Machine Audition:
Principles, Algorithms and Systems

Wenwu Wang
University of Surrey, UK

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Book Publications: Julia Mosemann
Acquisitions Editor: Lindsay Johnston
Development Editor: Joel Gamon
Publishing Assistants: Casey Conapitski and Travis Gundrum
Typesetter: Michael Brehm
Production Editor: Jamie Snavely
Cover Design: Lisa Tosheff

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Machine audition : principles, algorithms, and systems / Wenwu Wang, editor.
p. cm.
Includes bibliographical references and index.
Summary: "This book covers advances in algorithmic developments, theoretical frameworks, and experimental research findings to assist professionals who want an improved understanding about how to design algorithms for performing automatic analysis of audio signals, construct a computing system for understanding sound, and to learn how to build advanced human-computer interactive systems"--Provided by publisher. ISBN 978-1-61520-919-4 (hardcover) -- ISBN 978-1-61520-920-0 (ebook) 1. Computational auditory scene analysis. 2. Signal processing. 3. Auditory perception--Computer simulation. I. Wang, Wenwu, 1974- TK7881.4.M27 2010
006.4'5--dc22
2010010161

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 4
Audition:
From Sound to Sounds

Tjeerd C. Andringa
University of Groningen, Netherlands

ABSTRACT

This chapter addresses the functional requirements of auditory systems, both natural and artificial, to be able to deal with the complexities of uncontrolled real-world input. The demand to function in uncontrolled environments has severe implications for machine audition. The natural system has addressed this demand by adapting its function flexibly to changing task demands. Intentional processes and the concept of perceptual gist play an important role in this. Hearing and listening are seen as complementary processes. The process of hearing detects the existence and general character of the environment and its main and most salient sources. In combination with task demands these processes allow the pre-activation of knowledge about expected sources and their properties. Consecutive listening phases, in which the relevant subsets of the signal are analyzed, allow the level of detail required by task and system-demands. This form of processing requires a signal representation that can be reasoned about. A representation based on source physics is suitable and has the advantage of being situation independent. The demand to determine physical source properties from the signal imposes restrictions on the signal processing. When these restrictions are not met, systems are limited to controlled domains. Novel signal representations are needed to couple the information in the signal to knowledge about the sources in the signal.

INTRODUCTION

This chapter addresses machine audition and natural audition by carefully analyzing the difficulties and roles of audition in real-world conditions. The reason for this focus is my experience with the development of a verbal aggression detection system (van Hengel and Andringa, 2007). This system was first deployed in 2004 and 2005 in the inner city of Groningen (the Netherlands), by the

DOI: 10.4018/978-1-61520-919-4.ch004

company Sound Intelligence, and helps the police to prioritize camera feeds. It is the first commercial sound recognition application for a complex target in uncontrolled (city) environments.

Furthermore, the system is a prototypical and a rather idiosyncratic example of machine audition. It is prototypical because it must function, like its natural counter part, in realistic and therefore complex social environments. Inner cities are complex because they are full of people who speak, shout, play, laugh, tease, murmur, sell, run, fall, kick, break, whistle, sing, and cry. The same environment contains birds that sing, dogs that bark, cars that pass or slam with doors, police and ambulances that pass with wailing sirens and screeching tires, pubs that play music, wind that whines, rain that clatters, builders who build, and many, many other rare or common sound events.

What makes the system idiosyncratic is simple: *it must ignore all these sounds*. The simplest way of doing this is to make it deaf. However, there is one type of sound that should not be ignored: verbal aggression. Of the 2,678,400 seconds each week, the system is interested in about 10 seconds of verbal aggression and has to ignore the other 2.6 million seconds. Fortunately the situation is not as bleak as it seems. The police observers may graciously allow the system some false alarms, as long as the majority of them are informative and justifiable. This means that the system is allowed to select no more than about 50 seconds per month, which corresponds to 0.002% of the time. Ignoring almost everything, while remaining vigilant for the occasional relevant event, is an essential property of the natural auditory system. It requires the (subconscious) processing of perceptual information up to the point of estimated irrelevancy. That is exactly what the system aims to do.

After a considerable period of optimization the system worked (and works) adequately. However it has one major restriction: the system is not easily extended or adapted to other tasks and environments. Every migration to a new city or new operating environment requires some expert-time to readjust the system. Although this is a restriction the system has in common with other applications of machine learning and pattern recognition, it is qualitatively different from the performance of human audition. In general, the comparison between a natural and an artificial system is not favorable for the artificial system. In fact, it is quite a stretch to refer to the function of the verbal aggression detection system as similar to audition: I consider the comparison degrading for the richness, versatility, robustness, and helpfulness of the natural auditory system.

My experiences with the development of the verbal aggression detection system have led me to reconsider my approach to machine audition. This chapter aims at the functional demands of audition, both natural and artificial, because I consider the functional level the level where most progress can be made. The functional level is both beneficial for theories about the natural system and for the design of technology that can function on par with its natural counter-part.

Working with police-observers, who are not at all interested in the technology itself, but only in whether or not it actually helps them, was also revealing. Expectation management was essential to ensure a favorable evaluation and the eventual definitive deployment of the first system. This is why I use a common sense definition of audition as starting point and why I aim to develop systems that comply with common-sense expectations. Only these systems will be truly impressive for the end-user.

The chapter addresses four connected topics. The next section addresses the demands that operating in the real world poses on a system. It is followed by an investigation into the special options that sounds offer as source of information about the environments. This forms the basis for the longest section, which addresses how the natural system detects the relevance in the signal and can analyze it up to the desired degree. The flexible and knowledge intensive analysis requires signal

representations that are situation independent. The properties of these representations form the topic of the next section that suggests that sonic knowledge should represent source physics to some degree. The chapter is concluded with a short vision on the possible future of machine audition applications.

The chapter will not provide definite answers, nor ready to implement algorithms, but it will provide a number of design constraints that help to design and build systems that approach the performance of that of a human listener, and that, in my opinion, truly deserve the term machine audition.

REAL WORLD DEMANDS

Scientific Challenges

Research that is aimed at building systems endowed with cognitive functions normally associated with people or animals is becoming increasingly important to our information society. Such cognitive systems should exhibit a high degree of robustness and flexibility in coping with unpredictable situations and to handle simple everyday situations with common sense and without detailed pre-programming (http://cordis.europa.eu/fp7/ict/cognition/home_en.html). True cognitive systems prove their value in open-ended environments. A EU program manager stated this bluntly as: "Stop making toy-systems! We know you can do that. Impress us outside the lab!" (Roberto Cencioni at SAMT2006). This statement refers to an important limitation of current technology that becomes apparent whenever scientists and engineers attempt to scale-up laboratory demonstrators of intelligent systems to actual applications that can deal with real-world conditions.

For example, despite considerable technological progress, exponentially increasing computational power, and massive training databases, automatic speech recognition (ASR) has been a promising technology for more than 30 years (O'Shaughnessy, 2008). During all these years ASR-systems have never been able to impress a majority of potential users (the observation that many pc-users do not know or care that their computers has preinstalled ASR-software is testimony of this). It is likely that these ASR-systems constrain the majority of users more than they prefer. Although the performance in terms of recognition accuracy on benchmark tests has improved immensely, the basic probabilistic architecture and robustness of ASR systems has remained the same. It is possible this approach suffers from the fundamental limitations characteristic of modern intelligent systems. Since natural audition can deal with the complexities of the real world it is instructive to contrast the two types of systems.

Natural Intelligence

A natural intelligent system, such as a human listener, is extremely versatile; both in terms of function and operating environment. A single auditory system can perform a multitude of different functions that vary from detecting an animal rustling between leaves, determining the temperature of tea being poured in a cup, identifying a problem with a car, recognizing a similar melody when played on different instruments, to recognizing speech. In addition, the operating environment of a natural auditory system can be very complex in the sense that it may contain many different objects and processes. It may also be partially or largely unknown and variable. And finally it is unlimited in the sense that novel events can be introduced at any time and that the sonic world may be extended indefinitely. Basically, a natural intelligent system can dynamically adapt to the demands that complex, unknown, variable, and unlimited environments pose. Humans, therefore, only need a single auditory system for all sonic tasks and environments.

Artificial Intelligence

Although system designers aim to make systems as versatile as possible, the operating environment of engineering systems is fundamentally limited by the definition of the task to be executed: a dictation system is not a keyword-spotting device. The seemingly innocent choice to build a single-function system limits engineering to a different domain than natural systems. The rational is that dictation systems pose different demands than keyword spotting devices, and music genre detection poses different demands than verbal aggression detection. By focusing on task-specific demands it is assumed that effective technology can be developed more efficiently, because irrelevant functionality does not have to be developed. This is a very dangerous assumption because the existence of (natural) multipurpose systems that can perform these tasks is no proof that single purpose systems can perform individual tasks. It is quite possible that some level of "multi-purposeness" is essential for the robust execution of single tasks in real-world environments.

