Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 654914, 12 pages
doi:10.1155/2010/654914

Research Article

Development of the Database for Environmental Sound Research and Application (DESRA): Design, Functionality, and Retrieval Considerations

Brian Gygi[1] and Valeriy Shafiro[2]

[1] *Speech and Hearing Research, Veterans Affairs Northern California Health Care System, Martinez, CA 94553, USA*
[2] *Communications Disorders and Sciences, Rush University Medical Center, 600 S. Paulina Street, Chicago, IL 60612, USA*

Correspondence should be addressed to Brian Gygi, bgygi@ebire.org

Received 16 January 2010; Accepted 17 May 2010

Academic Editor: Harvey Thornburg

Theoretical and applied environmental sounds research is gaining prominence but progress has been hampered by the lack of a comprehensive, high quality, accessible database of environmental sounds. An ongoing project to develop such a resource is described, which is based upon experimental evidence as to the way we listen to sounds in the world. The database will include a large number of sounds produced by different sound sources, with a thorough background for each sound file, including experimentally obtained perceptual data. In this way DESRA can contain a wide variety of acoustic, contextual, semantic, and behavioral information related to an individual sound. It will be accessible on the Internet and will be useful to researchers, engineers, sound designers, and musicians.

1. Introduction

Environmental sounds are gaining prominence in theoretical and applied research that crosses boundaries of different fields. As a class of sounds, environmental sounds are worth studying because of their ability to convey semantic information based on complex acoustic structure. Yet, unlike other large meaningful sound classes, that is, speech and music, the information conveyed by environmental sounds is not linguistic, as in speech, and typically is not designed for its aesthetic value alone, as in music (e.g., [1, 2]). There are numerous practical and specific applications for environmental sounds, which include auditory rehabilitation for hearing aid users and cochlear implants recipients; nonlinguistic diagnostic tools for assessing auditory and cognitive deficits in prelinguistic children; noise control and design of acoustic environments; auditory icons in computer displays; sonification, the process of representing information with nonspeech sounds (see [3] for a recent review). However, the knowledge base still lags far behind that of the other major classes of naturally occurring everyday sounds, speech, and music. One of the major hurdles is the lack of a standardized database of readily available, free, tested, high quality, identifiable environmental sounds for users to work with. There are various resources for accessing environmental sounds, some of which were detailed in [4]; however, that paper noted several caveats for users who are looking for sounds on their own. Among them were redundancy in time and effort needed to find the necessary information (which may have been found by others before), and the possibility of idiosyncrasies or occasional unwanted "surprises" (e.g., clipping or artifacts) in otherwise suitable stimuli. To correct those may require the user to have sufficient expertise in several technical areas such as digital signal processing and recording techniques, which may not, in and of themselves, have any relevance to the goals of the intended project.

To remedy this, this paper relates some findings of an ongoing project on the part of the authors assisted by James Beller, programmer, the Database for Environmental Sound Research and Application (DESRA—website: http://www.desra.org/) which aims to collect, label, edit when necessary, norm, evaluate, and make available a large collection of environmental sounds comprising multiple tokens (exemplars) of a wide variety of common sound sources. The collection of sounds and development of the Web front end are ongoing. This paper will describe the structure and function of the database, which reflects and responds to the complex ways in which we think about and use environmental sounds.

2. Defining Environmental Sounds

In order to optimally design a useful multipurpose database for environmental sounds, it is necessary to have a fuller understanding of the nature of environmental sounds, what they represent for humans, factors in environmental sound perception, and how their perception may be similar or different for different listeners. This information will guide sound selection by database users: researchers, designers, and musicians. The ability to use experimentally obtained perceptual criteria in sound selection, in addition to a thorough description of technical characteristics of the sounds, constitutes a unique feature of the present database. Although what Gaver termed "everyday listening" [5] is a frequent activity, the nature of the experience has been remarkably underscrutinized, both in common discourse and in the scientific literature, and alternative definitions exist [6, 7]. This is changing, as the numerous articles in this volume will attest, but still even our basic understanding of environmental sounds has large *lacunae*.

Thus, unlike speech and music, there is no generally agreed upon formal structure or taxonomy for environmental sounds. Instead, there are several prominent approaches to environmental sound classification that have been advanced over the last several decades [5–7]. A major initial contribution to environmental sound research is contained within the framework of Acoustic Ecology advanced by Schafer [6] who advanced the notion of the soundscape as the totality of all sounds in the listener's dynamic environment. Further extended by Truax [7] in his Acoustic Communication model, speech, music, and soundscape (that includes all other sounds in the environment) are treated as part of the same acoustic communication continuum wherein sounds' acoustic variety increases from speech to soundscape, while sounds' rule-governed perceptual structure, temporal density of information, and specificity of meaning all increase from soundscapes to speech. Importantly, the Acoustic Communication approach also treats listening as an active process of interacting with one's environment and distinguishes among several different levels of listening such as listening-in-search (when specific acoustic cues are being actively sought in the sensory input), listening-in-readiness (when the listener is ready to respond to specific acoustic cues if they appear but is not actively focusing his/her attention on finding them), and background listening (when listeners are not expecting significant information or otherwise actively processing background sounds). The theoretical constructs of the Acoustic Communication model are intuitive and appealing and have been practically useful in the design of functional and aesthetically stimulating acoustic environments [8]. However, directed mostly toward more general aspects of acoustic dynamics of listener/environment interactions, as regards cultural, historical, industrial, and political factors and changes at the societal level, it is still the case that more specific perceptual models are needed to investigate the perception of environmental sounds in one's environment.

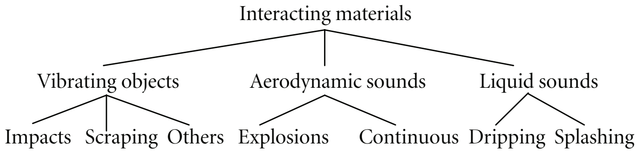


Figure 1: Hierarchy of sound producing events adapted from Gaver [5].

In his seminal piece, What Do We Hear in the World [5], Gaver attempted to construct a descriptive framework based on what we listen for in everyday sounds. He examined previous efforts, such as libraries of sound effects on CD, which were largely grouped by the context in which the sound would appear, for example, "Household sounds" or "Industry sounds." While this would be useful for people who are making movies or other entertainment, he found it not very useful for a general framework. "For instance, the categories are not mutually exclusive; it is easy to imagine hearing the same event (e.g., a telephone ringing) in an office and a kitchen. Nor do the category names constrain the kinds of sounds very much."

