OPEN ACCESS Freely available online
G. Sharma, K. Umapathy, and S. Krishnan, “Trends in Audio Texture Analysis, Synthesis, and Applications”
J. Audio Eng. Soc., vol. 70, no. 3, pp. 108–127, (2022 March).
DOI: https://doi.org/10.17743/jaes.2021.0060

Trends in Audio Texture Analysis, Synthesis, and Applications

GARIMA SHARMA, KARTHIKEYAN UMAPATHY, AND SRIDHAR KRISHNAN
(garima.sharma@ryerson.ca) (kumapath@ryerson.ca) (krishnan@ryerson.ca)

Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada

Audio signals are classified into speech, music, and environmental sounds. From the evolution of audio features, an adequate amount of work has been seen in speech and music processing. On the other hand, the environmental sounds have not been studied that much, and the major reason behind it is the lack of coherent information present in an environmental sound compared with the speech signal or a musical sound. The definition to express audio textures is imprecise and insufficient, so audio textures tend to be defined by drawing a comparison to the known sound source (e.g., “it sounds like a motor” or “like a fan”). Audio textures could be either natural or artificial. Natural audio textures, such as heavy rain, fire, and stream flowing, are very common. The artificial audio textures include sounds such as applause, a motor running, someone walking on gravel, babble, and many more. Although these audio textures have been used in virtual reality, music, screen saver sounds, and more, a considerable amount of possible work is still untouched. The aim of this study is to summarize the literature on audio textures, textural features, and their applications. In this survey, the texture synthesis and features are explained in detail.

0 INTRODUCTION

Humans have an extremely fine sense to feel, classify, and analyze various things. The modern machine learning and deep learning models try to mimic that fine sense of humans in order to design smart machines. Texture is one of the things that humans can classify very well. For example, one can easily classify texture of the surface as smooth or rough just by touching it. Similarly, just by viewing, one can classify the texture of an image as coarse or smooth, uniform or non-uniform, symmetric or non-symmetric, natural or artificial, etc., and in the same way, humans have a capability of classifying and analyzing textures present in the audio signals.

Audio textures are present everywhere in the environment, but they are largely not studied. The major reason behind not studying audio texture is that, unlike speech and music, there is no coherent information or organized message present in the audio texture [1]. But in the last few decades, textures have been widely used in background sounds in movies, augmented reality, virtual reality, mobile applications, screen saver sounds, music, lullabies, audio texture synthesis, and many more [2]. As the knowledge about audio texture grows, the more and more applications are coming up. Currently audio textures are being used in rejuvenation therapies where long duration calm audio textures like ocean waves, forest rain, etc. are played to release stress and tension from one’s brain and body. There are many mobile applications that help people sleep, get calm, and feel relaxed by using audio textures [3].

The most common textures in the environment include rainfall, wind blowing, cricket sounds, frog calls, a running tap, and many more. Despite having such a close connection with these audio textures, it is quite difficult to explain them in words unless it is heard. This explanation of textures is not precise and hence quite vague in nature (e.g., an audio sounds like wind, a machine running, etc.). For machine learning, audio textures are either captured from the audio itself or its corresponding time-frequency representation (TFR) such as a spectrogram. Fig. 1 shows the time-domain and corresponding spectrograms for a speech signal and audio texture. It can be interpreted from the time domain representation that audio textures have a uniform pattern compared with the speech signal. In the case of

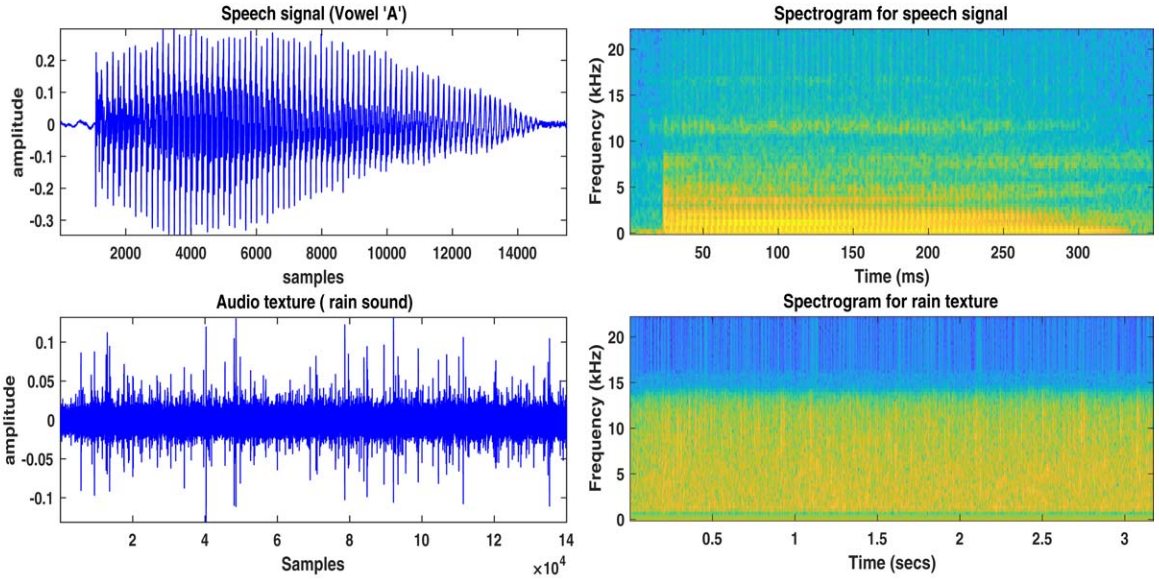

Fig. 1. Time domain and time-frequency representations for a speech signal and audio texture.

a speech spectrogram, the formants are somewhat visible. The frequency spread is concentrated at the lower end in this case. While on the other side, for audio textures, the spectrograms are pretty much uniform throughout the frequency axis, and no formants are clearly visible.

The broad categories into which audio textures could be grouped are natural audio textures (e.g., rain and wind), animal/bird audio textures (e.g., crickets humming, seagull, and frog calls), audio from activity (e.g., running on gravel, typing, and applauding), machine audio textures (e.g., air conditioner, lathe, and grinder), and human utterances (e.g., sounds in a restaurant, babble, and crowd noise). Sounds from musical instruments (e.g., bongo and guitar) are not considered textures because these sounds are rich in harmonic content.

This review paper discusses the trends of audio texture analysis methods, synthesis algorithms, and their various applications. The definitions of audio textures given by various researchers throughout the evolution of audio textures are reported. This paper summarizes the evolution history of audio textures and discusses each of its phases in detail. This work summarizes all the relevant literature on audio texture analysis and synthesis algorithms.

The rest of the paper is organized as follows. Sec. 1 discusses the audio texture definitions. Sec. 2 explains audio texture analysis methods including statistical, image-based, and timbre-based features. Sec. 3 describes the synthesis methods for audio textures in detail including the modular, granular, and deep learning–based methods. Sec. 4 highlights some of the application areas where audio textures are used, and the last section summarizes this survey article.

1 AUDIO TEXTURE DEFINITIONS

Various researchers have given various definitions to explain the textures. The earliest attempt to explain texture was in the context of visual textures and was proposed in 1962 by Julesz [4]. In this article, the author proposed a theory called "Julesz's conjecture" and stated that humans are not able to distinguish between visual textures if they have similar second-order statistics (i.e., variance of the image). However this theory was later proved false in [5]. Since then image or visual textures have been studied and explored deeply in various applications [6].

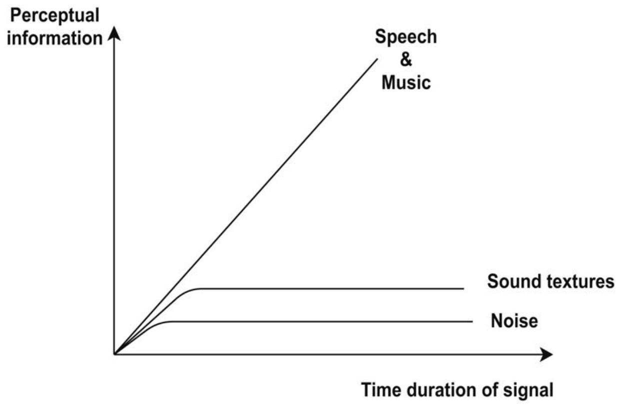

Fig. 2. Potential information in audio textures [1].

On the other side, the earliest attempt to explain audio textures was done by Nicholas Saint-Arnaud in 1995 [1]. In this the authors have given the following definition to the audio textures: "Sound textures are formed by the basic sound elements called atoms; atoms occur as per a high level pattern which could be periodic or random; the high level parameters must remain same over a long period of time; the parameters must be exposed within few seconds and high level randomness is acceptable as long as enough occurrences are present with in the attention span."

Also in [1] the relation between potential information is established with speech and music, texture, and noise. The relation is shown in Fig. 2. The relations explains that the

Table 1. Summary of the texture definitions based on initial work.

Year	Citation	Texture	Texture property
1962	[4]	Visual	Julesz conjecture
1978	[5]	Visual	Improved Julesz conjecture
1995	[1]	Audio	Sound has audio atoms
1998	[7]	Audio	Audio as two-level representation

information content present in the sound textures is more than the noise but way less than the speech or music signals.

Later in 1998 the audio texture synthesis was proposed for the first time by Saint-Arnaud and Popat in [7]. In this the audio textures were defined as a two-level representation, where the first level states that textures are made of atoms and second level explains all the probability-based transitions between atoms. The atom in the audio texture is the smallest unit present in the texture sound, for example, a fire crackle in a fire texture sound or raindrop in a rain texture sound. In [7] the authors proposed that there could be more than one atom present in an audio texture. During the synthesis of audio textures, the authors found it very difficult to separate and categorize the atoms from the texture sound [7]. Table 1 summaries the definitions provided for textures during initial works. The visual textures were defined well before audio textures.

2 AUDIO TEXTURE ANALYSIS

Plenty of work has been done in the field of audio signal processing using the various types of audio features, such as based on time domain, frequency domain, cepstral domain, textural features, and more [8]. These features are normally frame-level features, which means the audio clip is divided into frames, and features are extracted from each frame. There are direct and indirect methods to capture the temporal structure of an audio from its time-frequency representations. In recent classification algorithms, these frame-level features are integrated before being fed into the classifier. This method is called temporal feature integration (TFI) [9], and the frames from which these temporal features are extracted are called texture windows [10, 11]. This integration method reduces the within-class variability and hence improves the performance of the classifier. Textural features are actually those handcrafted features that define the texture present in an audio. In literature, the textural features are classified into the following classes: statistical, image-based, and timbre-based features. This section explores all these textural features.

2.1 Statistical Features

The statistical texture features are easy to extract and understand. The first four statistical parameters (mean, variance, skewness, and kurtosis) have been generally used in many classification applications, but extracting these parameters directly from the audio clip does not give a considerable amount of information because of the basic homogeneous nature of the audio textures. Hence these statistical parameters are extracted from other audio domains [12, 13]. The most explored statistical moments are from conditional domains, temporal correlations, spectral correlations, tempo-spectral correlation, mean instantaneous frequency [14], and mean crossing rate.

