Applied Acoustics 158 (2020) 107020

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Trends in audio signal feature extraction methods

Garima Sharma *, Kartikeyan Umapathy, Sridhar Krishnan

The Department of Electrical and Computer Engineering, Ryerson University, ON M5B 2K3, Canada

ARTICLE INFO

Article history:
Received 6 August 2019
Received in revised form 28 August 2019
Accepted 1 September 2019
Available online 23 September 2019

Keywords:
Audio
Speech
Signal
Feature extraction
Survey
Machine learning

ABSTRACT

Audio signal processing algorithms generally involves analysis of signal, extracting its properties, predicting its behaviour, recognizing if any pattern is present in the signal, and how a particular signal is correlated to another similar signals. Audio signal includes music, speech and environmental sounds. Over the last few decades, audio signal processing has grown significantly in terms of signal analysis and classification. And it has been proven that solutions of many existing issues can be solved by integrating the modern machine learning (ML) algorithms with the audio signal processing techniques. The performance of any ML algorithm depends on the features on which the training and testing is done. Hence feature extraction is one of the most vital part of a machine learning process. The aim of this study is to summarize the literature of the audio signal processing specially focusing on the feature extraction techniques. In this survey the temporal domain, frequency domain, cepstral domain, wavelet domain and time-frequency domain features are discussed in detail.

1. Introduction

In order to make machines and computer intelligent like humans, we are using machine learning and artificial intelligence methods. Ideally, in today's world we want computers to have capability to make decisions like humans. One of the important sense of humans is hearing. Fig. 1 helps us to understand how human auditory system behaves over the listening range of 20 Hz–20 kHZ. It is a graph between sound pressure level (SPL) in decibel and the audible frequency range [1,2]. The graph shows the absolute threshold of hearing for different frequencies. Absolute threshold of hearing is the minimum sound pressure level of a pure tone that can be heard by a normal ear in silence. It is nice to note that the human auditory system's sensitivity is best between frequency range 2 kHZ–5 kHZ where the threshold reaches as low as −9 db SPL [3]. Approximately, for music our ears are sensitive between range 50 Hz and 15 kHZ while for speech the ears are sensitive between 100 Hz and 4.5 kHZ.

Humans can easily classify between various sounds without putting an extra effort e.g. we can easily classify between speech and music, car and truck sounds, baby and adult speech quality, various speakers, noise and useful sound etc. We want machines would be able to classify between various sounds as humans can do effortlessly. This problem is also called as machine hearing [4].

This leads to the research area of acoustic scene classification (ASC) [15] where a machine is trained to classify between various sounds that are present in an acoustic scene. Fig. 2 represents the ASC model was first proposed by Gerhard in [5].

Regardless of any particular aim, a ML system requires robust and discriminatory features that helps a machine to learn accurately and quickly. The intelligence of a machine is defined by the amount of training given to it. Normally the whole dataset is not fed into the computer or machine during its training to learn its properties, rather a reduced in size representation of the signals is used to train the machine. This compact representation of a signal is called as feature. The challenge is to extract right features ensures the success of a ML algorithm. The features must be compact in size but must highlight the characteristics of signal. The Features are so selected that it reduces the size of a signal significantly while still describing a signal completely and accurately. The reduced version of the signal improves the computational complexity and time complexity of the ML algorithms which in turn make it more suitable for real time applications. So, we can say that feature extraction is a process of dimension reduction of a signal which makes the signal more suitable for ML algorithms.

In this paper, we focus only on the features extracted from speech signals and on the musical sound. The application area of speech signals is very vast, to name a few like speech recognition [6], speaker recognition [8], blind source separation [7,8], speech enhancement [9], pathological speech detection [10,11], improving pathological speech [11,12], noise reduction, noise cancellation, gender classification, human-computer interaction [13] like Apple's Siri or Amazon's alexa etc. While on the other hand, music

* Corresponding author.
E-mail addresses: garima.sharma@ryerson.ca (G. Sharma), kumapath@ryerson.ca (K. Umapathy), krishnan@ryerson.ca (S. Krishnan).

https://doi.org/10.1016/j.apacoust.2019.107020
0003-682X/

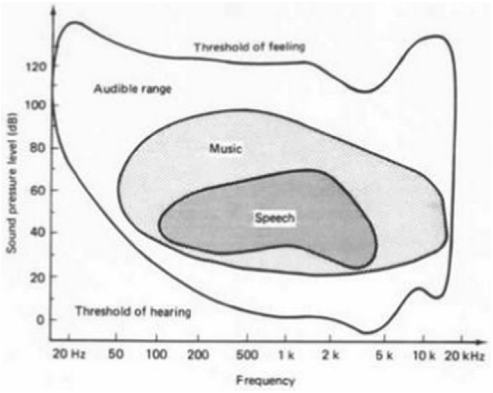


Fig. 1. Absolute threshold of hearing [1].

signals are analyzed in applications like mood detection [14], emotion prediction, genre classification [15], song tagging, singer classification etc.

The rest of the paper is organized as: Section 2explains the basic structure of a typical ML system and explains in brief about the audio types: speech, music and environmental sounds, Section 3discuss about the evolution of audio signal features and types of audio features. Section 4explains the time domain features, frequency domain features, time-frequency features, wavelet features, cepstral features, phase and eigen features and the last section discuss about the critical analysis, conclusion and future work.

2. Structure of machine learning system

The basic structure of any typical audio ML system is defined in Fig. 3. In the first stage, the pre-processing is done on the audio signal. The pre-processing may involve noise cancellation, silence reduction, normalization etc. The next stage is windowing of the signal that helps us to analyze the possible non-stationary signal as quasi-stationary signal. The whole signal can be studied and analyzed by sliding the window over the whole length of the signal. Using modern windowing methods, the size of the window can be made adaptive according to the characteristics of the signal. After that, the feature extraction and feature selection steps are taken. These steps decide the performance of the classifier. Then the selected features are fed into the classifier for training and testing and based on the classifier's prediction the decision is made.

2.1. Audio types

As discussed in Fig. 2, the audible audio signals are categorized into speech, music and environmental sounds. These are explained in brief below:

Speech: Speech is produced by human beings by using combination of various organs like lungs, mouth, nose, abdomen and the brain. The vocal tract and vocal cords play a major role in speech production. The speech production starts at frequency of 100 Hz and may go up to the frequency of 17 kHZ [17].

Music: Musical sounds are produced by musical instruments or humans in order to produce harmony and expression of emotions. Music can be described in various dimensions like genre, mood and sound characteristics. Traditionally, music has been classified into categories like rock, jazz, classical or pop. Ideally frequency range of music varies from as low as 40 Hz and may go as high as up to 19.5 kHZ [17].

Environmental sounds: In everyday life, we are surrounded by endless number of environmental sounds like sound of car or any other vehicle, running water, door bell, phone ring, factory noise, animal sounds etc. These sounds spread over the whole audible range. Fig. 4 shows the time domain structure of these 3 types of sounds. Fig. 4 shows the audio signal for human speech, guitar note and a car honk sound. The periodicity in the signal can be observed in speech and music sounds, but it is hard to find any periodicity in environmental sounds.

It is clear from Fig. 4 that speech is continuous in nature and has a smooth envelope, while guitar notes are of short duration and non-continuous in nature. The fire truck sound looks like noise and have very high amplitude. These three sounds not only differ in time domain, but also differ in frequency domain. Fig. 5 shows the frequency spectra of these 3 sounds. It can be noted from the Fig. 4 that the magnitude of the frequency spectrum of speech

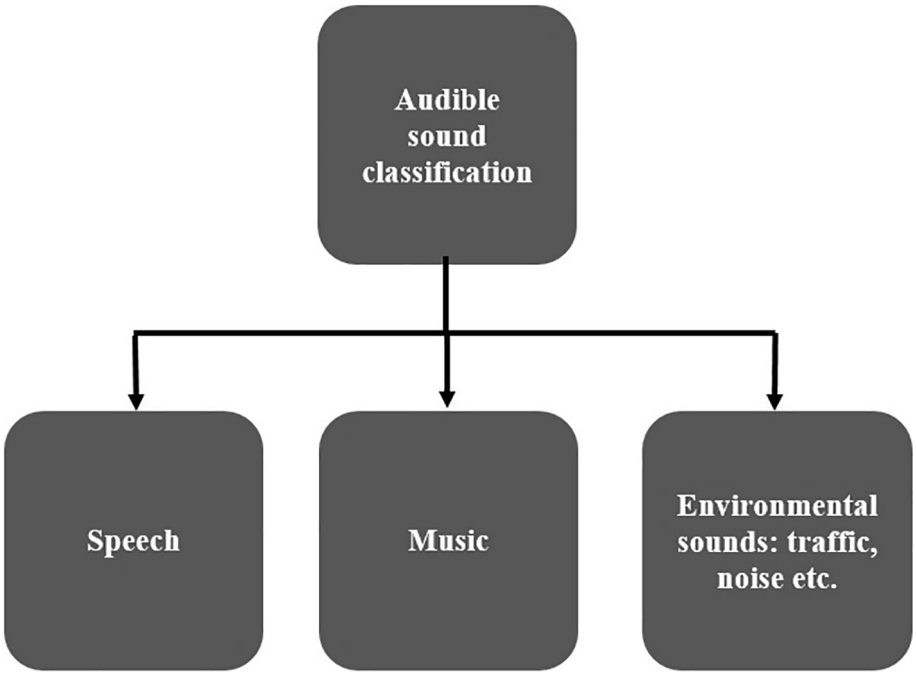


Fig. 2. Sound Classification.

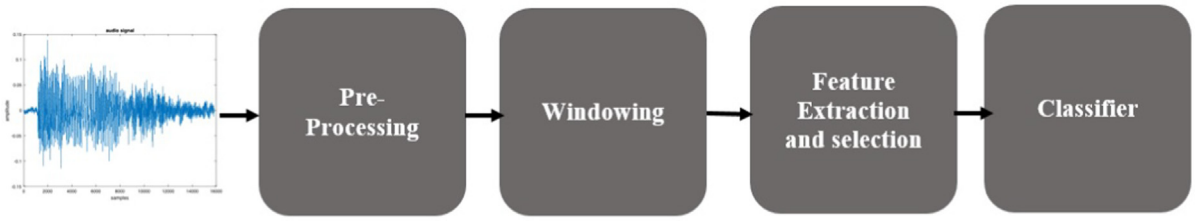


Fig. 3. Typical Machine Learning System.

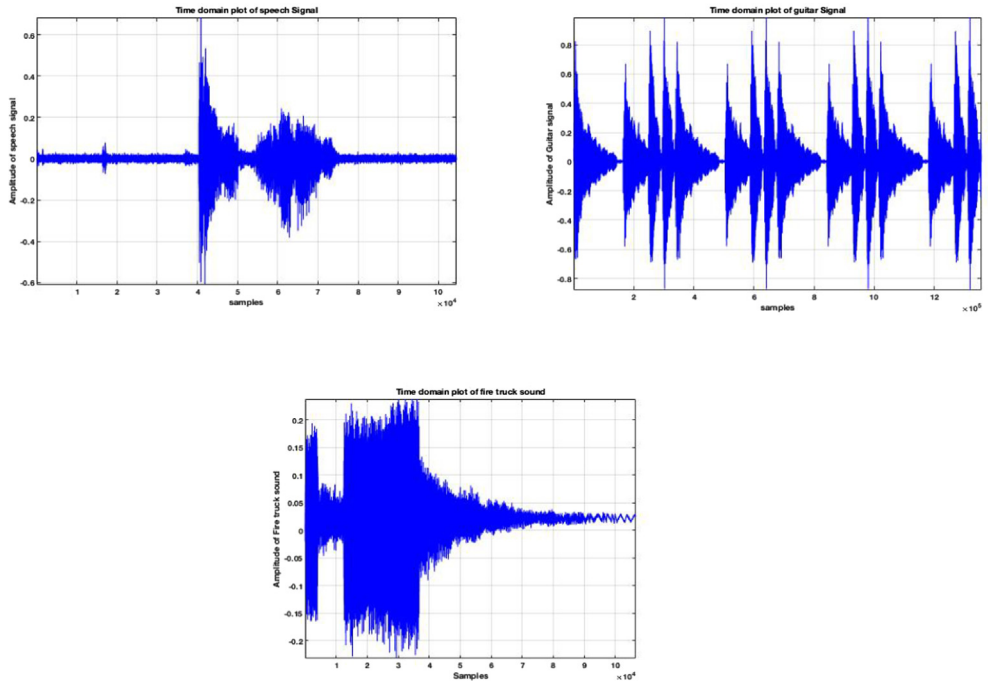


Fig. 4. Speech, Guitar and Fire Truck sound.

signal is very less in comparison to the music and environmental sound. This could be used speech/music classification.

3. Audio feature extraction methods: evolution and types

Evolution of audio features: In simple terms, feature extraction is a process of highlighting the most dominating and discriminating characteristics of a signal. A suitable feature mimics the properties of a signal in a much compact way. The evolution of audio signal features is explained in Fig. 6. The evolution of audio features can be sub-categorized into time domain, frequency domain, joint time-frequency domain, and deep features. The oldest and simplest features are extracted from the time domain. The time domain features evolved up to late 1950s [18,19,21]. Till date the time domain features plays an important role in audio analysis and classification. To analyze the spectrum of an audio signals, several features like pitch, formants etc. were evolved from frequency domain and employed in various application till date. The evolution of frequency domain features was around 1950s to 1960s [20,22]. In later 1960s, the joint time-frequency [23–25]feature extraction algorithms were developed. Since then these features are used in audio signal processing algorithms. Since the evolution of deep learning, the deep features are extensively used in various applications, in audio signal processing deep features have been used since 2010 in the area of acoustic scene classification [127,128], speaker recognition [130]and audio video analysis [129].