Instead, he looked at experimental results by himself and others [9–12] which suggested in everyday listening that we tend to focus on the sources of sounds, rather than acoustic properties or context. He reasoned that in a hierarchical framework, "Superordinate categories based on types of *events* (as opposed to contexts) provide useful clues about the sorts of sounds that might be subordinate, while features and dimensions are a useful way of describing the differences among members of a particular category." Inspired by the ecological approach of Gibson [13], he drew a sharp distinction between "musical listening", which is focusing on the attributes of the "sound itself", and "everyday listening" in which "... the perceptual dimensions and attributes of concern correspond to those of the sound-producing event and its environment, not to those of the sound itself."

Based on the physics of sound-producing events, and listeners' description of sounds, Gaver proposed a hierarchical description of basic "sonic events," such as impacts, aerodynamic events and liquid sounds, which is partially diagrammed in Figure 1. From these basic level events, more complex sound sources are formed, such as *patterned sources* (repetition of a basic event), *complex sources* (more than one sort of basic level event), and *hybrid sources* (involving more than one basic sort of material).

Gaver's taxonomy is well thought out, plausible, and fairly comprehensive, in that it includes a wide range of naturally occurring sounds. Naturally there are some that are excluded—the author himself mentions electrical sounds, fire and speech. In addition, since the verbal descriptions were culled from a limited sample of listener responses, one must be tentative in generalizing them to a wider range of sounds. Nevertheless, as a first pass it is a notable effort at providing an overall structure to the myriad of different environmental sounds.

Gaver provided very limited experimental evidence for this hierarchy. However, a number of experiments both previous and subsequent have supported or been consistent with this structuring [12, 14–18] although some modifications have been proposed, such as including vocalizations as a basic category (which Gaver himself considered). It was suggested in [16] that although determining the source of a sound is important, the goal of the auditory system is to enable an appropriate response to the source, which would also necessarily include extracting details of the source such as the size and proximity and contextual factors that would mitigate such a response. Free categorization results of environmental sounds from [16] showed that the most common basis for grouping sounds was on source properties, followed by common context, followed by simple acoustic features, such as Pitched or Rumbling and emotional responses (e.g., Startling/Annoying and Alerting). Evidence was provided in [19] that auditory cognition is better described by the actions involved from a sound emitting source, such as “dripping” or “bouncing”, than by properties of their causal objects, such as “wood” or “hollow”. A large, freely accessible database of newly recorded environmental sounds has been designed around these principles, containing numerous variations on basic auditory events (such as impacts or rolling), which is available at http://www.auditorylab.org/.

As a result, the atomic, basic level entry for the present database will be the source of the sound. In keeping with the definition provided earlier, the source will be considered to be the objects involved in a sound-producing event with enough description of the event to disambiguate the sound. For instance, if the source is described as a cat, it is necessary to include “mewing”, “purring”, or “hissing” to provide a more exact description. There are several potential ways to describe the source, from physical objects to perceptual and semantic categories. Although the present definition does not allow for complete specificity, it does strike a balance between that and brevity and allows for sufficient generalization that imprecise searches can still recover the essential entries.

Of course sounds are almost never presented in isolation but in an auditory scene in which temporally linear mixtures of sounds enter the ear canal and are parsed by the listener. Many researchers have studied the regularities of sound sources that can be exploited by listeners to separate out sounds, such as common onsets, coherent frequency transitions, and several other aspects (see, e.g., [20]). The inverse process, integration of several disparate sources into a coherent “scene”, has been much less studied, as has the effect of auditory scenes on perception of individual sounds [21–23]. As a result, the database will also contain auditory scenes, which consist of numerous sound sources bound together by a common temporal and spatial context (i.e., recorded simultaneously). Some examples are a street scene in a large city, a market, a restaurant or a forest. For scenes, the context is the atomic unit for the description.

Above these basic levels, multiple hierarchies can be constructed, based on the needs and desires of the users, which are detailed in the next section.

3. Projected Users of the Database

The structure and functionality of the database are driven by the users and their needs. The expected users of this database are described below.

3.1. Researchers. This is the largest expected group of users. Based on environmental sound research conducted in the past several decades, there are several common criteria in selecting sounds suitable for research. One of their main concerns has been how identifiable or familiar a sound is, since as noted above, identification of the source of a sound is the primary goal of the auditory system with respect to environmental sounds. Other researchers might also want to know acoustic attributes of sounds, such as the amount of high frequency information, the duration, and the temporal structure if they are undertaking studies in filtering environmental sounds (e.g., [2]) or looking at the role of temporal cues [1]. Many researchers have investigated semantic attributes of sounds, such as “harshness” or “complexity” (see Section 4 below for citations), or broader sound categories which can also be included, either from pre-existing data, if an associate (described below) has such data on that particular sound, or by aggregating ratings submitted on the website (see Section 8.4 below). Other researchers might be interested in emotional aspects of sounds, such as “pleasant”, or “chilling” [24]. Some more psychoacoustically oriented researchers would like several tokens of the same sound source that vary in only specific aspects, such as a ball bouncing on wood, on metal, or on plastic, or a hammer striking a large plate versus a small plate [25–27]. Finally, a citation history, listing studies which have employed this particular sound, would be very useful for cross-study comparisons.

3.2. Sound Designers. Aside from the source of the sound, sound designers may also want some metadata such as details of the recording: the location, the time, the distance of the microphone from the source, and the recording level. If they are planning to use it in a film or video, it would be useful for them to know what settings a sound would be appropriate for, for example, if a dog barking would seem out of place in an office. Such searches will be helped by recording background data as well as perceptual ratings data on sound congruency for different auditory scenes [28].

3.3. Musicians and Game Designers. There is also a large number of people who are looking for sounds to use as musical samples for songs or games. There are sites already geared towards these users, such as freesound.org and soundsnap.com. These sites and some of their limitations are described below. In addition to the above information, they might also like to know how musical a sound is (which is related to harmonic structure) or how rhythmic a sound is, which can be based on acoustic analyses.