Algorithm 1. Conditional spectral moments.

```
1. Result: conditional spectral moments.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Find $P(t, \omega)$, the spectrogram of $x_i(t)$.
4. Choose the value of moment $m$ between 1 and 4.
5. Calculate marginal distribution $P(t)$ of $x_i(t)$.
6. Calculate the conditional spectral moment using Eq. (1).
```

2.1.1 *Conditional Moments*

The moments first include fourth-order statistical parameters called mean, variance, skewness, and kurtosis. Mean describes the average, variance shows the spread range, skewness reflects the symmetry in the histogram, and kurtosis shows the sparsity present in the data [13, 15]. When analyzing statistics from the TFR, standard statistical moments do not make much sense. The conditional moments are calculated by keeping one parameter constant, either time or frequency. Such conditional moments are called conditional spectral moments, conditional temporal moments, and the joint time-frequency moment [16]. These conditional moments are currently being used in detecting faults in machines, bearings, or gears [17, 18]. The sounds produced by these motors could be considered audio textures.

2.1.1.1 Conditional spectral moments. The conditional spectral moments of a signal describe how the signal spectrum is evolving in time. The moments of the sub-bands histogram help to distinguish between sound textures that have fairly steady power (e.g., in classic filtered noise) in sub-bands versus the sub-bands having few, large, and sparse values (e.g., crackling fire). Mathematically conditional spectral moment is a function of frequency, given time and its marginal distribution. Assume there is a signal $x_i(t)$ and its spectrogram power spectrum is $P(t, \omega)$. For N number of sub-bands, the conditional spectral moment [19] is α when t is given and is described as

$$[\alpha^m|t] = \frac{1}{P(t)} \sum_{\omega=0}^{N} \omega^m P(t, \omega), \qquad (1)$$

where m is the order of the moment and $P(t)$ is the marginal distribution. For $m = 1$, it is the conditional spectral mean; if $m = 2$, it is conditional spectral variance; for $m = 3$, it is the conditional spectral skewness; and for $m = 4$, it is conditional spectral kurtosis. Higher-order moments could also be calculated by choosing a higher value of m. Algorithm 1 explains the steps to calculate conditional spectral moments.

2.1.1.2 Conditional temporal moments. The conditional temporal moment of an audio signal describes how group delay is evolving in time. The marginal temporal

Algorithm 2. Conditional temporal moments.

```
1. Result: conditional temporal moments.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Find $P(t, \omega)$, the spectrogram of $x_i(t)$.
4. Choose the value of moment $n$ between 1 and 4.
5. Calculate marginal distribution $P(\omega)$ of $x_i(t)$.
6. Calculate the conditional temporal moment using Eq. (2).
```

Algorithm 3. Joint time-frequency moments.

```
1. Result: joint time-frequency moments.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Find $P(t, \omega)$, the spectrogram of $x_i(t)$.
4. Choose the value of moment $m$ and $n$ between 1 and 4.
5. Calculate the joint time-frequency moment using Eq. (3).
```

Algorithm 4. Auto-correlation of signal $x_i(t)$.

```
1. Result: auto-correlation of signal $x_i(t)$.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Consider $\tau$ as lag.
4. Calculate auto-correlation using Eq. (4) or Eq. (5).
```

Algorithm 5. Spectral correlation between sub-bands.

```
1. Result: spectral correlation.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Consider two sub-bands: $\omega_a(t)$ and $\omega_b(t)$.
4. Calculate spectral correlation using Eq. (6).
```

moments, such as variance and kurtosis, help to distinguish between the amount of textures present in the audio. A high-textural audio has a symmetric histogram of these moments compared with the low-textural audio. Just like conditional spectral moments, the conditional temporal moment is also a function of frequency, time, and marginal distribution. The β is the conditional temporal moment when ω is given, M is the number of samples/time length, and the conditional temporal moment is described in Eq. (2). Algorithm 2 explains steps to calculate conditional temporal moments.

$$[\beta^n|\omega] = \frac{1}{P(\omega)} \sum_{t=0}^{M} t^n P(t, \omega), \tag{2}$$

2.1.1.3 Joint time-frequency moments. The joint time-frequency moments are related to the conditional spectral and temporal moments. It is a function of frequency, time, and marginal distribution. The joint time-frequency moment is defined below as

$$\beta^n \alpha^m = \sum_{t=0}^{M} \sum_{\omega=0}^{N} t^n \omega^m P(t, \omega), \tag{3}$$

Here n, m represent the order of the temporal and spectral moment. For example for joint time-frequency mean order of the system should be [1,1]. Similarly for joint time-frequency variance, skewness, and kurtosis, the order should be [2, 2], [3, 3], and [4,4] respectively. Algorithm 3 explains the steps to calculate the joint time-frequency moments.

2.1.2 Temporal Correlation

In few audio textural sounds, there is a temporal pattern present within the sounds. For example frog calls, cricket sounds, etc. have some kind of temporal structures. The temporal structure helps to capture the characteristic rhythm and smoothness from an audio texture. For example it captures the difference between fast/rough clapping and smooth/slow sea-wash. These temporal structures could also be analyzed from a TFR, such as a spectrogram [20]. In the spectrogram, the correlation among time axes is called temporal correlation. This gives an idea how the signal present in a small window is related to the other signal present in some other window. An auto-correlation function R_{xx} reflects the temporal correlation and is explained in Eq. (4) for the continuous signal $x_i(t)$. Eq. (5) shows the auto-correlation for the discrete signal $x_i(n)$. The early work of McDermott [12] includes the auto-correlation function, but later in [15] this parameter was dropped. Algorithm 4 explains the steps to calculate temporal correlation by the auto-correlation function.

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x_i(t)\overline{x_i(t-\tau)}, \tag{4}$$

$$R_{xx}(l) = \sum x_i(n)\overline{x_i(n-l)}. \tag{5}$$

2.1.3 Spectral Correlation

Spectral correlation is also known as cross-band correlation. This feature helps to identify sub-bands that exhibit synchronized energy maxima (e.g., crackling fire and speech), distinct from the independent variations in each band (e.g., water sounds). By analyzing the spectral correlation between the sub-bands, the more correlated sub-bands can be identified, and this could help in classification applications [20]. The spectral correlation is an important work in [12, 13, 15, 21]. Spectral correlation is the correlation between various frequency sub-bands in the TFR. The spectral correlation is defined by Eq. (6), where ω_a and ω_b are the two sub-bands. Algorithm 5 explains the spectral correlation in detail.

$$C_{\omega_a,\omega_b} = \sum_{t=-\infty}^{\infty} \omega_a(t)\bar{\omega}_b(t). \tag{6}$$

2.1.4 Spectro-Temporal Correlation

The temporal correlation characterizes the horizontal relationship in TFR, and on the other side, the spectral correlation defines the vertical relationship in TFR [22]. There are also spectro-temporal structures present in TFR. This slant relationship involves correlations in both time and frequency. This statistical parameter can describe those fre-

Algorithm 6. Spectro-temporal correlation by cross-correlation function.

```
1. Result: spectro-temporal correlation.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Choose a delay $\tau$.
4. Calculate spectro-temporal correlation using Eq. (7).
```

Algorithm 7. Mean instantaneous frequency.

```
1. Result: mean instantaneous frequency of a signal.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Consider $\phi(t)$ as unwrapped instantaneous phase of the signal
   $x_i(t)$ at any time $t$.
4. Calculate instantaneous frequency using Eq. (8).
```

Algorithm 8. Local binary pattern (LBP) from spectrogram.

```
1. Result: LBP feature vector.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Generate spectrogram image from the audio signal, and
   convert it into its gray-scale equivalent.
4. Choose the radius of the mask and type of normalization.
5. Extract LBP features from the gray-scale spectrogram image.
```

quencies, which appear in the audio textures after a time delay (e.g., chirps and vibrating whispers). This parameter shows the correlation between various time and frequency bands. It could be understood as a spectral correlation with time delay. In TFR it is characterized by the delayed cross-correlation. The tempo-spectral correlation is shown by Eq. (7). Algorithm 6 describes the spectro-temporal correlation in detail.

$$C_{\omega_a,\omega_b}(\tau) = \sum_{t=-\infty}^{\infty} \omega_a(t)\bar{\omega_b}(t+\tau). \tag{7}$$

2.1.5 Mean Instantaneous Frequency

The instantaneous frequency is also known as time-dependent frequency. Instantaneous frequency is a key parameter to distinguish bird calls, since most bird calls have ascending instantaneous frequency. The instantaneous frequency is the first derivative of the instantaneous phase and is calculated from each frequency sub-band generated by a TFR. Calculating the mean of that instantaneous frequency could help in classifying various audio textures. The instantaneous frequency is defined as

$$\omega(t) = \frac{d\phi(t)}{dt}, \tag{8}$$

where $\phi(t)$ is an unwrapped instantaneous phase angle.

2.1.6 Mean Crossing Rate

Similar to zero crossing rate (ZCR), the mean crossing rate (MCR) feature estimates the alternations of successive feature values inside a texture window [11]. The mean crossing rate is defined as

$$MCR = \frac{1}{N-1} \sum_{m=k-N+1}^{k-1} \frac{h_i[m] - h_i[m-1]}{2}, \tag{9}$$

where

$$h_i[m] = sgn(x_i[m] - \overline{x_i[m]}). \tag{10}$$

2.2 Image-Based Features

Image-based audio textural features are actually borrowed from the image-processing techniques. The audio signals are converted into an appropriate TFR, such as a spectrogram, and this TFR is considered an audio's image. The most used image-based features are local binary patterns, local ternary patterns, Histogram of Oriented gradients, and Haralick's features.

2.2.1 Local binary patterns (LBPs):

LBP is one of the most explored image texture features [23]. The LBP is primarily used in applications such as face recognition, face detection, and object detection. Now when audio TFRs are considered as an image, this LBP also becomes a part of an audio textural feature set [24]. In [25] LBP is used to classify sound effects by capturing the textural information from a spectrogram. In this work, authors also validate that a logarithm of the Gammatone-like spectrogram provides richer texture information than other spectrograms. The LBP measures the local spatial information and gray-scale contrast from the TFRs. The LBP tries to find the local uniform patterns present in an audio texture TFR. LBP performs thresholding and converts pixels into binary units 0 and 1. The thresholding is based on the Eq. (11).

$$\begin{cases} 1, p > c \\ 0, p < c \end{cases}, \tag{11}$$

Here c is the center pixel, and p is the neighboring pixel. It converts the pixel intensities present in a circular neighborhood into a binary pattern. The LBP is a 59-dimensional feature that contains 58 uniform patterns and one non-uniform pattern present in the audio TFR. There are several modified and improved versions of LBP present that are employed in image-processing applications [26]. The LBP is used in audio scene classification, snore analysis, emotion detection [27, 28], and analysis of pathological speech [29]. Algorithm 8 describes the extraction of LBP from the spectrogram images.

