Multimedia Tools and Applications (2023) 82:36143–36177
https://doi.org/10.1007/s11042-023-14734-1

Time-frequency visual representation and texture features for audio applications: a comprehensive review, recent trends, and challenges

Yogita D. Mistry[1] · Gajanan K. Birajdar[1] · Archana M. Khodke[1]

Received: 29 December 2021 / Revised: 13 July 2022 / Accepted: 4 February 2023 /
Published online: 16 March 2023

Abstract

The conventional audio feature extraction methods employed in the audio analysis are categorized into time-domain and frequency-domain. Recently, a new audio feature extraction approach using time-frequency texture image is developed and utilized for different applications. In this approach, the input audio signal is first converted into a time-frequency image, and then textural features are extracted from the visual representation. The distinctive two-dimensional time-frequency visualization textural descriptors can produce better features for improved audio detection and classification. In this article, a comprehensive review of state-of-the-art techniques used for audio detection and classification is presented. The generalized architecture of time-frequency texture feature extraction approaches in audio classification algorithms is presented first. Based on a review of over 70 papers, the key contributions in the area of time-frequency representations of various researchers are highlighted in addition to the textural features. This survey also compares and analyzes the existing experimental algorithms proposed for various audio classification tasks. Finally, the critical challenges and limitations with different visual representations are highlighted, along with potential future research directions.

Keywords Time-frequency representation · Time-frequency texture image · Spectrogram · Cochleagram · Textural features · Audio classification · Acoustic scene classifcation · Music genre classification

✉ Gajanan K. Birajdar
gajanan.birajdar@rait.ac.in

Yogita D. Mistry
yogita.mistry@rait.ac.in

Archana M. Khodke
archana.khodke@rait.ac.in

[1] Department of Electronics Engineering, Ramrao Adik Institute of Technology, DY Patil deemed to be University, Nerul, 400706, Navi Mumbai, Maharashtra, India

1 Introduction

Various forms of the interactions between machine and human have grown in the last decade, including smart home appliances, smart stores, automotive industry, security and mobile phones. The human interaction happens primarily using audio-visual form and gestures. Because of the recent technological advancements in the audio and music field, an enormous amount of data is available locally or over the network. The data content search and information retrieval is a demanding task in a variety of applications such as music mood classification, genre classification, melody identification, acoustic scene classification and cover song identification [66].

The audio analysis tasks such as audio surveillance, music genre recognition, sound event classification, acoustic scene classification require robust and most discriminating features. The conventional audio feature extraction methods are categorized as: (a) time-domain features and (b) frequency-domain features [66]. Zero-crossing rate (ZCR), signal energy, maximum amplitude, and auto-correlation based features are the few examples of time-domain features. The frequency-domain features include fundamental frequency, spectral centroid, spectral flux, spectral density, spectral roll-off, chroma features, Mel-frequency cepstral coefficients (MFFC) and linear predictive coding (LPC). More often, these features are combined to enhance the algorithm performance in various applications.

Recently, a new audio feature extraction approach using time-frequency texture image is developed and employed for different applications. In this technique, the input audio signal is first converted into a time-frequency image (such as spectrogram or MFCC or Cochleagram image) and then textural features are extracted from this visual representation. The distinctive two-dimensional time-frequency visualization can produce better features for audio detection and classification tasks. Such texture features are expected to be complementary to the conventional features to construct a robust audio classification system. Because of the non-uniformity of the textures in the visual image, usually, local feature feature extraction is considered during the feature construction phase.

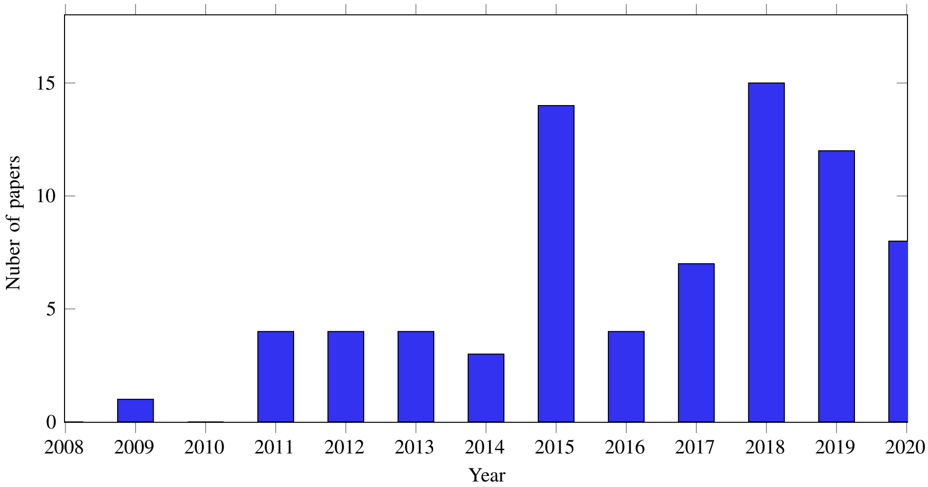


Fig. 1 A graph depicting number of papers published during 2009-2020

Researchers in 1970s initiated early efforts of understanding and analyzing visual information in the form of spectrogram image. Visual information of spectrogram image was employed in 1970s and 1980s for identifying the phonetic contents [86], continuous speech recognition [38], multi-speaker continuous speech recognition [22] and stop-consonants identification from continuous speech [87]. The manual analysis of visual information by different researchers was limited to spectrogram image and difficult because of the complex speech structure. More distinctive textural variations are observed in the time-frequency representation for the short duration audio sample. Different textural descriptors effectively capture these variations for a variety of audio classification task in the recent past.

A comprehensive survey of time-frequency image texture feature extraction algorithms in audio applications is presented in this article. To the best of our knowledge, this is the first attempt to survey different image texture feature extraction techniques for speech, music, audio and environment sound classification. A total of 77 papers from top-tier journals and conferences in the last twelve years are collected. All the articles collected are between the years 2009 to 2020, and year-wise number of papers appeared are illustrated in Fig. 1. This literature has been categorized based on (a) audio signal, (b) time-frequency representations and (c) applications.

Based on the audio signal type, different articles are divided in to: (a) speech (b) music (c) environment/acoustic sound and (d) other applications. Figure 2 depicts the percentage of papers from each of the audio type. The other category includes audio signals such as bird sounds, baby cry and bird vocalization identification. From Fig. 2 it is clear that most algorithms are focused on music and environment/acoustic sound signal analysis and classification. The second type of categorization is based on the use of time-frequency visual representation.

Different types of time-frequency images are used in the literature for texture feature extraction like spectrogram, cochleagram, Constant-Q transforms (CQT) and MFCC. Figure 3 illustrates the classification of various algorithms according to the time-frequency image. Spectrogram images are primarily employed for feature extraction by researchers in different application development. The spectrogram image provides more distinctive patterns for the classification or identification task compared to other image representations.

The last classification is based on applications of the time-frequency texture image. We have classified the proposed approaches in five broad application areas: (1) music genre

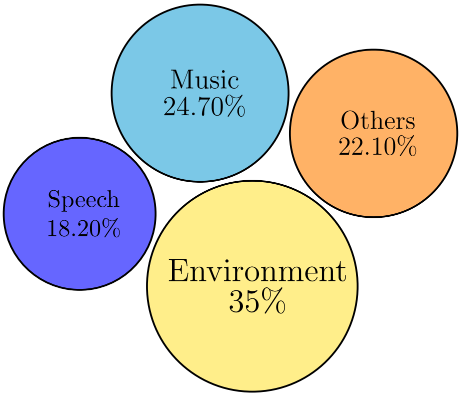


Fig. 2 Classification of various articles according to the audio signal type

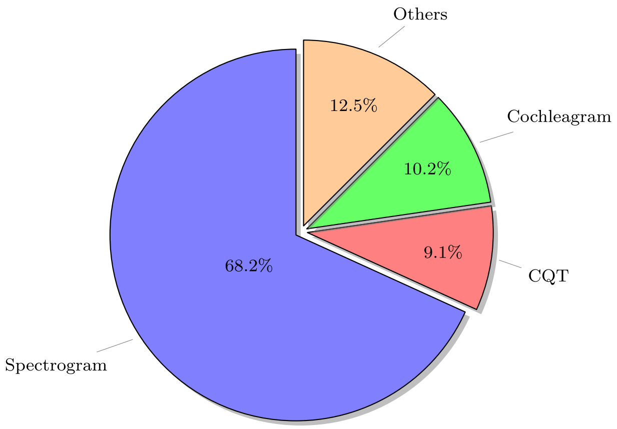


Fig. 3 Classification of various algorithms according to the time-frequency image

identification (2) acoustic scene classification (3) bird and animal species classification (4) sound event classification and (5) other applications. Figure 4 shows the classification and percentage of articles in each of the application. The time-frequency texture image features are first introduced for music genre classification and then extended for other application areas. Table 1 depicts list of abbreviations used in the article.

Generalized architecture of time-frequency texture feature extraction approaches in audio classification algorithms: The different audio classification approaches in the literature based on time-frequency texture image encompasses time-frequency image generation,

Fig. 4 Broad areas of application development classification of various algorithms

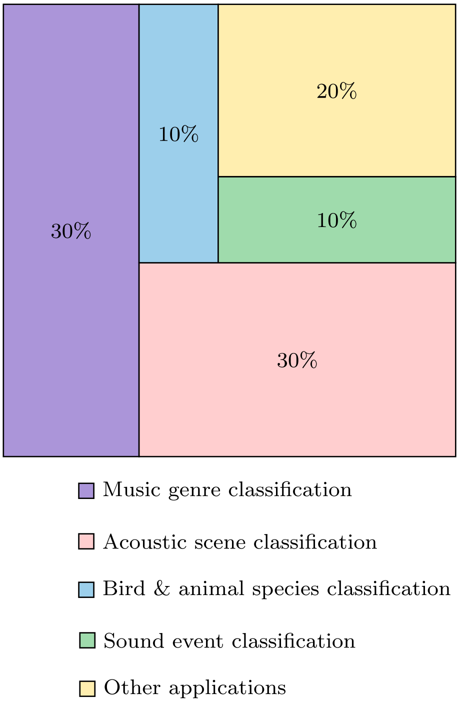


Table 1 List of abbreviations

Abbreviation	Definition
GLCM	Gray-level co-occurrence matrix
LBP	Local binary pattern
HOG	Histogram of Oriented Gradients
LPQ	Local phase quantization
LCP	Local configuration pattern
WLD	Weber local descriptors
RICLBP	Rotation invariant co-occurrence LBP
CoLBP	Co-occurrence of adjacent LBP
NTLBP	Noise tolerant LBP
RILBP	Rotation invariant LBP
LTP	Local ternary pattern
HASC	Heterogeneous auto-similarities of characteristics
LBPHF	LBP Fourier features
MLBP	Multiscale LBP
RLBP	Rotated LBP
LPC	Linear predictive coding
MFCC	Mel-frequency cepstral coefficients
CQT	Constat-Q transform
SIFT	Scale-invariant feature transform
SVM	Support vector machine

texture feature extraction and classification model. We have developed the generalized architecture of time-frequency texture feature extraction algorithms in audio classification tasks as illustrated in Fig. 5.

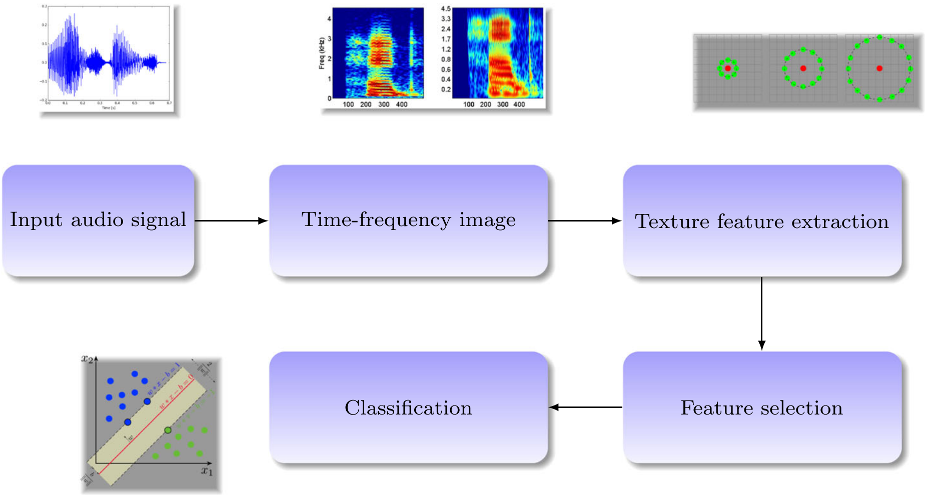


Fig. 5 Generalized architecture of time-frequency texture feature extraction approaches in audio classification algorithms

The first step is to generate visual representation such as spectrogram, cochleagram, CQT image from the input audio signal. In the second step, textural descriptors are extracted from the audio image. LBP, LTP, LPQ, and RLBP descriptors are commonly used for the feature extraction task. Due to the non-uniformity of textures in time-frequency image, local feature extraction is considered by employing zoning during the feature construction stage.