For current technology to function optimally, or even adequately, it is important to impose limitations on the operating environment. Typically, the operation environment must be simplified to one source that must be known in advance and of which the properties are assumed to be constant and representative of the training data. For modern ASR-systems one can even conclude that they have been designed for input of which everything except the word order is known. Essentially, the operating environment must be limited to conditions the system can handle. All additional uncertainty will reduce the reliability of the results. In other words, these systems require the user to keep the operating environment under control. Without the user's essential role in controlling the environment, the output of the system is unspecified. At best it withholds output (which might be erroneous as well), at worst it produces random results, i.e. results with no apparent relation to the input. All in all, the current approaches lead to a large number of task and domain specific systems in combination with an essential role for a human user to prevent nonsense output. Figure 1 depicts the scope of human and artificial tasks and operating environments. Modern technology is clearly limited to a subset of the human scope.

Each new prototype shows that scientists and engineers can develop task and domain specific systems and applications. Although the intellectual challenge of adding advanced new example systems and applications might be considerable, the true scientific challenge is to develop new strategies for intelligent systems that, by design, remove the limits of modern technology. A good starting point is to study the demands of functioning in a world without simplifications: the real world if you like.

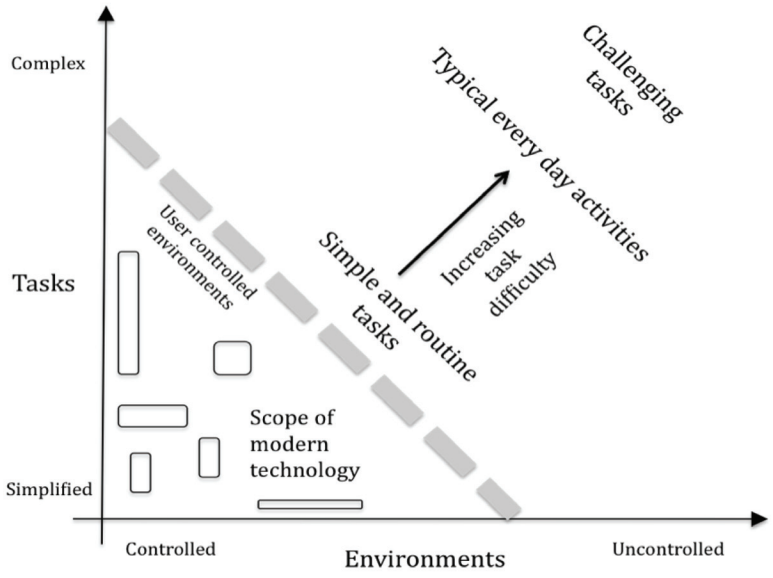
We will describe the flexibility of the natural perceptive system to change its tasks and to rely on a smart interplay between bottom-up and top-down processing. But we will start with a study of the functional role of the auditory system. This role will be approached from a common sense definition that reflects how non-experts think and talk about audition and secondly from an exploratory perspective on the opportunities offered by modern perception research.

THE FUNCTIONAL ROLE OF THE AUDITORY SYSTEM

Common Sense Definitions

Experts often use different and typically more constrained terminology than non-experts. This may be confusing in cases where expert terminology utilizes common words for a specific domain-dependent purpose. The term "machine audition" might have this confusing aspect because it suggests a strong relation between a class of applications and a highly developed, intimately familiar but partially understood natural ability.

Figure 1. The typical scope of human task operating environments extends from complex tasks in uncontrolled environments to simple tasks in controlled environments. Modern applications are limited to simple tasks in controlled environments. The dashed gray bar denotes the likely scope limitation of modern technology.



Researchers in the field of machine audition will always be confronted with user-expectations based on common sense notions of audition. Therefore, it is instructive to study common-sense definitions of audition and related words to make the expectations of the users of the technology explicit. Because the definitions within a single dictionary show often a high degree of circularity (a=b=a), the definitions selected below here have been derived from a range of (online) dictionaries. All selected definitions are quoted verbatim.

For the word "audition" one might find for example:

- **Audition (noun):** the act, sense, or power of hearing.

Audition therefore refers either to the activity of hearing like in the case of "involving an act of hearing", or secondly "the sense of hearing as similar to the sense of vision", and thirdly "the enabling capacity to hear". Because dictionaries define words using other words this leads to the problem of defining the word "hearing". Of this word there are two relevant variants:

- **To hear (verb):** perceive with the ear the sound made by someone or something.
- **Hearing (noun):** the faculty of perceiving sounds.

Of the two definitions the first refers to "the act of hearing", while the second refers to "the enabling capacity to hear". Both definitions combine variants of sound and perception that require a definition as well. Two different concepts of sound are introduced: 1) "a" or "the" sound and 2) sound (without article):

- **Sound (uncountable noun):** vibrations that travel through the air or other media

and can be heard when they reach a person's or animal's ear.
- **A sound (countable noun):** audible sound stemming from someone or something.

The uncountable noun refers to the physical phenomenon. The countable noun uses the uncountable noun with the adjective audible. With a focus on "sound" it refers to the audible subset of the physical phenomenon that originated from a single source, but with the focus on "audible" it refers to the role of the sound source in audition.

For the verb to perceive and its noun perception one can find:

- **To perceive (verb):** become aware or conscious of something [through the senses].
- **Perception (noun):** the ability to become aware of something through the senses.

Perception refers to more senses than hearing alone. However, when applied to hearing, the first definition specifies the "act of hearing", while the second specifies "the enabling capacity to hear". These definitions refer to the adjectives aware and conscious.

- **Aware (adjective):** noticing or realizing something, knowing that something exists because you notice it or realize that it is happening.
- **Conscious (adjective):** keenly aware, fully appreciating the importance of something.

Apparently "being aware" and "being conscious" differ in degree, but both refer to being able to include the existence of something in a reasoning process. "Being aware of something" refers to knowledge of its existence, while "conscious of something" refers to fully appreciating its consequences. The difference between the two degrees is reflected in the words hearing and listening. Hearing in the sense of "can you hear the boy?" may be a question directed at whether or not the boy was audible. The question "did you listen to the boy" presupposes a more detailed analysis of what you heard. Consequently, the difference between hearing and listening seems to correspond to the difference between a passive and a more active process. In fact listening and listener can be defined as:

- **Listening (verb):** making an effort to hear something
- **Listening to (verb):** to give one's attention to a sound
- **Listener (noun):** someone who listens, especially someone who does so in an attentive manner

These definitions introduce the term "attention" of which one dictionary definition reads:

- **Attention (noun):** concentration of mental powers, especially on a particular object

The "concentration of mental powers" in this definition involves memory and reasoning. When memory and reasoning are focused on a particular object we can use the memories we have stored about it to reason about the object and its consequences, In other words, we can determine its meaning:

- **Meaning (noun):** inner importance, psychological or moral sense, purpose, or significance

While the users of a language negotiate the meaning of a word, the meaning of something for someone is the individual importance one gives to a thing.

Elaborated Common Sense Definition of Audition

Most dictionary definitions so far have used words like object, something, and someone to denote

an event, thing, object, or person in the world. The perception process makes these available for the perceiver. It is now possible to formulate an elaborated common sense definition of audition based on the previous dictionary definitions.

Audition (noun): the capacity for, or act of sound-based processing in which the existence of something or someone becomes mentally available (in the case of awareness), this availability can be used in a reasoning process to discover the consequences of what has been perceived (in the case of consciousness).

Someone who is limited to the passive form of audition can be called "a hearer", but this term is rarely used in this context. Furthermore future systems for machine audition will typically be required to assist content based reasoning and automated decision-making, the active process is consequently very relevant for machine audition. If (human, animal, or natural) audition is to be contrasted to machine audition it is reasonable to contrast a listener to a system for machine audition.

The elaborated common sense definition is consistent with modern insights in perception (see section Audition = hearing + listening). It will be argued that the sound induced processes that lead to awareness, correspond to hearing and that the more detailed conscious analysis is typical of listening (Harding et al., 2007). The "problem" with the elaborated common sense definition of audition is that there are no easy short cuts to understand the phenomenon of audition. Awareness and conscious processing are intimately involved in audition. The design of systems that approach the common sense expectations of most users requires some engineering equivalent of these phenomena. This does not imply that we have to wait until computers are conscious. It only indicates that some of the essential functional consequences of both awareness and consciousness cannot be ignored. The properties of attention will be a guideline to formulate these ingredients.

Another apparent complication is the requirement that audition has to work reliably in the real world. However it will be argued that real-world constraints pose a number of very helpful restrictions to guide the development of models of audition. The next sections make a first step in that direction by addressing the special possibilities that sound affords as a channel to derive information from the environment.

The Specialties of Audition

Each sense has unique strength and limitations. Sound for example is caused by mechanically interacting objects and carries detailed information about the objects and their interaction (Gaver, 1993). A direct consequence of this is that it makes sense to ask the question "What caused the sound?" In the case of vision, light-sources like the sun or lamps produce or cause the light that we see. But in the visual domain the most relevant question is not "What caused the light?" but "Which objects reflected the light last?" The related question "What reflected the sound last?" is generally less important, but it is of particular relevance for echolocating animals and blind people.

