

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,800

Open access books available

142,000

International authors and editors

180M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Classification and Separation of Audio and Music Signals

Abdullah I. Al-Shoshan

Abstract

This chapter addresses the topic of classification and separation of audio and music signals. It is a very important and a challenging research area. The importance of classification process of a stream of sounds come up for the sake of building two different libraries: speech library and music library. However, the separation process is needed sometimes in a cocktail-party problem to separate speech from music and remove the undesired one. In this chapter, some existed algorithms for the classification process and the separation process are presented and discussed thoroughly. The classification algorithms will be divided into three categories. The first category includes most of the real time approaches. The second category includes most of the frequency domain approaches. However, the third category introduces some of the approaches in the time-frequency distribution. The approaches of time domain discussed in this chapter are the short-time energy (STE), the zero-crossing rate (ZCR), modified version of the ZCR and the STE with positive derivative, the neural networks, and the roll-off variance. The approaches of the frequency spectrum are specifically the roll-off of the spectrum, the spectral centroid and the variance of the spectral centroid, the spectral flux and the variance of the spectral flux, the cepstral residual, and the delta pitch. The time-frequency domain approaches have not been yet tested thoroughly in the process of classification and separation of audio and music signals. Therefore, the spectrogram and the evolutionary spectrum will be introduced and discussed. In addition, some algorithms for separation and segregation of music and audio signals, like the independent Component Analysis, the pitch cancelation and the artificial neural networks will be introduced.

Keywords: audio signal, music signal, classification, separation, time domain, frequency domain, time-frequency domain

1. Introduction

Audio signal processing is an important subfield of signal processing that is concerned with the electronic manipulation of audio signals [1–6]. The problem of discriminating music from audio has increasingly become very important as automatic audio signal recognition (ASR) systems and it has been increasingly applied in the domain of real-world multimedia [7]. Human's ear can easily distinguish audio without any influence of the mixed music [8–23]. Due to the new methods of the analysis and the synthesis processing of audio signals, the processing of musical signals has gained particular weight [16, 24], and therefore, the classical sound analysis methods may be used in the processing of musical signals [25–28]. Many

types of musical signals such as Rock music, Pop music, Classical music, Country music, Latin music, Arabic music, Disco and Jazz, Electronic music, etc. are existed [29]. The sound type signals hierarchy is shown in **Figure 1** [30].

Audio signal changes randomly and continuously through time. As an example, music and audio signals have strong energy content in the low frequencies and weaker energy content in the high frequencies [31, 32]. **Figure 2** depicts a generalized time and frequency spectra of audio signals [33]. The maximum frequency f_{max} varies according to type of audio signal, where, in the telephone transmission f_{max} is equal to 4 kHz, 5 kHz in mono-loudspeaker recording, 6 KHz in multi-loudspeaker recording or stereo, 11 kHz in FM broadcasting, however, it equals to 22 KHz in the CD recording.

Acoustically speaking, the audio signals can be classified into the following classes:

1. Single talker in specific time [34].
2. Singing without music.
3. Mixture of background music and single talker audio.
4. Songs that are a mixture of music with a singer voice.
5. May completely be music signal without any audio component.
6. Complex sound mixture like multi-singers or multi-speakers with multi-music sources.

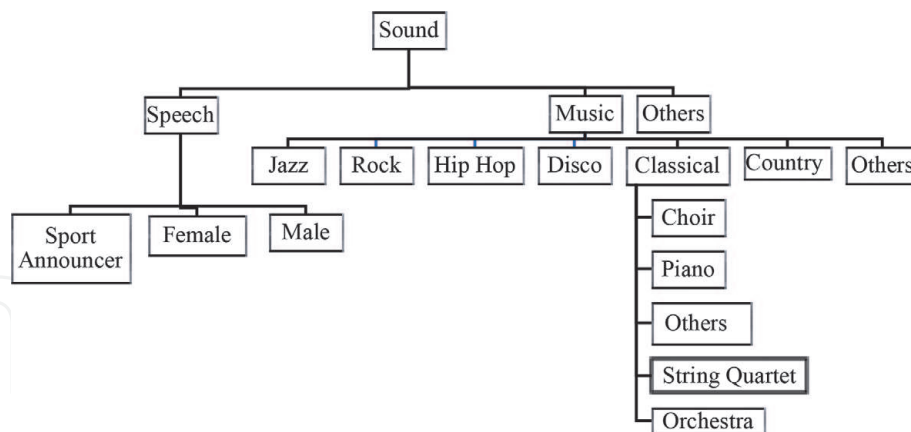


Figure 1.
Types of audio signals.

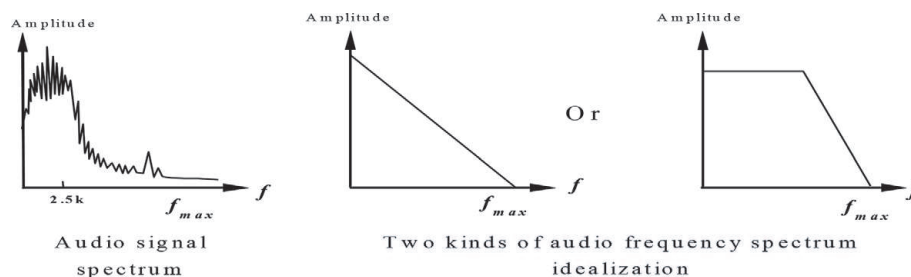


Figure 2.
Generalized frequency spectrum for audio signal [33].

7. Non-music and non-audio signals: like fan, motor, car, jet sounds, etc.
8. Audio signal that is a mixture of more than one speakers talking simultaneously at the same time [8].
9. Abnormal music can be single word cadence, human whistle sound, or opposite reverberation [4, 34–38].

2. Analysis of audio and music signals

2.1 Properties of audio signal

2.1.1 Representation of audio signal

The letters symbols used for writing are not adequate, as the way they are pronounced varies; for example, the letter “o” in English, is pronounced differently in words “pot” most“ and “one”. It is almost impossible to tackle the audio classification problem without first establishing some way of representing the spoken utterances by some group of symbols representing the sounds produced [39–43]. The phonemes in **Table 1** are divided into groups based on the way they are produced [44], forming a set of *allophones* [45]. In some tonal languages, such as Vietnamese and Mandarin, the intonation determines the meaning of each word [46–48].

2.1.2 Production of audio signal

Since the range of sounds that can be produced by any system is limited [39–44], the pressure in the lungs is increased by the reverse process. They push the air up the *trachea*; the larynx is situated at the top of the trachea. By changing the shape of the vocal tract, different sounds are produced, so the fundamental frequency will be changing with time. The spectrogram (or sonogram) for the sentence “What can I have for dinner tonight?” is shown in **Figure 3**.

Vowels	Diphthongs	Fricatives	Plosives	Semivowels	Nasals	Affricates
heed	bay	sail	bat	was	am	jaw
hid	by	ship	disc	ran	an	chore
head	bow	funnel	Goat	lot	sang	
had	bough	thick	pool	yacht		
hard	beer	hull	tap			
hod	doer	zoo	kite			
hoard	boar	azure				
hood	boy	that				
who'd	bear	valve				
hut						
heard						
the						

Table 1. Phoneme categories of British English and examples of words in which they are used [44].

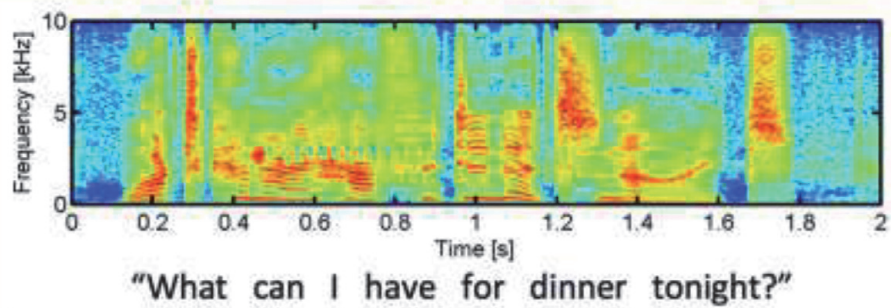


Figure 3.
A sonogram for the sentence “What can I have for dinner tonight?” [43].

The way that humans recognize and interpret audio signal has been considered by many researchers [1, 25, 39]. To produce a complete set of English vowels, many researchers have depicted that the two lowest formants are necessary, as well as that the three lowest formants in frequency are necessary for good audio intelligibility. As the number of formants increased, sounds that are more natural are produced. However, when we deal with continuous audio, the problem becomes more complex. The history of audio signal identification can be found in [1, 25, 39–48].

2.2 Properties of music signal

2.2.1 Representation of music signal

There are two kinds of tone structures in music signal. The first one is a simple tone formed of single sinusoidal waveform, however, the second one is a more complex tone consisting of more than one harmonic [31, 49–52]. The spectrum of music signal has twice the bandwidth of audio spectrum, and most of the power of audio signal is concentrated at lower frequencies. Melodists and musicians divide musical minor to eight parts and each part named octave, where each octave is divided into seven parts called tones [30]. For different instrument, a tempered scale is shown in **Table 2**. These tones, shown in **Table 2**, are named (Do, Re, Me, Fa, So, La and Se) or simply (A, B, C, D, E, F, and G). The tone (A₁) at the first octave has the fundamental frequency of the first tone in each octave, i.e., every first tone in each octave takes the reduplicate frequency of the first tone of previous one, (i.e., $A_n = 2^n A_1$ or $B_n = 2^n B_1$ and so on where $n \in \{2, 3, 4, 5, 6, 7\}$).

From **Table 2**, the highest tone C₈ occurs at the frequency of 4186 Hz, which is the highest frequency produced by human sound system, which leads musical

A Hz	B Hz	C Hz	D Hz	E Hz	F Hz	G Hz
A ₁ 27.5	B ₁ 30.863	C ₁ 32.703	D ₁ 36.708	E ₁ 41.203	F ₁ 43.654	G ₁ 48.99
A ₂ 55	B ₂ 61.735	C ₂ 65.406	D ₂ 73.416	E ₂ 82.407	F ₂ 87.307	G ₂ 97.99
A ₃ 110	B ₃ 123.47	C ₃ 130.81	D ₃ 146.83	E ₃ 164.81	F ₃ 174.61	G ₃ 196
A ₄ 220	B ₄ 246.94	C ₄ 261.63	D ₄ 293.66	E ₄ 329.63	F ₄ 349.23	G ₄ 392
A ₅ 440	B ₅ 493.88	C ₅ 523.25	D ₅ 587.33	E ₅ 659.26	F ₅ 698.46	G ₅ 783.9
A ₆ 880	B ₆ 987.77	C ₆ 1046.5	D ₆ 1174.7	E ₆ 1318.5	F ₆ 1396.9	G ₆ 1568
A ₇ 176	B ₇ 1975.5	C ₇ 2093	D ₇ 2349.3	E ₇ 2637	F ₇ 2793	G ₇ 3136
A ₈ 352	B ₈ 3951.1	C ₈ 4186				

Table 2.
Frequencies of notes in the tempered scale [3].

instrument manufactures to try their best to bound music frequency to human's sound system limits to achieve strong concord [35, 53, 54]. In the real world, musical instruments cover more frequencies than audible band, which is limited to 20 kHz).