4. Audio signal feature extraction

4.1. Time domain features

Before discussing about the time domain features it is important to discuss the concept of windowing in time domain. The simplest way to analyze a signal is to analyze it in its original form. All the sound signals we are discussing here is a time series signal i.e. these signals evolves with time. By visualizing a signal in time domain, we may analyze few key characteristics of a signal and this information can be used in predicting and analyzing similar signals. This time domain analysis is simple till the signal is of short time or reflects stationary properties over time. In real time audio signals are non-stationary over the time. To analyze such non-stationary signals windowing technique is employed and the long non-stationary signal is analyzed as short chunks of quasi-stationary signal.Windowing can be seen as multiplying a signal with a window function that is zero everywhere except the region of interest. The resultant windowed signal is the subset of the original signal which is passed though the window, for rest of the time the signal is zero. The simplest type of window is rectangular window which is defined by Eq. (1):

$$W_R(n) = \begin{cases} 1, & -\frac{M-1}{2} \leqslant n \leqslant \frac{M-1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

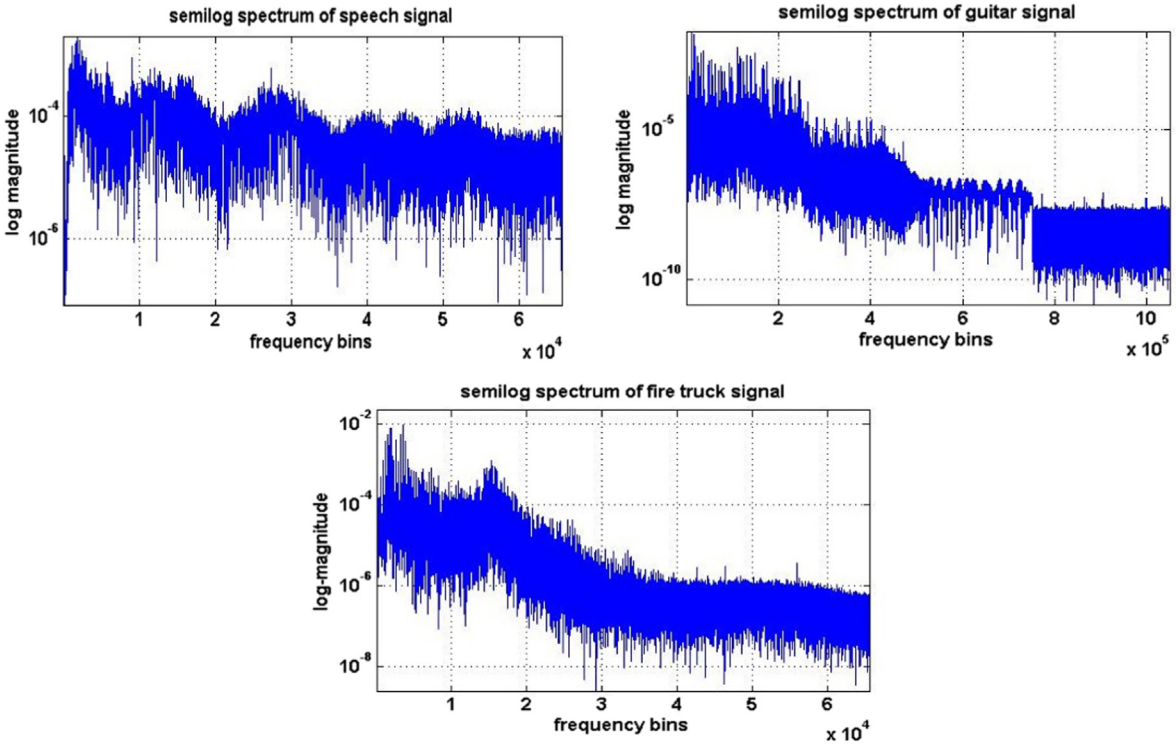


Fig. 5. Speech, Guitar and Fire Truck sound frequency spectrum.

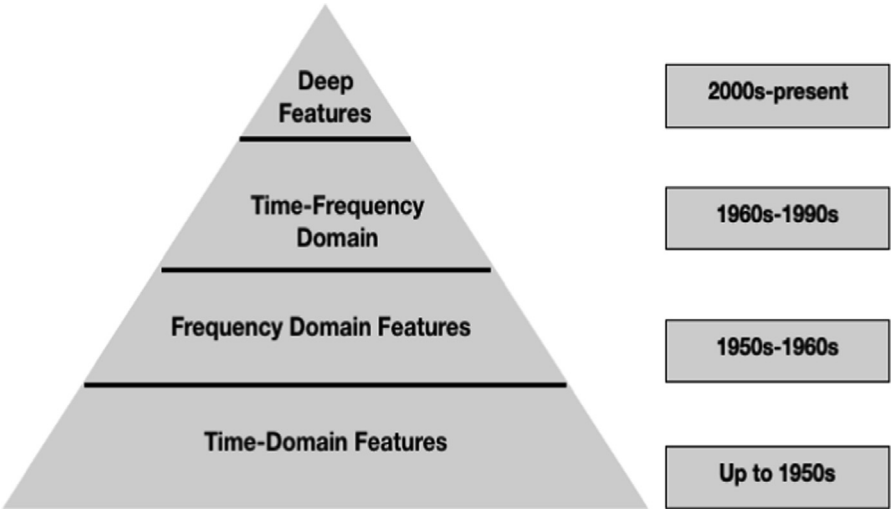


Fig. 6. Evolution of Audio Feature Extraction.

The Fig. 7 explains the concept of windowing a signal when using rectangular window as a function. In order to analyze the whole signal, the window is slided over the time and moves from the left most corner of the plot towards the right corner of the plot. The size of the window is made adaptive and is changed according to the characteristics of the original source signal in order to convert the long non-stationary signal into small quasi-stationary signal. Fig. 8 explains the sliding process of an adaptive rectangular window over a signal.

One problem with the rectangular window is the abrupt change in its shape at the edges, which may cause distortion when the signal is being analyzed. The distortion is the result of the Gibbs phenomenon. In order to handle this problem, we may use a window function with smooth curves like Hamming or Hanning window. These window functions are zero at the edges and rises gradually to be one in the middle of the window shape. With such window functions, the edges of the signal are downgraded and the edge effect because of Gibbs phenomenon is reduced.

Zero-crossing rate (ZCR): The ZCR of an audio frame is defined as the rate of change of sign of the signal during the frame. Mathematically, it is the number of times a signal changes its sign from positive to negative and vice versa, divided by the length of the frame. In simple words, ZCR is the number of times signal crosses the zero level in one second interval. The ZCR for ith frame with the length N is defined by Eq. (2) and explained in Fig. 9 as:

$$Z(i) = \frac{1}{2N}\sum_{n=1}^{N}|sgn[x_i(n)] - sgn[x_i(n-1)]| \tag{2}$$

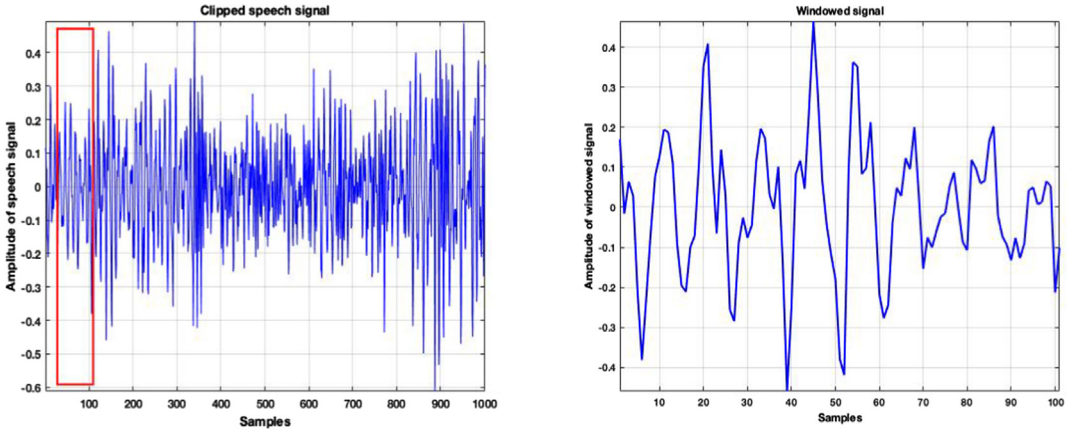


Fig. 7. Windowing in Time Domain.

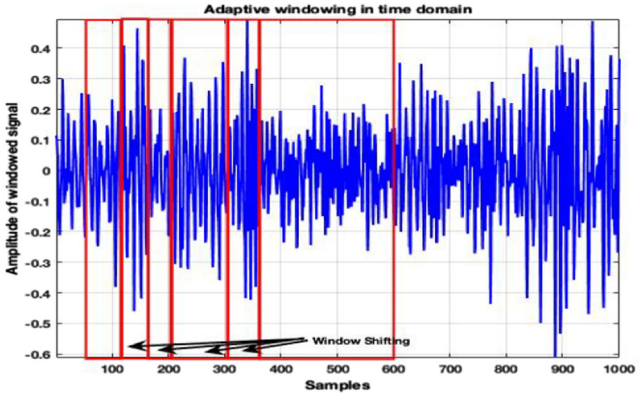


Fig. 8. Adaptive Windowing in Time Domain.

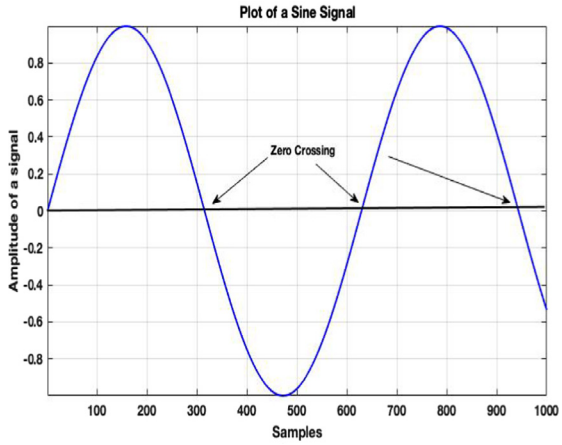


Fig. 9. Zero-Crossing in signal.

where sgn(.) is a sign function i.e

$$sgn[x_i(N)] = \begin{cases} 1, & x_i(n) \geqslant 0 \\ 0, & x_i(n) < 0 \end{cases} \quad (3)$$

ZCR is a very efficient way to detect voice activity that determines whether a speech frame is voiced, unvoiced or silent. The ZCR is higher for the unvoiced portions of the speech in comparison to the voiced portion of the speech. Fig. 10(a) represents the speech signal and its ZCR. It is clear that ZCR for unvoiced segments are very high than for the voiced segments. Of course, in ideal conditions the silence portion in a clean speech the ZCR must be equal to zero.

ZCR is also a technique to estimate fundamental frequency (FF) [28] of the speech. The ZCR is equal to the twice the frequency of the signal. Hence, we can say that ZCR gives indirect information about the frequency of the signal. Hence, this feature can be used to design discriminator and classifier [27]. Fig. 10(b) shows the ZCR for music clip and it would be interesting to note that, the ZCR for music is higher than that for the speech signal [56]. The MATLAB pseudo code for calculating ZCR is given below:

Algorithm 1: Zero crosing rate

1. **Result**: Zero Crossing Rate of a Signal
2. Initialization: mono-channel signal $x_i(n)$
3. ZCR= $sum(abs(diff(x_i(n) > 0)))/length(x_i(n))$

- Another type of ZCR based feature is called modified ZCR. The modified ZCR is the Zero-crossing function with detrending technique. In this method, Eq. (2) is modified and represented in Eq. (4) as:

$$Z(i) = \frac{1}{2N}\sum_{n=1}^{N}|sgn[\hat{y}_i(n)] - sgn[\hat{y}_i(n-1)]| \quad (4)$$

where

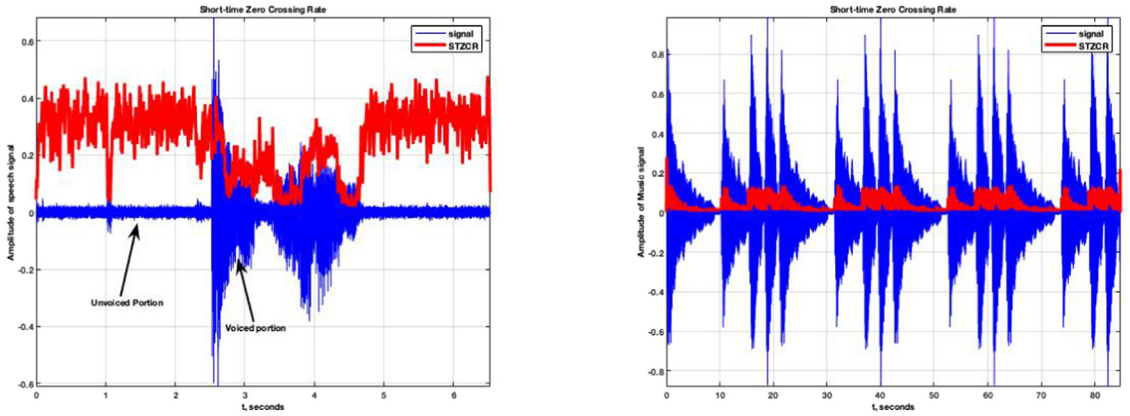
$$\hat{y} = \hat{x} - y_d, \hat{x} \text{ is mean value of x} \quad (5)$$

- Linear prediction ZCR: It is the ratio between ZCR of original signal and the ZCR of prediction error obtained from a linear prediction filter [33].

Hence, ZCR is used in many applications including music/speech discrimination [29,30], music genre classification [32], voice activity detection [31] and vowel detection and analysis [145].

Amplitude based features: These are based on very simple analysis of temporal envelop of the signal. The various type of amplitude based features are discussed below:

- Amplitude descriptor (AD): This feature helps to differentiate between different type of sound envelopes by considering the energy, variation of duration of signal segments from high and low amplitudes by the means of adaptive threshold. This feature is mainly used in environmental sound classification [34].
- Attach, Delay, Sustain, Release (ADSR) envelop: This ADSR feature is used in music analysis and classification between musical genres.



(a) ZCR for Speech Signal

(b) ZCR for Music Signal

Fig. 10. Comparison of ZCR for Speech and Music.

The ADSR envelope feature is not achievable for most real time sounds because the decay part is not clearly present, sustain part is not present in speech and environmental sounds (only present in music sounds). To deal with such a problem a modified envelope based on attack and rest is used, it is called as AR envelope. In this the decay part is not present and sustain and release part is merged. The ADSR and AR envelopes are used in timbre analysis of musical instruments [36]. Fig. 11 shows the ADSR and AR envelop for a signal. The MATLAB pseudo code for ADSR envelope is given in algorithm 2.

Algorithm 2: ADSR envelope detection

1. **Result**: ADSR envelope
2. Input: Musical Key number, duration of key pressing
3. $freq = 440 * 2((keynum - 49)/12)$ ▷calculate frequency for given key
4. $t = 0 : 1/sampling\ frequency : duration$
5. $tone = sin(2 * pi * freq * t)$ ▷generate sinusoidal output tone
6. $A = linspace(0, 1, 0, 1 * (length(tone))$ ▷rise 10 percent of signal
7. $D = linspace(1, 0.8, 0.15 * (length(tone))$ ▷drop of 15 percent of signal
8. $S = linspace(0.8, 0.8, 0.6 * (length(tone))$ ▷delay of 60 percent of signal
9. $R = linspace(0.8, 0, 0.15 * (length(tone))$ ▷drop of 15 percent of signal
10. $ADSR = [ADSR]$
11. Multiply ADSR with tone and plot on MATLAB

- log attack time (LAT): It is the logarithmic (base 10) of the time duration between the time starts to the time it reaches its stable part. It has been used for musical onset detection [61]and environmental sound recognition [39,40].

Algorithm 3: Log attack time

1. **Result**: Log attack time
2. LAT = $log_{10}(t_{attackend} - t_{attackstart})$ ▷t is time

- Shimmer: Shimmer computes cycle-to-cycle variations of the amplitude in a waveform. It is used in voice activity detection, speaker recognition, speaker verification [37], classify musical sounds [38].