The textural descriptors like LPQ, LPQ and LTP extracted from the time-frequency texture image produces large dimensional feature vector. To speed-up the computation and to reduce classification complexity, feature selection is often employed before the classifier stage. Feature selection stage removes redundant and less critical features leaving only relevant descriptors, hence creating small dimensional final feature set. Finally, classification is implemented using a support vector machine, k-nearest neighbor or neural network classifier. It is observed from the literature that, SVM is most popularly used classification approach because of its excellent performance even in noisy conditions.

A systematic methodology to compose a comprehensive record of the state-of-art algorithms focusing on time-frequency image texture features employed in audio applications is illustrated. The major contributions of the survey article can be summarized as:

- A comprehensive survey of the state-of-art algorithms focusing on time-frequency image texture features is demonstrated.
- Generalized architecture of time-frequency texture feature extraction approaches in audio classification algorithms is presented.
- Presents a critical review of different time-frequency representations with their features, advantages, and limitations employed in audio classification tasks.
- Furnishes a brief review of various textural descriptors with their advantages and disadvantages utilized for feature extraction.
- Presents limitations and challenges of existing techniques.

The article is organized as follows. Firstly, the published state-of-the art algorithms are categorized into three different classes. Section 2 describe time-frequency image representations in detail along with the applications. Different textural features and classification algorithms are discussed in Sections 3 and 4. Section 5 outlines challenges, advantages and limitations for implementing various audio applications using time-frequency texture image. Finally, Section 6 concludes the article.

2 Time-frequency visual representation

In time-frequency texture image based audio feature extraction technique, the input audio signal is first converted into a time-frequency image such as spectrogram or MFCC or cochleagram image, and then textural features are extracted from this visual representation. This section demonstrates different time-frequency visualizations employed for feature extraction and key aspects are compared.

2.1 Spectrogram

A spectrogram is a two-dimensional visual presentation of signal strength at a different frequency that varies with time. Since spectrogram provides profound attributes, it is popularly employed in a variety of speech and music processing applications. In a spectrogram, vertical axis illustrates frequency and the horizontal axis represents time and energy content present are depicted in the form of grayscale level. Recently, spectrogram image texture is characterized in various applications in order to capture the relevant details. The various types of spectrograms can be categorized as log-mel spectrogram [53], IIR-CQT spectrogram [14], and linear spectrogram [8, 16, 18, 19, 30]. Moreover, the spectrogram is also classified as narrowband and wideband based on the analysis window utilized.

To generate a spectrogram time-frequency image, firstly the input speech sample $x(i)$ is segmented into windows having length N frames. Later, these frames are transformed into frequency-domain by applying windowed Fourier transform as,

$$X_t(k) = \sum_{i=0}^{N-1} x(i)\omega(i)e^{-\frac{2\pi i}{N}ki} fork = 0, \ldots, (N-1) \tag{1}$$

where, $\omega(i)$ = Hamming window, k is frequency, $f(k) = kF_s/N$ and F_s is the sampling frequency. Finally, linear or log power is used to create the spectrogram as,

$$S_{Lin}(k,t) = |X_t(k)|$$

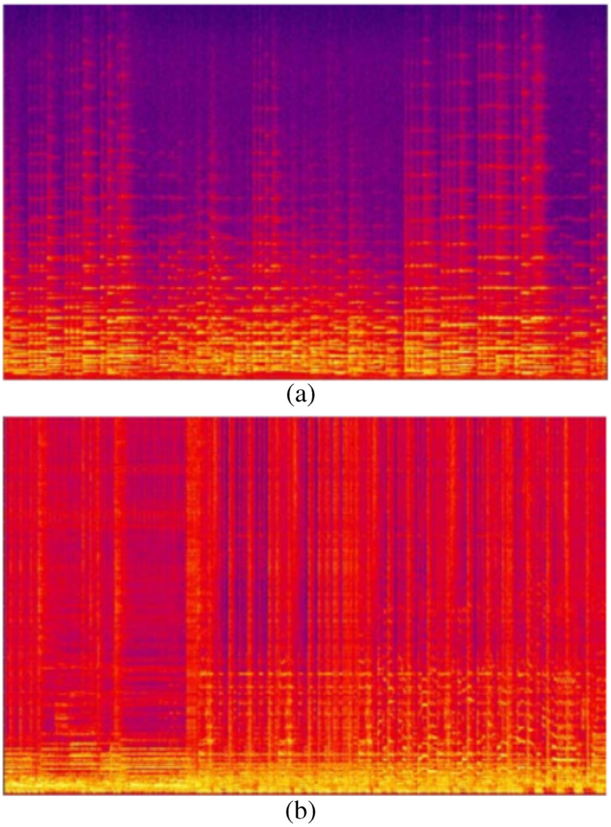
$$S_{\mathrm{Log}}(k,t) = \log S_{Lin}(k,t) \tag{2}$$

Some methods employ normalized grayscale spectrogram intensity image before extracting textural features [30].

An early attempt of time-frequency texture image feature extraction is presented in [81]. The proposed approach classifies different musical instruments using minimum-block matching of energy coefficients as features extracted from spectrogram visual image with 85% accuracy rate. Music genre classification approach is illustrated based on a scale-invariant feature transform (SIFT) keypoint features extracted from spectrogram time-frequency image [31, 43]. Local dynamic details are effectively represented using SIFT keypoint descriptors and classified using support vector machine (SVM) classifier attaining 82.7% classification rate. A method based on central moment features extracted from spectrogram image and one-against-one (OAO) multi-class SVM classifier for mismatched conditions sound event classification is illustrated in [30].

Music genre classification performance enhancement is observed by fusing acoustic and visual features (central moments) extracted from spectrogram image in [74]. The combined feature set is classified using SVM resulting in 86.1% average accuracy rate. Grey-level co-occurrence matrix (GLCM) features and classifier voting mechanism is introduced for music genre classification with 67.2% average accuracy [26]. 28-D GLCM and 59-D local binary pattern (LBP) textural descriptors are extracted from each spectrogram image zone for music genre classification in [25, 27]. Additionally, the effect of individual classifier assignment to each Mel scale zone and combination of different classifiers are investigated in [23]. SVM classification fusion rules such as min, max, sum rule are employed for different zones to increase the classification accuracy.

The performance of sound event classification scheme under noisy mismatched environment is enhanced in [29]. The method uses sub-band power distribution and spectrogram image features and classified using SVM resulting over 96% accuracy. Local phase quantization (LPQ) and Gabor filter features are extracted from a spectrogram image for music genre classification and classified using SVM with 80.78% [24]. It was observed that LPQ outperforms Gabor features when obtained globally from the spectrogram. Combination of nonlinear classifiers in addition to the Gabor and LBP features are explored for music genre classification with 84.9% accuracy rate in [72, 73]. Each music genre has unique spectrogram signature as illustrated in Fig. 6. A method to classify the music signal input into instrument and the song is designed using spectrogram visual intensity co-occurrence descriptors and random sample consensus (RANSAC) classification model [34].



(a)

(b)

Fig. 6 Sample music clips shown using spectrogram: (a) classical and (b) disco [72]

Audio surveillance in a noisy environment is analyzed using MFCC and central moment features extracted from the spectrogram image and multi-class SVM classifier [60]. It was observed that the linear grayscale descriptors are robust compared to log-grayscale features in a noisy environment. Music genre classification using ten different descriptors and three different spectrogram types (linear, global and mel scale zoning) is evaluated in [48]. An average accuracy of 86.1% is achieved using 45 SVMs trained for every texture features and combined using sum rule for the final decision.

Spectrogram image local statistics and SVM are utilized for environmental sound classification in [39] with an impressive accuracy rate of 98.62%. Besides, L2-Hellinger based feature normalization approach has proved enhanced robustness and added discriminating power. The codebook is created using the k-means clustering algorithm of LBP feature map from the spectrogram image and classified using SVM for acoustic context identification [9]. The bag-of-features (BoF) technique utilized reduces the computational complexity of the algorithm. Two different LBP variants RIC-LBP and μLBP in addition to LBP and SVM classifier are employed for music genre classification with 84% accuracy [5].

The amplitude histogram from each frequency band is extracted as subband power distribution (SPD) features and histogram of gradient (HoG) feature for acoustic scene classification in [17]. Moreover, earth mover distance is employed to compare histograms, and it was found that Sinkhorn kernel improved the classification performance. Prosodic cues in the language are effectively modeled using LPQ descriptors extracted from the spectrogram image of language utterance and employed for language identification [45].

The mean and standard deviation of central moments are computed from the linear grayscale spectrogram image and classified using one-against-all (OAA-SVM) classifier with an improved classification rate of 98.16% [62, 64]. However, higher training time is required for the OAA approach compared to other multi-class SVM techniques. In another approach, the authors used GLCM features and SVM classifier fusion to obtain the accuracy of 90.20% [63]. The sub-band frequency analysis has produced higher accuracy rate, however, generates large dimensional feature vector. A method based on entropy, third-order moments and directionality features and SVM classifier are developed for identification of ground moving targets [67]. Multilevel feature extraction from spectrogram time-frequency visual representation is introduced for music genre classification in [75]. A late classifier fusion of acoustic and visual descriptors is suggested with 88.60% classification accuracy.

A sound event identification in noisy conditions using LBP and HOG descriptors from spectrogram images is presented in [42]. Moreover, the global characteristics are exploited using bag-of-audio words and classified using SVM attaining 69.28% average accuracy. Music genre classification approach using spectrogram based gradient directional pattern is formulated using SVM classifier with 84.5% accuracy in [6]. Bird species identification algorithm using spectrogram based different texture features such as local ternary patterns (LTP) quantization, auto-similarities and LBP variants are implemented [52]. Combining textural features with acoustic has improved the classification rate up to 94.5%/.

Music genre classification algorithm combining different descriptors extracted from Mel-scaled spectrogram image and fusion of heterogeneous classifiers is presented in [50]. The proposed technique is evaluated over LMD, ISMIR 2004 and the GTZAN database with 84.9% highest classification accuracy. Robotic hearing sound event classification in noisy conditions using multi-channel band independent LBP textural descriptors is evaluated using RWCP and NTU-SEC database in [57, 70]. The study revealed that Gammatone spectrogram in the logarithm domain is more appropriate for textural analysis of sound. A combination of LBP, RLBP and LPQ textural features are constructed from spectrogram image representation for acoustic scene classification resulting 80.17% accuracy rate evaluated on DCASE2016 database [33]. Additionally, combining the left and right audio channel for feature extraction increases the classification performance.

Spectrogram texture descriptors using GLCM and SVM classification scheme are utilized for discriminating laryngeal mechanism with an average accuracy rate of 86.16% [40]. A set of texture features are extracted from the spectrogram, rhythm image and gammatonegram images after dividing it into sub-windows and trained SVM classifier in [49]. The proposed method is evaluated on different databases like GTZAN, ISMIR 2004 and LMD. An automatic method for bird and whale species identification using three different spectrograms and multiple texture descriptors is presented [47]. In addition to visual features, acoustic features are combined to enhance the identification rate measured using OAA-SVM. The class imbalance issue in music genre classification is addressed by applying oversampling and undersampling in [69]. LBP features are extracted after vertical splicing of spectrogram image and classified using several classifiers.

