J. King Saud Univ., Vol. 19, *Eng. Sci.* (1), pp. 95-133, Riyadh (1427H./2006)

Speech and Music Classification and Separation: A Review

Abdullah I. Al-Shoshan
Department of Computer Science, College of Computer,
Qassim University, Saudi Arabia
shoshan@myway.com

(Received 04 August 2003; accepted for publication 16 April 2006)

Abstract. The classification and separation of speech and music signals have attracted attention by many researchers. The purpose of the classification process is needed to build two different libraries: speech library and music library, from a stream of sounds. However, the separation process is needed in a cocktail-party problem to separate speech from music and remove the undesired one. In this paper, a review of the existing classification and separation algorithms is presented and discussed. The classification algorithms will be divided into three categories: time-domain, frequency-domain, and time-frequency domain approaches. The time-domain approaches used in literature are: the zero-crossing rate (ZCR), the short-time energy (STE), the ZCR and the STE with positive derivative, with some of their modified versions, the variance of the roll-off, and the neural networks. The frequency-domain approaches are mainly based on: spectral centroid, variance of the spectral centroid, spectral flux, variance of the spectral flux, roll-off of the spectrum, cepstral residual, and the delta pitch. The time-frequency domain approaches have not been yet tested thoroughly in literature; so, the spectrogram and the evolutionary spectrum will be introduced. Also, some new algorithms dealing with music and speech separation and segregation processes will be presented.

1. Introduction

The problem of distinguishing speech from music has become increasingly important as automatic speech recognition (ASR) systems and it has been applied to more and more "real-world" multimedia domains [1-6]. If we wish to build systems that perform ASR on soundtrack data, for example, it is important to be able to distinguish which segments of the soundtrack contain speech [7]. Humans can separate speech from music easily in their mind without any influence of the mixed music [8-23]. Due to the new techniques of analysis and synthesis of speech signals, the musical signal processing has gained particular weight [16, 24], and therefore, the classical sound analysis techniques are used in processing music signals. Music art has a long and distinguished history. It goes back to the time of Greek and is developed through centuries in both the musical instruments and melodies [25-28]. There are many kinds of music such as: Classical,

Rock, Pop, Disco, Jazz, Country, Latin, Electronic, Arabic, etc. [29]. Figure 1 shows the hierarchy of sound type signals [30].

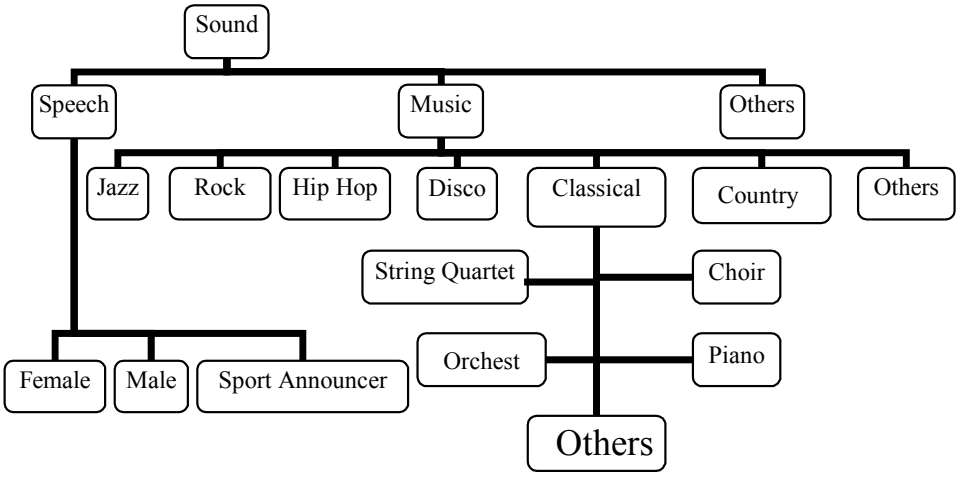


Fig. 1. The hierarchy of sound.

Audio signals change continuously and non-deterministically with time. Consequently, they are usually characterized as time averages, and their relative amplitude and frequency contents can be easily specified. As an example, speech and music typically have strong low-frequency energy and progressively weaker high-frequency content [31, 32]. Hence, generalized time and frequency audio signal spectra plots might look like those in Fig. 2 [33]. The frequency f_{max} varies according to audio signal kind; f_{max} equals 4 kHz in telephone transmitting quality, 5 kHz in mono-loudspeaker recording, 6 kHz in stereo or multi-loudspeaker recording, 11 kHz in FM stereo broadcasting, and 22 kHz in CD quality recording.

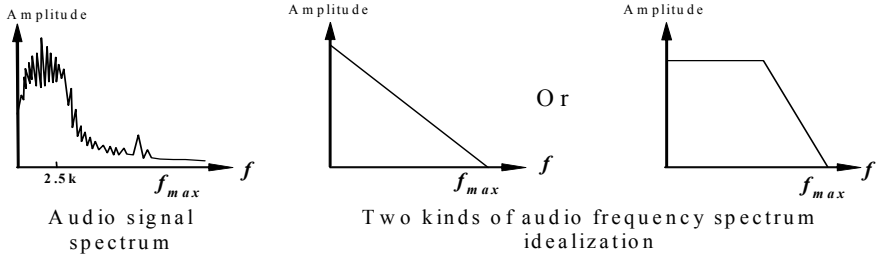


Fig. 2. Generalized frequency spectrum for audio signal.

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

1) Speech signal compounded of single talker in specific time period.
2) Completely music signal without any speech component.
3) Mixture of single talker speech and background music.
4) Songs; mixture of music with a singer voice.
5) Singing without music.
6) Abnormal music; uses acclaim cadence, single word cadence, human whistle sound, opposite reverberation or any non-music sound that has been inserted as a basic tone of the music melody. These cadences cannot be generated by any of the ordinary musical instrument except modern Organ and mainly processed by a major help of computers [4, 34-37].
7) Speech signal compounded of two or more speakers talking simultaneously in a specific time period. A good algorithm for separating the sounds of two talkers taking simultaneously is using the cepstrum analysis proposed by Subbs and Summerfield [8].
8) Non-speech and non-music signals: like car, motor, fan sounds, etc.
9) Complex sound mixture like multi-speakers or multi-singers with multi-music sources.

Many systems exist for content-based classification and retrieval of images; however, little work has been done on the audio portion of the multimedia stream [31, 32]. Although all fields of signal processing have attracted a considerable number of researchers and have enjoyed success in recent years, and as new software techniques ride the surging waves of ever faster computers [34-38], a few of them have touched the discrimination and separation of music and speech signals. The rest of this paper is organized as follows. In Section 2, an introduction to the analysis and properties of speech and music signals is given, with a summary of the characteristics and differences between the two types of signals. In Section 3, the main algorithms for discriminating speech from music are summarized. In Section 4, two approaches to the separation of speech from music are introduced.

2. Analysis of Speech and Music Signals

2.1. Speech signal

2.1.1. Representation of speech

It is almost impossible to tackle the speech classification problem without first establishing some way of representing the spoken utterances by some group of symbols representing the sounds produced [39-43]. The letters symbols we use for writing are obviously not adequate, as the way they are pronounced varies; for example, in English, the letter "o" is pronounced differently in words "pot", "most" and "one". One way of representing speech sounds is by using *phonemes*. Formally, we can define the phoneme as a linguistic unit such that, if one phoneme is substituted for another in a word, the meaning of that word could change. This will only hold for a set of phonemes in one language; within a single language, a finite set of phonemes will therefore exist.

However, when different languages are compared, there are differences; for example, in English, /l/ and /r/ (as in "lot" and "rot") are two different phonemes, whereas in Japanese, they are not. Similarly, the presence of individual sounds, such as the "clicks" found in some sub-Saharan African languages, or the velar fricatives (introduced later) found in Arabic, are readily apparent to listeners fluent in languages that do not contain these phonemes. Still, as the vocal apparatus used in the production of languages is universal, there is much overlap of the phoneme sets, and the total number of phonemes is finite. Table 1 shows how the phonemes are subdivided into groups based on the way they are produced [44]. The variation between different sets will be dealt with later. It is also possible to distinguish between speech sounds based solely on the way they are produced. The units in this case are known as the *phones*. There are many more phones than phonemes, as some of them are produced in different ways depending on context. For example, the pronunciation of the phoneme /l/ differs slightly when it occurs before consonants and at the end of utterances (as in "people"), as opposed to other positions (e.g. in "politics"). The two phones are called the valorized and the non-valorized "l" respectively. As they are both different forms of the same phoneme, they form a set of *allophones*. Any machine-based speech recognizer would need to be aware of the existence of allophone sets [45].

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

Vowels:	Diphthongs:	Semivowels:	Fricatives:	Nasals:	Plosives:	Affricates:
h**ee**d	b**ay**	**w**as	**s**ail	a**m**	**b**at	**j**aw
h**i**d	b**y**	**r**an	**sh**ip	a**n**	**d**isc	**ch**ore
h**ea**d	b**ow**	**l**ot	**f**unnel	sa**ng**	**g**oat	
h**a**d	b**ough**	**y**acht	**th**ick		**p**ool	
h**ar**d	b**ee**r		**h**ull		**t**ap	
h**o**d	d**oe**r		**z**oo		**k**ite	
h**oa**rd	b**oa**r		a**z**ure			
h**oo**d	b**oy**		**th**at			
wh**o**'d	b**ear**		**v**alve			
h**u**t						
h**ea**rd						
the						

It is not just the speech organs involved that influence the way an utterance is spoken and subsequently interpreted. The stress, rhythm and intonation of speech are its *prosodic* features [39-44]. Stress is used at two levels: in sentences, it indicates the most important words, while in words it indicates the prominent syllables. For example, the word "object" could be interpreted as either a noun or a verb, depending on whether the stress is placed on the first or second syllable. Rhythm refers to the timing aspect of utterances. Some languages (e.g., English) are said to be stress-timed, with approximately equal time intervals between stresses (experiments have shown that,

objectively, there is merely a tendency in this direction). The portion of an utterance beginning with one stressed syllable and ending with another is called a foot (by analogy with poetry). So, a four-syllable foot (1 stressed, 3 unstressed) would be longer than a single (stressed) syllable foot, but not four times longer. Other languages, such as French, are described as syllable-timed. Intonation, or pitch movement, is very important in indicating the meaning of an English sentence. In tonal languages, such as Mandarin and Vietnamese, the intonation also determines the meaning of individual words [46, 47].

2.1.2. Speech production

The range of sounds that we can produce is limited [39-44]. The pressure in the lungs is increased by the reverse process, which pushes the air up the *trachea* (wind pipe). The *larynx*, a bony structure covered by skin containing a slit-like orifice, the *glottis* (vocal cords), is situated at the top of the trachea. As the air flows through the glottis, local pressure falls, and eventually allows the laryngeal muscles to close the glottis, interrupting the flow of air. This in turn causes the pressure to rise again, forcing the vocal cords apart. The cycle repeats itself, producing a train of pulses. This process is known as *phonation*. The rest of the vocal tract, the oral and the nasal passages, acts as a filter, allowing the harmonics of the glottal waveform, which lie near the natural resonant frequencies of the tract to pass, whilst attenuating the others. Indeed, reasonable acoustic models of speech production have been created consisting of an excitation source driving a series of filters. So, what we get as a result of the above process is the acoustic wave radiated from the lips. To produce different sounds, we change the shape of the vocal tract by moving the jaw, tongue and lips so that the natural resonance occurs at different frequencies. In normal speech, the fundamental frequency will thus be changing all the time. However, the components of the larynx tone are always harmonics of the fundamental, and the effect of the resonance is to produce peaks in the spectrum of the output at the harmonics, which are the closest to the true resonance. This ensures that the spectrum of the resulting sound always has the same envelope (or general outline), although the fundamental frequency is continually changing. Thus, a certain sameness of quality is heard in a range of sounds with different fundamentals. If this was not the case, speech sounds could not fulfill the linguistic function that they in fact have. The peaks in the spectra described above thus correspond to the basic frequencies of the vibration of air in the vocal tract. These peaks depend on the shape of the vocal tract, and regions around them are called *formants*.

