Hybrid sound classification

Pascal Gaillard
CLLE UMR5263, University of Toulouse and CNRS, Toulouse, France.

Matt Coler
INCAS3, Assen, the Netherlands.

Julien Tardieu, Cynthia Magnen
MSHS-T USR3414, University of Toulouse and CNRS, Toulouse, France.

Summary
We posit a classification of sounds useful for studies of sound recognition and identification that accounts for both signal properties (source sound characteristics) and human perception (sound uses). This classification is split into four main branches: (1) systemic (speech and music) sounds, (2) environmental sounds, (3) warning sounds, and (4) animal sounds. We describe the differences between each in terms of criteria related to perception, production and goal. We outline the advantages of our classification, which considers the use of a sound within the context of a communication act, for example, within linguistics; or in harmonics, for musicology. Considering a sound both as a set of acoustic characteristics perceived by a human, and as having particular uses determined by a human, this classification permits a meaningful approach to the study of sound from object- and human-centered perspectives.

PACS no. 43.60.+d,43.90.+v, 43.75.Cd

1. Classification in the sciences

Child participants in an experiment were instructed to freely sort sounds of everyday life [1]. They tended to group together stimuli of human voices, a categorization that could be motivated from one or both of the perspectives below:

1. object-centered: children categorized voice stimuli with respect to shared characteristics among the voices
2. human-centered: children categorized voice stimuli together based on semantic criteria

In the same experiment, children also categorized together the sound of a doorbell and a closing door. The criteria used in categorizing these two stimuli together cannot be made with reference to some aspect of the signal. Instead, we must posit a semantic explanation (perhaps relating to the narrative aspects of a doorbell and door opening).

Thus, sometimes it is meaningful to consider object-centered criteria, other times human-centered criteria, and still other times both. As scientists, we must ask ourselves if we are always consistent with when we use one or another. This is no simple task as we lack tools to orient our interpretations, distinguishing criteria that can be attributed to the signal and that which can be attributed to the perception of that signal. Scientific classification tools are not always precise on the origins of the criteria.

Let us take a step back and explore what is meant by classification, a tool employed in many sciences to explain and understand aspects of a discipline. Here, we use the word *classification* to refer to the practice of arranging things into groups according to certain similarities and differences among those things with the goal of understanding the organization of the world and its natural laws. The word "practice" is key, i.e. there is no underlying classification – classification, like categorization, is an act performed to yield new insight and knowledge.

Classification strategies can be grouped into those that are object-centered, those that are human-centered, and those that are "hybrid", employing elements of both[1].

Object-centered classifications are common in the physical sciences, which mainly rely on criteria relating to physical measurements. These criteria are relevant because the goal of the description of the world is historically and intrinsically linked to these

[1] The division between object- and human-centered approaches is a heuristic, as neither is independent from the other; their relationship as dialectical.

sciences. Consider the classification of stars in astrophysics, according to which stars are classified by their spectra and temperature, two observable and measurable dimensions. Similarly, the biological taxonomy of Linnæus and other functional classifications can be considered object-centered as no interactions with humans are directly considered. This type of measurement is viewed as the optimal perspective from which to understand the variability and the (perceived) structure of the world.

While most object-centered classifications are typical to the physical sciences, it is not the case that human-centered classifications are exclusive to the human sciences. Some of the earliest biological classifications classified plants according to (medicinal, edible, or poisonous) purposes. In the domain of organology, dedicated to the classification of musical instruments, there are many proposed classifications, reflecting the diversity of perspectives which can be relevant [2], including mechanical structure, acoustics, use by a musician, or role in a certain arrangement. Each criterion implies a classification specific to a given task and use. In musical instrument classification, then, unlike that used for stars, interaction with and between humans is the basis for the classification.

This contribution is structured as follows. In §2 we provide an overview of sound classification,for speech (§2.1), music (§2.2), and environmental sounds (§2.3). In §3 we propose a reanalysis of these classifications with a hybrid classification that takes into account human- and object-centered characteristics of the sound. Finally, in §4, we conclude with some remarks on the relevance and implications of this approach for applied purposes.