Bird species identification is investigated using three different textural descriptors and SVM classification attaining 71% accuracy rate in [85]. The dissimilarity approach employed in the algorithm performs better even in case of a large number of input classes. Chinese regional folk-songs recognition approach using auditory features and visual textural descriptors is formulated in [78]. Ensemble SVM classification evaluated on three

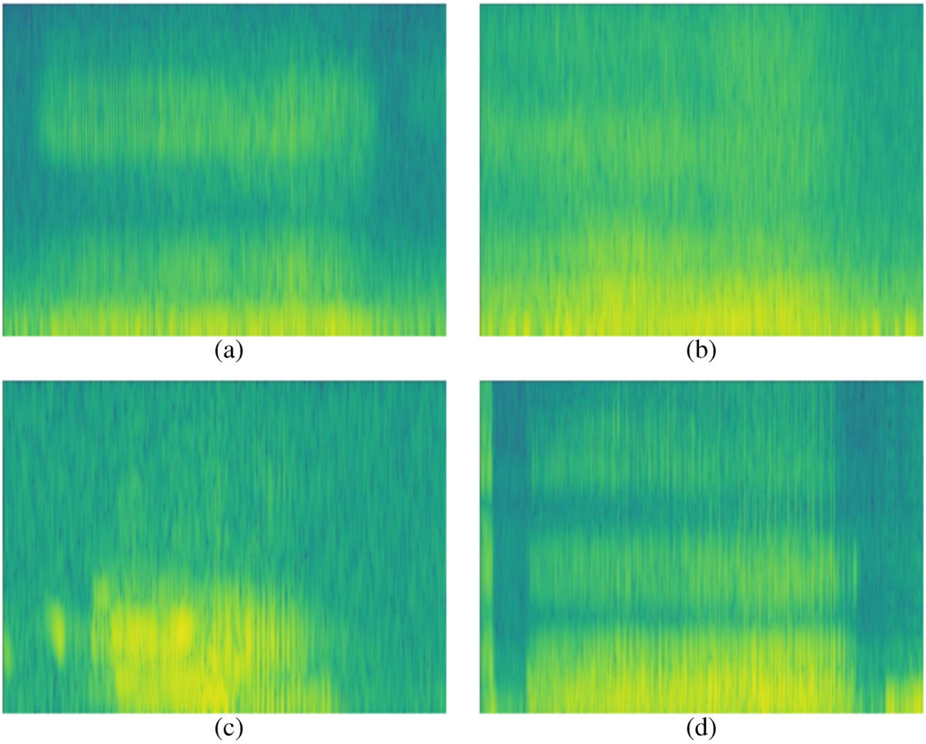


Fig. 7 Snore sound spectrogram images related to vibration point (a) Velum, (b) Oropharyngeal lateral walls (c) Tongue and (d) Epiglottis [28]

different Chinese folk-song databases achieved 89.29%. Emotion recognition from speech signal algorithm using spectrogram visual images LBP texture features and SVM classification is constructed in [54]. Highest identification rate of 84.5% is achieved using EMO-DP database. Initially, acoustic events are represented using a Gaussian mixture model (GMM) energy detection approach and acoustic and visual features are extracted for bird species identification [82]. Relief feature selection algorithm and SVM classifier applied on real-world bird species database resulted in 96.7% classification accuracy.

A method to discriminate snore sounds is designed based on HOG and LBP features from spectrogram visualization and SVM classification resulting in 72.6% accuracy in [28]. Figure 7 depicts various snore sound spectrogram images related to different vibration point such as velum, oropharyngeal lateral walls, tongue and epiglottis. Speech music classification algorithm is developed using major spectral-peak locations and identification of these sequences and three different classifiers SVM, GMM, and random forest classifier with 98% accuracy rate [10]. The periodicity, the average frequency and statistics of these peak sequences are finally used as features. The speech and music differences are clearly identified using the spectrogram shown in Fig. 8. Animal sound recognition approach using double spectrogram features: (a) projection features and (b) LBP variance features with random forest classifier is formulated in [41]. The combined feature set greatly enhances the classification performance attaining 98.02% accuracy.

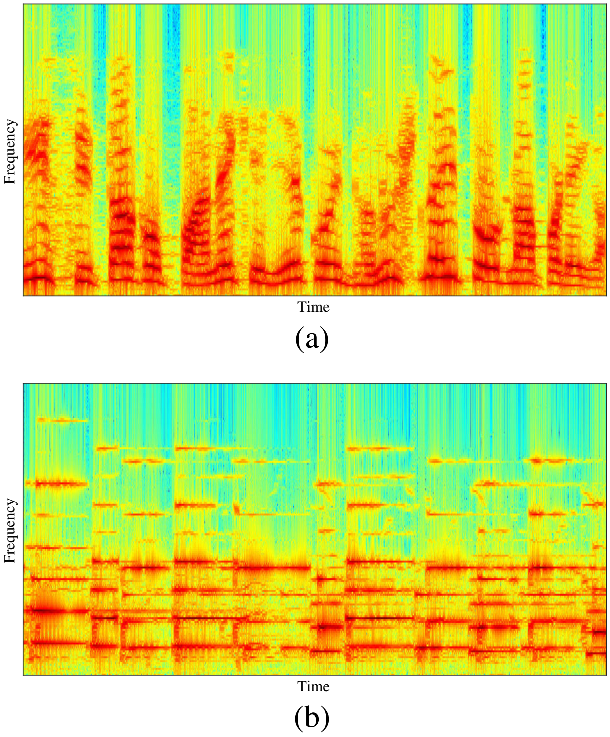


Fig. 8 Spectrograms representation of (a) Speech sample and (b) Music sample [10]

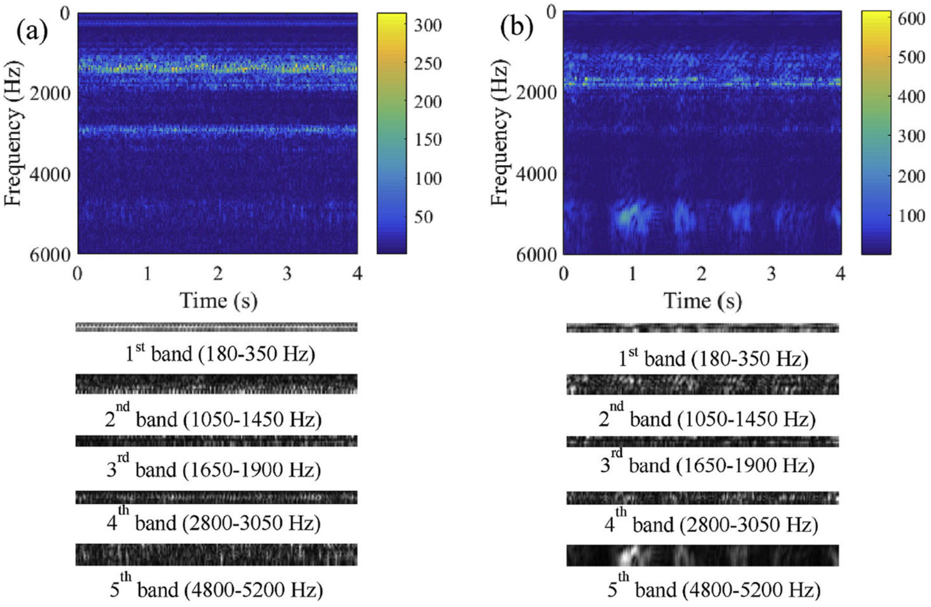
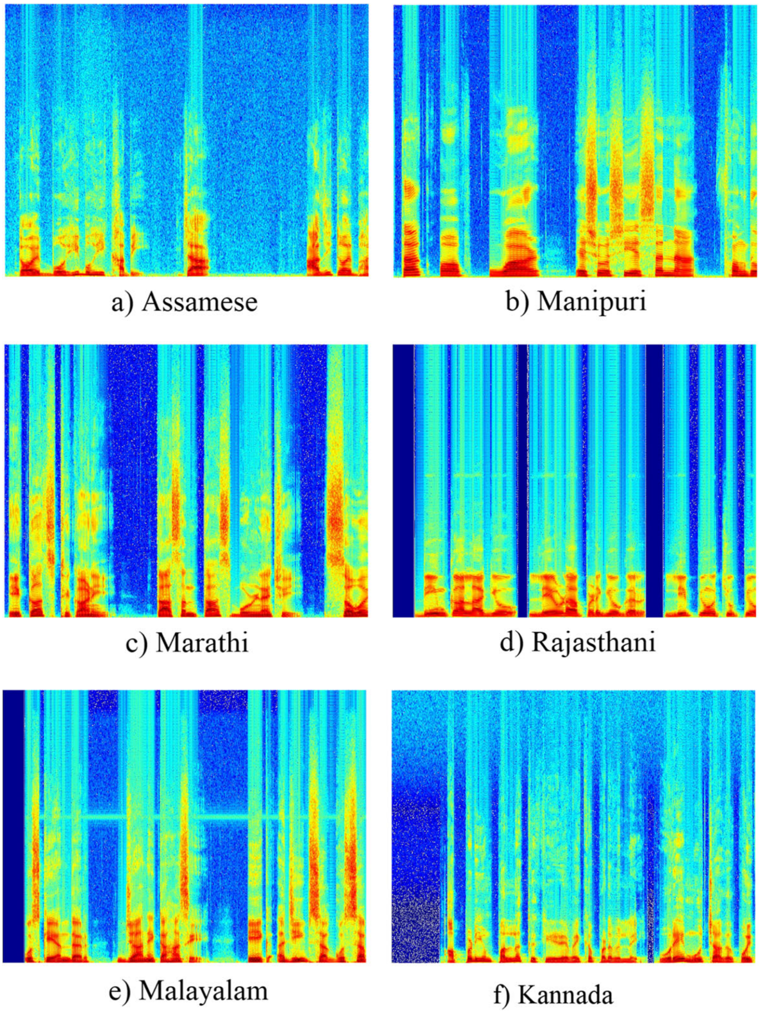


Fig. 9 Spectrogram images of vibration signals of (a) stable and (b) unstable [20]

Speech emotion recognition technique using bag-of-visual words extracted from spectrogram image is developed [68]. The visual vocabulary is constructed and classified using SVM classification evaluated on four different datasets. Figure 11 illustrates spectrogram images of various emotions created using EMO-DB database with and without noise. Chatter detection method using spectrogram time-frequency image and GLCM features is proposed in [20]. The machine condition using vibration signal analysis is performed by identifying high-energy dominant frequency bands from the spectrogram image. The vibration signal spectrograms are different at stable and unstable conditions that is shown in Fig. 9. A method to identify the motivation of infants' cry, i.e. because of feeling or pain is presented in [32] using textural features extracted from spectrogram image. Experiments are carried at different noise conditions and classifier fusion strategies.



a) Assamese

b) Manipuri

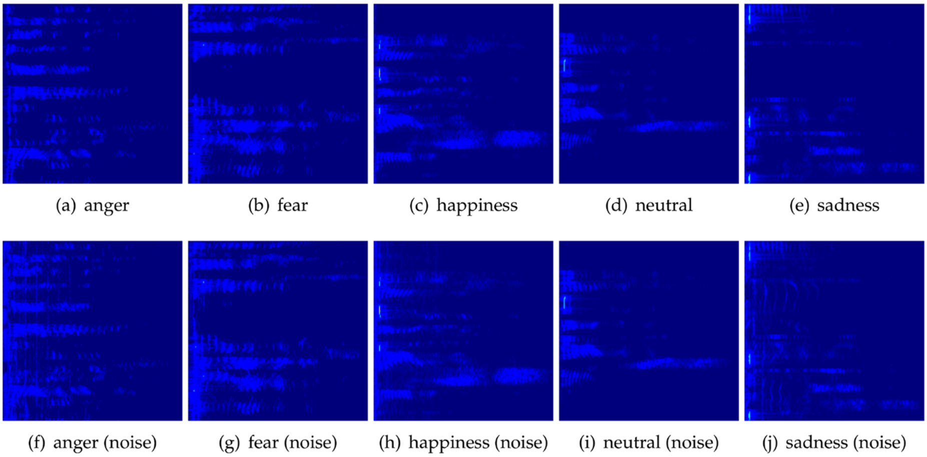
c) Marathi

d) Rajasthani

e) Malayalam

f) Kannada

Fig. 10 Spectrogram visualization of various Indian language speech samples. [21]


(a) anger (b) fear (c) happiness (d) neutral (e) sadness

(f) anger (noise) (g) fear (noise) (h) happiness (noise) (i) neutral (noise) (j) sadness (noise)

Fig. 11 Spectrogram images of various emotions created using EMO-DB database. First row depicts original audio samples whereas second row shows samples after noise addition [68]

Speech spoofing detection using spectrogram image LBP texture features and SVM classification is introduced in [55] with 71.67% average accuracy rate. The generalized Gaussian distribution (GGD) parameters are extracted from a non-subsampled Contourlet transform (NSCT) sub-bands for speech and music discrimination in [14]. The spectrogram image is decomposed using NSCT, and estimated parameters are employed for classification using extreme learning machine (ELM) classifier. The higher-order statistics are encoded using Fisher vectors from the spectrogram monochrome image and classified using SVM classifier resulting highest accuracy of 92.27% [46]. Spectrogram texture descriptor based Indian language identification technique is developed in [21, 35]. Each individual languages has different spectrogram visualizations as depicted in Fig. 10. CLBP, LBPHF and DWT texture features are extracted, and artificial neural network (ANN) classifier is used attaining 96.96% average identification rate (Fig. 11).