Formants are most easily seen in sonograms (also called spectrograms and spectrographs). Sonograms represented an important breakthrough in speech research when they were invented, because they could conveniently represent the way speech spectra vary with time. They are basically plots of frequency versus time, with the darkness of the trace showing the intensity of the sound at a particular frequency. Figure 3 shows a sonogram for the four semivowels /w/, /r/, /l/ and /j/, as in the syllables "wer", "rer", "ler" and "yer". It can be seen that all initial semivowels have a rising first formant. The second formant of /w/ rises, while that of /j/ falls. Different groups of phonemes (as shown in Table 1) produce different sonograms, but phonemes within a

given group will usually have similar formants. These groups may be used to distinguish between different languages by considering the frequency of occurrence of phones, which will vary for identical phones in different languages.

Fig. 3. A sonogram for the four semivowels /w/, /r/, /l/ and /j/ [43].

2.1.3. Speech perception

In speech research, a lot of effort has considered studying the way we as humans recognize and interpret speech [16, 39-43], which makes sense since the best and most accurate speech recognition (and language identification, for that matter) system in existence today is that which most of us possess. This field of study is still to answer many crucial questions, but a lot has been achieved to date. Research has shown that the two lowest formants are necessary to produce a complete set of English vowels, as well as that the three formants lowest in frequency are necessary for good speech intelligibility. More formants give more natural sounds. The situation is made more complex when dealing with continuous speech, as the speed at which some articulators can move is limited by their inertia. Consequently, there is sometimes no time for a steady vowel configuration to be reached before the tract must be changed for the next consonant, and the formants don't reach their target values. Other factors found to influence the perception of phonemes include duration and the frequency of the formants in the preceding utterance. Also, an interesting phenomenon, which has been called the "cocktail party effect", has been investigated. When a number of conversations are being carried on simultaneously, it is usually possible to follow one, even if the total loudness of the others seems greater. Experiments have shown that the continuity of the fundamental frequency group's events occurring at different times into the speech of a single speaker, and also that a common fundamental is a necessary (though not sufficient) condition for sounds to be grouped together as a stream.

2.1.4. History of speech identification

Over the past few decades, there have been important changes in the way the problem has been approached. They are briefly summarized below [39-44].

The acoustic approach (pre-1960)

The patterns of formant movements were analyzed in an attempt to recognize a word from a limited, predefined vocabulary (e.g., digits between 1 and 10). The systems performed well, but only when used by the speaker they were designed for. The usefulness of this method was limited by the fact that acoustic patterns of a word spoken on different occasions differs in duration and intensity, and the same

word produced by different persons produces patterns differing in frequency content as well.

The pattern-recognition approach (1960-1968)

Attempts were made to normalize the speech waveform in some way, so that comparisons with pre-defined patterns (words) could be made for a range of speakers. In particular, it was noted that the fundamental frequency could be used to normalize formant frequencies. Also, ways of normalizing the duration of patterns were investigated. The problem was still that such systems were only adequate for limited vocabularies.

The linguistic approach (1969-1976)

Early recognizers neglected the fact that, when two people communicate using speech, they must both use the same language. There are many sources of linguistic knowledge, which could be used to enhance various systems, such as pre-stored dictionaries, and the varying probabilities of a particular phoneme or word occurring after another one. This is referred to as *phonotactics*. Phonotactics deals with the rules that govern the combinations of the different phones in a language. There is a wide variance in such rules across languages. For example, the phone clusters /sr/ and /schp/ are common in Tamil and German respectively, but do not exist in English.

The pragmatic approach (1977-1980's)

The major advance that took place in isolated word recognizers was the use of dynamic programming algorithms, which enabled optimum non-linear timescale distortions to be introduced in the matching process. This improved the accuracy. Also, a number of more mathematically sophisticated algorithms were devised for other methods.

A simplified representation of the complete physiological mechanism for producing speech can be found in [44]. The lungs and the associated muscles act as the source of air for exciting the vocal mechanism [24, 33, 39-43]. The muscle force pushes air out of the lungs (shown schematically as a piston pushing up within a cylinder) and through the bronchi and trachea. When the vocal cords are tensed, the air flow causes them to vibrate, producing so-called voiced speech sound. When the vocal cords are released in order to produce a sound, the air flow either must pass through a constriction in the vocal tract and thereby become turbulent, producing so-called unvoiced sound, or it can build up pressure behind a point of total closure within the vocal tract, and when the closure is opened, the pressure is suddenly and abruptly released, causing a bright transient sound.

2.1.5. Speech properties

Speech is produced by the airflows from the lungs through the vocal folds and moves the larynx tube and vocal cords. Everyone has his own sound according to his organs physical dimensions. Using our human ear, we can distinguish between two

talkers talking simultaneously and may recognize them. Speech signal can be characterized by rapid rate of change of speech sounds. In other way, it can be considered as a noise like signal containing consonants [16, 33, 39-43]. We can figure out the speech signal as a continuous random signal [25]. Usually, 95% of the speech power is concentrated in frequencies below 4 kHz, and then it falls very fast through low-frequency values, and any components higher than 8 kHz [48].

2.2. Music signal

2.2.1. Music generation

Tone is the most basic component in music sound. There are two kinds of tone structures: a simple tone formed of single sinusoidal waveform and a complex tone formed of more than one harmonic [30, 49-51]. The tone quality depends on how little of non-harmonic frequency components. All classical musical tones come from a resonance frequency of moving or frictional parts of the musical instrument and some tones come from resonance tube except electronic music that produces its tone depending on electromagnetic force. Partials are any non-harmonic frequency components not a multiple of the fundamental frequency. Musical instrument manufacturers try to reduce partials. They try to make all musical sounds constructed of only harmonics with less-partial tones by producing wide band tones, covering all audible bands. Even though, all musical instruments usually produce partials that are not harmonically related to the fundamental frequency [52]. Furthermore, in correlating sound, spectra with instruments have led to another concept to explain differences in tone quality. This concept assumes that a given instrument has a sound spectrum characterized by a particular harmonic structure, which would ideally be the same for each tone of the instrument. Instead, the alternative new concept suggests that an instrument has a certain fixed frequency region or regions in which harmonics of a tone are emphasized, regardless of the frequency of the fundamental or partial components. A fixed frequency region of this kind is called a *formant*, and it is the location of these formant regions that characterizes the instrument [25]. However, there is no musical instrument restricting its formants boundaries. Finally, the most common concept of tone quality depends on subjective acoustic properties careless of partials or formants. For example, the violinists adjust the tension of violin's chord manually to reach the desired tone by just hearing, so there is no useful meaning of harmonics or fundamental frequency in this case [53, 54]. Musical production depends deeply on the musical instrument kind. The common kinds of music instruments can be summarized as follows:

1) **String instruments:** Their tones come from vibrating chords made from horsetail hair, cat's small intestine or other manufactured material like plastic or copper. This vibration is achieved by direct swing like guitar or by drawing, across chords, a properly constructed bow in a certain angle like violin. Every chord has a certain fundamental frequency so that a single musical instrument covers all audible bands. This kind produces complex tones.
2) **Woodwind instruments:** Mainly woodwind instruments consist of an open

cylindrical tube at both ends like flute. Openings in this kind's wall define the length of the standing resonance wave and allow it to radiate the sound. Some woodwind instruments use small-vibrated copper piece to produce musical tones like accordion. This kind of instrument produces harmonic tones.

3) **Brass instruments:** This kind depends on blowing like woodwind, but two main differences characterize these two kinds: the first difference is that the brass kind has a shape of animal horn like tuba; however, the second difference is that woodwind kind depends on the pressure of blowing to produce various tones while brass kind depends on manual valves to control cavity size like trumpet or special handle to vary tube length like trombone. Brass kind has a huge number of non-harmonics existed in its spectrum.
4) **Piano family instruments:** This kind uses vibrating strings as tone source by kicking it with a wooden hammer controlled by keyboard. Every single button of the keyboard is designed to produce a single tone. The tone amplitude depends on kicking force by pianist finger. Some manufactures put copper vibrating bars instead of strings. Pure harmonics have the majority of produced tone power.
5) **Percussion instruments:** Like drums, copper tam-tam, vibrating bars or carillon, kicking with baton plays all. Baton is a special wooden, plastic or metal rod. Specifications of the produced tones depend on the physical dimensions and the strength and position of the baton kick. Most of the power of tones produces non-harmonic components.
6) **Electronic production of music:** Organ is the most qualified, robust and accurate musical instrument; it has so many buttons in its large keyboard. Also, it has a memory in which it can store any note and use it frequently as a basic cadence or tone. It can play continually a background melody to let the music player add only some musical touches to complete the main melody. Rock, pop, disco and jazz cannot stand without organ help [29, 35, 36]. Although all organ special cadences are built up using computers after very precise calculations [34, 37], any sound from anywhere can be recorded then be used as a basic cadence. Although organ is the biggest and the most expensive musical instrument, it is not the only electronic musical producer. There are many of those instruments kinds with many sizes, names and brands in the market so that any person can easily buy and play his own electronic melodies. Finally, it is very important to mention that there is no need to think about harmonics or fundamental frequency to measure tone quality if the electronic musical instruments are used for music production.

2.2.2. Music properties

Music spectrum has twice the bandwidth of speech spectrum. In general, most of the signal power in audio waveform (speech or music) is concentrated at lower frequencies. Music specifications depend on the kind of played musical instruments and its physical dimensions. Musicians and melodists divide musical minor to eight parts, each part named octave and each octave is divided into seven parts (tones) [30]. These tones are named (Do, Re, Me, Fa, So, La and Se) or simply (A, B, C, D, E, F and G).

This deviation is made according to the tone frequency. For different instrument, a tempered scale is shown in Table 2. The tone (A_1) at the first octave has the fundamental frequency of the first tone in each octave, that means every first tone takes the reduplicate frequency of the first tone of previous octave, (i.e., $A_n = 2^n A_1$ or $B_n = 2^n B_1$) and so on where $n \in \{2, 3, 4, 5, 6, 7\}$. In one octave, every tone has its own frequency and the musical instrument has to produce a tone around its specific frequency. Looking at the tempered scale shown in Table 2, we note that the highest tone C_8 is at the frequency of 4186 Hz, which is the highest frequency used in human sound. This means that musical instrument manufactures try their best to bound music frequency to the human's sound limits to achieve strong concord [34, 53, 54]. In the actual world, the musical instruments cover more than audible band (approximately 20 kHz).

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

A Hz	B Hz	C Hz	D Hz	E Hz	F Hz	G Hz
A_1	B_1	C_1	D_1	E_1	F_1	G_1
27.5	30.863	32.703	36.708	41.203	43.654	48.999
A_2	B_2	C_2	D_2	E_2	F_2	G_2
55	61.735	65.406	73.416	82.407	87.307	97.999
A_3	B_3	C_3	D_3	E_3	F_3	G_3
110	123.47	130.81	146.83	164.81	174.61	196
A_4	B_4	C_4	D_4	E_4	F_4	G_4 3
220	246.94	261.63	293.66	329.63	349.23	92
A_5	B_5	C_5	D_5	E_5	F_5	G_5
440	493.88	523.25	587.33	659.26	698.46	783.99
A_6	B_6	C_6	D_6	E_6	F_6	G_6
880	987.77	1046.5	1174.7	1318.5	1396.9	1568
A_7	B_7	C_7	D_7	E_7	F_7	G_7
1760	1975.5	2093	2349.3	2637	2793	3136
A_8	B_8	C_8				
3520	3951.1	4186				

2.3. Characteristics and differences between speech and music

The speech signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time "between 5 and l00 *msec*," its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 seconds or more), the signal characteristics change to reflect the different speech sounds being spoken. A typical example of speech signal is shown in Fig. 4, which shows the time waveform corresponding to the sounds in the phrase ". . . *very good night* . . ." as spoken by a male speaker.