2. Sound classification

The typical approach to sound classification makes a distinction between three domains: (1) speech sounds; (2) musical sounds; and (3) environmental sounds. This division only arises as a consequence of the division of scientific domains: (1) is studied by phonologists and phoneticians; (2) by musicologists and musicians; and (3) by psychoacousticians. That is, the description of speech sounds is often performed for different scientific motivations than the description of musical and environmental sounds. This explains some of the heterogeneity of the approach to sound classification. For example, formant notion, characteristic of the acoustic signal, is relevant in phonetics to describe vowels but exists as impedance in musical sounds; it is a physical characteristic of the musical instrument, not of the sounds. In other words, in music, the format is relevant to describe the mechanics of the instrument. Formerly and acoustically the same concept is employed to describe two acoustic "realities" relevant to the science impulsing the classification.

The majority of studies on sound classification (and categorization) focus not on identification and semantic categories, but on the first step of perception. Exceptions are to be found in musical acoustics (see [3] for timbre) and within the domain of ecological psychology for environmental sounds [4, 5, 6].

2.1. Speech sounds

A meaningful unit of sound within a language, known as a phoneme (comparable to a "note" in music) is defined with reference to a set of articulatory and acoustic distinctive features, the value of which is determined by its contrastive relationship with other phonemes within a given language[2]. This could mean that the quantity and kind of meaningful sound distinctions in any language is limited by the available distinctive features, though it could also mean that the matter is functional (based on what humans perceive optimally among articulable sounds), at least in part.

Distinctive features, the most basic unit of phonological structure subject to analysis within phonological theory, are units that distinguish one phoneme from another. While there are many proposed models of features, nearly all of them are hybrid insofar as they employ a mix of articulatory [7] (human-centered) and acoustic (object-centered) reference [8]. Consider the aperture feature [±high], which can describe vowels that are produced with the tongue position high in the mouth, like /i, u/ as opposed to /ae, a/. Thus, [±high] is an articulatory feature, since it is defined with reference to articulatory movements. Comparatively, the class feature [±sonorant] describes specific aerodynamic qualities of the voicing of a phoneme commensurate with minimal airflow disruption during the vibration of the glottis. [±sonorant] can be viewed in articulatory and/or aerodynamic terms.

Several models of phonological representation group distinctive features in functional groupings, that can be shown to act together in phonological processes in the languages of the world. Most notably, autosegmental phonology [9][10], a formalism which depicts segments as vertical listings of features on separate tiers connected by association lines (see the top image in Figure 1, in which features are grouped into a tier-like structure, a three dimensional feature geometry).

[2] Consider the aspirated bilabial voiceless stop, [p^h]. Speakers of English classify the /p/ of [stop] 'stop' as the same sound as the /p^h/ in [p^hot] 'pot' whereas speakers of Aymara classify them into two distinct categories because they can distinguish word meanings: /p^haja/ 'cook' /paja/ 'two'. That is, the difference is English is non-contrastive whereas in Aymara it is contrastive. A class of speech sounds that are judged by a native speaker to the be the same sound are a *phoneme*. Each member of that class is called an *allophone*. Therefore, /p^h/ is an allophone of the phoneme /p/ in English, but is a phoneme in English.

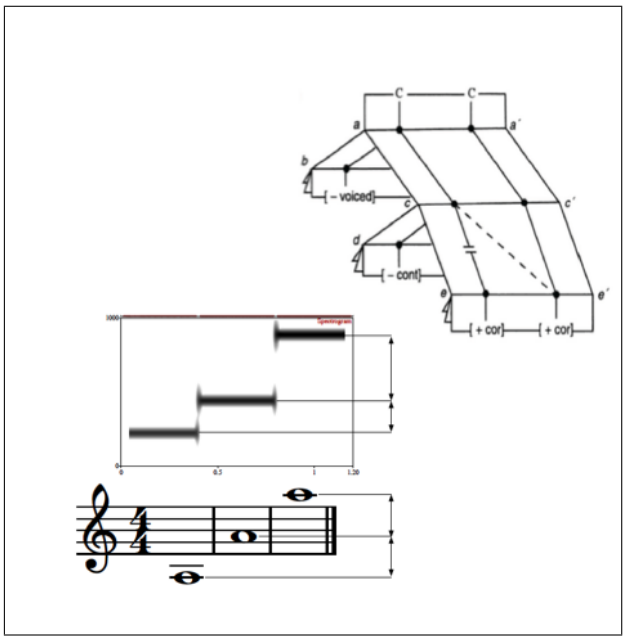

Figure 1. Autosegmental tiers (above) and octave notion (below)

Through the rule-governed manipulation of association lines, structural nodes and terminal features can be associated differently from the way they are represented in the mental lexicon, to account for the discrepancies between the abstract representation of sound sequences and the way they are pronounced.

