An overview on Perceptually Motivated Audio Indexing and Classification

Gaël Richard, *Senior Member, IEEE,*, Shiva Sundaram, *Member, IEEE* and Shrikanth (Shri) Narayanan, *Fellow, IEEE*

Abstract

An audio indexing system aims at describing audio content by identifying, labeling or categorizing different acoustic events. Since the resulting audio classification and indexing is meant for direct human consumption, it is highly desirable that it produces perceptually relevant results. This can be obtained by integrating specific knowledge of the human auditory system in the design process to various extent. In this paper, we highlight some of the important concepts used in audio classification and indexing that are perceptually motivated or that exploit some principles of perception. In particular, we discuss several different strategies to integrate human perception including 1) the use of generic audition models, 2) the use of perceptually-relevant features for the analysis stage that are perceptually justified either as a component of a hearing model or as being correlated with a perceptual dimension of sound similarity, and 3) the involvement of the user in the audio indexing or classification task. In the paper, we also illustrate some of the recent trends in semantic audio retrieval that approximate higher level perceptual processing and cognitive aspects of human audio recognition capabilities including affect-based audio retrieval.

Index Terms

Audio indexing, Audio classification, Music indexing, musical timbre recognition, semantic audio retrieval, music information retrieval, affect-based audio retrieval, perceptual audio features, perceptual signal representations.

I. Introduction

The development of the Internet and social media and the availability of affordable content-capture devices such as cameras and mobile phones have led to the availability of huge amounts of digital multimedia content that are both professionally produced and user generated. The need to quickly and efficiently access this content has inspired growing research effort in automatic processing of audio. An audio indexing system aims at describing the audio content by identifying and labeling, for example, the acoustic sources, their time of occurrence or automatically grouping audio clips into known recognizable categories.

There are two closely related domains attached with the term *audio processing*: *content-based audio processing* and *computational auditory scene analysis* (CASA). In CASA, as the term suggests, the final objective is to completely compute and even segregate the acoustic sources based on source type or location in a real environment or scene. The auditory scene is computed using partial or noisy recording of the scene. For instance, this may

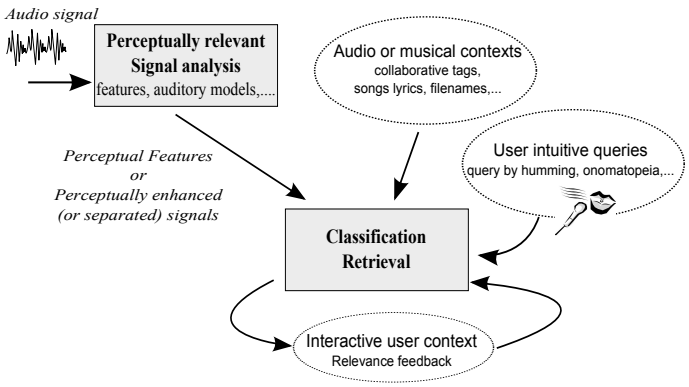

Fig. 1. Human perception and interaction in an audio classification system.

involve extracting speech emanating from a specific direction in the presence of other (noisy) sources to improve the robustness of an automatic speech recognition system.

In content-based audio processing the main objective is to automatically understand the high-level impression that an audio clip would create if a human listened to it. While like CASA, this also requires a deep understanding of perception of audio for estimating the source it also goes beyond that to understand and model the ontological organization of audio by humans. Content-based audio processing is therefore couched as a pattern recognition problem that involves automatic classification, segmentation and clustering of audio. The final objective here is to create efficient audio retrieval systems (e.g. searching Internet for similar urban sounds, ...) or Music Information Retrieval systems (e.g. searching for music songs of similar evoked emotion, ...). As humans are the primary end consumers of content, it is generally desirable that automatic audio content processing systems operate in a perceptually meaningful manner making *opaque* (without the sensory experience of listening) audio clips automatically more transparent to interpretation by the user.

There exist several strategies that can be combined to integrate human perception in the audio processing pipeline. First, it is known that there are complex differences between the physics of sounds and their perception by humans. A first strategy then is to exploit human perception knowledge in the audio analysis stage by means of generic audition models or through the use of perceptually-relevant features that can characterize specific aspects of the audio content (see Fig. 1).

A different strategy would be to build signal features that are correlated with perceptual test results. That means computing a low-level feature on the signal that allows one to explain a high-level perceptual dimension of sound similarity (as shown in subjective experiments) can also be defined as perceptually-motivated. As a consequence, these features may have no direct link with hearing models, however, since their perceptual relevance is justified by global similarity perception tests they often turn out to be particularly efficient features for audio indexing or similarity classification applications. In this paper, both strategies to build perceptually-motivated features will be

discussed.

Another current research trend (especially in audio retrieval) aims at directly involving the user to better take into account, in some sense, her/his perception of the involved sound similarities and their cognitive and affective influences. One of the strategies consists of using concurrent sources of information and in particular user-generated textual meta-information (filenames, captions, or more importantly user attributed comments or labels). The involvement of the user can also be in the form of intuitive queries such as, for instance, extending traditional text labels by more elaborate queries such as those based on onomatopoeia words or on generic language-level descriptions. Finally, the user may be directly involved in the retrieval process by iteratively exploiting user feedback on the analysis or retrieval results obtained at a previous stage (in a so-called *relevance feedback* paradigm). This permits to either perceptually adapt the classifier or the feature set at each iteration while performing the task at hand.

The goal of this paper is to highlight some of the important concepts used in audio classification and indexing that are perceptually motivated or that exploit some principles of perception. The focus of the paper is more on the early auditory processing (or low level initial perceptual stages) with limited concern to the higher stages of cognition (even if the discussion on semantic and affect-based retrieval does refer to some of these higher level perceptual processing). This overview is also limited herein to general audio and music signals and will not discuss approaches specifically dedicated to speech signals (recognition, understanding, translation, ...); topics that are considered in greater detail in other papers in this volume.

The paper is organized as follows. In Section II, we describe some of the basic perceptually motivated signal representations that can be used as the initial stage of any audio classification system. Then, in Section III, some perceptually-relevant solutions for audio classification are discussed with examples taken from two challenging applications namely audio categories recognition and musical timbre recognition.

This section concludes with a discussion of the two alternatives between a global analysis approach (justified in cases where the audio stream is perceived as a unique stream) and a segregated approach where the audio signal is analyzed "stream by stream" after some kind of source separation (justified in cases where individual streams can be perceived separately by the user). Then in Section IV, we present a discussion on specific aspects linked to user-based audio retrieval (user query by text/labels, humming, onomatopeia,...) and affect-based classification.

Finally, we conclude the paper by outlining future directions for better integration of perceptually motivated concepts in audio indexing schemes.

II. Perceptually Motivated Signal Representations

An audio signal is the time-varying voltage at a microphone input. The voltage values are a function of pressure changes caused by an acoustic wave generated by a sound source (for e.g., the surface of a table or the mouth of a speaker). Audio signals have rich time-varying characteristics, often diverse and heterogeneous, and hence analysis techniques both in the native time domain and transform Fourier domain have been developed to capture their dynamic details.

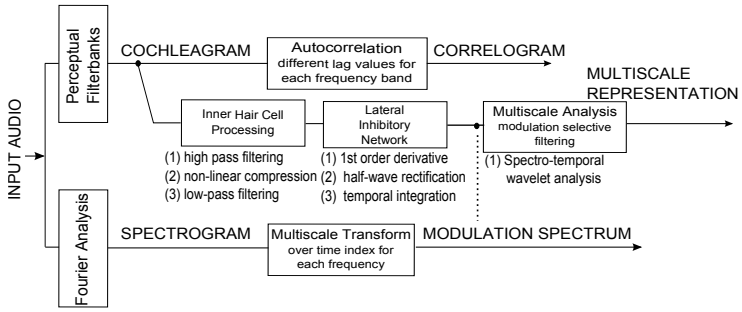

Fig. 2. A high-level overview of various transforms used to derive perceptually relevant acoustic features.

