So, it is commonly applied in speech processing. LPCC have lower susceptibility to noise [50] and gives lower error rate [51].The higher order cepstral coefficients are sensitive to noise [52] whereas the lower order cepstral coefficients are sensitive to the spectral slope. These problems can be handled using weighted coefficient [53].

## IV. Line Spectral Frequencies (LSF):

LSFs are the alternative parameters which are used to represent all-pole spectrum of speech signal [54]. LFSs are used to represent LPCs for transmission over a channel [55]. We can express the linear predictive (LP) polynomial as the mean of palindromic and antipalindromic polynomials [56]. The LP polynomial is given by:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \quad (7)$$

This polynomial can be expressed in terms of palindromic and antipalindromic polynomials as given below:

$$A(z) = 0.5[P(z) + Q(z)] \quad (8)$$

Where,

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \quad (9)$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \quad (10)$$

Speech signal →Frame blocking → Windowing → Auto Correlation Analusis → LPC Analysis → Decomposition → Root Finder → LPF

**Figure 4: Steps in LFS approach**

The LSF representation of linear predictive polynomial consists of the location of the roots of the palindromic and antipalindromic polynomials. LSFs are used in coding, recognition and synthesization of speech [57, 58]. LSFs have low sensitivity to quantization noise and can be interpolated [59]. LSFs have relatively uniform spectral sensitivity [60]. Since LSFs have excellent quantization properties, they are important in transmission of vocal tract information from speech coder to decoder system. A near-minimal data set is provided for subsequent classification by the LSP representation in many cases [59].

## V. Discrete Wavelet Transform (DWT):

DWT is a time scale representation of the signal. It is a special case of wavelet transform where the wavelets are sampled discretely. DWT composed of a high pass wavelet filter and a low pass scaling filter [61, 62].

$$\phi(t) = \sum_{n=0}^{N-1} h[n]\sqrt{2}\phi(2t - n) \quad (11)$$

$$\psi(t) = \sum_{n=0}^{N-1} g[n]\sqrt{2}\phi(2t - n) \quad (12)$$

Where$\phi(t)$is the scaling function,$\psi(t)$is the wavelet function and h[n] is low-pass filter with impulse response and g[n] is high-pass filter with impulse response.

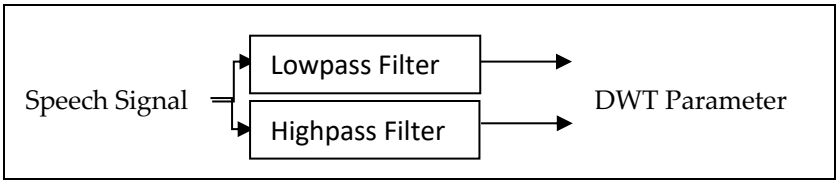


**Figure 5: Steps in DWT approach**

As it captures both frequency content and temporal content, it outperformed Fourier transformation. The DWT contains different scales of frequency information, thus DWT can help to get the speech information of respective frequency band [63]. Adequate numbers of frequency bands are provided by DWT for effective analysis of speech [64].

In this paper we have just included some commonly used feature extraction approaches. There are other feature extraction approaches for speech processing: Perceptual Linear Predictive (PLP) [65, 66], Relative Spectral Processing (RASTA) [67], Principle Component Analysis (PCA) [68, 69], Gammatone Frequency Cepstral Coefficient (GFCC) [70], Wavelet Packet Decomposition (WPD) [71, 72], Log-frequency power coefficient (LFPC) [25], etc.

## IV. Speech Emotion Recognition and Speaker Verification Databases

Suitable speech database is necessary for characterizing emotions for recognition or for verification. Quality of the database is an important issue to evaluate the performance of the system [24]. The proper design of the database is very much important for the classification task. Low-quality databases may yield incorrect conclusions. This section is divided into two sections: speech emotion recognition databases and speaker verification databases. We are not going to discuss all the existing databases; rather we will briefly discuss some interesting to us databases.

### I. Speech Emotion Recognition Databases:

**DESD:** Engberg et al. [73] described Danish Emotional Speech Database for evaluation of emotional state in emotional speech. It was a part of the VAESS project. It was recorder using microphone in Danish language with 4 speakers, out of which 2 were males and 2 were females. Ages of the speakers were between 34 and 52. They recorded the session on DAT tape @48kHz. It consists of 2 words, 9 sentences and 2 passages and five emotions: Neutral, surprise, anger, happiness and sadness.

**SUSAS:** Hansen et al. [74] developed Speech Under Simulated and Actual Stress (SUSAS) database for analysis and algorithm formulation in the field of speech recognition. The database is partitioned into five domains: a) styles of talking, b) speech with noise, c) speech associated with anxiety, depression, fear, d) actual subject motion-fear tasks and e) dual tracking computer response task. It has records of 32 speakers, ages between 22 and 76, out of which 19 are males and 13 are females and more than 16000 utterances. Samples are 1-channel 16-bit linear PCM with sample rate @8kHz.

**KISMET:** Breazeal et al. [75] used this database for explore the ability of recognition in robot directed speech. They used 2 female speakers and five communicative intents such as approval, attention, prohibition, soothing, and neutral. They used wireless microphone on Linux based system. There are 726 samples in .wav format with 16-bit single channel and 8 kHz signal.