Acoustic scene classification method using GLCM features extracted from log-mel spectrogram and SVM classifier is described in [53]. The dimensionality of the feature vector is reduced using principal component analysis, and the method achieved 83.2% classification rate evaluated on DCASE 2016 database. In [83] speech resampling manipulation algorithm based on spectrogram LBP features is presented. The forensic investigation of resampling operation is detected using SVM classifier. Tables 2 and 3 depicts summary of different techniques proposed in the literature based on spectrogram time-frequency visualization. Recently robust acoustic event recognition using gray scale spectrogram is presented [84].

2.2 Cocheleagram

The cochleogram also known as gammatonegram imitates outer and middle human ear components. It relies upon the gammatone warping function which fits empirical observations of frequency selectivity in the mammalian cochlea, with an impulse response $g(t)$ given by

$$g(t) = at^{P-1}\cos(2\pi fct + \phi)e^{2\pi bt} \tag{3}$$

where t is time, a is amplitude, P represents the filter order, ϕ is the phase shift, fc is the central frequency (in kHz).

Table 2 Summary of different algorithms based on spectrogram time-frequency representation

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy(%)	Description
Yu and Slotine [81]	2009	Musical instrument classification	Block energy	✗	420	Private	Nearest neighbor	85	Algorithm replies purely on time-frequency texture features
Matsui et al. [43]	2011	Genre classification	SIFT key points	✗	128	ISMIR 2004	SVM	82.7	Feature are robust against pitch and tempo variations
Dennis et al. [30]	2011	Sound event classification	Central moments	✗	486	RWCP	SVM	90.6	Algorithm is highly robust to noise in mismatched conditions
Wu et al. [74]	2011	Music genre classification	Gabor energy	✗	–	GTZAN, ISMIR 2004	SVM	86.1	Classification accuracy is improved by combining visual and acoustic features
Costa et al. [26]	2011	Music genre classification	GLCM	✗	28	Latin Music Database	SVM	67.2	Implementation of classifier voting mechanism for accuracy enhancement
Costa et al. [27]	2012	Music genre classification	GLCM, LBP	✗	87	Latin Music Database	SVM	80.3	Spectrogram zoning mechanism is used for performance enhancement
Wu et al. [73]	2013	Music genre classification	Gabor features, LBP	✗	–	GTZAN	Fisher, LDC & KNN	84.9	Combination of nonlinear classifiers enhances the performance
Costa et al. [25]	2012	Music genre classification	LBP	✗	59	LMD & ISMIR 2004	SVM	82.33	Spectrogram is divided into zones and classifier fusion is used for final decision
Ghosal et al. [34]	2012	Music/instrument classification	Intensity co-occurrence matrix	✗	10	Private	RANSAC	95	RANSAC performance is better compared to ANN

Table 2 (continued)

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy(%)	Description
Costa et al. [23]	2013	Music genre recognition	LBP	✗	59	Latin Music Database	KNORA	83	Mel scale zoning of spectrogram image enhanced the performance
Dennis et al. [29]	2013	Sound event classification	Sub-band power statistics	✗	600	RWCP	SVM	96	The method is robust against noisy conditions
Costa et al. [24]	2013	Music genre classification	LPQ	✗	256	LMD	SVM	80.78	Global feature extraction performs better compared to zoning
Sharan and Moir [60]	2014	Audio surveillance	MFCC, Central moments	✗	72	RWCP	SVM & KNN	96.59	OAA-SVM gives a better overall performance under noisy conditions
Nanni et al. [48]	2014	Music genre classification	HOG, LPQ, LBP, HARA, LCP, DENSE, WLD, RICLBP, CoALBP, NTLBP	✗	–	LMD,ISMIR 2004	SVM	86.1	Mel scale zoning performance is better compared to linear and global
Kobayashi and Ye [39]	2014	Environmental sound classification	Local statistics	✗	–	RWCP	SVM	98.62	L2-Hellinger normalization increases robustness of the descriptors
Battaglino et al. [9]	2015	Acoustic scene classification	MFCC, LBP codebook	Bag of features	30	DCASE, NXP	SVM	70	Codebook generation reducing the FV dimension and enhances the performance
Agera et al. [5]	2015	Music genre classification	LBP, μLBP & RILBP	✗	325	GTZAN	SVM	84	μLBP and LBP performance is better compared to RICLBP

Table 2 (continued)

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy(%)	Description
Bisot et al. [17]	2015	Acoustic scene classification	SPD & HOG	✗	–	LITIS Rouen	SVM	93.4	Use of earth mover's distance to impart knowledge of suitable distance beteen HOG and SPD features
Montalvo et al. [45]	2015	Language identification	LPQ	✗	256	NIST LRE-2009	SVM	86	This article shows how the spectrogram texture discriminates the musical genre and it is also useful to identify languages.
Sharan and Moir [62]	2015	Audio surveillance	Mean and standard deviation of the central moment	✗	216	RWCP & BBC	SVM & KNN	98.16	Method is noise independent
Sharan and Moir [63]	2015	Audio surveillance	GLCM	✗	16	RWCP & BBC	SVM & KNN	90.29	Features are robust at lower SNR compared to MFCC
Shi et al. [67]	2015	Moving target detection	Entropy, moments & directionality	✗	–	mirco-Doppler dataset	SVM	96	Method characterizes micro-motion targets effectively
Wu and Jang [75]	2015	Music genre classification	Gabor filter	✗	1170	MASD	SVM	88.60	The bag-of-features and late classifier fusion is employed for enhancing classification rate
Lim et al. [42]	2015	Acoustic event classification	LBP & HoG	✗	–	Private	SVM	69.28	A bag-of-audio words approach is used by combining LBP and HoG

Table 3 Summary of different algorithms based on spectrogram time-frequency representation

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy(%)	Description
Nanni et al. [52]	2016	Bird species classification	LBP, LTP & HASC	✗	–	/SACBase	SVM	94.5	Combining multiple spectrogram images improves the performance
Ahmed et al. [6]	2016	Music genre classification	Gradient directional pattern	✗	–	GTZAN	SVM	84.5	Local texture encoding by quantizing the gradient vector is effective for music genre classification
Nanni et al. [50]	2016	Music genre classification	LBP variants, LCP, DENSE, HOG, WLD	SVM ranker	–	LMD, GTZAN, ISMIR 2004	SVM, AdaBoost	86.1	The random subspace approach enhances AdaBoost performance compared to SVM
Ren et al. [57]	2017	Sound event classification	LBP	PCA	200C	RWCP & NTU-SEC	MC-BILBP	96.32	The proposed MC-BILBP and MC-BDLBP shows significant performance enhancement compared SIF & MFCC
Felipe et al. [33]	2017	Acoustic scene classification	LBP, LPQ, RLBP	✗	–	DCASE 2016	SVM	80.17	Performance enhancement is possible when left and right audio channels are used for feature extraction
Lacerda and Mello [40]	2017	Laryngeal mechanisms classification	GLCM	✗	28	LYRICS	SVM	86.16	The method works without any parameter optimization
Nanni et al. [49]	2017	Audio classification	LBP, LPQ, RICLBP, LBPHF, MLPQ, HASC, ELHF, GABOR	✗	–	LMD, ISMIR and GTZAN	SVM	86.2 (LMD)	Visual and acoustic descriptor combination increases classification accuracy

Table 3 (continued)

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy(%)	Description
Nanni et al. [47]	2018	Bird and whale identification	LBP, LPQ, RICLBP, LBPHF, MLPQ, HASC, BSIF, GABOR, AHP, LEN	✗	–	BIRD & WHALE	SVM	93.3	The method also evaluated using LMD, ISMIR and GTZAN database
Valerio et al. [69]	2018	Music genre classification	LBP	✗	–	FMA	C4.5, kNN, MLP, NB, RF and SVM	0.622 (F-score)	The method address data imbalancenes in songs database
Zottesso et al. [85]	2018	Bird species identification	LBP, RLBP, LPQ	✗	374	LifeClef 2015	SVM	71	The dissimilarity approach employed does not require to retrain when new class is added
Yang et al. [78]	2018	Chinese folk song recognition	LBP	Coefficient of Variance	597	Chinese folk song	SVM	89.29	Textural patterns and the intensity information features are used for the classification
Özseven [54]	2018	Speech emotion recognition	GF, HOG, GLCM, WD	✗	816	EMO-DB, eNTERFACE05 & SAVEE	SVM	82.8	Unevenness of classification rate for different database
Demir et al. [28]	2018	Snore sound discrimination	LBP & HOG	✗	2891	MPSSC	SVM	72.6	The method generates large feature vector dimension
Bhattacharjee et al. [10]	2018	Speech-music classification	Spectral peak locations	✗	–	GTZAN	GMM, SVM & RF	98	A method proposes two stage feature extraction approach
Zhang et al. [82]	2018	Bird species identification	ULBP, LBPHF, LPQ, HASC, GF	ReliefF	400	Private	SVM	96.7	The method performs better compared to other techniques

Table 3 (continued)

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy(%)	Description
Li et al. [41]	2019	Animal sound recognition	Projection feature & LBPV	✗	–	Freesound	SVM & RF	98.02	Double spectrogram approach enhances the performance
Spyrou et al. [68]	2019	Emotion recognition	SURF & BoW	✗	Vocabulary size 1500	EMOVO, SAVEE, EMO-DB & MOVIES	SVM	49	The method collects non-linguistic characteristics
Chen et al. [20]	2019	Chatter detection	GLCM	✗	–	Private	SVM	–	The high-energy dominant frequency-bands are extracted using the squared energy operator
Felipe et al. [32]	2019	Identification of infants cry	LPQ, LBP, RLBP	✗	–	Private	SVM	69.03	Classifier fusion is experimented at different noise conditions
Rahmeni et al. [55]	2019	Speech spoofing detection	LBP	✗	10	ASVspoof	SVM	71.67	Voice textures enhances the classification rate
Birajdar et al. [14]	2019	Speech and music classification	GGD statistics	Chaotic crow search algorithm	62	GTZAN, S & S	ELM	100	The algorithm generates better performance without any zoning mechanism and classifier ensemble
Mulimani and Koolagudi [46]	2019	Acoustic event classification	Fisher vectors	PCA	–	UPC-TALP	SVM	94.22	Method is robust to noise
Chowdhury et al. [21]	2020	Indian language identification	LBPHF, CLBP & DWT	GWO	203	Indic TTS	ANN	96.96	Several experiments are conducted with different training-testing configurations
Oo and Oo [53]	2020	Acoustic scene classification	GLCM	PCA	112	SVM	DCASE 2016	83.2	Reduced computation time
Zhang et al. [83]	2020	Speech resampling detection	LBP	✗	59	SVM	TIMIT	99	Capable of detecting various resampling operations