4. Sources of Information

(a) As mentioned above, a central concern for many researchers will be how identifiable a sound is, while others may be interested in typical or atypical sound tokens. Thus, the database should designate which sounds are “easily identifiable”, “very familiar”, or “highly typical.” These classifications will be based on empirical data where it exists [1, 2, 18]. Researchers who have gathered such data on these will be encouraged to submit it. In addition, the site will have results of online identification experiments which the users will be encouraged to participate in (see below), and those results will be made available to users. Users will also want to know whether the clip is of a sound in isolation or of a scene. A coding scheme will be used where 1 = single source in isolation, 2 = scene with many sources. This judgment will be made at the time of submission and cross-checked for accuracy by the maintainers of the site.

(b) Waveform statistics: file format, file size, sampling rate, quantization depth, duration, number of channels, dc offset, number of clipped samples, rms (in dB), and peak (in dB). Most users would want to know such details of the recording as the location, the time, the distance of the microphone from the source, and the recording level. This information will need to be entered by the associate submitting the sound, and everyone submitting a sound will be encouraged to supply these data.

(c) Contextual information, such as whether the sound is occurring outdoors or indoors, in a large space or a small space, or in an urban or rural setting. Again, a good deal of this information is recoverable from the acoustics (from reverberation or from higher order acoustic features [29]), but if the precise data are known, they should be included in the database.

(d) Qualitative aspects: in addition to properties of the source, sounds elicit semantic associations for listeners. Some sounds can be chilling, some sounds are considered pleasant, and some sounds are judged to be tense. Several studies have investigated these qualities using the semantic differential method [30–33] introduced by Osgood [34] (described below) and then tried to correlate those qualities with various acoustic features of the sounds. Some consistent results have emerged. For instance, perceived size is reliably associated with loudness [24], low frequencies with heaviness [24], tenseness correlates with an energy peak around 2500 Hz [35], and pleasant sounds tend to lack harmonics [30]. In perhaps the most comprehensive study [31], ratings of 145 environmental sounds were obtained representing various categories of naturally occurring environmental sounds (e.g., impact sounds, water sounds, ambient sounds) on 20 7-point bipolar scales. A principal components analysis of the rating data showed that the judgments of the listeners could be associated with four dimensions, accounting for 89% of the variance. The four dimensions roughly corresponded (in descending order of r^2) to “harshness”, “complexity”, “size”, and “appeal”. Since it is anticipated that some users will be interested in the semantic attributes of sounds, the four attributes mentioned in [31] as well as “tenseness” will be included as part of a sound’s entry. If the values on those dimensions are known (i.e., have been established by previous research), they will be included. Otherwise users of the system will have an opportunity to rate these sounds, as described below.

There are some qualitative features of sounds that can be calculated directly from the waveforms, such as roughness (as defined in [36]), sharpness [37], and loudness (ANSI loudness) if the recording level SPL is known. The appropriate algorithms for calculating these values will be applied to sounds as they are entered into the database and the resulting values attached to the sounds as part of a sound’s entry.

(e) Musical features: some users of the database may be musicians looking for sounds to use in musical compositions and would be concerned with how the sounds will fit in both harmonically and rhythmically. Therefore acoustic variables will be included for both aspects of the sounds.

Harmonically Related Variables

(1) Spectral centroid (closely related to the pitch, and will be expressed both in Hz and note scale value).

(2) Spectral spread (the bandwidth in Hz).

(3) Pitch salience (level of harmonicity of a sound—from [38]).

(4) Estimated pitch. Environmental sounds are not homogeneous with regard to pitch. Some sounds, primarily vocalizations, have a harmonic structure and thus have a pitch that can be calculated using common pitch estimation methods (such as in [39]). However, some, such as impacts or scraping, have a spectrum that is more akin to a broadband noise, and thus most algorithms fail at extracting a reliable pitch. Since the pitch salience is a measure of the degree of harmonicity, for sounds with a pitch salience above 0.7 (on a scale of 0-1) the system will attempt to extract a pitch. For the remaining sounds, it will just report “N/A”.

Rhythmically and Temporally Related Variables

(1) Amplitude slope (reflects the initial attack and decay of a sound).

(2) Autocorrelation peaks (indicating the degree and period of the rhythmicity of a sound).

These values can be automatically calculated for a sound upon entry [2].

5. Existing Sounds Online

5.1. Search Engines. There are a few existing search engines for environmental sounds on the Internet, an overview of which was provided in [4]. Most of these are geared towards sound effects for use in movies and music. Some of them are attached to libraries (including the excellent LABROSA site, http://labrosa.ee.columbia.edu/dpwe-bin/sfxlist.cgi); others just provide links to other web sites that contain the sounds (http://findsounds.com/, http://sound-effects-library.com/, and http://sonomic.com/). All of these engines allow searches by keywords, and some also allow specification of file format (.wav,.mp3), sampling rate, bit size, stereo/mono, and file size. Some of them provide schematics of the waveform and previews before downloading. The engines that simply search the Internet and provide links to other sites usually just give access to low-quality mp3s (in part to discourage users from recording them through their soundcards). Additional problems are the keywords are usually not standardized (so a search of "kitten" and "cat" would yield different sounds), and the copyright status of these clips is often not clear. In contrast, the search engines that are attached to dedicated libraries are usually not free and can be quite expensive if ordering a number of sounds (and the sounds are usually copyrighted and thus not freely distributable).

In the intervening years since the Shafiro and Gygi overview [4] some new sites have sprung up which more closely match the model being proposed here. Two examples are freesound.org and soundsnap.com. These are both collections of sounds donated by members, who are largely sound enthusiasts, both amateur and professional, which means the sounds are usually recognizable, and they are guaranteed to be freely sharable. Freesound.org requires sound submitters to abide by the Creative Commons license, which is described in the copyright notice above. The search engines allow searches on keywords (called tags in http://www.freesound.org/), descriptions, duration, bit rate, bit depth, and sampling rate or by the name of member who submitted the sound. The results of the search can be sorted by various criteria, such as relevance, and most recent, or the most popular. Related sounds can be organized into packs and downloaded as a group. People who are browsing the sounds can add tags and comments. Finally, and most interestingly, for a given sound, users can request to see "similar sounds", in which similarity is defined using the Wordnet taxonomy [40]. This is an instantiation of Query By Example (QBE) which is described in Section 10.