2.2.2 Local ternary patterns (LTPs):

LTP is a modified version of LBP, where the thresholding results in the three levels (1, 0, and –1) rather than two levels, which are 0 and 1 in the case of LBP. The thresholding in the case of LTP is defined in Eq. (12) below:

$$\begin{cases} 1, p > c + k \\ 0, p > c - k, \quad p < c + k \\ -1, p < c - k \end{cases}, \tag{12}$$

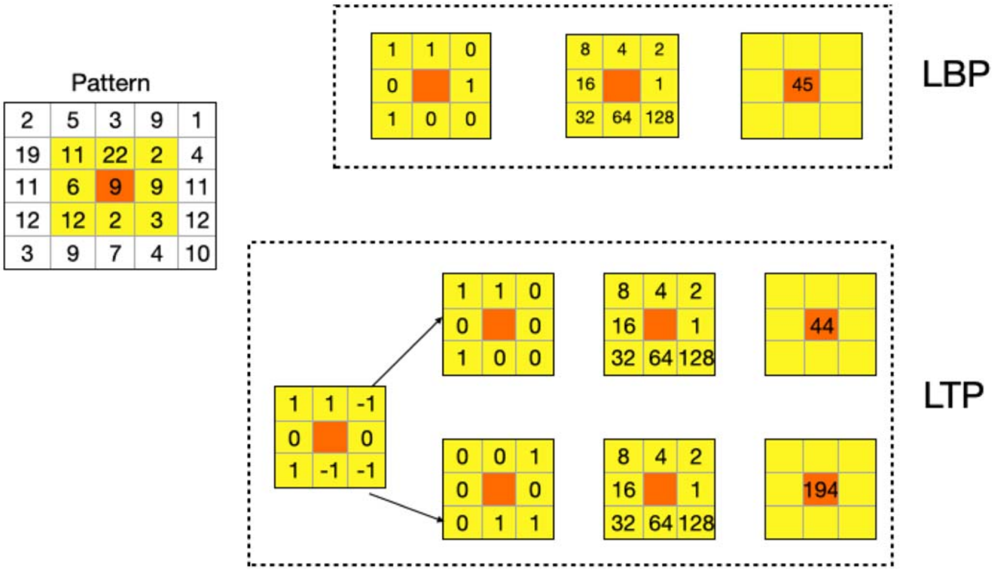

Fig. 3. Local binary patterns (LBPs) and local ternary patterns (LTPs).

Algorithm 9. Local ternary pattern (LTP) extraction.

```
1. Result: local ternary patterns.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Generate spectrogram from the audio signal $x_i(t)$, and convert it into its gray-scale equivalent image.
4. Choose the size of image mask and find the upper and lower bits using ternary function and threshold values –1, 0, and 1.
5. Calculate upper and lower values using upper and lower bits.
6. Construct upper and lower signals.
7. Generate histogram from these signals.
8. Join the histograms to get LTPs.
```

Fig. 3 shows the difference between LBP and LTP. The steps to calculate LTP are explained in Algorithm 9.

2.2.3 Histogram of oriented gradients (HOGs):

HOG is a feature descriptor initially explored for detection in the human body [30]. The HOG has been widely used in image-processing–based medical applications [31]. It describes the local appearance and shape of an object using the distribution of intensity gradients or edge directions. The extraction process of HOG is as follows. First 1D derivative masks (i.e., $[1, 0, 1]$ and $[1, 0, 1]^T$) are applied to each pixel value in both the horizontal and vertical direction respectively. Then the orientation and magnitude of a gradient are computed with both of the gradients. The orientations are normalized to several bins, which are equally spaced in the range of 0 to 360 degrees. Finally the HOG is obtained by summing the magnitude of the orientations over the whole range of the spectrogram image. HOG is widely used in audio scene classification [32–34], vehicle detection [35], and emotion detection [36].

2.2.4 Haralick's features:

Algorithm 10. Histogram of oriented gradient (HOG) feature extraction.

```
1. Result: HOG feature set.
2. Initialization: mono-channel audio signal $x_i(t)$.
3. Generate the spectrogram image of the audio signal, and convert it into its equivalent gray-scale image
4. Chose name-value parameters, such as cell and block size.
5. Extract HOG features from each cell of the gray-scale spectrogram.
```

Algorithm 11. Haralick's feature extraction.

```
1: Result: 14-dimensional Haralick's features.
2: Initialization: mono-channel audio signal $x_i(t)$.
3: Generate spectrogram of the audio signal.
4: Convert RGB spectrogram into its gray-scale equivalent image.
5: Calculate gray-level co-occurrence matrix from the gray-scale spectrogram image.
6: Calculate basic 14 statistical features from the matrix.
```

Haralick's features are a set of 14 features that describe the correlation between the intensity of a pixel to the adjacent pixel in the space. These 14 features are derived from the co-occurrence matrix. They describe the texture of an image. The 14 features include angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measure of correlation 1 and 2, and maximum correlation coefficient [29].

2.3 Timbre-Based Features

Timbre refers to the texture, character, and color of a sound that defines it. Timbre is actually the perceived sound quality of a musical note or sound [37]. The timbral texture features are based on the standard features proposed for music-speech discrimination and speech recognition [8,

Algorithm 12. Spectral centroid of an audio signal.

```
1. Result: spectral centroid $\mu_1$.
2. Initialization: mono-channel audio signal $x_i(t)$ and sampling frequency $f_s$.
3. Convert the signal into frequency domain.
4. Compute spectral centroid using Eq. (13).
```

Algorithm 13. Spectral roll-off point of an audio signal.

```
1. Result: spectral roll-off.
2. Initialization: audio signal $x_i(t)$ and sampling frequency $f_s$.
3. Transform the signal in frequency domain.
4. Compute spectral centroid using Eq. (15).
```

Algorithm 14. Spectral flatness value of an audio signal.

```
1. Result: spectral flatness value.
2. Initialization: Mono-channel audio signal $x_i(t)$.
3. Calculate the periodogram power spectral density of the audio signal.
4. Calculate the ratio of the geometric to arithmetic mean of the periodogram signal.
```

Algorithm 15. Short-time energy of an audio signal [8].

```
1. Result: short-time energy of the audio signal.
2. Initialization: mono-channel audio signal $x_i(t)$ and window type, amplitude, and length.
3. Calculate $xnew = x^2$.
4. Find $STE = xnew \bigotimes win$ (convolution of window and signal square).
```

38–40]. A representative set of timbre-based audio texture features are discussed below.

2.3.1 Spectral centroid:

This feature indicates where the center of mass of the spectrum is located. The spectral centroid indicates the brightness of the sound signal. The higher centroid values correspond to the "brighter" textures with more high-frequency components. It is defined as

$$\mu_1 = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k}, \tag{13}$$

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 [8, 38, 40].

2.3.2 Spectral flux:

The spectral flux is defined as 2-norm of the frame-to-frame spectral amplitude difference vector. It points out the sudden changes in the frequency-energy distribution of sounds. In simple words, spectral flux measures how quickly the power spectrum of a signal is changing. For high-texture audio signals, the value of the flux would be comparatively lower than that for low-texture signals, such as music. The spectral flux is defined by Eq. (12) as

$$Flux = \sum_{k=b_1}^{b_2} (N_k - N_{k-1})^2, \tag{14}$$

where N is the normalized magnitude of the Fourier transform for the bin k and $k - 1$.

2.3.3 Spectral roll-off:

The spectral roll-off point is the frequency below which 95% of the signal's energy is contained. The roll-off point is i if Eq. (15) holds true [8, 38].

$$\sum_{k=b_1}^{i} s_k = 0.95 \times \sum_{k=b_1}^{b_2} s_k, \tag{15}$$

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

2.3.4 Spectral flatness:

Spectral flatness indicates the uniformity in the frequency distribution of the power spectrum. Mathematically spectral flux is the ratio of the geometric to arithmetic mean of the periodogram signal [8, 38, 40]. The low-textural audio signals, such as harmonic sounds, have spectral flatness close to zero, and high-texture noise, like signals, have spectral flatness close to one. Steps to calculate spectral flatness are explained in Algorithm 14.

2.3.5 Short-time energy:

The energy throughout an audio signal is variable, and hence it is not feasible to predict a value. For this, the short-time energy, which is energy from a frame, is calculated [8, 38–40]. High-texture audio signals have uniform short term energy, while less-textural audio signals show high variation in short-time energy values. Algorithm 15 describes the steps to calculate short-time energy of an audio signal.

2.3.6 Zero crossing rate:

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 [8, 38, 40]. The ZCR for ith frame with the length N is defined as

$$Z(i) = \frac{1}{2N} \sum_{n=1}^{N} |sgn[x_i(n)] - sgn[x_i(n-1)]|, \tag{16}$$

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

$$sgn[x_i(N)] = \begin{cases} 1, & x_i(n) \geq 0 \\ 0, & x_i(n) < 0 \end{cases}. \tag{17}$$

Algorithm 16. Mel frequency cepstral coefficient (MFCC) features from an audio signal [8].

```
1. Result: mel frequency cepstral coefficients.
2. Initialization: mono-channel audio signal $x_i(t)$.
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 the power spectrum, and sum the energy in each filter.
6. Take the logarithm of the filter-bank energies.
7. Take discrete cosine transform (DCT) of the log filter-bank energies.
8. Keep 2–13 DCT coefficients, and discard the rest.
```

2.3.7 Mel frequency cepstral coefficients (MFCCs):

MFCC represents the short-time power spectrum of an audio clip based on the discrete cosine transform of log power spectrum on a nonlinear mel scale. In MFCCs the frequency bands are equally spaced on a mel scale, which mimics the human auditory system. The MFCC is one of the most popular audio features that are used in many applications such as speech recognition, speech enhancement, music genre classification, vowel detection, mood detection, and many more [8, 13, 41, 40]. The first and second derivative of MFCCs, delta MFCC and delta-delta MFCC, have been used as textural features in various applications [13].

Table 2 summarizes the textural features used in analysis of audio textures in various applications. The three main categories are statistical, image-based, and selected timbre-based textural features.

2.4 Deep Features for Texture Analysis

The deep learning algorithms have gained popularity in many audio applications [8]. Audio texture analysis and synthesis is one of those applications. The lower layer in a deep network captures the style and melody-based features from an audio signal, while the higher layers in the network capture the overall style of the audio sample. The features passed from one layer to another are called as deep features. In [42] and [43] deep features are extracted and analyzed from a convolution neural network (CNN).

In [44] timbre-based and rhythm-based features are extracted from two blocks of CNN structure. A three-stage deep model is designed to perform audio style transfer. A textural audio is synthesized by analyzing the deep features generated by this three-stage deep model. In [45] the authors have analyzed statistical features extracted from McDermott's model. In this work, the model is considered a three-layer hand-crafted neural network where the first layer has 30 band-pass filters to decompose audio signals into frequency bands. The second layer captures the envelope features from each frequency band and applies a non-linearity to it. The third layer decomposes the envelope using 20 band-pass modulation filters. This model extracts the statistical features up to fourth order.

Another popular deep learning technique in audio texture analysis is recurrent neural network (RNN) [11, 46]. In [47] a three-tier RNN is used. This network extracts features, such as MFCCs, spectral centroid, spectral flatness, and pitch, and employs these features to synthesize other audio textural signals. In recent works, long short-term memory (LSTM) is used for audio texture analysis in many applications, such as event detection [48], music generation, texture synthesis, music genre classification, and more. In a few of the works, LSTM is combined with other algorithms to generate a hybrid model, such as LSTN-CNN and Bi-LSTM [48]. In [48] the authors demonstrated that the texture analysis using the LSTM-CNN approach has outperformed other customized classifiers, such as support vector machines (SVM), K-nearest neighbors (KNN) decision tree, and stand-alone LSTM. The hybrid model LSTM-CNN has also been used for speech emotion detection [49, 50]. In [51] the authors analyzed deep textural features using LSTM. In that work, the textural features, such as ZCR and MFCC, are fed into the LSTM model to generate deep features. These deep features are used for music genre classification.