**INTERFACE:** Hozjan el al. [76] recorded a database, as a part of the IST project Interface, to study emotional speech and to analysis the emotional characteristics for automatic emotion classification. They recorded for six emotions: anger, disgust, fear, joy, sadness and surprise, in English, French,

Slovenian and Spanish language. They used 2 male and 2 female speakers for English and 1 male and 1 female speaker for other languages. The number of sentences for English was 8928, Slovenian was 6080, French was 5600 and Spanish was 5520. They recorded the samples @16 kHz in linear format using condenser microphones.

**ESMBS:** New et al [25] used ESMBS database for text independent emotion classification of speech. It consists of 6 speakers (3 male and 3 females) with Burmese language and 6 speakers (3 male and 3 females) with Mandarin language. It has 10 utterances * 6 emotions * 12 speakers = 720 utterances generated by speakers. They recorded by using microphone in a quiet environment.

**MPEG-4:** Schuller et al [77] used MPEG-4 database for speaker independent speech emotion recognition. The emotion they were used are fear, anger, joy, sadness, disgust, surprise and neutrality. There were 2440 samples from 35 speakers. A condenser microphone was used to record the samples. There were 1,144 phrases and 1,507 utterances with textual contents extracted from 7 US movies.

**Berlin Emotional Database:** Burkhardt et al [78] described a database of emotional speech in German language. It consisted of 5 male speakers and 5 female speakers who produced 10 utterances, out of which 5 were short and 5 were long. They used seven emotions: neutral, anger, fear, joy, sadness, disgust and boredom and used about 800 sentences. They recorded the samples with frequency @48 kHz and later they downsampled them to 16 kHz.

**CLDC:** Zhou et al [79] used Chinese Linguistic Data Consortium (CLDC) for speech emotion recognition. It consists of 4 speakers: 2 male and 2 female speakers and six emotions: normal, happiness, surprise, anger, sadness and fear. There were 200 utterances for each emotion which yield 1200 utterances in total. Samples are recorded using 16-bit channel with sample rate of 16 kHz.

**KES:** Kim et al [80] used KES database designed by Professor C.Y. Lee for emotion recognition. The database contains context independent short, medium, and long sentences and four emotions: neutrality, joy, sadness and anger. The database has 10 speakers compromised of 5 male and 5 female speakers and 5400 sentences. They recorded the data @16kHz and in 32bits format over 30dB SN in a silent experimental environment.

**FAU Aibo:** Batliner et al [81] designed FAU Aibo Emotion Corpus to collect spontaneous and emotional speech of children. It consists of data from 51 children, compromised of 21 male children and 30 female children, aged ranges 10-13 years. They recorded the sample data with sampling rate @48 kHz in 16-bit format and downsampled them to 16 kHz using wireless headset and DAT-recorder. It has 48401 words and 9.2 hours of recordings.

**IITKGP-SESC:** Koolagudi et al. [82] introduced IITKGP-SESC, a Emotion Speech Corpus simulated by Indian Institute of Technology, Kharagpur for speech emotion analysis. The database was recorded using 10 speakers where there were 5 males and 5 females, aged ranges from 25 to 40 years, in Telegu language. They recorded for the eight emotions: Neutral, happy, anger, surprise, compassion, disgust, fear and sarcastic. They sampled the signals @16 kHz in 16-bit format using dynamic cardioid microphone. The database has 12000 utterances in total; 1500 utterances per emotion.

**TURES:** Oflazoglu et al [83] constructed Turkish Emotional Speech (TURES) Database for emotion recognition. They labeled each utterance with happy, surprised, sad, angry, fear, neutral and other emotional states and with 3- dimensional emotional space such as activation, valence and dominance. The database contains 582 speakers, compromised of 394 male speakers and 188 female speakers, and 5100 utterances extracted from movies. The audio channels were saved as mono 16-bit, PCM-wave format, sampled @48 kHz.

**EMOVO:** Costantini et al. [84] described EMOVO which is the first Italian emotional database. It has six emotional states; they are anger, disgust, fear, joy, neutral, surprise and sadness, and 6 speakers, 3 males and 3 females, age ranges from 23 to 30 years. They recorded the speech samples @48 kHz with 16-bit stereo in .wav format using two microphones and a digital reorder.

**BAUM-1:** Zhalehpour et al [85] discussed BAUM-1 which is a audio-visual database in Turkish language. It contains six emotional states: happiness, sadness, surprise, disgust, anger, fear, boredom and contempt, and mental states: unsure, thinking, concentrating and bothered. The data were collected from 31 speakers where it compromised of 18 males and 13 females. They recorded 1222 video clips with stereo and mono cameras.

**RAVDESS:** Livingstone et al. [86] developed an Audio-Visual Database of Emotional Speech and Song known as RAVDESS in North American English. The dataset contains eight emotional states: anger, calm, disgust, fearful, happy, neutral, sad and surprised. It has 24 speakers, compromised of 12 males and 12 females, and 7356 recordings. All speakers were recorded for 60 spoken utterances and 44 sung utterances, in total 104 utterances. They recorded the speech samples @48 kHz sampling rate with 16-bit in .wav format using condenser microphone.