There are several advantages to these sites. They are open, the sounds are freely distributable, and users can create their own keywords. However, the lack of standardization of keywords can lead to difficulty in searches, and some of the sounds may be of dubious quality since the uploaded sounds are not moderated. The search engine itself is a bit clumsy when trying to handle and organize large numbers of sounds, and the only metadata on the sounds concern the audio type (mp3, wav, bit size, and sampling rate). Soundsnap suffers from similar problems, plus they seem to be moving towards a pay-to-download model. The database under construction will attempt to alleviate these problems.

6. Structure of the Proposed Database

The structure for the basic entries is shown below. It is similar to a structure that was suggested in [4], with some additional information added. For a single source, an entry is illustrated using one token of a baby crying sound see Table 1.

For an auditory scene example, an entry for a train station sound is used (see Table 2).

7. Sounds Accepted for Uploading

Sounds will be uploaded using the screen shown in Figure 4. The sounds accepted for uploading will all be high quality—at least 16-bit 22 kHz sampling rate for wav files, at least 196 kbps per channel bit rate for mp3s, with little or no clipped samples. The sounds must be recordings of physical sources—no synthesized sounds will be accepted. The sounds can either represent single sources or scenes. This will be designated by the originator upon uploading and verified by the admin. If the sounds represent single sources, a determination will be made as to the isolation of the source, that is, whether only the source is present or whether there are background sounds present.

8. Front End to the Database

There are four main front end functions that are essential to the functioning of the database. They are user enrollment and access; uploading sounds to the database; the search engine; and, perceptual testing.

8.1. User Levels. There will be three different user levels.

Admin would have all rights to the database and be limited to people working on the project.

(1) Associates would be verified sound workers, whether researchers, designers, or recordists. They would be able to upload sounds without vetting, add research-related metadata (e.g., identifiability, acoustic analyses, and citations), and create new keywords.

(2) Participants can download sounds, submit sounds for uploading, attach existing keywords to sounds, and suggest additional keywords.

8.2. The Search Engine. The portal to accessing sounds is the Sound Omnigrid shown in Figure 2. When users first enter the database, they are presented with the Omnigrid plus options for searching on and managing the sounds.

If the user selects "Search" a search screen will come up. Users will be able to search upon any of the sound data and/or keywords. For example, a search on the keyword "rooster" returned the screen shown in Figure 3.

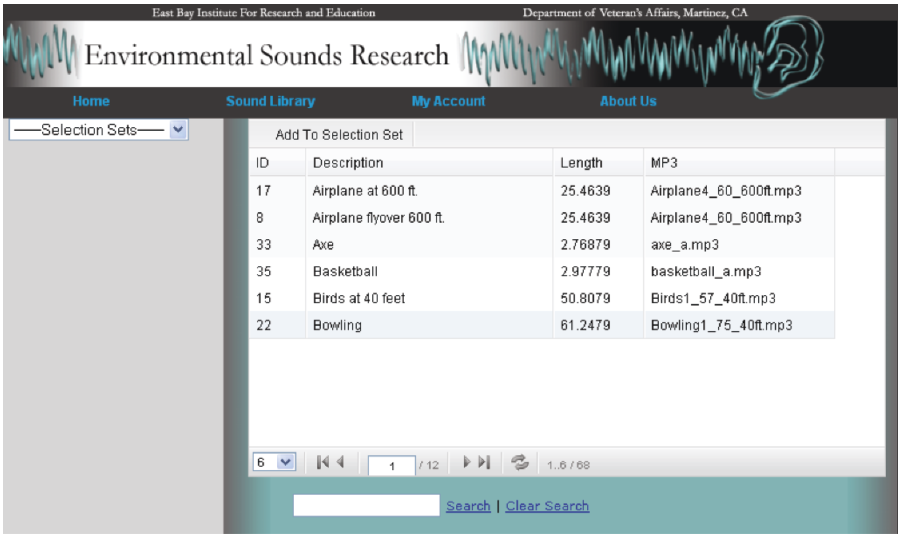
Furthermore users can specify multiple criteria on any of the fields (e.g., search for sound types "horse" and "train" in either mp3 or wav format), and the number of tokens returned for each criterion, allowing users to easily create sound lists for further use. Where multiple sound tokens fit

Table 1

Sound file name	Baby3.wav
Sound label(s)	Baby crying
Sound keywords	
(1) baby	
(2) infant calls	
(3) human	
(4) home	
More keywords can be created by the administrators as necessary	
Sound source(s)	Baby Crying
Source isolation	1 (isolated single source)
Contextual information	Home recording of a male child a few weeks old.
Recording quality (on a 1 to 7 scale)	7
File origin	The file was obtained from Freesound.org
Submitted by	Brian Gygi (bgygi@ebire.org)
Recording details	
Type of equipment used	N/A
Distance from the source	N/A
Recording environment	N/A
Recording date/time	N/A
Recording level (SPL)	N/A
Usage history	
Citation	None
Behavioral data available	Yes
Identifiability p(c)	1.00
Familiarity (1-7)	5.92
Typicality (1-7)	2.57
Number of downloads	0
File and waveform statistics	
File format (current)	PCM.wav
File size	5.88 MB
Sampling rate	44,100 Hz
Quantization depth	16 bits
Bit rate	1411 kbps
Duration	34.967 sec (34967 ms)
Number of Channels	Stereo
DC offset	L: 1.245% R: −1.244%
Number of clipped samples	L: 0 R:0
Mean rms (in dB)	L: −3.99 dB R: −3.86 dB below maximum (0 dB)
Peak (in dB)	L: −0.61 dB R: −1.0 dB below maximum (0 dB)
Acoustic analysis	
Loudness (sones)	22.66
Sharpness (acum)	1.76
Spectral centroid (Hz, scale value)	L: 27.88 Hz (A0, +23 cents) R: 23.63 Hz
Spectral spread (Hz)	780.3
Pitch salience	0.74 (moderately high)
Pitch	535.3880 Hz
Autocorrelation peaks (No. and Hz)	None

Table 1: Continued.