3 AUDIO TEXTURE SYNTHESIS

Most of the work done in audio textures is in the field of analysis/re-synthesis. In this the synthetic audio textures are augmented by analyzing the properties of the original audio texture. The most common ways to synthesize audio textures are by using granular-based, physical model–based, or deep learning approaches. In this section, all the approaches are explained in detail. Fig. 4 shows the evolution of audio texture synthesis methods. In [52] the summary of various audio texture synthesis algorithms is explained.

3.1 Granular-Based Synthesis

Granular-based audio texture synthesis is based on the process called "granulation" [53–55]. For the first time the granular-based audio synthesis approach were used in music synthesis in [56]. In this process the audio texture is divided into little slices called "grains." The slicing of audio signal is quite similar to the sampling of the signal. As per the initial assumptions, the size of the grain could be anywhere between 1 and 100 ms [57], but in a few practical applications the size of the grain could be as small as 20 ms [58]. These grains are played, synthesized, or overlapped to generate new audio textures. The order of the grains in the synthetic audio is user dependent. Depending on the size of the grain and number of grains used for synthesis, the perceptual quality of the audio could be controlled [59]. Presently granular-based synthesis is highly used at the commercial level to generate synthetic audio signals/textures.

Using this granular-based synthesis, audio textures have been synthesized for various applications. In [58] for the first time, audio textures were synthesized using wavelet tree learning. That work was based on the fact that techniques such as wavelets and short-time Fourier transform

Table 2. Summary of audio textural features.

Audio textural feature	Feature name
Statistical features	1. Conditional moments a) Spectral moments b) Temporal conditional moments c) Joint temporal-spectral moments 2. Temporal correlations 3. Spectral correlations 4. Spectro-temporal correlations 5. Mean instantaneous frequency
Image-based textural features	1. Local binary patterns 2. Local ternary patterns 3. Histogram of oriented gradients 4. Haralick's features
Representative set of timbre-based features	1. Spectral centroid 2. Spectral flux 3. Spectral roll-off 4. Spectral flatness 5. Short-time energy 6. Zero crossing rate 7. Mean crossing rate 8. Mel frequency cepstral coefficients

provide the local representation or grains of an audio texture signal [55, 60]. This method is equally good for periodic and stochastic audio textures.

In [61] and [62] the authors synthesized audio textures by using a similarity index between frame-based MFCC audio features. The grain/frame size chosen here was 32 ms. Extending their current work, the authors proposed a very interesting application of audio texture synthesis in which the missing part of an audio is determined by analyzing the audio textures [2]. That work was based on the concept of self-similarity. Also the concept of constrained texture synthesis was explored for the first time, where the missing part should be perceptually smooth at the joint points with the original audio clip.

Another development in granular synthesis comes into picture when time and frequency frames are considered as the grains. In [63] the authors proposed a cascade time-frequency linear prediction model, where the linear prediction in time and frequency are cascaded to capture the spectral and temporal information respectively. During synthesis of audio textures, white noise was chosen as a starting seed. The residual parameters extracted from the analysis process were used to design the filter coefficients to synthesize the audio texture. There are two major restrictions on this work, which are first, while analyzing the textures, the representation of the audio signal is compressed in a lossy way, and second, the audio clips of arbitrary length cannot be generated by this synthesis model.

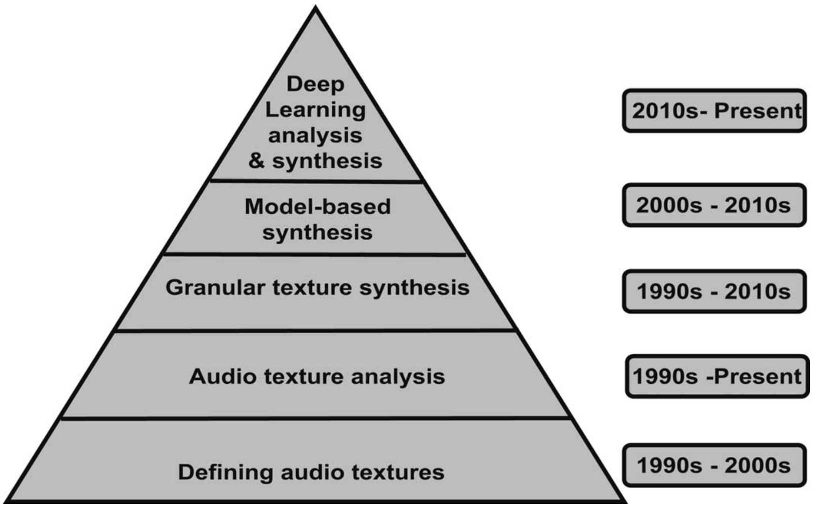

Fig. 4. Evolution of audio texture's understanding, analysis, and synthesis.

The work in [63] was expanded by Zhu and Wyse in [64]. The analysis and synthesis of audio textures was done by extracting the linear prediction coefficients (LPC) and residues from the time-frequency domain of the signal. By using LPC filters in both time and frequency domain, the perceptual quality of the sound textures can be preserved. Another advantage of this method is that the generated audio textures could be of arbitrary length.

Later after 2002 physics-based granular synthesis methods were employed to generate texture-like sounds. In [65] the author introduced physically informed stochastic event modeling for generation of texture-like sounds, such as ice cubes in a glass or rain drops. In [66] the author also developed an analysis/synthesis system for walking sounds. Similar experiments were conducted to generate wind chime sounds [67], but these physics-based methods were not based on audio signal processing techniques. In [68] another method called overlap-add granular synthesis is proposed. This is comparatively easy to implement granular synthesis for audio textures. The blocks/grains are extracted from the original stochastic source and then these blocks are overlapped and added to synthesize a new texture audio. Fig. 5 shows the process of creating new textures by adding the overlapped blocks.

In 2010 Schwarz et al. proposed a corpus-based cognitive synthesis method that uses the audio descriptors such as pitch, loudness, and brilliance of an audio signal to synthesize the audio textures of various intensities [69]. In this method, authors have synthesized audio textures for light rain, medium rain, and heavy rainfall. In 2011 the corpus-based granular synthesis method was implemented in interactive audio-graphic 3D scenes [70]. Later in 2013 Schwarz experimented the audio texture synthesis by using 1D continuous manual annotation of environmental recording.

Another method of synthesizing audio texture is based upon a two-level montage approach where the low-level and high-level details are preserved. This approach has been used to identify events in the audio textures and find the missing part of an audio clip [71, 72]. Schwarz et al. investigated three different approaches for sound texture synthesis: concatenated synthesis with descriptor controls, montage synthesis, and a new method called “Audio Texture.” The timbre-based audio descriptors such as Loudness, Fundamental Frequency, Noisiness, Spectral Centroid, Spectral Spread, and Spectral Slope are extracted and used [73, 38].

Also some work has been done in synthesizing the musical audio textures. The musically expressive audio textures have been generated by using generalized audio. The generalized audio is produced by using a statistical decorrelation technique, such as principle component analysis and dynamic time warping algorithm [74, 75]. In these methods, a single musical texture is generated by mixing many musical textures captured from various songs and radio stations.

An interesting approach is given by Zheng et al. in [76]. In this work, the authors have synthesized an audio scene texture by combining more than one audio texture. A decorative sound texture for soundscapes has been generated by this method. A rainy day on a street is generated by combining three sound textures: rain dropping on the ground, rain dropping on umbrellas, and birds chirping nearby [76].

3.2 Model-Based Synthesis

Most of the work about texture sound synthesis is done by implementing model-based synthesis methods. The model-based methods are based on the physiological models of human hearing systems, statistical models, and improved time-frequency representation models. The model-based synthesis methods are explored mostly after the year 2000 and includes many variations. In 2002 the environmental sound textures were generated by iterating non-linear functions based on the perceptual modeling arising from the system dynamics [77, 78]. Later in 2003 a spectral and statistical model was proposed to synthesize time-scale modified noise like audio signals. This method uses the standard synthesis-by-analysis approach. The main advantage of this method is that it permits high-quality synthesis of noisy signals and allows infinite time-scaling without degrading the original sound [79].

A data-driven framework for analyzing, transforming, and synthesizing sound textures is proposed in [80]. In this work, the background audio scene is analyzed to find the audio textures present in this, and then transformation of these audio textures is used to synthesize new audio textures for generating audio scenes. In 2009, Josh McDermott proposed a model for audio texture sound synthesis that is inspired by the image texture analysis model [21, 12]. In this work, higher order statistics and spectral correlations between sub-band signal envelopes were extracted and super imposed on Gaussian noise signal during synthesis process. The filters used in the analysis process are based on the band-pass filters that are present in the human ear. It has been concluded that imposing only the marginal statistics (variance and kurtosis) of the sub-bands was sufficient to generate synthetic examples of many audio textures, such as rain, streams, etc.

Later in 2011 the model proposed by McDermott in [12] is further modified. The new model used an equivalent rectangular bandwidth filter-bank to divide the signal into various perceptual bands. These perceptual bands were further divided into sub-bands called modulation bands. The first four statistical parameters (mean, variance, skewness, and kurtosis) were extracted from each modulation and perceptual band, and the cross-correlations between adjacent bands were also calculated. Because this model was based on the human auditory system, it soon becomes one of the most popular models for texture sound synthesis. Josh McDermott proposed a similar human auditory-based model to extract textural information from the audio signals. This method is very simple and uses MFCCs to extract statistical features [13]. It is one of the popular algorithms that is based on the model proposed by McDermott in 2009 [12]. Fig. 6 represents the proposed methodology.

Inspired by McDermott’s work, in 2013 another method of audio texture synthesis was proposed that was based on short-time Fourier transform representation of an audio texture [20]. In this method, each bin is considered a sub-

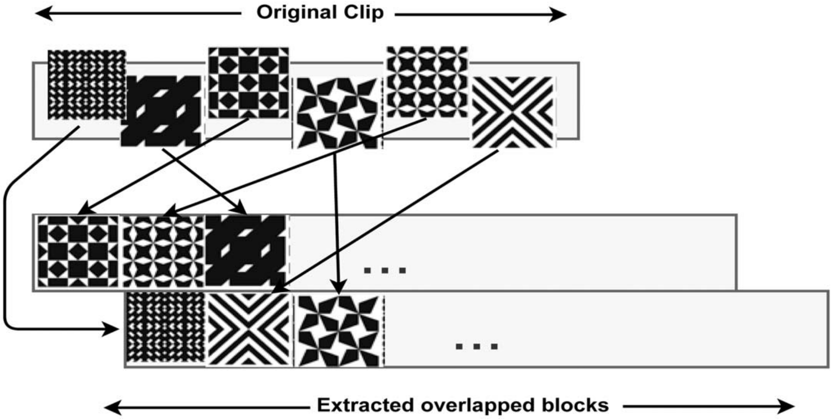

Fig. 5. Audio texture synthesis using overlap-add method.

band. The statistical parameters, such as auto-correlation cross-correlation and moments, are extracted from each bin and used in the synthesis process. Another way of modeling, classifying and recognizing sound textures, is through the empirical mode decomposition (EMD). The audio signal is first decomposed using EMD, and then those EMDs are used for texture analysis and recognition [81–83]. An ensemble-based hand clapping audio texture synthesis is done in past keeping room acoustics in mind. The synthesis process is tuned with respect to the room parameters for a small or medium-sized room [84, 85]. Another attempt to synthesize long term audio, such as airplane cabin noise and piano sounds, has been made in [86].