**Table 2:***Databases of speechemotion recognition*

| Database | Year | Language | #Speakers (male/female) | #Utterance | Emotions |
|---|---|---|---|---|---|
| DESD | 1996 | Da | 4(2/2) | 10 minutes of speech | An, Ha, Ne, Sa, Su |
| SUSAS | 1999 | En | 32(19/13) | 16000 | An, Lo, Lm, Ne |
| KISMET | 2002 | En | 2(0/2), | 1002 | Ap, At, Ne, Pr, So |
| INTERFACE | 2002 | En, Fr, Sl, Sp | 10(5/5) | 26128 | An, Di, Fe, Jo, Sa, Su |
| ESMBS | 2003 | Bu, Ma | 12(6/ 6) | 720 | An, Di, Fe, Jo, Sa, Su |
| MPEG-4 | 2005 | En | 35 | 2440 | An, Di, Fe, Jo, Ne, Sa, Su |
| Berlin Emotional Database | 2005 | Ge | 10(5/ 5) | 10 | An, Bo, Di, Fe, Jo, Ne, Sa |
| CLDC | 2006 | Ch | 4 | 1200 | An, Fe, Jo, Ne, Sa, Su |
| KES | 2007 | Ko | 10(5/5) | 5400 | An, Jo, Ne, Sa |
| FAU Aibo | 2008 | Ge | 51(21/30) | | An, Bo, Em, He, Jo, Mo, Ne, Re, Rs Su, To |
| IITKGP-SESC | 2009 | Te | 10(5/5) | 12000 | An, Cm, Di, Fe, Ha, Ne, Sr, Su |
| TURES | 2013 | Tu | 582(394/188) | 5100 | An, Fe, Ha, Ne, Sa, Su, Other |
| EMOVO | 2014 | It | 6(3/3) | 588 | An, Di, Fe, Jo, Ne, Sa, Su |
| BAUM-1 | 2017 | Tu | 31(18/31) | | An, Bo, Bt, Co, Cn, Di, Fe, Ha, Sa, Su, Th, Un |
| RAVDESS | 2018 | En | 24(12/12) | 7356 | An, Ca, Di, Fe, Ha, Ne, Sa, Su |

**Abbreviation for languages:** Bu: Burmese, Ch: Chinese, Da: Danish, En: English, Fr: French, Ge: German, It: Italian, Ko: Korean, Ma: Mandarin, Pe: Persian, Sl: Slovenian, Sp: Spanish, Te: Telegu, Tu: Turkish.

**Abbreviation for emotions:** An: Anger, Ap: Approval, At: Attention, Bo: Boredom, Bt: Bothered, Ca: Calm, Cm: Compassion Cn: Concentration, Co: Contempt, Di: Disgust, Em: Emphatic Fe: Fearful, Ha: Happiness, He: Helpless, Jo: Joy, Lm: Lombard, Lo: Loud, Mo: Motherese, Ne: Neutral, Pr: Prohibition, Re: Reprimanding, Rs: Rest, Sa: Sadness,Sr: Sarcastic, So: Soothing,Su: Surprise, Th: Thingking, To: Touchy, Un: Unsure.

## II. Speaker Verification Databases:

**YOHO:** Campbell et al. [87] designed a YOHO database for text-dependent speaker verification based on combination-lock phrase. It was collected by International Telephone & Telegraph under US Government contract. It consisted of 138 speakers where there were 108 males and 30 females. There were 4 enrollment sessions per subject of 24 phrases each and 10 test sessions per subject of 4 phrases each. They used sample rate and sample coding of 8 kHz and 16-bit word respectively. The size of the data was 15 GB and device used foe recording was microphone.

**MAT-2000:** Wang et al. [88] describe a database, MAT-2000 (Mandarin speech data Across Taiwan), of Mandarin Chinese spoken in Taiwan. It was produced by ACLCLP and Philips Research East-Asia and collected through telephone networks. MAT-2000 consisted of 2232 speakers, out of which 1227 were female and 1005 were male, with 83.7h of recording and 641,936 spoken syllables. Sampled data were recorded in binary format at sampling rate 8 kHz and encoded as 16-bit linear PCM.

**BANCA:** Bailliere et al. [89] described BANCA database, a multi-modal database for multi-modal verification systems. It contains recordings of 208 subjects in four different European languages in controlled, degraded and adverse scenarios. There are 12 sessions in total, 4 for each scenario, spanned over 3 months. They used a digital camera and a webcam and recorded audio in 16 bit and 12 bit @32 kHz.

**VidTIMIT:** Sanderson et al. [90] created VidTIMIT database, which is an audio-visual multi-modal database for face and speech recognition, identification and verification. It has 44 files with recordings of 43 peoples, 24 being male and 19 being female, in 3 sessions. They recorded audio with 16 bit mono @32 kHz in WAV format and video with resolution 512 x 384 in JPEG format. It has 10 sentences per person and collected from NTIMIT database.

**MIT-MDSVC:** Woo al el. [91] discussed about MIT-MDSVC, which was collected for speaker verification research at MIT. It consisted of a set of enrolled users and a set of imposters. Data were collected in two sessions and 54 samples per user were recorded in each session, which resulted in 5,184 examples from enrolled users and 2,700 examples from imposter users. They used 48 speakers in the enrollment set and 40 speakers in imposter set.

**BioSecure:** Ortega-Garcia et al. [92] described a multimodal database, BioSecure, which is a collection of biometric data such as voice, iris, face, fingerprint, hand and signature modalities. It consists of three datasets: internal, desktop and mobile dataset. Internal dataset has 2 sessions with 971 donors, desktop dataset has 2 sessions with 667 donors and the mobile dataset has 2 sessions with 713 donors.