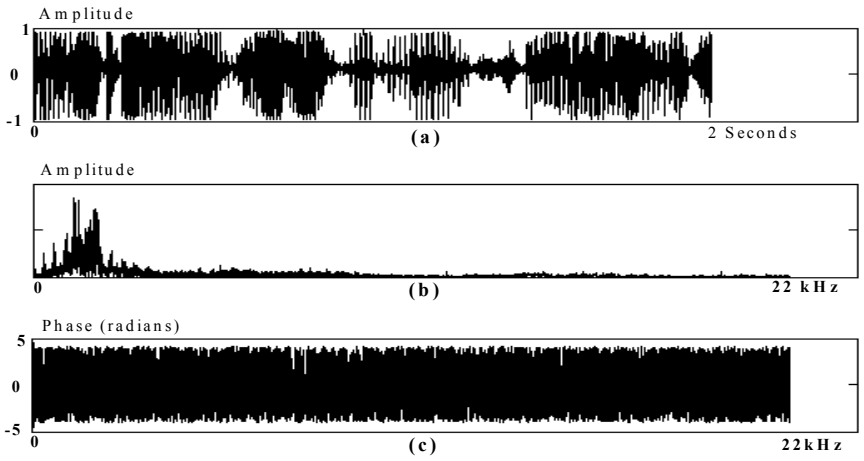


Fig. 4. A typical example of speech signal of speaking the two-second long phrase "very good night": (a) Time domain (b) Spectrum, and (c) Phase.

Figure 5 is a typical example of music portion; it is clear from the spectrum that it is possible to distinguish if the spectrum is originally speech or music.

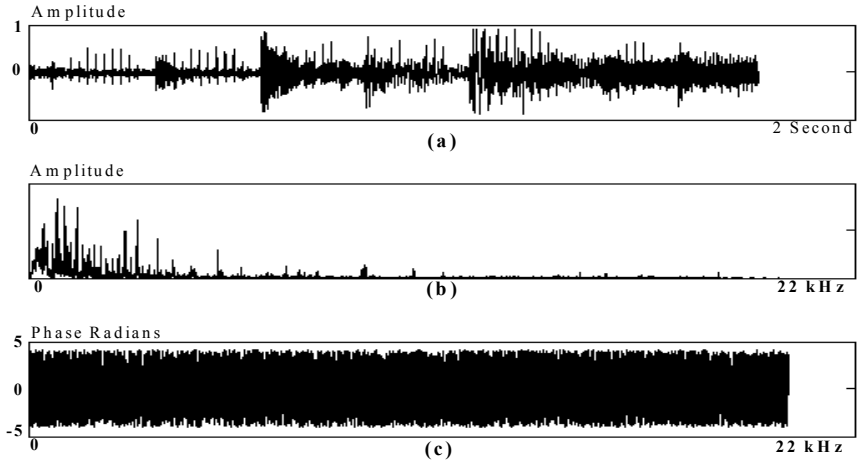


Fig. 5. A typical example of music signal for two-second long: (a) Time domain, (b) Spectrum, and (c) Phase.

The frequency spectrum and the evolutionary spectrum of the average of 500 different speech and music specimens are shown in Figs. 6 and 7, respectively.

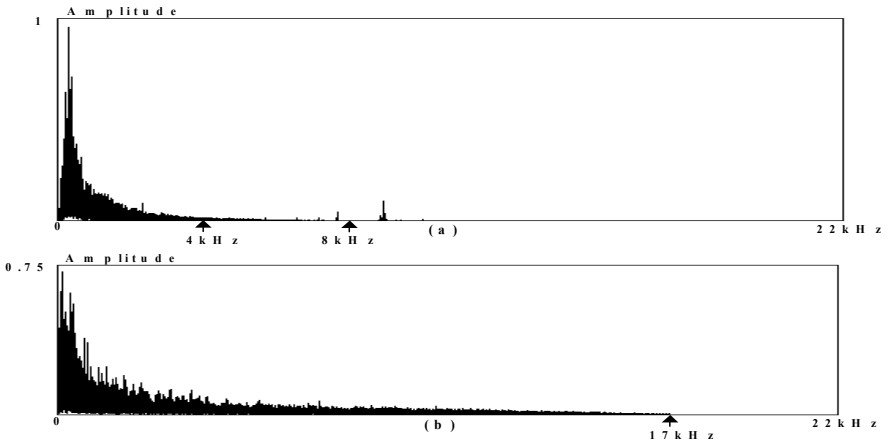
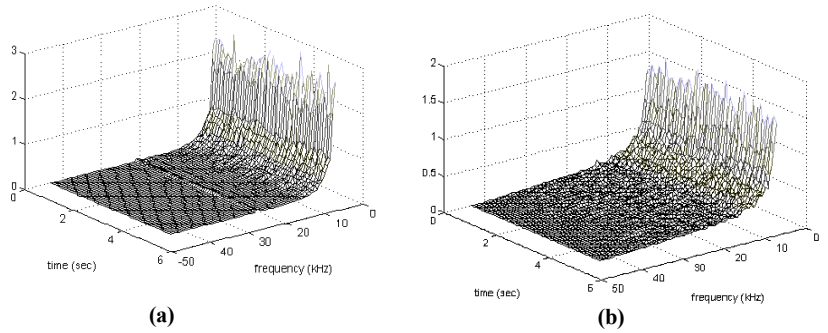


Fig. 6. Speech and music spectrum computed from averaging 500 different specimens: (a) Speech, (b) Music.



(a) **(b)**

Fig. 7. Evolutionary spectrum of speech and music computed from averaging 500 different specimens: (a) Speech, (b) Music.

In this section, we will briefly discuss some of the main differences between speech and music. These differences can be summarized as follow:

Tonality: Tone means pure sinusoidal waveform, or a single harmonic, without pitches, of a periodical signal. Music tends to be composed of a multiplicity of tones, each with a unique distribution of harmonics. This pattern is consistent regardless of the type of music or instruments [47, 52, 55-57], while in speech care much about his voice tonality.

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

Bandwidth: Speech has usually 90% of its power concentrated in frequencies lower than 4 kHz (and limited to 8 kHz), whereas music can extend through the upper limits of the ear's response at 20 kHz. In general, most of the signal power in music waveforms is concentrated at lower frequencies [52, 58].

Power distribution: The power distribution in the power spectral density of speech signal is different from music signal; usually the power of speech concentrates at low frequencies then collapses very fast trough the higher values of frequency without DC value while there is no specific shape of the music spectrum and perhaps it contains DC. Therefore, we can say that the power of the speech concentrates at some areas different than the case of music [59].

Fundamental frequency: If a specific person talks alone, we configure almost accurate fundamental frequency. However, it is not the case for a specific music instrument.

Dominant frequency: For a specific person talking alone, we can accurately assign his unique dominant. In a single play of a specific musical instrument, we can only determine average dominant. That is because the amplitude reaches its maximum in wide spectrum range. The case will be worst for multi-musical instruments.

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

Tonal duration: The duration of vowels in speech 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.

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

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

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

We can easily note that the strong temporal variation is the amplitude of the speech signals. Also, the short-terms (peaks and valleys) change over a short period of time. The main differences between speech and music signals can be summarized in Table 3.

Table 3. The main differences between speech and music signals

Key difference	Speech	Music
Units of analysis	Phonemes	Notes Finite
Spectral structure	• Largely harmonic (vowels, voiced consonants). • Tend to group in formants. • Some inharmonic stops.	• Largely harmonic, some inharmonic (percussion).
Temporal structure	• Short (40ms–200ms). • More steady state than dynamic. • Timing unstrained but variable. • Amplitude modulation rate for sentences is slow (~ 4 Hz)	• Longer (600-1200 ms). • Mix of steady-state (strings, winds) and transient (percussion). • Strong periodicity.
Syntactic/Semantic structure	• Symbolic, • Productive, • Can be combined in grammar	• Symbolic • Productive, • Combined in a grammar

Overlap in speech and music signal is, in general, very strong so that there is no ordinary filter that can separate them from each other. Speech covers the spectrum from near zero to 3.5 kHz with an average dominant frequency =1.8747 kHz. However, from the classical theorem of music, the lowest fundamental frequency (A1) is about 27.5 Hz and the highest tone C8 is around the frequency of 4186 Hz. Therefore, a musical instrument manufacture tries to bound music frequency to the human's sound limits to achieve strong consonant and also strong frequency overlap. Moreover, music propagates over all the audible spectrum and covers more than audible band (20 kHz), with an average dominant frequency = 1.9271 kHz. Also, a music instrument in general has many fundamental frequencies while the speech of a specific person has only unique fundamental frequency, and we can assign more accurately a dominant frequency for a specific person speech while it is not the case in a specific music instrument. For a music instrument, every tone kind (i.e., A) has a special fundamental and for a peace of music played by many instruments we will get too many fundamentals. So, we cannot think about reconstructing the music sound by depending only on fundamentals [25].

3. Speech and Music Classification

In this section, the main classification approaches are discussed. These approaches can be classified into three categories: (1) time-domain, (2) frequency domain, and (3) time-frequency domain types. El-Maleh [61, 62] has developed a two-level music and speech classifier and used long-term features such as differential parameters, variance, time-averages of spectral parameters, and zero crossing rate (ZCR). Saunders [60] has also proposed a two-level algorithm for discrimination based on the average ZCR and the short-time energy (STE) features, and applied a simple threshold procedure. Matityaho and Furst [63] have developed a neural network based model for classification of music type. They have designed their model based on human cochlea functional

performance. Hoyt and Wecheler [64] have developed a neural network base model for speech detection, but they have used the Hamming filter, Fourier transform and a logarithmic function as pre-processing before neural network input and used a simple threshold algorithm to detect speech from music, traffic, wind or any interfering sound. Also, they have suggested another wavelet transform feature as an option of pre-processing to improve the performance. Their work is similar to the work done by Matityaho and Furst's [63]. Scheirer and Slaney [65] examined 13 features, some of them are modifications of each other intended to measure conceptually distinct properties of speech and/or music signals, and combined them in several multidimensional classification frameworks. For the datasets they used, the best classifier classifies with 5.8% error on a frame-by-frame basis, and 1.4% error when integrating long (2.4 seconds) segments of sound. Using long-term features, like cepstrum pitch or spectral centroid, consumes large delay without worth increase in overall discrimination precision. It was observed that the most powerful discrimination features are the ZCR and the STE; therefore, they will be discussed in more details. In general, the music and speech discrimination process found in literature can be classified into the following algorithms:

I. **Time-domain approaches**:
   1) The ZCR [38, 60-62, 65, 66, 124-130]:
      (a) Standard deviation of first order difference of ZCR.
      (b) Third central moment about the mean of ZCR.
      (c) Total number of zero crossing exceeding a threshold.
   2) The STE [60, 62, 65, 66].
   3) ZCR and STE positive derivative [66].
   4) Variance of the roll-off feature [31, 59].
   5) Pulse metric [31, 59, 67-69].
   6) Number of silent segments [32].
   7) Hidden Markov Model (HMM) unit [70-72].
   8) Neural networks [12, 49, 58, 63, 72-108].
   9) Number of silent segments [60, 62, 65, 66].

II. **Frequency-domain approaches [31, 32, 59, 99, 100, 109, 124-130]**:
   1) Spectrum:
      (a) Spectral centroid.
      (b) Mean and variance of the spectral flux.
      (c) Mean and variance of the spectral centroid.
      (d) Variance of the spectral flux.
      (e) Roll-off of the spectrum.
      (f) Bandwidth of signal.
      (g) Amplitude.
      (h) Delta amplitude.
   2) Cepstrum [110]:
      (a) Cepstral residual [111, 112].
      (b) Variance of the cepstral residual [111, 112].

(c) Cepstral feature [111, 112].
(d) Pitch [82, 95, 96, 105-107, 113, 114].
(e) Delta pitch [76, 107].

III. **Time-frequency domain approaches**:

1) Spectrogram [13, 19, 74, 115].
2) Evolutionary spectrum [116, 117].

3.1. Time-domain approaches

3.1.1. The zero crossing rate

The ZCR of the time domain waveform is one of the most indicative and robust measures to discern voiced speech. It has widely used in practice as a strong measure to discern fricatives from voiced speech [38]. The ZCR is simply the count of crossing the zero throw fixed window size. It is said to occur if successive samples have different algebraic signs. Equation (1) defines the required computation for ZCR.