In this section, we mainly describe signal features i.e., signal transforms common in many audio processing applications. Using such transformed representations, both statistical and perceptually motivated signal features can be derived. The principles of human perception are based on quantitative results obtained by extensive subjective experiments performed in the field of psychoacoustics, as well as direct observations and modeling of the physiology of hearing in animals and humans. These principles are incorporated algorithmically by appropriate modifications to time-frequency signal representations. In the first part of this section, the basic time-frequency analysis is presented. Subsequently, specific signal measures that are popular in audio indexing and content processing tasks are also discussed. Although we are far from our end goal in having a single universally applicable solution to perceptually motivated processing of audio, there are notable commonalities in the pipeline used to model human audition (illustrated in Fig.2). Furthermore, it is important to note that the signal analysis schemes presented here are also typical in other audio applications such as audio coding or compression [1]. In compression, the objective is to derive a compact representation of audio that contains perceptually relevant information so as to allow reconstruction of the audio signal with minimum perceptual difference to the original audio signal. For audio indexing, the feature representations do not need to necessarily allow effective signal reconstruction but they are designed to robustly discriminate various acoustic sources or desired attributes.

A. Time-Frequency Analysis

Due to wide variability and heterogeneity in its properties, an audio signal is generally characterized as a non-stationary random process. For example, a spoken word utterance such as *soup* is a sequence of smaller voiced (the medial vowel /u/ in the example) and unvoiced segments (the noise-like word initial fricative /s/ and the transient word-final /p/) that have distinct statistical signal properties. Similarly, an audio recording of a street scene can include vehicle sounds, sirens, human speech, footsteps, music and animal/bird sounds, leading to a complexly composed signal. To simplify analysis, audio is assumed to be comprised of different stationary processes of short duration that are indexed together on the same time line. For this reason, analysis or feature transformation is carried out on a shorter frame-by-frame basis; the discrete Fourier transform (DFT) is applied to a single frame of audio that is few tens of milliseconds long and subsequently the frame is then shifted by few tens of milliseconds

and the analysis is repeated again. This analysis scheme allows the processing system to access the local time and frequency components as a sequence of signal measures (multidimensional frequency measures) that extract local properties of the audio signal.

It is important to note that the DFT analyses the signal in a linear frequency scale. with variable resolution. In the low and mid frequencies (up to $\approx$ 1 kHz) frequency components are grouped more evenly whereas in the higher frequencies, the bands have larger frequency spread leading to fewer bands covering a larger part of the auditory spectrum. As discussed in the next subsection, this aspect of non uniform analysis is typically implemented as a filter bank, i.e., a bank of band pass filters with different center frequencies and width that measure the collective energy output of their respective frequency components.

B. Cochleagram

Extensive noise masking experiments in the field of psychoacoustics have shown that the auditory system, in particular the cochlea in the inner ear acts as a bank of overlapping band pass filters (the reader is referred to the chapter 2 in the book by Moore [2] for details). The center frequency of the filters follows a log scale and the bandwidth of these filters (the *critical bands*) are narrow for low frequencies and significantly wider at higher frequencies. Interestingly, this allows for higher time resolution for high-frequency components and high frequency resolution for low-frequency components. Signals that are present within a critical band have been known to activate the same region in the basilar membrane in the cochlea [2]. Furthermore, it has also been determined that perceived intensity of noise signals' spread within a critical band remains constant and the loudness perception increases as the energy is spread to the neighboring bands [1], [2]. Using these empirical principles, perceptually motivated signal measure computations begin by collecting the magnitude of frequency components within a critical bank of an imposed *auditory filter bank*. This results in analysis of the time domain audio signal in to individual frequency components that have tonotopic correspondence to processing in the basilar membrane in the cochlea. Specifically, the center frequency and the bandwidth of such an auditory filter bank follow a *bark* frequency scale [3] where the auditory spectrum is divided in to 24 overlapping band pass filters[1]. Another scale that is based on the relationship of pitch sensation and frequency is the *mel* scale [3], [4]. Like the bark scale, the mel scale is also based on subjective experiments; in most signal processing implementations, the mel scale is designed to be approximately linear up to 1 kHz and logarithmic above that [5].

There are a handful of popular schemes for implementing the filter bank analysis model of the cochlea. These include using Gammatone filters [6]–[8], the cascade/parallel filter bank model originally proposed by Lyon [9] or those that simply impose band-pass filtering to the DFT [10]. As this approach models the frequency selectivity of the cochlea, the resulting time-frequency output of the filterbanks is referred to as a *cochleagram*. It mainly models the mechanical frequency response of the basilar membrane.

[1]It is important to note that the time-frequency analysis using DFT by FFT typically results in 512 or 1024 frequency components per frame of audio. Filter bank analysis leads to 24 components.

C. Correlogram

The filterbank analysis or the modulation representation performs coarse analysis of the frequency spectrum and actually smooths over finer pitch structure evident in the firing activity of the auditory nerve. Using the autocorrelation between the different time domain outputs of the filterbank, a *correlogram* can be used to model the perception of the fundamental frequency (F0)[2]. This results in a *lag-frequency* map of a windowed audio segment that also extracts the higher order harmonics present in the signal. A typical approach to extract the information regarding the fundamental and higher order harmonics is by summing the normalized outputs across various frequencies of analysis. A comprehensive treatment of correlogram and its application in audio segmentation and source separation can be found in the book by Wang and Brown [11].

D. Modulation Spectrum

While the above time-frequency schemes are based on directly measuring the local time energies at the outputs of the filter banks, it is also possible to extract long duration variations in the signal by applying another transform independently on each frequency component along the time axis. The second transform can be performed at different scales to represent the audio segment at different temporal resolution while maintaining the frequency resolution [12]. Furthermore, features for indexing or classification can be extracted by retaining only the non-zero low frequency *modulation* components [13]. Such modulation spectrum based features are perceptually relevant as the low frequency modulation components are important for signal intelligibility and can retain slow changing information such as phonemic or syllabic structure of speech [12], [13]. It has also been shown that the auditory system is sensitive to modulations of signals in addition to the spectral frequency. Psychoacoustic evidence can be found that support the use of features that extract modulation frequencies [14]. Modulation features have been well studied in speech processing [15], [16], audio identification/retrieval [13] and problems in music information retrieval such as genre classification [17].

E. Multiscale Representations

All the representations discussed above illustrate the sensory processes of the early auditory system. More advanced signal measures that model the central stages of the auditory system leading to a multi-scale time-frequency representation have also been explored by Yang et al. [18] and Mesgarani et al. [19]. After the cochlear processing stages (the analysis in the cochleagram), the transduction (i.e., the conversion of mechanical frequency responses in the basilar membrane to firing activity in the auditory nerve) is modeled as a series of transforms that functionally model the processing in the inner hair cells [18], [19] followed by a first order time derivative of the spatial (frequency) axis in the cochlea. Finally, a halfwave rectification and an integrator is applied to model the lateral inhibitory network (LIN) found in the auditory system. The LIN processing is also present in the vision

[2]The term pitch is related to perception of harmonics whereas the fundamental frequency (F0) is the frequency at which a physical source vibrates.

system where it highlights regions of fast variations in an image. Along this vein, in the auditory nerve, LIN processing has shown to highlight perceptually relevant features [18]. The additional processing in the LIN also includes the envelope (or modulation) information similar to modulation spectrum discussed in Section II-D.

Further processing by wavelet-like analysis using a two dimensional Gabor function that results in a multiscale cortical representation has been shown to perform well in speech/ non-speech discrimination [19] and perceptually close noise classification [20]. The *auditory attention model* approach by Kalinli et al. [21] adopts the visual attention model by Siagian et al. [22] to select a subset of salient acoustic events within an audio clip. This is obtained by mapping the time-frequency spectrogram of an audio signal (taken to be the *scene*) into a multi-scale representation using receptive filters that model intensity, frequency contrast, temporal contrast and orientation maps in the central auditory system.