**MOBIO:** S. Marcel [93] et al. in 2010 used MOBIO database to evaluate the performance of speaker verification methods in mobile environment. MOBIO consisted of diverse set of bi-modal data i.e. audio and video data. The data were captured from 152 participants in a ratio of 1:2 i.e. 100 male participants and 52 female participants using a mobile phone and a laptop. There were 12 sessions for each client divided in two phases; 6 sessions for each phase. The 1$^{st}$ phase consisted of 21 questions while the 2$^{nd}$ phase consisted of 11 questions.

**UNMC-VIER:** Wong et al. [94] created UNMC-VIER database for robust audio-visual recognition systems. It contains video as well as audio recordings of 123 subjects, out of which 74 are males

and 49 are females. Recordings were done in both controlled and uncontrolled environments. They used a camcorder and a webcam for audio and video recording. In controlled environment, with camcorder, they recorded audio with 16 bit stereo @48 kHz in MP2 format and with webcam, they recorded audio with 16 bit mono @22 kHz in PCM format. In uncontrolled environment, audio recording was same as in controlled environment for camcorder; but for webcam, it was 16 bit mono @32 kHz with WMV2 format.

**AusTalk:** Burnham et al. [95] described AusTalk, a Australian speech database that included component of emotional speech. It was recorded in English with 1000 speakers in 3 one-hour each per participant sessions. A total of 322 words and a set of 58 sentences were selected for design the test set.

**RSR2015:** Larcher [96] et al. from Human Language Technology department, Institute for Infocomm Research designed a database named RSR2015, stands for Robust Speaker Recognition 2015, for text-depended speaker verification. It was based on fixed-size pass-phrases. It has 300 speakers aged between 17 to 42 years where number of male is 157 and number of female is 143. They recorded more than 151 hours of speech data in English language in 9 sessions, each of consisted of 30 short sentences. They used 6 smart phones and 1 tablet and selected the sentences from TIMIT database.

**SAS:** Wu et al [97] developed SAS (spoofing and anti-spoofing) database for text-independent automatic speaker verification. SAS has two sub datasets: SAS-VCTK based on Voice Cloning Toolkit database and SAS-RSR based on RSR2015 database. It starts with data from VCTK database with 45 males and 61 females. They divided the data into: Part A, Part B, Part C, Part D, Part E with 24 parallel utterances (for spoofing), 20 non-parallel utterances (for spoofing), 50 non-parallel utterances (for verification), 100 non-parallel utterances (for verification) and 200 non-parallel utterances (for verification) per speaker respectively. In Part A and Part B, all signals were sampled to 48 kHz and 16 kHz respectively, while in Parts C, Part D and Part E, all signals are sampled to 16 kHz.

**VoxCeleb:** Nagrani [98, 99] et al. developed an audio-visual dataset named VoxCeleb that contains short clips of speech collected from YouTube for speaker identification and verification. It was released in two stages: VoxCeleb1 and VoxCeleb2. VoxCeleb1 contains 1251 speakers of which male being 690 and female being 561 and 153516 utterances while VoxCeleb2 contains 6112 speakers of which male being 3761 male and female being 2351 and 1,128,246 utterances. Data are extracted from 22496 videos for VoxCeleb1 and 150480 videos for VoxCeleb2.

**DeepMine:** Zeinali et al. [100] designed DeepMine database in English and Persian language for text-independent, text-dependent and text-prompted speaker verification. It consists of 1969 speakers with 1149 males and 820 females and more than 540000 hours of recordings with 22742 sessions. It has three parts: fixed phrase for text-dependent verification, random sequence of words for text-prompted verification and random phonetic level transcription phrase for text-independent verification.

**HI-MIA:** X. Qin [101] et al. in 2020 introduced HI-MIA database for text-dependent speaker verification in far-field conditions in both English and Chinese. It includes two sub databases: AISHELL-wakeup dataset and AISHELL2019B-eval dataset, with utterances from 254 and 86 speakers respectively. The AISHELL-wakeup consisted of 3,936,003 utterances with 131 male and 123 female speakers. They recorded 160 utterances for each speaker with 120 in noisy environment and 40 in clean environment. The AISHELL2019B-eval dataset consisted of with 44 male and 42 female speakers. They set up the dataset with 40 in noisy environment and 120 in clean environment.