Within the last decade, the scattering transforms have been introduced and successfully employed in proving the state of the art results for texture synthesis, texture discrimination, and genre recognition. The scattering transforms iterate on complex wavelet transform. Scattering moments provide general representations of stationary processes computed as expected values of scattering coefficients [87–90].

3.3 Deep Learning for Texture Synthesis

The audio texture analysis and synthesis is no longer away from deep learning algorithms. The concept of audio texture analysis or synthesis using deep learning is borrowed from the image style transfer concept [91, 92]. Even if the concept of audio style transfer is borrowed from the image style transfer, there is still some differences between the two. These differences are explained by Dieleman in [93]. In image style, lower layers shows simple visual patterns, such as lines and corners, and higher layers represent the complex features, such as human or animal faces, automobiles, etc. In audio style transfer, lower layers identify local stylistic and melodic features, and higher layers represent the overall style.

In [91] Gatys et al. used the cross correlation between the feature maps of 2D CNNs as parameters to analyze and synthesize image textures. In the last couple of years, audio texture analysis has been done by using deep algorithms mostly by CNNs and RNNs [94]. In [95] audio texture has been synthesized using 2D CNNs, and the synthesis process is back-propagated until the temporal cross-correlations of the feature maps resemble those of the target textures. Authors have shown that the synthesized audio textures are better in quality than the original audio samples.

A different approach of synthesizing audio texture has been introduced by Antognini et al. in [42]. Two new terms, an autocorrelation term and diversity term, have been introduced. These two terms contribute to the loss function of the CNN. In that work, the authors showed that there is a tradeoff between diversity and quality of the audio texture. In [47] the authors implemented multi-tier conditioning RNNs that synthesized the multi-level paradigm of the constant fine structure of audio textures. The researchers used the pre-trained networks VGG-19 [96], SoundNet [97], connectionist temporal classification [98], and ImageNet for audio texture analysis and synthesis. One of the CNN architectures used is shown in Fig. 7. These techniques are used for audio style transfer and audio texture synthesis [99].

There are several initial attempts at audio style transfer [99, 100, 45, 101]. In a blog, authors have recommended the use of shallow networks instead of deep pre-trained

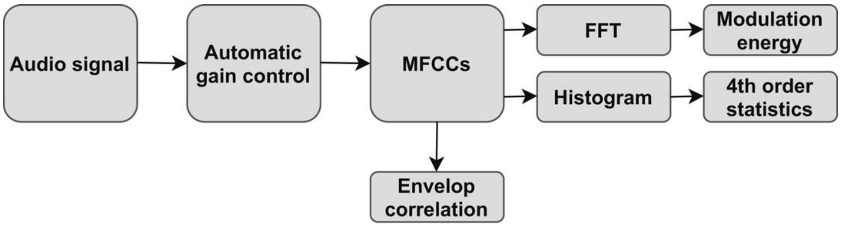

Fig. 6. Human auditory based textural features. FFT, fast Fourier transform; MFCC, mel frequency cepstral coefficient.

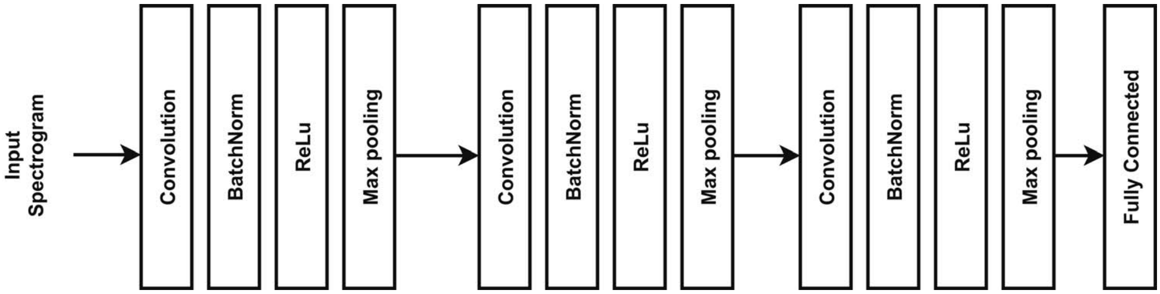

Fig. 7. Convolution neural network (CNN) architecture.

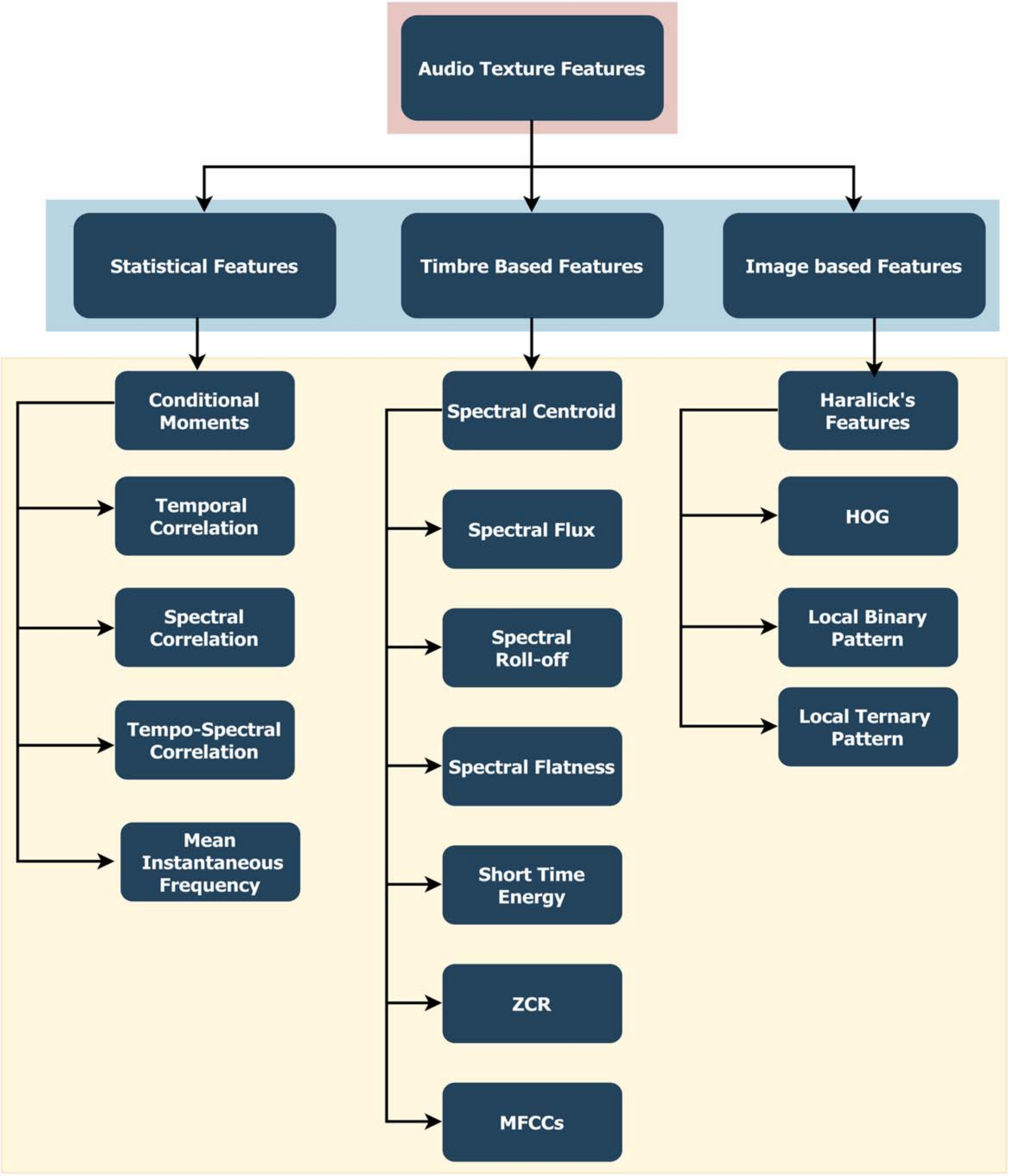

Fig. 8. Summary of audio texture features.

networks such as in the case of image processing [102]. On the other hand, few researchers keep working on the deep and pre-trained networks [103]. There was also an attempt to isolate prosodic speech using the style transfer approach, although only low-level textural properties of the voice were successfully transferred [104].

Table 3 summarizes the audio texture synthesis methods including granular-based, model-based, and deep learning–based algorithms.

4 APPLICATIONS OF AUDIO TEXTURES

The audio textures have been explored by researchers mainly in the applications related to audio scene analysis, speech and music, biomedical signals, creative arts, and recommendation systems. These applications are explained below in detail.

- **Audio scene analysis (ASA)**: The ASA describes the application area where the audio signals are used to identify the environment where they are produced. ASA is required in many applications, such as information indexing and retrieval in multimedia databases, video editing, and more. Audio textures and its associated features, such as LBP and HOG, are used in ASA [106]. In many application MFCCs have been used for ASA [107]. Recently deep learning algorithms have taken the TFR of the textural audio as an image and performed various classification tasks [32, 33, 108, 24, 109, 110]. The most popular ASA database available is Detection and Classification of Acoustic Scenes and Events, and recently a new database SARdb has been released [111].
- **Speech and music**: In the case of speech and music, audio textures have been vastly explored for emotion detection via speech and music genre classification. In [112] the authors detected applause sound textures in music by using MFCCs and low-level descriptors. Emotion estimation from an audio/speech signal is not new, but very few attempts have been made to detect emotions using textural features [27, 36]. Another speech-based application is the analysis of pathological speech. Recent studies have shown that the most-used features for the screening of pathological speech signals are acoustic features, such as jitter, skewness, kurtosis, peak frequency, MFCCs, and linear prediction coefficients [8]. Recently few attempts have been made to analyze and screen pathological speech signals using texture-based features [29, 113]. Another popular application is audio synthesis. In this, the audio clip or audio texture is synthesized by analyzing the existing audio textures. It has been highly explored by many researchers in last decade or so [14–37]. Music genre classification is a very popular application. In this, the musical clips are classified according to their genre such as classical, electronic, jazz/blues, metal/punk, rock, and more. Most of the work in this application is based on standard audio features, such as MFCCs [41], but recent studies have shown that textural features, such as LBP and HOG, could be used to classify these genres by using TFRs [13, 114, 115, 40, 116].
- **Creative arts**: Most of the work in this domain is done for synthesizing audio signals using style transfer. The audio style transfer is inspired by the concept of image style transfer. Audio style transfer is one of the most recent and popular audio texture applications, especially on social media. This application looks for how to transfer the style of the reference audio signal to a target content. The resultant sound is often the mixture of the style and content sound [99, 100, 45].
- **Industry/machine health**: Currently in the industry, a machine or gear fault diagnosis is done by analyzing the vibration sounds using image-based texture features and machine learning. In state of the art, the fault diagnosis using vibration signals is done in induction motors, helicopter machine, or oil pumping machine [117–119, 17].
- **Recommendation systems**: In this domain, the recommendations to find the missing part in the audio signal has been made. This is another application where a missing part of the audio clip is regenerated based on the audio textures present in the rest of the signal. This application is based on the concept of self-similarity. Also the concept of constrained texture synthesis is explored for the first time where the missing part should be perceptually smooth at the joint points with the original audio clip [2].