The number of analysis components in the cortical representation is fairly large and redundant in comparison to the "feature" dimensions of say a cochleagram. To make machine learning tractable on these features and reduce redundancy, dimension reduction techniques are used to map the features to a lower dimensional representation. In Mesgarani et al. [19] the authors use tensor decomposition to obtain a lower dimensional feature set. In Kalinli et al. [21], a summary of the intensity, feature contrast etc. maps is first derived by averaging. Subsequently, the mean vectors are augmented together and a lower dimensional vector is derived by principal component analysis resulting in the *gist* of the input time-frequency acoustic scene. The gist extraction essentially models the pre-attentive process that guides the higher cognitive processing (typically implemented as task-dependent pattern recognition algorithm) to salient parts of the scene that stand out [21], [22].

The auditory signal representations presented above can be directly used to extract perceptually-relevant audio features or as an element of more sophisticated operations. A number of physical factors such as extraneous sounds (noise), early and late reflections due to surrounding walls or windows affect the purity and perception of sounds captured by the microphone. These in turn change the spectral and temporal spread of an acoustic source. The general idea in audio indexing and classification task or computational auditory scene analysis is to precisely identify, segment or separate predefined acoustic sources in spite of these variabilities. The complexity of information extraction depends on the levels of, and how closely the auditory stages are being modeled. As discussed in sections III and IV, successful experiments in speech and audio classification tasks have consistently confirmed the motivation for using perceptually motivated spectral features that predominantly model the processes in the early stages of the auditory system.

III. Perceptually Motivated Audio Classification

In this section, we first describe audio processing in general terms and present a classical scheme for building an audio classification system. Then, to illustrate the relevance of perception for audio indexing and classification, we will focus on two applications namely *audio sound categories recognition* and *music timbre recognition*. Finally, following results on the perception of individual audio sources in an audio scene, we discuss the two alternative

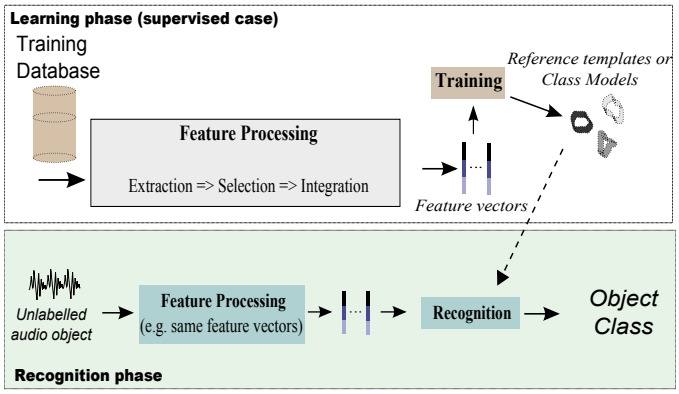

Fig. 3. General scheme for an audio classification system

strategies for audio and music classification that consider processing the signal either globally or as a mixture of sources that can be separated from the original signal.

A. General Scheme

The perception and interpretation of daily sounds (including music) involve very complex and yet largely unexplained processes in our auditory system. A complex acoustic scene can be, for example, described in terms of their elementary constituents, namely the sound sources (their nature and position in space, their semantic meaning if any) and the acoustical environment where the scene takes place (concert hall, small room, outdoors, etc). However, the perception or interpretation of an audio stream goes far beyond and involves a large number of aspects such as personal experience, knowledge and memory, social position or lifestyle [23], [24]. For example, numerous high level concepts are at the heart of music perception including instrumentation, melody, rhythm or harmony[3]. While most automatic processing systems already consider simple models of the human auditory system, significant gains in human-like processing have been achieved by considering intuitive applications combined with sophisticated pattern recognition algorithms.

The most widespread approaches for recognizing audio categories or classes rely on a rather classic pattern classification scheme which includes two major components: A feature extraction module and a classification or decision module (see Fig. 3).

For audio, designing appropriate features for a specific classification task is largely informed guess work. In many cases, the feature sets may be of very large dimensions ($\approx 10^{2-3}$ for example for the multi-scale feature sets discussed in section II-E). To make the concomitant pattern classifiers trainable, dimension reduction techniques such as principal component analysis (PCA) or linear discrimination analysis (LDA) are subsequently applied.

[3]*Instrumentation* refers to the musical instruments used in a given musical piece; *Melody* refers to the main musical line, for example the musical part sung by a lead singer; *Rhythm* encompasses a number of aspects related to the relative length and accentuation of the successive musical notes; and *Harmony* refers to the organisation of music in chords which are built from simultaneously sounding notes. More precise and detailed definitions may be found in many references including [25].

Another strategy is to rely on *feature selection* techniques which permit to obtain a reduced set of efficient features for the classification task at hand [26]–[28]. In some other cases, feature integration process can be further adopted to find a set of features that perform best [29].

Many studies aim at integrating perception principles directly in the design of acoustic features. Such features, called *perceptual features*, are either built

1) by including knowledge of the human ear in their design;
2) or by selecting signal characteristics that are best correlated with perception tests.

In the first case, the perceptual features can be directly computed from a generic audition model (see section II). For example, they can be directly obtained from perceptually-motivated filter banks such as filter banks on bark scales, Gammatone filter banks, Constant-Q transforms (see [30]–[32]), or from multi-scale or muti-linear representations (see [19], [33]). However, in many studies, the features are built by only using some concepts of perception and do not explicitly rely on a generic audition model. Since the generic audition models were already discussed in section II, this section will mostly focus on this latter design strategy.

In the second case, the objective is to find features which are correlated with perceptual judgment of similarity and dissimilarity of a set of sounds. Although such features do not model human audition directly, they are particularly interesting for perceptually-based indexing and classification tasks.

Concomitantly, perception principles from higher stages of cognition can also be integrated in the classification and decision modules. This is not further discussed herein, but for example, Neural networks, and more recently Bayesian networks and Hierarchical Temporal Models aim to represent the complex and hierarchical processing of information in the brain [34], [35].

B. Audio Categories Recognition

In this section, we highlight the use of some perceptually-motivated audio features using examples of distinct audio recognition problems.

A significant number of features used for audio categories recognition are signal driven and statistical. However, features such as spectral roll-off, spectral centroid and spectral flux carry perceptual significance [36]. Spectral centroid, for instance, has been shown to be directly linked to one of the important dimensions of sound similarity perception (see section III-C for additional details on musical instrument sounds). Other interesting features related to perception proposed by Scheirer et al. [37] includes the 4 Hz modulation energy features (using Mel scale filter banks). The 4 Hz modulation frequency is known to reflect the syllabic rate of speech [13].

However, the most common signal features that have consistently shown good performance are the Mel-frequency Cepstral Coefficients (MFCCs). MFCCs[4] were initially introduced for speech recognition [10] and characterize the

[4]Starting from the output of an auditory filter bank that uses a set of triangular filters designed on the mel scale, MFCC features are extracted by taking the logarithm of the magnitude square of the filter banks outputs and by applying a subsequent transformation such as the Discrete Cosine Transform (DCT). The DCT essentially gives a set of real-valued components that capture various local statistics of the filter bank outputs.

filter part (i.e, the resonances and anti-resonances of the vocal and nasal tracts) of the underlying source-filter model of speech production [38]. But MFCCs are often classified as perceptual features because they integrate certain aspects of perception. Indeed, the magnitude square of the filter bank outputs represent the instantaneous signal energy in a given band, and combined with the logarithm compression, these two transformations crudely approximate the processing in the inner hair cells in the cochlea. The use of MFCCs for music signals is further discussed in section III-C.

Other features, such as those obtained from cortical representations (see section II-E) have also been used with success by Mesgarani et al. [19]. Here the authors describe a system for discriminating speech from a variety of non-speech sounds of animal vocalizations, environmental sounds, and even music, and showed that the cortical representations are more robust than simpler signal-level features.

Speech/non-speech discrimination is one significant example where we seek to isolate components (mainly speech) from an audio stream [19], [37], [39]–[42]. However, audio classification systems can easily deal with a significantly larger number of acoustic categories. Notable examples include [43] and [44]. Here the audio processing systems are trained on categories such as *animals, bells, crowds, female, laughter, machines, male voices, percussion instruments, telephone, water* sounds etc. The main assumption here is that acoustic sources in an audio clip are homogeneous and therefore the content belongs to only one of the pre-defined categories. Nevertheless, they serve as a proof of efficacy of various perceptually motivated signal features [36].