2.2.2 Production of music signal

The concept of tone quality that is most common depends on the subjective acoustic properties, regardless of partials or formants and the production of music depends mainly on the kind of musical instruments [53, 54]. These instruments can be summarized as follows:

1. **The string musical instrument.** Its tones is produced by vibrating chords made from horsetail hair, or other manufactured material like copper or plastic. Every vibrating chord has its own fundamental frequency, producing complex tones so that it covers most of the audible bands. **Figure 4** shows string instruments.
2. **The brass musical instrument.** The Brass musical instrument depends on blowing air like woodwind. Its shape looks like an animal horn and has manual valves to control cavity size. Brass musical instrument has huge number of nonharmonic signals existed in its spectrum. **Figure 5** shows brass instruments.
3. **The woodwind musical instrument.** Woodwind instrument consists of an open cylindrical tube at both ends. Some woodwind instruments may use small-vibrated piece of copper to produce tones. It produces many numbers of harmonic tones. **Figure 6** shows woodwind instruments.
4. **The percussion musical instrument.** Examples of percussion instruments are piano, snare drum, chimes, marimba, timpani, and xylophone. Most of the power of tones in percussion instruments produces non-harmonic components. **Figure 7** shows some percussion instruments.
5. **The electronic musical instrument.** The most qualified robust and accurate electronic musical instrument is the organ. It has a large keyboard, a memory that can store notes and use their frequencies as basic cadences or tones. Without organ help, disco, pop, rock and jazz cannot stand [29, 35–38]. Organ is not the only electronic musical producer. If the electronic musical

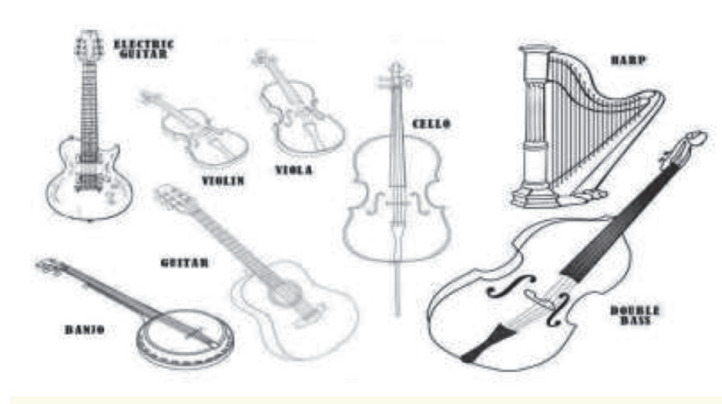


Figure 4.
String instruments.



Figure 5.
Brass instruments.



Figure 6.
Woodwind instruments.

instruments are used for producing music, the tone quality measure of the fundamental frequency or harmonics is not needed. **Figure 8** shows an example of organ electronic instrument.

2.3 Characteristics and differences between audio and music

The audio signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time “between 5 and 100 msec. Therefore, its characteristics are stationary within this period of time. A simple example of an audio signal is shown in **Figure 9**.

Figure 10 is a typical example of music portion. It is very clear from the two spectrums in **Figures 9** and **10** that we can distinguish between the two types of signals.

Figures 11 and **12** depict the evolutionary spectrum of two different types of signals, audio and music.

Now, let us discuss some of the main similarity and differences between the two types of signals.



Figure 7.
Percussion instruments.



Figure 8.
Electronic organ.

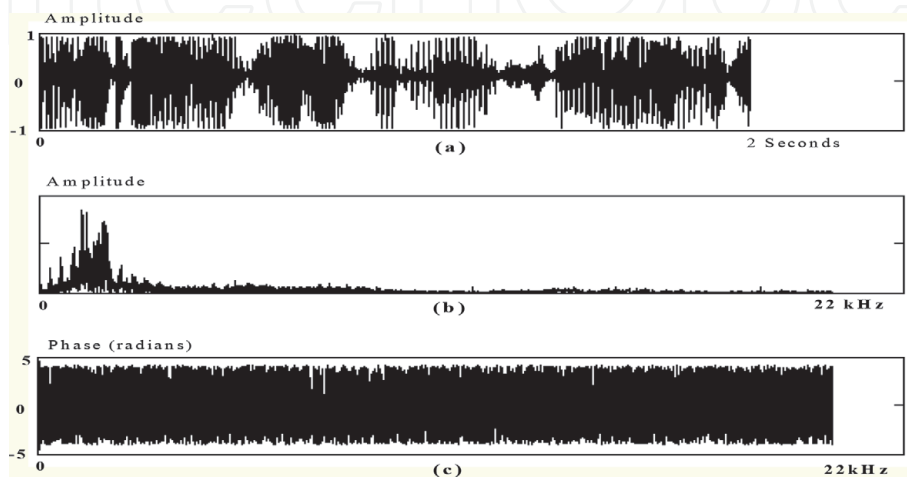


Figure 9.
An example of audio signal of speaking the two-second long phrase “Very good night”: (a) time domain (b) magnitude. (c) Phase.

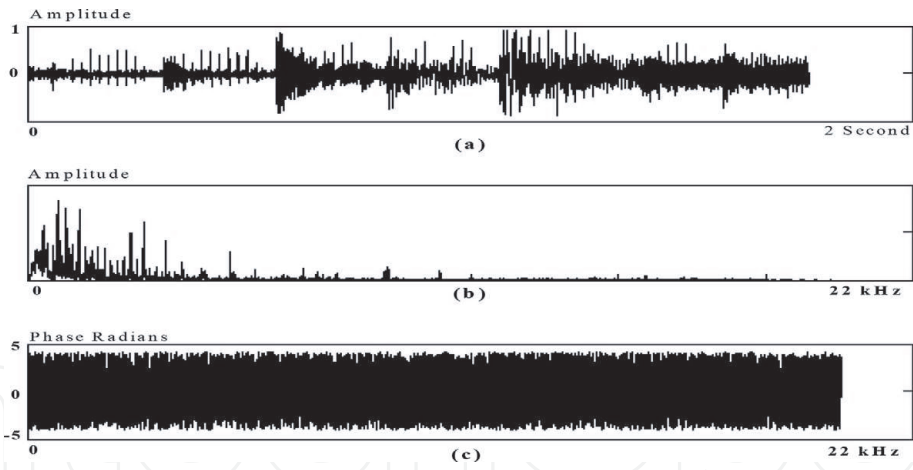


Figure 10. A 2-second long music signal: (a) time domain. (b) Spectrum. (c) Phase.

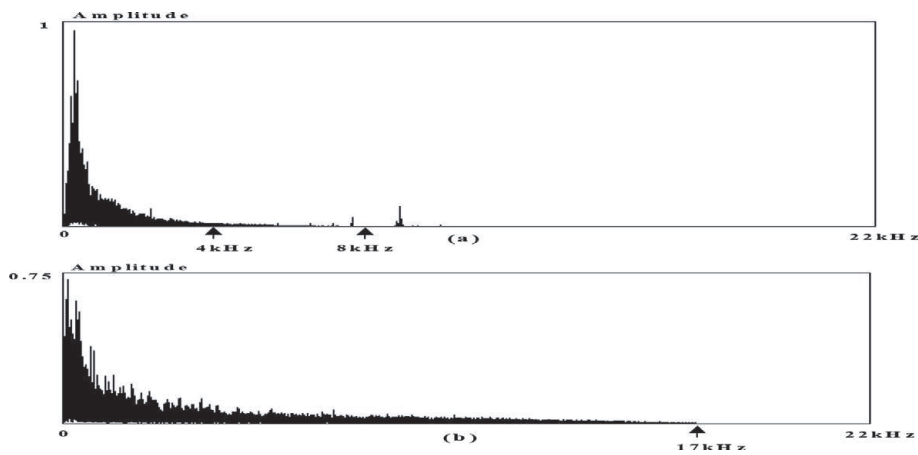


Figure 11. The spectrum of an average of 500 specimens: (a) audio, (b) music.

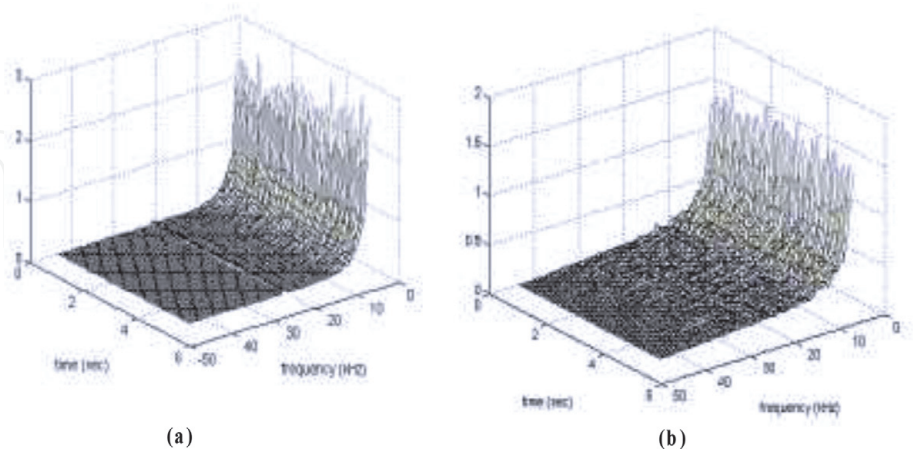


Figure 12. Evolutionary spectrum of an average of 500 specimens: (a) audio, (b) music.

Tonality. By tone, we mean a single harmonic of a pure periodical sinusoid. Regardless of the type of instruments or music, the musical signal is composed of a multiple of tones; however, this is not the case in the voice signal [47, 52, 55–57].

Bandwidth. Normally, the audio signal has 90% of its power concentrated within frequencies lower than 4 kHz and limited to 8 kHz; however, music signal can extend its power to the upper limits of the ear’s response, which is 20 kHz [52, 58].

Alternative sequence. Audio exhibits an alternating sequence of noise-like segment while music alternates in more tonal shape. In other words, audio signal is distributed through its spectrum more randomly than music does.

Power distribution. Normally, the power distribution of an audio signal is concentrated at frequencies lower than 4 kHz, and then collapsed rapidly above this frequency. On the other hand, there is no specific shape of the power of music spectrum [59].

Dominant frequency. For a single talker, his dominant frequency can accurately be determined uniquely, however, in a single musical instrument only the average dominant frequency can be determined. In multiple musical instruments, the case will be worst.

Fundamental frequency. For a single talker, his fundamental frequency can be accurately configured. However, this is not the case for a single music instrument.

Excitation patterns. The excitation signals (pitch) for audio are usually existed only over a span of three octaves, while the fundamental music tones can span up to six octaves [60].

Energy sequences. A reasonable generalization is that audio follows a pattern of high-energy conditions of voicing followed by low energy conditions, which the envelope of music is less likely to exhibit.

Tonal duration. The duration of vowels in audio is very regular, following the syllabic rate. Music exhibits a wider variation in tone lengths, not being constrained by the process of articulation. Hence, tonal duration would likely be a good discriminator.

Consonants. Audio signal contains too many consonants while music is usually continuous through the time [33].

Zero crossing rate (ZCR). The ZCR in music is greater than that in audio. We can use this idea to design a discriminator [60].