5 SUMMARY AND DISCUSSION

This review paper discusses the audio texture's basics and how audio textures are different from the speech and music signals in terms of their time domain and time-frequency representation. This review also covers the baseline work and evolution of audio textures right from the 1990s to their progress to date. This paper discusses the various synthesis algorithms, such as model-based, granular-based, and modern deep learning–based algorithms. This work also explores the types of textural features employed in various applications. The statistical, image-based, and timbre-based textural features are discussed in detail. Fig. 8 gives a bird's eye view of the various textural features. For example, the most popular audio textural image-based features are LBPs, LTPs, HOG, and Haralick's. Similarly, the timbre-based textural features are spectral flatness, spectral roll-off, spectral flux, spectral centroid, short-time energy, ZCR, and MFCCs.

In the future, a more comprehensive analysis of audio textures could be done. A detailed discussion on texture versus speech/music could also be done. A change in trends of audio textures and its applications is expected when more machine learning and deep learning algorithms are explored.

Table 3. Summary of audio texture synthesis methods.

Synthesis method	Year	Citation	Algorithm
Granular-based synthesis	1988, 1992	[54, 53]	Define grains
	1988	[56]	Music synthesis
	2001	[55]	Natural grains
	2002, 2003	[61, 62]	MFCC
	2002	[58]	Wavelet tree learning
	2003	[63]	Cascade time-frequency LPC
	2004	[57]	Grain size is defined
	2004	[64]	LPC
	2002, 2002, 2004	[65–67]	Physics-based synthesis
	2003	[75]	Karhunen-Loève transform
	2009	[68]	Overlap-add granular method
	2010, 2011	[69, 70]	Corpus-based cognitive synthesis
	2013	[105]	Manual annotation–based
	2014, 2016	[71, 72]	Montage approach
	2015, 2016	[38, 73]	Audio descriptor–based
	2020	[76]	Concatenation of textures
Model-based synthesis	1999, 2002	[77, 78]	Perceptual modeling
	2003	[79]	Synthesis-by-analysis
	2006	[80]	Data-driven framework
	2009	[12]	Statistical model
	2011	[13]	MFCC-based
	2010, 2012	[81, 82]	Empirical mode decomposition
	2013	[20]	STFT-based
	2012, 2012, 2013, 2014	[87–90]	Scattering transform
	2009, 2020	[84, 85]	Ensemble-based approach
Deep learning–based synthesis	2018	[86]	Long-term textures
	2014, 2015, 2016	[96, 91, 92]	Image-based methods
	2014	[93]	Audio textures for music
	2016	[97]	SoundNet
	2016, 2017, 2017, 2018	[102, 100, 104, 45]	Audio style transfer
	2018	[42]	Autocorrelation, diversity in CNN
	2019, 2020, 2021	[95, 47, 94, 101]	2D CNN and RNN
	2020	[99]	Visual style texture synthesis

Note: CNN, convolution neural network; LPC, linear prediction coefficients; MFCC, mel frequency cepstral coefficient; RNN, recurrent neural network; STFT, short-time Fourier transform.

6 REFERENCES

[1] N. Saint-Arnaud, *Classification of Sound Textures*, Master's thesis, Massachusetts Institute of Technology, Cambridge, MA (1995 Sep.).

[2] L. Lu, L. Wenyin, and H. J. Zhang, "Audio Textures: Theory and Applications," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 156–167 (2004 Mar.). https://doi.org/10.1109/TSA.2003.819947.

[3] S. V. Shushardzhan and S. V. Petoukhov, "Engineering in the Scientific Music Therapy and Acoustic Biotechnologies," in Z. Hu, S. Petoukhov, and M. He (Eds.), *Advances in Artificial Systems for Medicine and Education III*, Advances in Intelligent Systems and Computing, vol. 1126, pp. 273–282 (Springer, Cham, Switzerland, 2019).

[4] B. Julesz, "Visual Pattern Discrimination," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 84–92 (1962 Feb.). https://doi.org/10.1109/TIT.1962.1057698.

[5] T. Caelli and B. Julesz, "On Perceptual Analyzers Underlying Visual Texture Discrimination: Part I," *Biol. Cybernetics*, vol. 28, no. 3, pp. 167–175 (1978 Sep.). https://doi.org/10.1007/BF00337138.

[6] A. Humeau-Heurtier, "Texture Feature Extraction Methods: A Survey," *IEEE Access*, vol. 7, pp. 8975–9000 (2019 Jan.). https://doi.org/10.1109/ACCESS.2018.2890743.

[7] D. F. Rosenthal and H. G. Okuno (Eds.), *Computational Auditory Scene Analysis* (CRC press, Boca Raton, FL, 1998).

[8] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in Audio Signal Feature Extraction Methods," *Appl. Acoust.*, vol. 158, paper 107020 (2020 Jan.). https://doi.org/10.1016/j.apacoust.2019.107020.

[9] V. Bountourakis, L. Vrysis, K. Konstantoudakis, and N. Vryzas, "An Enhanced Temporal Feature Integration Method for Environmental Sound Recognition," *Acoustics*, vol. 1, no. 2, pp. 410–422 (2019 May). https://doi.org/10.3390/acoustics1020023.

[10] L. Vrysis, N. Tsipas, C. Dimoulas, and G. Papanikolaou, "Extending Temporal Feature Integration for Semantic Audio Analysis," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), paper 9808.

[11] L. Vrysis, L. Hadjileontiadis, I. Thoidis, C. Dimoulas, and G. Papanikolaou, "Enhanced Temporal Feature Integration in Audio Semantics via Alpha-Stable Model-

ing," *J. Audio Eng. Soc.*, vol. 69, no. 4, pp. 227–237 (2021 Apr.). https://doi.org/10.17743/jaes.2021.0001.

[12] J. H. McDermott, A. J. Oxenham, and E. P. Simoncelli, "Sound Texture Synthesis via Filter Statistics," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 297–300 (New Paltz, NY) (2009 Oct.). https://doi.org/10.1109/ASPAA.2009.5346467.

[13] D. P. W. Ellis, X. Zeng, and J. H. McDermott, "Classifying Soundtracks With Audio Texture Features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5880–5883 (Prague, Czech Republic) (2011 May). https://doi.org/10.1109/ICASSP.2011.5947699.

[14] A. P. Mishra, N. S. Harper, and J. W. H. Schnupp, "Exploring the Distribution of Statistical Feature Parameters for Natural Sound Textures," *PLoS One*, vol. 16, no. 6, paper e0238960 (2021 Jun.). https://doi.org/10.1371/journal.pone.0238960.

[15] J. H. McDermott and E. P. Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence From Sound Synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940 (2011 Sep.). https://doi.org/10.1016/j.neuron.2011.06.032.

[16] K. L. Davidson and P. J. Loughlin, "Instantaneous Spectral Moments," *J. Franklin Inst.*, vol. 337, no. 4, pp. 421–436 (2000 Jul.). https://doi.org/10.1016/S0016-0032(00)00034-X.

[17] P. Loughlin, F. Cakrak, and L. Cohen, "Conditional Moments Analysis of Transients With Application to Helicopter Fault Data," *Mech. Syst. Signal Process.*, vol. 14, no. 4, pp. 511–522 (2000 Jul.). https://doi.org/10.1006/mssp.1999.1287.

[18] I. Yesilyurt, "The Application of the Conditional Moments Analysis to Gearbox Fault Detection—A Comparative Study Using the Spectrogram and Scalogram," *NDT E Int.*, vol. 37, no. 4, pp. 309–320 (2004 Jun.). https://doi.org/10.1016/j.ndteint.2003.10.005.

[19] S. Ghofrani, D. C. McLernon, and A. Ayatollahi, "On Conditional Spectral Moments of Gaussian and Damped Sinusoidal Atoms in Adaptive Signal Decomposition," *Signal Process.*, vol. 85, no. 10, pp. 1984–1992 (2005 Oct.). https://doi.org/10.1016/j.sigpro.2005.04.005.

[20] W.-H. Liao, A. Roebel, and A. Su, "On the Modeling of Sound Textures Based on the STFT Representation," in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx)*, p. 33 (Maynooth, Ireland) (2013 Sep.).

[21] J. Portilla and E. P. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 49–70 (2000 Oct.). https://doi.org/10.1023/A:1026553619983.

[22] W. H. Liao, *Modelling and Transformation of Sound Textures and Environmental Sounds*, Ph.D. thesis, Université Pierre et Marie Curie, Paris, France (2015 Jul.).

[23] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification With Local Binary Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987 (2002 Jul.). https://doi.org/10.1109/TPAMI.2002.1017623.

[24] O. K. Toffa and M. Mignotte, "Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration," *IEEE Trans. Multimedia*, vol. 23, pp. 3978–3985 (2021 Nov.). https://doi.org/10.1109/TMM.2020.3035275.

[25] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann, "Sound-Event Classification Using Robust Texture Features for Robot Hearing," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 447–458 (2017 Mar.). https://doi.org/10.1109/TMM.2016.2618218.

[26] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local Binary Patterns and Its Application to Facial Image Analysis: A Survey," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 41, no. 6, pp. 765–781 (2011 Nov.). https://doi.org/10.1109/TSMCC.2011.2118750.

[27] Y. Ü. Sönmez and A. Varol, "A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns," *IEEE Access*, vol. 8, pp. 190784–190796 (2020 Oct.). https://doi.org/10.1109/ACCESS.2020.3031763.

[28] C.-Y. Fang, C.-W. Ma, M.-L. Chiang, and S.-W. Chen, "An Infant Emotion Recognition System Using Visual and Audio Information," in *Proceedings of the IEEE 4th International Conference on Industrial Engineering and Applications (ICIEA)*, pp. 284–291 (Nagoya, Japan) (2017 Apr.). https://doi.org/10.1109/IEA.2017.7939223.

[29] G. Sharma, D. Prasad, K. Umapathy, and S. Krishnan, "Screening and Analysis of Specific Language Impairment in Young Children by Analyzing the Textures of Speech Signal," in *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 964–967 (Online) (2020 Jul.). https://doi.org/10.1109/EMBC44109.2020.9176056.

[30] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893 (San Diego, CA) (2005 Jun.). https://doi.org/10.1109/CVPR.2005.177.

[31] A. Bhargava, P. Gairola, G. Vyas, and A. Bhan, "Computer Aided Diagnosis of Cervical Cancer Using HOG Features and Multi Classifiers," in R. Singh, S. Choudhury, and A. Gehlot (Eds.), *Intelligent Communication, Control and Devices*, Advances in Intelligent Systems and Computing, vol. 624, pp. 1491–1502 (Springer, Singapore, Singapore, 2018).

[32] H. Lim, M. J. Kim, and H. Kim, "Robust Sound Event Classification Using LBP-HOG Based Bag-of-Audio-Words Feature Representation," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pp. 3325–3329 (Dresden, Germany) (2015 Sep.).