An extension of this scheme considers structured categories, organised in a meaningful hierarchy from a perceptual or cognitive point of view (for example a meta category "urban sounds" that contains "car engine noise", "car horn" or "train passing" as basic categories). In a broader perspective, the acoustic sources belonging to different scenes (e.g. a train station or a park) can be assumed to be different. This allows the audio processing system to recognize *context* instead of the individual acoustic sources [45]–[48]. These aspects will be further discussed in section III-D.

C. Musical Timbre Recognition

To better understand the perception of music, numerous studies have focused on isolated sounds with an objective to highlight a few major characteristics that will permit to link the perceptual properties of sounds with their physical properties as measured by traditional instrumentation. This has permitted over the years to define basic dimensions of sounds amongst which Loudness, Pitch and Timbre are undoubtedly the most important for isolated sounds (see for example [23], [49]–[54]).

a) Timbre: Timbre, is a rather (computationally) elusive concept despite the numerous existing definitions[5]. Reference to the standard definition by ANSI *"Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar"* also reveals the difficulty of the problem since it is only described as what it is not [54], [55]. This difficulty of defining timbre is largely attributed to its multidimensional nature. Not surprisingly, numerous studies have

[5]http://acousticslab.org/psychoacoustics/PMFiles/Timbre.htm

concentrated on finding meaningful timbre spaces [56]–[59] that would provide a geometrical representation of the perceived similarity of the tested sounds. Most studies have obtained such timbre spaces by means of Multi Dimensional Scaling (MDS) [56], [60] and have permitted to highlight perceptually meaningful acoustical features. The results of these studies are not fully consistent due to a number of limitations concerning the sounds stimuli and experimental protocol, but it can still be argued that a number of features obtained within these studies do explain perceptual dimensions of timbre and have been successfully used in automatic instrument timbre recognition experiments. Since most of these studies have considered monophonic stimuli, some of the obtained features are not meaningful for more complex audio sounds composed of multiple sound sources and it will be important to increase the research effort on polyphonic timbre to better understand the perception of such complex signals [59]. Nevertheless, numerous features have been proposed for analysing polyphonic music, especially for music genre or music emotion analysis (see for example [61]–[63]).

We further discuss below common audio features that have already been used in automatic timbre recognition system with a specific focus on perceptual features.

b) Timbre features: The most widespread approaches for recognizing monophonic or polyphonic timbre (in music follow the general scheme provided in Fig. 3. As far as the features are concerned, MFCCs remain very popular for monophonic timbre recognition studies [32], [64], [65], and surprisingly also for polyphonic timbre recognition [31], [59], [66]. This is surprising since the initial design of MFCC was inspired by speech recognition. Notably, in music, the pitch range is much wider than in speech, and for high pitches the deconvolution property of MFCCs does not hold anymore. The consequence is that they are usually quite dependent on the pitch values as shown for example by Mesaros et al. in [67]. Their use in music signal processing is therefore not well justified but it seems that their capacity of MFCCs to capture "global" spectral envelope properties is the main reason of their success in such studies. Another deficiency for polyphonic music is that this global characterization prevents MFCCs to describe localised time-frequency information and in that sense they fail to model well-known masking properties of the ear. Numerous authors have confirmed that these features are not sufficient for describing the polyphonic timbre justifying the use of other, more specific, features. Alluri and Toiviainen [59] recently argued that they are indeed not highly correlated with any of the perceptual dimensions of polyphonic timbre despite their widespread use as predictors of perceived similarity of timbre.

Many of the acoustical features drawn from numerous studies on monophonic timbre perception proposed for automatic timbre recognition may not be easily used for polyphonic timbre recognition. This is the case for example of the Log-Attack time, the spectral harmonic structure, attack synchrony and spectral irregularity which are all linked to perceptual dimensions of timbre (see for example [30], [68], [69] for extended lists of features with their detailed description).

For complex or polyphonic audio signals, most of the above features are not suitable. The problem itself is even less well defined than for monophonic sounds and also depends on the complexity and nature of the audio scene. Similarly, to audio categories recognition, many studies consider that global features (e.g, features computed on the complex polyphonic music signal) are sufficient to recognize the polyphonic timbre, even if the timbre is usually

assumed to be built as an agglomerate of individual sound entities [70]. This aspect will be further discussed in the following section.

D. Global versus Separation-based Processing

When listening to a complex audio scene, humans tend to identify separate streams, or in other words to separate the different sound sources that they can then identify. However, in other situations the sound sources will tend to blend and therefore be perceived as a single complex audio category (or equivalently "polyphonic timbre" for music signals). This property is obviously largely used by composers (and especially in classical or electroacoustic music) to either let the listener perceive the global sound or the addition of separate sound objects [71], [72].

Most of the methods discussed above follow this so-called *global approach* but using simple categories of comparable semantic level. An alternative scalable scheme that targets wider applications is by organizing clips and audio categories in a meaningful hierarchy. In Zhang et al. [73], the authors present a system of hierarchical categories such as silence portions and portions with and without music components. The approach by Liu et al. [74] also uses a similar hierarchical scheme by grouping sources into speech, music or environmental sounds. In such a scheme, a given segment is first labeled as speech or non-speech. Then, speech segments are analysed to detect speaker changes while non-speech segments are further classified as music, environment sound or silence.

Another example in music is provided by [66] where musical instruments and group of instruments are organised in an hierarchical taxonomy that is automatically obtained from the training data. Obtained categories for jazz quartet music include for example a blend class describing combined sounds from bass, drum and a wind instrument or "pure" classes involving sounds from a single instrument.

If global approaches remain very popular in the literature, it is in many cases perceptually relevant to follow an alternative strategy where the individual sources or streams are first isolated and identified in a subsequent stage.

The source separation principles involved may be rather rudimentary in comparison to traditional CASA-based approaches. In this case, they do not aim at precisely recovering the different sources but rather aim at enhancing parts of the involved sources to help the subsequent stages of the classification engine. The most illustrative example of this is the use of (perceptual) filterbanks which can approximately separate sources that are well localised in frequency. Another example is the so-called Harmonic/noise decomposition [75], [76] which performs a rough separation of the analysed signal into a tonal component (i.e., sum of periodic sub-content) and an aperiodic component which then captures the stationary and impulsive noises such as sound attacks, door slam, gun shots etc.

However, a number of studies have investigated more sophisticated source separation principles to retrieve the individual streams especially for recognising predominant sources (see [77] for singing voice extraction, [78], [79] for music instrument recognition, [31] for monophonic source detection in complex audio streams, [80], [81] for Drum extraction or [82] for sound events detection).

IV. Semantic and Affect-based Retrieval

The previous sections described audio and music classification systems which can be used as core technologies for audio retrieval applications (e.g. searching Internet for similar water sounds) and Music Information Retrieval applications (e.g. searching music sounds of similar timbral characteristics, similar evoked emotion). To illustrate other recent trends in audio and music classification retrieval that align with higher level perceptual processing and cognitive aspects of human audio recognition capabilities, general problems in semantic audio retrieval are discussed in this section with a specific focus on affect-based retrieval. Note though, that even if it is not detailed in this section, systems discussed herein also exploit perceptually motivated acoustic features in their processing stages as those described in section II or those given in table I.

A. Semantic Audio Retrieval

Current trends in audio research (especially in audio retrieval) aims at directly involving the user to better take into account, in some sense, his/her perception of the involved sounds similarity. For example, it is possible to involve the user in an iterative recognition process (in a so-called relevance feedback paradigm). This permits to either perceptually adapt the classifier or the feature set at each iteration while performing the task at hand and improving results to the query [83]–[88].