The answer to the question "What caused the sound?" requires a transformation from sound as a physical phenomenon to sounds as an interpreted sound. The interpreted sound refers either to an interpretation as a real-world cause (an explanation like "I hear a car") or to the subsets of the signal that belong to the sources that the sounds refer to ("This part of the sound is caused by a car"). The answer to "What caused the sound" can be used as further specification of the common sense definition of audition. Audition may even be defined as the process that makes sounds from sound, hence the subtitle of this chapter.

Compared to light, sound carries less far, which limits sonic information to the proximal environment. Furthermore, visual processing is spatially oriented while auditory processing is spectrally oriented. Both the visual and the auditory system are computationally limited and the whole system

needs to make choices concerning what to process superficially and what to process in detail. The visual field covers only a subset of all spatial directions and only a tiny fraction with a cross-section of 3˚ can be analyzed in detail. Consequently, it is not trivial to aim detailed visual analysis at the most relevant region of space. Because auditory information is spectrally oriented, spectral information from all directions is pooled, which makes auditory sensitivity omnidirectional. This omnidirectionality, in combination with sounds being informative of mechanical interactions in the proximal environment, implies that audition has properties that make it very suitable to monitor the proximal environment.

A special and highly relevant aspect of the interaction with an unpredictable environment is the ability to detect and analyze unexpected but potentially important events. The combination with the directional sensitivity of binaural integration allows audition to guide visual attention to the correct spatial region. Responding to unexpected events requires audition to interrupt ongoing mental activities in favor of analyzing the unexpected event. This has a negative consequence when the unexpected event is not worth the interruption. Irrelevant sonic events that interrupt mental activities and interfere with the tasks and goals of the individual are annoying. The prominence of noise annoyance in our society forms strong support for audition's role in detecting unexpected and potentially relevant events.

Masking and Reverberation

While listeners move around, they are exposed to a wealth of different sound sources in a wealth of acoustically different spaces. Some of the sources may be intimately familiar, others completely novel. Some may be proximal and with a favorable signal-to-noise ratio, others are partially masked by louder or more proximal sources. As a consequence much of the information of a source is masked by other sources. Whatever is still available as reliable source information is distributed in varied ways over the time-frequency plane.

There is an important additional complication. Close to the source within the reverberation radius, direct sounds such as smoothly developing harmonics may be much more prominent than the sum of the indirect reflections of objects and walls. Outside reverberation radius the indirect sound dominates. Because the indirect sound consists of delayed copies of the direct sounds that recombine with random phase, the resulting signal shows fluctuations and temporal and spectral smearing that changes smoothly developing harmonics in fluctuating narrow-band noise. Consequently, the indirect contributions are quite different from direct sounds.

Hence reverberation will ensure that even when the relevant ranges of the time-frequency plane have been found the information it represents will appear quite different depending on the ratio between direct and indirect sounds. However, it takes a trained ear, or paired comparisons of speech samples recorded without and with reverberation, to detect these prominent effects on the signal (Nábělek & Robinson, 1982). The same amount of indirect sound poses ASR-systems with serious fluctuations on the input parameterization that impair the recognition process. This is the reason that most ASR-systems require the use of a close-talking microphone.

Everyday vs. Musical Listening

The qualitative difference in sensitivity to reverberation is indicative of the relevance of the question "What caused the sound?" Human auditory processing seems to focus on the cause of the sounds, while modern ASR-systems seem to focus on the detailed properties of the sound and not its probable cause. The difference between cause and signal properties is reflected in Gaver's (Gaver, 1993) distinction between everyday listening and musical listening. Everyday listening refers to a description of the sounds (as countable noun) in

terms of the processes or events that produced them. For example, we do not hear a noisy harmonic complex in combination with a burst of noise; instead we hear a passing car. Likewise we do not hear a double pulse with prominent energy around 2.4 and 6 kHz, but we hear a closing door. In contrast, musical listening focuses on the properties of sound (as uncountable noun) and couples sensations like pitch and loudness to physical properties like frequency and amplitude. A typical aspect of musical listening is that we can focus on melodies, hear differences in the timbre of different instruments, and determine that someone sings at the wrong pitch.

Controlled vs. Uncontrolled Sounds

Considering all sources of variability in acoustic environments and the problems they pose to modern engineering systems, it makes sense to introduce the term uncontrolled sound as a sound of which no a priori knowledge is available: uncontrolled sounds can be any sound out of the set of all possible sound combinations. All knowledge about its contents must be derived from the signal itself. Arbitrary sounds can be contrasted to controlled sounds, in which some essential intervention has ensured that some of the problems of arbitrary sounds have been defined away. Recording with a close-talking microphone or telephone constrains the signal in a similar way as careful recording by a researcher does. In both cases it results in a limitation to a (convenient) subset of all possible sounds. For machine audition, it is important to distinguish approaches for controlled and uncontrolled sounds. In the case of controlled input the type of control needs to be defined as precise as possible. For example, the results reported by Cowling & Sitte (2003) on environmental sound recognition presuppose one typical instance of a sound source per recording. This constraint is quite severe for signals described as environmental sounds.

The Role of Meaning

The answer to the question "What caused the sound?" is only part of the task of a listener. According to the elaborated common sense definition of audition, the possible consequences of the sound producing events should be investigated in a reasoning process. Each audible sound contributes information. If this information is not included, behavioral options may be suboptimal and sometimes dangerously inadequate. Therefore, the auditory system should predict the consequences of the events in a proper context and in doing so give meaning to the event.

The term meaning is still ill defined, but the "meaning of something to someone" denotes the personal importance of something. The importance of something is of course personal and highly situation dependent. Hence the importance is defined through the interactions of the individual with its environment. The meaning of something for someone can be therefore defined as the difference between mental and behavioral states with and without it. If something affords strict behavioral options it is obviously meaningful. However if not including it does not change behavioral options, it is meaningless. Audition, and especially the process of listening, is about figuring out the most effective behavioral options afforded by the sounds. This process, which can be described as maximizing the meaning of sound, requires an intimate interaction between (auditory) perception, the rest of cognition, and the environment.

Note that the *meaning of something*, without a direct reference to a person, refers typically to the linguistic meaning of a thing. Linguistic meaning is a common denominator of personal meanings and is, as all words, the result of a negotiation between the users of a language to ensure that the use of the word leads to a predictable interpretation by a listener or reader. This form of meaning is not referred to in this chapter.

Maximizing the meaning of sound in terms of the effective behavior it affords can be considered

the goal of machine audition as well. Take the example of a verbal aggression detection system that guides the attention of police observers to the most informative surveillance camera feeds (van Hengel & Andringa, 2007). Commercial systems like this wait for very rare, but highly significant, events by computing a moving average of signal evidence indicative of verbal aggression. The moment a threshold is exceeded, a possible aggressive event is indicated to the police observer for more detailed inspection. Because observers do not want to be bothered with a large number of "meaningless" events, the improvement of systems like these is aimed at making the output of the system as meaningful as possible. Ideally the systems should indicate, with explicit justification, why they considered this event as more relevant than all ignored events. Something similar is the case in dictation systems: meaningless or bizarre recognition results are less appreciated than an equal number of word-errors that do not change the meaning of the dictated message.

Summarizing the Role of Audition

The analysis of the functional role of audition in this section was based on the special properties of sound in combination with the demands a complex and uncontrolled operating environment pose. This led to a number of conclusions about the role of audition that are equally valid for human and machine audition. These can be summarized as follows:

- From sound to sounds: one purpose of audition is to separate sound in a way that explains the causes of the sounds that constitute it.
- Uncontrolled input: nothing is known in advance from the signal, consequently the signal itself (including its full history) must inform the system how it should be analyzed
- Work everywhere and always: internal representation must be based on knowledge that is always and everywhere applicable.
- Detect the unexpected: compare expectations with the actual signal and detect mismatches.
- Listening: the search for the most meaningful interpretation of the signal.

These conclusions will be elaborated in the course of this chapter. The next section focuses on the way attention helps to estimate task relevance.

ESTIMATING TASK-RELEVANCE

In uncontrolled environments the input is unconstrained. Therefore, an unknown fraction of the input will be relevant for the system's task. When the response of a system is based on either a random selection of evidence or on an arbitrary mixture of sources, it has no relation to the information in the signal and is extremely unlikely to be correct. Therefore, determining the relevant part of the signal is a prerequisite to correct task performance. One strategy to deal with the clutter of task-irrelevant contributions is to process each part of the input up to the moment it can be ignored without the risk of discarding essential information.

Attentive Listening

This leads to a central design guideline, namely that all input needs to be processed up to the point of estimated task-irrelevance. The naive approach is to hope that the target is much louder than the background so that the background can be easily ignored or discarded. This approach is helpful in many situations and in particular in situations in which the noise can be assumed to be stationary or known. However, extensive research on human audition in noise, called the cocktail-party effect (Cherry, 1953, Bronkhorst, 2000), has shown that

listeners reliably detect and recognize speech (and other sources) in situations where the target speech is neither louder nor otherwise different from a "background" of babble sounds. In other words listeners can detect and recognize the target whenever there is a minimum of reliable evidence unmasked by the background (Allen, 1994). However, listeners must focus more and more attention on the task if the sonic environment becomes more challenging.