In the frequency domain, there is a strong overlapping between audio and music signals, so no ordinary filter can separate them. As mentioned before, audio signal may cover spectrum between 0 and 4 kHz with a dominant frequency of an average = 1.8747 kHz. However, the lowest fundamental frequency (A1) of a music signal is about 27.5 Hz and the highest frequency of the tone C8 is around 4186 Hz. The reason behind this is that musical instrument manufacturers try to bound music frequency to human's sound limits in order to achieve a strong consonant and a strong frequency overlap. Moreover, music may propagate over the audible

Key Difference	Audio	Music
Units of Analysis	Phonemes	Notes Finite
Temporal Structure	<ul style="list-style-type: none"> • Short sample (40 ms–200 ms). • More steady state than dynamic. • Timing unstrained but variable. • Amplitude modulation rate for sentences is slow (~ 4 Hz) 	<ul style="list-style-type: none"> • Longer sample: 600–1200 ms. • Mix of steady state (strings, winds) and transient (percussion). • Strong periodicity.
Spectral Structure	<ul style="list-style-type: none"> • Largely harmonic (vowels, voiced consonants). • Tend to group in formants. • Some inharmonic stops. 	<ul style="list-style-type: none"> • Largely harmonic and some inharmonic (percussion).
Syntactic / Semantic Structure	<ul style="list-style-type: none"> • Symbolic • Productive • Can be combined in grammar 	<ul style="list-style-type: none"> • Symbolic • Productive • Combined in a grammar

Table 3.
 The main differences between audio and music signals.

spectrum to cover more than the audible band of 20 kHz, with a dominant frequency of an average = 1.9271 kHz [25].

Table 3 summarizes the main similarity and differences between music and audio signals.

3. Audio and music signals classification

The main classification approaches will be discussed in this section. They can be categorized into three different approaches: (1) time domain approaches, (2) frequency domain approaches, and (3) time-frequency domain approaches. A two-level music and audio classifier was developed by El-Maleh [61, 62]. He used a combination of long-term features such as the variance, the differential parameters, the zero crossing rate (ZCR), and the time-averages of spectral parameters. Saunders [60] proposed another two-level classifier. His approach was based on the short-time energy (STE) and the average ZCR features. In addition, Matityaho and Furst [63] have developed a neural network based model for classifying music signals. Their model was designed based on human cochlea functional performance.

For audio detection, Hoyt and Wecheler [64] have developed a neural network base model using Fourier transform, Hamming filtering, and a logarithmic function as pre-processing then they applied a simple threshold algorithm for detecting audio, music, wind, traffic or any interfering sound. In addition, to improve the performance, they suggested wavelet transform feature for pre-processing. Their work is much similar to the work done by Matityaho and Furst's [63, 64]. 13 features were examined by Scheirer and Slaney [65]. Some of these features were simple modification of each other's. They also tried combining them in several multidimensional classification forms. From these previous works, the most powerful discrimination features were the STE and the ZCR. Therefore, the STE and the ZCR will be discussed thoroughly. Finally, the common classifiers of the audio and the music signals can be divided into the following approaches:

I. The Time domain algorithms:

1. The ZCR algorithm [1, 34, 66–77]:
 - a. The standard deviation of first order difference of the ZCR.
 - b. The 3rd central moment of the mean of ZCR.
 - c. The total number of zero crossings exceeding a specific threshold.
2. The STE [60–65, 78].
3. The ZCR and the STE positive derivative [78, 79].
4. The Pulse Metric [31, 59, 80–82].
5. The number of silence [32, 60].
6. The HMM (Hidden Markov Model) [83–85].
7. The ANN (Artificial neural networks) [12, 49, 58, 63, 79, 83–120].
8. The Roll-Off Variance [31, 59].

II. The Frequency-domain algorithms [32, 33, 35, 59, 112, 66–77, 121]:

1. The Spectrum [31, 111]:

- a. The Spectral Centroid.
- b. The Spectral Flux Variance.
- c. The Spectral Centroid Mean and Variance.
- d. The Spectral Flux Mean and Variance.
- e. The Spectrum Roll-Off.
- f. The Signal Bandwidth.
- g. The Spectrum Amplitude.
- h. The Delta Amplitude.

2. The Cepstrum [122]:

- a. The Cepstral Residual [122–124].
- b. The Variance of the Cepstral Residual [122–124].
- c. The Cepstral feature [122–124].
- d. The Pitch [94, 107, 108, 117–119, 125, 126].
- e. The Delta Pitch [88, 119].

III. The Time-Frequency domain algorithms:

1. The Spectrogram (or Sonogram) [13, 19, 86, 127].

2. The Evolutionary Spectrum and the Evolutionary Bispectrum [81, 128, 129].

3.1 Time domain algorithms

3.1.1 The ZCR algorithm

The ZCR algorithm can be defined as the number of crossing the signal the zero axis within a specific window. It is widely used because its simplicity and robustness [34]. We may define the ZCR as in the following equation.

$$Z_n = \frac{1}{2N} \sum_{m=n-N+1}^N | \operatorname{sgn} [x(m)] - \operatorname{sgn} [x(m-1)] | \quad (1)$$

where Z_n is the ZCR, N is the number of samples in one window, and sgn is the sign of the signal such that $\operatorname{sgn} [x(n)] = 1$ when $x(n) > 0$, $\operatorname{sgn} [x(n)] = -1$,

when $x(n) < 0$. An essential note is that the sampling rate must be high enough to catch any crossing through zero. Another important note before evaluating the ZCR is to normalize the signal by subtracting its average value. It is clear from Eq. (1) that the value of the ZCR is proportional to the sign change in the signal, i.e., the dominant frequency of $x(n)$. Therefore, we may find that the ZCR of music is, in general, higher than that of audio, but not sure at the unvoiced audio.

Properties of ZCR:

The ZCR properties can be summarized as follow.

1. The Principle of Dominant Frequency

The dominant frequency of a pure sinusoid is the only value in the spectrum. This value of frequency is equal to the ZCR of the signal in one period. If we have a non-sinusoidal periodic signal, its dominant frequency is frequency with the largest amplitude. The dominant frequency (ω_0) can be evaluated as follow.

$$\omega_0 = \frac{\pi E\{D_0\}}{N - 1} \quad (2)$$

where N is the number of intervals, $E\{.\}$ is the expected value, and D_0 is the ZCR per interval.

2. The Highest frequency

Since D_0 denotes the ZCR of a discrete-time signal $Z(i)$, let us assume that D_n denotes the ZCR of the n^{th} derivative of $Z(i)$, i.e., D_1 is the ZCR of the first derivative of $Z(i)$, D_2 is the ZCR of the second derivative of $Z(i)$, and so on. Then, the highest frequency ω_{max} in the signal can be evaluated as follow.

$$\omega_{max} = \lim_{i \rightarrow \infty} \frac{\pi E\{D_i\}}{N - 1} \quad (3)$$

where N is the number of samples. If the sampling rate equals 11 KHz, then the change in ω_{max} can be ignored for $i > 10$.

3. The Lowest frequency

Assuming that the time period between any two samples is normalized to unity, the derivative of $Z(i)$ can be defined as $Z(i) = Z(i) - Z(i-1)$. Then, the ZCR of the n^{th} derivative of $Z(i)$ is defined as D_n . Now, let us define ∇^+ as the +ve derivative of $Z(i)$, then $\nabla^+ [Z(i)]$ can be defined as follow.

$$\nabla^+ [Z(i)] = Z(i) + Z(i - 1) \quad (4)$$

Now, let us define the ZCR of the n^{th} +ve derivative of $Z(i)$ by the symbol ${}_n D$. Then we can find the lowest frequency ω_{min} of a signal as follow.

$$W_{min} = \lim_{i \rightarrow \infty} \frac{\pi E\{D_i\}}{N - 1} \quad (5)$$

4. Measure of Periodicity

A signal is said to be purely periodic if and only if.

$$E\{D_1\} = E\{D_2\} \quad (6)$$

Using Eq. (6), it was found that music is more periodic or than audio [44–47, 55–57, 130].

The Ratio of High ZCR (RHZCR)

It was found that the variation of the ZCR is more discriminative than the exact ZCR, so the RHZCR can be considered as one feature [78]. The RHZCR is defined as the ratio of the number of frames whose ZCR are above 1 over the average ZCR in one-window, and can be defined as follow.

$$\text{RHZCR} = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sgn}(\text{ZCR}(n) - \text{ZCR}_{av}) + 1] \quad (7)$$

$$\text{ZCR}_{av} = \sum_{n=0}^{N-1} \text{ZCR}(n) \quad (8)$$

where N is the number of frames per one-window, n is the index of the frame, $\text{sgn}[\cdot]$ is a sign function and $\text{ZCR}(n)$ is the zero-crossing rate at the n^{th} frame. In general, audio signals consist of alternating voiced and unvoiced sounds in each syllable rate, while music does not have this kind of alternation. Therefore, from Eq. (7) and Eq. (8), we may observe that the variation of the ZCR (or the RHZCR) in an audio signal is greater than that of a music, as shown in **Figure 13**.

3.1.2 The STE algorithm

The amplitude of the audio signal varies appreciably with time. In particular, the amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments. The STE of the audio signal provides a convenient representation

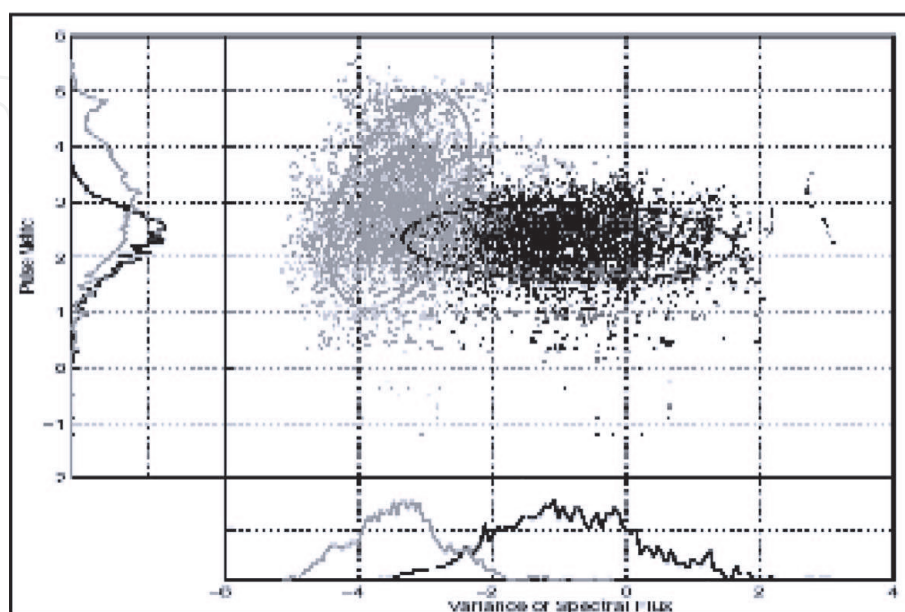


Figure 13.
 Music and audio sharing some values [65].

that reflects these amplitude variations. Unlike the audio signal, since the music signal does not contain unvoiced segments, the STE of the music signal is usually bigger than that of audio [60]. The STE of a discrete-time signal $s(n)$ can define as.

$$\text{STE}_s = \sum_{n=-\infty}^{\infty} |s(n)|^2 \quad (9)$$

where STE_s in Eq. (9) is the total energy of the signal. The average power of $s(n)$ is defined as.

$$P_s = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N |s(n)|^2 \quad (10)$$

Signals can be classified into three types, in general: an energy signal, which has a non-zero and finite energy, a power signal, which has a non-zero and finite energy, and the third type is neither energy nor power signal, see **Table 4**. Now, let us define another sequence $\{f(n;m)\}$ as follow.

$$f_s(n, m) = s(n)w(m-n) \quad (11)$$

where $w(n)$ is just a window with a length of N with a value of zero outside $[0, N-1]$. Therefore, $f_s(n, m)$ will be zero outside $[m-N+1, m]$.