A further level of sophistication is semantic audio retrieval where the content processing system is able to retrieve audio clips by approximate language-level descriptions which can be seen as a generalization of direct categorization by labeling discussed earlier. Furthermore, semantic context and affective factors for retrieval have also been considered. In most cases, the problem is to map a long audio clip of heterogenous sound sources to a single context or language-level label. This is generally achieved by empirical time integration of acoustic features obtained from short duration of audio segments that are a few milliseconds long or by observing longer term repeating patterns that can serve as both parametric or non-parametric acoustic model of the context or label.

Semantic Audio Retrieval attempts to model the largely unknown cognitive aspects of human auditory capabilities. Notable examples of methods that deal with semantic descriptions, and not just category labels of audio retrieval include the systems described by Slaney [89] and Barrington et al. [90]. These methods present a probabilistic approach to modeling the signal features of unstructured audio clips directly with terms in the text description instead of explicit classes or categories. Slaney [89] proposed an improvement on the direct labeling scheme by creating a mapping from each node of a hierarchical model in the abstract semantic space to the acoustic signal feature space. The nodes in the hierarchical model (represented probabilistically as words) are mapped onto their corresponding acoustic models. Barrington et al. [90], describe a similar approach of modeling features with text labels in the captions. As with many others, both techiques use MFCCs as their acoustic feature. Furthermore, longer duration information is incorporated by either stacking the short-term features together [89] or by using first and second order derivatives of the feature sets. By integrating temporal information, such techniques allow for discrimination based on both spectral content and temporal patterns of audio. This has been empirically found to be beneficial to the overall classsification or retrieval performance. Another line of research in semantic audio

retrieval is the text-like processing of audio where audio clips are assumed to be composed of acoustic units similar to words that compose a text document. The main idea here is to leverage the information in unit-document co-occurrence that is similar to word-document co-occurrence for text [91], [92]. By modeling audio as a collection of units, the approach is able to scale to (arbitrary) collection of audio clips. Notable examples include latent perceptual indexing by Sundaram et al. [93], [94], the related anchor-space model by Lu et al. [95] and Lee et al. [96] and the bag-of-patterns representation used by Lyon et al. [97]. These techniques formalize the method discussed in Slaney et al. [98] where a form of unit-document co-occurrence is implicitly used for semantic information extraction from audio. MFCCs are used at different scale levels to represent and measure changes in the overall timbral qualities of the audio track. This is combined with the video and text features, and the time series of these features after dimensionality reduction is used to segment media clips.

In Lu et al. [95] the authors use a document view of audio clips and proposed a method to connect semantic categories to signal features and subsequently device a similarity metric between audio clips using Kullback Leibler divergence (KLD). Here, probabilistic models for representing membership of audio clips in the anchor space (the semantic class) are used to map the clips as vectors in the anchor space. The models are built using acoustic features that is comprised of perceptual correlates such as spectral centroid, sub-frequency band energy ratio (see table I) along with MFCC. Herein, the second order statistics over a longer one second duration of features are used as a coarse form of temporal integration. Subsequently, the divergence measure is used to determine similarity between audio clips. This is advantageous (and common with the latent indexing framework) as it provides a mechanism to measure similarity between audio clips irrespective of the number of unknown acoustic sources present in them. The approach is predominantly unsupervised where the anchors are automatically discovered using the spectral clustering method described by Cai et al. [99].

In Sundaram et al. [93], the authors use the centroids of clusters of signal features as acoustic units and subsequently derive unit-document frequencies between the centroids and features extracted from audio clips. While the initial work was performed on semantic categories, this approach has also been investigated for onomatopoeic categorization of audio [94]. Another work that is based on this idea is by Chechik et al. [100] where the motivating application was audio retrieval using text-queries. Their extensive experiments on large audio data-sets reveal the potency of the acoustic information in unit-document frequencies.

In the work by Lyon et al. [97], a bag of repeated patterns from a modified form of autocorrelation (termed the *stabilized auditory image*) applied to the output of a filterbank simulating the cochlea. Collecting repeating patterns This was then combined with a passive-aggressive learning algorithm [101] to develop a robust sound retrieval system.

A sparse code or sparse representation of audio content can be derived by collecting repeating patterns (using a very large number of known patterns). Sparse codes are also known to be analogous to the coding mechanism in neural sensory system (see [97] and references therein). Interestingly, mathematical analogy between sparse representation of data and dimension reduction using matrix factorization (used in the bag-of-units representation by [93]) has also been observed in the context of speech recognition (the reader is referred to [102] for an overview).

Starting from perceptual features, retrieval techniques using these representations can therefore be used to further render higher level sensory processes in the auditory system.

For music signals, a specific line of research was also devoted to more sophisticated user entries such as under the form of a hummed query [103], [104], a sequence of onomatopeia [105] or finger-tapped rhythmic query [106]. Systems that generate sophisticated user tags that jointly describe high-level aspects of music listening such as mood, vocal quality, instrumentation and genre have also been successfully implemented [107].

B. Affect-based Retrieval

It is well known in the speech and audio research community that a large part of (expressive) speech or music signals affect the emotional state of the listener. This is also true for general audio [108]. However, emotions involve very complex phenomena and cannot be described without involving multiple dimensions including physiological and cognitive aspects. There is therefore a long history of research to better understand emotions and to build relevant theoretical models even if there is not always a consensus on the different models [109]. More recently, emotions have started to be used in machine-based systems leading to the so-called *Affective Computing* domain [110].

Note, that it is rather common to distinguish between expressed emotion and perceived (or evoked) emotions [111]. Only the latter is usually modeled in audio retrieval systems since it is (relatively) less influenced by situational factors (environment, mood, ...) of listening [112].

An audio or music emotion system can be built in the same way as any audio classification system (as described above) where the different categories to be recognized will be the different emotions (happy, sad, disgust, etc). This approach however supposes that emotions can be charaterised by broad and distinctive categories (which remains a strong assumption). An alternative to this approach is to view emotions in a 2D, or higher dimensional, continuous space (termed *emotion space*) using *valence* (referring to the pleasantness, positive or negative affective state) and *arousal* (referring to the activation or stimulation level) [113].

Then, the problem in affect-based retrieval is to predict the affect score (using the continuous valence and arousal scales) or descriptive labels such as 'happy' or 'annoying' of audio clips if a category based approach is followed. Such descriptions could be used for indexing movie clips [114] or automatically categorizing large collections of multimedia on the Internet.

Relative to the various problems in content-based audio retrieval, work on affect-based retrieval has commenced only recently. The main efforts have focused on discovering signal features that are correlated with the affective scores. In the work by Schuller et al. [115], a collection of over three hundred audio clips have been used to evaluate a large set of perceptual and statistical spectral measures. In Malandrakis et al. [116], a larger collection of over 1400 audio clips were subjectively annotated and feature selection was performed using regression-based models. In other work applied on music signals [112], the goal was to predict the perceived emotion of a sound clip as a probability distribution in the emotion plane using a total of 46 features including lyrics of the songs under the form of latent topics.

In spite of its perceptual implications and its potential in indexing multimedia content, work on affect-based retrieval has been limited. First, absence of comprehensive datasets that can support this research work has been a major hurdle. Additionally, it is difficult to establish a single ground truth to affective ratings amongst users. In fact, recent experiments have shown that relatively simple classification schemes and signal features can already perform better than user expectations [116]. Furthermore, in contrast to classical measures such as precision and recall to evaluate classifiers, performance evaluation of affect-based retrieval is difficult [117] as affective ratings are continuous scores and highly subjective.

V. Conclusion

Audio indexing and classification is a vibrant field of research, fueled by the development of Internet and social media. Historically in the shadow of the speech processing domain, the field is growing and gaining independent momentum in particular in the area of music signal processing [25].

Nevertheless, if it is not surprising that audio indexing has been strongly influenced by speech indexing and recognition advances, and an analysis of the current situation reveals that more intricate synergies do exist between the two domains [118]. Developing further cross-fertilization and exchange of ideas definitely opens the path for clear progress in both fields.

Rather surprisingly, the integration of auditory perception in most systems of both domains is limited. As highlighted in this paper, a number of perceptually-relevant concepts have been exploited in audio indexing research. But, it still seems that it is difficult to build a fully perceptually-relevant system that outperforms efficient machine-learning based methods that use only some principles of perception.