The relationship between human perception and auditory objects in speech sounds is apparent when one considers the definition of vowels solely from the acoustic characteristics of its formants. The variability of production prohibits fixing the measurement of each formant. Despite the differences in the observed formant structures among speakers of the same language and even withing different realizations of a given vowel by the same speaker. The concept of prototype or percept magnet for example [11] can provide insight. In brief: It is insufficient to define a vowel (or any segment) as a signal with a set of necessary and sufficient acoustic characteristics, disregarding the linguistic system in which the segment contrasts. This system, which is in the perceiver's mind (and not in the signal) drives a perceiver's attention to be sensitive to relevant cues in the signal for a particular goal.

2.2. Musical sounds

In classical Western music, the acoustic characteristics of a note are defined by their relation within an octave. The musical notation extracts the pitch, the duration and, in some case, the loudness from the note. These dimensions are expressed in a representation that trained musicians can read. For example, an octave distance between two sounds is noted in music with an equal distance gap (see Figure 1) regardless of the octave. In frequency measurement, the distance is always different due to the necessity to double the frequency between each octave.

Musical notation, then, takes into account perception but not acoustics of the sound. As for speech, the perception of musical sounds is driven by the musical system: we do not try to hear third tones in Western music, where it is irrelevant, even if its possible to measure some third tones in a performance. The goal of the listener (to hear music) depends of the musical system used and drives perception.

Timbre is another musical dimension, which provides an interesting vantage point from which to frame the link between acoustic parameters and human perception. Timbre is never used to scientifically classify instruments by itself, apart from the orchestration used by composers. In this case, the classification of timbre (in reference to an instrument) depends on several factors, be they linked to religious symbolism or the role of an instrument/s within an orchestra but not necessary to an acoustic definition[3]. From an aesthetic perspective, timbre is the quality of something making sound. From a musical perspective, it is the type of musical instrument that one can hear as a sound family (like a piano) or as a sub-type of instrument (as a concert piano or electric piano). From a scientific perspective, a systematic classification taking into account the acoustic mechanism of the instrument and, at times, the sound resulting from this mechanism, without taking in account the musician itself, was developed as early as the 19th century. But this acoustic classification (and more exactly, this mechanical classification) never used the timbre notion, because this concept, used by musicians, is relevant only in musical context[4].

2.3. Environmental sounds

Unlike musical and speech sounds, environmental sounds[5] do not belong to a system. Therefore, the categorical membership of environmental sounds is considerably less consensual than it is for speech (where sounds can organized into phonetic inventories and described in terms of distinctive features) or music

[3] In fact, no acoustic definition exists to describe timbre acoustically in music.

[4] Outside the three first parameters of a note, timbre seems to be more complicated to define only with physical parameters. Despite a very large variability in the measurement of acoustic dimensions, humans can group various physicals features into one timbre, and change the level of integration according to the goal of the perception (between string and wind instruments in an orchestra, kinds of strings within a string quartet, or two violins to choose the best). The timbre perception as a human notion, certainly based on acoustic features but structured by humans. It is a good representative of a human dimension not always considered in studies of the sound.

[5] Otherwise referred to as everyday sounds or non-specific sounds, "environmental sounds" are those sounds which are not speech or musical sounds.

(where sounds, at least in the western tradition, can be organized into notes, octaves, etc.).