**Table 2:***Databases for speaker verification*

| Database | Year | #Speaker (Male/Female) | Phrase | Age Group | Language | Session | Environment/ Condition |
|---|---|---|---|---|---|---|---|
| YOHO | 1995 | 138(108/30) | Combination lock | -- | En | 14 | Office |
| MAT-2000 | 2000 | 2232(1005/1227) | Random | -- | Ma | -- | Noisy |
| BANCA | 2003 | 208(104/104) | -- | -- | En, Fr, It, Sp | 12 | Quiet & Noisy |
| VidTIMIT | 2003 | 43(24/19) | -- | -- | En | 3 | Noisy |
| MIT-MDSVC | 2006 | 88(49/39) | -- | -- | En | 2 | Quiet & Noisy |
| BioSecure | 2008 | 2351 | -- | 18-75 | En | 6 | Quiet & Noisy |
| MOBIO | 2010 | 152(100/52) | -- | -- | En | 12 | Office |
| UNMC-VIER | 2010 | 123(74/49) | -- | -- | En | 2 | Quiet & Noisy |
| AusTalk | 2011 | 1000(500/500) | -- | <25->50 | En | 3 | Quiet |
| RSR2015 | 2012 | 300 (157/143) | Fixed-size pass | 17 to 42 | En | 9 | Office |
| SAS | 2015 | 106(45/61) | Fixed, random | -- | En | -- | Clean |
| VoxCeleb1 | 2017 | 1251(690/561) | Random | -- | Multi | -- | Multi-Media |
| VoxCeleb2 | 2018 | 6112(3761/2351) | Random | -- | Multi | -- | Multi-media |
| DeepMine | 2019 | 1969(1149/820) | Fixed, random | 10 to 60 | En, Pe | 22742 | -- |
| HI-MIA | 2020 | 340(175/165) | -- | 10 to 50+ | En, Ch | -- | Quiet & Noisy |

**Abbreviation for Language:**Ch: Chinese, En: English, Fr: French, It: Italian, Ma: Mandarin, Pe: Persian, Sp:

## V. Conclusion:

Emotion recognition from speech signal is quite difficult because speaking styles, speaking rates of the speakers is different from person to person and it also changes from place to place i.e. different for native speakers and non-native speakers. Hence it is more important to select particular speech features which are not affected by the culture, region, and speaking style of the speaker. Frequency analysis of a signal can provide more relevant information than the time domain analysis. There are different spectral, prosodic, and acoustic properties of the signal which contains the information which are helpful in extracting features from speech signal. After feature extraction, feature selection is also very important and then followed by a suitable classifier to recognize the emotions. In this paper, we have briefly discussed some aspects of speaker and speech emotion recognition system. We have discussed features to characterize different emotional state, different feature extraction approaches and some databases for emotion detection and speaker verification. Performance of emotion recognition and speaker verification system depends on the proper selection of these aspects as different combinations of these may produce different results. We have presented some features with some basic merits and demerits of feature extraction approaches, information of emotional and verification databases; but have not provided any comparison among them.

References

[1] Soltani, K. and Ainon, R. N. (2007), Speech emotion detection based on neural networks, *9th International Symposium on Signal Processing and Its Applications*, 1-3.

[2] Mandal, A. K., Das, S. K. and Bhattacharjee, U. (2021), Speech recognition classifiers: a literature review, *Solid State Technology*, 64(2): 5215-5230.

[3] Bosch L. (2003), Emotions, speech and the ASR framework, *Speech Communication.* 40(1-2): 213–225.

[4] Busso, C., Lee, S. and Narayanan S. (2009), Analysis of emotionally salient aspects of fundamental frequency for emotion detection, *IEEE Trans. Audio Speech Language Process*, 17(4): 582–596.

[5] Cowie R., Douglas-Cowie E., Tsapatsoulis, N., Kollias, S., Fellenz, W. and Taylor, J. (2001), Emotion recognition in human–computer interaction, *IEEE Signal Processing Magazine,* 18(1): 32–80.

[6] Cowie R., and Cornelius, R. R. (2003), Describing the emotional states that are expressed in speech, *Speech Communication*, 40 (1–2): 5–32.

[7] Johnstone, T. and Scherer, K. R. (2000), Vocal Communication of Emotion, second ed., Guilford, New York, 226–235.

[8] Williams, C. and Stevens, K. (1981), Vocal correlates of emotional states, *Speech Evaluation in Psychiatry, Grune and Stratton,* 189–220.

[9] Banse, R. and Scherer, K. (1996), Acoustic profiles in vocal emotion expression, *Journal of Personality and Social Psychology*, 70(3): 614–636.

[10] Murray, I. and Arnott, J. (1993), Toward a simulation of emotions in synthetic speech: A review of the literature on human vocal emotion, *Journal of the Acoustic Society of America*, 93(2): 1097–1108.

[11] Lee, C. and Narayanan, S. (2005), Towards detecting emotions in spoken dialogs, *IEEE Trans. Speech Audio Process*ing, 13(2): 293–303.

[12] Cowie, R. and Douglas-Cowie, E. (1996), Automatic statistical analysis of the signal and prosodic signs of emotion in speech, *Proc. Fourth International Conference on Spoken Language*, 3: 1989–1992.

[13] Oster, A. and Risberg, A. (1986), The identification of the mood of a speaker by hearing impaired listeners, *Speech Transmission Lab. Quarterly Progress Status Report 4, Stockholm, 1986*, 79–90.

[14] Beeke, S., Wilkinson, R. and Maxim, J. (2009), Prosody as a compensatory strategy in the conversations of people with agrammatism, *Clinical Linguistic & Phonetics*, 23(2): 133–155.

[15] Borchert, ]. M. and Dusterhoft, A. (2005), Emotions in speech—experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments, *Proc. 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 147–151.

[16] Tao, J., Kang, Y. and Li, A. (2006), Prosody conversion from neutral speech to emotional speech, *IEEE Trans. Audio Speech Language Processing*, 14(4): 1145–1154.

[17] Cahn, J. (1990), The generation of affect in synthesized speech, *Journal of the American Voice I/O Society*, 8: 1–19.

[18] Rabiner, L. and Schafer, R., Digital Processing of Speech Signals, first ed., Pearson Education, 1978.