Deriving short term features

The silence and unvoiced period in audios can be considered a stochastic background noise. Now, let us define F_s as a feature of $\{s(n)\}$, mapping its values of the Hilbert space, H , to a set of complex numbers C such that.

$$F_s : H \rightarrow C \quad (12)$$

The long-term feature of $\{s(n)\}$ may be defined as follow.

$$L\{s(n)\} = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N s(n) \quad (13)$$

The long-term average, when applied to energy signals, will have zero values, however, it is appropriate for power signals. Eq. (13) can be re-written as follow.

$$L\{s(n)\} = \frac{1}{2N} \sum_{n=-\infty}^{\infty} s(n) \quad (14)$$

Energy Signal $0 < E_s < \infty$	Transient	$S(n) = \alpha^n u(n) \quad \alpha < 1$
	Finite Sequence	$e^{\beta t} [u(n) - u(n-255)] \quad \beta < \infty$
Power Signal $0 < P_s < \infty$	Constant	$s(n) = \alpha \quad -\infty < \alpha < \infty$
	Periodic	$s(n) = \alpha \sin(n\omega_0 + \varphi) \quad -\infty < \alpha < \infty$
	Stochastic	$S(n) = \text{rand}(\text{seed})$
Neither Energy nor Power Signal	Zero	$s(n) = 0$
	Blow up	$s(n) = \alpha^n u(n) \quad \alpha > 1$

Table 4.
Types of signals.

Resulting a family of mappings. If each member of the family is selected to be a λ , then we can use the notation $F_s(\lambda)$. The discrete-time Fourier transform is an example of a parametric long-term feature. The long-term feature can be of the form.

$$L\{M(\lambda)\{s(n)\}\} \quad (15)$$

where M in Eq. (15) is the mapping sequence. It maps $\{s(n)\}$ to another sequence. The long-term feature $F_s(\lambda)$ is defined as $L^o M$, a composition of function L and M . If $F_s(\lambda)$ is the long-term feature of Eq. (12), then the short-term feature $F_s(\lambda, m)$ of time period m can be constructed as follows:

- Define a frame as in Eq. (11).
- Apply the long-term feature transformation to the frame sequence as in Eq. (16).

$$\begin{aligned} F_s(\lambda, m) &= L\{M(\lambda)\{f_s(n, m)\}\} \\ &= L\{M(\lambda)\{s(n)w(m-n)\}\} \\ &= \frac{1}{N} \sum_{n=-\infty}^{\infty} M(\lambda)\{s(n)w(m-n)\} \end{aligned} \quad (16)$$

Low Short Time Energy Ratio (LSTER)

As done in the ZCR, the variation is selected [33]. Here, the LSTER is used to represent the variation of the STE. LSTER is defined as the ratio of the number of frames whose STE are less than 0.5 times of the average STE in a one-second window, as in Eq. (17).

$$\text{LSTER} = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sgn}(0.5\text{STE}_{av} - \text{STE}(n)) + 1] \quad (17)$$

where.

$$\text{STE}_{av} = \sum_{n=0}^{N-1} \text{STE}(n) \quad (18)$$

N is the total number of frames, $\text{STE}(n)$ is the STE at the n^{th} frame, and STE_{av} in Eq. (18) is the average STE in a one-window.

3.1.3 The effect of positive derivation

Figure 14 shows the preprocessing flow on $Z(i)$ using the positive derivation concept (+), which provided some improvement in the discrimination process [78].

This pre-processing increased the ZCR of music and reduced the ZCR of the audio with the expenses of some delay. The averages of the ZCR in speech, mixture, and music are shown in **Figure 15**, after applying the +ve derivative of order 50.



Figure 14.
 The preprocessing using the +ve derivative before evaluating the ZCR.

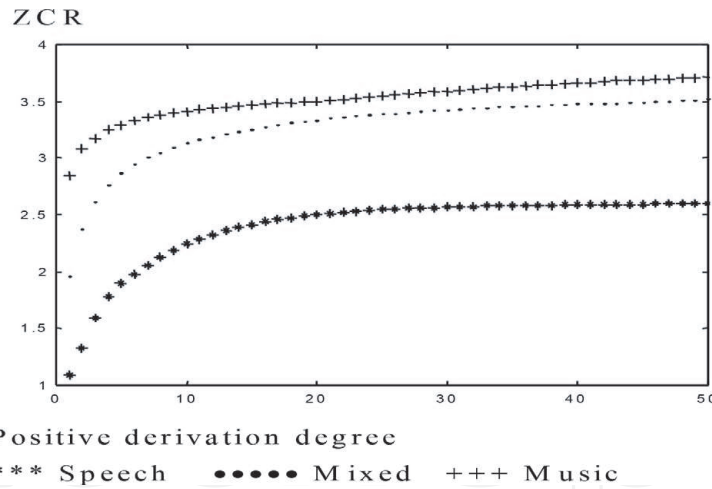


Figure 15. The average ZCR of speech, mixture, and music, after pre-processing with the +ve derivative [78].

3.1.4 Artificial neural network (ANN) approach

The ANN approach is a multipurpose technique that was used for implementing many algorithms [14, 36, 63, 79, 86–105, 110, 125], especially in classification issues [16, 49, 107–111, 119, 120, 131, 132]. A multi-layer ANN approach was used in many classification tools since it can represent nonlinear decision support systems.

3.2 Algorithms in the frequency domain

3.2.1 The spectrum approaches

3.2.1.1 Spectral flux mean and variance

This feature characterizes the change in the shape of the spectrum so it measures frame-to-frame spectral difference. Audio signals go through less frame-to-frame changes than music. The spectral flux values in audio signal is lower than that of music.

The spectral flux, sometimes called the *delta spectrum magnitude*, is defined as the *second norm* of the spectral amplitude of the difference vector and defined as in Eq. (19).

$$SF = \sqrt{\sum_k |X(k) - X(k+1)|^2} \quad (19)$$

where $X(k)$ is the signal power and k is the corresponding frequency. Another definition of the SF is also described as follow.

$$SF = \frac{1}{(N-1)(M-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{M-1} [\log(A(n, k) + \delta) - \log(A(n-1, k) + \delta)]^2 \quad (20)$$

where $A(n, k)$ in Eq. (20) is the discrete Fourier transform (DFT) of the n^{th} frame of the input signal and can be described as in Eq. (21).

$$A(n, k) = \left| \sum_{m=-\infty}^{\infty} x(m)w(nL - m)e^{j\frac{2\pi}{L}km} \right| \quad (21)$$

and $x(m)$ is the original audio data, L is the window length, M is the order of the DFT, N is the total number of frames, δ is an arbitrary constant, and $w(m)$ is the

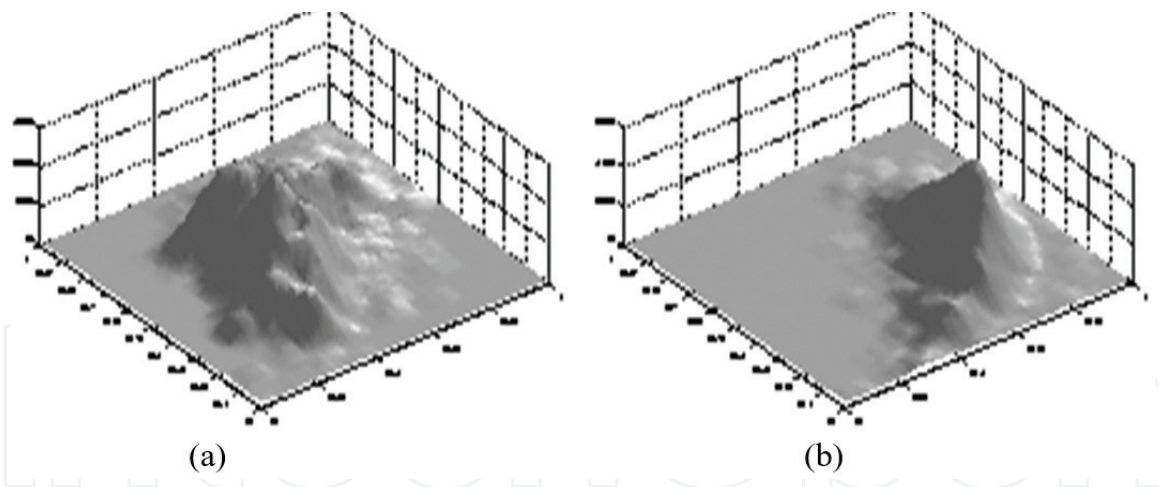


Figure 16. 3D histogram normalized features (the mean and the variance of spectral flux) of: (a) music signal, (b) audio signal [133].

	Training	Testing	Cross-validation
GMM	8.0%	8.1%	8.2%
kNN	X	6.0%	8.9%
ANN	6.7%	6.9%	11.6%

Table 5. Percentage of misclassified segments [133].

window function. Scheirer and Slaney [65] has found that *SF* feature is very useful in discriminating audio from music. **Figure 16** depicts that the variances are lower for music than for audio, and the means are less for audio than for music signal. Rossignol and others [133] have computed the means and variances of a one-second segment using frames of length 18 milliseconds.

Rossignol and others [133] have tested three classification approaches to classify the segments. They used the k-nearest-neighbors (kNN) with $k =$ seven, the Gaussian mixture model (GMM), and the ANN classifiers. **Table 5** shows their results are shown in **Table 5**, using the mean and the variance of the SF.

3.2.1.2 The mean and variance of the spectral centroid

In the frequency domain, the mean and variance of the spectral centroid feature describes the center of frequency at which most of the power in the signal is found. In audio signals, the pitches of the signals are concentrated in narrow range of low frequencies. In contrast, music signals have higher frequencies that result higher spectral means, i.e., higher spectral centroids. For a frame at time t , the spectral centroid can be evaluated as follows.

$$SC = \frac{\sum_k kX(k)}{\sum_k X(k)} \quad (22)$$

where $X(k)$ is the power of the signal at the corresponding frequency band k . When the mean and the variance of the SP are combined with the mean and the variance of the SC in Eq. (22), and the mean and the variance of the ZCR, the results of **Table 6** are found.

	Training	Testing	Cross-validation
GMM	7.9%	7.3%	22.9%
kNN	X	2.2%	5.8%
ANN	4.7%	4.6%	9.1%

Table 6.
Percentage of misclassified segments [133].

3.2.1.3 Energy at 4 Hz modulation

Audio signal has an energy peak centered on the 4 Hz syllabic rate. Therefore, a 2nd order band pass filter is used, with center frequency of 4 Hz. Although audio signals have higher energy at that 4 Hz, some music bass instruments was found to have modulation energy around this frequency [65, 133].

3.2.1.4 Roll-off point

In the frequency domain, the roll-off point feature is the value of the frequency that has 95% of the power of the signal. The value of the roll-off point can be found as follow [65, 133].

$$\sum_{k < v} X(k) = (0.95) \sum_k X(k) \quad (23)$$

where the left hand side of Eq. (23) is the sum of the power at the frequency value V , and the right hand side of Eq. (23) is the 95% of the total power of the signal of the frame, and $X(k)$ is the DFT of $x(t)$.

3.2.2 Cepstrum

The cepstrum of a signal can be defined as the inverse of the DFT of the logarithm of the spectrum of a signal. Music signals have higher cepstrum values than that of speech ones. The complex cepstrum is defined in the following Equation [122–124].

$$\hat{X}(e^{j\omega}) = \log [X(e^{j\omega})] = \log |X(e^{j\omega})| + j \arg [X(e^{j\omega})] \quad (24)$$

and then.

$$\hat{x}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}(e^{j\omega n}) d\omega \quad (25)$$

where $X(e^{j\omega})$ is the DFT of the sequence $x(n)$.