Although physical parameters of sound (fundamental frequency, duration, intensity, etc.) can yield some insight into particular characteristics, a descriptive overview is also required to more adequately characterize the perception of environmental sounds. Varied attempts have been made to provide different kinds of classifications. There are several taxonomies, typically divided into hierarchies > features > dimensions. One well-known example is that of Gaver [12], who classifies sound-producing events into a hierarchical taxonomy. At the highest level are broad classes of materials, then interactions that can cause them to sound (level 1), and into one of three sub-categories: vibrating solids, aerodynamics, and liquids (level 2). Subsequent categorization is defined by simple interactions that can cause the above categories to sound.

While Gaver claims to consider the human in justifying his approach of everyday sounds perception, the taxonomy he proposes offers no human psychological aspect (like "familiarity"), apart from the concept of "event".

Another example of descriptive overview of environmental sounds is the domain-based organization of sound effects on sound effect CDs. This is an apt illustration of how the meaning of environmental sounds is also contingent on so-called "contextual" factors. In most cases, those factors relate to human activities. For example, the sound of a coffee-maker may be heard as a motor operating, an electrical appliance, a certain model of coffee-maker, an alarm indicating that it is time to awaken, or the recognition of one's own coffee-maker. The perceiver often integrates (in)congruent features from memory and/or other sense modalities into an (ordinary) multimodal cognitive representation. What then is "context" if not part of the holistic stimulation?

3. Synthetic classification of human- and object-centered sound classifications

In the psychophysical tradition, physical descriptions of sound are dominated by frequency, amplitude, phase, and duration. According to this view, perception can be characterized with reference to combinations of these measurements. Such analyses disregard the cognitive aspects of perception like attention, memory, familiarity, or other human-centered so-called "higher-level" qualities. Moreover, scientific knowledge of sonic objects is based on the properties of the human senses. After all, the study of acoustic phenomena is primarily concerned with the range of sounds audible to the human ear. Despite being unified within the domain of acoustics, audition gives rise to a diversity of types of sonic objects which do not solely rely on acoustic properties: audition requires reference to semantic characterizations for different types of categories of sonic objects [13]

One human-centered division made in everyday language is that between *sound* and *noise* (see [14] for the same distinction in scientific language - but both *noise* and *sound* are polysemic words in everyday and scientific discourse), the distinction of which is not acoustic but human-centered. In general, sounds intentionally produced by devices purposely designed for such a goal are typically considered "sounds". Sounds that are the byproduct, as opposed to the purpose, of the functioning of some artifact, are typically considered "noises". The case is not always so clear. A closing door does not have musical components. But, with overlay and repetition, as is common to the "musique concrête" genre, one may perceive musical characteristics[6]. The difference is in the use of the sound by the perceiver, not in the sound itself. This use of the stimulation suggests than the context is not necessarily exogenous, but mainly endogenous, linked with our actions. Addressing the question with which we ended the last section, contexts do not exist *a priori*, but are constructions.

In this remainder of this section, we outline a hybrid classification of sounds. However, it is important first to specify what sort of object we refer to by "sound". We propose a primary distinction between inaudible and audible sound. The former being an object of study only to physicists (who can describe inaudible sound with reference only to acoustic parameters) and the latter being an object of study for physicists, too - but also psychophysicists, human scientists, and others. Audible sounds fit the commonsense understanding of sound as vibrations that travel through a medium which are audible when they reach a human or animal ear. Casting aside audible vibrations which occur in the absence of any perceiver[7] and perceivers with different hearing capabilities, we are struck with the observation that the study of (audible) sound requires reference to a perceiver. Accordingly, we cannot use exclusively acoustic characteristics as the main criteria on which to classify audible sounds and to describe, for example, the differences between e.g. whispered speech, salsa music, and chainsaws.

Tables I and II provide a useful tool for researchers studying how different individuals categorize the "same" stimuli differently. The former typically refers to the descriptions of sounds as a signal

[6] Cf. the speech-to-song effect described by Deutsch [15] and exploited in the form of tape loops by many pioneering minimalist composers like Reich, whose 17'56" piece "It's gonna rain" (1965) is created entirely from the looped speech of a Pentecostal preacher. This track is considered fundamental in what would later become the genre of ambient music.

[7] This definition providing an interesting answer for the zen question of whether a tree falling in the ones makes a sound even if no one hears it

where the latter to the sound perceived by humans. Each column provides an example of the use of these criteria. These examples are not exclusive and do not take in account the diversity of speech, music, etc.