[19] Davitz, J. R., The Communication of Emotional Meaning, McGraw-Hill, New York, 1964.

[20] Gobl, C. and Chasaide, A. N. (2003), The role of voice quality in communicating emotion, mood and attitude, *Speech Communication*, 40(1–2): 189–212.

[21] Schlosberg, H. (1954), Three dimensions of emotion, *Psychological Rev.*, 61(2): 81–88.

[22] Scherer, K. R. (1996), Vocal affect expression: a review and a model for future research, *Psychological Bull.*, 99(2): 143–165.

[23] Wu, S., Falk, T. H. and Chan, W. (2011), Automatic speech emotion recognition using modulation spectral features, *Speech Communication*, 53(5): 768-785.

[24] Ayadi, M. E., Kamel, M. S. and Karray, F. (2011), Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition,* 44(3): 572–587.

[25] Nwe, T., Foo, S. and De Silva, L. (2003) Speech emotion recognition using hidden markov models, *Speech Communication*, 41(4): 603–623.

[26] Kaiser, L. (1962), Communication of affects by single vowels, *Synthese*, 14(4): 300–319.

[27] Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cetin, O., Bourlard, H. and Athineos, M. (2005), Pushing the envelope – aside, *IEEE Signal Processing Magazine*, 22(5): 81–88.

[28] Beyramienanlou, H. and Lotfivand, N. (2018), An efficient teager energy operator-based automated QRS complex detection, *Journal of Healthcare Engineering*, 2018: 1-11.

[29] Teager, H. and Teager, S. (1990), Evidence for nonlinear production mechanisms in the vocal tract, *Hardcastle W.J., Marchal A. (eds) Speech Production and Speech Modelling, NATO Advanced Institute*, 55: 241–261.

[30] Teager, H. (1990), Some observations on oral air flow during phonation, *IEEE Trans. Acoustic Speech Signal Processing*, 28(5): 599–601.

[31] Zhou, G., Hansen, J. and Kaiser, J. (2001), Nonlinear feature based classification of speech under stress, *IEEE Trans. Speech Audio Processing*, 9(3): 201–216.

[32] Caims, D. and Hansen, J. (1994), Nonlinear analysis and detection of speech under stressed conditions, *Journal of the Acoustic Society of America*, 96(6): 3392–3400.

[33] Alim, S. A. and Rashid, N. K. A. (2018), Some commonly used speech feature extraction algorithms, *From Natural to Artificial Intelligence - Algorithms and Applications,* R. Lopez-Ruiz, Ed., IntechOpen.

[34] Ravikumar, K. M., Reddy B. A., Rajagopal, R. and Nagaraj, H. C. (2008), Automatic detection of syllable repetition in read speech for objective assessment of stuttered Disfluencies, *Proc. World Academy Science, Engineering and Technology*, 36: 270-273.

[35] Chakroborty, S., Roy, A. and Saha, G. (2006), Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification, *IEEE International Conference on Industrial Technology*, 387-390.

[36] Chu, S., Narayanan, S. and Kuo, C. C. (2008), Environmental sound recognition using MP-based features, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1-4.

[37] Shah, S. A. A., ul Asar A, Shaukat SF. Neural network solution for secure interactive voice response, *World Applied Sciences Journal*, 6(9):1264-1269.

[38] Mubarak, O. M., Ambikairajah, E. and Epps, J. (2005), Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources, *8th international symposium on signal processing and its applications, Sydney, Australia*.

[39] Corrnaz C, Hunkeler U, Velisavljevic V., An automatic speaker recognition system. Switzerland: Lausanne; 2003. Retrieved from: http://read.pudn.com/downloads60/sourcecode/multimedia/audio/209082/asr_project.pdf

[40] Rao, T. B., Reddy, P. and Prasad, A. (2011), Recognition and a panoramic view of Raaga emotions of singers-application Gaussian mixture model, *International Journal of Research and Reviews in Computer Science*, 2(1): 201-204.

[41] Narang, S. and Gupta, M. D. (2015), Speech feature extraction techniques: a review, *International Journal of Computer Science and Mobile Computing*, 4(3): 107-114.

[42] Al-Sarayreh, K. T., Al-Qutaish, R. E. and Al-Kasasbeh, B. M. (2009), Using the sound recognition techniques to reduce the electricity consumption in highways, *Journal of American Science*, 5(2): 1-12.

[43] Kumar, P. and Chandra, M. (2011), Speaker identification using Gaussian mixture models. *MIT International Journal of Electronics and Communication Engineering,* 1(1): 27-30.

[44] Kumar, P. P., Vardhan. K. and Krishna, K. (2010), Performance evaluation of MLP for speech recognition in noisy environments using MFCC & wavelets, *International Journal of Computer Science & Communication*, 1(2): 41-45.

[45] Agrawal, S., Shruti, A. K. and Krishna, C. R. (2010), Prosodic feature based text dependent speaker recognition using machine learning algorithms, *International Journal of Engineering Science and Technology,* 2(10): 5150-5157.

[46] Kumar, R., Ranjan, R., Singh, S. K., Kala, R., Shukla, A. and Tiwari, R. (2009), Multilingual speaker recognition using neural network, *Proceedings of the Frontiers of Research on Speech and Music*, 1-8.