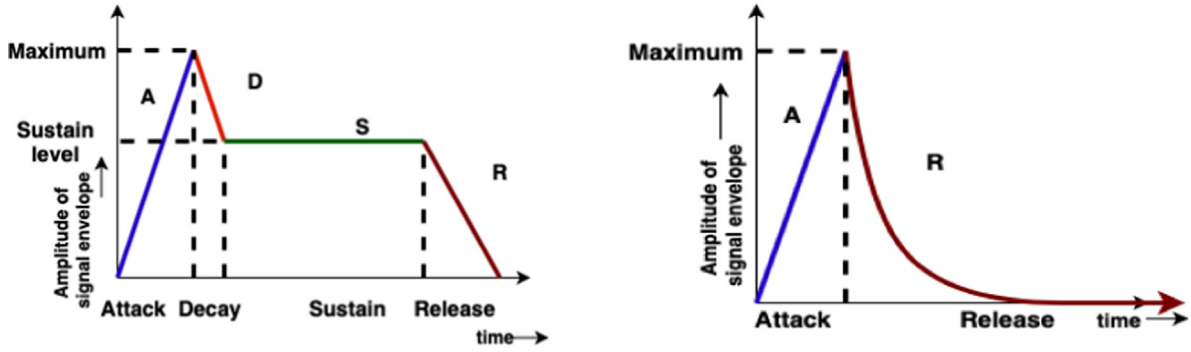
Energy based features:

- Short time energy (STE): The sound signals are non-stationary in nature. As explained above, the non-stationary signal can be transformed into small portions of quasi-stationary signals using framing/windowing method. The energy through out the signal is variable and hence it is not feasible to predict a value. For this, the short time energy which is the energy from a frame is calculated. STE [56] is defined as average energy per frame. The STE is low for unvoiced segments and high for voiced segments. Fig. 12 shows the STE for speech and music. STE is used to detect the voiced-unvoiced segments [31], music onset detection [61], environmental sound detection [42], vowel detection and analysis [145] and audio based surveillance systems [41]. The pseudo code for calculating STE is given in algorithm 4.

Algorithm 4: Short time energy calculation

1. **Result**: Short time energy value
2. Inputs: Audio signal (x), window type, window amplitude, window length
3. win = select a window function ▷rectangular, hamming, hanning, blackmann etc.
4. $xnew = x^2$
5. $STE = xnew \otimes win$ ▷convolution of window and signal square

- Volume: Volume or Loudness of a sound is the one of the most promising feature of a human auditory system. Mathematically, volume is defined as the root mean square (RMS) value of the magnitude of the signal within a frame. It is used in speech/music discrimination [52], speech segmentation and acoustic scene classification [43]. algorithm 5 explains the pseudo code for calculating volume of an audio signal.



(a) ADSR envelope (b) AR envelope

Fig. 11. Amplitude based envelope feature.

Algorithm 5: Volume or loudness of audio signal

Result: Loudness of audio signal
1. Input: Audio signal (x), sampling frequency (Fs)
2. loudness = integratedLoudness(x,Fs) ▷in-build function of MATLAB

- Temporal centroid (TC): The temporal centroid is the time averaged over the energy envelope. The temporal centroid has been used in environmental sound recognition [42] and acoustic scene classification [43]. The algorithm for the TC is given below:

Algorithm 6: Temporal centroid

1. **Result**: TC
2. find $e(x)$ = energy envelope of the Audio signal (x)
3. Multiply $e(x)$ by the signal itself
4. Find sum of the products of signal and energy envelope of signal
5. TC = Divide the result by total energy envelope

Auto-correlation Based Features: Auto-correlation is a measure of self-similarity of a signal in time domain. In simple words, it measures the similarity between the signal and its delayed version. The auto-correlation value of +1 represents strong positive association, −1 represents a negative association and 0 shows no association. The auto-correlation at lag zero is always 1, this is because the signal is always perfectly correlated with itself. Fig. 13 shows the values of auto-correlation of a speech signal with itself with a time lag of 20. For example at time lag 1 the auto-correlation value is 0.8 that represents the 80% similarity of the signal to itself when the signal is lag by 1 unit [35,44].

Auto-correlation function is used to determine the periodicity present in a signal. It is used to analyze musical beats and their

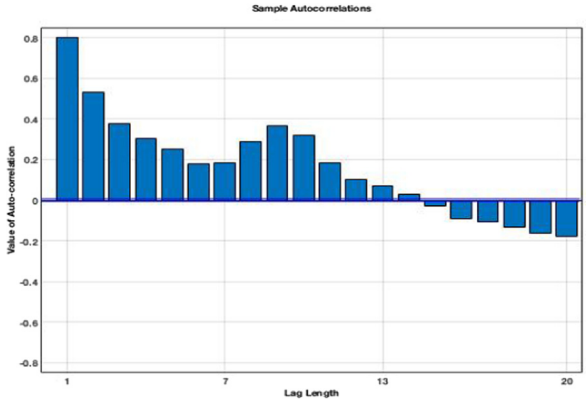
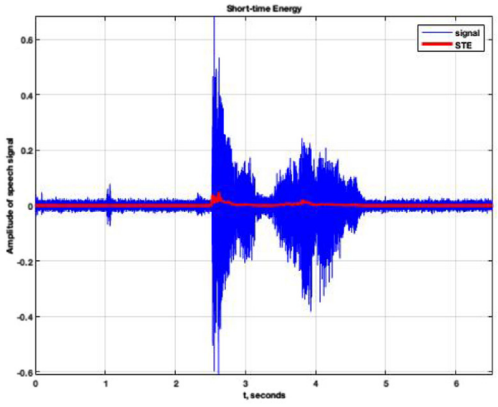
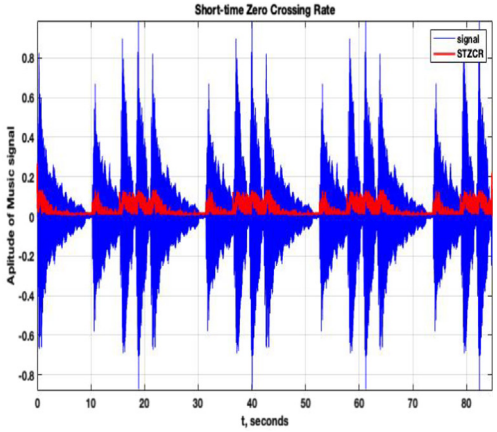


Fig. 13. Auto-correlation values of a Signal with Lag of 20.



(a) STE for Speech



(b) STE for Music

Fig. 12. STE from audio clips.

tempo. It is also used to estimate the pitch (fundamental frequency) of a signal.

Rhythm Based: Rhythm in general is regular recurrence of pattern over the time. The rhythm is found in musical instruments, poetry (speech) and environmental sounds (e.g. bird chirping). Some rhythm based features are speech duration, articulation rate, phoneme duration, pause ratio, total duration, total pause duration, total vowel duration, pulse metric, pulse clarity, band periodicity, beat tracker, beat histogram etc. Pulse metric is a measure that uses the long-time band pass auto-correlation over a window of 5 s. This feature is used in speech/music discrimination [33,52], analysis of pathological speech [45], music genre classification [46], music instrument classification.

4.2. Frequency domain features

Time domain graph shows the signal variation with respect to time. To analyze a signal in terms of frequency, the time-domain signal is converted into frequency domain signal by using Fourier transform or auto-regression analysis. Frequency domain analysis is a tool of utmost importance in audio signal processing. The major frequency domain features are discussed below:

- *Auto-regression based:* Auto-regression based features are extracted from linear prediction analysis of a signal. The most common auto-regression based features are:
  - *Linear Predictive Coding (LPC) Coefficients:* LPC removes the redundancy from a signal and tries to predict next values by linearly combining the previous known coefficients. LPC is the all pole filter that represents the spectral envelope of a digital speech in compressed form using linear prediction model. LPC coefficients are use for audio segmentation and audio retrieval. In MATLAB there is a in-build function "lpc" having parameters audio signal *x* and order of linear predictor.

Algorithm 7: Extraction of LPC coefficients

Result: LPC coefficients
1. Input: Audio signal *x*.
2. Perform normalization and preemphasis on the signal *x*.
3. Implement frame blocking.
4. Perform windowing on the frame blocked signal.
5. Do auto-correlation analysis.
6. Analyze using liner prediction by using levinson-durbin algorithm.

  - *Code Excited Linear Prediction (CELP):* The CELP is based on three techniques:
    1. Use of linear prediction model to mimic vocal tract.
    2. Use of adaptive or fixed code-book entries as the excitation signal to the linear prediction model.
    3. The search is performed in a closed loop and in a perceptually weighted domain. CELP is a speech coding algorithm and provides better quality than low bit rate algorithms such as linear predictive coding Vocoders and residual-exited linear prediction algorithms. This features is the combination of Linear spectral frequency and features related to pitch and residual signal. The CELPs are used in environmental sound recognition [50].

  - *Linear Spectral Frequency:* It is also called as linear spectral pairs and useful in speech coding. LSF are used to represent linear prediction coefficients for the transmission over the channel. A linear prediction polynomial is represented as the average of pallindromic polynomial and antipallindromic polynomial. The pallindromic polynomial represents the vocal tract when glottis is closed and the anti-pallindromic polynomial represents the vocal tract when glottis is open. The roots of the pallindromic and anti-pallindromic polynomials are conjugate in nature and hence half of the roots are transmitted. The LSF representation of the Linear Prediction polynomial consists simply of the location of the roots. This feature shows variation in value when the glottis is closed and open. Hence, this feature is used in voiced/unvoiced segment detection, speaker segmentation [51] and speech/music discrimination [52].
- **Peak Frequency:** Peak frequency is simply the frequency of maximum power. It gives an estimate about the most dominant frequency present in the signal and also helps to calculate fundamental frequency of the signal. Peak frequency is a useful parameter when we are classifying music and speech, gender classification etc. Fig. 14 shows the peak frequency from a single sided FFT spectrum.

Algorithm 8: Peak Frequency

1. Input: Audio signal (x)
2. Covert the signal into frequency domain by using FFT.
3. Calculate the absolute value of the transformed signal.
4. Find the maximum value from the result of step 3.

- **Method of Selection of Amplitudes of Frequency Multi-expanded Filter based features:** The Method of Selection of Amplitudes of Frequency Multi-expanded (MSAF MULTI-EXPANDED) filter based features are the hand-crafted features specially used to detect faults in electric motors used for drilling or grinding [147] or in detecting faulty commutator motor [151]. This is an acoustic feature that extracts the differences between FFT spectra. The generalized algorithm to extract MSAF-MULTIEXPANDED acoustic features is given below:

Algorithm 9: MSAF-MULTIEXPANDED acoustic features

1. Input: Acoustic signal captured from electric motors.
2. Compute FFT spectra of good and faulty acoustic signals.
3. Calculate difference between good and faulty spectra.
4. Compute the absolute value of the difference calculated.
5. Find common frequency components from the absolute difference values.
6. Form groups of frequency components.
7. Form bandwidth of the frequencies to construct classes for classification.
8. Form a feature vector from those bandwidths.

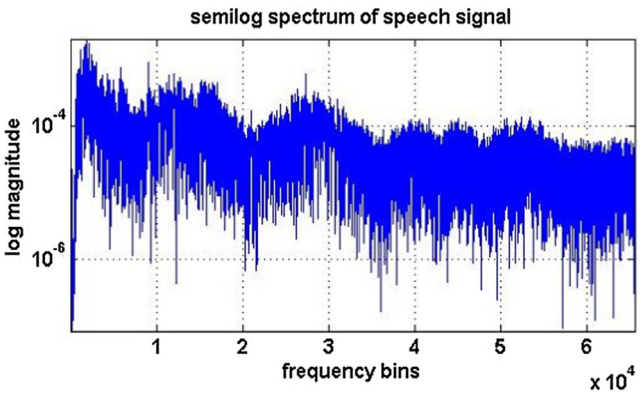


Fig. 14. Peak Frequency of a Signal from single sided spectrum.

- **Shortened method of frequencies selection MULTIEXPANDED:** Very similar to MSAF-MULTIEXPANDED feature, Shortened method of frequencies selection MULTIEXPANDED (SMOFS-MULTIEXPANDED) is another hand crafted feature that is too employed in industry applications specifically to diagnose fault in motors [149]. The algorithm below explains how SMOFS-MULTIEXPANDED features are extracted and how these are different from MSAF-MULTIEXPANDED features.

Algorithm 10: SMOFS-MULTICRAFTED acoustic feature

1. Input: Vibrations captured from motors.
2. Extract FFTs from various healthy and faulty motor's vibration sounds.
3. Calculate absolute difference of the various FFTs.
4. Select the frequency components greater than a threshold.
5. Compute the variable threshold of selection of frequency components using iterative methods.
6. Set the parameter: Threshold of common frequency components MULTIEXPANDED extension.
7. Find common frequency components.
8. Form a final feature vector from common frequency components.

- **STFT based features/Time-Frequency Features:** A time-frequency transform of a signal is the way of looking into the signal having time on one axis and frequency on another. The time-frequency analysis could be obtained by using time-frequency distribution (TFD). Generally the time-frequency domain is called as time-frequency representation (TFR) obtained by TFD. The time-domain shows the variations in signal amplitude over the period of time, while in frequency domain the magnitude of the frequency content gives only frequency information but no time information. A TFR bridges this gap and provide the time and frequency resolution. STFT is the most common way to have a TFR. Also, a TFR features are effective for analysis non-stationary aspects of a signal such as trends, discontinuities and patterns which is normally missed by the time or frequency domain features [135].

  – *Time-Frequency Matrix:* By using the STFT, the time-domain signal could be converted into TFR. However, this representation contains a huge amount of data and information. For example of the Guitar signal, sampled at 44.1 kHZ the TFD with a resolution of 128×1024 gives 131,072 TF samples. To reduce the dimension of the TF matrix the various decomposition techniques could be employed to get the relevant and compact TFR. Some well known TFD techniques are:

  1. linear TFDs: These are the simplest form of TFDs and is equal to the squared modulus of the STFT signal [139]. The original size of this TFD is quite huge.
  2. Quadratic TFDs: To make the TFD adaptive, the window size is made adaptive with respect to the signal. The wigner ville distribution (WVD) is the simplest quadratic TFD. It is a great method to analyze signals in time-frequency domain. This distribution doesn't suffers from leakage effects as STFT does. Hence it gives the best spectral resolution. This feature has been used in analysis of audio signals [119] and detecting industrial gear failures [120]. It is also used in seismic data processing [121]. The major drawback of the WVD is the generation of cross terms due to its quadratic nature. This drawback is eliminated by Cohens class of bilinear TF representation.
  3. Positive TFDs: The positive TFDs gives positive terms and are free from cross terms. The positive TFDs [138] are based on the signal dependent kernels.
  4. Matching pursuit TFDs: Matching pursuits TFD is based on the matching pursuit decomposition that uses non-orthogonal basis functions to decompose a signal into gabor atoms with a variety of possible translations, modulations and scaling. These are highly used in environmental sound classification [135,139].

This Time-frequency matrix could directly be used in various applications like classification of environmental sounds [135,137]. The TFD is used in industry too. For instance the vibrations generated by the motors is analyzed using TFD to find the faulty motor bearing present in the motor [148,150].

  – *Sub-band energy ratio:* The sub-band coding breaks the signal into different frequency bands typically by using FFT or STFT and encodes each one independently. The sub-band energy ratio is defined as the measure of normalized signal energy along these different frequency bands. It has been used for audio segmentation, environmental audio recognition [42] and music analysis [43].