3.2.3 Summary

Table 7 summarizes the percentage error of a simulation done per each feature. Latency refers to the amount of past input data required to calculate the feature.

Scheirer and Slaney [65] have evaluated their models using 20 minutes long data sets of music and audio. Their data set consists of 80 samples, each with 15-second-long audio. They collected their samples using a 16-bit monophonic FM tuner with a sampling rate of 22.05 kHz, from a variety of stations, with different content styles

Features	The 4 Hz Mod Energy	The Low Energy	The Roll off	The Roll off Var	Spec Centroid	Spec Centroid Var	The Spec Flux	Spec Flux Var	The ZCR	The Var of the ZC Rate	The Cepstrum Resid	Cepstrum Res Var	The Pulse Metric
Latencies	1 sec	1 sec	1 frame	1 sec	1 frame	1 sec	1 frame	1 sec	1 frame	1 sec	1 frame	1 sec	5 sec
Errors	12 +/- 1.7%	14 +/- 3.6%	46 +/- 2.9%	20 +/- 6.4%	39 +/- 8.0%	14 +/- 3.7%	39 +/- 1.1%	5.9 +/- 1.9%	38 +/- 4.6%	18 +/- 4.8%	37 +/- 7.5%	22 +/- 5.7%	18 +/- 2.9%

Table 7.
 Latency and univariate discrimination performance for each feature [65].

and different noise levels, over a period of three days in the San Francisco Bay Area. They also claimed that they have audios from both male and female.

They also recorded samples of many types of music, like pop, jazz, salsa, country, classical, reggae, various sorts of rock, various non-Western styles [29, 65]. They also used several features in a spatial partitioning classifier. **Table 8** summarizes their results.

The features used in Best 8 are the plus the 4 Hz modulation, the variance features, the pulse metric, and the low-energy frame [80, 134]. In the Best 3, they used the pulse metric, the 4 Hz energy, and the variance of spectral flux. In the Fast 5, they used the 5 basic features. From results shown in **Table 8**, we conclude that it is not necessary to use all features in order to have a good classification, so in real time a good performance system may be found using only few features. A more detailed discussion can be found in [29, 65, 80, 134].

3.3 Algorithms in the time-frequency domain

3.3.1 Spectrogram (or sonogram)

The spectrogram is an example of time-frequency distribution and this method was found to be a good classical tool for analyzing audio signal [13, 19, 86, 127]. The spectrogram (or sonogram) of a signal $x(n)$ can be defined as follow.

$$X(n, \omega) = \sum_{m=-N}^N W(n+m)x(m)e^{-j\omega m} \quad (26)$$

where N is the length of the sequence $x(n)$, and $W(n)$ is a specific window.

The method of spectrogram can be used in discriminating audio from music signal, however, it may have a high percentage error. That is because it depends on the strength of the frequency in the tested samples. **Figure 17** depicts two examples of spectrograms of audio and music signals.

3.3.2 Evolutionary spectrum (ES)

The spectral representation of a stationary signal may be viewed as an infinite sum of sinusoids with random amplitudes and phases as described in Eq. (27).

$$e(n) = \int_{-\pi}^{\pi} e^{j\omega n} dZ(\omega) \quad (27)$$

where $Z(\omega)$ is the process with orthogonal increments i.e.

$$E\{dZ^*(\omega)dZ(\Omega)\} = \frac{S(\omega)d\omega}{2\pi}\delta(\omega - \Omega) \quad (28)$$

Subset	All features	Best 8	Best 3	VS Flux only	Fast 5
Audio % Error	5.8 +/- 2.1	6.2 +/- 2.2	6.7 +/- 1.9	12 +/- 2.2	33 +/- 4.7
Music % Error	7.8 +/- 6.4	7.3 +/- 6.1	4.9 +/- 3.7	15 +/- 6.4	21 +/- 6.6
Total % Error	6.8 +/- 3.5	6.7 +/- 3.3	5.8 +/- 2.1	13 +/- 3.5	27 +/- 4.6

Table 8.
Performance for various subsets of features.

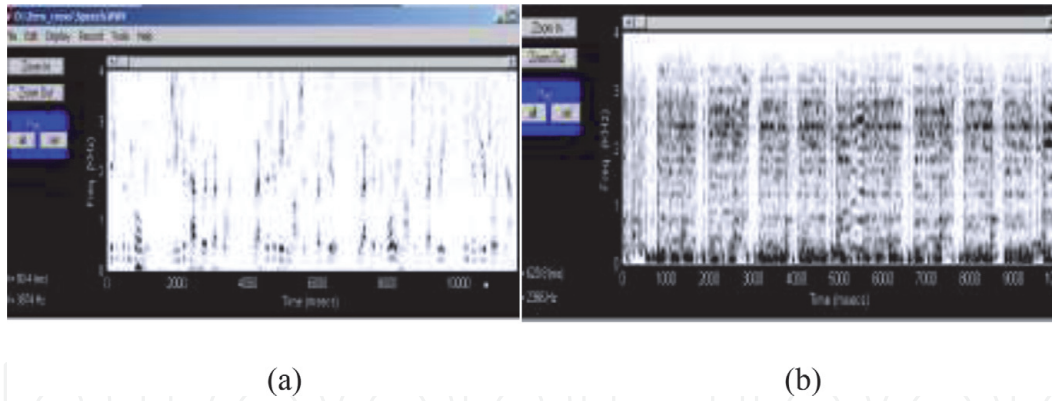


Figure 17.
 (a) Audio spectrogram, (b) music Spectrum.

and $S(\omega)$ in Eq. (28) is the spectrum of $e(n)$ [81]. Since the audio signal is, in general nonstationary, we will use the Wold-Cramer (WC) representation of a nonstationary signal. WC considers the discrete-time non-stationary process $\{x(n)\}$ as the output of a casual, linear, and time-variant (LTV) system with a white noise input $e(n)$ that has a zero-mean, unit-variance, i.e.,

$$x(n) = \sum_{m=-\infty}^n h(n, m)e(n - m) \quad (29)$$

where $h(n, m)$ is defined as the unit impulse response of an LTV system. Substituting $e(n)$ into $x(n)$ of Eq. (29) (assuming $S(\omega) = 1$ for white noise) we get.

$$x(n) = \int_{-\pi}^{\pi} H(n, \omega)e^{j\omega n} dZ(\omega) \quad (30)$$

where $H(n, \omega)$ in Eq. (30) is the time-frequency transfer function of the LTV system defined as

$$H(n, \omega) = \sum_{m=-\infty}^n h(n, m)e^{-j\omega m} \quad (31)$$

and the instantaneous power of $x(n)$ is given by

$$E\{|x(n)|^2\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(n, \omega)|^2 d\omega \quad (32)$$

and then, the Wold-Cramer ES is defined as

$$S(n, \omega) = \frac{1}{2\pi} |H(n, \omega)|^2 \quad (33)$$

The ES $S(n, \omega)$ in Eq. (33) was found to be a good classifier for the distinction of audio from music signals [81, 129]. Because of the extensive math calculation of the time-frequency spectrum, they may be very useful in off-line classification and analysis. The ESs of music and audio signals are shown in **Figure 18(a)** and **(b)**, respectively. The suppression of the amplitude for audio might be due to gaussianity.

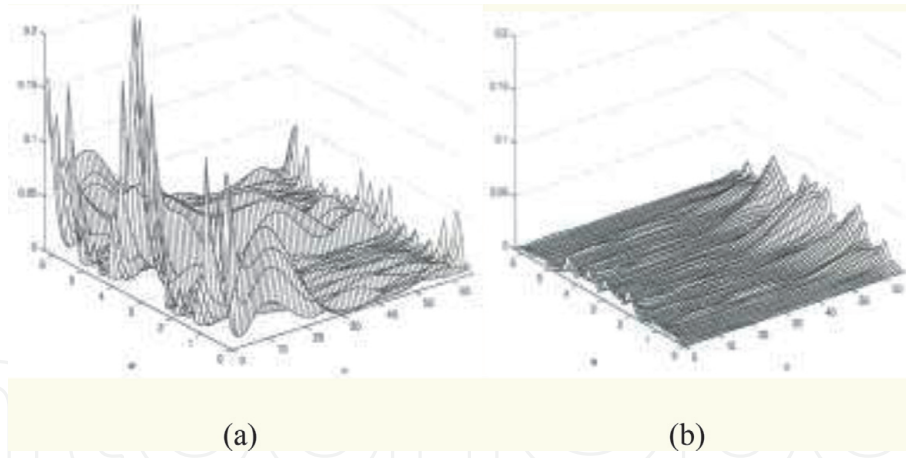


Figure 18.
(a) The ES of a music signal, (b) the ES of an audio signal [81].

4. Separation of audio and music signals

Since the separation of audio and music signals is more complicated than classification, in this section we will introduce only two approaches [7–13, 22, 76, 77, 86, 135]. The first approach is the approach of independent component analysis (ICA) with ANN. The second classifier is the pitch cancellation approach. A block diagram of a classifier integrated with a separator is depicted in **Figure 19**.

4.1 ICA with ANN separation approach

In [13, 20, 21, 127, 136], Wang and Brown proposed a model for audio segregation algorithm. His model consists of preprocessing using cochlear filtering, gammatone filtering, and correlogram forming autocorrelation function and feature extraction. The impulse response of the gammatone filters is represented as.

$$h_i(t) = t^{n-1} e^{[-2\pi b_i t] \cos(2\pi f_i t + \phi_i)} U(t) g(i), l \leq i \leq N \quad (34)$$

where n is the filter order, N is the number of channels, and U is the unit step function. Therefore, the gammatone system can be considered as a causal, time invariant system with an infinite response time. For the i^{th} channel, f_i is the center

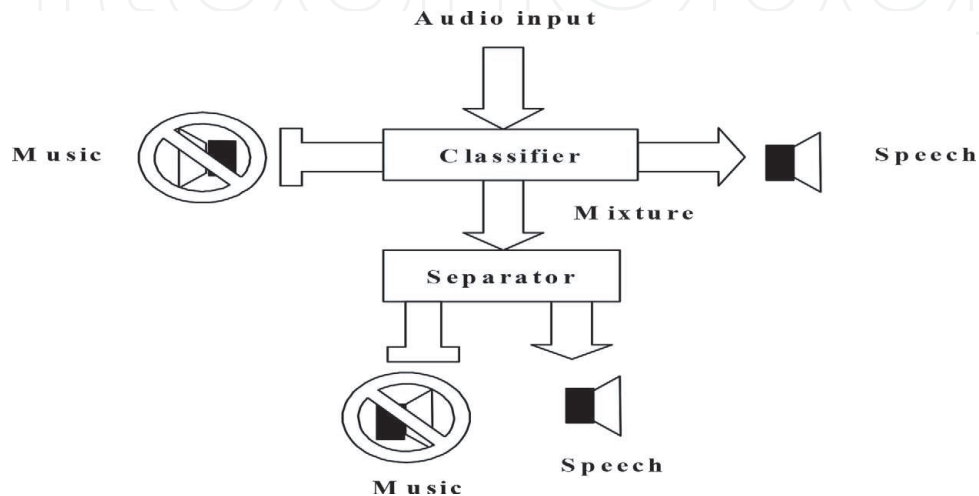


Figure 19.
A block diagram of a classifier integrated with a separator.