Qualitative ratings	
Harshness	N/A
Complexity	N/A
Size	N/A
Appeal	N/A
Comments	This sound was originally titled 31527_Erdie_baby3.wav on Freesound. It was uploaded to Freesound by user Erdie and is covered under the Creative Commons License


East Bay Institute For Research and Education
Department of Veteran's Affairs, Martinez, CA
Environmental Sounds Research
Home
Sound Library
My Account
About Us
——Selection Sets——
Add To Selection Set
6
1
/ 12
1..6 / 68
Search | Clear Search

ID	Description	Length	MP3
17	Airplane at 600 ft.	25.4639	Airplane4_60_600ft.mp3
8	Airplane flyover 600 ft.	25.4639	Airplane4_60_600ft.mp3
33	Axe	2.76879	axe_a.mp3
35	Basketball	2.97779	basketball_a.mp3
15	Birds at 40 feet	50.8079	Birds1_57_40ft.mp3
22	Bowling	61.2479	Bowling1_75_40ft.mp3

Figure 2: Omnigrid browser for DESRA.

a particular criterion, a random one will be returned (or several random ones, if multiple tokens are requested) and this will be noted in the sound's search history so that usage statistics can be calculated and to prevent the same sound tokens from always being used. The users will be able to save the search criteria and the sound lists returned as part of their user profile. The users will also be able to organize sounds into selection sets to use in the experiment module of the database program (not discussed here) or for downloading and managing the sounds.

8.3. Adding/Editing Sound Data. As mentioned, admin and associates will be able to freely upload sounds, edit all sound data, and create keywords. Users will submit sounds for uploading and suggestions for new keywords, which will be vetted by the admin. Anyone who uploads or submits a sound for uploading will become the originator of that sound. Only admin will be able to delete sounds, with the exception that any originator will have the option to remove one of their sounds from the database.

8.4. Database. Participants can add or edit audio data on sounds they originate and can make ratings on other sounds for such metadata as "loud/soft", "harsh/smooth", familiarity, and typicality and make comments on the sounds.

8.5. Collecting User Supplied Data. Much of the desired data for a sound cannot be calculated and will have to be supplied. These data include behavioral data, such as identification and typicality, semantic attributes such as harshness and complexity, and subjective assessment such as the overall recording quality. There are a few avenues for obtaining these data. The preferred method would be to have access to experimentally obtained data from either the submitter of the sound or from researchers who have used the sound in studies. If that is not available, users can take part in online experiments available on the site which will require them to identify and rate a number of sounds on the various desired attributes under relatively controlled conditions. In addition, the main access page for each sound will provide an opportunity for users to provide ratings for this sound on several dimensions (e.g., via drop boxes to rate a sound for size or appeal). This is an extension to what is already present on the Freesound site, where users can judge the overall quality of a sound from the main screen for that sound. Sound ratings obtained on line for a representative subset of database sounds will also be validated in laboratory experiments under more tightly controlled listening conditions.

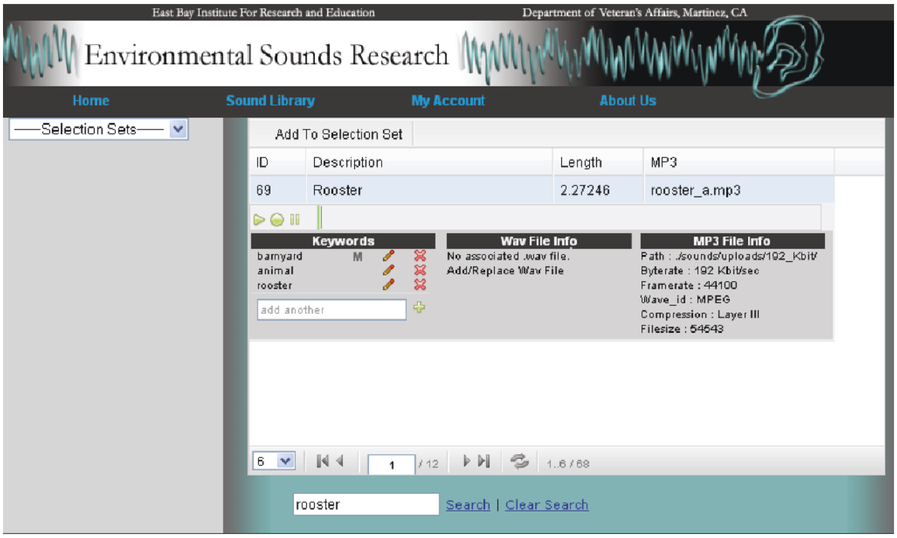
8.6. Other Front End Options. Users will be able to see the most recently uploaded sounds, the most frequently

Table 2

Sound file name	StephenSykes\bartStation_1.WAV
Sound label(s)	Train Station
Sound keywords	
(1) Train station	
(2) Transportation	
(3) City scene	
Sound source(s)	Indoor sounds train coming up, people talking, announcer, stopping, air releasing, doors opening, and conductor speaking
Source isolation	2 (Scene with multiple sources)
Contextual information	Recorded at a BART train station, San Francisco
Recording quality (on a 1 to 7 scale)	5
File origin	The file was obtained from sound recordist Stephen Sykes. It was originally submitted in wave format
Submitted by	Brian Gygi (bgygi@ebire.org)
Recording details	
Type of equipment used	N/A
Distance from the source	N/A
Recording environment	N/A
Recording date/time	N/A
Recording level (SPL)	N/A
Usage history	
Citation	Gygi, B.: Parsing the Blooming Buzzing Confusion: Identifying Natural Auditory Scenes. In Speech Separation and Comprehension in Complex Acoustic Environments Montreal, Quebec, Canada (2004).
Behavioral data available	Yes
Identifiability p(c)	0.95
Familiarity (1-7)	N/A
Typicality (1-7)	N/A
Number of downloads	0
File and waveform statistics	
File format (current)	PCM.wav
File size	5.19 MB (5,446,816 bytes)
Sampling rate	44,100 Hz
Quantization depth	16 bits
Duration	30.877 sec (308777 msec)
Number of Channels	Stereo
DC offset	0
Number of clipped samples	0
Mean rms (in dB)	L: −22.24 dB below maximum (0 dB) R: −21.65 dB
Peak (in dB)	L: −10.87 dB below maximum (0 dB) R: −8.57 dB
Acoustic analysis	
Loudness (sones)	27.31
Sharpness (acum)	1.34
Spectral centroid (Hz, scale value)	L: 113.68 Hz (A#2 -43) R: 108.91 Hz
Spectral spread (Hz)	3136.4
Pitch salience	0.42 (average-low)
Pitch	N/A
Autocorrelation peaks (No. and Hz)	None

Table 2: Continued.