  – *Spectrum envelope:* The spectrum envelope is a log-frequency power spectrum of a signal and can be used to generate reduced spectrogram of the audio signal. The spectral envelope when generated by linear prediction method, it is called as linear prediction spectral envelope. Due to error optimized by linear prediction, the spectral peaks of an audio signal are more accurate and emphasis on the envelope as they are in auditory system. This feature has been used in music genre classification [29,32]and environmental sound recognition [42].

  – *Stereo panning spectrum feature (SPSF):* In audio signal processing, the stereo audio is converted into to mono channel audio before proceeding for the processing [68]. Hence, the information content present due to stereo panning is not fully utilized. In order to answer this, stereo panning spectrum is considered. The frequency-domain source identification system based on cross-channel metric is called panning index. The stereo panning spectrum holds the signals between -1 to $+1$ (0 is center). The main aim of this feature is to calculate the stereo panning information for different frequencies based on the STFT of left and right channel. The basic idea behind the stereo panning spectrum is to compare the left and right channel signals in the time-frequency [70] representation to derive a 2-dimensional map that identifies the different panning gains associated with each time-frequency bin. Many statistical features can be derived from the stereo panning spectrum like: derivative of panning index, panning root mean square for a particular frequency band, panning for low medium and high frequencies etc. This feature has been used in audio classification, separating audio sources [71], music information retrieval [69] and music classification [72].

  – *Group delay function (GDF):* During frequency domain analysis of an audio signal, we generally avoid the phase information and do analysis based on the real values. The information in STFT phase function is extracted by calculating the derivative of the phase, this is called as group delay function [73,75]. It is also known as negative derivative of the phase. Group delay function reveals information about the temporal events in a signal e.g. identifying spectral peaks. The GDF can be: minimum phase GDF, maximum phase GDF or mixed phase GDF. Generally formants are extracted from GDFs and are used in various audio processing algorithms. Another type of GDF is modified GDF,

in which the cepstral smoothing is done prior to GDF computation which smooths the intrinsic spikes of the signal. GDF is used in speaker identification [74], speech segmentation [73], language identification etc.

- **Envelope Modulation Spectrum (EMS) based**: The envelope modulation spectra represents the slow amplitude modulations in the signal and distribution of energy in amplitude fluctuations across certain frequencies. The original audio signal is first filtered into 9 octave bands with center frequencies of approximately 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680 Hz, using eight-order Butterworth filters. Then the envelope of 10 signals: one original and 9 octave filtered signals is extracted using Hilbert transform. Once the envelope of each signal is extracted, the mean is removed and the power spectrum for each of the bands is estimated by evaluating the DFT using the Goertzel algorithm at frequencies 0–10 Hz. From the power spectrum, six EMS metrics are computed for each of the 9 octave bands, and the full signal and it makes a 60-dimensional feature vector. The 6 EMS features are:
  - *Peak frequency 0–10 Hz:* It is the frequency corresponding to the spectral peak with the maximum magnitude in the spectrum.
  - *Peak amplitude:* It is the absolute amplitude of the spectral peak described above and it is seen as a measure of rhythm.
  - *Energy in spectrum 3–6 Hz:* Energy in band 3–6 Hz is the normalized energy present in the band and it corresponds to the segment 167–333 ms which captures majority of syllable duration in normal speech.
  - *Energy in spectrum 0–4 Hz:* Human auditory system is most sensitive around 4 Hz of modulation, this peak sensitivity corresponds to the segment 0–250 ms which is quite close to common syllable rate for speech.
  - *Energy in spectrum 4–10 Hz:* This region of frequencies focus on super segmental variations in the rhythm.
  - *Energy ratio between 0–4 Hz and 4–10 Hz:* The energy below 4 Hz is uncorrelated with the energy in band 4–10 Hz. Hence the ratio of these energies becomes a vital feature.

  The EMS features is best used in classifying pathological speech and control speech [47,48].

Algorithm 11: Envelope modulation energy in desired frequency band

Result: Energy in desired frequency band
1. Input: Audio signal x
2. Find the spectrogram and cyclic frequencies of the signal.
3. Select the range of cyclic frequencies.
4. Calculate square of the spectrogram values.
5. Find summation of the squared spectrogram values.
6. Multiply the result number by 2, this gives energy in a desired band.

- **Long-term Avearge Spectrum (LTAS):** LTAS captures atypical spectral information from the signal. LTAS [49] is used in classifying pathological speech (like dysarthria, dyphasia etc.) from the controlled normal speech. The intelligibility of a speech is determined by the variations in nasality, breath control and loudness of the speech. LTAS tries to capture these cues from each octave filtered speech signal. There are 9 octave filters used at center frequencies 30, 60, 120, 240, 480, 960, 1920, 3840 and 7680 Hz. From each octave filtered speech signal the following parameters are determined and joined together to form a 99-dimensional feature vector.
  - RMS value normalized/full band RMS value
  - Normalized mean frame RMS
  - Standard deviation of frame RMS
  - Frame standard deviation normalized by full-band RMS
  - Frame standard deviation normalized by band RMS
  - Skewness of frame RMS
  - Kurtosis of frame RMS
  - Range of frame RMS
  - Normalized range of frame RMS
  - Pairwise variability of RMS energy between ensuing frames
- **Chroma related features:** Chroma features are interesting and powerful representation of music audio in which entire spectrum is mapped into 12 bins that represents the 12 semitones (or chroma) of musical octave. It can be computed from the logarithmic short-time fourier transform of the sound signal. It is also called as chromagram. Another chroma based feature is chroma energy distribution normalized statistics. This feature is used to identify similarity between different interpretations of the music given.
- **Tonality based features:** The tonal sounds are actually the fundamental frequency of a harmonic stationary audio signal. In music, tonality organizes the notes of the musical scale. The main tonality based audio features are:
  - *Fundamental Frequency (FF):* The fundamental frequency F0 is the first peak of the local normalized spectro-temporal auto-correlation function. In simple terms, fundamental frequency is the lowest frequency of a periodic waveform. For music, the fundamental frequency is the musical pitch of a note that is perceived as the lowest partial present. Fundamental frequency is an important feature and used in music onset detection [61], environmental sound recognition [39,40]and audio retrieval [63].

Algorithm 12: Fundamental frequency F0

1. **Result**: Fundamental frequency (F0)
2. Input: Audio signal x and sampling frequency fs.
3. Find the pitch of the audio signal by auto-correlation or cepstral methods.
4. Pitch will give an estimate of the fundamental frequency of the signal x.

  - *Pitch Histogram:* Pitch histogram explains the pitch of a signal in a more compact form. It has been highly used in music genre classification [15].
  - *Pitch Profile:* This feature is more accurate representation of musical pitch and it has been used for musical key detection [64].
  - *Harmonicity:* Harmonicity is a feature used to distinguish between tonal and noise like sounds. It uses auto-correlation function to find periodicity in sound in time or frequency domain.
  - *Harmonic-to-noise ratio:* It is computed as the ratio between the harmonic part of the signal to the rest of the signal. It has been used in analyzing pathological voices [65] and music-related applications.
  - *Jitter:*It computes the variations of fundamental frequency, that is, the average absolute difference between consecutive periods of speech. It is used in speaker recognition [67], analyzing pathological voices, determine vocal and non-vocal sounds [66].
- **Spectrum shape based features:**
  - *Spectral Centroid:* Spectral Centroid indicates the where the center of mass of the spectrum is located. It describes

the brightness of a sound signal and hence also called brightness feature of a sound. It is computed considering the spectrum as a distribution which values are the frequencies and the probabilities to observe these are the normalized amplitude. Spectral centroid is an excellent measure of brightness of sound signal and used to measure of timbre of music [53], music classification [29,30], music mood classification [14,54]. Spectral Centroid of the music is higher than the spectral centroid of speech [56]. The basic algorithm for calculating spectral centroid is given below.

Algorithm 13: Spectral centroid of an audio signal

1. **Result**: Spectral Centroid μ_1
2. Input: Audio signal x and sampling frequency fs.
3. Convert the signal into frequency domain.
4. $\mu_1 = (\sum_{k=b_1}^{b_2} f_k s_k)/(\sum_{k=b_1}^{b_2} s_k)$

where f_k is the frequency corresponding to bin k and s_k is the spectral value at bin k, b_1 and b_2 are band edges

– *Spectral Center:* This feature is closely related to the spectral centroid. It is the measure of the median frequency present in the signal spectrum. This is a median frequency hence it balances the higher and lower energies. This feature is applied in tracking rhythm in musical signals [55].
– *Spectral roll-off:* The spectral roll-off point is the frequency so that 95% of the signal energy is contained below this frequency. Spectral roll of is used in speech/music classification [56], music genre classification [14,15,29,30], musical instrument classification and audio-based surveillance systems [41].

Algorithm 14: Spectral Roll-off point of an audio signal

1. **Result**: Spectral Roll-off
2. Input: Audio signal x and sampling frequency fs.
3. Transform the signal in frequency domain.
4. $roll\ off\ point = i$ such that $\sum_{k=b_1}^{i} s_k = 0.95(\sum_{k=b_1}^{b_2} s_k)$

where s_k is the spectral value at bin k and b_1 and b_2 are band edges

– *Spectral Spread:* It is also called as Spectral Dispersion. This feature is closely related to the bandwidth of the signal [55]. It can be described as average deviation of the rate-map around its centroid. Noise like signals have wide spectral spread than the pure tonal sounds which has small spectral spread. It is generally, wide for music and environmental sounds and comparatively narrow for speech like sounds [56].

Algorithm 15: Spectral Spread of an audio signal

1. **Result**: Spectral Spread μ_2
2. Input: Audio signal x and sampling frequency fs.
3. Transform the signal in frequency domain.
4. $\mu_2 = \sqrt{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^2 s_k / \sum_{k=b_1}^{b_2} s_k}$

where f_k is the frequency corresponding to bin k and s_k is the spectral value at bin k,
b_1 and b_2 are band edges and μ_1 is spectral centroid

– *Spectral Skewness:* Spectral skewness is the 3rd order statistical value and it measure the symmetry of the spectrum around its arithmetic mean value. This feature would be equal to zero for silent segments and high for voiced parts.The Skewness is the 3rd order statistical feature. Skewness equal to zero describes symmetric distribution, skewness less than zero indicates more energy to the right side of spectral distribution and skewness greater than zero describes more energy components are present on the left side of the spectrum. This feature is used in mood detection [14,26] and music genre classification [29,32,46,57], fault detection in motor bearings [150] and Parkinson's disease detection from speech [144]. Fig. 15 represents the skewness for different spectrum.

Algorithm 16: Spectral skewness of an audio signal

1. **Result**: Spectral Skewness
2. Input: Audio signal x and sampling frequency fs.
3. Transform the signal into frequency domain.
4. $skewness = \sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k / (\mu_2)^3 \sum_{k=b_1}^{b_2} s_k$

where f_k is the frequency corresponding to bin k and s_k is the spectral value at bin k,
b_1 and b_2 are band edges, μ_1 is spectral centroid and μ_2 is spectral spread

– *Spectral Kurtosis:* On the other hand, Kurtosis is the 4th order statistical measure and describes the flatness of the spectrum around its mean value. For gaussian distribution the spectral kurtosis has value 0, if kurtosis is less than 0, we observe flat distribution and if spectral kurtosis is greater than 0, we observe sharp peaked spectral. Just like spectral skewness, spectral Kurtosis is also used as a feature in music genre classification [29,32,57] and mood classification [14,26], fault detection in bearing of electric motors [150] and Parkinson's disease detection from speech [144]. Fig. 16 explains the kurtosis for different type of spectrum for a function $f(x)$ or audio signal x.

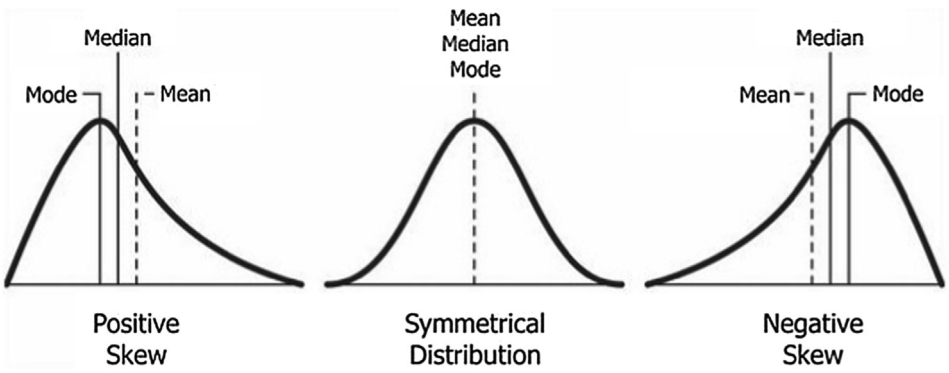


Fig. 15. Skewness for different type of spectrum.

Algorithm 17: Spectral kurtosis of an audio signal

1. **Result**: Spectral Kurosis
2. Input: Audio signal x and sampling frequency fs.
3. Transform the signal into frequency domain.
4. $kurtosis = \sum_{k=b_1}^{b_2}(f_k - \mu_1)^4 s_k / (\mu_2)^4 \sum_{k=b_1}^{b_2} s_k$

where f_k is the frequency corresponding to bin k and s_k is the spectral value at bin k,
b_1 and b_2 are band edges, μ_1 is spectral centroid and μ_2 is spectral spread

– *Spectral Slope:* It is the measure of slope of the amplitude of the signal and it is computed by linear regression. This feature is used in speech analysis [58] and in identifying speaker from a speech signal [59].

Algorithm 18: Spectral slope of an audio signal

1. **Result**: Spectral slope
2. Input: Audio signal x and sampling frequency fs.
3. Transform the signal into frequency domain.
4. $slope = \sum_{k=b_1}^{b_2}(f_k - \mu_f)(s_k - \mu_s) / \sum_{k=b_1}^{b_2}(f_k - \mu_f)^2$

where f_k is the frequency corresponding to bin k
s_k is the spectral value at bin k
b_1 and b_2 are band edges
μ_s is mean spectral value
μ_f is the mean frequency

– *Spectral Decrease:* It explains the average spectral-slope of the rate-map representation, putting a strong emphasis on low frequencies [60].

Algorithm 19: Spectral slope of an audio signal

1. **Result**: Spectral decrease
2. Input: Audio signal x and sampling frequency fs.
3. Transform the signal into frequency domain.
4. $spectral\ decrease = \sum_{k=b_1+1}^{b_2}((s_k - s_{b_1})/(k-1)) / \sum_{k=b_1+1}^{b_2} s_k$

where s_k is the spectral value at bin k and b_1 and b_2 are band edges

– *Bandwidth:* Spectral bandwidth is the second order statistical value the determines the low bandwidth sounds from the high frequency sounds. It is used in music classification [15] and environmental sound recognition [16].
– *Spectral Flatness:* Spectral flatness is the measure of the uniformity in the frequency distribution of the power spectrum. It is calculated as the ratio of the geometric mean to the arithmetic mean. It can be used to distinguish between harmonic and noise like sounds. For harmonic sounds the spectral flatness is close to zero and for noise like sounds the value of spectral flatness is close to one. It is employed in music onset detection [61], music classification and audio fingerprinting [59].