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

where Z_n is the ZCR, $sgn[x(n)] = 1$ when $x(n) > 0$, $sgn[x(n)] = -1$, when $x(n) < 0$, and N is the number of samples in one window. Obviously, the ZCR is a time-domain algorithm and it deeply depends upon the frequency of the input signal $x(n)$. Moreover, the sampling rate should be high enough to illustrate any crossing through zero. In addition, the most important thing that must be thought about before starting counting the crossing is normalizing the signal so that the amplitude average through the window should be equal to zero using constant shift to every sample in the amplitude axis. This constant should be equal to the actual amplitude average [38]. That means readjusting the balance of the ZCR. This will ensure that every widow has only one specific ZCR. From Eq. (1), it is clear that the ZCR is proportional to the dominant frequency of $x(n)$, so it can be concluded that the ZCR of music is usually bigger than that of speech. However, there is an abrupt increase of ZCR in speech due to strongly unvoiced speech. This is a pitfall in this informer because, in some regions, the ZCR of speech exceeds the ZCR of music.

Properties of the ZCR:

The ZCR of a signal has many properties [38], which can be summarized as follows:

1) The dominant frequency (*DF*) principle:

If a signal is a pure sinusoidal waveform, the dominant frequency is the only one in the spectrum. This frequency equals the number of zero crossings of the signal in one second. In other words, it equals the value of the ZCR if the rate is taken every second. For non-sinusoidal periodic waveform, the dominant frequency has the largest amplitude. An accurate dominant frequency (w_0) can be calculated using the formula:

$$w_o = \frac{\pi E\{D_o\}}{N-1} \tag{2}$$

where D_o is the ZCR per second, N is the number of intervals and $E\{.\}$ always denotes the expected value.

2) Highest frequency detection:

Since D_0 denotes the ZCR of a discrete-time signal $Z(i)$, let us assume that D_1 is the ZCR of the first derivative of $Z(i)$, and D_n denotes the ZCR of the n^{th} derivative of $Z(i)$. The highest frequency W_{max} in the signal can be approached using the following equation:

$$W_{max} = \lim_{i \to \infty} \frac{\pi E\{D_i\}}{N-1} \tag{3}$$

where N is the number of samples. If the index i reaches 10, then the change in W_{max} can be ignored if the sampling rate equals 11 kHz.

3) Lowest frequency detection:

In discrete-time signals, an approximate derivative ∇ of a signal $Z(i)$ can be evaluated by simply subtracting the amplitude of a sample from the amplitude of the previous sample, assuming that the time between any two samples is normalized to unity, so:

$$\nabla Z(i) = Z(i) - Z(i-1)$$

Let D_n denotes the ZCR of the n^{th} derivative of $Z(i)$. If we define ∇^+ as the positive derivative operator, then $\nabla^+ [Z(i)]$ can be defined as:

$$\nabla^+ [Z(i)] = Z(i) + Z(i-1) \tag{4}$$

Also, let ${}_nD$ be the ZCR of the n^{th} positive derivative of the signal $Z(i)$. Then the lowest frequency W_{min} of a signal can be described as:

$$\lim_{i \to \infty} \frac{\pi E\{{}_iD\}}{N-1} = W_{min} \tag{5}$$

From the previous information about dominant, maximum and minimum frequencies of a signal, we can have a brief shape of the signal spectrum.

4) Periodicity measure:

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

$$\frac{E\{{}_1D\}}{N-1} = \frac{E\{{}_2D\}}{N-1} \tag{6}$$

and as these two sides close to each other as the signal becomes more periodic. Using this measure, music was found to be more periodic or more tonal than speech [46, 47, 55-57, 118].

High zero-crossing rate ratio (HZCRR)

The ZCR has proved to be very useful in characterizing different audio signals. It was used in many previous speech/music classification algorithms. It was experimentally found [66] that the variation of the ZCR is more discriminative than the exact value of the ZCR, so the HZCRR can be considered as one feature. The HZCRR is defined as the ratio of the number of frames whose ZCR are above 1 fold average zero-crossing rate in one-second window, and can be expressed as:

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(ZCR(n) - ZCR_{av}) + 1 \tag{7}$$

$$ZCR_{av} = \sum_{n=0}^{N-1} ZCR(n) \tag{8}$$

where n is the frame index, N is the total number of frames in a one-second window, *sgn*[.] is a sign function and $ZCR(n)$ is the zero-crossing rate at the n^{th} frame. In general, speech signals are composed of alternating voiced sounds and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure. Hence, for speech signal, the variation of the ZCR (or the HZCRR) will be greater than that of music, as shown in Fig. 8.

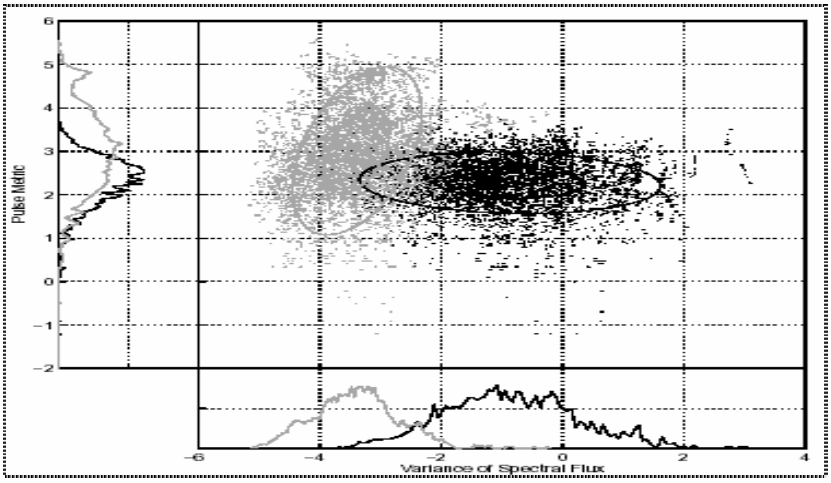


Fig. 8. Music and speech sharing some values [65].

3.1.2. Short-time energy

The amplitude of the speech signal varies appreciably with time. In particular, the amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments. The short-time energy (STE) of the speech signal provides a convenient representation that reflects these amplitude variations. Since the music signal does not contain unvoiced segments, its STE is usually bigger than that in speech case **[60]**. Given a discrete-time signal $s(n)$, we define its energy as:

$$E_S = \sum_{n=-\infty}^{\infty} \left| s(n) \right|^2 \tag{9}$$

The power of a discrete-time signal $s(n)$ is defined as:

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

Table 4. Types of signals

Energy $0<E_s<\infty$	Transient	$S(n)= \alpha^n u(n)$ $\lvert\alpha\rvert<1$
	Finite Sequence	$e^{\beta t}[u(n)-u(n-255)]$ $\lvert\beta\rvert<\infty$
Power $0<P_s<\infty$	Constant	$s(n)= \alpha$ $-\infty< \alpha<\infty$
	Periodic	$s(n)=\alpha \sin(n\omega_o+\varphi)$ $-\infty< \alpha<\infty$
	Stochastic	$S(n)$=rand (seed)
Neither	Zero	$s(n)=0$
	Blow up	$s(n)=\alpha^n u(n)$ $\lvert\alpha\rvert>1$

In general, signals can be classified into three types: 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, as described in Table 4. Now, if we are given a sequence $\{s(n)\}$we can form another sequence $\{f(n;m)\}$ as follows:

$$f_S(n;m) = s(n)w(m-n) \tag{11}$$

where $w(n)$ is a window sequence with a length of N. It is zero outside the range [0, N-1]. The new sequence $f_s(n;m)$ is therefore zero outside the range [m-N+1,m]. It should also be noted that N and w are implicit arguments to $f_s(n;m)$. f_s is called a frame. If, for example, the window is chosen as rectangular, then f_s is simply the last N points of $s(n)$ ending with $s(m)$.

Energy signals and frames problems:

Speech is most like a power signal. During voiced speech, it is periodic and during unvoiced speech it is filtered white noise and therefore stochastic. During the silence

between speeches, it can be modeled as stochastic background noise. Speech is also not stationary and when we frame the speech to isolate a stationary bit, the framed sequence is an energy signal.

Define $\boldsymbol{F_s}$ as a long-term feature on the sequence $\{\boldsymbol{s(n)}\}$. It maps elements of the Hilbert space $\boldsymbol{H}$ to the set of complex numbers $\boldsymbol{C}$, i.e.,

$$\boldsymbol{F_s: H \rightarrow C} \tag{12}$$

The Hilbert space can be considered as the space of all sequences, or an infinite dimensional vector space. The long-term feature on a signal can be defined as follows:

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

The long-term average is appropriate for power signals. However, when applied to energy signals, the result will be zero. For energy signals resulting from the application of a window, a more appropriate definition is:

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

A family of mapping can also be considered. If each member of the family is selected to be a parameter $\boldsymbol{\lambda}$, the notation $\boldsymbol{F_s(\lambda)}$ can be used. An example of a parametric long-term feature is the discrete-time Fourier transforms. We wish to consider the long-term feature of the form:

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

where M is the sequence mapping. It maps a sequence $\{\boldsymbol{s(n)}\}$ to another sequence. The long-term average feature $\boldsymbol{F_s(\lambda)}$ is therefore $\boldsymbol{L^oM}$, a composition of function $\boldsymbol{L}$ and $\boldsymbol{M}$. If $\boldsymbol{F_s(\lambda)}$ is a long-term feature of the form (12), then a short-term feature $\boldsymbol{F_s(\lambda;m)}$ relative to time $\boldsymbol{m}$ can be constructed as follows:

Construction principle:

- Define a frame sequence as:

$$f_s(n;m) = s(n)w(m-n) \tag{16}$$

- Apply the long-term feature transformation to the frame sequence.

$$\begin{aligned} F_s(\lambda;m) &= L\{M(\lambda)\}\{f_s(n;m)\} \\ &= L\{M(\lambda)\}\{s(n)w(m-n)\} \\ &= \frac{1}{N}\sum_{n=-\infty}^{\infty} M(\lambda)\{s(n)w(m-n)\} \end{aligned} \tag{17}$$

Low short time energy ratio (LSTER)

As in the ZCR, we also select the variation, not the exact value of short-time energy as one component of our feature vector. Here, we use the LSTER [33] 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 follows:

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(0.5STE_{av} - STE(n)) + 1] \tag{18}$$

where

$$STE_{av} = \sum_{n=0}^{N-1} STE(n) \tag{19}$$

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

3.1.3. Positive derivation effect

Preprocessing the input signal with positive derivation concept $(\nabla+)$, as shown in Fig. 9, has proved some improvement [66] in the discrimination process, where ${}_nD$ denotes the ZCR of the n^{th} positive derivative of the signal $Z(i)$.

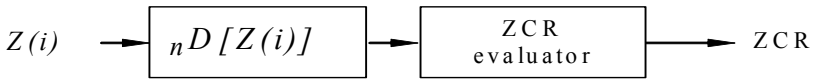


Fig. 9. Positive derivative before applying the ZCR.

This preprocessing has reduced the ZCR of the speech and increased the ZCR of music but caused some delay. Figure 10 shows the averages of the ZCR in speech, music and mixture after positive derivation of order 50.

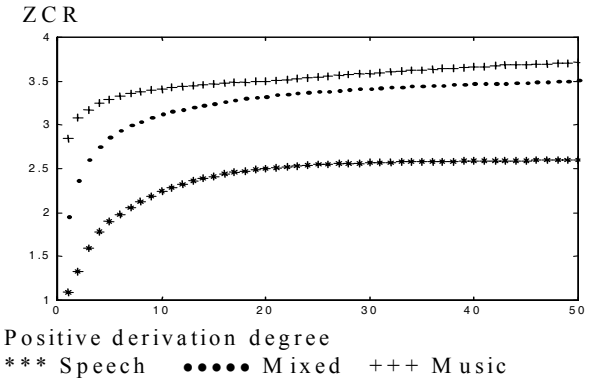


Fig. 10. The average ZCR of music, speech and mixture.