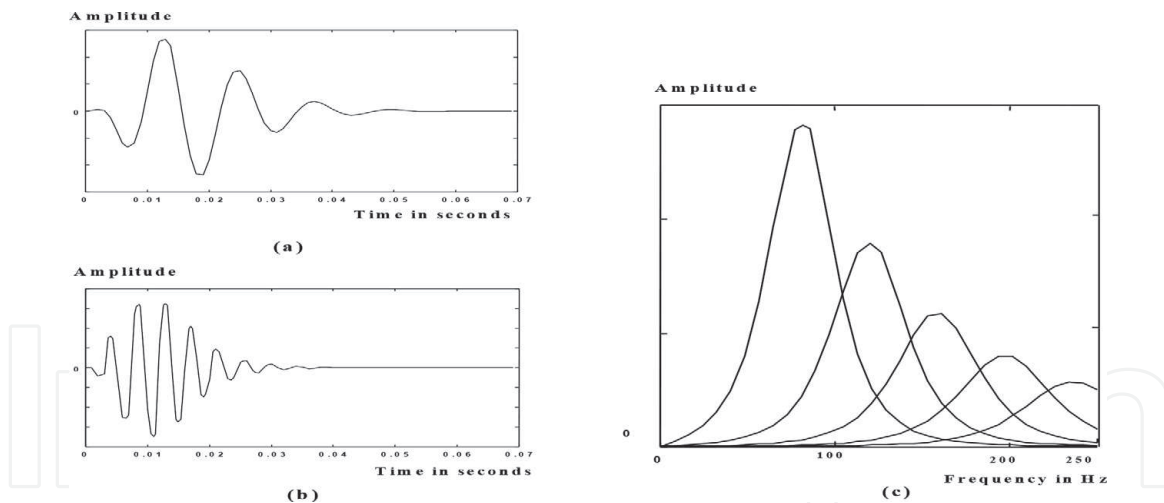


Figure 20.
 4th order impulse response Gammatone system: (a) In time domain when $i = 1$, $f_i = 80$ Hz. (b) In time domain when $i = 5$, $f_i = 244$ Hz. (c) In the frequency domain for the 1st five filters (i.e $i = 1$ to $i = 5$) with gain $g(i)$ set to unity.

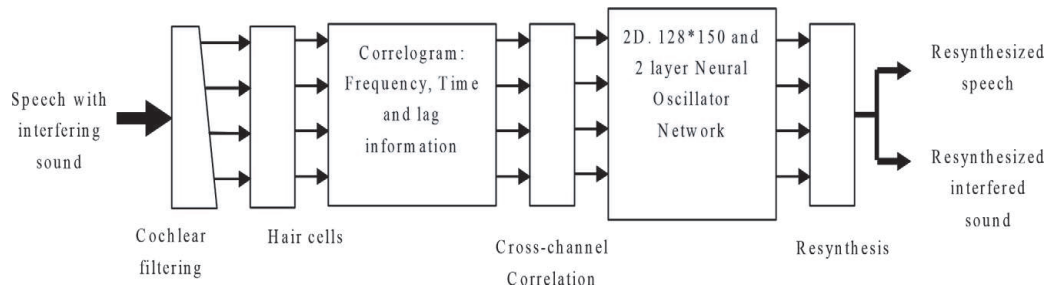


Figure 21.
 A block diagram of Wang and Brown model.

frequency of the channel, ϕ_i is the phase of the channel, b is the rate of decay of the impulse response and $g(i)$ is an equalizing gain adjust for each filter. **Figure 20** depicts the impulse response of the gammatone system, where **Figure 21** depicts the block diagram of the Wang and Brown model.

Wang and Brown model has some drawbacks. The first drawback is its complexity. Their model needs a high specification hardware to perform the calculations. In [20], Andre reported that Wang and Brown model needs to be improved. The ICA method can be used for separation if two sources of mixture are available assuming that the two signals from the two different sources are statistically independent [66, 74, 75, 121, 137]. In [19], Takigawa tried to improve the performance of W & B model. He used the short time Fourier transform (STFT) in the input stage and used the spectrogram values instead of correlogram, however, they have not reported the amount of improvement. A similar work for separating the voiced audio of two talkers speaking simultaneously at similar intensities in a single channel, using pitch peak canceling in cepstrum domain, was done by Stubbs [8].

4.2 The pitch cancelation

The pitch cancelation method is widely used in noise reduction. A good try to separate two talkers speaking simultaneously at similar intensities in a single channel, or by other words, separation of two talkers without any restriction was introduced by Stubbs [8]. For a certain person, the letters A and R have lot of consonant. These consonants, in the frequency domain, have low amplitudes, however, they appear as long pitch peak in the cepstrum domain. If these consonants are deleted

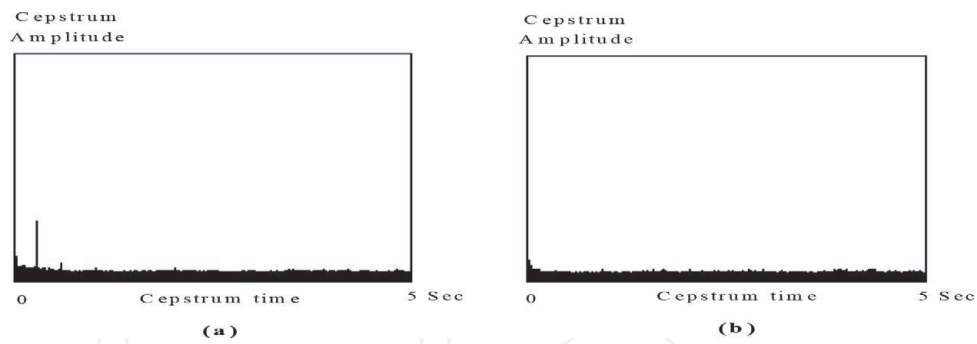


Figure 22. (a) A typical 5 seconds audio signal in cepstrum domain, the pitch peak appears near zero. (b) a typical 5 seconds music signal in cepstrum domain.

by replacing the five-cepstral samples centered at the pitch peak by zeros, the audio segment may be attenuated or distorted completely. A typical example of the cepstrum of two audio and music signals is depicted in **Figure 22** for 5 seconds signals. The logarithmic effect will increase low amplitude reduce high one, and the values near zero will be very large after the logarithm.

5. Conclusions

In this chapter, a general review of the common classification and separation algorithms used for speech and music was presented and some were introduced and discussed thoroughly. The approaches dealt with classification were divided into three categories. The first category included most of the real-time approaches. In the real-time approaches, we introduced the ZCR, the STE, the ZCR and the STE with positive derivative, with some of their modified versions, and the neural networks. The second category included most of the frequency domain approaches such as the spectral centroid and its variance, the spectral flux and its variance, the roll-off of the spectrum, the cepstral residual, and the delta pitch. However, the last category introduced two time-frequency approaches, mainly the spectrogram and the evolutionary spectrum. It has been noticed that the time-frequency classifiers provided an excellent and a robust discrimination result in discriminating speech from music signals in digital audio. Depending on the application, the decision of which feature should be chosen is selected. The algorithms of the first category are faster since the processing is made in the real time; however, those of the second

Approaches	Time domain	Frequency domain (Spectrum) (Cepstrum)	Time-Frequency domain	
Algorithms	ZCR	Spectral Centroid	Cepstral Residual	Spectrogram (Sonogram)
	STE	Spectral Flux	Variance of the Cepstral Residual	Evolutionary Spectrum
	Roll-Off Variance	Spectrum Roll-Off	Cepstral feature	Evolutionary Bispectrum
	Pulse Metric	Signal Bandwidth	Pitch	
	Number of Silence	Spectrum Amplitude	Delta Pitch	
	HMM	Delta Amplitude		
	ANN			

Table 9. Summary of the classification and separation algorithms.

one are more precise. The time-frequency approaches has not been discussed thoroughly in literature and they still need more research and elaboration. Lastly, we may conclude that many classification algorithms were proposed in literature, however, few ones were proposed for separation. The algorithms introduced in this chapter can be summarized in **Table 9**.

IntechOpen

IntechOpen

Author details

Abdullah I. Al-Shoshan
Computer Engineering, Qassim University, Saudi Arabia

*Address all correspondence to: ashoshan@qu.edu.sa

IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Al-Shoshan A. I. Speech and Music Classification and Separation: A Review. *Journal of King Saud University-Engineering Sciences*. 2006; 19(1): 95–133. doi:10.1016/S1018-3639(18)30850-X
- [2] Martin, K. Towards Automatic Sound Source Recognition: Identifying Musical Instruments. In: *Proceedings of NATO Computational Hearing Advanced Study Institute, Italy, July 1998*.
- [3] Herrera-Boyer P., Amatriain X., Batlle E., Serra X. Towards Instrument Segmentation for Music Content Description: a Critical Review of Instrument Classification Techniques. In: *Proceedings of International Society for Music Information Retrieval (ISMIR), 2000*.
- [4] Gjerdingen R.O. Using Connectionist Models to Explore Complex Musical Patterns. *Computer Music Journal*. 1989; 13(3):67–75. DOI: 10.2307/3680013
- [5] Hörnel D., Menzel W. Learning Musical Structure and Style with Neural Networks. *Computer Music Journal*. 1998; 22(4):44–62. doi:10.2307/3680893
- [6] Leman, M., Van Renterghem P. Transputer Implementation of the Kohonen Feature Map for a Music Recognition Task. In: *Proceedings of the Second International Transputer Conference; Antwerpen: BIRA, 1989*. pp. 1–20
- [7] Al-Atiyah A. Music and Speech Separation [thesis]. King Saud University; 2002.
- [8] Stubbs R., Summerfield Q. Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms. *J. Acoustical Society of America*. 1991; 89:1383–1393. DOI: 10.1121/1.400539
- [9] Lee T-W., Koehler B-U. Blind Source Separation of Nonlinear Mixing Modes. In: *Proceedings of the IEEE Signal Processing Society Workshop (1997)*. Amelia Island, FL. USA: IEEE; 1997. pp. 406–415. doi: 10.1109/NNSP.1997.622422
- [10] Lee T-W., Orglmeister R. A Contextual Blind Separation of Delayed and Convolved Sources. *IEEE ICASSP'97; (1997)*. pp. 1199–1202. DOI: 10.1109/ICASSP.1997.596159
- [11] Lee T-W., Bell A., Lambert R. Blind Separation of Convolved and Delayed Sources. *Advance in Neural Information Processing System*. MIT Press. 1997.
- [12] Lee T-W., Bell A. J., Orglmeister R. Blind Source Separation of Real Word Signals. *IEEE ICNN*. Houston, USA; (1997). 2129–2134. DOI: 10.1109/ICNN.1997.614235
- [13] Wang D. L., Brown G. J. Separation of Speech From Interfering Sounds Based on Oscillatory Correlation. *IEEE Transaction on Neural Networks*. Vol. 10: No. 3. (May 1999), 684–697. DOI: 10.1109/72.761727
- [14] Leman M. The Theory of Tone Semantics: Concept, Foundation, and Application, *Minds and Machines*. 2(4): (1992); pp. 345–363. doi.org/10.1007/BF00419418
- [15] Patel A.D., Gibson E., Ratner J., Besson M., Holcomb P.J. Processing Grammatical Relations in Music and Language: An Event-Related Potential (ERP) Study. *Proceedings of the Fourth International Conference on Music Perception and Cognition*. Montreal: McGill University. (1996). 337–342.
- [16] Stevens C., Latimer C. A Comparison of Connectionist Models of Music Recognition and Human

Performance. *Minds and Machines*, 2 (4): (1992); pp. 379–400.