Algorithm 20: Steps to calculate spectral flatness

1. **Result**: Spectral flatness value
2. Input: Audio signal x.
3. calculate the periodogram power spectral density of x.
4. Find the geometric mean of the periodogram signal.
5. Find arithmetic mean of the periodogram signal.
6. Calculate the ratio of geometic and arithmetic mean.
7. Result is the value of spectral flatness.

– *Spectral Crest Factor:* In contrast to the spectral flatness, spectral crest factor determines how peaked is the power spectrum of the sound signal. It is also used to distinguish between harmonic/tonal sounds and noise like sounds. It is higher for harmonic/tonal sounds and lower for noise like sounds. This feature is also used for audio fingerprinting [59] and music classification [29].

Algorithm 21: Spectral crest factor algorithm

1. **Result**: Spectral crest factor of an audio
2. Input: Audio signal x.
3. Calculate peak amplitude of the signal x.
4. Calculate root mean square value of x.
5. Crest factor is the ratio between peak amplitude and RMS value.
6. Convert the crest factor in decibles (if required).

– *Entropy:* Entropy is also the measure of uniformity of flatness, and it is computed as Shannon's entropy or Renyi entropy. This feature has been used for automatic speech recognition [62].
The Shannon's entropy for a signal could be calculated by using formula $-sum(P_i log P_i)$ where P_i is the sample class probabilities.
– *Spectral Flux:* The spectral flux is defined as 2-norm of the frame-to-frame spectral amplitude difference vector. It points the sudden changes in the frequency energy distribution of sounds. This feature is used in speech/music discrimination [27], music genre classification [29,32]and environmental sound classification [39,40].
– *Octave based spectral contrast (OBSC):* It is the difference between peaks and valleys measured in sub-bands by octave scale filters. It has been used for music classification [29]and music mood classification [14,26].

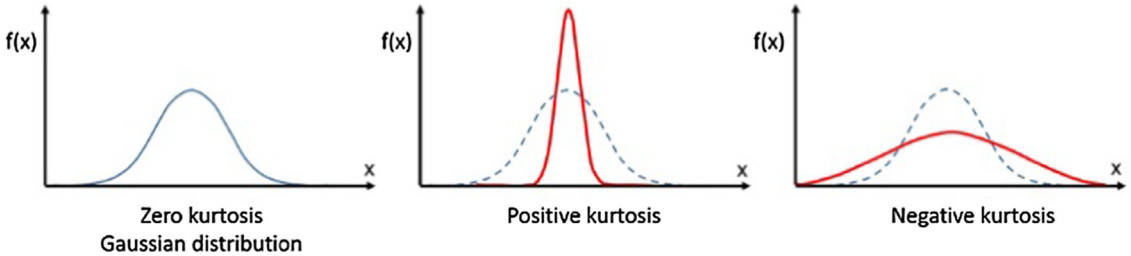


Fig. 16. Kurtosis for different type of spectrum.

Algorithm 22: OBSC extarction algorithm

1. **Result**: OBSC feature
2. Input: Audio signal x.
3. Perform framing or windowing on the signal.
4. Take FFT of each framed signal.
5. Divide the transformed signal into 6 octave scaled sub-bands.
6. For each band, calculate peak and valley values.
7. Calculate the contrast of the peaks and valleys.
8. Valleys and contrast gives OBSC feature set.

4.3. Cepstral domain features

A cepstrum is obtained by taking the inverse Fourier transform of the logarithm of the spectrum of the signal. There is a complex, power, phase and real cepstrum. Among all of these, power cepstrum is the one most relevant to the speech signal processing. The analysis of the cepstrum is called as cepstrum analysis, quefrency analysis (equivalent to frequency analysis in spectrum domain) or liftering [76](equivalent to the filtering in spectrum domain). The cepstrum features are mainly used in pitch detection [77,78], speech recognition and speech enhancement [79]. The cepstrum/cepstral features are discussed below:

- *Mel Frequency Cepstral Coefficients (MFCCs):* MFCCs are derived from the cepstral representation of an audio clip. MFCCs represents short-time power spectrum of an audio clip based on the discrete cosine transform of log power spectrum on a non-linear mel scale. In MFCCs the frequency bands are equally spaced on mel-scale, which mimics the human auditory system very closely, making MFCCs key feature in various audio signal processing applications. MFCCs has been widely used in speech recognition [80,83], speech enhancement [84], speaker recognition [81], music genre classification [82], music information retrieval [85], audio similarity measurement [82], vowel detection [145] etc.

Algorithm 23: MFCC extraction

1. **Result**: Mel frequency cepstrum coefficients
2. Input: Audio signal x
3. Frame the signal into short frames. ▷Use windowing
4. For each frame, calculate the periodogram estimate of the power spectrum.
5. Apply the mel filter-bank to power spectrum, sum the energy in each filter.
6. Take logarithm of filter-bank energies.
7. Take Discrete cosine transform (DCT) of the log filter-bank energies.
8. Keep 2–13 DCT coefficients, discard the rest.

- *Linear Prediction Cepstral Coefficients (LPCCs):* The cepstrum has number of advantages like soucre-filter separation, orthogonality, compactness etc. These properties makes cepstrum coefficients robust and suitable for machine learning. On the other hand, linear prediction coefficients (LPC) are too sensitive to numerical precision, hence it is desirable to transform the LPC to the cepstral domain. The resultant transformed coefficients are called as LPCCs. We can say that LPCCs are derived from LPCs. LPCCs has been used in various application areas like speech recognition [86], speech analysis [87], noise removal [88], music genre classification [89] etc. Algorithm below describes the steps to calculate LPCCs from an audio signal.

Algorithm 24: LPCCs extraction

1. **Result**: LPCCs feature
2. Input the audio signal x
3. Perform pre-emphasis on original signal.
4. Frame the signal using windowing method.
5. Find the auto-correlation of the signal with itself.
6. Calculate LPC parameters.
7. Convert the LPC parameters in cepstral domain.
8. Result is LPCCs.

- *Preceptual linear prediction (PLP) cepstral coefficients:* The PLP coefficients is based on the concepts: critical band spectral resolution, equal-loudness curve and intensity loudness power law [90]. The PLP coefficients are derived from the linear prediction coefficients (LPC) by performing perceptual processing before auto-regressive modelling. After this processing, the linear coefficients are converted into cepstral coefficients. It has been widely employed in emotion recognition [91], speech recognition [92], baby crying sound analysis [93], animal sound vocalization analysis [94] etc.

Algorithm 25: PLP cepstral extraction

1. **Result**: PLP cepstral coefficients
2. Input the audio signal x
3. Perform windowing on signal.
4. Do critical band analysis.
5. Perform equal loudness and pre-emphasis.
6. Do intensity-loudness conversion.
7. Apply linear prediction algorithm.
8. Convert into cepstarl domain.

- *Relative-spectral PLP (RASTA-PLP) feature:* RASTA is a technique that applies a band-pass filter to the energy in each frequency sub-band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech signal. The RASTA-PLP features are robust to noise version of PLP features. The advantage of the RASTA-PLP feature is that it tries to incorporate noise cancellation feature of human auditory system. This hybrid feature is widely employed in speech recognition [95], gender classification [96], speaker verification [97] etc. The process to extract RASTA-PLP features is explained in algorithm below.

Algorithm 26: RASTA-PLP feature extarction

1. **Result**: RASTA-PLP feature set
2. Input the audio signal x
3. Perform pre-emphasis and windowing on original signal.
4. find DFT of the windowed signal.
5. Perform critical bank analysis on DFT signal.
6. Take logarithm of the result of step 5.
7. Perform RASTA-filtering. 8. Perform equal loudness and pre-emphasis on RASTA filtered signal.
9. Apply Intensity loudness power law.
10. Take inverse of logarithm of the result of step 9.
11. Perform auto-regressive modelling.
12. Convert into cepstral domain.
13. Result is the RASTA-PLP coefficients.

- *Greenwood function cepstral coefficients (GFCC):* GFCCs [98] were introduced as a generalized form of MFCCs. GFCCs use mel-scale features that mimics the properties of HAS and theoretically well-founded for nearly all terrestrial mammals and give good vocal representation for nearly all species. GFCCs can be implemented using very basic knowledge of the minimum and maximum frequency range for a particular species and is derived from greenwood equation. Greenwood equation nearly maps the cochlear-frequency position for all species. This feature is highly employed in environmental sound recognition specially for animals and bird sound classification [99].

Algorithm 27: GFCCs extraction

1. **Result**: Greenwood function cepstral coefficients
2. Input: Audio signal (x)
3. Frame the signal into short frames. ▷Use windowing
4. For each frame, calculate the periodogram estimate of the power spectrum.
5. Apply the Greenwood-function scaled filter-bank to power spectrum, sum the energy in each filter.
6. Take logarithm of filter-bank energies.
7. Take Discrete cosine transform (DCT) of the log filter-bank energies.
8. Result is GFCCs.

- *Gammatone cepstral coefficients (GTCCs):* The main problem with the automatic speech recognition (ASR) systems is noise reduction. In recent years, the GTCCs has shown noise robustness in many ASR systems. GTCCs are based on gammatone filter banks, these filter banks give cochleagram as the output which is actually the frequency-time representation of a sound signal. The extraction process of GTCCs is similar to the extraction process of MFCCs except the mel-filter bank is replaced by gammatone filter bank. Just like MFCCs, GTCCs can also have additional features like delta GTCCs, delta-delta GTCCs which are actually the first and second order derivatives of GTCCs. These group of cepstral features are used in environmental sound recognition [100] and speech recognition [101].

Algorithm 28: GTCCs extraction

1. **Result**: Gammatone cepstral coefficients
2. Input: Audio signal (x)
3. Frame the signal into short frames. ▷Use windowing
4. For each frame, calculate the periodogram estimate of the power spectrum.
5. Apply the gammatone filter-bank to power spectrum, sum the energy in each filter.
6. Take logarithm of filter-bank energies.
7. Take Discrete cosine transform (DCT) of the log filter-bank energies.
8. Result is GTCCs.

4.4. Discrete wavelet transform domain features

The wavelet transform is another way to transform the time-domain audio signal into a time-frequency representation. It computes the inner product of the signal with a member from family of wavelets. There are two types of wavelets: continuous wavelet transform (CWT) and discrete wavelet transform (DWT). The wavelet specially DWT has the capacity to extract information from non-stationary signals like audio. It overcomes the shortcomings of the STFT that provides uniform time-frequency resolution. DWT gives high time resolution and low frequency resolution for higher frequencies and high frequency resolution and low time resolution for lower frequencies. The approximations and detailed coefficients are generated by the wavelet transform that gives the information about a signal. These approximations and detailed coefficients are called as wavelet features. These wavelet features could be extracted either from wavelet transform or from wavelet packet decomposition. In wavelets the approximation coefficients are decomposed while in wavelet packet decomposition/transform both approximation and detailed components are decomposed. Below Fig. 17 explain the difference between wavelet transform and wavelet packet transform.

The conventional features like MFCCs, PLPC etc could be extracted from the wavelet packet decomposition. Or these coefficients could be directly used as wavelet features. These features or coefficients are used in audio analysis [122], audio classification [123,136], audio fingerprinting [124], content-based audio retrieval [125], music classification [126], vowel detection in speech signals [145], snore sound analysis [133] audio-visual emotion recognition [134], detecting fault bearing in electric motors [150] etc.

4.5. Image/texture based features

- *Local Binary Pattern (LBP):* Local binary pattern is primary used for computer vision applications like face detection, face recognition, object detection etc. LBPs measures the local spatial information and gray scale contrast. In audio signal processing, the LBPs are extracted from the spectrograms of the signals and used in audio scene classification [112,113], depression analysis from speech [114] snore sound discrimination [115], emotion detection [117], and pathological voice (Cordectomy and frontolateral resection diseases) detection [143].

Algorithm 29: Local Binary pattern from spectrogram

1. **Result**: LBP feature set
2. Input: Audio signal x.
3. Generate spectrogram from the audio signal.
4. Convert the RGB image of spectrogram into gray scale image.
5. Choose the radius of the mask and type of normalization.
6. Extarct LBP features from the gray scale image.
7. LBP are generated from each masked image subset.

- *Local Ternary Patterns (LTPs):* LTPs are the extension of the LBPs. Just like LBPs the LTPs are extracted from the image description of a signal like a spectrogram. But unlike LBPs, it doesn't threshold the pixels into binary pattern of 0's and 1's, rather it uses a threshold constant to threshold pixels into three values i.e. −1, 0 and 1. In this manner each threshold pixel could have any of these three values and neighbouring pixels could be combined after thresholding into a ternary patterns.The LTPs are used in audio scene clas-

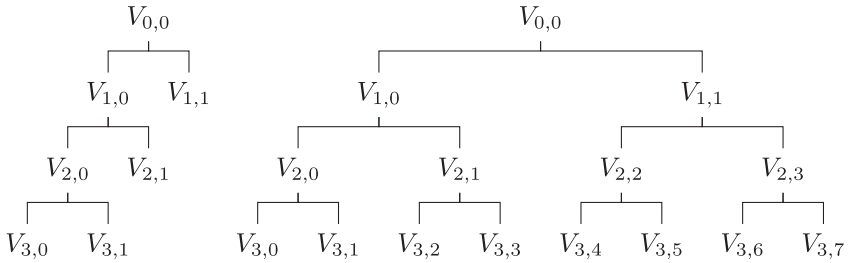


Fig. 17. Left: Wavelet Decomposition, Right: Wavelet packet decomposition.

sification [140], fall detection in elderly people by analyzing environment sounds [141], and in health care monitoring using speech analysis [142]. The algorithm to extract LTPs is explained below:

Algorithm 30: Local Ternary Pattern Extraction

1. **Result**: Local ternary patterns
2. Input: Audio signal *x*.
3. Generate spectrogram from the audio signal.
4. Convert the RGB image of spectrogram into gray scale image.
5. Consider the image mask from where LTPs are to be extracted.
6. Calculate the lower and upper bits using ternary function and three threshold values −1, 0 and 1.
7. Calculate lower and upper values using lower and upper bits.
8. Construct lower and upper signals using lower and upper bits.
9. Generate histogram from these signals.
10. Join the histograms to get LTPs.

- *Histogram of gradients (HOG) feature:* Just like LBPs, HOG features are used to extract the time-frequency information from the spectrograms. This feature also has been used in acoustic scene classification [113,116],snore sound discrimination [115], emotion detection [117].

Algorithm 31: HOG descriptor from spectrogram

1. **Result**: HOG descriptor
2. Input: Audio signal *x*.
3. Generate spectrogram from the audio signal.
4. Convert the RGB image of spectrogram into gray scale image.
5. Choose name-value pair arguments such as cell size, block size etc.
6. Extaract HOG features from spectrogram.
7. Output is HOG feature from each cell or block of the image.

- *Scale invariant feature transform (SIFT):* SIFT is also a feature extraction algorithm in computer vision used initially for detecting local information in images. This feature is also employed on spectrograms of audio signals to detect the local information. It has been highly used in emotion detection [117] and audio-video concept classification [118].