[47] Paulraj, M. P., Sazali, Y., Nazri, A. and Kumar, S. (2009), A speech recognition system for Malaysian English pronunciation using neural network, *Proc. of the International Conference on Man-Machine Systems*.

[48] Tan, C. L. and Jantan, A. (2004), Digit recognition using neural networks, *Malaysian Journal of Computer Science*, 17(2): 40-54.

[49] Anusuya, M. A. and Katti, S. K. (2009), Speech Recognition by Machine: A Review, *International journal of computer science and Information Security*, 6(3): 181-205.

[50] El Choubassi, M. M., El Khoury, H. E., Alagha, C. E. J, Skaf, J. A. and Al-Alaoui, M. A. (2003), Arabic speech recognition using recurrent neural networks. *Proc. 3rd IEEE International Symposium on Signal Processing and Information Technology*, 543-547.

[51] Wu, Q. Z., Jou, I. C. and Lee, S. Y. (1997), On-line signature verification using LPC cepstrum and neural networks, *IEEE Transactions on Systems, Man, and Cybernetics*, *Part B: Cybernetics,* 27(1): 148-153.

[52] Rabiner, L. and Juang, B., Fundamentals of Speech Recognition, PTR Prentice Hall, NJ, 1993.

[53] Zhu, L. and Yang, Q. (2012), Speaker Recognition System Based on weighted feature parameter, *Physics Procedia*, 25: 1515-1522.

[54] Itakura, F. (1975), Line spectrum representation of linear predictive coefficients of speech signals, The Journal of the Acoustic Society of America, 57(S1): S35.

[55] Sahidullah, Md., Chakroborty, S. and Saha, G. (2010), On the use of perceptual Line Spectral pairs Frequencies and higher-order residual moments for Speaker Identification, *International Journal of Biometrics*. 2(4): 358–378.

[56] Ma, Z., Taghia, J., Kleijn, Bastiaan, W., Leijon, A. and Guo, J. (2015), Line spectral frequencies modeling by a mixture of von Mises–Fisher distributions, *Signal Processing,* 114:, 219–224.

[57] de Alencar V. F. S. and Alcaim A. (2005), Transformations of LPC and LSF Parameters to Speech Recognition Features. *In: Singh S., Singh M., Apte C., Perner P. (eds) Pattern Recognition and Data Mining. ICAPR 2005. Lecture Notes in Computer Science*, 3686. Springer, Berlin, Heidelberg.

[58] Saito, S. (1992), Speech Science and Technology, Ohmsha, Tokyo, Japan, 81–86.

[59] McLoughlin, I. V. (2008), Line spectral pairs, *Signal processing*, 88: 448-467.

[60] Li, J., Chaddha, N. and Gray, R. M. (1999), Asymptotic performance of vector quantizers with a perceptual distortion measure, *IEEE Trans. Information Theory,* 45(May): 1082–1091.

[61] Hao, Y. and Zhu, X. (2000), A new feature in speech recognition based on wavelet transform, *Proc. IEEE 5th Inter Conf on Signal Processing,* 3(3): 1526-1529.

[62] Soman, K. P. and Ramchandran, K. I. (2005), Insight into Wavelets from Theory to Practice, 2nd edition, Prentice-Hall of India, New Delhi.

[63] Nehe, N. S. and Holambe, R. S. (2012), DWT and LPC based feature extraction methods for isolated word recognition, *EURASIP Journal on Audio, Speech, and Music Processing*, 7(2012): 1-7.

[64] Gałka, J. and Ziółko, M. (2009), Wavelet speech feature extraction using mean best basis algorithm, *International Conference on Nonlinear Speech Processing Berlin*, 128-135.

[65] Hermansky, H. (1990), Perceptual linear predictive (PLP) analysis of speech, The Journal of the Acoustic Society of America, 87(4): 1738–1752.

[66] Honig, F., Stemmer, G., Hacker, C. and Brugnara, F. (2005), Revising perceptual linear prediction (PLP), *INTERSPEECH 2005*, 2997-3001.

[67] Sing, R. and Rao, P. (2007), Spectral subtraction speech enhancement with RASTA filtering, *Proc. National Conference on Communications*, Kanpur, India.

[68] Ahmed, A. I., Chiverton, J. P., Ndzi, D. L. and Becerra, V. M. (2019), Speaker recognition using PCA-based feature transformation, *Speech Communication*, 110: 33-46.

[69] Jing, X., Ma, J., Zhao, J. and Yang, H. (2014), Speaker recognition based on principal component analysis of LPCC and MFCC, *IEEE Int. Conf. on Signal Processing, Communications and Computing*, 403-408.

[70] Kumaran, U., Rammohan, R., Nagarajan, S., et al. (2021), Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN, *Int Journal of Speech Technology*, 24: 303–314.

[71] Kotnik, B. and Kacic, Z. (2007), A comprehensive noise robust speech parameterization algorithm using wavelet packet decomposition-based denoising and speech feature representation techniques, *EURASIP Journal on Advances in Signal Processing*, 1-20.

[72] Mini, P. P., Thomas, T. and Gopikakumari, R. (2021), Wavelet feature selection of audio and imagined/vocalized EEG signals for ANN based multimodal ASR system, *Biomedical Signal Processing and Control*, 63.