Attention is a core cognitive process that allows animals to selectively focus on one aspect of the environment while ignoring the rest. Attention is intimately related with the solution to dealing with uncontrolled environments because it ensures the efficient allocation and application of the brain's algorithmic and knowledge resources. For these reasons it is very useful to study the algorithmic properties of attentional processes in some detail.

Bottom-Up Attention

Attention has been a target of research for many decades, which has led to a consensus on a number of its key aspects. For example, it is possible to differentiate signal driven (bottom-up) and knowledge driven (top-down) attentional processes (Koch & Tsuchiya, 2007, Knudsen, 2007). In the first form, attention is captured by salient parts of the input, which can suspend the current mental task in favor of the analysis of the salience sound. Attention can be captured involuntarily by either sudden and/or unexpected changes in the situation or by well-trained stimuli (Gopher & Iani, 2003). The saliency of sudden or unexpected stimuli allows attention to be captured by mismatches between expected and actual input. Alternatively, the saliency of well-trained and personally relevant stimuli makes a conversation more difficult when you hear your own name mentioned in the background. Moreover, emotional stimuli like angry faces are easier to respond to than neutral faces in the same conditions (Whalen et al., 1998). Your name and emotional individuals are both of high potential relevance for you as a system, which justifies a strong effect towards involuntarily suspending the current mental task or activity in favor of analyzing the unexpected or otherwise relevant stimulus.

Top-Down Attention and Consciousness

The top-down variant of attention is a prerequisite for advanced cognitive processes like reasoning, language, and the (strategic) analysis of complex input (Dehaene et al., 2006). Top-down attention is said to govern task execution through the flow of internal and external information. Consequently, it governs the planning and selection of responses (Gopher & Iani, 2003). As such it is also involved in the algorithmic processing of sound. Top-down attention is intimately related to conscious processing. Consciousness can be interpreted as our current best summary of the knowledge pertaining to the mental current state of the organism, its environment, and the behavioral options it affords (Koch & Tsuchiya, 2007). Top-down attention is a process that actively structures the input by applying stored knowledge in memory.

The result of this attentive structuring is, at least in part, a configuration of interacting discrete entities (objects, concepts) that describes the state-of-the-world in so far it is relevant for the individual and its goals. For example, during the analysis of a picture, observers have constant access to the best current interpretation, while attentional processes ensure that suitable knowledge is made available to improve the current interpretation more and more. The analysis continues up to the point that the individual's goals and task demands do not benefit further. Generally, the estimation of the relevance in the input is adequate, but errors, accidents, and misunderstandings do occur when relevance has been judged inadequately. For example while writing it is easy to miss a typo, especially when you know what the text should be. While driving it is quite possible to miss an important sign with far reaching consequences.

And when you fail to pick up the irony in a voice it is easy to misunderstand each other.

Inattentional Blindness

Structuring novel and complex input and formulating verbal reports are only possible when the perceptual input is 1) sufficiently strong and informative, and 2) if task specific top-down attention is present (Dehaene et al., 2006). The combination of demands is important for the phenomenon of inattentional blindness (Mack, 2003). This phenomenon suggests that when top-down attention is engaged in a sufficiently demanding task we are blind for task-irrelevant information, even when it is very strong and informative. Inattentional blindness is typically demonstrated with a visual task where two teams pass balls and the subject has to count the number of times the ball is passed within each team. Most task-engaged subjects fail to "see" clearly visible task-irrelevant objects like a woman with an open umbrella or a man in a gorilla suit. A recent study (Cartwright-Finch & Lavie, 2007) shows that more participants fail to notice the presence of task-irrelevant stimuli when the perceptual load is increased through increasing the number of items or by requiring more subtle perceptual discriminations. Cartwright-Finch concludes that stimulus complexity and not task-complexity, is the main limiting factor of attentional blindness.

Inattentional Deafness?

Similar effects are well known in audition: it is usually easy to focus on one perceptual stream while being completely unaware of the contents of other streams. In early work on the 'cocktail party effect', Cherry (1953) found that, when listeners attend to one of two talkers, they might not be aware of the meaning or even the language of the unattended speech. But they are aware of its presence and basic properties such as pitch range and the end of the message. Similarly, it is possible that novel details in music and soundscapes only become noticeable after multiple exposures. And from that moment on it is difficult to understand how these details could be missed. Hence, it is often impossible to analyze a signal fully in a single presentation, because task demands and the complexity of the signal will lead to a partial analysis. The more information and knowledge a listener has about the situation and the signal, the more detailed analysis is possible. Which demonstrates that activating task-relevant knowledge through attentional processes allows us to select matching evidence from the signal that can be missed without knowledge preactivation.

Perceiving Task-Relevant Information

The results in vision and audition substantiate the conclusion that perceptual systems process input up to the point of estimated task-irrelevance. For example, in the case of the demanding visual task, irrelevant information did not become consciously available even though the balls passed frequently behind the gorilla or the woman with the umbrella. The task-irrelevant persons were treated as 3-D objects behind which balls can pass, so they were clearly represented at some fairly advanced level of visual processing. However, the task-irrelevant object was not consciously accessible for most participants, although information about it was available. In the auditory example the situation is similar. Task-relevant speech can be tracked and understood, while only the existence and basic properties of task-irrelevant speech can be estimated. The existence and basic properties of the interfering speech are task-relevant for the auditory system because we use pitch-based grouping (Bregman, 1990) to separate speakers. The awareness of only these task-relevant properties is a particular convincing example that processing up to the point of task-irrelevance is performed. Note that this is only the case in demanding tasks. Less demanding tasks leave processing capacity

for task irrelevant, but possibly system relevant, processing (Cartwright-Finch & Lavie, 2007).

Connect Signal Subsets to Interpretations and Vice Versa

Attention and estimating task-relevance in the input are closely related. Signal-driven attention capturing occurs through salient subsets of the input. It corresponds to either a mismatch between expected and actual input or to well-trained stimuli with system relevance. Salient stimuli forces the system to process the salient subset up to the conscious interpretation of the stimulus at the cost of other mental tasks. In task-driven attention, attentional processes lead to the selection of task-relevant subsets of the input by task-relevant interpretations. Hence bottom-up attentional processes connect subsets of the stimulus to (potentially conscious) interpretations, while top-down attention connects conscious interpretations to subsets of the stimulus. The combination forms a very flexible process because it can, to good approximation at least, be assumed that whatever is relevant in the input can be coupled to the correct interpretation. This process is depicted in Figure 2.

The bottom-up processes are qualitatively different from the top-down processes. The bottom-up processes lead to possible interpretation hypotheses, while the top-down processes capture the signal evidence consistent with generated interpretation hypotheses. The interpretation hypotheses that are able to capture sufficient and consistent evidence remain active and can become conscious. A similar combination of bottom-up hypothesis generation and top-down checking, but exclusively based on functional arguments, was described for speech recognition in Andringa (2002). This proposed a system that leads only to recognition results that are sufficiently supported by a combination of signal evidence and the state of the recognition system.

This section addressed the need for a task-optimized analysis of the signal. Since task-optimization requires task-related knowledge, the next sections will address how suitable knowledge can be made serviceable to capture relevant information from the output. However, the first question is if and how suitable knowledge can be activated at all. A suitable form of knowledge activation has been extensively studied in vision research as the gist of a scene, a notion that seems extendable to the auditory domain.

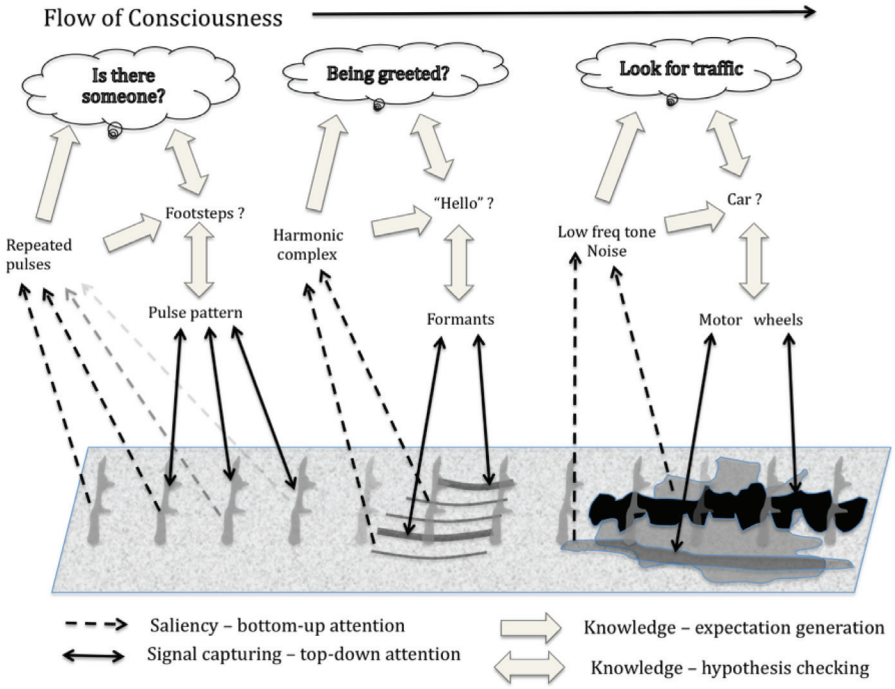
VISUAL AND AUDITORY GIST

Visual Gist

A study performed more than 30 years ago (Potter, 1976) has shown that a preliminary meaningful interpretation of a complex scene occurs within only 100 ms after stimulus onset. This preliminary semantic interpretation is independent on whether or not the scene is expected and occurs independently of the clutter and the variety of details in the scene. This fast and preliminary interpretation is called the 'gist' of a scene (Oliva, 2005). The gist includes all levels of visual information, among which low-level features such as color and contours, intermediate-level features as shapes and texture regions, and high-level information such as the activation of a meaningful interpretation. The gist estimation process is strongly related to bottom-up attention. The gist is also related to top-down attention, because it connects, like top-down attention, perceptual input to interpretations.