Qualitative ratings	
Harshness	N/A
Complexity	N/A
Size	N/A
Appeal	N/A
Comments	This sound is freely distributable. The recording quality is decent but not outstanding



ID	Description	Length	MP3
69	Rooster	2.27246	rooster_a.mp3

Keywords: barnyard M, animal, rooster

No associated .wav file.
Add/Replace Wav File

Path : ./sounds/uploads/192_Kbit/
Byterate : 192 Kbit/sec
Framerate : 44100
Wave_id : MPEG
Compression : Layer III
Filesize : 54543

rooster

Figure 3: Search screen for DESRA.

downloaded, and the highest or lowest on various criteria, for example, most identifiable, sounds rated loudest or softest, or sounds with specific acoustic attributes, such as pitch strength or rhythmicity.

9. Search Examples

(a) A researcher wants to test some common environmental sounds in hearing impaired people. He wants 20 easily identifiable sounds with limited high frequency content that are not rated as being too "harsh" (and thus, unpleasant for hearing aid users), under 3 s in duration and three tokens of each sound.

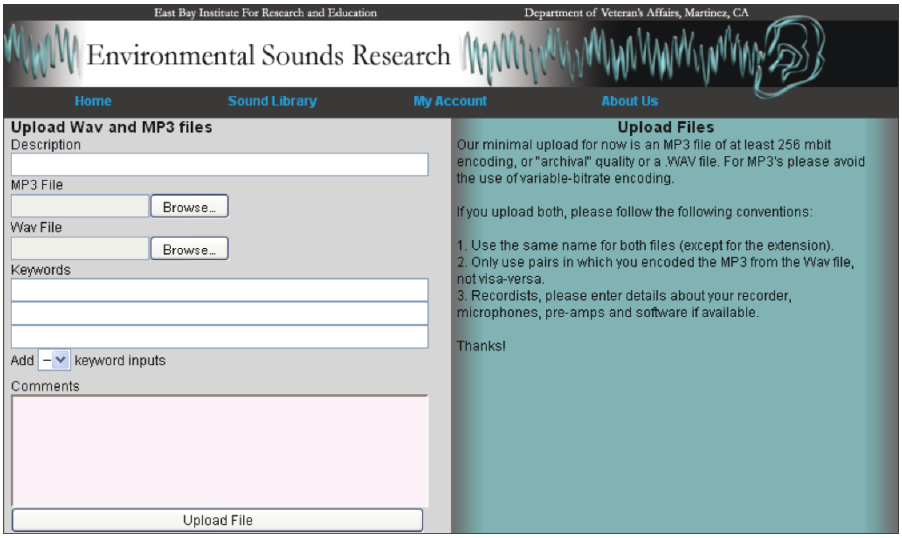
(b) A sound designer wants an unspecified number of sounds to include in a horror film, to take place in daytime and nighttime settings in a rural location. She wants the sounds to have a range of rms values, that is, some high intensity, some medium, and some low intensity, and she wants the sounds to be rated as chilling or intense, while being somewhat difficult to identify.

(c) A musician wants some sound samples to drop in a song to match the lyrics. The samples should be short (under 500 ms), identifiable, and have a certain pitch to match the key of the song. He also wants some longer samples (around 1 s) with a strong rhythm to take the place of scratches or drum breaks.

10. Future Additions: Query by Example

A current feature of many music search engines is "Query by Example" (QBE) in which a user can search for a song that "sounds like" something, either a selection from another song or by some descriptive terms, such as "light and romantic." One example is the Shazam application for the iPhone which can recognize a song based upon a sample submitted to it [41]. It would be useful to apply this paradigm to environmental sounds, so that users could search for a sound that "sounds like" a sample submitted to it (e.g., if a user had a car crash sound they liked and wanted more that sound like it, they could retrieve those) or to identify a hard to identify sound sample based upon returned matches.

However, extending the technology that is currently used in most Music QBE searches is problematic for environmental sounds. Most musical QBE searches do an encoding of the song signal using some compression algorithm, such as Mel Frequency Cepstral Coefficients (MFCCs) or projections onto basis functions, which is the MPEG-7 standard. The compressed version is compared to stored examples and the closest match returned via a distance metric, a common one being a Gaussian Mixture Model [42, 43], which is one of the options in the MPEG-7 standard [44, 45]. These programs are greatly aided by the fact that nearly all musical examples have a similar structure. They are harmonic, which makes



Upload Files

Our minimal upload for now is an MP3 file of at least 256 mbit encoding, or "archival" quality or a .WAV file. For MP3's please avoid the use of variable-bitrate encoding.

If you upload both, please follow the following conventions:

1. Use the same name for both files (except for the extension).
2. Only use pairs in which you encoded the MP3 from the Wav file, not visa-versa.
3. Recordists, please enter details about your recorder, microphones, pre-amps and software if available.

Thanks!

Figure 4: Upload interface for DESRA.

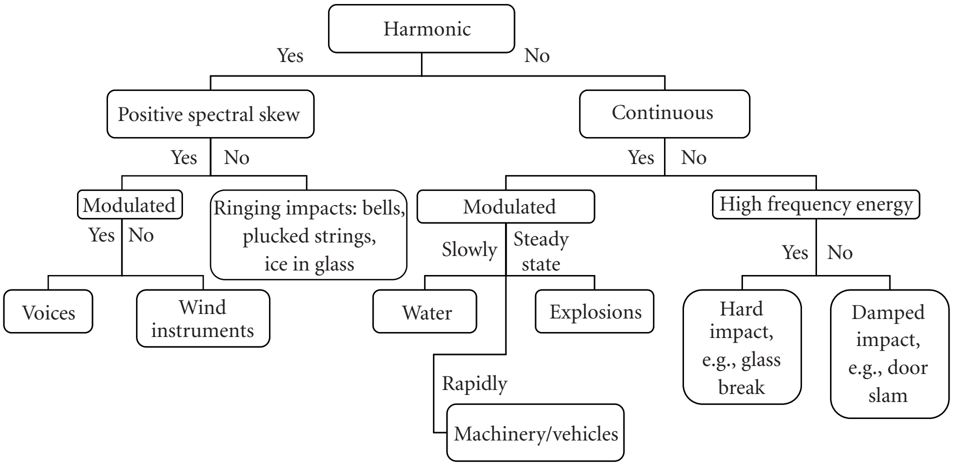


Figure 5: Proposed decision tree for automatic environmental sound recognition.

encoding by MFCCs particularly effective, extended in time, continuous (not many silent intervals), and nearly all have a few basic source types (strings, percussion, wood winds, and brass).