[17] Weigend A.S. Connectionism for Music and Audition. In J. Cowan, G. Tesauro & J. Alspecter (Eds.), *Advances in Neural Information Processing Systems 6*, San Francisco: Morgan Kaufmann. (1994); pp. 1163–1164.

[18] Anagnostopoulou C., Westermann G. Classification in Music: A Computational Model for Paradigmatic Analysis. *Proceedings of the International Computer Music Conference*, San Francisco, (1997), 125–128.

[19] Takigawa I., Toyama J., Shimbo M. A Modified LEGION using a spectrogram for speech segregation. *IEEE*. (1999); I526-I531. DOI: 10.1109/ICSMC.1999.814147

[20] Andre J. W., Kouwe V. D., Wang D., Brown G. J. A Comparison of Auditory and blind Separation Techniques for speech segregation. *IEEE transaction on speech and audio processing*. 9(3): (March 2001); pp. 189–195. DOI: 10.1109/89.905993

[21] Wang D. L., Brown G. J. Speech Segregation on Sound Localization. *IEEE*, (2001), 2861–2866.

[22] Belouchrani A., Aben-Meraim K., Cardoso J. F., Moulines E. A Blind Source Separation Technique Using Second Order Statistics. *IEEE Trans. Signal processing*, vol. 45, (Feb. 1997), pp. 434–444. DOI: 10.1109/78.554307

[23] Govindarajan K.K., Grossberg S., Wyse L.L., Cohen M.A. A Neural Network Model of Auditory Scene Analysis and Source Segregation. *Technical Report CAS/CNS-TR-94-039*, Boston University, Dept. of Cognitive and Neural Systems, 1994.

[24] Kahrs M., Brandenburg K. Application of digital signal processing

to audio and acoustics. Kluwer Academic Puplicher, Bosten/ Dordrecht/ London, 1998.

[25] Backus J. *The Acoustical Foundations of Music*. 2nd, W. W. Norton & Company, 1977.

[26] Gang D., Lehmann D., Wagner N. Harmonizing Melodies in Real-Time: The Connectionist Approach. *Proceedings of the International Computer Music Conference*, San Francisco, (1997), pp. 27–31.

[27] Kaipainen M., Toiviainen P., Louhivuori J. A Self-Organizing Map that Recognizes and Generates Melodies. In P. Pylkkänen & P. Pylkkö (Eds.), *New Directions in Cognitive Science*, (1995), 286–315.

[28] Port R., Anderson S. Recognition of Melody Fragments in Continuously Performed Music. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum Associates, (1989), 820–827.

[29] Toiviainen P. Modeling the Target-Note Technique of Bebop-Style Jazz Improvisation: An Artificial Neural Network Approach. *Music Perception*, 12(4), (1995), 399–413.

[30] Cook, N. *A Guide to Musical Analysis*. Oxford University Press, 1987.

[31] Roy, D. and Malamud, C. Speaker Identification Based Text to Audio Alignment for an Audio Retrieval System. *IEEE ICASSP'97*, vol. 2, Munich, Germany, (April 1997), 1099–1102.

[32] Beigi H., Maes S., Sorensen J., Chaudhari U. A Hierarchical Approach to Large-Scale Speaker Recognition. *IEEE ICASSP'99*, Phoenix, Arizona, March 1999.

[33] Rabiner L., Juang B. H. *Fundamentals of Speech Recognition*.

Prentice-Hall, Englewood Cliffs, NJ, 1993.

[34] Kedem B. Spectral Analysis and Discrimination by Zero-Crossings. *Proceedings of IEEE*, Vol. 74, NO. 11, (Nov. 1986), 1477–1492.

[35] Bateman W. *Introduction to Computer Music*. John Wiley&sons, 1984.

[36] Fedor P. Principles of the Design of D-Neuronal Networks I: Net Representation for Computer Simulation of a Melody Compositional Process. *International Journal of Neural Systems*, 3(1), (1992), 65–73.

[37] Horner A., Goldberg D.E. Genetic Algorithms and Computer-Assisted Music Composition. In B. Alphonse & B. Pennycock (Eds.), *Proceedings of the 1991 International Computer Music Conference*, San Francisco, (1991), 479–482.

[38] McIlwain P. The Yuri Program: Computer Generated Music for Multi-Speaker Sound Systems. *Proceedings of the ACMA Conference*, Melbourne, Australia, (1995), 150–151.

[39] Ainsworth W. A. *Speech Recognition by Machine*. Peter Peregrinus Ltd., 1988.

[40] Muthusamy Y. K., Barnard E., Cole R. A. Reviewing Automatic Language Identification. *IEEE Signal Processing Magazine*, (October 1994), 33–41.

[41] Ladefoged P. *Elements of Acoustic Phonetics*. University of Chicago Press, 1962.

[42] Fry D. B. *The Physics of Speech*. Cambridge University Press, 1979.

[43] Beck D.L., Callaway S.L. Breakthroughs in signal processing and feedback reduction lead to better speech

understanding. *Hearing Review*. 2019; 26(4) [Apr]:30–31.

[44] Linster C. Rhythm Analysis with Backpropagation. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulie & L. Steels (Eds.), *Connectionism in Perspective*, North-Holland: Elsevier Science Publishers B.V, (1989), 385–393.

[45] Jakobsson M. Machine-Generated Music with Themes. *Proceedings of the International Conference on Artificial Neural Networks (Vol. 2)* Amsterdam: Elsevier, (1992), 1645–1646.

[46] Griffith N.J.L. Connectionist Visualization of Tonal Structure. *AI Review*, 8, (1995), 393–408.

[47] Stevens C., Wiles J. Representations of Tonal Music: A Case Study in the Development of Temporal Relationships. In M.C. Mozer, P. Smolensky, D.S. Touretzky, J.E. Elman & A.S. Weigend (Eds.), *Proceedings of the Connectionist Models Summer School*, Hillsdale, NJ: Erlbaum, (1993), 228–235.

[48] Young P. H. *Electrical Communication Techniques*. 2nd, MERRILL, 1990.

[49] Laine P. Generating Musical Patterns Using Mutually Inhibited Artificial Neurons. *Proceedings of the International Computer Music Conference*, San Francisco, (1997), 422–425.

[50] Leman M. Symbolic and Subsymbolic Description of Music. In G. Haus (Ed.), *Music Processing*, New York: Oxford University Press, (1993), 119–164.

[51] Lischka C. Understanding Music Cognition: A Connectionist View. In G. De Poli, A. Piccialli & C. Roads (Eds.), *Representations of Musical Signals*, Cambridge, MA: MIT Press, (1991), 417–445.

- [52] Griffith N., Todd P. M. *Musical Networks*. Bradford Books The MIT Press, 1999.
- [53] Pierce J. R. *The Science of Musical Sound*. 3rd Ed., W.H. Freeman and company, 1996.
- [54] Lerdahl F. and Jackendoff, R., *A Generative Theory of Tonal Music*. MIT Press, Cambridge, 1983.
- [55] Monelle R. *Linguistics and Semiotics in Music*. Harwood Academic Publishers, 1992.
- [56] Gang D., Berger J. Modeling the Degree of Realized Expectation in Functional Tonal Music: A Study of Perceptual and Cognitive Modeling Using Neural Networks. In D. Rossiter (Ed.), *Proceedings of the International Computer Music Conference*, San Francisco, (1996), 454–457.
- [57] Bharucha J. Tonality and Expectation. In R. Aiello (Ed.), *Musical Perceptions*, New York: Oxford University Press, (1994), 213–239.
- [58] Feiten B., Ungvary T. “Organizing Sounds with Neural Nets”, *Int. Computer Music Conference*, San Francisco, (1991), 441–443.
- [59] Foote J. T. Content-Based Retrieval of Music and Audio. *SPIE’97*, (1997), 138–147.
- [60] Saunders J. Real-Time Discrimination of Broadcast Speech/Music. *IEEE ICASSP’96*, (1996), 993–996.
- [61] El-maleh K., Samoulian A., Kabal P. Frame-Level Noise Classification in Mobile Environment. *proc. IEEE Int. Conf. On Acoustics, Speech, Signal processing*, Phoenix, Arizona, (March 1999), 237–240.
- [62] El-maleh K., Klein M., Petrucci G., Kabal P. *Speech/Music Discriminator for Multimedia Application*. *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Istanbul)*, (June 2000), 2445–2448.
- [63] Benyamin M., Miriam F. Neural Network Based Model for Classification of Music Type. *IEEE Cat.*, No. 95, (1995), 640–645.
- [64] Hoyt J. D., Wechsler H. Detection of Human Speech Using Hybrid Recognition Models. *IEEE*, (1994), 330–333.
- [65] Scheirer E., Slaney M. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. *Proceedings of the 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP97)*, Munich, Germany, April 1997.
- [66] Chien J-T. *Source Separation and Machine Learning*. Elsevier Inc. 2019. <https://doi.org/10.1016/C2015-0-02300-0>
- [67] Pope S. T., Holm F., Kouznetsov A. Feature extraction and database design for music software. *Proceedings of the International Computer Music Conference*, (2004), 596–603.
- [68] McKay C., Fujinaga I. Automatic genre classification using large high-level musical feature sets, *Proceedings of the International Conference on Music Information Retrieval*, (2004), 525–530.
- [69] Essed S., Richard G., David B. Musical instrument recognition based on class pairwise feature selection. *Proceedings of the International Conference on Music Information Retrieval*, (2004), 560–568.
- [70] Downie J. The scientific evaluation of music information retrieval systems: Foundations and future. *Computer Music Journal*, 28, 2, (2004), 12–33.
- [71] West K., Cox, S. Finding an Optimal Segmentation for Audio Genre

Classification. Proceedings of the 6th Int. Symposium on Music Information Retrieval, University of London, (2005), 680–685.

[72] Tzanetaki G. Music Information Retrieval. ICASSP2005, Tutorial TUT-5, Philadelphia, 2005.

[73] West C., Cox S. Features and classifiers for the automatic classification of musical audio signals. Proceedings of the International Conference on Music Information Retrieval, (2004), 531–537.

[74] Yang X-S. Introduction to Algorithms for Data Mining and Machine Learning. Elsevier Inc. 2019. <https://doi.org/10.1016/C2018-0-02034-4>

[75] Kotu V. Data Science: Concepts and Practice. Elsevier Inc. 2019. <https://doi.org/10.1016/C2017-0-02113-4>

[76] Mimitakis S. I., Drossos K., Cano E. Schuller G. Examining the Mapping Functions of Denoising Autoencoders in Singing Voice Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. vol. 28, pp. 266–278, 2020. DOI: 10.1109/TASLP.2019.2952013

[77] Sharma G., Umapathy K., Krishnan S. Trends in audio signal feature extraction methods. Applied Acoustics. Elsevier Ltd. Volume 158, 15 January 2020. <https://doi.org/10.1016/j.apacoust.2019.107020>

[78] Al-Shoshan A., Al-Atiyah A., Al-Mashouq K. A Three-Level Speech, Music, and Mixture Classifier. Journal of King Saud University {Engineering Sciences (No. 2)}, Volume 16, (1424), 319–332. [https://doi.org/10.1016/S1018-3639\(18\)30794-3](https://doi.org/10.1016/S1018-3639(18)30794-3)

[79] Al-Shoshan, A.I., “A Classification of Music, Speech and Mixture Signals Via Fuzzy Logic,” The 28th International Conference on Computers

and Their Applications, (CATA-2013), Honolulu, Hawaii, USA, pp. 117–122, March 4–6, 2013.