This remains surprising because from a pure acoustical point of view it intuitively appears that it may be unnecessary to capture similarity or dissimilarity information that is not perceived by humans. However, it is clear that all efficient audio classification systems do exploit some principles of perception either at the feature level (the vast majority of efficient systems do use at least some of the perceptually-motivated features given in table I) or at the classification or decision levels. This clearly shows the importance of perception in the design of such systems and motivates further investigation.

Our belief is that our understanding of audio sounds and music perception (especially of complex audio and music mixtures) is still limited and greater effort should be devoted to research to allow better exploitation of auditory models in audio indexing and retrieval approaches.

A recent trend in complex audio sounds and music signals analysis is to rely on source separation to pre-process and simplify the subsequent analysis of the constituent acoustic sources or events. Similarly, progress in perception should also allow to bridge the gap between the current capabilities of machine-based source separation and human performance to identify and separate the different audio sources involved in an audio scene.

Another interesting topic is the study of the temporal or dynamical evolution of an audio scene, an aspect often neglected or only roughly addressed in audio indexing and retrieval systems. Indeed, the audio features discussed in this paper are most often extracted on a frame-by-frame basis (the so-called bag-of-frames approach) leaving

little room to capture contextual information. However, as shown for example by many studies in musical timbre perception, context is important [119], [120]. One possibility is to observe the signal of interest at different time scales in order to capture these higher level semantics [121], and especially in the case of polyphonic music and complex audio sounds categories [122]. Temporal characteristics are indeed essential for human perception and source identification. Temporal integration [29], [31], the use of dynamical length sonic units [29], [123] or the design of novel multi-resolution perceptual filterbanks or redundant signal representations [124] are amongst the most promising directions to appropriately capture such temporal characteristics. This in conjunction with co-occurrence models can further help capture context in audio modeling [125].

As we have seen in the paper, an alternative to the direct integration of perception and hearing findings in a fully automatic audio processing system is to involve the user in the retrieval or indexing task in an iterative process. This permits design of user-aware retrieval systems where user context can be used or where context-based feature extraction (and similarity estimation) can be designed [126]. This also represents one of the most promising directions for building perceptually relevant semantic audio retrieval systems.

When applicable, the user can also be involved in the process of source separation in a so-called *informed source separation* paradigm where the user is able to enter a variety of information that would help the source separation engine to obtain better performance and to render more perceptually meaningful results [127], [81], [128]. Designing novel and efficient interaction scenarios is another interesting research direction.

In some cases, the perceptual space is squeezed into general categories (musical genres, emotions categories,...). In such cases, one faces the problem of pertinence of the ground-truth annotation which directly affects the difficulty of performance evaluation and thus of building a system that is judged perceptually relevant by users. Building audio categories classification systems based on continuous perceptual spaces with overlapping "soft categories" represents a current promising trend in affective computing research.

In parallel, user-aware audio and music indexing may serve other content creation applications. For example, a promising application in that direction will allow the user to generate new content from exploiting musical elements and ideas of existing music or video. This perceptually relevant sound creation process called N-th order derivative creation can be further developed by combining it with efficient perceptually relevant audio indexing systems [129], [130].

Another direction of research is to use multiple sources of information, for example coming from multiple modalities. This would allow not only to have access to a diversity of information sources but also to combine important perceptual cues from each modality [131]. In particular, building affect-based recognition engines that would jointly exploit information coming from classical sensors (microphones, cameras, ...) and physiological information such as from Electroencephalography (EEG), Electrocardiogram (ECG), blood pressure, skin conductance etc. is becoming one of the current targeted goal of the community. Building perceptually-relevant audio (and by extension multimodal) retrieval systems integrating high-level affect-based information is indeed one of the major challenges in this general field of research.

TABLE I

SOME PERCEPTUALLY-MOTIVATED FEATURES

Features	**Description**	**Citations**
Loudness	Is the subjective impression of the intensity of a sound.	[30]
Spectral Centroid	Spectral centroid is the weighted mean of the magnitude frequency spectrum; it is commonly described as one of the main dimension of timbre perception.	[56], [57]
Sharpness	Can be interpreted as a perceptual spectral centroid.	[30]
Perceptual spread	Is a measure of the timbral width of a given sound.	[30]
Signal to Mask ratio	Is the difference between the signal intensity and the intensity of the signal perceptual mask.	[1], [66]
Local energy in Bark scale	Represents the relative importance of local energy distribution in Bark bands.	[31]
Spectral Flux	Is the spectral magnitude Euclidean distance between neighboring audio frames.	[36], [37]
Sub-band flux	Represents the fluctuation of frequency content in ten octave-scaled bands.	[59]
High energy / low Energy	Represents the ratio of energy above and below a given frequency.	[59]
Roughness	Is a basic psychoacoustical sensation for rapid amplitude variations.	[132]
Relative entropy	Provides an estimate of the whiteness of a signal.	[59]
MFCC	Mel-Frequency Cepstral coefficients; Estimate the spectral envelope using (limited) perception principles.	[5], [10]
Cortical Representations	Multiscale or Multi linear representations that model various spectro-temporal properties in the central auditory system.	[19], [33]

APPENDIX

Two tables are given in this appendix. Table I provides some of the most common features with some indication of their links with perception. Table II provides some of the audio processing tools that are freely available to researchers and that have been used in many studies in audio and music signals processing.

REFERENCES

[1] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451 – 515, apr 2000.

[2] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Elsevier Academic Press, 2007, vol. 3.

[3] X. D. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, May 2001, vol. 1.

[4] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, 1937.

[5] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *In Proceedings of the SPECOM-2005*, 2005, pp. 191–194.

[6] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and A. M., *Complex Sounds and Auditory Images*. Pergamon, Oxford, 1992, pp. 3–48.

[7] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," *Apple Computer Technical Report # 35*, no. 35, 1993.

[8] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "Practical Gammatone-Like Filters for Auditory Processing," *EURASIP Journal on Audio Speech and Music Processing*, no. 4, pp. 1–16, 2007.

TABLE II

SOFTWARE TOOLS IN CONTENT-BASED AUDIO PROCESSING

Name	**Description**	**URL**
Sonic Visualizer	Application for GUI-based analysis and processing with built in plugin features that provides a variety of music and audio processing capabilities	http://www.sonicvisualizer.org
OpenAUDIO	A collection of tools capable of extracting and processing large number of low-level signal features and their functionals. See openSMILE, openEAR tools.	http://openaudio.eu/
CQT Toolbox	Matlab toolbox for computing the constant-Q transform (CQT) of a time-domain signal [133]	http://www.elec.qmul.ac.uk/people/anssik/cqt/
MIRtoolbox	Matlab toolbox for low-level and high-level feature extraction for music signals [134]	http://r.jyu.fi/1Ku2
MARSYAS	Software framework for rapid prototyping and experimentation with audio analysis [135]	http://marsyas.info/
jAudio	Java-based audio feature extractor library [136]	http://sourceforge.net/projects/jaudio/
FEAPI	A low-level Features Extraction Plugin API	[137] http://feapi.sourceforge.net/
YAAFE	*Yet Another Audio Features Extractor*: Efficient computation of many audio features simultaneously [138]	http://www.tsi.telecom-paristech.fr/aao/?p=155

[9] R. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea," in *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 7. IEEE, 1982, pp. 1282–1285.

[10] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences ," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[12] Thompson, J., K. and Atlas, L., E., "A Non-Uniform Modulation Transform for Audio Coding with Increased Time Resolution," *IEEE International Conference on Acoustics Speech and Signal (ICASSP)*, vol. 5, pp. 397–400, 2003.

[13] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-Scale Analysis for Content Identification," *IEEE Trans. on Signal Processing*, vol. 52, no. 10, pp. 3023 – 3035, Oct. 2004.

[14] Atlas, L. and Shamma, S. A., "Joint Acoustic and Modulation Frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[15] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech And Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[16] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust Speech Recognition using the Modulation Spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, 1998.