Environmental sounds, on the other hand, since they are produced by a much wider variety of sources, basically encompassing every sound-producing object in the environment, are much more varied in terms of their spectral-temporal structure, some being continuous and harmonic (a cow mooing), some continuous and inharmonic (wind blowing), some impulsive and inharmonic (basketball bouncing), and some impulsive and harmonic (ice dropping in glass). Finding a common coding scheme that can encompass all of these has proven quite difficult, and most systems that classify and recognize music well do quite poorly with a wide range of environmental sounds [46, 47]. It should be noted that this refers to individual environmental sounds in isolation. When multiple sound sources are combined in a soundscape the envelope tends to be smoother, and the long term spectrum approaches a pink noise [48, 49]. In this case, algorithms used for musical classification also perform quite well [50].

In musical QBE systems it is often the case that a musical sample is first associated with a certain genre, such as “rock” or “classical” due to gross acoustic characteristics common to members of that genre. Some algorithms for environmental sounds will similarly initially classify a sound clip based on a semantic taxonomy and then use signal features to narrow the search. An example of this is Audioclas [40, 51, 52] which uses Wordnet semantic classifiers [53] and was incorporated into Freesound.org’s similarity search procedure. However, in [40] it was reported that the probability of retrieving conceptually similar sounds using this method was only 30%. An alternate taxonomy put forth in this issue by [54] is based on the one formulated by Gaver described earlier. A comparison of the two can be found in the Roma et al. piece in this issue [54].

However, there is another way to restrict the search space and thus enable better automatic recognition of environmental sounds which uses only signal properties. In [16] strong correlations were found between the ranking of sounds in a multidimensional similarity space and acoustic features of these sounds. For example, sounds that were grouped together on one dimension tended to be either strongly harmonic or inharmonic. A second dimension reflected the continuity or discontinuity of the sound. Based on this finding, a decision tree can be proposed for automatic classification of sounds, as shown in Figure 5. While this does not cover all the sounds, it is a fairly simple structuring that does account for a large percentage of the sounds necessary for an effective classification system and would greatly enable the development of a true automatic classification scheme for environmental sounds.

11. Summary

The structure of a database of environmental sounds has been outlined, which will relate to the way people listen to sounds in the world. The database will be organized around the sources of sounds in the world and will include a wide variety of acoustic, contextual, semantic, and behavioral data about the sounds, such as identifiability, familiarity, and typicality as well as acoustic attributes such as the spectral centroid, the duration, the harmonicity, semantic attributes of sounds, such as "harshness" or "complexity", and details of the recording, for example the location, the time, the distance of the microphone from the source, and the recording level, along with a citation history. A flexible search engine will enable a wide variety of searches on all aspects of the database and allow users to select sounds to fit their needs as closely as possible. This database will be an important research tool and resource for sound workers in various fields.

Acknowledgments

The authors would like to thank Kim Ho for her work in collecting and annotating a large number of sounds from the Internet. This work was partially supported by a Merit Review Training Grant from the Department of Veterans Affairs Research Service, VA File no. 06-12-00446 to B. Gygi and by a grant from the National Institute of Health/National Institute of Deafness and Communication Disorders no. DC008676 to V. Shafiro.

References

[1] V. Shafiro, "Identification of environmental sounds with varying spectral resolution," *Ear and Hearing*, vol. 29, no. 3, pp. 401–420, 2008.

[2] B. Gygi, G. R. Kidd, and C. S. Watson, "Spectral-temporal factors in the identification of environmental sounds," *Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1252–1265, 2004.

[3] B. Gygi and V. Shafiro, "Environmental sound research as it stands today," in *Proceedings of the Meetings on Acoustics*, vol. 1, p. 050002, 2008.

[4] V. Shafiro and B. Gygi, "How to select stimuli for environmental sound research and where to find them," *Behavior Research Methods, Instruments, and Computers*, vol. 36, no. 4, pp. 590–598, 2004.

[5] W. W. Gaver, "What in the world do we hear? An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[6] R. M. Schafer, *The Tuning of the World*, Knopf, New York, NY, USA, 1977.

[7] B. Truax, *Acoustic Communication*, Ablex, Westport, Conn, USA, 2001.

[8] M. Droumeva, "Understanding immersive audio: a historical and socio-cultural exploration of auditory displays," in *Proceedings of the 11th International Conference on Auditory Display (ICAD '05)*, pp. 162–168, 2005.

[9] J. A. Ballas and J. H. Howard, "Interpreting the language of environmental sounds," *Environment and Behavior*, vol. 19, no. 1, pp. 91–114, 1987.

[10] W. W. Gaver, "Everyday listening and auditory icons," in *Cognitive Science and Psychology*, p. 90, University of California, San Diego, Calif, USA, 1998.

[11] J. J. Jenkins, "Acoustic information for objects, places, and events," in *Persistence and Change: Proceedings of the 1st International Conference on Event Perception*, W. H. Warren and R. E. Shaw, Eds., pp. 115–138, Lawrence Erlbaum, Hillsdale, NJ, USA, 1985.

[12] N. J. Vanderveer, "Ecological acoustics: human perception of environmental sounds," *Dissertation Abstracts International*, vol. 40, no. 9, p. 4543, 1980.

[13] J. J. Gibson, "Survival in a world of probable objects," in *The essential Brunswik: Beginnings, Explications, Applications*, J. J. Gibson, Ed., pp. 244–246, Oxford University Press, Oxford, England, 2001.

[14] T. L. Bonebright, "Perceptual structure of everyday sounds: a multidimensional scaling approach," in *Proceedings of the International Conference on Auditory Display*, Laboratory of Acoustics and Audio Signal Processing and the Telecommunications Software and Multimedia Laboratory, Helsinki University of Technology, Espoo, Finland, 2001.

[15] B. L. Giordano and S. McAdams, "Material identification of real impact sounds: effects of size variation in steel, glass, wood, and plexiglass plates," *Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 1171–1181, 2006.