3.1.4. Neural network (NN)

The human brain has roughly 10 trillion nerve cells (neurons) that communicate with one another through synaptic connections. In the developing brains of very young mammals -- and probably humans, although scientists don't know that for certain -- these neural messages often fail to get through from one synapse to the next. As the brain matures, the synaptic connections also mature, becoming more functional and numerous, and the mammal develops cognitive skills [14, 35, 63, 73, 75-93, 98, 113]. The NN is a multipurpose technique and it can be used to implement many algorithms. It has been used widely, for example, in classification matter [16, 49, 95-99, 107, 108, 119, 120]. It is assumed that speech and music representation is based on known physiological facts of the human ear, followed by a non-linear function that connects the decision about the sound type and representation. A multi-layer NN, as a classification tool, is an applicable approach for representing a non-linear decision making system. In order to apply the decision process with a NN, it is essential to choose a reliable representation of the input signal. Time domain representation of the musical tone requires a huge amount of input nodes when considering the fact that a decision is obtained from a few second long time interval. The human ears process the auditory signal by performing Fourier analysis and the frequency components are transferred to the brain by independent channels [14, 74, 120].

3.2. Frequency-domain approaches

3.2.1. Spectrum

Mean and variance of the spectral flux

This feature measures frame-to-frame spectral difference so it characterizes the change in the shape of the spectrum. Music has a higher rate of range, and goes through more drastic frame-to-frame changes than speech. It can be noted that music alternates periods of transition (consonant-vowel boundaries) and periods of relative stasis (vowels), where speech typically has a more constant rate of change. As a result, the spectral flux value is higher for music, particularly unvoiced signals, than it is for speech. But, the value of spectral flux for speech signals is much smaller than the value of spectral flux of environment signals because in the environment signals there will be much frame-to-frame changes than speech signals. Spectral flux (or "Delta Spectrum Magnitude") is defined as "the 2nd *norm* of the frame-to-frame spectral amplitude difference vector:

$$SF = \| | X(k) - | X(k+1) | \|$$

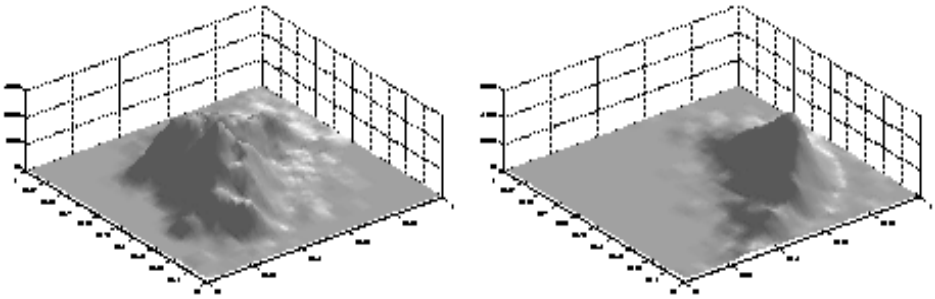
where k is an index corresponding to a frequency and $X(k)$ is the power of the signal at the corresponding frequency band. SF can also be described as:

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

where $A(n, k)$ is the discrete Fourier transform (DFT) of the n^{th} frame of the input signal,

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

$x(m)$ is the original audio data, $w(m)$ is the window function, L is the window length, M is the order of the DFT, N is the total number of frames, and δ is an arbitrary constant. Spectral flux is a feature that Scheirer and Slaney [65] found to be useful in discriminating music from speech. Music can be regarded as a succession of periods of relative stability, notes and phones, in spite of the presence of short signals such as percussions (inducing high-frequency noise). Speech is rather a rapid succession of noise periods, such as unvoiced consonants, and of periods of relative stability, like vowels. Then, the selected features have interesting properties. They give very different values for voiced and unvoiced speech; and they are relatively constant within a segment of musical sound. The variances are higher for speech than for music, whereas the means tends to be greater for music than for speech, as shown in Fig. 11. Rossignol *et al*. [109] have used frames of 18 *ms* long. Means and variances were computed in a one-second segment.



(a) (b)

Fig. 11. Normalized features 3D histogram (mean and variance of spectral flux) of (a) music, (b) speech [109].

In order to identify the segments, they have tested three classification methods: a Gaussian mixture model (GMM) classifier; the k-nearest-neighbors (kNN) classifier, with k=7; and NN classifier. Table 5 shows their results when using only the mean and variance of the spectral "flux".

Table 5. Percentage of misclassified segments [109]

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

Mean and variance of the spectral centroid:

This feature describes the center of frequency at which most of the power in the signal (at the time frame examined) is found. Music signals have high frequency noise

and percussive sounds that result in a high spectral mean. On the other hand, in speech signals the pitch of the audio signal stays in a more narrow range of low values. As a result, music has a higher spectral centroid than speech. The spectral centroid for a frame occurring at time t is computed as follows:

$$SC = \frac{\sum_{k} kX(k)}{\sum_{k} X(k)} \tag{22}$$

where k is an index corresponding to a frequency and $X(k)$ is the power of the signal at the corresponding frequency band. When using a combination of the mean and variance of the spectral flux, the mean and variance of the spectral centroid, and the mean and variance of the ZCR, we get the results of Table 6.

Table 6. Percentage of misclassified segments [109]

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

Roll-off point:

This feature is the value of the frequency that 95% of the power of the signal resides under. As mentioned before, the power in music is concentrated in the higher frequencies; however, speech has a range of low frequency power. The mathematical expression for finding this value of frequency is as follows [65, 109]:

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

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

4-Hz modulation energy:

Speech has a characteristic energy modulation peak around the 4 Hz syllabic rate. To use this method, we use a second order band pass filter with center frequency of 4 Hz, then calculate the energy at that modulation frequency with respect to the overall signal energy. Speech signals have higher energy at that frequency. However, some music bass instruments were found [65, 109] to lead this test to error because they also have modulation energy around 4 Hz.

3.2.2. Cepstrum

Many research papers refer to cepstrum as discretion of the spectrum shape. It is defined as the inverse DFT of the logarithm of the power spectrum of a signal. The bases

of this approach are that a signal with an echo will have an added periodic component to the logarithm of its power spectrum, and the Fourier transform of that logarithm should exhibit a peak at the echo delay. Music has higher cepstrum values than that of speech. The mathematical expression of the complex cepstrum is as follows [110-112]:

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

and

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{jwn})dw \tag{24}$$

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

3.2.3. Summary

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

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

Feature	Latency	Error
4 Hz Mod Energy	1 sec	12 +/-1.7%
Low Energy	1 sec	14 +/-3.6%
Roll off	1 frame	46 +/- 2.9%
Var Roll off	1 sec	20 +/- 6.4%
Spec Cent	1 frame	39 +/- 8.0%
Var Spec Cent	1 sec	14 +/- 3.7%
Spec Flux	1 frame	39 +/- 1.1%
Var Spec Flux	1 sec	5.9 +/- 1.9%
Zero-Cross Rate	1 frame	38 +/- 4.6%
Var ZC Rate	1 sec	18 +/- 4.8%
Ceps Resid	1 frame	37 +/- 7.5%
Var Ceps Res	1 sec	22 +/- %5.7
Pulse Metric	5 sec	18 +/- %2.9

They evaluated their models using labeled data sets, each 20 minutes long, of speech and music data. Each set contained 80 15-second-long audio samples. The samples were collected by digitally sampling an FM tuner (16-bit monophonic samples at a 22.05-kHz sampling rate), using a variety of stations, content styles, and noise levels, over a three-day period in the San Francisco Bay Area. They claimed they have both male and female speakers, both "in the studio" and telephonic, with quiet conditions and with varying amounts of background noise in the speech class; and samples of jazz, pop, country,

salsa, reggae, classical, various non-Western styles, various sorts of rock, and new age music, both with and without vocals, in the music class [29]. They also tested several feature subsets using the spatial partitioning classifier; the results are summarized in Table 8. The "best 8"' features are the variance features, plus the 4 Hz modulation, low-energy frame percentage, and pulse metric [67, 68, 121]. The "best 3" are the 4 Hz energy, variance of spectral flux, and pulse metric. The "fast 5" features are the 5 basic features which look only at a single frame of data, and thus have low latency. It could be seen from these results that not all features are necessary to perform accurate classification, and so a real-time system may gain improved performance by using only some of the features.

Table 8. Performance for various subsets of features

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

The "Var Spec Flux only"' data is not directly comparable to that in Table 7, since in Table 7 "cannot classify" points were ignored, but here they are treated as errors. They also calculated a long-term classification by averaging the results of the frame-by-frame spatial partitioning classification in non-overlapping 2.4 second windows. Using this testing method, the error rate drops to 1.4%. Thus, the frame-by-frame errors, while not distributed independently, are separate enough that long-term averaging can eliminate many of them.

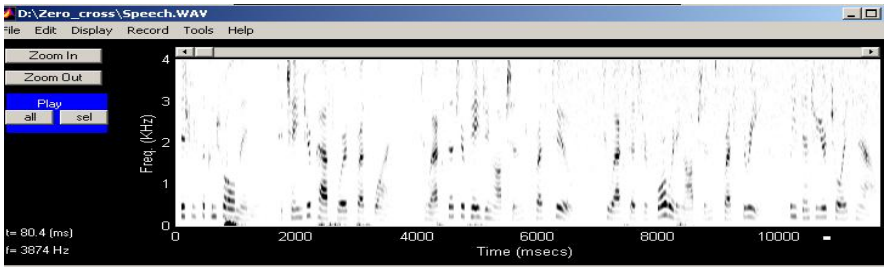
3.3. Time-frequency domain approaches

3.3.1. Spectrogram

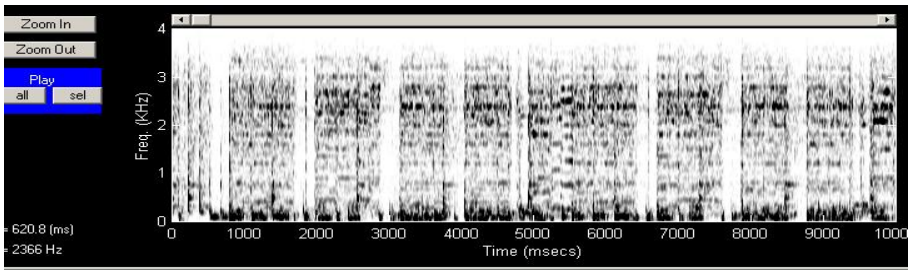
The spectrogram is an example of time-frequency distribution and this method was found to be a good classical tool for analyzing speech signal [13, 19, 74, 115]. The spectrogram kernel of a sequence $x(n)$ is defined as follows:

$$X(n,w) = \sum_{m=-N}^{N} W(n+m)x(m)e^{-jwm} \tag{25}$$

where $W(n)$ is a window and N is the length of the sequence $x(n)$. This method can be used to discriminate music from speech signal, but it has a high percentage error because it depends on the strength of the frequency in a range of time. Figure 12 shows two examples of speech and music spectrograms.



(a)



(b)

Fig. 12. (a) Speech spectrogram, (b) Music spectrum.

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:

$$e(n) = \int_{-\pi}^{\pi} e^{jwn} dZ(w) \tag{26}$$

where Z(ω) is the process with orthogonal increments; i.e.

$$E\{dZ^*(w)dZ(\Omega)\} = \frac{S(w)dw}{2\pi}\delta(w-\Omega) \tag{27}$$

and $S(\omega)$ is the spectral density function of *e(n)*. The family of constant amplitude sinusoids is, however, not appropriate for characterizing non-stationary processes, like speech. In the Wold-Cramer decomposition, a discrete-time non-stationary process {*x(n)*} is considered the output of a casual linear, and time-variant (LTV) system with a zero-mean, unit-variant white noise input *e(n)*; i.e.

$$x(n) = \sum_{m=-\infty}^{n} h(n,m)e(e-m), \tag{28}$$

where *h(n,m)* is the impulse response of the LTV system. Substituting *e(n)* from Eq. (26) into Eq. (31) (*S(ω)*=1 for white noise) we get:

$$x(n) = \int_{-\pi}^{\pi} H(n,\omega)e^{j\omega n} dZ(\omega) \tag{29}$$

where the generalized transfer function of the LTV system is defined as:

$$H(n,w) = \sum_{m=-\infty}^{n} h(n,m)e^{-jwm} \tag{30}$$

A non-stationary process can thus be expressed as an infinite sum of sinusoids with random, time-varying amplitudes and phases. Since the instantaneous variance of *x*(*n*) is given by:

$$E\{|x(n)|^2\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(n,\omega)|^2 d\omega \tag{31}$$

the Wold-Cramer ES is defined as:

$$S(n,\omega) = \frac{1}{2\pi}|H(n,\omega)|^2 \tag{32}$$

It was also found that the ES could be used for distinctions music from speech signals [117]. Although the time-frequency distributions, like the spectrogram and the ES, are good in discriminating music from speech signals, the main disadvantage of these tools is the cost of their computations; therefore, they may be used in off-line analysis. Figure 13 shows the ES of speech and music signals. The suppression of the amplitude for speech might due to gaussianity.