[17] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "On Modulation Spectral Analysis of Spectral and Cepstral Features," *IEEE Trans. on Multimedia*, vol. 11, no. 4, pp. 670–682, 2009.

[18] Y. X., K. Wang, and A. S. Shamma, "Auditory Representations of Acoustic Signals," *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 824–839, March 1992.

[19] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 920 – 930, may 2006.

[20] S. Sundaram and S. Narayanan, "Discriminating Two Types of Noise Sources using Cortical Representation and Dimension Reduction Technique," *IEEE, International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 1, pp. I–213 – I–216, 2007.

[21] O. Kalinli and S. Narayanan, "Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 5, pp. 1009 –1024, July 2009.

[22] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300 –312, feb. 2007.

[23] D. Levitin, *This is your brain on music*. The science of a human obsession. London, UK: Atlantic Books, 2008.

[24] D. Hargreaves and A. North, "The functions of music in everyday life: Redefining the social in music psychology." *Psychology of Music*, vol. 27, pp. 71–83, 1999.

[25] M. Mueller, D. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, oct. 2011.

[26] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *International Conference on Machine Learning*. Morgan Kaufmann, 1994, pp. 121–129.

[27] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[28] M. Ramona, G. Richard, and B. David, "Multiclass Feature Selection With Kernel Gram-Matrix-Based Criteria," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 1611 – 1623, 2012.

[29] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 174–186, 2009.

[30] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," IRCAM, Paris, France, Tech. Rep., Apr. 2004.

[31] F. Fuhrmann, "Automatic musical instrument recognition from polyphonic music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, April 2012.

[32] S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1401 –1412, july 2006.

[33] O. Kalinli, S. Sundaram, and S. Narayanan, "Saliency-driven Unstructured Acoustic Scene Classification using Latent Perceptual Indexing," in *IEEE Multimedia Signal Processing (MMSP) Workshop*, October 2009, pp. 1 – 6.

[34] J. Hawkins and S. Blakeslee, *On Intelligence*. Times Books, 2004.

[35] Y. Wang, Y. Wang, S. Patel, and D. Patel, "A layered reference model of the brain (LRMB)," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews.*, vol. 36, no. 2, pp. 124 –133, March 2006.

[36] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of General Audio Data for Content-Based Retrieval," *Pattern Recognition Letters*, vol. 22, pp. 533–544, 2001.

[37] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," *IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP) 1997. Munich, Germany.*, vol. 2, pp. 1331–1334, April 1997.

[38] G. Fant, *Acoustic theory of Speech Production*. Mouton, La Hague, 1960.

[39] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings," *IEEE Trans. on Multimedia*, vol. 7, no. 1, February 2005.

[40] Z. J., N. Pavesic, and F. Mihelic, "Speech/Non-Speech Segmentation Based on Phoneme Recognition Features," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–13, February 2006.

[41] M. Ramona and G. Richard, "Comparison of different strategies for a svm-based audio segmentation," in *European Signal Processing Conference (EUSIPCO)*, Glasgow, UK, Sep. 2009.

[42] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust Voice Activity Detection using Long-term Signal Variability," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[43] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-Based Classification, Search, and Retrieval of Audio," *IEEE Trans. on Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[44] G. Guo and S. Z. Li, "Content-Based Audio Classification and Retrieval by Support Vector Machines," *IEEE Trans. on Neural Networks*, vol. 14, no. 1, pp. 209–215, January 2003.

[45] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *, IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, jan. 2006.

[46] S. Chu, S. Narayanan, C. C. Kuo, and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," *International Conference on Multimedia and Expo (ICME)*, July 2006.

[47] S. Chu, S. Narayanan, and C. C. Kuo, "Environmental Sound Recognition With Time-Frequency Audio Features," *IEEE Trans. on Audio Speech and Language Processing*, vol. 17, pp. 1142–1158, August 2009.

[48] B. Ghoraani and S. Krishnan, "Time-Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2197 –2209, September 2011.

[49] J. Licklider, *Basic correlates of the auditory stimulus*. Wiley, New York, 1951, pp. 985–1035.

[50] S. Stevens, "On the psychophysical law," *Psychological Review*, vol. 64, no. 3, pp. 153–181, 1957.

[51] J. F. Schouten, "The perception of timbre," in *6th International Congress on Acoustics*, Tokyo, Japan, 1968, pp. GP–6–2.

[52] R. Shepard, "Circularity in judgments of relative pitch," *Journal of the Acoustical Society of America (JASA)*, vol. 36, pp. 2346–2353, 1964.

[53] B. C. J. Moore, *An introduction to the psychology of hearing, 5th ed.* Academic Press, 2003.

[54] A. Bregman, *Auditory scene analysis: the perceptual organization of sound*. Cambridge, Mass.: The MIT Press, 1990.

[55] Americal National Standards Institute, "USA standard acoustical terminology," no. S1.1–1960, 1960.

[56] S. McAdams, S. Winsberg, G. de Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes," *Psychological Research*, vol. 58, pp. 177–192, 1995.

[57] P. Iverson and C. Krumhansl, "Isolating the dynamic attributes of musical timbre," *Journal of the Acoustic Society of America*, vol. 94, pp. 2595–2603, 1993.

[58] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 471–482, 2005.

[59] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Perception*, vol. 27, no. 3, pp. 223–241, 2010.

[60] J. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America (JASA)*, vol. 61, pp. 1270–1277, 1977.

[61] J. Barbedo and A. Lopes, "Automatic genre classification of musical signals," *EURASIP Journal on Advances in Signal Processing*, pp. 157–168, 2007.

[62] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293 – 302, jul 2002.

[63] D. Liu, L. Lu, and H. J. Zhang, "Automatic mood detection from acoustic music data." in *4th International Conference on Music Information Retrieval*, 2003, pp. 81–87.

[64] P. Herrera-Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006, pp. 163–200.

[65] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs," in *Seventh International Symposium on Signal Processing and its Applications*, Paris, France, 2003, pp. 133–136.

[66] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 68–80, 2006.

[67] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, p. 546047, 2010.

[68] K. D. Martin, "Sound-source recognition : A theory and computational model," Ph.D. dissertation, Massachusets Institute of Technology, June 1999.

[69] G. Richard, *Audio Indexing*. Information Science Reference - IGI Global, 2008.

[70] R. Gjerdingen and D. Perrot, "Scanning the dial: The rapid recognition of music genres," *Journal of New Music Research*, vol. 37, pp. 93–100, 2008.

[71] J.-J. Aucouturier, "Dix expériences sur la modélisation du timbre polyphonique [Ten experiments on the modelling of polyphonic timbre]," Ph.D. dissertation, University Paris 6, France, 2006.

[72] C. Reuter, "The role of formant positions and micro-modulations in blending and partial masking of musical instruments," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2237–, 2009.

[73] T. Zhang and C.-C. J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, May 2001.

[74] L. Liu, H. J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, October 2002.

[75] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 12 –23, jan 1998.

[76] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, aug 1986.

[77] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

[78] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 116–128, 2008.

[79] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, 2003, pp. 553–556.

[80] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 529–540, Mar. 2008.

[81] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, Sep. 2011.

[82] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen., "Sound event detection in multisource environments using source separation," in *International Workshop on Machine Listening in Multisource Environments*, Florence, Italy, 2011.

[83] C. Rijsbergen, *Information retrieval*. London: Butterwoth-Heinmann, 2nd edition, 1979.

[84] G. Salton, *Automatic Information Organization and Retrieval*. McGraw-Hill, 1968.

[85] K. Hoashi, K. Matsumoto, and N. Inoue, "Personalization of user profiles for content-based music retrieval based on relevance feedback," in *ACM international conference on Multimedia*. New York, NY, USA: ACM Press, 2003, pp. 110–119.

[86] M. Mandel, G. Poliner, and D. Ellis, "Support vector machine active learning for music retrieval," *Multimedia Systems*, vol. 12, no. 1, pp. 3–13, 2006.

[87] G. Chen, T.-J. Wang, and P. Herrera, "A novel music retrieval system with relevance feedback," in *International Conference on Innovative Computing Information and Control*, june 2008, p. 158.