The gist can be separated in a perceptual gist, which refers to the structural representation of a scene built during perception, and a conceptual gist, which includes the semantic information that is inferred during or shortly after the scene has disappeared from view. Conceptual gist is enriched and modified as the perceptual information bubbles up from early stages of visual processing (Oliva, 2005) and develops from a fast initial indicative interpretation to a fully reliable interpretation.

Figure 2. Hearing and Listening. A schematic representation of the attentional processes associated with an event sequence involving footsteps, someone saying "Hello", and a passing car. The clouds in the upper part reflect the flow of consciousness and the situation dependent tasks. The lower part depicts a time-frequency plane with pulse-like, tonal and noise-like events. Unexpected sounds (like the onset of the footsteps) or highly trained sound (like the word "hello") can be salient so that attention is directed to subsets of the input. This attentional change leads to task changes and top-down attentional processes that capture task-relevant subsets of signal evidence and that connect it to the explanation the listener becomes conscious of. The listener's state and task changes from determining the presence of a passer-by, to whether or not one is greeted, and to the question if a passing car is relevant. Together these form a meaningful narrative describing real-world events.



In general, speed and accuracy in scene (gist) recognition are not affected by the quantity of objects in a scene when these objects are grouped (Ariely, 2001). While gist information about the type and position of individual items is minimal, the gist represents accurate statistical information about groups of items. Additionally, scene information outside the focus of (top-down) attention becomes consciously accessible in the form of ensemble representations that lack local detail. Nevertheless they carry a reliable statistical summary of the visual scene in the form of group centroids of similar objects (Alvarez and Oliva, 2008). Without specific expectations of the scene, the gist is based on an analysis of the low spatial frequencies in the image that describe the scene only in coarse terms (Oliva, 2005) and that may even be insufficient for a reliable interpretation after prolonged visual analysis.

Gist and Scene Recognition

The rapid activation of the conceptual gist is in marked contrast with prominent views of scene recognition that are based on the idea that a scene is built as a collection of objects. This notion

has been influenced by seminal approaches in computational vision, which have treated visual processing as a strictly bottom-up hierarchical organization of modules of increasing complexity (edges, surfaces, objects), with at the highest level, object identification, and scene schema activation (Marr, 1982).

However, modern empirical results seem more consistent with perceptual processes that are temporally organized so that they proceed from an initial global structuring towards more and more fine-grained analysis (Navon, 1977). The role of the visual gist suggests that visual scenes may initially be processed as a single entity, e.g. a sunny beach, and that segmentation of the scene in objects, e.g. palm trees and tourists, occurs at a later stage. This holistic, whole first, approach does not require the use of objects as an intermediate representation and it is not based on initial stages of segmentation in regions and objects.

Additionally a Reverse Hierarchy Theory (Ahissar and Hochstein, 2004) has been formulated, in which perceptual task learning stems largely from a gradual top-down-guided increase in the usability of first coarse, then more detailed task-relevant information. A cascade of top-to-bottom modifications that enhance task-relevant, and prune task-irrelevant, information serves this process. The result is an ever more efficient perceptual process that is guided by conceptual gist based expectations about input signal detail.

Task Optimized Analysis

The success of a top-down and task-adapted approach depends crucially on the relation between the gist contents and the actual content of a scene. The process of visual gist content activation was modeled by Torralba and Oliva (2003), who reported that eight perceptual dimensions capture most of the three-dimensional structures of real-world scenes (naturalness, openness, perspective or expansion, size or roughness, ruggedness, mean depth, symmetry, and complexity). They observed that scenes with similar perceptual dimensions shared the same semantic category. In particular, scenes given the same base-level name, e.g., street, beach, (Rosch et al., 1976) tend to cluster within the same region of a multidimensional space in which the axes are the perceptual properties. Torralba's and Oliva's results show that (simple) signal properties are able to activate the correct semantic evaluation.

Algorithmically, the notion of a rapidly available conceptual gist allows a task-optimized analysis in which the focus of the analysis shifts stepwise to regions where task-relevant information can be derived from. By first attending to the coarse scale, the visual system acquires a rough interpretation of the input that activates the conceptual part of the gist. The conceptual gist represents scene schemas in memory (Yeh & Barsalou, 2006), which represent knowledge about how situations can develop. Subsequently attending to task relevant salience may provide information to refine or refute the initial estimate. If a scene is unknown and must be categorized very quickly, highly salient, though uncertain, information is very efficient for an initial rough estimate of the scene's gist. However, if one already knows the content of the scene or knows what the appropriate spatial scale for a visual task is (Oliva, 2005), it is possible to initiate a fast verification task at the spatial scale that may lead to a selection of expected details (Schyns & Oliva, 1994)

Is There an Auditory Gist?

The perceptual gist has the algorithmic properties required to make task-relevant information available to the system and as such it is ideal for audition. In contrast to visual gist, the concept of auditory gist has not yet had much scientific attention. Nevertheless Harding et al. (2007) concluded there is ample evidence that auditory processing complies with the ideas proposed for vision. Their paper addresses a number of pro-

posals for an auditory gist, but due to the lack of focused research the auditory gist has not been defined and described in scientifically satisfying terms. In particular they found auditory (and visual) domain evidence that:

- Only the gist of the scene or object is initially processed;
- Processing of the gist is rapid;
- The focus of attention is deployed according to prior knowledge and the perception of the gist;
- Conscious detailed analysis is possible on the part of the scene within the focus of attention;
- Only limited processing of the unattended parts of the scene occurs.

These are all properties consistent with gist-guided expectation-based processing. Completely in line with this evidence is the hierarchical decomposition model (Cusack et al., 2004). In this model of sound perception, basic streaming is performed on the whole input, but only a single stream can be attended and subdivided further. Unattended streams cannot be fragmented further. Because the studies were conducted on (meaningless) tones instead of complex real-world sounds, the effects of task-specific knowledge might not be maximally prominent. Nevertheless it was concluded that if a general idea about the whole signal is obtained, the unattended parts of the signal do not need to be subdivided. For example, during a conversation at a street corner café with speech, music, and traffic noise, it is not necessary that the auditory system segregates music into instruments or speech into individual words if the listener is interested in the sound of people entering in a bus. This is depicted in Figure 3.

Audition = Hearing + Listening

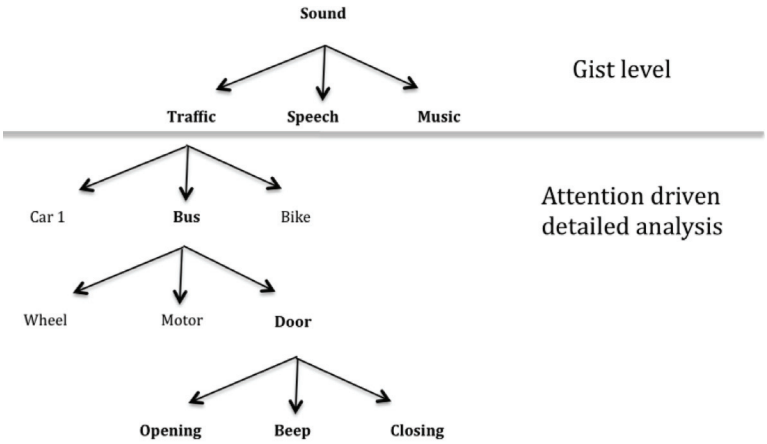
Harding et al. (Harding, Cooke, & Konig, 2007) suggest a 'hearing' stage as an initial bottom-up gist processing stage which provides an overview of the whole auditory scene suitable for higher level processes. The initial processing indicates the likely number of sources and the source categories. Additional, task-specific top-down processes can focus on the attended source and analyze its detail, which they suggest is the 'listening' stage. This stage determines the features of the attended stream. Details of the signal outside the focus of attention will not be consciously perceived, although some limited processing of these regions might occur, typically in ways consistent with processing capabilities of the hearing-stage. Note that these suggestions dovetail nicely with the differences between hearing and listening in the elaborated common sense definition of audition as formulated earlier:

Audition (noun): the capacity for, or act of sound-based processing in which the existence of something or someone becomes mentally available (in the case of awareness), this availability can be used in a reasoning process to discover the consequences of what has been perceived (in the case of consciousness).