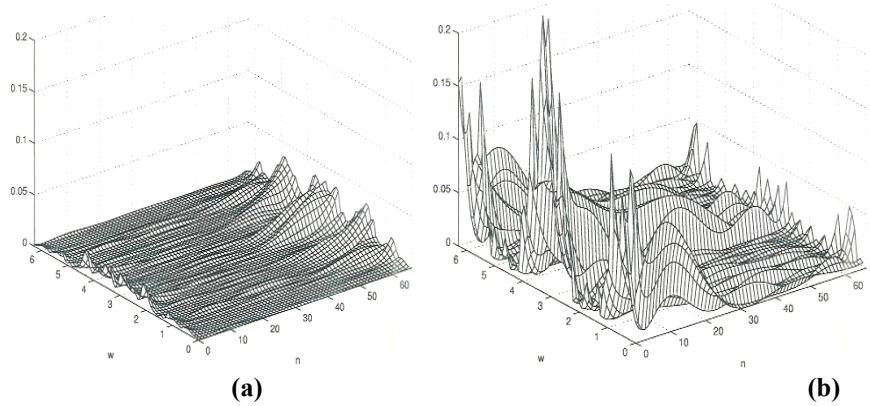


Fig. 13. (a) Speech ES, (b) Music ES.

4. Speech and Music Separation

In this section, two approaches for separating speech form music are presented. Figure 14 shows how a classifier works together with a separator [7, 22].

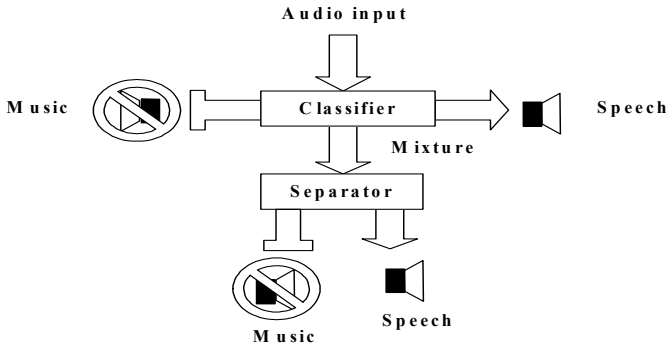


Fig. 14. Separator integrated with classifier.

4.1. Independent component analysis and neural networks

Source separation is still an open research area [7-13, 22, 74]. Wang and Brown (W&B) [13, 21] have proposed a model for speech segregation using LEGION oscillatory neurons network [122]. LEGION is the acronym for locally excitatory globally inhibitory oscillator network. They proposed their model for human speech segregation from random noise, noise bursts, cocktail party sound male or female speech, siren, trill telephone, or rock music. They presented their model with only single mixture input of signals [20], and they based their analysis on physiological studies of auditory nerve tuning curves, depending on the fundamental frequency (F_0) features for

the target speech. Their 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:

$$h_i(t) = t^{n-1} e^{[(-2\pi b_i t)\cos(2\pi f_i t + \phi_i)]}\, U(t) g(i),\ 1 \le i \le N \tag{33}$$

where N is the number of filter channels, n is the filter order and U is the unit step function (i.e., $U(t) = 1$ for t greater or equal zero and zero otherwise). Hence, the gammatone is a time invariant causal filter with an infinite response time. For the i^{th} filter channel, f_i is the center frequency of the filter (in Hz), ϕ_i is the phase (in radians), b determines the rate of decay of the impulse response which is related to bandwidth and $g(i)$ is an equalizing gain for each i filter which is related to gain adjust. The impulse response of the gammatone filters is depicted in Fig. 15.

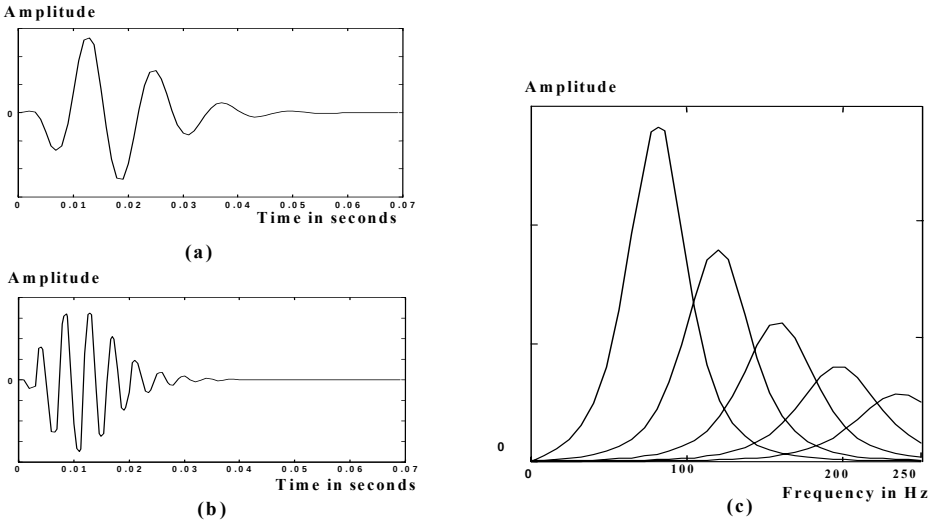


Fig. 15. Gammatone filter fourth order phase-less impulse response: (a) in time domain for *i=1*, *fi=80Hz*, (b) *i=5*, *fi=244Hz* and (c) in frequency domain for the first five filters (i.e *i=1 to i=5*) with *g(i)* set to unity.

Then, *grouping* will be achieved by passing the preprocessed input to a two-layer 128*150, two-dimensional oscillator neural networks. The first layer is a network of relaxation oscillators. Wang coined the word "LEGION" for this network because the property of its connection weight is locally excitatory globally inhibitory. The two-dimensional network, with respect to time and frequency, are derived from the cross-correlation values computed in the preprocessing stage while all connection weights along the time axis remain constant. Synchronized blocks of oscillators, called *segments*, correspond to the connected regions of acoustic energy in the time-frequency plane,

while different segments are desynchronized. Segments are the atomic elements of a represented auditory scene; they capture the evolution of perceptually relevant acoustic components in time and frequency. In the second layer, these segments are appropriately grouped. These groups correspond to each perceptual auditory stream. Different segments within the same time frame are grouped if their corresponding frequencies are either both agree or both disagree with the F_0 (fundamental frequency of target speech extracted in the preprocessing stages). Accordingly, this layer groups a collection of oscillators to form a foreground stream that corresponds to a synchronized population of oscillators and puts the remaining segments into a background stream that also corresponds to a synchronized population. The last stage of this model is *resynthesizing* each group that has been grouped by neural oscillator network by reversing auto-correlation functions that has been done through preprocessing stages. A block diagram of the W&B model is shown in Fig. 16.

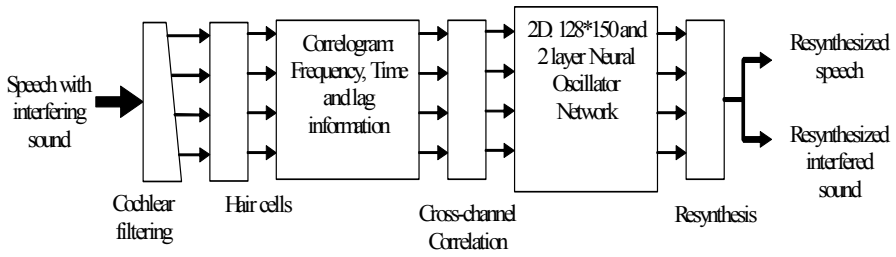


Fig. 16. Block diagram of W&B model.

The main drawback of W&B model is the complexity. Very high specification hardware is needed to perform correlogram and two-dimensional neural network oscillator during the training phase. Also, nothing has been mentioned about the separation delay. 75% of the total target speech power was recovered when the interference sound was rock music and the signal to noise ratio of the overall output did not go below 15 *db*. Kouwe [20] reported that Wang and Brown model needs to be improved to reach an acceptable level like an independent component analysis method if two sources of mixture sound are available and assumes that signals from different sources are statistically independent [123]. Also, their model does not have any mechanism for grouping segments that do not overlap in time domain. Takigawa and his partners [19] have also tried to improve the performance of Wang and Brown model by using short time Fourier transform (STFT) in the input stage and using spectrogram values instead of correlogram with a larger two-dimensional oscillator neural network (256x256), but they did not report how much was the improvement. Mu and Wang [115] extended the Wang and Brown model by adding further processing based on psychoacoustics evidence to improve the performance. They included the estimation of the F_0 of target speech and refined the generation of target speech stream with the estimated F_0. Their model increased the recovered target speech power by about 5%. A

similar work has been also done by Stubbs and Summerfield [8] to separate the voiced speech of two talkers speaking simultaneously at similar intensities in a single channel using pitch peak canceling in cepstrum domain.

4.2. Pitch cancellation

This method is widely used in noise reduction. Stubbs **[8]** introduced 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. He used pitch canceling in the cepstrum domain. The main point in this method is enhancing one of the sound components, like music, by zeroing the five cepstral samples centered on the pitch peak of the other interfering components, like speech. Input sound of 11 kHz is low-pass filtered at about 4.25 kHz using recursive filter of five poles. It is then normalized to insure that the average is zero, maximum absolute amplitude of one and unity standard deviation. The fast Fourier transform (FFT) is computed at each 51.2 *ms* fragment to produce amplitude and phase spectrum. This period is long enough to create an amplitude spectrum in which the harmonic structure of the signal is observed [8]. The main goal here is to delete speech signal from the input mixture sound using its pitches, then recover it again by directly subtracting the separated music signal from the mixed.

The logarithmic effect will reduce a high amplitude and increase a low one, and the values near zero will be very large after the logarithm. In the speech signal, it is well known that the physical parts of generating speech (lungs, vocal folds, larynx tube) are not solid all the time, during speech generation. Physical dimensions of these organs are changing and they stop some milliseconds to produce different sounds, producing consonants. It is usually produced in almost fixed duration and is repeated periodically at almost constant frequency. Every human at most has unique consonant frequency and duration. For example, the letter *A* has a lot of consonant and the letter *R* has too for a certain person. In frequency domain, these speech consonants appear as a very weak sample (near zero), but in cepstrum domain it will appear as long pitch peak. If we try to delete this consonant by zeroing the five-cepstral samples centered at the pitch peak, we hope that the speech signal will be distorted completely. Figure 17 shows a typical example of speech and music in cepstrum domain computed from five-second time signal.

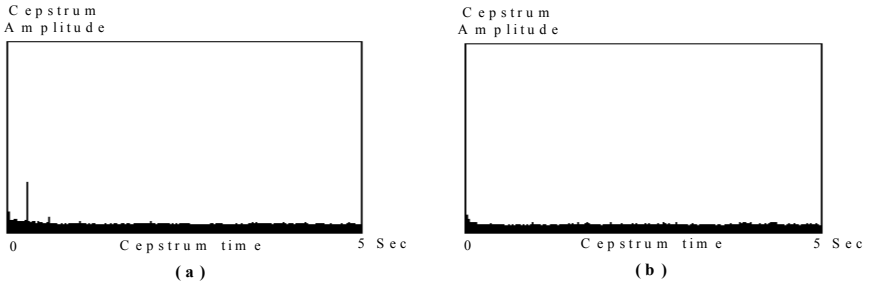


Fig. 17. (a) is a typical 5 seconds speech signal in cepstrum domain, the pitch peak appears near zero and (b) is a typical 5 seconds music signal in cepstrum domain.