[16] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception and Psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[17] R. E. Pastore, J. D. Flint, J. R. Gaston, and M. J. Solomon, "Auditory event perception: the source-perception loop for posture in human gait," *Perception and Psychophysics*, vol. 70, no. 1, pp. 13–29, 2008.

[18] M. M. Marcell, D. Borella, M. Greene, E. Kerr, and S. Rogers, "Confrontation naming of environmental sounds," *Journal of Clinical and Experimental Neuropsychology*, vol. 22, no. 6, pp. 830–864, 2000.

[19] L. M. Heller and B. Skerrit, "Action as an organizing principle of auditory cognition," in *Proceedings of the Auditory Perception, Action and Cognition Meeting*, Boston, Mass, USA, 2009.

[20] A. S. Bregman, "Auditory scene analysis: hearing in complex environments," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, S. McAdams and E. Bigand, Eds., pp. 10–36, Clarendon Press, Oxford, UK, 1991.

[21] J. A. Ballas and T. Mullins, "Effects of context on the identification of everyday sounds," *Human Performance*, vol. 4, no. 3, pp. 199–219, 1991.

[22] B. Gygi and V. Shafiro, "The incongruency advantage for environmental sounds presented in natural auditory scenes," *Journal of Experiemental Psychology*, In press.

[23] R. Leech, B. Gygi, J. Aydelott, and F. Dick, "Informational factors in identifying environmental sounds in natural auditory scenes," *Journal of the Acoustical Society of America*, vol. 126, no. 6, pp. 3147–3155, 2009.

[24] L. N. Solomon, "Search for physical correlates to psychological dimensions of sounds," *The Journal of the Acoustical Society of America*, vol. 31, no. 4, p. 492, 1959.

[25] C. Carello, K. L. Anderson, and A. J. Kunkler-Peck, "Perception of object length by sound," *Psychological Science*, vol. 9, no. 3, pp. 211–214, 1998.

[26] D. J. Feed, "Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events," *Journal of the Acoustical Society of America*, vol. 87, no. 1, pp. 311–322, 1990.

[27] A. J. Kunkler-Peck and M. T. Turvey, "Hearing shape," *Journal of Experimental Psychology*, vol. 26, no. 1, pp. 279–294, 2000.

[28] B. Gygi and V. Shafiro, "The incongruency advantage for sounds in natural scenes," in *Proceedings of the 125th Meeting of the Audio Engineering Society*, San Francisco, Calif, USA, 2008.

[29] B. Gygi, "Parsing the blooming buzzing confusion: identifying natural auditory scenes," in *Speech Separation and Comprehension in Complex Acoustic Environments*, Montreal, Quebec, Canada, 2004.

[30] E. A. Bjork, "The perceived quality of natural sounds," *Acustica*, vol. 57, pp. 185–188, 1985.

[31] G. R. Kidd and C. S. Watson, "The perceptual dimensionality of environmental sounds," *Noise Control Engineering Journal*, vol. 51, no. 4, pp. 216–231, 2003.

[32] G. von Bismarck, "Timbre of steady sounds: a factorial investigation of its verbal attributes," *Acustica*, vol. 30, no. 3, pp. 146–159, 1974.

[33] L. N. Solomon, "Semantic approach to the perception of complex sounds," *The Journal of the Acoustical Society of America*, vol. 30, pp. 421–425, 1958.

[34] C. E. Osgood, "The nature and measurement of meaning," *Psychological Bulletin*, vol. 49, no. 3, pp. 197–237, 1952.

[35] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *Journal of Experimental Psychology*, vol. 19, no. 2, pp. 250–267, 1993.

[36] P. Daniel and R. Weber, "Psychoacoustical roughness: implementation of an optimized model," *Acustica*, vol. 83, no. 1, pp. 113–123, 1997.

[37] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, Berlin, Germany, 1999.

[38] M. Slaney, "Auditory Toolbox: a Matlab toolbox for auditory modeling work," Tech. Rep. Apple Computer no. 45, 1995.

[39] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679–688, 1982.

[40] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, N. Wack, and P. Herrera, "Nearest-neighbor automatic sound annotation with a WordNet taxonomy," *Journal of Intelligent Information Systems*, vol. 24, no. 2-3, pp. 99–111, 2005.

[41] A. Wang, "The Shazam music recognition service," *Communications of the ACM*, vol. 49, no. 8, pp. 44–48, 2006.

[42] D. P. W. Ellis, "Audio signal recognition for speech, music, and environmental sounds," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, p. 2424, 2003.

[43] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames 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.

[44] M. Casey, "General sound classification and similarity in MPEG-7," *Organised Sound*, vol. 6, no. 2, pp. 153–164, 2001.

[45] M. Casey, "MPEG-7 sound-recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 737–747, 2001.

[46] K. Hyoung-Gook, et al., "Enhancement of noisy speech for noise robust front-end and speech reconstruction at back-end of DSR system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 716–725, 2003.

[47] D. Mitrovic, M. Zeppelzauer, and H. Eidenberger, "Analysis of the data quality of audio descriptions of environmental sounds," *Journal of Digital Information Management*, vol. 5, no. 2, pp. 48–54, 2007.

[48] B. De Coensel, D. Botteldooren, and T. De Muer, "1/f noise in rural and urban soundspaces," *Acta Acustica*, vol. 89, no. 2, pp. 287–295, 2003.

[49] H. F. Boersma, "Characterization of the natural ambient sound environment: measurements in open agricultural grassland," *Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2104–2110, 1997.

[50] J.-J. Aucouturier and B. Defreville, "Sounds like a park: a computational technique to recognize soundscapes holistically, without source identification," in *Proceedings of the 19th International Congress on Acoustics*, Madrid, Spain, 2007.

[51] P. Cano, et al., "Knowledge and content-based audio retrieval using wordNet," in *Proceedings of the International Conference on E-business and Telecommunication Networks (ICETE)*, 2004.

[52] F. Gouyon, et al., "Content processing of music audio signals," in *Sound to Sense, Sense to Sound: A State of the Art in Sound and Music Computing*, P. Polotti and D. Rocchesso, Eds., pp. 83–160, Logos, Berlin, 2008.

[53] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[54] G. Roma, et al., "Ecological acoustics perspective for content-based retrieval of environmental sounds," *EURASIP Journal on Audio, Speech, and Music Processing*, submitted.