Algorithm 32: SIFT descriptor

1. **Result**: SIFT feature set
2. Input: Audio signal *x*.
3. Generate spectrogram from the audio signal.
4. Convert the RGB image of spectrogram into gray scale image.
5. Perform scale-space extrema detection on gray scale image.
6. Localize the keypoints.
7. Assign orientation to each keypoint.
8. Create keypoint descriptor called as SIFT descriptor.

4.6. Deep features

Deep learning has been proven to be a powerful technique to extract high level features from low level information. The features extracted from the hidden layers of various deep learning models is known as *deep features*. The deep features could be extracted from any deep leaning model like convolutional neural networks (CNNs), deep neural networks (DNNs), recurrent neural networks (RNNs), Deep stacked auto-encoder (SAE), unidirectional long short term memory network (LSTM), bi-directional long short term memory (BLSTM) and other similar models.

Deep features are extracted from the deep neural networks (DNNs). The MFCCs or any other relevant audio feature is fed to the DNNs as the input. The deep features depend on how deep the neural network is. If we have shallow neural network, the deep features given by lower layers can be thought of as speaker-adapted features. And from the upper layers class-based discriminatory features could be extracted. The deep features could also be extracted from the bottleneck layer of a DNN. Fig. 18 shows the extraction of deep features from bottleneck layer of a DNN. In this figure DAF stands for deep audio features).

Any CNN architecture consists of three major components: convolution layers, pooling layers and fully connected layers as shown in Fig. 19. Convolution layers apply definite number of convolutional filters on the spectrogram of an audio signal. The output of this layer is called as feature map. Pooling layer decrease the dimensions of the feature maps generated by convolutional layers and hence reduce the processing time. Fully connected layers extract global features from the local feature maps. The deep features could be extracted from any of these layers, the initial layers of the CNN gives deep features which is nothing but the information about pixels and edges of the spectrogram. The higher layers gives deep features which have highly discriminative. while the deep features extracted from the fully connected layer gives global

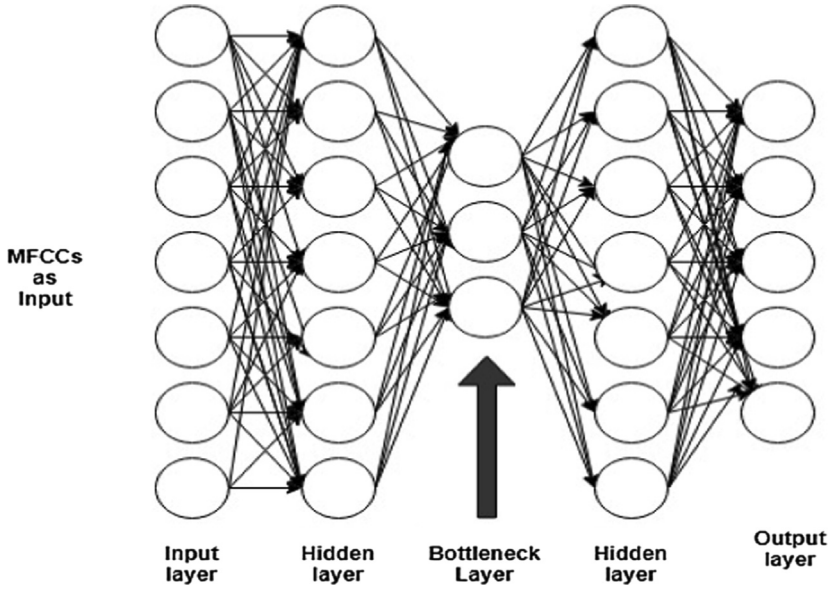


Fig. 18. Deep feature from bottleneck layer.

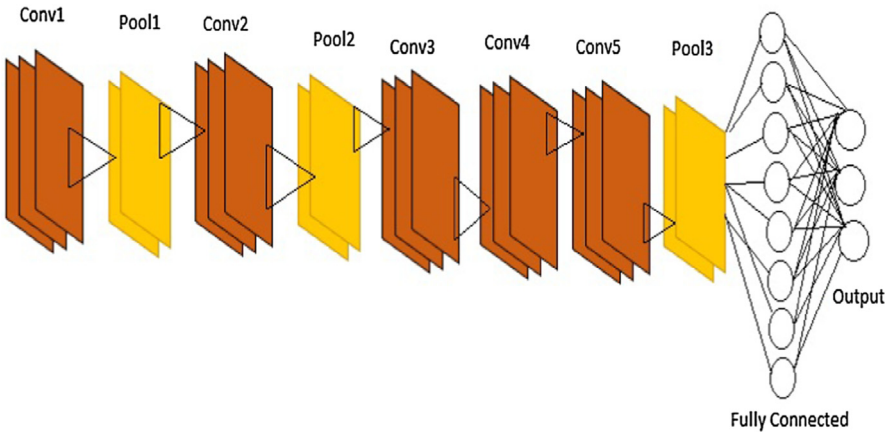


Fig. 19. Basic structure of CNN.

information. Hence, the significance of the deep features highly depends on the layer from which these are extracted.

Another type of artificial neural network is auto-encoder, which is basically used for dimension reduction and works in an unsupervised manner. The stacked auto-encoder (SAE) is a neural network having multiple layers of sparse auto-encoders in such a manner that the output of the each layer is fed as the input to its successive layer. Once the SAE is trained, the deep features could be extracted from the hidden deep layers of the unsupervised trained SAE and can be used for various applications.

In order to encode the sequential knowledge in the model, RNNs are used. RNNs have advantages over DNNs as it has flexibility to process sequential knowledge and have ability to memorize information internally. The memory unit is called as LSTM. Unidirectional or bi-directional LSTM are the types memory element. In unidirectional LSTM the flow of information is in just forward direction while in BLSTM, one LSTM layer process the information in forward direction and another LSTM layer process the information in backward direction. The deep features could be extracted from any deep layer of LSTM-RNN.

In audio signal processing, deep features have been used in acoustic scene classification [127,128], speaker recognition [130], audio-video analysis [129], gender recognition [146], emotion recognition [131] and spoofing detection [132].

4.7. Sparsity in features

In numerical analysis, a sparse matrix is the one in which most of the elements is zero and very few elements are non-zero. The same concept can be applied to audio signals, which may have very few non-zero components when represented in a particular domain. For example, a pure tonal signal can be represented by one single spike in frequency domain, hence we can say that signal is sparse in frequency domain. Hence, very less number of features are required to represent a signal if it has sparsity. Sparsity can be achieved in any domain say time domain, frequency domain, time-frequency domain, wavelet domain, cepstral domain etc. Once the sparsity is achieved in any of the domains, the domain specific features can be extracted and analyzed. For example, if sparsity is present in cepstral domain, cepstral features can be extracted from the compact sparse signal. This sparsity of a signal may finally leads to compressing sensing. The sparsity can be realized by many methods, to name a few are:

- Matching Pursuit (MP)
- Orthogonal matching pursuit
- Stage wise greedy method
- Basic pursuit
- Coordinate descent etc.

The sparsity plays an important role in compressive sensing of speech signals [108], speech/music separation [109] and scalable audio classification [110].

4.8. Other domains

- *Eigen Domain features:* The eigen domain features are the set of features extracted from the eigen vectors of an audio signal. Eigen vector of an audio signal is the most dominant vector/dimension present in the signal. The most dominant vector [111] can be obtained by various techniques, the most exploited is principal component analysis (PCA). Other techniques are independent component analysis (ICA) and singular value decomposition (SVD). These techniques project the original audio signal to the eigen-vector space. Most relevant eigen domain features are MPEG-7 audio spectrum basis feature and distortion discriminant analysis feature.
- *Phase Domain:*
  – *Modified group delay function (MODGDF):* The modified group delay function gives better spectral smoothing than the standard group delay function. In context of speech, zeros of the slowly varying envelope of speech represents the nasal sounds. The zeros in speech are either within or outside the unit circle since the zeros also have nonzero bandwidth and produce spikes in the spectrum. In MODGDF these zeros are suppressed leading to a smoother spectrum of the sound. From this MODGDF, cepstral coefficients could be extracted and are called as MODGDF cepstral coefficients. This feature has been used in speech recognition [102], speaker verification [103] and synthetic speech detection [104,105].
  – *cosine normalized phase cepstral feature:* These cepstral features are derived by um-wrapping the phase spectrum of a signal. The cosine normalization is done on this unwrapped signal and the cepstral coefficients are extracted

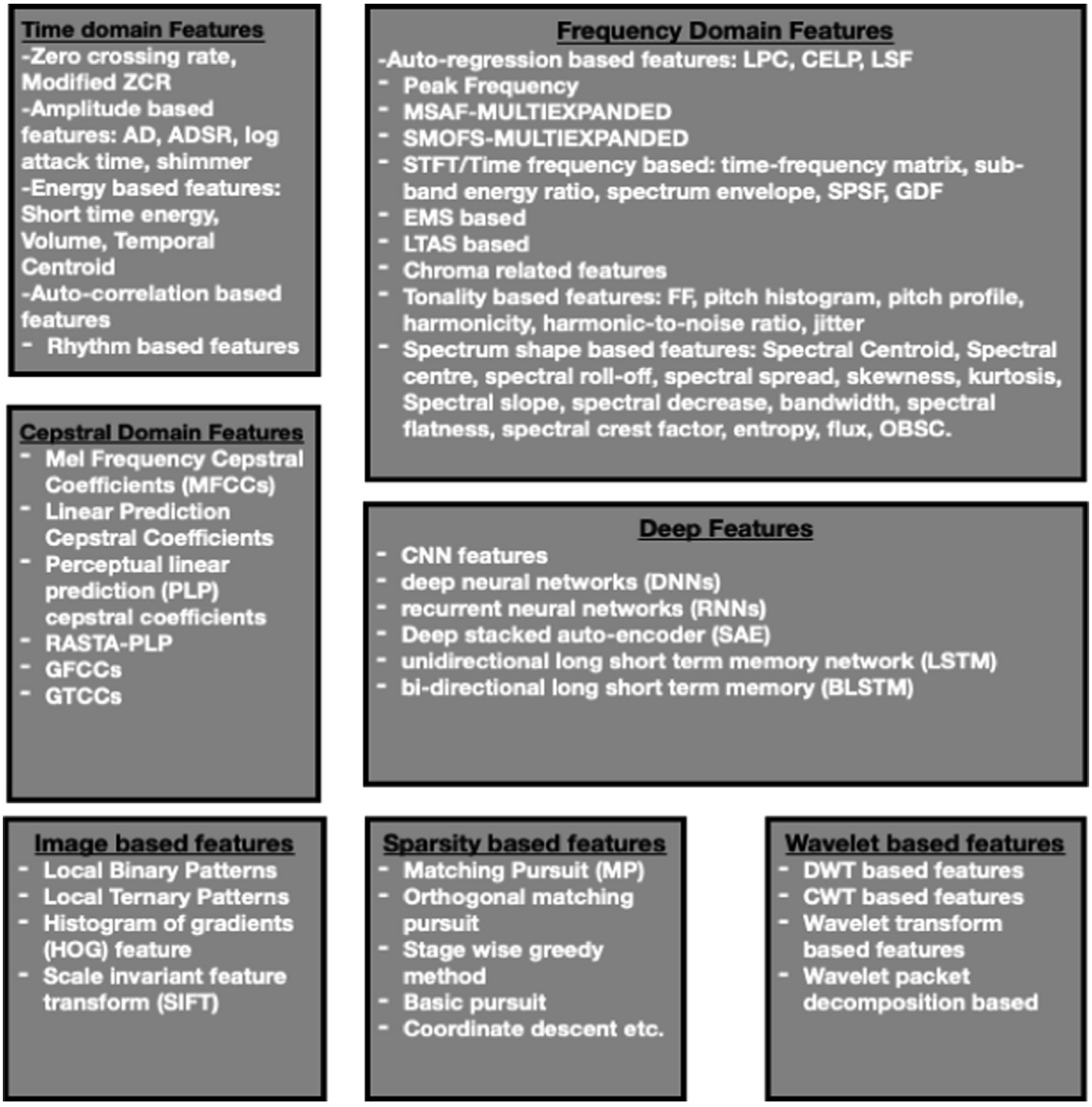


Fig. 20. Benchmarking of audio features discussed.

by performing discrete cosine transform on the normalized signal. This feature is used in spoofing detection [106] and speaker recognition [107].

5. Critical analysis and conclusion

In this work we have discussed about the various thresholds of human auditory system and the classification of sounds into speech, music and environmental sounds. This review also covers the evolution of the audio features in early 1950's and their progress till date. This paper discuss the features from time domain, frequency domain, time-frequency domain, cepstral domain, phase domain,sparse domain, eigen domain, wavelet domain and image/-texture based features. Fig. 20 gives the benchmarking of the audio features discussed in this review article.

In this review work, we try to relate the current audio features and their extraction algorithms which are suitable for machine learning or machine hearing in the most popular domains of

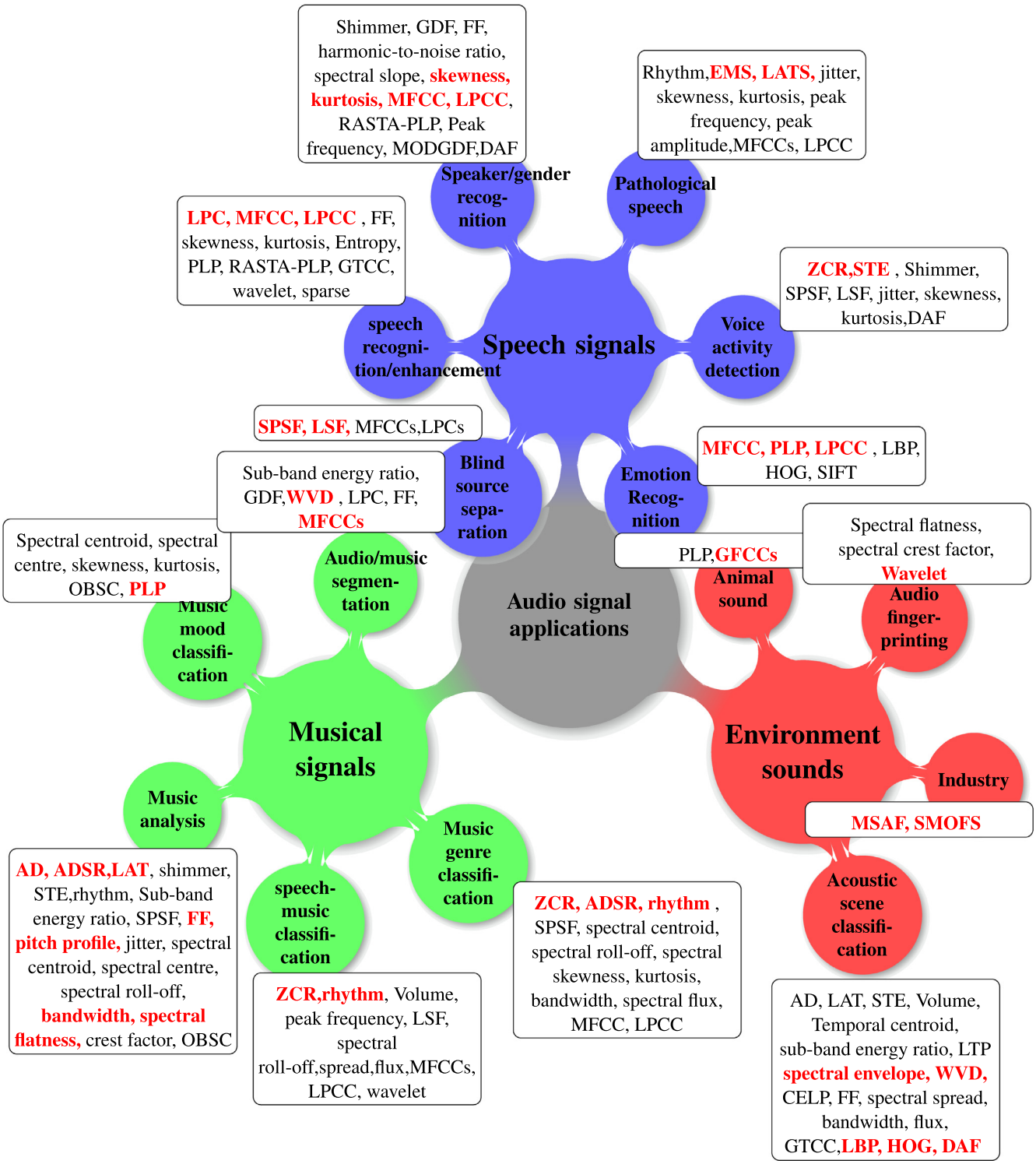


Fig. 21. Summary of audio signal features according to their application area (Red color features shows most prominent features).

audio i.e. speech, music and environmental audio and their applications.