[33] A. Rakotomamonjy and G. Gasso, "Histogram of Gradients of Time–Frequency Representations for Audio Scene Classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 1, pp. 142–153 (2015 Jan.). https://doi.org/10.1109/TASLP.2014.2375575.

[34] V. Bisot, S. Essid, and G. Richard, "HOG and Subband Power Distribution Image Features for Acoustic Scene Classification," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, pp. 719–723 (Nice, France) (2015 Sep.). https://doi.org/10.1109/EUSIPCO.2015.7362477.

[35] T. Wang and Z. Zhu, "Multimodal and Multi-Task Audio-Visual Vehicle Detection and Classification," in *Proceedings of the IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance*, pp. 440–446 (Beijing, China) (2012 Sep.). https://doi.org/10.1109/AVSS.2012.47.

[36] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 495–499 (San Francisco, CA) (2016 Sep.). https://doi.org/10.21437/Interspeech.2016-1124.

[37] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302 (2002 Jul.). https://doi.org/10.1109/TSA.2002.800560.

[38] D. Schwarz and S. O'Leary, "Smooth Granular Sound Texture Synthesis by Control of Timbral Similarity," in *Proceedings of the 12th International Conference on Sound and Music Computing (SMC)*, p. 6 (Maynooth, Ireland) (2015 Jul.).

[39] D. Schwarz and B. Caramiaux, "Interactive Sound Texture Synthesis Through Semi-Automatic User Annotations," in M. Aramaki, O. Derrien, R. Kronland-Martinet, and S. Ystad (Eds.), *Sound, Music, and Motion: 10th International Symposium, CMMR 2013, Marseille, France, October 15–18, 2013. Revised Selected Papers*, Lecture Notes in Computer Science, pp. 372–392 (Springer, Cham, Switzerland, 2014).

[40] B. K. Baniya, D. Ghimire, and J. Lee, "Automatic Music Genre Classification Using Timbral Texture and Rhythmic Content Features," in *Proceedings of the 17th International Conference on Advanced Communication Technology (ICACT)*, pp. 434–443 (PyeongChang, South Korea) (2015 Jul.). https://doi.org/10.1109/ICACT.2015.7224907.

[41] G. Vyas and M. K. Dutta, "Automatic Mood Detection of Indian Music Using MFCCs and K-means Algorithm," in *Proceedings of the 7th International Conference on Contemporary Computing (IC3)*, pp. 117–122 (Noida, India) (2014 Aug.). https://doi.org/10.1109/IC3.2014.6897159.

[42] J. Antognini, M. Hoffman, and R. J. Weiss, "Synthesizing Diverse, High-Quality Audio Textures," *arXiv preprint arXiv:1806.08002v1* (2018).

[43] J. M. Antognini, M. Hoffman, and R. J. Weiss, "Audio Texture Synthesis With Random Neural Networks: Improving Diversity and Quality," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3587–3591 (Brighton, UK) (2019 May). https://doi.org/10.1109/ICASSP.2019.8682598.

[44] M. Tomczak, C. Southall, and J. Hockman, "Audio Style Transfer With Rhythmic Constraints," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx)*, pp. 45–50 (Aveiro, Portugal) (2018 Sep.).

[45] E. Grinstein, N. Q. K. Duong, A. Ozerov, and P. Pérez, "Audio Style Transfer," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 586–590 (Calgary, Canada) (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8461711.

[46] L. Wyse and M. Huzaifah, "Audio Textures in Terms of Generative Models," in *Proceedings of the 13th International Workshop on Machine Learning and Music*, pp. 36–40 (Online) (2020 Sep.).

[47] M. Huzaifah and L. Wyse, "MTCRNN: A Multi-Scale RNN for Directed Audio Texture Synthesis," *arXiv preprint arXiv:2011.12596v1* (2020).

[48] S. Pandya and H. Ghayvat, "Ambient Acoustic Event Assistive Framework for Identification, Detection, and Recognition of Unknown Acoustic Events of a Residence," *Adv. Eng. Inform.*, vol. 47, paper 101238 (2021 Jan.). https://doi.org/10.1016/j.aei.2020.101238.

[49] T. Swain, U. Anand, Y. Aryan, et al., "Performance Comparison of LSTM Models for SER," in S. K. Sabut, A. K. Ray, B. Pati, and U. R. Acharya (Eds.), *Proceedings of International Conference on Communication, Circuits, and Systems*, Lecture Notes in Electrical Engineering, pp. 427–433 (Springer, Singapore, Singapore, 2021).

[50] O. Atila and A. Şengür, "Attention Guided 3D CNN-LSTM Model for Accurate Speech Based Emotion Recognition," *Appl. Acoust.*, vol. 182, paper 108260 (2021 Nov.). https://doi.org/10.1016/j.apacoust.2021.108260.

[51] Y. Yi, X. Zhu, Y. Yue, and W. Wang, "Music Genre Classification With LSTM Based on Time and Frequency Domain Features," in *Proceedings of the IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pp. 678–682 (Chengdu, China) (2021 Apr.). https://doi.org/10.1109/ICCCS52626.2021.9449177.

[52] D. Schwarz, "State of the Art in Sound Texture Synthesis," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx)*, pp. 221–232 (Paris, France) (2011 Sep.).

[53] I. Xenakis, *Formalized Music: Thought and Mathematics in Composition* (Pendragon Press, Hillsdale, NY, 1992), 2nd ed.

[54] C. Roads, "Introduction to Granular Synthesis," *Comput. Music J.*, vol. 12, no. 2, pp. 11–13 (1988). https://doi.org/10.2307/3679937.

[55] R. Hoskinson and D. Pai, "Manipulation and Resynthesis With Natural Grains," in *Proceedings of the International Computer Music Conference (ICMC)* (Havana, Cuba) (2001 Sep.).

[56] B. Truax, "Real-Time Granular Synthesis With a Digital Signal Processor," *Comput. Music J.*, vol. 12, no. 2, pp. 14–26 (1988). https://doi.org/10.2307/3679938.

[57] C. Roads, *Microsound* (MIT press, Cambridge, MA, 2004).

[58] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Synthesizing Sound Tex-

tures Through Wavelet Tree Learning," *IEEE Comput. Graph. Appl.*, vol. 22, no. 4, pp. 38–48 (2002 Aug.). https://doi.org/10.1109/MCG.2002.1016697.

[59] T. Kastner, "The Influence of Texture and Spatial Quality on the Perceived Quality of Blindly Separated Audio Source Signals," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), paper 8299.

[60] A. Kokaram and D. O'Regan, "Wavelet Based High Resolution Sound Texture Synthesis," in *Proceedings of the AES 31st International Conference: New Directions in High Resolution Audio* (2007 Jun.), paper 17.

[61] L. Lu, S. Li, L. Wenyin, H. Zhang, and Y. Mao, "Audio Textures," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 1761–1764 (Orlando, FL) (2002 May). https://doi.org/10.1109/ICASSP.2002.5744963.

[62] L. Lu, Y. Mao, L. Wenyin, and H.-J. Zhang, "Audio Restoration by Constrained Audio Texture Synthesis," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, vol. 3, pp. 405–408 (Baltimore, MD) (2003 Jul.). https://doi.org/10.1109/ICME.2003.1221334.

[63] M. Athineos and D. P. W. Ellis, "Sound Texture Modelling With Linear Prediction in Both Time and Frequency Domains," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 648–651 (Hong Kong, China) (2003 Apr.). https://doi.org/10.1109/ICASSP.2003.1200054.

[64] X. Zhu and L. Wyse, "Sound Texture Modeling and Time-Frequency LPC," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx)*, vol. 4, 345–349 (Naples, Italy) (2004 Oct.).

[65] P. R. Cook, "FOFs, Wavelets, and Particles," in P. R. Cook, *Real Sound Synthesis for Interactive Applications*, pp. 149–168 (A K Peters/CRC Press, New York, NY 2003). https://doi.org/10.1201/b19597.

[66] P. R. Cook, "Modeling Bill's Gait: Analysis and Parametric Synthesis of Walking Sounds," in *Proceedings of the AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio* (2002 Jun.), paper 000222.

[67] T. Lukkari and V. Välimäki, "Modal Synthesis of Wind Chime Sounds With Stochastic Event Triggering," in *Proceedings of the IEEE 6th Nordic Signal Processing Symposium*, pp. 212–215 (Espoo, Finland) (2004 Jun.).

[68] M. Fröjd and A. Horner, "Sound Texture Synthesis Using an Overlap–Add/Granular Synthesis Approach," *J. Audio Eng. Soc.*, vol. 57, no. 1/2, pp. 29–37 (2009 Jan.).

[69] D. Schwarz and N. Schnell, "Descriptor-Based Sound Texture Sampling," in *Proceedings of the 7th International Conference in Sound and Music Computing (SMC)*, pp. 510–515 (Barcelona, Spain) (2010 Jul.).

[70] D. Schwarz, R. Cahen, F. Brument, H. Ding, and C. Jacquemin, "Sound Level of Detail in Interactive Audiographic 3D Scenes," in *Proceedings of the International Computer Music Conference (ICMC)*, pp. 312–315 (Huddersfield, UK) (2011 Aug.).

[71] S. O'Leary and A. Roebel, "A Two Level Montage Approach to Sound Texture Synthesis With Treatment of Unique Events," in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx)*, pp. 123–128 (Erlangen, Germany) (2014 Sep.).

[72] S. O'Leary and A. Röbel, "A Montage Approach to Sound Texture Synthesis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 6, pp. 1094–1105 (2016 Jun.).

[73] D. Schwarz, A. Roebel, C. Yeh, and A. Laburthe, "Concatenative Sound Texture Synthesis Methods and Evaluation," in *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx)*, pp. 217–224 (Brno, Czech Republic) (2016 Sep.).

[74] J. C. Stapleton and S. C. Bass, "Synthesis of Musical Tones Based on the Karhunen-Loeve Transform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 3, pp. 305–319 (1988 Mar.). https://doi.org/10.1109/29.1527.

[75] B. Recht and B. Whitman, "Musically Expressive Sound Textures From Generalized Audio," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx)* (London, UK) (2003 Sep.).

[76] J. Zheng, S.-H. Hung, K. Hiebel, and Y. Zhang, "Real-Time Rendering of Decorative Sound Textures for Soundscapes," *ACM Trans. Graph.*, vol. 39, no. 6, paper 271 (2020 Dec.). https://doi.org/10.1145/3414685.3417875.

[77] A. Di Scipio, "Synthesis of Environmental Sound Textures by Iterated Nonlinear Functions," in *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx)* (Trondheim, Norway) (1999 Dec.).

[78] A. D. Scipio, "The Synthesis of Environmental Sound Textures by Iterated Nonlinear Functions, and Its Ecological Relevance to Perceptual Modeling," *J. New Music Res.*, vol. 31, no. 2, pp. 109–117 (2002 Jan.). https://doi.org/10.1076/jnmr.31.2.109.8090.

[79] P. Hanna and M. Desainte-Catherine, "Time Scale Modification of Noises Using a Spectral and Statistical Model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6, pp. 181–184 (Hong Kong, China) (2003 Apr.). https://doi.org/10.1109/ICASSP.2003.1201648.

[80] A. Misra, P. R. Cook, and G. Wang, "A New Paradigm for Sound Design," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, pp. 319–324 (Montreal, Canada) (2006 Sep.).