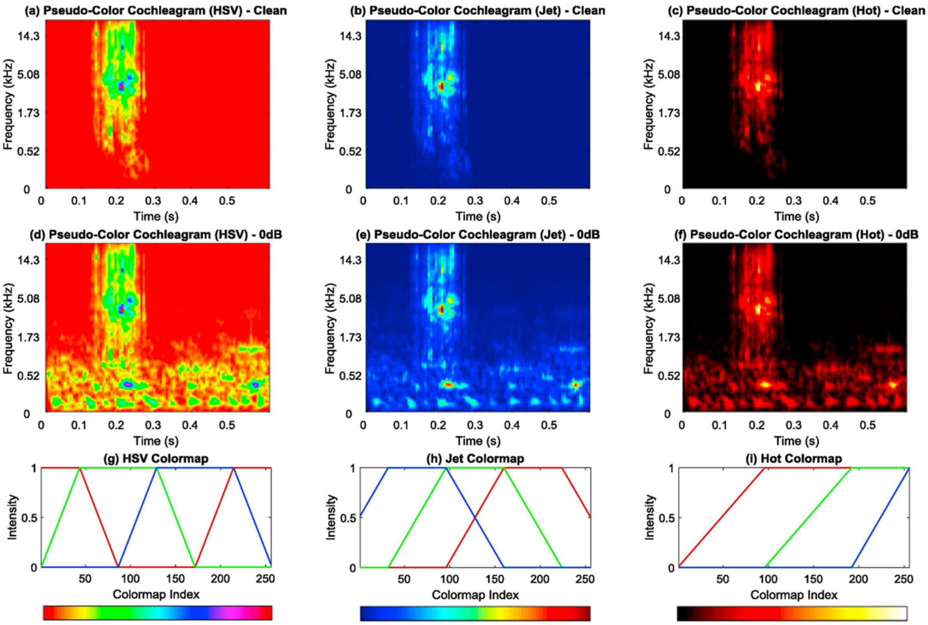


Fig. 12 Different color map representation of a pseudo-color cochleagram of sample sound [65]

In [61], sound signal time-frequency representation based on cochleagram, which uses a gammatone filter, was found very effectual than spectrogram image. Comprehensive classification performance is shown using all three equivalent rectangular bandwidth (ERB) filter models. It is also observed that cochleagram image features at low signal-to-noise ratios (SNRs) give better results. For feature extraction, the work presented in [65] utilizes

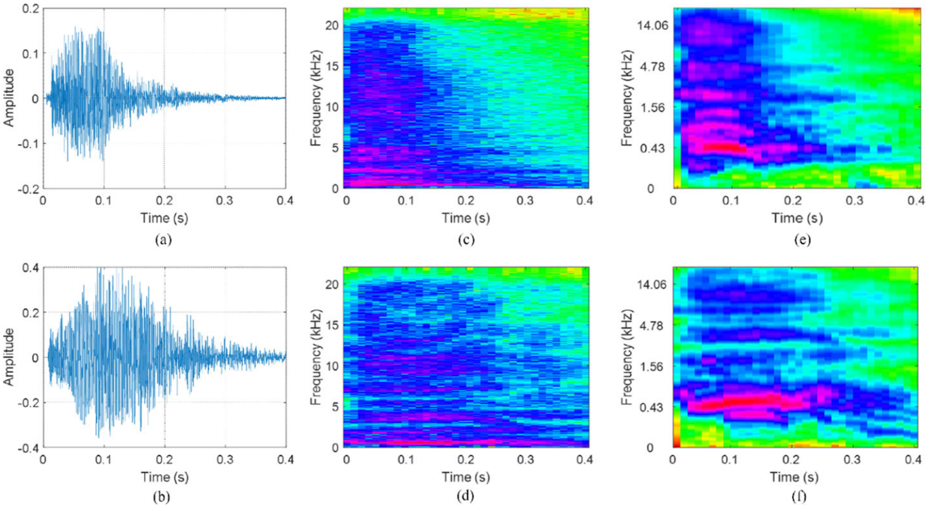


Fig. 13 (a) Time domain normal cough (b) time domain croupy cough (c) spectrogram of normal cough (d) spectrogram of croupy cough (e) cochleagram of normal cough (f) cochleagram of croupy cough [59]

pseudo-color cochleagram image of sound signals for robust acoustic event recognition as illustrated in Fig. 12. For improving characterization from environmental noise, the author mapped grayscale cochleagram image to higher-dimensional color space. The result shows notable improvement at low signal to noise ratios.

An automated cough sounds analysis methods to diagnose croup properly is presented in [59]. In this article, the authors used cochleagram visual representation frequency components based on selectivity property of human cochlea, as shown in Fig. 13. The proposed algorithm results in a sensitivity and selectivity 92.31% and 85.29%, respectively, for croup and non-croup patient classification. In [44], deep neural network back-end classifiers are explored using three different 2-D time-frequency features for audio event classification. Along with the cochleagram, authors utilized spectrogram and CQT based images. Significant improvement in the results are achieved, which shows cochleagram image feature performs well in extreme noise cases of -5dB and -10dB SNR. Indian Language identification using cochleagram image texture descriptors and ANN classifier with 95.36% average accuracy is illustrated in [37]. Cochleagram image-based algorithms and applications are summarized in Table 4.

2.3 Chromagram

In the chromagram, for the music signal shifting of time window results in a chroma features sequence. Each pitch content represents spread over 12 chroma bands within time window [71]. This time-frequency representation is known as Chromagram. A chroma feature vector is also known as pitch class profile (PCP). This is a well-built tool for analyzing music whose tuning closes to equal-tempered scale and whose pitches can be meaningfully sorted. The important property of chroma features is that it can capture harmonic features and melodic features of music, on the contrary of changes in timbre and instrumentation. The chromagram feature vector consists of the 12-dimensional short-time energy distribution of a music signal. These 12 PCPs achieves a frame-wise spectral energy mapping onto spectral bins which correspond to the twelve semi-tones of the chromatic scales for each analysis frame.

Chroma vector utilizes an octave invariance principle which states that there is no functional difference between musical notes separated by doubling of frequency. It is computed with the help of grouping the discrete Fourier transform (DFT) coefficients of a short-term window into 12 bins. Each bin represents one of the 12 equal-tempered pitch classes of Western-type music (semi-tone spacing) [58].

Music specific chromagram representation and ULBP textural feature are utilized for speech/music signal classification in [15]. The use of chromagram representation based visual and spectral features efficiently extracts melodic and harmonic details of music signal otherwise absent in speech as depicted in Fig. 14. In this study, the eigenvector centrality feature selection is used that enhances the detection performance. It was observed 24 bin chromagram representation is sufficient to explorer music tonality features for the speech/music classification.

2.4 Constant-Q Transform (CQT)

The CQT transform provides time-domain to frequency-domain signal transformation producing a log-scale frequency resolution similar to perception of auditory delivering fine

Table 4 Summary of different algorithms based on Cocheleagram time-frequency visual representation

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy (%)	Comments
Sharan and Moir [61]	2015	Sound recognition	Central moments	✗	–	RWCP	SVM	89.03	Significant improvement in accuracy using spectogram compared to cochleagram at low SNR
Sharan and Moir [65]	2018	Acoustic event recognition	Statistical	Sequential backward	132	RWCP, SSD	KNN and SVM	88.80	Works well under noisy environment
Sharan et al. [59]	2019	Cough sound recognition	Central moments	Backward sequential	–	Private	SVM	91.16	Algorithm is solely based on automate croup diagnosis on cough sound analysis
McLoughlin et al. [44]	2019	Audio event classification	SIF, CIF	✗	–	RWCP	DNN	99	Method performs well in noisy conditions
Jog et al. [37]	2019	Indian language identification	CLBP, LPQ, GLCM, BSIF	✗	638	Indic-TTS	ANN	95.36	Large dimensional feature vector is generated

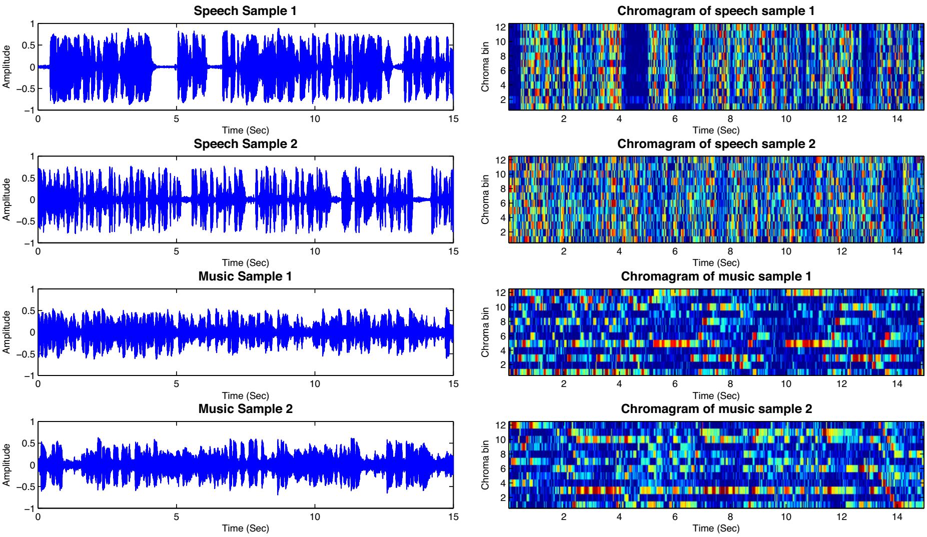


Fig. 14 Chromagram visual representation obtained using 12 bin music and speech signals from Scheirer and Slaney database [15]

resolution at low frequency [1]. In the constant-Q transform (CQT), the $c(k, t)$ over k frequency bins of time domain signal $s(t)$ is defined as

$$c(k,t) = \sum_{n=t-w_s/2}^{t+w_s/2} s(n)ak^*(n - t + w_s/2) \tag{4}$$

where $ak^*(n)$ is the complex conjugate of time–frequency atoms which are defined by and $w(n)$ a window function over length w_s. The major difference between CQT and a spectrogram is that w_s is itself a variable rather than a constant.

The algorithm demonstrated in [80] employs distinct models to use sound textures and events in acoustic scenes. The framework achieved superior results in real data evaluation. With the Rouen dataset, the proposed algorithm performed better compared to other existing approaches. Novel features obtained in [56] by constant Q-transform followed by appropriate pooling. Experimentally it is proved that HOGs computed from constant Q-transform were useful capturing specific features present in time-frequency (TF) representation. This HOG based feature proved globally efficient. Novel zoning approach, along with time-frequency representation (TFR) to improve the classification performance for acoustic scene classification is described in [1]. The technique achieved accuracy up to 95.2%. Abidin et al. [2] presented an algorithm which fuses spectral and temporal features for acoustic scene classification. For the generation of T-F representation, variable Q-transform is used which improved the classification rate by 5.2%. Figure 15 clearly depicts the difference between beach and cafe scenes using CQT visual representation.

In [4], audio signal is converted to CQT representations first and later LBP textural features are extracted from CQT T-F representations. The proposed system achieved an accuracy of 85% on the DCASE 2016 datatset. Joint T-F image-based feature representations are found effective in [3]. These joint features produce better results across a wide range of low and middle frequencies in the audio signal attaining the classification accuracy

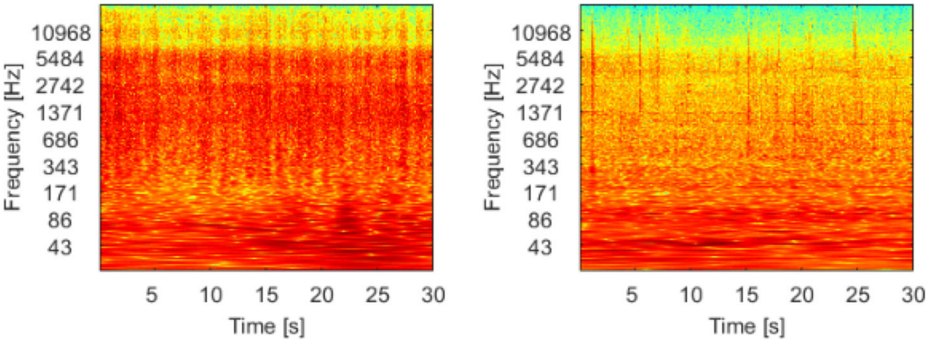


Fig. 15 Constant-Q transform representation for beach sound (left) and cafe sound (right) scenes. [2]

of 83.4%. TFR with zoning technique in combination with image-based features is very productive and computationally efficient for the ASC. In [76], acoustic and visual feature are fused with a various set of features for acoustic scene classification. RelieF algorithm, correlation-based feature (CFS) and principal component analysis (PCA) techniques are used for feature selection. Use of feature selection improved the algorithm performance and reduced the feature vector dimensionality. Table 5 shows a summary of different techniques based on CQT image feature extraction.