When considering speech (in a generic sense), we can consider more than criteria generally used to explain the "perception" of speech. From the object-centered side, we can take into account how speech is adapted for human hearing: a speaker attempts to adapt production (e.g. frequency or speed) to the hearing of the perceiver. This criterion is related to perception (of the perceiver) but it can also be described in terms of the speech signal. We can also consider the high attentional level used in speech perception as a necessity to adequately understand the signal. But this criterion is not "in" the signal. It is a construction of the mind, generating attention to something not necessarily wanted by the speaker or salient in the signal. In these two cases, we speak about perception, but in one side "in" the signal, and in other side "in" the perceiver.

Continuing with this distinction, the goal in Table I refers to the communicative goal in the speaker's production. Salience measured in the speech signal (intonations and accents for example) are observable in the signal itself and produced voluntarily by the speaker. Comparatively, the goal in Table II can be used to describe how those "same" intonations and accents are used by humans to construct meaning in the speech signal, as a necessity. There is evidence (e.g. [16]) of an automatic and low level of interpretation of the speech signal. This type of process is in the receiver, not the signal.

The use of these criteria, separating object and subject can be instrumental in distinguishing what we are observing, be it something in the mind of the receiver or in the signal.

4. Discussion

The application of sound event recognition is broad, but highly challenging. Most research into sound event recognition [17], [18], [19] relies on feature extraction techniques coming from automatic speech recognition, like Mel-frequency cepstral coefficients to describe a signal, along with hidden Markov models to classify "something" into predefined categories. Other approaches employ a bag-of-frames method [20], which uses long-term statistics of the spectral range to identify auditory scenes from real-world recordings. For more on these approaches, see Niessen [21]. Regardless, without human- and object-based classification methods, they face similar difficulties.

Any automatic sound classification technique must rely primarily on the sound signal as input. Psychological aspects like "actions" or "goal" are not easily convertible into algorithms. Nonetheless, even if the automatic classification uses a model of recognition based of signal analysis, it can use human goals. In other words, it could use something like the script concept [22] to achieve sound recognition. For example, upon hearing a car door open and close, an engineer turn and start, one may expect the sounds that follow to be those of a car driving away, the volume of the engine sounds decreasing over distance. The variety in the engine sounds, due to the different phase of the process of starting, pulling out, and departing can be recognized by an automatic classification as the same thing: a car leaving.

Methodologies used for sound event recognition are based only on signal analysis. They do not take into account parameters like context and necessity, which cannot be detected only with reference to fluctuation in air pressure. Sensors, then, are at once more sensitive than humans (insofar as sensors have greater sensitivity and are consistent) and less sensitive than humans (insofar as sensors are incapable of recognizing something based on previous knowledge). The latter inability arises from the fact that such "information" are not "encoded" "in" the signal, but rather that the signal provides some cue that enables a physical characteristic to be constructed "in" the perceiver [13].

Although some there are some attempts [8] to address this by including human parameters like expectation into automatic indexation, even these approaches lack a clear, replicable and efficient classification for scientific purposes.

For human processing, the signal as information points to previous knowledge or categorical representation that gives meaning through categorical membership to the acoustic stimulation. Given the impossibility of getting at a single systematic classification of (environmental) sounds, it seems as if the criteria of categorization of sounds are so diverse and knowledge/context-dependent that an exhaustive formal structure will remain beyond our grasp. Nonetheless, the authors believe that awareness of object and human-centered aspects of acoustic/auditory phenomena can stimulate progress in an approach to acoustic event detection that respects the diversity of human hearing in an attempt to understand the cognitive processes and maybe to more efficiently convert them into an algorithm.

References

[1] A. Berland, P. Gaillard, M. Guidetti, P. Barone: Perception of Everyday Sounds: A Developmental Study of a Free Sorting Task. PloS One, 10(2), 2015, e0115557.

[2] G. Dournon: Organology. In: H. Myers (eds.). Ethnomusicology, an introduction. MacMillan, 1992.

[3] J. M. Grey : Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America, 61(5), 1977, 1270–1277.