[81] D. Van Nort, J. Braasch, and P. Oliveros, "Sound Texture Analysis Based on a Dynamical Systems Model and Empirical Mode Decomposition," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), paper 8251.

[82] D. Van Nort, J. Braasch, and P. Oliveros, "Sound Texture Recognition Through Dynamical Systems Modeling of Empirical Mode Decomposition," *J. Acoust. Soc. Am.*, vol. 132, no. 4, pp. 2734–2744 (2012 Oct.). https://doi.org/10.1121/1.4751535.

[83] S. Kersten and H. Purwins, "Fire Texture Sound Re-Synthesis Using Sparse Decomposition and Noise Modelling," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx)* (York, UK) (2012 Sep.).

[84] S. Farner, A. Solvang, A. Sæbo, and U. P. Svensson, "Ensemble Hand-Clapping Experiments Under the In-

fluence of Delay and Various Acoustic Environments," *J. Audio Eng. Soc.*, vol. 57, no. 12, pp. 1028–1041 (2009 Dec.)

[85] J. R. R. Lee and J. D. Reiss, "Real-Time Sound Synthesis of Audience Applause," *J. Audio Eng. Soc.*, vol. 68, no. 4, pp. 261–272 (2020 Apr.). https://doi.org/10.17743/jaes.2020.0006.

[86] V. Välimäki, J. Rämö, and F. Esqueda, "Creating Endless Sounds," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx)*, pp. 32–39 (Aveiro, Portugal) (2018 Sep.).

[87] S. Mallat, "Group Invariant Scattering," *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398 (2012 Jul.). https://doi.org/10.1002/cpa.21413.

[88] L. Sifre and S. Mallat, "Combined Scattering for Rotation Invariant Texture Analysis," in *Proceedings of the 20th European Symposium on Artificial Neural Networks*, vol. 44, pp. 68–81 (Bruges, Belgium) (2012 Apr.).

[89] J. Bruna and S. Mallat, "Invariant Scattering Convolution Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886 (2013 Aug.). https://doi.org/10.1109/TPAMI.2012.230.

[90] J. Andén and S. Mallat, "Deep Scattering Spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128 (2014 Aug.). https://doi.org/10.1109/TSP.2014.2326991.

[91] L. Gatys, A. S. Ecker, and M. Bethge, "Texture Synthesis Using Convolutional Neural Networks," in *Proceedings of the 29th International Conference on Neural Information Processing Systems*, pp. 262–270 (Montreal, Canada) (2015 Dec.).

[92] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman, "Preserving Color in Neural Artistic Style Transfer," *arXiv preprint arXiv:1606.05897v1rXiv preprint arXiv:1606.05897v1* (2016).

[93] S. Dieleman, "Recommending Music on Spotify With Deep Learning," https://benanne.github.io/2014/08/05/spotify-cnns.html (accessed Jun. 7, 2021).

[94] Z. Mushtaq, S.-F. Su, and Q.-V. Tran, "Spectral Images Based Environmental Sound Classification Using CNN With Meaningful Data Augmentation," *Appl. Acoust.*, vol. 172, paper 107581 (2021 Jan.). https://doi.org/10.1016/j.apacoust.2020.107581.

[95] H. Caracalla and A. Roebel, "Sound Texture Synthesis Using Convolutional Neural Networks," *arXiv preprint arXiv:1905.03637v1* (2019).

[96] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556v6* (2014).

[97] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations From Unlabeled Video," in *Proceedings of the 30th Conference on Neural Information Processing Systems*, pp. 892–900 (Barcelona, Spain) (2016 Dec.).

[98] J. Chorowski, R. J. Weiss, R. A. Saurous, and S. Bengio, "On Using Backpropagation for Speech Texture Generation and Voice Conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2256–2260 (Calgary, Canada) (2018 Apr.). https://doi.org/10.1109/ICASSP.2018.8461282.

[99] M. H. bin Md Shahrin and L. Wyse, "Applying Visual Domain Style Transfer and Texture Synthesis Techniques to Audio: Insights and Challenges," *Neural Comput. Appl.*, vol. 32, no. 4, pp. 1051–1065 (2020 Feb.). https://doi.org/10.1007/s00521-019-04053-8.

[100] P. K. Mital, "Time Domain Neural Audio Style Transfer," *arXiv preprint arXiv:1711.11160v1* (2017).

[101] J. Chen, G. Yang, H. Zhao, and M. Ramasamy, "Audio Style Transfer Using Shallow Convolutional Networks and Random Filters," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 15043–15057 (2020 Jun.).

[102] L. V. Ulyanov D, "Audio Texture Synthesis and Style Transfer," https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/. (accessed Jun. 7, 2021).

[103] P. Verma and J. O. Smith, "Neural Style Transfer for Audio Spectograms," *arXiv preprint arXiv:1801.01589v1* (2018).

[104] A. Perez, C. Proctor, and A. Jain, "Style Transfer for Prosodic Speech," Tech. Rep., Stanford University (2017). https://web.stanford.edu/class/cs224s/project/reports_2017/Anthony_Perez.pdf.

[105] D. Schwarz, "Retexture — Towards Interactive Environmental Sound Texture Synthesis Through Inversion of Annotations," in *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research*, pp. 283–290 (Marseille, France) (2013 Oct.).

[106] J. Ye, T. Kobayashi, N. Toyama, H. Tsuda, and M. Murakawa, "Acoustic Scene Classification Using Efficient Summary Statistics and Multiple Spectro-Temporal Descriptor Fusion," *Appl. Sci.*, vol. 8, no. 8, paper 1363 (2018 Aug.). https://doi.org/10.3390/app8081363.

[107] M. Green and D. Murphy, "Environmental Sound Monitoring Using Machine Learning on Mobile Devices," *Appl. Acoust.*, vol. 159, paper 107041 (2020 Feb.). https://doi.org/10.1016/j.apacoust.2019.107041.

[108] Z. Ren, V. Pandit, K. Qian, et al., "Deep Sequential Image Features on Acoustic Scene Classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, pp. 113–117 (Munich, Germany) (2017 Nov.).

[109] W. Yang and S. Krishnan, "Combining Temporal Features by Local Binary Pattern for Acoustic Scene Classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1315–1321 (2017 Jun.). https://doi.org/10.1109/TASLP.2017.2690558.

[110] S. Abidin, R. Togneri, and F. Sohel, "Spectrotemporal Analysis Using Local Binary Pattern Variants For Acoustic Scene Classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 11, pp. 2112–2121 (2018 Nov.). https://doi.org/10.1109/TASLP.2018.2854861.

[111] M. Nigro and S. Krishnan, "SARdB: A Dataset for Audio Scene Source Counting and Analysis,"

Appl. Acoust., vol. 178, paper 107985 (2021 Jul.). https://doi.org/10.1016/j.apacoust.2021.107985.

[112] C. Uhle, "Applause Sound Detection," *J. Audio Eng. Soc.*, vol. 59, no. 4, pp. 213–224 (2011 Apr.).

[113] G. Sharma, X.-P. Zhang, K. Umapathy, and S. Krishnan, "Audio Texture and Age-Wise Analysis of Disordered Speech in Children Having Specific Language Impairment," *Biomed. Signal Process. Control*, vol. 66, paper 102471 (2021 Apr.). https://doi.org/10.1016/j.bspc.2021.102471.

[114] Y. M. G. Costa, L. S. Oliveira, A. L. Koericb, and F. Gouyon, "Music Genre Recognition Using Spectrograms," in *Proceedings of the 18th International Conference on Systems, Signals and Image Processing*, pp. 1–4 (Sarajevo, Bosnia and Herzegovina) (2011 Jun.).

[115] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins, "Music Genre Classification Using LBP Textural Features," *Signal Process.*, vol. 92, no. 11, pp. 2723–2737 (2012 Nov.). https://doi.org/10.1016/j.sigpro.2012.04.023.

[116] J. H. Foleis and T. F. Tavares, "Texture Selection for Automatic Music Genre Classification," *Appl. Soft Comput.*, vol. 89, paper 106127 (2020 Apr.). https://doi.org/10.1016/j.asoc.2020.106127.

[117] J. Uddin, M. Kang, D. V. Nguyen, and J.-M. Kim, "Reliable Fault Classification of Induction Motors Using Texture Feature Extraction and a Multiclass Support Vector Machine," *Math. Probl. Eng.*, vol. 2014, paper 814593 (2014 Jun.). https://doi.org/10.1155/2014/814593.

[118] S. Zhao, K. Chang, E. Wang, B. Li, K. Wang, and Q. Wu, "Fault Diagnosis of Oil Pumping Machine Retarder Based on Sound Texture-Vibration Entropy Characteristics and Gray Wolf Optimization-Support Vector Machine," *Shock Vib.*, vol. 2020, paper 2709384 (2020 May). https://doi.org/10.1155/2020/2709384.

[119] W. Sun and X. Cao, "Curvature Enhanced Bearing Fault Diagnosis Method Using 2D Vibration Signal," *J. Mech. Sci. Technol.*, vol. 34, pp. 2257–2266 (2020 May).

THE AUTHORS

Garima Sharma

Karthikeyan Umapathy

Sridhar Krishnan

Garima Sharma is a Ph.D. candidate in the Department of Electrical and Computer Engineering at Ryerson University, Toronto. She received her Master's degree in 2010 and Bachelor's degree in 2007 in Electronics and Communication Engineering from Kurukshetra University, India. She has teaching experience of over eight years working as an Assistant Professor in the Department of Electronics and Communication Engineering at Amity University, India. Her research interests include audio signal processing and signal analysis, including pathological speech. She has been awarded the Ryerson Graduate Fellowship since 2018.

•

Karthikeyan Umapathy is a Professor and Program Director of Biomedical Engineering in the Department of Electrical, Computer, and Biomedical Engineering at Ryerson University, Toronto, Canada. He has been working in the area of biomedical signal processing and imaging with applications in cardiac electrophysiology, sleep neurophysiology, and functional cardiac MRI imaging. A majority of his research is in collaboration with clinician researchers at major hospitals in Toronto. He has published more than 70 articles in refereed journals and conferences and 19 refereed abstracts in journals. Many of the works involve the use of machine learning and pattern classification methods for decision making. He is a senior member of IEEE and Affiliate Scientist at St. Michael's Hospital (Institute for Biomedical Engineering, Science and Technology), Toronto, Canada. He has been a recipient of prestigious Canadian federal research grants as Principal Investigator and Co-Principal Investigator toward his research.

•

Sridhar Krishnan joined Ryerson University, Toronto, Canada in 1999, and he is now a Professor of Electrical, Computer, and Biomedical Engineering. His research interests are in biomedical signal analysis, audio signal analysis, and explainable machine learning. He is a Fellow of the Canadian Academy of Engineering. From 2007–2017 he was a Canada Research Chair in Biomedical Signal Analysis. Sridhar Krishnan has published 385 articles in refereed journals and conferences, filed 16 invention disclosures/patent applications, and been a scientific advisor to five AI/wearables start-ups. Sridhar Krishnan is a recipient of the Outstanding Canadian Biomedical Engineer Award, Achievement in Innovation Award from Innovate Calgary, Sarwan Sahota Distinguished Scholar Award, Young Engineer Achievement Award from Engineers Canada, New Pioneers Award in Science and Technology, and Exemplary Service Award from IEEE.