2.5 Other time-frequency representations

In this subsection, all remaining time-frequency representation methods are presented. Mel-frequency cepstral coefficients (MFCCs) is one of the most popularly used feature extraction schemes for audio analysis. MFCC filter banks mimic human auditory producing discriminating features in speech processing applications. In [77], the temporal dynamics present in the audio sample is extracted using subband MFCC time-frequency image and LBP texture features for acoustic sound classification. The work explores three frequency bands spanning from 0 to 11 kHz with 23 mseconds of a time window for each frame. The method developed in [77] achieved an improvement of 8% using a D3C ensemble classifier.

Harmonic and percussion images are produced using harmonic-percussion separation (HPSS) algorithm and various texture descriptors extracted from these images are employed for music genre classification in [51]. Application of median filtering across frequency axis percussive occurrence is highlighted, whereas, across the time-bin application of median filtering, the harmonic regions are enhanced. In the same work, authors presented scattergram image-based textural features for music genre classification. The ScatNet scattering framework is employed for generating the scattergram. The speed of the audio can be measured by tempo in beats per minute. The tempogram based feature set is exploited using the novelty curve from the input audio signal for speech and non-speech signal classification in [79]. ThehHighest classification rate of 99.20% is achieved by the proposed approach using multi-layer perceptron classifier and correlation-based feature selection.

Two-dimensional neurogram is generated using physiological computation model of the auditory periphery for phoneme classification and voice activity detection. DCT coefficients from a neurogram image are extracted as features and classified using multi-layer

Table 5 Summary of different algorithms based on Constant-Q transform time-frequency visual representation

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy (%)	Comments
Ye et al. [80]	2015	Acoustic Scene Classification	LBP, HOG	✗	304	Litis Rouen	SVM	96.08	Two-channel information is incorporated to improve scene classification performance.
Rakotomamonjy and Gasso [56]	2015	Audio Scene Classification	HOG	✗	1024 to 3584	Litis Rouen & D-Case Challenge	SVM	89	HoG features efficiently encodes the local variation of power spectrum which is not possible using MFCC.
Abidin et al. [1]	2017	Acoustic Scene Classification	LBP, HOG	✗	1536	LITIS Rouen	SVM	95.2	HoG performance is better compared to LBP.
Abidin et al. [2]	2018	Acoustic Scene Classification	LBP	✗	–	DCASE 2016	SVM	83.4	Score level fusion of the CQT and Mel-spectrum image-based features improved the classification accuracy.
Abidin et al. [4]	2018	Acoustic Scene Classification	LBP, HOG	Random forest	1500	DCASE 2016	SVM	85	The random forest offers flexible control on the feature dimension selection to provide robust and compact features.
Abidin et al. [3]	2018	Acoustic Scene Classification	ECLBP	✗	2700	DCASE 2016	SVM	85.5	The use of variable-Q transform provides finer control over the resolution compared to CQT.
Xie and Zhu [76]	2019	Acoustic Scene Classification	LBP, HoG, Moments	Relief-F, CFS, PCA	3270	TUT Acoustic scenes 2016	KNN, SVM, RF, NN, ELM, KELM	87.44	Enhancement in the accuracy by combining acoustic and visual features is observed.

perceptron for voice activity detection in [36]. A new phoneme classification technique based on discrete Radon transform features is illustrated in [7]. The method exhibited better performance evaluated under noisy conditions compared to other conventional techniques (Table 6).

3 Texture descriptors

Textural descriptors can extract the presence of prominent visual content in time-frequency image. Texture analysis is a process of distinguishing different textures into separate classes by identifying key features. Discerning an effective texture feature is a crucial step for enhancing the algorithm performance. Several texture descriptors such as LBP, LPQ, GLCM, HOG, Gabor filters, central moments, and other LBP variants are used in the literature for an audio classification task. This section briefly summarizes most widely descriptors.

3.1 Local Binary Pattern (LBP) and LBP variants

Local Binary Pattern (LBP) is the most widely texture encoding scheme use in literature. LBP imparts remarkable performance in all audio application algorithms, including music genre recognition, bird species classification, and acoustic scene classification [51, 77]. LBP operates on the local neighbourhood of a central pixel to find a local binary pattern. This is important because of the non-uniformity of the textures in a visual image; usually, local feature extraction is considered during the feature construction phase. The feature vector which describes the textural content of the image corresponds to the histogram of local binary patterns found in all pixels of the image. Two parameters are important during the LBP feature extraction: the first one is the number of neighbouring pixels that will be taken into account for the central pixel (P), the second one is related to the distance between the central pixel and its neighbours (R) [2, 76].

Local binary pattern (LBP) is used in [1, 2, 5, 23, 25, 27, 28, 32, 33, 42, 47–49, 51, 51, 52, 55, 57, 69, 73, 76–78, 80, 83, 85] attaining better performance in various applications. Different LBP variants are also used such as RICLBP, CoALBP, and NTLBP [47–51], RLBP [32, 33, 47, 49, 51, 85], ULP [82], LBPHF [21, 82], ECLBP [3], CLBP [21, 37, 51], and μLBP and RILBP [5, 51].

3.2 Grey-Level Co-occurrence Matrix (GLCM)

The spatial relationship among local pixels is examined in grey-level co-occurrence matrix (GLCM) textural descriptor. This is also known as the gray-level spatial dependence matrix. The GLCM specifies the texture of an image by computing pairs of pixel and characterizing spatial relationship present in an image. Different statistical measures include energy, correlation, energy and homogeneity. GLCM is widely used feature extraction scheme after LBP and HOG in different algorithms [20, 26, 27, 34, 37, 40, 53, 54, 63, 79, 82].

3.3 Histogram of Oriented Gradients (HOG)

Histogram of oriented gradients (HOG) effectively extracts shape and appearance from an image using edge directions or intensity distribution. Similar to LBP. HOG is obtained by dividing input image into small regions and concatenation histogram. HOG descriptors

Table 6 Summary of different algorithms based on MFCC, neurogram, scattergram, and tempogram time-frequency visual representation

Method	Year	Application	Features	Feature selection	FV dimension	Database	Classifiers	Accuracy (%)	Comments
Yang et al. [77]	2017	Acoustic scene classification	LBP, Statistical	✗	–	TUT Acoustic scenes 2016	D3C	80.30	Temporal evolution of short-time features are constructive in classifying environmental sounds.
Nanni et al. [51]	2018	Music genre classification	LBP, LBP-Variant, LPQ, LPQ-Variant, BSIF	✗	500	LMD, GTZAN, ISMIR, BIRD, WHALE	SVM	95.7 (GZAN)	Fusion of various texture descriptors performs better.
Yasmin and Das [79]	2019	speech & non-speech classification	GLCM, skewness, ZCR, tempo	CFS	80	–	SVM, MLP, DT	99.20	The accuracy is improved significantly by utilizing the feature selection algorithm.
Jassim and Harte [36]	2018	Voice activity detection	DCT	✗	320	QUT-NOISE-TIMIT	MLP	EER- 22.63	Generates robust features because of phase-locking characteristics of the neuron in the auditory periphery.
Alam et al. [7]	2018	Phoneme classification	DRT	Laplacian, Fisher, CFS	350	HTIMIT	SVM	83.53	For noisy speech Gammatone ceptral coefficients performance is better.

are popularly used after LBP features. Several algorithms in which HOG descriptors are employed includes [1, 4, 17, 28, 42, 48, 50, 54, 56, 76, 80].

3.4 Local Phase Quantization (LPQ)

A robust symmetric blur descriptor is developed by estimating the phase angles of Fourier transform at different frequencies [24, 45]. LPQ feature extraction approach effectively characterizes underlying textural variations existing in an image. In the literature, various techniques based on LPQ descriptors are [24, 32, 33, 37, 45, 47, 49, 51, 85].

3.5 Other texture features

In addition to the textural features described above, few other descriptors are employed by different researchers. Mostly these features are combined to the LBP, LPQ, GLCM and HOG before the classification. Enhanced algorithm performance is attained by fusing these textural descriptors. These feature are: Block energy [43], Central moments [30, 59–62, 67, 76], Gabor features [47, 49, 73–75, 82], LTP and HASC [52], and WLD: [50].

4 Classifiers

The suitable choice of a classifier is one of the dominant factors in classification. Most widely methodologies used for classification found in various works are support vector machine (SVM), linear discriminant analysis(LDA), artificial neural network (ANN), k-nearest neighbor (KNN) and random forest (RF). This section briefly summarizes the different classification algorithms used.

Support vector machine (SVM) is one of the most widely supervised machine learning algorithms in audio classification goal. The SVM technique is the most common among linear separation algorithms since it is virtually parameter free and has shown that it can have the same or better performance than other more complex algorithms. Variety of kernel functions are employed in SVM such as, linear, polynomial, Gaussian and RBF kernels. SVM implementation using Libsvm is a popular choice of various researchers [32, 83]. However, LibLinear package is also employed in some studies as illustrated in [28]. Almost over 70% algorithms employed SVM as a classification algorithm. For real-world recognition tasks, SVM based multi-class classification method appears to be very appropriate. Various works that employed SVM are [5, 9, 17, 20, 24–30, 32, 39–43, 45, 47, 48, 52, 53, 55, 57, 59, 61–63, 65, 67, 68, 74, 75, 78, 82, 83, 85].

K-nearest neighbors (KNN) is a simple non-parametric algorithm used in pattern recognition. In the KNN algorithm, the training instances of the dataset are extracted as data points in the feature space and divided into several separate classes. To predict the class of a new instance point, initially, it is evaluated in the proposed feature space. KNN is utilized by different researchers in [60, 62, 63, 65, 69, 76, 81].

Artificial neural networks (ANN) are motivated by the functioning of the excitatory or inhibitory neuron connections in the human brain. Multilayer perceptron (MLP) is a feed-forward network, with an input layer, an output layer, and one or more hidden layers. Usually, this network uses the backpropagation technique for training, where the error of prediction is propagated from the output layer to the input layer, modifying interconnection weights trained their models. ANN classifier is used in [21, 36, 37, 69, 76, 79]. Few authors also presented their experiential evaluation using extreme learning machine (ELM) classification [14, 76].

Random forest (RF) is a fast, highly precise, noise resistant ensemble classification algorithm. Random forest classification is utilized in [4, 10, 41, 69, 76]. Additionally, authors used Gaussian mixture model (GMM) [10], linear discriminant analysis (LDA) [73], and RANSAC [34] for audio classification task.

5 Discussions

The survey presents a systematic methodology to compose a comprehensive record of current research dynamics and algorithms focusing on time-frequency image texture features employed in audio applications. Initial attempts in the filed of time-frequency texture image algorithm development were primarily focused on spectrogram based music genre classification [5, 23–27, 43, 73, 74] and acoustic event classification. [9, 29, 30, 39]. Later, this trend is extended for the development of bird and animal sound detection techniques [41, 47, 52, 82, 85]. In addition to this, the application area spans different audio applications like language identification [21, 45], Chinese folk song recognition [78], speech emotion recognition [54], snore sound discrimination [28], speech-music classification [10, 14] and identification of infants cry [32]. Overall, the new time-frequency visualization and texture feature approach is found suitable and efficient in speech, music and audio applications.

Spectrogram image textures are most widely used (almost more than 70%) by the researchers for algorithm development (Fig. 3). Spectrogram representation efficiently characterizes and provides profound attributes present in an audio sample. Moreover, spectrogram image texture is identified in various applications in order to capture the relevant details. Apart from the spectrogram, CQT and cochleagram visualization are also utilized widely. It is also found that CQT representation is better suited for acoustic scene classification as evident from the Table 5. This might be because the CQT image is able to learn the sound texture from acoustic sound efficiently compared to other representations. Whereas, chromagram image texture descriptors are effective in music applications [15, 58].