[80] Berger J., Gang D. A Neural Network Model of Metric Perception and Cognition in the Audition of Functional Tonal Music. Proceedings of the 1997 International Computer Music Conference, San Francisco, (1997), 23–26.

[81] Al-Shoshan, A.I., “Audio Signal Discrimination Using Evolutionary Spectrum,” International Journal of Computers and Applications, Volume 31, No. 2, pp. 69–73, 2009.

[82] Toivainen P., Kaipainen M., Louhivuori J. Musical Timbre: Similarity Ratings Correlate with Computational Feature Space Distances. Journal of New Music Research, 24(3), (1995), 282–298.

[83] Jin H., Kubala F., Schwartz R. Automatic Speaker Clustering. Proc. of the Speech Recognition Workshop, (1997), 108–111.

[84] Meddis R., Hewitt M. Modeling the Identification of Concurrent Vowels with Different Fundamental Frequency. J. Acoust. Soc. Am., vol. 91, (1992), 233–245.

[85] Raphael C. “Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.21, No. 4, April 1999.

[86] Hyvarinen, A. and Oja, E. Independent Component Analysis: Algorithms and Applications. Int. J. of Neural Networks, April 1999.

[87] Akarte N.J. Music Composition Using Neural Networks. Master’s thesis, University of Nevada, Reno, 1992.

[88] Barnard E., Cole R.A., Vea M.P., Alleva F.A. Pitch Detection with a Neural-Net Classifier. IEEE

Transactions on Signal Processing, 39 (2), (1991), 298–307.

[89] Bellgard M.I., Tsang C.P. Harmonizing Music Using a Network of Boltzmann Machines. In Proceedings of the Fifth Annual Conference of Artificial Neural Networks and Their Applications (Neuro-Nimes), (1992), 321–332.

[90] Bellgard M.I., Tsang C.P. “On the use of an Effective Boltzmann Machine for Musical Style Recognition and Harmonization”, Proceedings of the International Computer Music Conference, San Francisco, (1996), 461–464.

[91] Berger J., Gang D. Modeling Musical Expectations: A Neural Network Model of Dynamic Changes of Expectation in the Audition of Functional Tonal Music. Proceedings of the Fourth International Conference on Music Perception and Cognition, Montreal: McGill University, (1996), 373–378.

[92] Bharucha J. Neural Net Modeling of Music. Proceedings of the First Workshop on Artificial Intelligence and Music, Menlo Park, CA, (1988), 173–182.

[93] Bharucha J. Neural Networks and Perceptual Learning of Tonal Expectancies. Proceedings of the First International Conference on Music Perception and Cognition, Kyoto: Kyoto City University of Arts, (1989), 81–86.

[94] Bharucha J. Pitch, Harmony and Neural Nets: A Psychological Perspective. In P.M. Todd & D.G. Loy (Eds.), *Music and Connectionism*, Cambridge, MA: MIT Press, (1991), 84–99.

[95] Bharucha J., Olney K.L. Tonal Cognition, Artificial Intelligence and Neural Nets. *Contemporary Music Review*, Vol. 4, (1989), 341–356.

[96] Bresin R., Vedovetto A. Neural Networks for Musical Tones

Compression, Control, and Synthesis. In Proceedings of the International Computer Music Conference, San Francisco, (1994), 368–371.

[97] Bresin R., Vedovetto A. Neural Networks for the Compression of Musical Tones and for the Control of Their Resynthesis. Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, 1994.

[98] Carpinteiro O. A Neural Model to Segment Musical Pieces. In E. Miranda (Ed.), Proceedings of the Second Brazilian Symposium on Computer Music, Fifteenth Congress of the Brazilian Computer Society, (1995), 114–120.

[99] Ciaccia P., Lugli F., Maio D. Using Neural Networks to Perform Harmonic Analysis in Music. The Fifth Italian Workshop on Neural Nets, WIRN VIETRI-92, Singapore, (1992), 273–279.

[100] Cosi P., DePoli G., Lauzzana G. Auditory Modeling and Self-Organizing Neural Networks for Timbre Classification. *Journal of New Music Research*, 23(1), (1994), 71–98.

[101] Fedor P. Principles of the Design of D-Neuronal Networks II: Composing Simple Melodies. *International Journal of Neural Systems*, 3(1), (1992), 75–82.

[102] Feiten B., Guenzel S. Automatic Indexing of a Sound Data Base Using Self-Organizing Neural Nets. *Computer Music Journal*, 18(3), (1994), 53–65.

[103] Feulner J. Learning the Harmonies of Western Tonal Music Using Neural Networks. Proceedings of the International Symposium on Computer and Information Sciences VII, Paris: EHEI Press, (1992), 303–307.

[104] Feulner J. Neural Networks that Learn and Reproduce Various Styles of Harmonization. Proceedings of the

International Computer Music Conference, San Francisco, (1993), 236–239.

[105] Gang D., Lehmann D. An Artificial Neural Net for Harmonizing Melodies. Proceedings of the International Computer Music Conference, San Francisco, (1995), 440–447.

[106] Gjerdingen R.O. Categorization of Musical Patterns by Self-Organizing Neuronlike Networks. *Music Perception*, 7(4), (1990), 339–370.

[107] Laden B. A Parallel Learning Model for Pitch Perception. *Journal of New Music Research*, 23(2), (1994), 133–144.

[108] Laden B., Keefe B.H. The Representation of Pitch in a Neural Net Model of Pitch Classification. *Computer Music Journal*, 13(4), (1989), 12–26. Also in P.M. Todd & D.G. Loy (Eds.), *Music and Connectionism*, Cambridge, MA: MIT Press, (1991), 64–78.

[109] Leman M. Artificial Neural Networks in Music Research. In A. Marsden & A. Pople (Eds.), *Computer Representations and Models in Music*, London: Academic Press, (1991), 265–301.

[110] Mencl W.E. Effects of Tuning Sharpness on Tone Categorization by Self-Organizing Neural Networks. Proceedings of the Fourth International Conference on Music Perception and Cognition, Montreal: McGill University, (1996), 217–218.

[111] Mourjopoulos J.N., Tsoukalas D.E. Neural Network Mapping to Subjective Spectra of Music Sounds. *Journal of the Audio Engineering Society*, 40(4), (1992), 253–259.

[112] Cohen M.A., Grossberg S., Wyse L.L. A Spectral Network Model of Pitch Perception. *Journal of the Acoustical Society of America*, 498(2), (1995), 862–879. <https://doi.org/10.1121/1.413512>.

[113] Ohya K. A Sound Synthesis by Recurrent Neural Network. In E. Michie (Ed.), *Proceedings of the International Computer Music Conference*, San Francisco, (1995), 420–423.

[114] Palmieri F. Learning Binaural Sound Localization through a Neural Network. Proceedings of the IEEE Seventeenth Annual Northeast Bioengineering Conference, (1991), 13–14.

[115] Röbel A. Neural Networks for Modeling Time Series of Musical Instruments. In E. Michie (Ed.), *Proceedings of the International Computer Music Conference*, San Francisco, (1995), 424–428.

[116] Röbel A. Neural Network Modeling of Speech and Music Signals. In M.C. Mozer, M.I. Jordan & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9*, Cambridge, MA: MIT Press, 1997.

[117] Sano H., Jenkins K.B. A Neural Network Model for Pitch Perception. *Computer Music Journal*, 13(3), (1989), 41–48. Also in P.M. Todd & D.G. Loy (Eds.), *Music and Connectionism*, Cambridge, MA: MIT Press, (1991), 42–49.

[118] Taylor I. Artificial Neural Network Types for the Determination of Musical Pitch. Unpublished doctoral thesis, University of Wales, College of Cardiff, Dept. of Physics, 1994.

[119] Taylor I. J. Greenhough, M. Neural Network Pitch Tracking Over the Pitch Continuum. In E. Michie (Ed.), *Proceedings of the International Computer Music Conference*, San Francisco, (1995), 432–435.

[120] Trubitt D.R., Todd P.M. The Computer Musician: Neural Networks and Computer Music. *Electronic Musician*, 7(1), (1991), 20–24.

[121] Walpole R. E., Myers R. H. *Probability and Statistics for Engineer*

and Scientists. 5th Ed., Macmillan Publishing, 1993.

[122] Bogert B. P., Healy M. J. R., Tukey J. W. The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-Cepstrum, and Saphe Cracking. John Wiley and Sons, New York, (1963), 209–243.

[123] Eronen A., Klapuri A. Musical Instrument Recognition using cepstral coefficients and temporal features", Proc. ICASSP 2000.

[124] Cosi P., DePoli G., Prandoni P. "Timbre characterization with mel-cepstrum and neural nets", Proceedings of the International Computer Music Conference, (1994), 42–45.

[125] Griffith N. J. L. Modeling the Influence of Pitch Duration on the Induction of Tonality from Pitch-use. Proceedings of the International Computer Music Conference, San Francisco, (1994), 35–37.

[126] Taylor I., Greenhough M. An Object Oriented ARTMAP System for Classifying Pitch. Proceedings of the International Computer Music Conference, San Francisco, (1993), 244–247.

[127] Mu G., Wang D. L. An Extended Model for Speech Segregation. Proceeding of IEEE, (2001), 1089–1094.

[128] Priestley M. B. Non-linear and Non-stationary Time Series Analysis. New York, NY:Academic Press, 1988.

[129] Al-Shoshan A.I. LTV System Identification Using the Time-Varying Autocorrelation Function and Application to Audio Signal Discrimination. ICSP02, Beijing, China, 2002. DOI: 10.1109/ICOSP.2002.1181036

[130] Scarborough D.L., Miller B.O., Jones, J.A. Connectionist Models for

Tonal Analysis. Computer Music Journal, 13(3), (1989), 49–55. Also in P. M. Todd & D.G. Loy (Eds.), Music and Connectionism, Cambridge, MA: MIT Press, (1991), 54–60.

[131] Shuttleworth T., Wilson R., A Neural Network for Triad Classification. In E. Michie (Ed.), Proceedings of the International Computer Music Conference, San Francisco, (1995), 428–431.

[132] Sergent J. Mapping the Musician Brain, Human Brain Mapping, (1993), 20–38.

[133] Rossignol S., Rodet X., Soumagne J., Collette L., Depalle P. Feature extraction and temporal segmentation of acoustic signals. Proceedings of the International Computer Music Conference, 1998.

[134] Scarborough D.L., Miller B.O., Jones J.A. On the Perception of Meter. In M. Balaban, K. Ebcioğlu & O. Laske (Eds.), Understanding Music with AI: Perspectives in Music Cognition, Cambridge, MA: MIT Press, (1992), 427–447.

[135] Magron P., Virtanen T. Online Spectrogram Inversion for Low-Latency Audio Source Separation. IEEE Signal Processing Letters. vol. 27, pp. 306–310, 2020. DOI: 10.1109/LSP.2020.2970310

[136] Wang D. L. Primitive Auditory Segregation Based on Oscillator Correlation. Cognit. Sci., vol. 20, (1996), 409–456.

[137] Israr M., Khan M. S., Khan K. Speech Sources Separation Based on Models of Interaural Parameters and Spatial Properties of Room. International Journal of Engineering Works. 7(1): January 2020; pp. 22–26. <https://doi.org/10.34259/ijew.20.7012226>