[8] Including CIESS, a project run by some of the authors, thanks to support from an ANR grant.

Table I. from the object, with some but not exclusive example of use (general speech, western music, warning, environmental sounds and sound from animals)

related to	criterion	speech	music	warning	environ	animal
Perception	specifically adapted to human hearing	+	+	+	-	-
	actively controlled by humans	+	+	-	-	-
Production	produced intentionally	+	+	+	-	+
	produced by mechanical parts	-	-	+	-	-
	produced by natural causes	-	-	-	+	-
	produced by sth internal to the system	-	-	+	+	-
Goal	communicative goal	+	-	+	-	+
	aesthetic goal	-	+	-	-	-
	designed to do sth acoustic	+	+	+	-	+
	explains system state	/	/	+	-	/
Other	results from the measurement of system state	/	/	+	-	/
	bilateral communication	+	/	-	/	-
	monolateral communication	-	/	+	/	+

Table II. from the human perceived auditory object, with some but not exclusive example of use (general speech, western music, warning, environmental sounds and sound from animals)

related to	criterion	speech	music	warning	environ	animal
Perception	high attentional level	+	+	+	-	-
	learned implicitly	+	+	+	+	-
	organized in a system	+	+	-	-	+
Production	by humans for humans	+	+	-	-	-
Goal	expresses warning	+	-	+	+	+
	indicates environ. state	-	-	+	+	-

[4] N. van der Veer: Ecological acoustics: Human perception of environmental sounds. 1980.

[5] J. Ballas: Common factors in the identification of an assortment of brief everyday sounds. Journal of experimental psychology 19(2). 1993, 250–267.

[6] D. Dubois: Categories as acts of meaning: The case of categories in olfaction and audition. Cognitive Science Quaterly, 1, 2000 35–68.

[7] N. Chomsky, M.Halle: The Sound Pattern of English. Cambridge: MIT Press. 1968.

[8] R. Jakobson, G. Fant, M. Halle: Preliminaries to Speech Analysis. Cambridge MA: MIT Press. 1952.

[9] J. Goldsmith: An Overview of Autosegmental Phonology. Linguistic Analysis, 2, 1976, 23–68.

[10] N. Clements: Does sonority have a phonetic basis. Contemporary views on architecture and representations in phonological theory, 2009, 165–175.

[11] P. K. Kuhl: Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. Perception & Psychophysics, 50, 1991, 93–107.

[12] W. Gaver: How Do We Hear in the World? Explorations in Ecological Acoustics. Ecological Psychology. 1993.

[13] D. Dubois, M. Coler, H. Wörtche: Knowledge, sensory experience, and sensor technology. In: C. Rangacharyulu, E. Haven, B. Juurlink (eds.). The World in Prismatic Views. World Scientific, 2013, 97–133.

[14] M. Niessen: Sound and Noise in Sonic Environmental Studies: Comparing Word Meaning in Discourses of Community Noise and Soundscape Research. Soundscape Research, 2013, 1âĂŞ-10.

[15] D. Deutsch: Musical Illusions and Paradoxes. Philomel Records, 1995.

[16] R. Näätänen, P. Paavilainen, T. Rinne, K. Alho: The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clinical Neurophysiology, 118(12), 2007, 2544–2590.

[17] M. Cowling, R. Sitte: Acoustic signal processing,Audio signal processing,Environmental sound recognition,Joint time-frequency feature extraction, Non-speech sound recognition. Pattern Recognition Letters. 15(24) 2003, 2895–2907

[18] B. Defréville, P. Roy, C. Rosin, F. Pachet: Automatic recognition of urban sound sources. Proc. of the 120th AES Convention, 2006.

[19] A. Mesaros, T. Heittola, A. Eronen, T. Virtanen: Acoustic event detection in real life recordings, Proc. EUSIPCO, 2010.

[20] J.J. Aucouturier, B. Defreville, 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, 2007.

[21] M, Niessen, C. Cance, & D. Dubois. Categories for soundscape: Toward a hybrid classification. Proc. Internoise 2010.

[22] Roger C. Schank, R. Abelson: Representation. Lawrence Erlbaum Associates, Hillsdale. 1977.