Furthermore, we have tried to explain the pseudo code or algorithm whatever applicable to explain the extraction process of an audio feature. The MATLAB (version 2019) commands are mentioned wherever it is directly available for use, otherwise the extraction algorithm is provided to have better understanding.

Fig. 21 gives the bird's eye view to the audio application areas and the list of respective audio features which are most suitable for that very particular application. The features highlighted in red color shows the most prominent features for that very particular application. The application areas exploited by the researchers the most are: human speech based applications, applications involving music signals and applications related to environmental sounds. For speech signals, the most researched areas include speech recognition, speech enhancement, speaker/gender recognition, voice activity detection, pathological speech analysis, blind source separation and emotion recognition. Similarly, for music based applications, the main focus of the state-of-the-art is on music segmentation, music mood classification, music analysis, speech-music classification and music genre classification. For environmental sounds, the most of the research converges into applications such as acoustic scene classification, audio fingerprinting, Industry applications and animal sound classification.

For instance say under the category speech the one of the most popular area of research is speaker/gender recognition. According to the literature survey only few audio features have proven to be most discriminatory and suitable for machine learning. In this case shimmer, group delay function (GDF), fundamental frequency (FF), harmonic-to-noise ratio, spectral slope, skewness, kurtosis, MFCCs, LPCC RASTA-PLP and peak frequency are mostly used features. It could be concluded that, not every feature provides good results for every application. A researcher must look for the best hand crafted features which are suitable for a particular application.

In future, a more comprehensive analysis of the audio signals, acoustic signals and vibrations could be done. A detailed discussion on feature and its behaviour with audio, acoustic or vibration signal would be an interesting analysis. We expect the change in trends in audio signal feature extraction methods in future and would like to analyze those new and emerging features used in machine hearing and their relevant application areas.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.apacoust.2019.107020.

References

[1] Sound M. Naerum. Bruel and Kjaer Co.; 1984.
[2] Miller GA. The perception of speech McGraw-Hill. New York, NY: McGraw-Hill; 1951. https://doi.org/10.1037/11135-003. Retrieved fromhttp://ezproxy.lib.ryerson.ca/login?url=https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/614171122?accountid=13631.
[3] Johnson Keith. Acoustic and auditory phonetics. 3rd ed. Wiley-Blackwell; 2015.
[4] Lyon RF. Machine hearing: an emerging field [exploratory dsp]. IEEE Signal Process Mag 2010;27(5):131–9.
[5] Gerhard D. Audio signal classification: history and current techniques. Department of Computer Science, University of Regina; 2003.
[6] Pieraccini R. The voice in the machine: building computers that understand speech. MIT Press; 2012.
[7] Belouchrani A, Abed-Meraim K, Cardoso JF, Moulines E. A blind source separation technique using second-order statistics. IEEE Trans Signal Process 1997;45(2):434–44.
[8] Campbell JP. Speaker recognition: a tutorial. Proc IEEE 1997;85(9):1437–62.
[9] Loizou PC. Speech enhancement: theory and practice. CRC Press; 2007.
[10] Lansford KL, Liss JM. Vowel acoustics in dysarthria: speech disorder diagnosis and classification. J Speech Lang Hearing Res 2014.
[11] Dibazar AA, Narayanan S, Berger TW. Feature analysis for automatic detection of pathological speech. In: Proceedings of the second joint 24th annual conference and the annual fall meeting of the biomedical engineering society. Engineering in medicine and biology, vol. 1. IEEE; 2002. p. 182–3.
[12] Yilmaz E, Ganzeboom MS, Cucchiarini C, Strik H. Combining non-pathological data of different language varieties to improve DNN-HMM performance on pathological speech; 2016. .
[13] Card SK. The psychology of human-computer interaction. CRC Press; 2018.
[14] Lu L, Liu D, Zhang HJ. Automatic mood detection and tracking of music audio signals. IEEE Trans Audio Speech Lang Process 2006;14(1):5–18.
[15] Tzanetakis G, Cook P. Musical genre classification of audio signals. IEEE Trans Speech Audio Process 2002;10(5):293–302.
[16] Liu Z, Huang J, Wang Y, Chen T. Audio feature extraction and analysis for scene classification. In: Proceedings of first signal processing society workshop on multimedia signal processing. IEEE; 1997. p. 343–8.
[17] Snow WB. Audible frequency ranges of music, speech and noise. Bell Syst Tech J 1931;10(4):616–27.
[18] Smith CP. A phoneme detector. J Acoust Soc Am 1951;23(4):446–51.
[19] Goldman-Eisler F. Speech analysis and mental processes. Lang Speech 1958;1(1):59–75. https://doi.org/10.1177/002383095800100105.
[20] Howard CR. Speech analysis-synthesis scheme using continuous parameters. J Acoust Soc Am 1956;28(6):1091–8. https://doi.org/10.1121/1.1908565.
[21] Stevens KN. Autocorrelation analysis of speech sounds. J Acoust Soc Am 1950;22(6):769–71. https://doi.org/10.1121/1.1906687.
[22] Potter RK, Steinberg JC. Toward the specification of speech. J Acoust Soc Am 1950;22(6):807–20. https://doi.org/10.1121/1.1906694.
[23] Gambardella G. A contribution to the theory of short-time spectral analysis with nonuniform bandwidth filters. IEEE Trans Circuit Theory 1971;18(4):455–60. https://doi.org/10.1109/TCT.1971.1083298.
[24] Rihaczek A. Signal energy distribution in time and frequency. IEEE Trans Inf Theory 1968;14(3):369–74. https://doi.org/10.1109/TIT.1968.105457.
[25] Gambardella G. Time scaling and Short-Time spectral analysis. J Acoust Soc Am 1968;44(6):1745–7. https://doi.org/10.1121/1.1911332.
[26] Bhat AS, Amith VS, Prasad NS, Mohan DM. An efficient classification algorithm for music mood detection in western and hindi music using audio feature extraction. In: 2014 fifth international conference on signal and image processing. p. 359–64.
[27] Saunders J. Real-time discrimination of broadcast speech/music. 1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings, vol. 2. IEEE; 1996. p. 993–6.
[28] Kedem B. Spectral analysis and discrimination by zero-crossings. Proc IEEE 1986;74(11):1477–93.
[29] Li T, Ogihara M, Li Q. A comparative study on content-based music genre classification. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2003. p. 282–9.
[30] Bergstra J, Casagrande N, Erhan D, Eck D, Kégl B. Aggregate features and adaboost for music classification. Mach Learn 2006;65(2–3):473–84.
[31] Yang X, Tan B, Ding J, Zhang J, Gong J. Comparative study on voice activity detection algorithm. In: 2010 International conference on electrical and control engineering. IEEE; 2010. p. 599–602.
[32] Ahrendt P, Meng A, Larsen J. Decision time horizon for music genre classification using short time features. In: 2004 12th European signal processing conference. IEEE; 2004. p. 1293–6.
[33] El-Maleh K, Klein M, Petrucci G, Kabal P. Speech/music discrimination for multimedia applications. 2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100), vol. 4. IEEE; 2000. p. 2445–8.
[34] Mitrovic D, Zeppelzauer M, Breiteneder C. Discrimination and retrieval of animal sounds. In: 2006 12th international multi-media modelling conference. IEEE; 2006. p. 5.
[35] Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project; 2004. .
[36] Burred JJ, Robel A, Sikora T. Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. IEEE Trans Audio Speech Lang Process 2010;18(3):663–74.
[37] Farrús M, Hernando J, Ejarque P. Jitter and shimmer measurements for speaker recognition. In: Eighth annual conference of the international speech communication association.
[38] Jensen K. Pitch independent prototyping of musical sounds. In: 1999 IEEE third workshop on multimedia signal processing (Cat. No. 99TH8451). IEEE; 1999. p. 215–20.
[39] Muhammad G, Alghathbar K. Environment recognition from audio using MPEG-7 features. In: 2009 Fourth international conference on embedded and multimedia computing. IEEE; 2009. p. 1–6.
[40] Valero X, Alías F. Applicability of MPEG-7 low level descriptors to environmental sound source recognition. In: Proceedings 1st Euroregio Conference, Ljubjana.

[41] Rabaoui A, Davy M, Rossignol S, Ellouze N. Using one-class SVMs and wavelets for audio surveillance. IEEE Trans Inf Forensics Secur 2008;3 (4):763–75.
[42] Peltonen V, Tuomi J, Klapuri A, Huopaniemi J, Sorsa T. Computational auditory scene recognition. In: 2002 IEEE international conference on acoustics, speech, and signal processing, vol. 2. IEEE; 2002. p. II–1941.
[43] Jiang H, Bai J, Zhang S, Xu B. SVM-based audio scene classification. In: 2005 international conference on natural language processing and knowledge engineering. IEEE; 2005. p. 131–6.
[44] Ando Y. Autocorrelation-based features for speech representation. J Acoust Soc Am 2013;133(5). https://doi.org/10.1121/1.4805418. pp. 3292–3292.
[45] Sztaho D, Tulics MG, Vicsi K, Valalik I. Automatic estimation of severity of parkinson's disease based on speech rhythm related features. Paper presented at the 000011-000016; 2017. https://doi.org/10.1109/CogInfoCom.2017.8268208. .
[46] Tzanetakis G, Cook P. Musical genre classification of audio signals. IEEE Trans Speech Audio Process 2002;10(5):293–302.
[47] Berisha V, Sandoval S, Utianski R, Liss J, Spanias A. Selecting disorder-specific features for speech pathology fingerprinting. Paper presented at the 7562-7566; 2013. https://doi.org/10.1109/ICASSP.2013.6639133. .
[48] Liss JM, LeGendre S, Lotto AJ. Discriminating dysarthria type from envelope modulation spectra. J Speech Lang Hear Res 2010.
[49] Mendoza E, Valencia N, Muñoz J, Trujillo H. Differences in voice quality between men and women: use of the long-term average spectrum (LTAS). J Voice 1996;10(1):59–66. https://doi.org/10.1016/S0892-1997(96)80019-1.
[50] Tsau E, Kim SH, Kuo CCJ. Environmental sound recognition with CELP-based features. In: ISSCS 2011-international symposium on signals, circuits and systems. IEEE; 2011. p. 1–4.
[51] Sarkar A, Sreenivas TV. Dynamic programming based segmentation approach to LSF matrix reconstruction. In: Ninth European conference on speech communication and technology.
[52] Fu Z, Lu G, Ting KM, Zhang D. A survey of audio-based music classification and annotation. IEEE Trans Multimedia 2011;13(2):303–19.
[53] Agostini G, Longari M, Pollastri E. Musical instrument timbres classification with spectral features. EURASIP J Adv Signal Process 2003;2003(1): 943279.
[54] Wang F, Wang X, Shao B, Li T, Ogihara M. Tag integrated multi-label music style classification with hypergraph. In: ISMIR. p. 363–8.
[55] Sethares WA, Morris RD, Sethares JC. Beat tracking of musical performances using low-level audio features. IEEE Trans Speech Audio Process 2005;13 (2):275–85.
[56] Al-Shoshan AI. Speech and music classification and separation: a review. J King Saud Univ-Eng Sci 2006;19(1):95–132.
[57] Baniya BK, Lee J, Li ZN. Audio feature reduction and analysis for automatic music genre classification. In: 2014 IEEE international conference on systems, man, and cybernetics (SMC). IEEE; 2014. p. 457–62.
[58] Shukla S, Dandapat S, Prasanna SRM. Spectral slope based analysis and classification of stressed speech. Int J Speech Technol 2011;14(3):245.
[59] Murthy HA, Beaufays F, Heck LP, Weintraub M. Robust text-independent speaker identification over telephone channels. IEEE Trans Speech Audio Process 1999;7(5):554–68.
[60] Peeters G, Giordano BL, Susini P, Misdariis N, McAdams S. The timbre toolbox: extracting audio descriptors from musical signals. J Acoust Soc Am 2011;130 (5):2902–16.
[61] Smith D, Cheng E, Burnett IS. Musical onset detection using MPEG-7 audio descriptors. Proceedings of the 20th international congress on acoustics (ICA), Sydney, Australia, vol. 2327. p. 1014.
[62] Misra H, Ikbal S, Bourlard H, Hermansky H. Spectral entropy based feature for robust ASR. In: IEEE international conference on acoustics, speech, and signal processing, vol. 1. IEEE; 2004. p. I–193.
[63] Wold E, Blum T, Keislar D, Wheaten J. Content-based classification, search, and retrieval of audio. IEEE Multimedia 1996;3(3):27–36.
[64] Zhu Y, Kankanhalli MS. Precise pitch profile feature extraction from musical audio for key detection. IEEE Trans Multimedia 2006;8(3):575–84.
[65] Lee JW, Kim S, Kang HG. Detecting pathological speech using contour modeling of harmonic-to-noise ratio. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2014. p. 5969–73.
[66] Murthy YS, Koolagudi SG. Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations. In: 2015 IEEE 28th canadian conference on electrical and computer engineering (CCECE). IEEE; 2015. p. 1271–6.
[67] Farrús M, Hernando J, Ejarque P. Jitter and shimmer measurements for speaker recognition. In: Eighth annual conference of the international speech communication association.
[68] Tzanetakis G, Jones R, McNally K. Stereo panning features for classifying recording production style. In: ISMIR. p. 441–4.
[69] Tzanetakis G, Martins LG, McNally K, Jones R. Stereo panning information for music information retrieval tasks. J Audio Eng Soc 2010;58(5):409–17.
[70] Härmä A. Classification of time-frequency regions in stereo audio. J Audio Eng Soc 2011;59(10):707–20.
[71] Avendano C. Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications. In: 2003 IEEE workshop on applications of signal processing to audio and acoustics (IEEE Cat. No. 03TH8684). IEEE; 2003. p. 55–8.
[72] Fu Z, Lu G, Ting KM, Zhang D. A survey of audio-based music classification and annotation. IEEE Trans Multimedia 2011;13(2):303–19.
[73] Murthy HA, Yegnanarayana B. Group delay functions and its applications in speech technology. Sadhana 2011;36(5):745–82.
[74] Hegde RM, Murthy HA, Rao GR. Application of the modified group delay function to speaker identification and discrimination. 2004 IEEE international conference on acoustics, speech, and signal processing, vol. 1. IEEE; 2004. p. I–517.
[75] Smits R, Yegnanarayana B. Determination of instants of significant excitation in speech using group delay function. IEEE Trans Speech Audio Process 1995;3(5):325–33.
[76] Bogert BP. The quefrency alanysis of time series for echoes; Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. Time Ser Anal 1963:209–43.
[77] Noll AM, Schroeder MR. Short-time "Cepstrum pitch detection. J Acoust Soc Am 1964;36(5). 1030–1030.
[78] Noll AM. Short-time spectrum and "cepstrum techniques for vocal-pitch detection. J Acoust Soc Am 1964;36(2):296–302.
[79] Moir TJ, Barrett JF. A kepstrum approach to filtering, smoothing and prediction with application to speech enhancement. Proc R Soc London Ser A 2003;459(2040):2957–76.
[80] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;28(4):357–66. https://doi.org/10.1109/TASSP.1980.1163420.
[81] Sahidullah M, Saha G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun 2012;54(4):543–65.
[82] Müller M. Information retrieval for music and motion, vol. 2. Heidelberg: Springer; 2007.
[83] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;28(4):357–66.
[84] Krueger A, Haeb-Umbach R. Model-based feature enhancement for reverberant speech recognition. IEEE Trans Audio Speech Lang Process 2010;18(7):1692–707.
[85] Hu N, Dannenberg RB, Tzanetakis G. Polyphonic audio matching and alignment for music retrieval. In: 2003 IEEE workshop on applications of signal processing to audio and acoustics (IEEE Cat. No. 03TH8684). IEEE; 2003. p. 185–8.
[86] Bernard A, Alwan A. Source and channel coding for remote speech recognition over error-prone channels. 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), vol. 4. IEEE; 2001. p. 2613–6.
[87] Kinjo T, Funaki K. On hmm speech recognition based on complex speech analysis. In: IECON 2006–32nd annual conference on IEEE industrial electronics. IEEE; 2006. p. 3477–80.
[88] Chen J, Huang Y, Li Q, Paliwal KK. Recognition of noisy speech using dynamic spectral subband centroids. IEEE Signal Process Lett 2004;11(2):258–61.
[89] Maddage NC, Xu C, Wang Y. A SVM C based classification approach to musical audio; 2003. .
[90] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 1990;87(4):1738–52.
[91] Glodek M, Tschechne S, Layher G, Schels M, Brosch T, Scherer S, Schwenker F. Multiple classifier systems for the classification of audio-visual emotional states. In: Affective computing and intelligent interaction. Berlin, Heidelberg: Springer; 2011. p. 359–68.
[92] Dave N. Feature extraction methods LPC, PLP and MFCC in speech recognition. Int J Adv Res Eng Technol 2013;1(6):1–4.
[93] Protopapas A, Eimas PD. Perceptual differences in infant cries revealed by modifications of acoustic features. J Acoust Soc Am 1997;102(6):3723–34.
[94] Clemins PJ, Johnson MT. Generalized perceptual linear prediction features for animal vocalization analysis. J Acoust Soc Am 2006;120(1):527–34.
[95] Koehler J, Morgan N, Hermansky H, Hirsch HG, Tong G. Integrating RASTA-PLP into speech recognition. Proceedings of ICASSP'94. IEEE international conference on acoustics, speech and signal processing, vol. 1. IEEE; 1994. p. I–421.
[96] Zeng YM, Wu ZY, Falk T, Chan WY. Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. In: 2006 international conference on machine learning and cybernetics. IEEE; 2006. p. 3376–9.
[97] Hardt D, Fellbaum K. Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification. 1997 IEEE international conference on acoustics, speech, and signal processing. IEEE; 1997. p. 867–70.
[98] Greenwood DD. A cochlear frequency-position function for several species-29 years later. J Acoust Soc Am 1990;87(6):2592–605.
[99] Valero X, Alias F. Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification. IEEE Trans Multimedia 2012;14 (6):1684–9.
[100] Valero X, Alias F. Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification. IEEE Trans Multimedia 2012;14 (6):1684–9.
[101] Yin H, Hohmann V, Nadeu C. Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency. Speech Commun 2011;53(5):707–15.
[102] Hegde RM, Murthy HA, Gadde VRR. Significance of the modified group delay feature in speech recognition. IEEE Trans Audio Speech Lang Process 2007;15 (1):190–202. https://doi.org/10.1109/TASL.2006.876858.