In this interpretation, hearing and listening are complementary processes. Hearing detects the existence and general character of the environment and its main and salient sources. In combination with task demands this allows the pre-activation of knowledge about expected sources and their properties. Consecutive listening phases, in which the task-relevant subsets of the signal are analyzed, allow the level of detail required by task- and system-demands.

As was outlined in the section addressing the estimation of relevance, one of the functions of top-down attention is to capture relevant subsets of the output by connecting it to suitable knowledge. This poses several demands on the way the signal is processed, the way information about sounds is stored and accessible, and the way the interpretation is connected to the signal. These will be addressed in the next sections, of which the first subsection focuses on knowledge and

Figure 3. The focus of attention. Not all sonic events are analyzed in similar detail. The gist of the scene with a general analysis of the content is always available. In this case the gist represents the sounds of a street-corner café. Only a single stream, in this case the one belonging to a stopping bus with a focus on the door, is analyzed in detail. Especially when the door-events are partially masked by the other sounds, attentive listening is required to detect them. This reduces the awareness of other sounds in the environment, which might lead to the conscious accessibility of only the events in bold. (Conform Cusack et al., 2004).



signal representations suitable for audition in uncontrolled environments.

THE PHYSICAL CHARACTER OF AUDITORY KNOWLEDGE

Physical Realizability

There is an often ignored but very important and useful constraint on real-world input that makes the task of interaction in the real world considerably less daunting by constraining both top-down expectations and bottom-up signal representations. This constraint follows from the acknowledgement that all input stems from a physically realizable world. Gaver (Gaver, 1993), who studied everyday listening from an ecological perspective, used this constraint implicitly when he stressed the relation between source physics and perception as follows:

"Taking an ecological approach implies analyses of the mechanical physics of source events, the acoustics describing the propagation of sound through an environment, and the properties of the auditory system that enable us to pick up such information. The result of such analyses will be a characterization of acoustic information about sources, environments, and locations, which can be empirically verified. This information will often take the form of complex, constrained patterns of frequency and amplitude which change over time: These patterns, not their supposedly primitive components, are likely to provide listeners with information about the world" (Gaver, 1993, p. 8)

Gaver argues that sonic input, if suitably processed, leads to complex but constrained patterns that are informative of the sources that produced the sounds. Top-down attentional processes should be aimed at the detection and capturing of these patterns.

If individual sources are subject to physical constraints, by extension a natural sonic environment consisting of individual sources is also subject to physical constraints. In fact the whole

sonic environment is physically realizable in the sense that the sounds it produces stem from a physically allowed configuration. This is an extremely important property because it entails that the system can limit the set of possible signal interpretations to those that might actually describe a real-world situation and as such does not violate the physical laws that shape reality. Although this set is still huge, it is at least not contaminated with a majority of physically impossible, and therefore certainly incorrect, signal interpretations. Recognition systems that are based on computationally convenient manipulations that do not use this constraint have no way to decide which of a number of possible interpretations is an allowed state of reality. Without methods to limit the output of engineering systems to the physically probable, these systems cannot be extended from limited and controlled domains to uncontrolled domains.

Physics and Knowledge

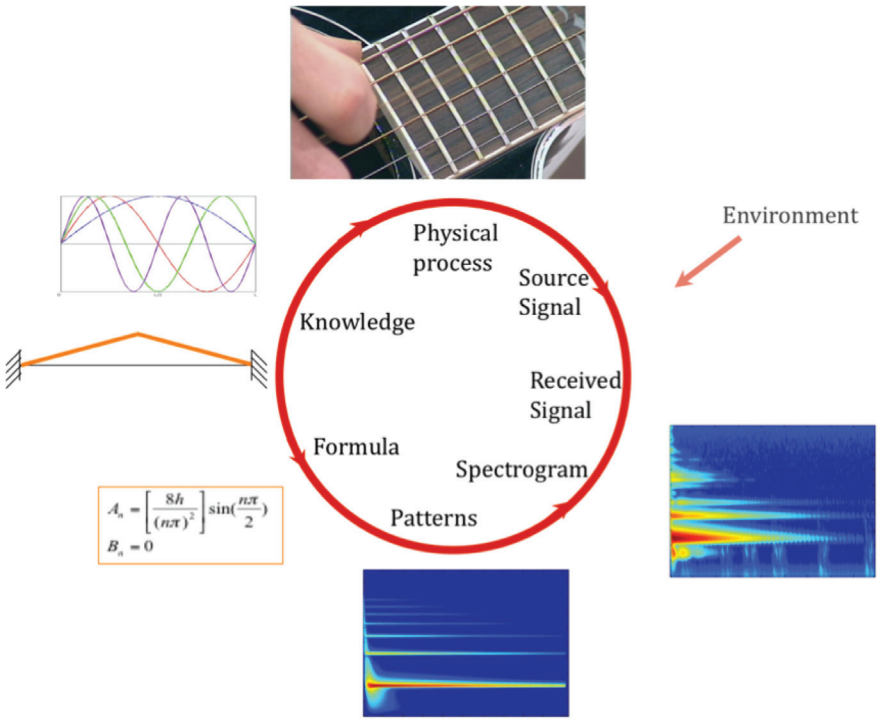
The strong relation between physics and knowledge can be demonstrated by a thought experiment, adapted from Andringa (2002), in the form of the question: "Which sound source cannot be recognized?" We might perform an actual experiment by hiding the sound source in question behind an opaque screen. First you hear a sound and you say "a violin". "That is correct" we say. Then you hear another sound. You hear again a violin and you report that. "Wrong" we say. But you definitely heard the violin. We remove the screen and you see both a violin player and a HiFi-set with very good loudspeakers. The loudspeakers are definitely sound sources and they tricked you the second time. This might not seem particular informative because this is exactly what loudspeakers are used for. However, the point is that the violin will always produce 'the sound of a violin'; it is the only sound that physics allows it to produce and that our auditory system allows us to interpret. The same hold for all other "normal" sound sources. The (ideal) HiFi-set in contrast can reproduce any sound at will. It has no audible physical limitations and as a consequence it has no sound of itself. It will always be interpreted as another sound source as long as the listener is naïve about the true origin of the sound. And even then it is effortless to interpret the sounds it produces as the sound sources it reproduces.

Gaver's argument, generalized to arbitrary modalities, is that the sources (e.g. the sun, sound sources, and surfaces with evaporating odor molecules) and the transmission properties (e.g. reflecting surfaces, decay with distance, wind) do not lead to arbitrary structures, but on the contrary, lead to highly structured patterns that can be estimated by a perceptive system. These patterns can be stored and used as top-down expectations.

This argument leads to the relations in Figure 4 that couples physical representations via two routes to the patterns refered to in Gaver (1993). The counter-clockwise route is via knowledge and expectations; the clockwise route is via a real-world signal and a suitable form of signal processing. For example a guitar sound stems from a physical process involving a string being plucked that can be modeled as a differential equation of which the solutions correspond to a number of modes that can be summarized in a formula. The formula corresponds to the expectation of a pattern of damped sinusoidal contributions. The brain computes something functionally similar, but it uses generalized memories of previous exposures to expect the pattern of damped sinusoidal contributions. The clockwise route is via the real world in which a guitar sound is mixed with other sounds and transmitted through a reverberant environment. The resulting sound can be analyzed and compared with the generalized and idealized expectation. The mismatches can be used to refine the knowledge driven expectation; in this case for example by including the guitar's resonances around 1000-1500 Hz.

This example was idealized in the sense that it was trivial to assign signal evidence to the correct source. Competing sounds makes this more

Figure 4. Two different routes to connect a physical process to a pattern. The counter clockwise route is via knowledge; the clockwise route is via a real world signal and suitable preprocessing.



difficult in normal situations. The signal processing should therefore be optimized to form units of evidence that are highly likely to stem from a single source and that capture the information needed for the counter-clockwise route. The next section addresses this problem.