5. Conclusion

In this paper, a general review of the common classification algorithms was presented and some were discussed in details. The approaches dealt with classification were divided into three categories: time-domain, frequency-domain, and time-frequency domain approaches. The time-domain approaches were: the ZCR, the STE, the ZCR and the STE with positive derivative, with some of their modified versions, and the neural networks. The frequency-domain approaches were: spectral centroid, variance of the spectral centroid, spectral flux, variance of the spectral flux, roll-off of the spectrum, cepstral residual, and the delta pitch. The time-frequency domain approaches were: the spectrogram and the evolutionary spectrum. We observed that the multidimensional classifiers that we have reviewed provide excellent and robust discrimination between speech and music signals in digital audio. The final decision of which feature should be selected depends on the application. The algorithms of the first category are faster; however, those of the second one are more precise. Finally, we conclude that more research is needed on the methods humans use to solve these sorts of classification problems, and what is the best way of implementing those or other strategies in pattern-recognition systems. Moreover, some separation algorithms were also introduced.

References

[1] Tzanetakis, G. and Cook, P. "Multifeature Audio Segmentation for Browsing and Annotation." *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 1999.

[2] Martin, K. "Towards Automatic Sound Source Recognition: Identifying Musical Instruments." *Proc. NATO Computational Hearing Advanced Study Institute*, Italy, 1998.

[3] Herrera, P.; Amatriain, X.; Batlle, E. and Serra, X. "Towards Instrument Segmentation for Music Content Description: A Critical Review of Instrument Classification Techniques." *Proceedings of the International Symp. on Music Information Retrieval (ISMIR),* Plymouth, MA, October 2000.

[4] Gjerdingen, R.O. "Using Connectionist Models to Explore Complex Musical Patterns." *Computer Music Journal*, 13, No. 3 (1989), 67-75; also in: P.M. Todd & D.G. Loy (Eds.), *Music and Connectionism*. Cambridge, MA: MIT Press, 1991.

[5] Hörnel, D. and Ragg, T. "Learning Musical Structure and Style by Recognition, Prediction and Evolution." In: D. Rossiter (Ed.), *Proceedings of the International Computer Music Conference*. San Francisco, 1996.

[6] Leman, M. and Van Renterghem, P. "Transputer Implementation of the Kohonen Feature Map for a Music Recognition Task." *Proceedings of the Second International Transputer Conference: Transputers for Industrial Applications II,* Antwerp: BIRA, 1989.

[7] Al-Atiyah, A. *Music and Speech Separation*. MS Thesis, King Saud University, Riyadh, Saudi Arabia, 2002.

[8] Stubbs, R. and Summerfield, Q. "Two Voice-separation Algorithms." *J. Acoustical Society of America*, 89 (March 1991), 1383-1393.

[9] Lee, T-W. and Koehler, B-U. "Blind Source Separation of Non-linear Mixing Modes." *IEEE*, (1997), 406-415.

[10] Lee, T-W. and Orglmeister, R. "A Contextual Blind Separation of Delayed and Convolved Sources." *IEEE ICASSP'97*, Munich, Germany, 1997, 1199-1202.

[11] Lee, T-W.; Bell, A. and Lambert, R. "Blind Separation of Convolved and Delayed Sources." *Advance in Neural Information Processing System,* Cambridge, MA, USA: MIT Press, 1997.

[12] Lee, T-W.; Bell, A. J. and Orglmeister, R. "Blind Source Separation of Real Word Signals." *IEEE ICNN,* Houston, USA, (1997), 2129-2134.

[13] Wang, D. L. and Brown, G. J. "Separation of Speech from Interfering Sounds Based on Oscillatory Correlation." *IEEE Transaction on Neural Networks,* 10, No. 3 (May 1999), 684-697.

[14] Leman, M. "The Theory of Tone Semantics: Concept, Foundation, and Application." *Minds and Machines*, 2, No. 4 (1992), 345-363.
[15] Patel, A.D.; Gibson, E.; Ratner, J.; Besson, M. and 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. and Latimer, C. "A Comparison of Connectionist Models of Music Recognition and Human Performance." *Minds and Machines*, 2, No. 4 (1992), 379-400.
[17] Weigend, A.S. "Connectionism for Music and Audition." In: J. Cowan, G. Tesauro and J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6*. San Francisco: Morgan Kaufmann, 1994.
[18] Anagnostopoulou, C. and 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. and Shimbo, M. "A Modified LEGION Using a Spectrogram for Speech Segregation." *IEEE International Conference on Systems, Man, and Cybernetics (SMC'99),* Tokyo, Japan, 1999, I526-I531.
[20] Andre, J. W.; Kouwe, V. D.; Wang, D. and Brown, G. J. "A Comparison of Auditory and Blind Separation Techniques for Speech Segregation." *IEEE Transaction on Speech and Audio Processing,* 9, No. 3 (March 2001), 189-195.
[21] Wang D. L. and Brown G. J. "Speech Segregation on Sound Localization." *IEEE 2001 Proceedings (IJCNN '01), International Joint Conference on Neural Networks,* Washington, DC, 15-19 July 2001, 2861-2866.
[22] Belouchrani, A.; Aben-Meraim, K.; Cardoso, J. F. and Moulines, E. "A Blind Source Separation Technique Using Second Order Statistics." *IEEE Trans. Signal Processing,* 45 (Feb. 1997), 434-444.
[23] Govindarajan, K.K.; Grossberg, S.; Wyse, L.L. and 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, USA, 1994.
[24] Kahrs, M. and Brandenburg, K. *Application of Digital Signal Processing to Audio and Acoustics*. Bosten/Dordrecht/London: Kluwer Academic Puplisher, 1998.
[25] Backus, J. *The Acoustical Foundations of Music*. 2nd ed., W.W. Scranton, Pennsylvania, U.S.A.: Norton & Company, 1977.
[26] Gang, D.; Lehmann, D. and Wagner, N. "Harmonizing Melodies in Real-time: The Connectionist Approach." *Proceedings of the International Computer Music Conference*, San Francisco, 1997, 27-31.
[27] Kaipainen, M.; Toiviainen, P. and Louhivuori, J. "A Self-organizing Map that Recognizes and Generates Melodies." In: P. Pylkkänen and P. Pylkkö (Eds.), *New Directions in Cognitive Science*. Finland: Publications of the Finnish Artificial Intelligence Society (FAIS), 1995, 286-315.
[28] Port, R. and 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, No. 4 (1995), 399-413.
[30] Cook, N. *A Guide to Musical Analysis*. UK: 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. and Chaudhari, U. "A Hierarchical Approach to Large-scale Speaker Recognition." *IEEE ICASSP'99*, Phoenix, Arizona, March 1999.
[33] Rabiner, L. and Juang, B. H. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[34] Bateman, W. *Introduction to Computer Music*. New York: John Wiley & Sons, 1984.
[35] 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, No. 1 (1992), 65-73.
[36] Horner, A. and Goldberg, D.E. "Genetic Algorithms and Computer-assisted Music Composition." In: B. Alphonce and B. Pennycock (Eds.), *Proceedings of the 1991 International Computer Music Conference*. San Francisco, 1991, 479-482.
[37] McIlwain, P. "The Yuri Program: Computer Generated Music for Multi-speaker Sound Systems." *Proceedings of the ACMA Conference,* Melbourne, Australia, 1995, 150-151.

[38] Kedem, B. "Spectral Analysis and Discrimination by Zero-crossings." *Proceedings of IEEE*, 74, No. 11 (Nov. 1986), 1477-1492.
[39] Ainsworth, W. A. *Speech Recognition by Machine*. London: Peter Peregrinus Ltd., 1988.
[40] Muthusamy, Y. K.; Barnard, E. and Cole, R. A. "Reviewing Automatic Language Identification." *IEEE Signal Processing Magazine*, (October 1994), 33-41.
[41] Ladefoged, P. *Elements of Acoustic Phonetics*. Chicago, IL, USA: University of Chicago Press, 1962.
[42] Fry, D. B. *The Physics of Speech*. Chicago, IL, USA: Cambridge University Press, 1979.
[43] Simon, J. C. *Spoken Language Generation and Understanding: Proceedings of the NATO Advanced Study Institutes*. Hingham, MA, USA: D. Reidel Publi. Co., 1980.
[44] Linster, C. "Rhythm Analysis with Backpropagation." In: R. Pfeifer, Z. Schreter, F. Fogelman-Soulie and L. Steels (Eds.), *Connectionism in Perspective*. North-Holland: Elsevier Science Publishers B.V., 1989.
[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. and 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 and A.S. Weigend (Eds.), *Proceedings of the Connectionist Models Summer School*. Hillsdale, NJ: Erlbaum, 1993.
[48] Young, P. H. *Electrical Communication Techniques*. 2nd ed., New York: Merrill Pub. Co., 1990.
[49] Laine, P. "Generating Musical Patterns Using Mutually Inhibited Artificial Neurons." *Proceedings of the International Computer Music Conference,* San Francisco, 1997.
[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 and C. Roads (Eds.), *Representations of Musical Signals*. Cambridge, MA: MIT Press, 1991.
[52] Griffith, N. and Todd, P. M. *Musical Networks*. Bradford Books, Cambridge, MA, USA: The MIT Press, 1999.
[53] Pierce, J. R. *The Science of Musical Sound*. 3rd ed., New York, USA: W.H. Freeman and Company, 1996.
[54] Lerdahl, F. and Jackendoff, R. *A Generative Theory of Tonal Music*. Cambridge: MIT Press, 1983.
[55] Monelle, R. *Linguistics and Semiotics in Music*. Chur, Switzerland: Harwood Academic Publishers, 1992.
[56] Gang, D. and 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.
[57] Bharucha, J. "Tonality and Expectation." In: R. Aiello (Ed.), *Musical Perceptions*. New York: Oxford University Press, 1994.
[58] Feiten, B. and Ungvary, T. "Organizing Sounds with Neural Nets*." Int. Computer Music Conference*, San Francisco, 1991.
[59] Foote, J. T. "Content-based Retrieval of Music and Audio." *SPIE'97*, San Diego, California, 1997, 138-147.
[60] Saunders, J. "Real-time Discrimination of Broadcast Speech/Music." *IEEE ICASSP'96*, Atlanta, Georgia, 1996, 993-996.
[61] El-Maleh, K.; Samoulian, A. and Kabal, P. "Frame-level Noise Classification in Mobile Environment." *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing,* Phoenix, Arizona, March 1999.
[62] El-Maleh, K.; Klein, M.; Petrucci, G. and Kabal, P. "Speech/Music Discriminator for Multimedia Application." *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing,* Istanbul, June 2000.
[63] Matityaho, Benyamin and Furst, Miriam. "Neural Network Based Model for Classification of Music Type." *IEEE Cat. No. 95*, 1995.
[64] Hoyt, J. D. and Wechsler, H. "Detection of Human Speech Using Hybrid Recognition Models." *IEEE Conference B: Computer Vision and Image Processing, Proceedings of the 12th IAPR International Conference on Pattern Recognition,* 2, No. 9-13 (1994), 330-333.
[65] Scheirer, E. and 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] Al-Shoshan, A.; Al-Atiyah, A. and Al-Mashouq, K. "A Three-level Speech, Music, and Mixture Classifier." *Journal of King Saud University* (*Engineering Sciences),* 16, No. 2 (1424), 319-332.