[73] Engberg, I. S., Hansen, A. V. (1996), Documentation of the Danish Emotional Speech database (DES), *Internal AAU report, Center for Person Kommunikation, Aalborg Univ., Denmark (Online)*.

[74] Hansen, J. H. L., and Bou-Ghazale, S. E. (1997), Getting started with SUSAS: a speech under simulated and actual stress database, *Proc. EuroSpeech 97*, 4: 1743–1746.

[75] Breazeal, C. and Aryananda, L. (2002), Recognition of affective communicative intent in robot-directed speech, *Autonomous Robots*, 12: 83–104.

[76] Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., and Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database, *Proc. 3rd international conference on language*, 2019–2023.

[77] Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M. and Rigoll, G. (2005) Speaker independent speech emotion recognition by ensemble classification, *IEEE International Conference on Multimedia and Expo*, 864–867.

[78] Burkhardt, F., Paeschke, A. Rolfes, M., Sendlmeier, W. and Weiss, B. (2005), A database of German emotional speech, *Proc. INTERSPEECH 2005*, 1517–1520.

[79] Zhou, J. Wang, G., Yang, Y. and Chen, P. (2006), Speech emotion recognition based on rough set and svm, *5th IEEE International Conference on Cognitive Informatics*, 53–61.

[80] Kim, E., Hyun, K., Kim, S. and Kwak, Y. (2007), Speech emotion recognition using eigen-fft in clean and noisy environments, *16th IEEE International Symposium on Robot and Human Interactive Communication*, 689–694.

[81] Batliner, A., Steidl, S. and Noth, E. (2008), Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus, *Proc. of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect*, 28-31.

[82] Koolagudi S. G., Maity S., Kumar V. A., Chakrabarti S. and Rao K. S. (2009), IITKGP-SESC: Speech database for emotion analysis, *Ranka S. et al. (eds) Contemporary Computing. IC3 2009. Communications in Computer and Information Science,* 40: 485-492.

[83] Oflazoglu, C. and Yildirim, S. (2013), Recognizing emotion from Turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing*, 1: 1-11.

[84] Costantini, G. Iaderola, I., Paoloni, A. and Todisco, M. (2014), Emovo corpus: an italian emotional speech database, *Int. Conf. on Language Resources and Evaluation*, 3501–3504.

[85] Zhalehpour, Z., Onder, O., Akhtar, Z. and Erdem, C. E. (2017), Baum-1: A spontaneous audio-visual face database of affective and mental states, *IEEE Transactions on Affective Computing*, 8(3):300–313, 2017.

[86] Livingstone, S. R. and 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).

[87] Campbell, J. and Higgins, A., YOHO Speaker Verification LDC94S16. Web Download. Philadelphia: Linguistic Data Consortium, 1994.

[88] Wang H. C., Seide F., Tseng C. Y. and Lee L. S. (2000), MAT2000 – design, collection, and validation of a Mandarin 2000-speaker telephone speech database," Proc. ICSLP-2000, 4: 460-463.

[89] Bailly-Bailliére E. et al. (2003) The BANCA database and evaluation protocol, *In: Kittler J., Nixon M.S. (eds) Audio- and Video-Based Biometric Person Authentication. AVBPA 2003. Lecture Notes in Computer Science*, 2688: 625–638.

[90] Sanderson, C. (2002), Automatic person verification using speech and face information", *PhD Thesis, School of Microelectronic Engineering, Griffith University, Brisbane, Australia*.

[91] Woo, R. H., Park, A. and Hazen, T. J. (2006), The MIT mobile device speaker verification corpus: data collection and preliminary experiments, *Proc. of Odyssey, The Speaker & Language Recognition Workshop*, 1-5.

[92] Ortega-Garcia, J., et al. (2010), The multiscenario multienvironment biosecure multimodal database, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6): 1097-1111.

[93] Marcel, S., et al. (2010), Mobile biometry (MOBIO) face and speaker verification evaluation, *ICPR 2010*, 210-255.

[94] Wong, Y., Ch'ng, S., Seng, K., Ang, L., Chin, S., Chew, W. and Lim, K. (2011), A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities, *Pattern Recognition Letters*, 32(13), 1503–1510.

[95] Burnham, D., et al. (2011), Building an audio-visual corpus of Australian English: Large corpus collection with an economical portable and replicable black box, *Proc. ISCA, 2011*.

[96] Larcher, A., Lee, K., Ma, B. and Li, H. (2014), Text-dependent speaker verification: classifiers, databases and RSR2015, *Speech Communication*, 60: 56-77.

[97] Wu, Z., *et al*. (2015), SAS: A speaker verification spoofing database containing diverse attacks, *IEEE International Conference on Acoustics, Speech and Signal Processing, 2015*, 4440-4444.

[98] Nagrani, A., Chung, J. S. and Zisserman, A. (2017), VoxCeleb: A Large-Scale Speaker Identification Dataset, *Proc. INTERSPEECH 2017*, 2616-2620.

[99] Chung, J. S., Nagrani, A. and Zisserman, A. (2018), VoxCeleb2: Deep Speaker Recognition, *Proc. INTERSPEECH 2018*, 1086-1090.

[100] Hossein, Z, Jan, C. and Lukáš, B. (2019), A multipurpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database, *Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 2019*, 397-402.

[101] Qin, X., Bu, H. and Li, M. (2020), HI-MIA: a far-field text-dependent speaker verification database and the baselines, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, 7609-7613.