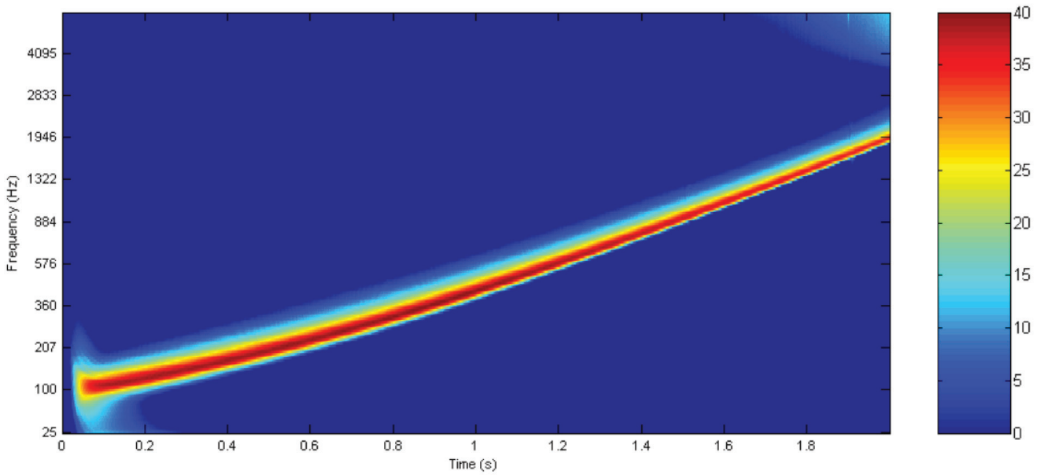
Representing Sounds Physically

Suppose you are presented with a test sound consisting of a tone that starts low and ends at a high pitch: a signal that can be visualized as in Figure 5. The question you are asked is "How many sounds did you hear?"

You are likely to report that you heard a single sound. But why is it a *single* sound? During the interval with sonic energy there was obviously sound, but how many sounds? The justification to call this a single sound is that the signal does not provide any evidence that somewhere during its development it stopped and one or more other sounds took over. While this is not impossible, the probability is vanishingly small that one sound stopped and was smoothly extended by an *uncorrelated* new sound that had exactly the correct phase and energy to ensure no discontinuity whatsoever. This suggests that that our auditory system uses a continuous source development to form a single, and continuous, representation of the sound. This basic assumption formed the basis for Continuity Preserving Signal Processing (Andringa, 2002).

Continuity Preserving Signal Processing (CPSP) is a form of Computational Auditory Scene Analysis (CASA) (Rosenthal and Okuno, 1998) that aims to track the development of sound sources as reliable as possible. CPSP was developed to allow recognition systems to function as often as possible in varying and uncontrollable acoustic environments. CPSP aims to start from the weak-

Figure 5. A single sound in the form of a log-sweep represented as cochleogram according to Andringa, 2002. The signal starts at 100 Hz and ends at 2000 Hz two seconds later. The cochleogram was computed with a transmission line model of the basilar membrane that does not bias special frequencies or points in time like frame-based approaches like an FFT do. As a consequence the development of the sweep is, like its representations in the human cochlea, localized and very smooth.



est (most general) possible basic assumptions. For sounds, the weakest possible basic assumption is that sounds consist of signal components that each shows an onset, an optional continuous development and an offset. The sine-sweep in Figure 5 is an example of a signal component.

Quasi-Stationarity

The inertia of sound sources entails that they cannot change infinitely fast. This entails that sound sources can be approximated with a quasi-stationarity assumption that assumes that the source can be modeled as originating from a process that is assumed to be stationary over short intervals. This is similar to the sample-and-hold process used to transform continuous signals into discrete signals that are suitable for computerized analysis and storage. For speech a quasi-stationarity period of 10 ms is often assumed (Young and Bloothooft, 1997). Quasi-stationarity is a perfectly reasonable assumption, but because it depends on a source dependent stationarity interval, it holds exclusively for the signal of a single and (partially) known source type. If, however, a signal is produced by two speakers, it will change more rapidly and certainly differently than is allowed by the physics of a single vocal tract. Consequently, a form of quasi-stationarity that is only valid for a single source is not justified for mixtures of sources and should be avoided. The same holds for sources outside the reverberation radius.

In uncontrolled environments, the situation is even worse, since a suitable stationarity interval may be impossible to choose. If quasi-stationarity is nevertheless applied, the induced approximation errors will degrade the combined signal irreparably and therefore reduce the probability to reach a correct recognition result. This leads to the conclusion that quasi-stationarity, with a proper time-constant must either be applied to individual signal components or to complex signals, like the speech of a single speaker, for which it holds. As long as the signal, or some selection of it, is not positively identified as suitable for the quasi-stationarity assumption, the application of

quasi-stationarity is not justified and may lead to suboptimal or incorrect results.

The Danger of Frame Blocking

Unfortunately this is the way quasi-stationarity is usually applied. All common approaches to ASR (O' Shaughnessy, 2008), sound recognition (Cowlin & Sitte, 2003), and most approaches to CASA (Hu & Wang, 2006, Wang & Brown, 2006) apply quasi-stationarity, but make no effort to apply it safely. The most common way to apply quasi-stationarity is frame-blocking as essential step before the application of a short term Fourier Transform. Frame-blocking determines that whatever the contents of the resulting window is, it is treated as quasi-stationary with a period equal to the time-shift between blocks and with a spectro-temporal resolution determined by the effective window size. Since this may or may not be appropriate for physical information in the signal, it limits these approaches to controlled domains in which the user can ensure that the detrimental effects are not dominant.

The Safe Application of Quasi-Stationarity

It is possible that the auditory system takes great care to apply quasi-stationarity safely. At least there are no known violations during cochlear processing. In auditory modeling is possible to preserve continuity as long as possible and to postpone the application of quasi-stationarity to the moment it can be justified. The use of a transmission line model of the basilar membrane (or suitable approximation as for example the gammachirp filterbank (Irino and Patterson, 1997) can lead to *cochleogram* as spectrogram variant in which it is possible to apply quasi-stationarity in some subsets of the time-frequency plane when transmission effects are not too prominent. (Andringa, 2002). In general the problems associated with the safe application of quasi-stationarity, and therefore of signal component estimation, are not yet solved. Note that work on adaptive sparse coding (Smith & Lewicki, 2006) or sinusoidal modeling approaches (Daudet, 2006, Davies & Daudet, 2006) avoid frame-blocking altogether. But likewise these approaches cannot guarantee the formation of representation consisting of single source evidence.

Tones, Pulses, and Noises

The two-dimensional cochleogram can be augmented with periodicity information to yield a three dimensional *Time Normalized Correlogram* (Andringa, 2002). The Time Normalized Correlogram reflects always a superposition of two qualitatively different stable patterns: one associated with the *aperiodic excitation* of the corresponding BM region, the other associated with a *periodic excitation*. Furthermore the aperiodic excitation has two variants, one associated with a pulse-like excitation and one associated with broadband noise stimulation. This results in three qualitatively different excitations of the basilar membrane: tonal, pulse-like, and noise-like.

Interestingly these three patterns reflect the different sound production mechanisms, and as such the source production physics described by (Gaver, 1993). Recently it was shown (Gygi & Watson, 2007) that environmental sounds appear to be grouped perceptually in harmonic sounds with predominantly periodic contributions, impact sounds with predominantly pulse-like contributions, and "continuous sounds" with prominent aperiodic contributions. This entails that signal-processing, source physics, and perceptual experiments all suggest that tone, pulses, and noises should be treated as qualitatively different types of signals that are represented by different types of signal components.

Initial experiments to measure the fractions of tonal, pulse-like and aperiodic contributions indicate that the distribution of these contributions correlates with the perceptual results of Gygi

(Andringa, 2008). Additionally, the perceptual evaluation of a highly reverberant bouncing ball, which was strongly aperiodic in terms of signal content, was scored as a typical impact sound by the listeners. This perceptual insensitivity to transmission effects suggests, again, that listeners use sound production physics to represent the sound and ignore, in the source evaluation at least, much of the signal if it is the result of transmission effects.

CONCLUSION

This chapter argued that machine audition, as the rest of intelligent systems, is currently trapped in application domains in which a human user must ensure that the system is exposed to input it can process correctly. Apparently something essential is missing in modern systems, which is provided by the human user. Since a human listener is a multi-purpose system, it is able to assign its computational resources very flexibly to the ever-changing demands of real world environments. By processing all input only up to the point of estimated irrelevance, human audition processes only a relevant subset of all input in detail. This efficiency is the result of interplaying bottom-up hearing and top-down listening. The hearing stage keeps track of the general properties of the physical environment. The listening stage leads to a knowledge guided strategic analysis of subsets of the signal. The strategic analysis requires a signal representation that is closely related to the physical limitations imposed on the signal by sources and environments that lead to the demand to interpret the signal as a physically realizable configuration. This demand poses restrictions on the form of signal processing that are not met by most modern signal processing approaches, but that seem to be realized in the human auditory system.

THE FUTURE OF MACHINE AUDITION

The moment machine audition is able to make the transition from simplified tasks in controlled domains to uncontrolled real-world input (Andringa & Niessen, 2006, Krijnders, Niessen & Andringa, 2010), it extends its application scope considerably. First, it will no longer be necessary to develop a large number of different single-purpose applications. A single, but flexible, multi-purpose system will suffice. This system will, like the natural auditory system, not be able to analyze every sound in detail, but if it happens to have the knowledge required for the task, it can produce a reliable and well-founded result, which it can justify to the user. While the recognition of unconstrained sonic environments might be well outside our current reach, due to the huge amount of knowledge required, it will be possible to implement all kinds of expert knowledge domains into, for example, a mobile phone. Devices like this can be used as a singing coach to give feedback on pitch and singing style. The next day they might download the knowledge required to analyze irregular sounds of a car from the web. In the evening they can be used as a smart baby phone, and during the night they function as a device that detects and diagnoses apnea (prolonged suspension of breathing during sleep with serious medical consequences).