[67] Berger, J. and 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.
[68] Gasser, M.; Eck, D. and Port, R. “Meter as Mechanism: A Neural Network that Learns Metrical Patterns.” In: M. Lynch (Ed.), *The Cognitive Science of Prosody*. North-Holland: Elsevier, 1997.
[69] Toiviainen, P.; Kaipainen, M. and Louhivuori, J. “Musical Timbre: Similarity Ratings Correlate with Computational Feature Space Distances.” *Journal of New Music Research*, 24, No. 3 (1995), 282-298.
[70] Jin, H.; Kubala, F. and Schwartz, R. “Automatic Speaker Clustering.” *Proc. of the Speech Recognition Workshop*, 1997.
[71] Meddis, R. and Hewitt, M. “Modeling the Identification of Concurrent Vowels with Different Fundamental Frequency.” *J. Acoust. Soc. Am.,* 91 (1992), 233-245.
[72] Raphael, C. "Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21, No. 4, April 1999.
[73] Ross, T.J. *Fuzzy Logic with Engineering Applications.* NJ: John Wiley & Sons Canada Ltd., 1995.
[74] Hyvarinen, A. and Oja, E. “Independent Component Analysis: Algorithms and Applications.” *Int. J. of Neural Networks*, 13, No. 4-5 (2000), 411-430.
[75] Akarte, N.J. *Music Composition Using Neural Networks*. Master's Thesis, University of Nevada, Reno, 1992.
[76] Barnard, E.; Cole, R.A.; Vea, M.P. and Alleva, F.A. “Pitch Detection with a Neural-net Classifier.” *IEEE Transactions on Signal Processing*, 3, No. 2 (1991), 298-307.
[77] Bellgard, M.I. and Tsang, C.P. “Harmonizing Music Using a Network of Boltzmann Machines.” *Proceedings of the Fifth Annual Conference of Artificial Neural Networks and Their Applications (Neuro-Nimes)*, Nimes, France, 1992, 321-332.
[78] Bellgard, M.I. and 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.
[79] Berger, J. and 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.
[80] Bharucha, J. “Neural Net Modeling of Music.” *Proceedings of the First Workshop on Artificial Intelligence and Music*, Menlo Park, CA, 1988.
[81] 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.
[82] Bharucha, J. “Pitch, Harmony and Neural Nets: A Psychological Perspective.” In: P.M. Todd and D.G. Loy (Eds.), *Music and Connectionism.* Cambridge, MA: MIT Press, 1991.
[83] Bharucha, J. and Olney, K.L. “Tonal Cognition, Artificial Intelligence and Neural Nets.” *Contemporary Music Review*, Vol. 4 (1989).
[84] Bresin, R. and Vedovetto, A. “Neural Networks for Musical Tones Compression, Control, and Synthesis.” In: *Proceedings of the International Computer Music Conference*, San Francisco, 1994.
[85] Bresin, R. and 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,* Phyladilphia, USA, 1994.
[86] Carpinteiro, O. “A Neural Model to Segment Musical Pieces.” In: E. Miranda (Ed.), *Proceedings of the Second Brazilian Symposium on Computer Music, 15th Congress of the Brazilian Computer Society*, 1995.
[87] Ciaccia, P.; Lugli, F. and Maio, D. “Using Neural Networks to Perform Harmonic Analysis in Music.” *The Fifth Italian Workshop on Neural Nets, WIRN VIETRI-92,* Singapore, 1992.
[88] Cosi, P.; DePoli, G. and Lauzzana, G. “Auditory Modeling and Self-organizing Neural Networks for Timbre Classification.” *Journal of New Music Research*, 23, No. 1 (1994), 71-98.
[89] Fedor, P. “Principles of the Design of D-neuronal Networks II: Composing Simple Melodies.” *International Journal of Neural Systems*, 3, No. 1 (1992), 75-82.
[90] Feiten, B. and Guenzel, S. “Automatic Indexing of a Sound Data Base Using Self-organizing Neural Nets.” *Computer Music Journal*, 18, No. 3 (1994), 53-65.
[91] 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.

[92] Feulner, J. "Neural Networks that Learn and Reproduce Various Styles of Harmonization." *Proceedings of the International Computer Music Conference,* San Francisco, 1993.
[93] Gang, D. and Lehmann, D. "An Artificial Neural Net for Harmonizing Melodies." *Proceedings of the International Computer Music Conference*, San Francisco, 1995.
[94] Gjerdingen, R.O. "Categorization of Musical Patterns by Self-organizing Neuronlike Networks." *Music Perception*, 7, No. 4 (1990), 339-370.
[95] Laden, B. "A Parallel Learning Model for Pitch Perception." *Journal of New Music Research*, 23, No. 2 (1994), 133-144.
[96] Laden, B. and Keefe, B.H. "The Representation of Pitch in a Neural Net Model of Pitch Classification." *Computer Music Journal*, 13, No. 4 (1989), 12-26; also in: P.M. Todd & D.G. Loy (Eds.), *Music and Connectionism.* Cambridge, MA: MIT Press, 1991.
[97] Leman, M. "Artificial Neural Networks in Music Research." In: A. Marsden & A. Pople (Eds.), *Computer Representations and Models in Music.* London: Academic Press, 1991.
[98] 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.
[99] Mourjopoulos, J.N. and Tsoukalas, D.E. "Neural Network Mapping to Subjective Spectra of Music Sounds." *Journal of the Audio Engineering Society*, 40, No. 4 (1992), 253-259.
[100] Cohen, M.A.; Grossberg, S. and Wyse, L.L. "A Spectral Network Model of Pitch Perception." *Journal of the Acoustical Society of America*, 498, No. 2 (1995), 862-879.
[101] 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.
[102] Palmieri, F. "Learning Binaural Sound Localization through a Neural Network." *Proceedings of the IEEE Seventeenth Annual Northeast Bioengineering Conference,* 1991.
[103] 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.
[104] Röbel, A. "Neural Network Modeling of Speech and Music Signals." In: M.C. Mozer, M.I. Jordan and T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press, 1997.
[105] Sano, H. and Jenkins, K.B. "A Neural Network Model for Pitch Perception." *Computer Music Journal*, 13, No. 3 (1989), 41-48; also in P.M. Todd & D.G. Loy (Eds.), *Music and Connectionism.* Cambridge, MA: MIT Press, 1991.
[106] 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.
[107] Taylor, I.J. and Greenhough, M. "Neural Network Pitch Tracking over the Pitch Continuum." In: E. Michie (Ed.), *Proceedings of the International Computer Music Conference*. San Francisco, 1995.
[108] Trubitt, D.R. and Todd, P.M. "The Computer Musician: Neural Networks and Computer Music." *Electronic Musician*, 7, No. 1 (1991), 20-24.
[109] Rossignol, S.; Rodet, X.; Soumagne, J.; Collette, L. and Depalle, P. "Feature Extraction and Temporal Segmentation of Acoustic Signals." *Proceedings of the International Computer Music Conference (ICMC-98)*, San Francisco, 1998.
[110] Bogert, B. P.; Healy, M. J. R. and Tukey, J. W. *The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-cepstrum, and Saphe Cracking.* New York: John Wiley and Sons, 1963.
[111] Eronen, A. and Klapuri A. "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features." *Proc. ICASSP,* Istanbul, Turkey, 2000, 753-756.
[112] Cosi, P.; DePoli, G. and Prandoni, P. "Timbre Characterization with Mel-cepstrum and Neural Nets." *Proceedings of the International Computer Music Conference,* San Francisco, 1994, 42-45.
[113] 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.
[114] Taylor, I. and Greenhough, M. "An Object Oriented ARTMAP System for Classifying Pitch." *Proceedings of the International Computer Music Conference,* San Francisco, 1993.
[115] Mu, G. and Wang D. L. "An Extended Model for Speech Segregation." *Proceedings of INNS-IEEE International Joint Conference on Neural Networks (IJCNN01),*Washington, DC, 2001, 1089-1094.
[116] Priestley, M. B. *Non-linear and Non-stationary Time Series Analysis*. New York, NY: Academic Press, 1988.

[117] Al-Shoshan, A.I. “LTV System Identification Using the Time-varying Autocorrelation Function and Application to Audio Signal Discrimination.” *ICSP02*, Bejing, China, 2002.
[118] Scarborough, D.L.; Miller, B.O. and Jones, J.A. “Connectionist Models for Tonal Analysis.” *Computer Music Journal*, 13, No. 3, (1989), 49-55; also in: P.M. Todd & D.G. Loy (Eds.), *Music and Connectionism.* Cambridge, MA: MIT Press, 1991.
[119] Shuttleworth, T. and Wilson, R. “A Neural Network for Triad Classification.” In: I: E. Michie (Ed.), *Proceedings of the International Computer Music Conference.* San Francisco, 1995.
[120] Sergent, J. “Mapping the Musician Brain.” *Human Brain Mapping*, 1 (1993), 20-38.
[121] Scarborough, D.L.; Miller, B.O. and Jones, J.A. “On the Perception of Meter.” In: M. Balaban, K. Ebcioglu & O. Laske (Eds.), *Understanding Music with AI: Perspectives in Music Cognition*. Cambridge, MA: MIT Press, 1992.
[122] Wang, D. L. “Primitive Auditory Segregation Based on Oscillator Correlation.” *Cognit. Sci.,* 20 (1996), 409-456.
[123] Walpole, R. E. and Myers, R. H. *Probability and Statistics for Engineer and Scientists*. 5th ed., London, UK: Macmillan Publishing, 1993.
[124] West, C. and Cox, S. "Features and Classifiers for the Automatic Classification of Musical Audio Signals." *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004, 531-537.
[125] Pope, S. T.; Holm, F. and Kouznetsov, A. "Feature Extraction and Database Design for Music Software." *Proceedings of the International Computer Music Conference*, Miami, FLA, USA, 2004, 596–603.
[126] McKay, C. and Fujinaga, I. "Automatic Genre Classification Using Large High-level Musical Feature Sets.” *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004, 525-530.
[127] Essed, S.; Richard, G. and David, B. "Musical Instrument Recognition Based on Class Pairwise Feature Selection." *Proceedings of the International Conference on Music Information Retrieval,* Barcelona, Spain, 2004, 560-568.
[128] Downie, J. “The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future.” *Computer Music Journal*, 28, No. 2 (2004), 12-33.
[129] West, K. and Cox, S. “Finding an Optimal Segmentation for Audio Genre Classification." *Proceedings of the 6th Int. Symposium on Music Information Retrieval*, University of London, 2005.
[130] Tzanetaki, George. “Music Information Retrieval.” *ICASSP2005,* Tutorial TUT-5, Philadelphia, 2005.

مسح شامل لطرق فصل وتصنيف الإشارات الصوتية و الموسيقية

عبدالله إبراهيم الشوشان

قسم علوم الحاسب، كلية الحاسب، جامعة القصيم، القصيم، المملكة العربية السعودية

(قدِّم للنشر في ٢٠٠٣/٠٨/٠٤م؛ وقبل للنشر في ٢٠٠٦/٠٤/١٦م)

ملخص البحث. شهد تصنيف وفصل الإشارات الصوتية والموسيقية تقدما ملحوظا خلال العقد الماضي، ونظرا لأهمية تصنيف أو فصل الإشارات الصوتية إلى مكتبتين مختلفتين: مكتبة صوتية، ومكتبة موسيقية، فقد تم خلال هذا البحث مراجعة شاملة لطرق التصنيف وطرق الفصل لهذين النوعين من الإشارات، حيث أن الفصل بينهما قد تدعو الحاجة إليه في حالة الحديث المختلط مع وجود عدة إشارات في آن واحد. ولقد تم خلال هذا البحث دراسة مقارنة وتحليل رياضي لطرق التصنيف، حيث تم تقسيمها إلى ثلاثة أجزاء رئيسة كالتالي: (١) طرق التصنيف في مجال الزمن الحقيقي، (٢) طرق التصنيف في مجال التردد، (٣) طرق التصنيف في مجال الزمن-التردد. ونشير هنا إلى أن طرق التصنيف في مجال الزمن-التردد لم يحصل على بحث مكثف خلال الفترة الماضية، لذا فقد تم خلال هذا البحث اقتراح طريقتين للتصنيف: طريقة باستخدام Spectrogram والطريقة الثانية باستخدام Evolutionary Spectrum. هذا بالنسبة للتصنيف، أما عملية الفصل بين النوعين المذكورين من الإشارات فلم تحصل أيضا على بحث مكثف خلال الفترة الماضية، لذا تم عرض طريقتين مبسطتين في هذا المجال.