[88] M.-K. Shan, M.-F. Chiang, and F.-F. Kuo, "Relevance feedback for category search in music retrieval based on semantic concept learning," *Multimedia Tools Appl.*, vol. 39, no. 2, pp. 243–262, 2008.

[89] M. Slaney, "Semantic Audio Retrieval," *International Conference on Acoustic Speech and Signal Processing (ICASSP), Orlando, USA.*, pp. 13–17, May 2002.

[90] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, " Audio Information Retrieval using Semantic Similarity," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Honolulu, Hawaii, USA.*, vol. 2, pp. II–725–II–728, 2007.

[91] S. Deerwester, S. T. Dumais, G. W. Furnas, T. Landaurer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 6, no. 41, pp. 391–407, 1990.

[92] J.R.Bellegarda, "Latent Semantic Mapping," *IEEE Signal Processing Magazine*, vol. 22, pp. 70–80, September 2005.

[93] S. Sundaram and S. Narayanan, "Audio Retrieval by Latent Perceptual Indexing," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Las Vegas, USA.*, 2008.

[94] ——, "Classification of Sound Clips by Two Schemes: Using Onomatopoeia and Semantic labels," *IEEE International Conference on Multimetia and Expo (ICME), Hannover, Germany.*, pp. 1341–1344, June 2008.

[95] L. Lu and A. Hanjalic, "Unsupervised Anchor Space Generation for Similarity Measurement of General Audio," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Las Vegas,USA*, pp. 53 – 56, 2008.

[96] K. Lee and D. Ellis, "Audio-Based Semantic Concept Classification for Consumer Video," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1406 –1416, aug. 2010.

[97] R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik, "Sound retrieval and ranking using sparse auditory representations." *Neural Computation*, vol. 22, no. 9, pp. 2390–2416, 2010.

[98] M. Slaney, D. Ponceleon, and J. Kaufman, "Multimedia Edges: Finding Hierarchy in all Dimensions," *ACM international conference on Multimedia*, pp. 29–40, 2001.

[99] R. Cai, L. Lu, and A. Hanjalic, "Unsupervised Content Discovery in Composite Audio," in *13th annual ACM international conference on Multimedia MULTIMEDIA 05*. ACM Press, december 2005, p. 628.

[100] G. Chechik, E. le, M. Rehn, S. Bengio, and D. Lyon, "Larget-Scale Content-Based Audio Retrieval from Text Queries," *MIR' 08, Vancouver, Canada.*, vol. 2, pp. 105–112, October 2008.

[101] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 8, pp. 1371 –1384, Aug. 2008.

[102] T. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. Gemmeke, J. Bellegarda, and S. Sundaram, "Exemplar-Based Processing for Speech Recognition: An Overview," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 98–113, Nov. 2012.

[103] H.-H. Shih, S. Narayanan, and C.-C. Kuo, "A statistical multidimensional humming transcription using phone level hidden markov models for query by humming systems," in *International Conference on Multimedia and Expo (ICME)*, vol. 1, july 2003, pp. 61–65.

[104] C.-C. Wang, J.-S. R. Jang, and W. Wang, "An improved query by singing/humming system using melody and lyrics information," in *ISMIR*, 2010, pp. 45–50.

[105] O. Gillet and G. Richard, "Drum loops retrieval from spoken queries," *Journal of Intelligent Information Systems - Special issue on Intelligent Multimedia Applications*, vol. 24, no. 2/3, pp. 159–177, Mar. 2005.

[106] P. Hanna and M. Robine, "Query by tapping system based on alignment algorithm," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, april 2009, pp. 1881 –1884.

[107] E. Coviello, Y. Vaizman, A. Chan, and G. Lankreit, "Multivariate Autoregressive Mixture Models for Music Auto-Tagging," *International Symposium on Music Information Retrieval (ISMIR) 2012, Porto, Portugal.*, pp. 547–552, October 2012.

[108] M. M. Bradley and P. J. Lang, "Affective Reactions to Acoustic Stimuli," *Psychophysiology*, vol. 37, no. 2, pp. 204–215, 2000.

[109] C. Pelachaud(Editor), *Emotional Interaction System*. John Wisley, 2011.

[110] R. Picard, *Affective Computing*. The MIT Press, 1997.

[111] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, pp. 123–147, 2002.

[112] Y.-H. Yang and H. Chen, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2184–2196, sept. 2011.

[113] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, "Primitives based estimation and evaluation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, November 2007.

[114] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A Supervised Approach to Movie Emotion Tracking," *IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2376 –2379, May 2011.

[115] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic Recognition of Emotion Evoked by General Sound Events," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012.

[116] N. Malandrakis, S. Sundaram, and A. Potamianos, "Affective Classification of Generic Audio using Regression Models," *Submitted ICASSP 2013*, 2012.

[117] *Towards Evaluation of Example-based Audio Retrieval System using Affective Dimensions*. IEEE, June 2010.

[118] F. Weninger, B. Schuller, C. C. Liem, F. Kurth, and A. Hanjalic, "Music Information Retrieval: An Inspirational Guide to Transfer from Related Disciplines," in *Multimodal Music Processing*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, vol. 3, pp. 195–216.

[119] R. Kendall, "The role of acoustic signal partitions in listener categorization of musical phrases," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 4, no. 2, pp. 185–213, 1986.

[120] G. Sandell, "Identifying musical instruments from multiple versus single notes," *Journal of the Acoustical Society of America*, vol. 100, no. 4, p. 2752, 1996.

[121] M. Casey and M. Slaney, "The Importance of Sequences in Musical Similarity," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, p. V, may 2006.

[122] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[123] M. Lagrange, M. Raspaud, R. Badeau, and G. Richard, "Explicit modeling of temporal dynamics within musical signals for acoustical unit similarity," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1498–1506, 2010.

[124] E. Ravelli, G. Richard, and L. Daudet, "Audio signal representations for indexing in the transform domain," *IEEE Trans. on Audio, Speech and Language Processing*, Mar. 2010.

[125] S. Kim, P. Georgiou, and S. Narayanan, "Latent Acoustic Topic Models for Unstructured Audio Classification," *APSIPA Transactions on Signal and Information Processing*, vol. 1, 11 2012.

[126] M. Schedl, S. Stober, E. Gómez, N. Orio, and C. C. Liem, "User-Aware Music Retrieval," in *Multimodal Music Processing*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, vol. 3, pp. 135–156.

[127] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, 2009, pp. 69–72.

[128] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *, IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721 – 1733, 2011.

[129] M. Goto, "Grand Challenges in Music Information Research," in *Multimodal Music Processing*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, vol. 3, pp. 217–226.

[130] T. Nakano, S. Murofushi, M. Goto, and S. Morishima, "Dancereproducer: An automatic mashup music video generation system by reusing dance video clips on the web." in *Sound and Music Computing Conference (SMC 2011)*, Glasgow, UK, 2011, pp. 183–189.

[131] S. Essid and G. Richard, "Fusion of Multimodal Information in Music Content Analysis," in *Multimodal Music Processing*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, vol. 3, pp. 37–52.

[132] P. Daniel and R. Weber, "Psychoacoustical roughness: Implementation of an optimized model," *Acustica*, vol. 83, pp. 113–123, 1997.

[133] C. Schoerkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *Sound and Music Computing Conference*, 2010.

[134] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *International Conference on Digital Audio Effects*, 2007, pp. 1–8. [Online]. Available: http://dafx.labri.fr/main/papers/p237.pdf

[135] G. Tzanetakis and P. Cook, "Marsyas: a framework for audio analysis," *Organised Sound*, vol. 4, pp. 169–175, 2000.

[136] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle, "jaudio: A feature extraction library," in *Proc. of the International Conference on Music Information Retrieval*, 2005.

[137] A. Lerch, G. Eisenberg, and K. Tanghe, "Feapi, a low level features extraction plugin api," in *International Conference on Digital Audio Effects*, 2005.

[138] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *ISMIR*, 2010, pp. 441–446.