Texture features play an important role in the classification. From the literature, it is found that, the local binary pattern is the most widely used descriptors for feature extraction from the time-frequency image. It is also evident from the studies that combining different textural descriptors attained superior classification performance as compared to individual feature. For example, GLCM and LBP [27], Gabor and LBP [73], HOG, LPQ, LBP, HARA, LCP, DENSE, WLD, RICLBP, CoALBP, NTLBP [48], LBP, μLBP and RILBP [5], LBP and HoG [42], LBP, LTP and HASC [52], LBP variants, LCP, DENSE, HOG, WLD [50], LBP, LPQ, RICLBP, LBPHF, MLPQ, HASC, ELHF, GABOR [49], and LBP, RLBP and LPQ [85] .

Use of classification algorithm is dominated by support vector machine classifier in different audio application areas (Section 4). Although, some of the works presented evaluation utilizing an artificial neural network, random forest, and K-nearest neighbors classifiers [11–13]. Ensemble classifiers are also explored in [23, 77]. SVM showed good potential even in a noisy environment because of its robustness against such conditions.

It is important to note that, local feature extraction approach attained exceptional performance as compared to global extraction scheme. To accomplish this zoning technique is employed in which the time-frequency image is divided into different regions and then from each region descriptors are extracted [25, 47–50].

The major issue with all the algorithms is feature vector dimensionality. Different texture descriptors such as LBP, LPQ, and their variants generate large dimension vector. This hugely impacts of the training and testing times of the classifier in addition to the classification performance. However, feature selection is rarely addressed in the literature like PCA [46, 57], coefficient of variance [78], ReliefF [82], chaotic crow search algorithm [14], and GWO [21]. In future, it is worthy of exploring the effect of feature selection when a combination of textural descriptors are employed in the application development.

6 Conclusion

In this survey, a comprehensive overview of state-of-the-art research works on time-frequency texture image features in audio classification algorithms is presented. Firstly, we identified salient characteristics from the existing literature, and a generalized architecture of time-frequency texture feature extraction approach in audio classification algorithms is presented which we believe helps to new researchers in this area to comprehend overall composition. Later, key characteristics and categories of time-frequency visual representations are identified along with dominant texture feature extraction algorithms. Various time-frequency visualization algorithms in diverse audio applications are categorized and compared using their key aspects. A brief discussion of feature selection approaches utilized in several applications are also explored. Finally, some open research challenges and future trends in these fields are outlined.

Funding Funding No funding was received to assist with the preparation of this manuscript.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Abidin S, Togneri R, Sohel F (2017) Enhanced lbp texture features from time frequency representations for acoustic scene classification. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 626–630. https://doi.org/10.1109/ICASSP.2017.7952231
2. Abidin S, Togneri R, Sohel F (2018) Acoustic scene classification using joint time-frequency image-based feature representations. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6. https://doi.org/10.1109/AVSS.2018.8639164
3. Abidin S, Togneri R, Sohel F (2018) Spectrotemporal analysis using local binary pattern variants for acoustic scene classification. IEEE/ACM Trans Audio Speech Lang Process 26(11):2112-2121. https://doi.org/10.1109/TASLP.2018.2854861
4. Abidin S, Xia X, Togneri R, Sohel F (2018) Local binary pattern with random forest for acoustic scene classification. In: 2018 IEEE international conference on multimedia and expo, ICME 2018. IEEE, institute of electrical and electronics engineers, United States, vol 2018-July. https://doi.org/10.1109/ICME.2018.8486578
5. Agera N, Chapaneri S, Jayaswal D (2015) Exploring textural features for automatic music genre classification. In: 2015 International conference on computing communication control and automation, pp 822–826. https://doi.org/10.1109/ICCUBEA.2015.164

6. Ahmed F, Paul PP, Gavrilova M (2016) Music genre classification using a gradient-based local texture descriptor. In: Czarnowski I, Caballero AM, Howlett RJ, Jain LC (eds) Intelligent decision technologies 2016. Springer international publishing, Cham, pp 455–464. https://doi.org/10.1007/978-3-319-39627-9-40
7. Alam MS, Jassim WA, Zilany MSA (2018) Radon transform of auditory neurograms: a robust feature set for phoneme classification. IET Sig Process 12(3):260-268. https://doi.org/10.1049/iet-spr.2017.0170
8. Ashfaque Mostafa T, Soltaninejad S, McIsaac TL, Cheng I (2021) A comparative study of time frequency representation techniques for freeze of gait detection and prediction. Sensors, vol 21(19). https://doi.org/10.3390/s21196446
9. Battaglino D, Lepauloux L, Pilati L, Evans N (2015) Acoustic context recognition using local binary pattern codebooks. In: 2015 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA), pp 1–5. https://doi.org/10.1109/WASPAA.2015.7336886
10. Bhattacharjee M, Prasanna SRM, Guha P (2018) Time-frequency audio features for speech-music classification
11. Bhatti UA, Huang M, Wu D, Zhang Y, Mehmood A, Han H (2019) Recommendation system using feature extraction and pattern recognition in clinical care systems. Enterpr Inf Syst 13(3):329–351. https://doi.org/10.1080/17517575.2018.1557256
12. Bhatti UA, Ming-Quan Z, Qing-Song H, Ali S, Hussain A, Yuhuan Y, Yu Z, Yuan L, Nawaz SA (2021) Advanced color edge detection using clifford algebra in satellite images. IEEE Photon J 13(2):1–20. https://doi.org/10.1109/JPHOT.2021.3059703
13. Bhatti UA, Zhaoyuan Y, Linwang Y, Zeeshan Z, Ali NS, Mughair B, Anum M, Ul AQ, Luo W (2020) Geometric algebra applications in geospatial artificial intelligence and remote sensing image processing. IEEE Access 8:155783–155796. https://doi.org/10.1109/ACCESS.2020.3018544
14. Birajdar GK, Patil MD (2019) Speech and music classification using spectrogram based statistical descriptors and extreme learning machine. Multimed Tools Appl 78(11):15141–15168. https://doi.org/10.1007/s11042-018-6899-z
15. Birajdar GK, Patil MD (2020) Speech/music classification using visual and spectral chromagram features. J Ambient Intell Humanized Comput 11:329–347. https://doi.org/10.1007/s12652-019-01303-4
16. Birajdar GK, Raveendran S (2022) Indian language identification using time-frequency texture features and kernel ELM. J Ambient Intell Humanized Comput:1–12. https://doi.org/10.1007/s12652-022-03781-5
17. Bisot V, Essid S, Richard G (2015) HOG and subband power distribution image features for acoustic scene classification. In: 2015 23rd European signal processing conference (EUSIPCO), pp 719–723. https://doi.org/10.1109/EUSIPCO.2015.7362477
18. Breve B, Cirillo S, Cuofano M, Desiato D (2020) Perceiving space through sound: mapping human movements into MIDI. In: 26th International conference on distributed multimedia systems, virtual conference center, USA, pp 49–56. https://doi.org/10.18293/DMSVIVA20-011
19. Breve B, Cirillo S, Cuofano M, Desiato D (2022) Enhancing spatial perception through sound: mapping human movements into MIDI. Multimed Tools Appl 81(1):73–94. https://doi.org/10.1007/s11042-021-11077-7
20. Chen Y, Li H, Hou L, Bu X (2019) Feature extraction using dominant frequency bands and time-frequency image analysis for chatter detection in milling. Precis Eng 56:235–245. https://doi.org/10.1016/j.precisioneng.2018.12.004
21. Chowdhury AA, Borkar VS, Birajdar GK (2020) Indian language identification using time-frequency image textural descriptors and gwo-based feature selection. J Exp Theor Artif Intell 32(1):111–132. https://doi.org/10.1080/0952813X.2019.1631392
22. Connolly J, Edmonds E, Guzy J, Johnson S, Woodcock A (1986) Automatic speech recognition based on spectrogram reading. Int J Man-Mach Stud 24(6):611–621. https://doi.org/10.1016/S0020-7373(86)80012-8 . http://www.sciencedirect.com/science/article/pii/S0020737386800128
23. Costa Y, Oliveira L, Koerich A, Gouyon F (2013) Music genre recognition based on visual features with dynamic ensemble of classifiers selection. In: 2013 20th International conference on systems, signals and image processing (IWSSIP), pp 55–58
24. Costa Y, Oliveira L, Koerich A, Gouyon F (2013) Music genre recognition using gabor filters and LPQ texture descriptors. In: Ruiz-Shulcloper J, Sanniti di Baja G (eds) Progress in pattern recognition, image analysis, computer vision, and applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 67–74. https://doi.org/10.1007/978-3-642-41827-3-9
25. Costa Y, Oliveira L, Koerich A, Gouyon F, Martins J (2012) Music genre classification using LBP textural features. Sig Process 92(11):2723–2737. https://doi.org/10.1016/j.sigpro.2012.04.023
26. Costa YMG, Oliveira LS, Koericb AL, Gouyon F (2011) Music genre recognition using spectrograms. In: 2011 18th International conference on systems, signals and image processing, pp 1–4

27. Costa YMG, Oliveira LS, Koerich AL, Gouyon F (2012) Comparing textural features for music genre classification. In: The 2012 international joint conference on neural networks (IJCNN), pp 1–6. https://doi.org/10.1109/IJCNN.2012.6252626
28. Demir F, Sengür A, Cummins N, Amiriparian S, Schuller BW (2018) Low level texture features for snore sound discrimination. In: 2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 413–416. https://doi.org/10.1109/EMBC.2018.8512459
29. Dennis J, Tran HD, Chng ES (2013) Image feature representation of the subband power distribution for robust sound event classification. IEEE Trans Audio Speech Lang Process 21(2):367–377. https://doi.org/10.1109/TASL.2012.2226160
30. Dennis J, Tran HD, Li H (2011) Spectrogram image feature for sound event classification in mismatched conditions. IEEE Sig Process Lett 18(2):130–133. https://doi.org/10.1109/LSP.2010.2100380
31. Dutta A, Sil D, Chandra A, Palit S (2022) Cnn based musical instrument identification using time-frequency localized features. Int Technol Lett 5(1):e191. https://doi.org/10.1002/itl2.191
32. Felipe GZ, Aguiar RL, Costa YMG, Silla C, Brahnam S, Nanni L, McMurtrey S (2019) Identification of infants' cry motivation using spectrograms. In: 2019 International conference on systems, signals and image processing (IWSSIP), pp 181–186. https://doi.org/10.1109/IWSSIP.2019.8787318
33. Felipe GZ, Maldonado Y, Costa DG, Helal LG (2017) Acoustic scene classification using spectrograms. In: 2017 36th International conference of the chilean computer science society (SCCC), pp 1–7. https://doi.org/10.1109/SCCC.2017.8405119
34. Ghosal A, Chakraborty R, Dhara BC, Saha SK (2012) Song/instrumental classification using spectrogram based contextual features. In: Proceedings of the CUBE international information technology conference, CUBE '12. Association for computing machinery, New York, NY, USA, pp 21–25. https://doi.org/10.1145/2381716.2381722
35. Godbole S, Jadhav V, Birajdar G (2020) Indian language identification using deep learning. ITM Web Conf 32:01010. https://doi.org/10.1051/itmconf/20203201010
36. Jassim WA, Harte N (2018) Voice activity detection using neurograms. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5524–5528. https://doi.org/10.1109/ICASSP.2018.8461952
37. Jog AH, Jugade OA, Kadegaonkar AS, Birajdar GK (2018) Indian language identification using cochleagram based texture descriptors and ann classifier. In: 2018 15th IEEE India council international conference (INDICON), pp 1–6. https://doi.org/10.1109/INDICON45594.2018.8987167
38. Klatt D, Stevens K (1973) On the automatic recognition of continuous speech:implications from a spectrogram-reading experiment. IEEE Trans Audio Electroacoustics 21(3):210–217. https://doi.org/10.1109/TAU.1973.1162453
39. Kobayashi T, Ye J (2014) Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3052–3056. https://doi.org/10.1109/ICASSP.2014.6854161
40. Lacerda EB, Mello CA (2017) Automatic classification of laryngeal mechanisms in singing based on the audio signal. Procedia Comput Sci 112:2204–2212. https://doi.org/10.1109/ICASSP.2014.6854161
41. Li Y, Huang H, Wu Z (2019) Animal sound recognition based on double feature of spectrogram. Chinese J Electron 28(4):667–673. https://doi.org/10.1049/cje.2019.04.005
42. Lim H, Kim MJ, Kim H (2015) Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation. In: INTERSPEECH, pp 3325–3329
43. Matsui T, Goto M, Vert J, Uchiyama Y (2011) Gradient-based musical feature extraction based on scale-invariant feature transform. In: 2011 19th European signal processing conference, pp 724–728
44. McLoughlin IV, Xie Z, Song Y, Phan H, Palaniappan R (2020) Time-frequency feature fusion for noise-robust audio event classification. Circ Syst Sig Process 39:1672–1687. https://doi.org/10.1007/s00034-019-01203-0
45. Montalvo A, Costa YMG, Calvo JR (2015) Language identification using spectrogram texture. In: Pardo A, Kittler J (eds) Progress in pattern recognition, image analysis, computer vision, and applications. Springer international publishing, Cham, pp 543–550. https://doi.org/10.1007/978-3-319-25751-8-65
46. Mulimani M, Koolagudi SG (2019) Robust acoustic event classification using fusion fisher vector features. Appl Acoust 155:130–138. https://doi.org/10.1016/j.apacoust.2019.05.020
47. Nanni L, Aguiar RL, Costa YMG, Brahnam S, Silla CN, Brattin RL, Zhao Z (2018) Bird and whale species identification using sound images. IET Comput Vis 12(2):178–184. https://doi.org/10.1049/iet-cvi.2017.0075
48. Nanni L, Costa Y, Brahnam S (2014) Set of texture descriptors for music genre classification
49. Nanni L, Costa Y, Lucio D, Silla C, Brahnam S (2017) Combining visual and acoustic features for audio classification tasks. Pattern Recog Lett 88:49–56. https://doi.org/10.1016/j.patrec.2017.01.013