[103] Liu Y, Tian Y, He L, Liu J, Johnson MT. Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing. In: Sixteenth annual conference of the international speech communication association.

[104] Sahidullah M, Kinnunen T, Hanilçi C. A comparison of features for synthetic speech detection; 2015. .

[105] Wu Z, Chng ES, Li H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Thirteenth annual conference of the international speech communication association.

[106] Das KA, George KK, Kumar CS, Veni S, Panda A. Modified gammatone frequency cepstral coefficients to improve spoofing detection. Paper presented at the 50-55; 2016.https://doi.org/10.1109/ICACCI.2016.7732024. .

[107] Wu Z, Chng ES, Li H. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Thirteenth annual conference of the international speech communication association.

[108] Sreenivas TV, Kleijn WB. Compressive sensing for sparsely excited speech signals. In: 2009 IEEE international conference on acoustics, speech and signal processing. IEEE; 2009. p. 4125–8.

[109] Grais EM, Erdogan H. Single channel speech-music separation using matching pursuit and spectral masks. In: 2011 IEEE 19th signal processing and communications applications conference (SIU). IEEE; 2011. p. 323–6.

[110] Henaff M, Jarrett K, Kavukcuoglu K, LeCun Y. Unsupervised learning of sparse features for scalable audio classification. In: ISMIR, vol. 11. p. 445.

[111] Gu J, Lu L, Cai R, Zhang HJ, Yang J. Dominant feature vectors based audio similarity measure. In: Pacific-Rim conference on multimedia. Berlin, Heidelberg: Springer; 2004. p. 890–7.

[112] Abidin S, Togneri R, Sohel F. Spectrotemporal analysis using local binary pattern variants for acoustic scene classification. IEEE/ACM Trans Audio Speech Lang Process 2018;26(11):2112–21. https://doi.org/10.1109/TASLP.2018.2854861.

[113] Yang W, Krishnan S. Combining temporal features by local binary pattern for acoustic scene classification. IEEE/ACM Trans Audio Speech Lang Process 2017;25(6):1315–21. https://doi.org/10.1109/TASLP.2017.2690558.

[114] He L, Cao C. Automated depression analysis using convolutional neural networks from speech. J Biomed Inf 2018;83:103–11. https://doi.org/10.1016/j.jbi.2018.05.007.

[115] Demir F, Sengur A, Cummins N, Amiriparian S, Schuller B. Low level texture features for snore sound discrimination. In: Conference proceedings: annual international conference of the IEEE engineering in medicine and biology society. IEEE engineering in medicine and biology society. Annual conference. p. 413.

[116] Rakotomamonjy A, Gasso G. Histogram of gradients of time-frequency representations for audio scene classification. IEEE/ACM Trans Audio Speech Lang Process 2015;23(1):142–53. https://doi.org/10.1109/TASLP.2014.2375575.

[117] Sun B, Li L, Zuo T, Chen Y, Zhou G, Wu X. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In: Proceedings of the 16th international conference on multimodal interaction. ACM; 2014. p. 481–6.

[118] Jiang W, Cotton C, Chang SF, Ellis D, Loui A. Short-term audio-visual atoms for generic video concept classification. In: Proceedings of the 17th ACM international conference on multimedia. ACM; 2009. p. 5–14.

[119] Preis D, Georgopoulos VC. Wigner distribution representation and analysis of audio signals: an illustrated tutorial review. J Audio Eng Soc 1999;47(12):1043–53.

[120] Baydar N, Ball A. A comparative study of acoustic and vibration signals in detection of gear failures using Wigner-Ville distribution. Mech Syst Signal Process 2001;15(6):1091–107.

[121] Boles P, Boashash B. Application of the cross-Wigner-Ville distribution to seismic data processing. Longman Cheshire; 1992.

[122] Tzanetakis G, Essl G, Cook P. Audio analysis using the discrete wavelet transform. Proc. conf. in acoustics and music theory applications, vol. 66.

[123] Lambrou T, Kudumakis P, Speller R, Sandler M, Linney A. Classification of audio signals using statistical features on time and wavelet transform domains. Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, ICASSP'98 (Cat. No. 98CH36181), vol. 6. IEEE; 1998. p. 3621–4.

[124] Baluja S, Covell M. Waveprint: efficient wavelet-based audio fingerprinting. Pattern Recogn 2008;41(11):3467–80.

[125] Li G, Khokhar AA. Content-based indexing and retrieval of audio data using wavelets. 2000 IEEE international conference on multimedia and expo. ICME2000. Proceedings. Latest advances in the fast changing world of multimedia (Cat. No. 00TH8532), vol. 2. IEEE; 2000. p. 885–8.

[126] Liu Y, Xiang Q, Wang Y, Cai L. Cultural style based music classification of audio signals. In: 2009 IEEE international conference on acoustics, speech and signal processing. IEEE; 2009. p. 57–60.

[127] Li Y, Zhang X, Jin H, Li X, Wang Q, He Q, Huang Q. Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection. Multimedia Tools Appl 2018;77(1):897–916. https://doi.org/10.1007/s11042-016-4332-z.

[128] Li Y, Li X, Zhang Y, Wang W, Liu M, Feng X. Acoustic scene classification using deep audio feature and BLSTM network. Paper presented at the 371-374; 2018.https://doi.org/10.1109/ICALIP.2018.8455765. .

[129] Takahashi N, Gygli M, Van Gool L. AENet: learning deep audio features for video analysis; 2017. .

[130] Rahmani MH, Almasganj F, Seyyedsalehi SA. Audio-visual feature fusion via deep neural networks for automatic speech recognition. Digital Signal Process 2018;82:54–63. https://doi.org/10.1016/j.dsp.2018.06.004.

[131] Badshah AM, Rahim N, Ullah N, Ahmad J, Muhammad K, Lee MY, Kwon S, Baik SW. Deep features-based speech emotion recognition for smart affective services. Multimedia Tools Appl 2019;78(5):5571–89. https://doi.org/10.1007/s11042-017-5292-7.

[132] Qian Y, Chen N, Yu K. Deep features for automatic spoofing detection. Speech Commun 2016;85:43–52. https://doi.org/10.1016/j.specom.2016.10.007.

[133] Qian K, Schmitt M, Janott C, Zhang Z, Heiser C, Hohenhorst W, Herzog M, Hemmert W, Schuller B. A bag of wavelet features for snore sound classification. Ann Biomed Eng 2019;47(4):1000–11. https://doi.org/10.1007/s10439-019-02217-0.

[134] Noor S, Dhrubo EA, Minhaz AT, Shahnaz C, Fattah SA. Audio visual emotion recognition using cross correlation and wavelet packet domain features. Paper presented at the 233-236; 2017.https://doi.org/10.1109/WIECON-ECE.2017.8468871. .

[135] Ghoraani B, Krishnan S. Time-frequency matrix feature extraction and classification of environmental audio signals. IEEE Trans Audio Speech Lang Process 2011;19(7):2197–209. https://doi.org/10.1109/TASL.2011.2118753.

[136] Umapathy K, Krishnan S, Rao RK. Audio signal feature extraction and classification using local discriminant bases. IEEE Trans Audio Speech Lang Process 2007;15(4):1236–46. https://doi.org/10.1109/TASL.2006.885921.

[137] Umapathy K, Krishnan S, Jimaa S. Multigroup classification of audio signals using time-frequency parameters. IEEE Trans Multimedia 2005;7(2):308–15. https://doi.org/10.1109/TMM.2005.843363.

[138] Cohen L, Posch T. Positive time-frequency distribution functions. IEEE Trans Acoust Speech Signal Process 1985;33(1):31–8.

[139] Umapathy K, Ghoraani B, Krishnan S. Audio signal processing using time-frequency approaches: coding, classification, fingerprinting, and watermarking. EURASIP J Adv Signal Process 2010;2010(1):1–28. https://doi.org/10.1155/2010/451695.

[140] Tuncer T, Dogan S. Novel dynamic center based binary and ternary pattern network using M4 pooling for real world voice recognition. Appl Acoust 2019;156:176–85. https://doi.org/10.1016/j.apacoust.2019.06.029.

[141] Adnan SM, Irtaza A, Aziz S, Ullah MO, Javed A, Mahmood MT. Fall detection through acoustic local ternary patterns. Appl Acoust 2018;140:296–300. https://doi.org/10.1016/j.apacoust.2018.06.013.

[142] Hossain MS. Patient state recognition system for healthcare using speech and facial expressions. J Med Syst 2016;40(12):1–8. https://doi.org/10.1007/s10916-016-0627-x.

[143] Tuncer T, Dogan S, Ertam F. Automatic voice based disease detection method using one dimensional local binary pattern feature extraction network. Appl Acoust 2019;155:500–6. https://doi.org/10.1016/j.apacoust.2019.05.023.

[144] Tuncer T, Dogan S. A novel octopus based Parkinson's disease and gender recognition method using vowels. Appl Acoust 2019;155:75–83. https://doi.org/10.1016/j.apacoust.2019.05.019.

[145] Korkmaz Y, Boyací A, Tuncer T. Turkish vowel classification based on acoustical and decompositional features optimized by genetic algorithm. Appl Acoust 2019;154:28–35. https://doi.org/10.1016/j.apacoust.2019.04.027.

[146] Ertam F. An effective gender recognition approach using voice data via deeper LSTM networks. Appl Acoust 2019;156:351–8. https://doi.org/10.1016/j.apacoust.2019.07.033.

[147] Glowacz A. Fault detection of electric impact drills and coffee grinders using acoustic signals. Sensors (Basel, Switzerland) 2019;19(2):269. https://doi.org/10.3390/s19020269.

[148] Lu S, Wang X, Liu F, He Q, Liu Y, Zhao J. Fault diagnosis of motor bearing by analyzing a video clip. Math Probl Eng 2016;2016:1–11. https://doi.org/10.1155/2016/8139273.

[149] Glowacz A. Fault diagnosis of single-phase induction motor based on acoustic signals. Mech Syst Signal Process 2019;117:65–80. https://doi.org/10.1016/j.ymssp.2018.07.044.

[150] Duan Z, Wu T, Guo S, Shao T, Malekian R, Li Z. Development and trend of condition monitoring and fault diagnosis of multi-sensors information fusion for rolling bearings: a review. Int J Adv Manuf Technol 2018;96(1):803–19. https://doi.org/10.1007/s00170-017-1474-8.

[151] Glowacz A. Acoustic-based fault diagnosis of commutator motor. Electronics 2018;7(11):299. https://doi.org/10.3390/electronics7110299.