50. Nanni L, Costa YM, Lumini A, Kim MY, Baek SR (2016) Combining visual and acoustic features for music genre classification. Expert Syst Appl 45:108–117. https://doi.org/10.1016/j.eswa.2015.09.018
51. Nanni L, Costa YMG, Aguiar RL, Jr CNS, Brahnam S (2018) Ensemble of deep learning, visual and acoustic features for music genre classification. J New Music Res 47(4):383–397. https://doi.org/10.2018.1438476
52. Nanni L, Costa YMG, Lucio DR, Silla C, Brahnam S (2016) Combining visual and acoustic features for bird species classification. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI), pp 396–401. https://doi.org/10.1109/ICTAI.2016.0067
53. Oo MM, Oo LL (2020) Fusion of Log-Mel spectrogram and GLCM feature in acoustic scene classification. Springer international publishing, Cham, pp 175–187. https://doi.org/10.1007/978-3-030-24344-9-11
54. Özseven T (2018) Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. Appl Acoust 142:70–77. https://doi.org/10.1016/j.apacoust.2018.08.003
55. Rahmeni R, Ben Aicha A, Ben Ayed Y (2019) On the contribution of the voice texture for speech spoofing detection. In: 2019 19th International conference on sciences and techniques of automatic control and computer engineering (STA), pp 501–505
56. Rakotomamonjy A, Gasso G (2015) Histogram of gradients of time-frequency representations for audio scene classification. IEEE/ACM Trans Audio Speech Lang Process 23(1):142–153. https://doi.org/10.1109/TASLP.2014.2375575
57. Ren J, Jiang X, Yuan J, Magnenat-Thalmann N (2017) Sound-event classification using robust texture features for robot hearing. IEEE Trans Multimed 19(3):447–458. https://doi.org/10.1109/TMM.2016.2618218
58. Sell G, Clark P (2014) Music tonality features for speech/music discrimination. 2014. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP) pp 2489–2493. https://doi.org/10.1109/ICASSP.2014.6854048
59. Sharan RV, Abeyratne UR, Swarnkar VR, Porter P (2019) Automatic croup diagnosis using cough sound recognition. IEEE Trans Biomed Eng 66(2):485–495. https://doi.org/10.1109/TBME.2018.2849502
60. Sharan RV, Moir TJ (2014) Audio surveillance under noisy conditions using time-frequency image feature. In: 2014 19th International conference on digital signal processing, pp 130–135. https://doi.org/10.1109/ICDSP.2014.6900815
61. Sharan RV, Moir TJ (2015) Cochleagram image feature for improved robustness in sound recognition. In: 2015 IEEE international conference on digital signal processing (DSP), pp 441–444. https://doi.org/10.1109/ICDSP.2015.7251910
62. Sharan RV, Moir TJ (2015) Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM. Neurocomputing 158:90–99. https://doi.org/10.1016/j.neucom.2015.02.001
63. Sharan RV, Moir TJ (2015) Robust audio surveillance using spectrogram image texture feature. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1956–1960. https://doi.org/10.1109/ICASSP.2015.7178312
64. Sharan RV, Moir TJ (2015) Subband spectral histogram feature for improved sound recognition in low SNR conditions. In: 2015 IEEE international conference on digital signal processing (DSP), pp 432–435. https://doi.org/10.1109/ICDSP.2015.7251908
65. Sharan RV, Moir TJ (2018) Pseudo-color cochleagram image feature and sequential feature selection for robust acoustic event recognition. Appl Acoust 140:198–204. https://doi.org/10.1016/j.apacoust.2018.05.030
66. Sharma G, Umapathy K, Krishnan S (2020) Trends in audio signal feature extraction methods. Appl Acoust 158:107020. https://doi.org/10.1016/j.apacoust.2019.107020
67. Shi X, Zhou F, Liu L, Zhao B, Zhang Z (2015) Textural feature extraction based on time-frequency spectrograms of humans and vehicles. IET Radar Sonar Navig 9(9):1251–1259. https://doi.org/10.1049/iet-rsn.2014.0432
68. Spyrou E, Nikopoulou R, Vernikos I, Mylonas P (2019) Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms. Technologies, vol 7(1). https://doi.org/10.3390/technologies7010020
69. Valerio VD, Pereira RM, Costa YMG, Bertolini D, Silla CN (2018) A resampling approach for imbalanceness on music genre classification using spectrograms. In: Thirty-first international florida artificial intelligence research society conference (FLAIRS), pp 500–505
70. Vyas S, Patil MD, Birajdar GK (2021) Classification of heart sound signals using time-frequency image texture features, Chapter 5, Wiley, pp 81–101. https://doi.org/10.1002/9781119818717.ch5

71. Wakefield GH (1999) Mathematical representation of joint time-chroma distributions. pp 3807–3807-9. https://doi.org/10.1117/12.367679
72. Wu H, Zhang M (2012) Gabor-lbp features and combined classifiers for music genre classification. In: Proceedings of the 2012 2nd international conference on computer and information application (ICCIA 2012), pp 419–423. Atlantis Press. https://doi.org/10.2991/iccia.2012.101
73. Wu HQ, Zhang M (2013) Gabor-lbp features and combined classifiers for music genre classification. In: Information technology applications in industry, computer engineering and materials science, advanced materials research, vol 756, pp 4407-4411. Trans Tech Publications Ltd. https://doi.org/10.4028/www.scientific.net/AMR.756-759.4407
74. Wu M, Chen Z, Jang JR, Ren J, Li Y, Lu C (2011) Combining visual and acoustic features for music genre classification. In: 2011 10th International conference on machine learning and applications and workshops, vol 2, pp 124–129. https://doi.org/10.1109/ICMLA.2011.48
75. Wu MJ, Jang JSR (2015) Combining acoustic and multilevel visual features for music genre classification. ACM Trans Multimed Comput Commun Appl, vol 12(1). https://doi.org/10.1145/2801127
76. Xie J, Zhu M (2019) Investigation of acoustic and visual features for acoustic scene classification. Expert Syst Appl 126:20–29. https://doi.org/10.1016/j.eswa.2019.01.085
77. Yang W, Krishnan S, Yang W, Krishnan S (2017) Combining temporal features by local binary pattern for acoustic scene classification. IEEE/ACM Trans Audio Speech Lang Proc 25(6):1315–1321. https://doi.org/10.1109/TASLP.2017.2690558
78. Yang X, Luo J, Wang Y, Zhao X, Li J (2018) Combining auditory perception and visual features for regional recognition of chinese folk songs. In: Proceedings of the 2018 10th international conference on computer and automation engineering, ICCAE 2018. Association for computing machinery, New York, NY, USA, pp 75–81. https://doi.org/10.1145/3192975.3193006
79. Yasmin G, Das AK (2019) Speech and non-speech audio files discrimination extracting textural and acoustic features. In: Bhattacharyya S, Mukherjee A, Bhaumik H, Das S, Yoshida K (eds) Recent trends in signal and image processing. Springer Singapore, Singapore, pp 197–206. https://doi.org/10.1007/978-981-10-8863-6_20
80. Ye J, Kobayashi T, Murakawa M, Higuchi T (2015) Acoustic scene classification based on sound textures and events. In: Proceedings of the 23rd ACM international conference on multimedia. Association for computing machinery, New York, NY, USA, pp 1291–1294. https://doi.org/10.1145/2733373.2806389
81. Yu G, Slotine JJE (2009) Audio classification from time-frequency texture. In: 2009 IEEE international conference on acoustics, speech and signal processing pp 1677–1680. https://doi.org/10.1109/ICASSP.2009.4959924
82. Zhang S, Zhao Z, Xu Z, Bellisario K, Pijanowski BC (2018) Automatic bird vocalization identification based on fusion of spectral pattern and texture features. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 271–275. https://doi.org/10.1109/ICASSP.2018.8462156
83. Zhang Y, Dai S, Song W, Zhang L, Li D (2020) Exposing speech resampling manipulation by local texture analysis on spectrogram images. Electronics 9(1):1–23. https://doi.org/10.3390/electronics9010023
84. Zhang Y, Zhang K, Wang J, Su Y (2021) Robust acoustic event recognition using AVMD-PWVD time-frequency image. Appl Acoust 178:107970. https://doi.org/10.1016/j.apacoust.2021.107970
85. Zottesso RH, Costa Y, Bertolini D, Oliveira L (2018) Bird species identification using spectrogram and dissimilarity approach. Ecol Inform 48:187–197. https://doi.org/10.1109/ICASSP.1979.1170735
86. Zue V, Cole R (1979) Experiments on spectrogram reading. In: ICASSP '79. IEEE international conference on acoustics, speech, and signal processing, vol 4, pp 116–119. https://doi.org/10.1109/ICASSP.1979.1170735
87. Zue V, Lamel L (1986) An expert spectrogram reader: a knowledge-based approach to speech recognition. In: ICASSP '86. IEEE international conference on acoustics, speech, and signal processing, vol 11, pp 1197–1200. https://doi.org/10.1109/ICASSP.1986.1168798

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yogita D. Mistry She received B.E. degree in Electronics & Telecommunication Engineering from Govt. College of Engineering, Pune in 2002 and ME in Electronics Engineering from Prof. Ram Meghe Institute of Technology & Research Badnera in 2010 and completed Ph.D. in Electronics & Telecommunication Engineering from Amravati University in 2018. Currently she is the Associate Professor in the department of Electronics Engineering in Ramrao Adik Institute of Technology, Nerul in Mumbai, India. Her current research interests include CBIR, image processing, signal processing and image compression. She is a Life Member of the Indian Society for Technical Education (ISTE).

Gajanan K. Birajdar obtained his M. Tech. (Electronics and Telecommunication Engineering) from Dr. Babasaheb Ambedkar Technological University, Maharashtra, India, in 2004 and Ph. D. in the area of blind image forensics from Nagpur University, India, in 2018. He is working in the Department of Electronics Engineering, Ramrao Adik Institute of Technology Nerul, Navi Mumbai, University of Mumbai. He has published over 25-refereed papers and several copyrights/patents, most in the areas of signal and image processing. He has served on the program committees of various conferences/workshops and member of several prestigious professional bodies. His current research interests are multimedia, speech processing and soft computing.

Archana M. Khodke is presently working as Assistant Professor in Ramrao Adik Institute Of Technology, Nerul, Navi Mumbai since 2014. She also worked as Assistant Professor in Padmabhushan Vasantdada Patil Pratisthan's College of Engineering, Sion, Mumbai for nine years from July 2005 to July 2014. She has teaching experience of 16 years. She has completed her B.E in Electronics Engineering from Nagpur University and M.E in Electronics Engineering from Mumbai University.