A second range of novel applications is related to environmental monitoring and especially noise monitoring. Current noise regulations rely exclusively on noise levels, which have only a strong correlation with annoyance above 70 dB(A). Listeners are exquisitely sensitive to the source composition of signals, while being bad dB-meters. An approach that mimics human perception well can be used to detect the sounds that attract attention and as such demand processing time. Typically, sounds that attract attention but do not contribute to the tasks and goals of the listener are not appreciated, because they steal time from higher valued activities. Systems that

are able to measure, and even better, are able to predict, level independent noise disturbance can be used in urban planning procedures to design and monitor regions where the combination of sound and human activities are least likely to disturb.

Both examples share a vision of the ubiquitous application of the next generation of machine audition and are by no means exhaustive. The imminent technological transition from controlled domains to uncontrolled domains is a major technological breakthrough that is likely to lead to applications and new generations of technology that cannot yet be foreseen. This makes the future of machine audition seem bright indeed.

ACKNOWLEDGMENT

I thank Maria Niessen, Dirkjan Krijnders, Ronald van Elburg for helpful comments. I thank INCAS3 for supporting this work.

REFERENCES

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464. doi:10.1016/j.tics.2004.08.011

Allen, J. (1994). How do humans process and recognize speech? *Speech and Audio Processing*, *2*(4), 567–577. doi:10.1109/89.326615

Alvarez, G., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392–398. doi:10.1111/j.1467-9280.2008.02098.x

Andringa, T. C. (2002). Continuity Preserving Signal Processing. *Dissertations University of Groningen,* http://dissertations.ub.rug.nl/FILES/faculties/science/2001/t.c.andringa/thesis.pdf.

Andringa, T. C. (2008). The texture of natural sounds. *Proceedings of Acoustics'08, Paris* (pp. 3141-3146).

Andringa, T. C., & Niessen, M. E. (2006). *Real World Sound Recognition, a Recipe. Learning the Semantics of Audio Signals*. Athens, Greece: LSAS.

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science*, *12*(2), 157–162. doi:10.1111/1467-9280.00327

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Bronkhorst, A. (2000). The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acoustica – acta acoustica, 86*, 117-128.

Cartwright-Finch, U., & Lavie, N. (2007). The role of perceptual load in inattentional blindness. *Cognition*, *102*, 321–340. doi:10.1016/j.cognition.2006.01.002

Cherry, E. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. doi:10.1121/1.1907229

Cowling, M., & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, *24*, 2895–2907. doi:10.1016/S0167-8655(03)00147-8

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(4), 643–656. doi:10.1037/0096-1523.30.4.643

Daudet, L. (2006). A review on techniques for the extraction of transients in musical signals (LNCS). In Kronland-Martinet, R., Voinier, T., & Ystad, S. (Eds.), *Springer-Verlag Berlin Heidelberg, Jan 2006*.

Davies, M., & Daudet, L. (2006). Sparse audio representations using the mclt. *Signal Processing*, *86*(3), 457–470. doi:10.1016/j.sigpro.2005.05.024

Dehaene, S., Changeux, J., Naccache, L., & Sackur, J. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. doi:10.1016/j.tics.2006.03.007

Gaver, W. (1993). What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, *5*(1), 1–29. doi:10.1207/s15326969eco0501_1

Gopher and Iani. (2002). Attention. Encyclopedia of Cognitive Science L. Nadel (Ed). (pp. 220-226).

Gygi, B., Kidd, G., & Watson, C. (2007). *Similarity and categorization of environmental sounds*. Perception & Psychophysics.

Harding, S., Cooke, M., & Konig, P. (2007). Auditory gist perception: an alternative to attentional selection of auditory streams? *In Lecture Notes in Computer Science: Attention in Cognitive Systems.* [Springer-Verlag Berlin Heidelberg.]. *Theories and Systems from an Interdisciplinary Viewpoint*, *4840*, 399–416. doi:10.1007/978-3-540-77343-6_26

Haykin, S., & Chen, Z. (2005). The Cocktail Party Problem. *Neural Computation*, *17*, 1875–1902. doi:10.1162/0899766054322964

Hu, G., & Wang, D. (2006). An auditory scene analysis approach to monaural speech segregation. *Topics in acoustic echo and noise control* (pp. 485-515).

Irino, T., & Patterson, R. D. (1997). A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America*, *101*(1), 412–419. doi:10.1121/1.417975

Knudsen, E. (2007). Fundamental Components of Attention. *Annual Review of Neuroscience*, *30*, 57–78. doi:10.1146/annurev.neuro.30.051606.094256

Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends in Cognitive Sciences*, *11*(1), 16–22. doi:10.1016/j.tics.2006.10.012

Krijnders, J.D., Niessen, M.E. & Andringa, T.C. (2010). Sound event identification through expectancy-based evaluation of signal-driven hypotheses. Accepted for publication in Pattern Recognition Letters.

Mack, A. (2003). Inattentional blindness: Looking without seeing. *Current Directions in Psychological Science*, *12*(5), 180–184. doi:10.1111/1467-8721.01256

Marr, D. (1982). *Vision*. New York: Henry Holt and Co., Inc.

Nábělek, A. K., & Robinson, P. K. (1982). Monaural and binaural speech perception in reverberation for listeners of various ages. *The Journal of the Acoustical Society of America*, *71*(5), 1242–1248. doi:10.1121/1.387773

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383. doi:10.1016/0010-0285(77)90012-3

O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*, 2965–2979. doi:10.1016/j.patcog.2008.05.008

Oliva, A. (2005). *Gist of a scene* (pp. 251–256). Neurobiology of Attention.

Potter, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory*, *2*(5), 509–522. doi:10.1037/0278-7393.2.5.509

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. doi:10.1016/0010-0285(76)90013-X

Rosenthal, D. F., & Okuno, H. G. (1998). *Computational Auditory Scene Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Schyns, P., & Oliva, A. (1994). Evidence for Time-and Spatial-Scale-Dependent Scene Recognition. *Psychological Science*, *5*(4), 195–200. doi:10.1111/j.1467-9280.1994.tb00500.x

Smith, E., & Lewicki, M. (2006). Efficient auditory coding. *Nature*, *439*(23), 978–982. doi:10.1038/nature04485

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network (Bristol, England)*, *14*, 391–412. doi:10.1088/0954-898X/14/3/302

Van Hengel, P. W. J., & Andringa, T. C. (2007). Verbal aggression detection in complex social environments. In *IEEE Conference on Advanced Video and Signal Based Surveillance* (pp. 15-20).

Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, Algorithms, and Applications*. New York: IEEE Press/Wiley-Interscience.

Whalen, P. J., Rauch, S. L., & Etcoff, N. L. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience*, *18*(1), 411–418.

Yeh, W., & Barsalou, L. (2006). The situated nature of concepts. *The American Journal of Psychology*, *119*(3), 349–384. doi:10.2307/20445349

Young, S., & Bloothooft, G. (Eds.). (1997). *Corpus-Based Methods in Language and Speech Processing. Text, Speech and Language Technology*. Dordrecht, Netherlands: Kluwer.

ADDITIONAL READING

Alain, C., & Izenberg, A. (2003). Effects of attentional load on auditory scene analysis. *Journal of Cognitive Neuroscience*, *15*(7), 1063–1073. doi:10.1162/089892903770007443

Aucouturier, J., Defreville, B., & Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic msic. *The Journal of the Acoustical Society of America*, *122*(2), 881–891. doi:10.1121/1.2750160

Chu, S., Narayanan, S., & Kuo, C. (2008). *Environmental sound recognition using MP-based features*. Acoustics.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *34*, 267–285.

Guastavino, C. (2007). Categorization of environmental sounds. *Canadian Journal of Experimental Psychology*, *61*(1), 54–65. doi:10.1037/cjep2007006

Guastavino, C., Katz, B., Polack, J., Levitin, D., & Dubois, D. (2005). Ecological validity of soundscape reproduction. *Acta Acustica united with Acustica, 91* (2), 333-341.

Nahum, M., Nelken, I., & Ahissar, M. (2008). Low-level information and high-level perception: The case of speech in noise. *PLoS Biology*, *6*(5), 978–991. doi:10.1371/journal.pbio.0060126

Niessen, M. E., van Maanen, L., & Andringa, T. C. (2009). Disambiguating Sounds through Context. *International Journal of Semantic Computing*, *2*(3), 327–341. doi:10.1142/S1793351X08000506

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. doi:10.1016/S0079-6123(06)55002-2

Recanzone, G., & Sutter, M. (2008). The Biological Basis of Audition. *Annual Review of Psychology*, *56*, 119–142. doi:10.1146/annurev.psych.59.103006.093544

Shinn-Cunningham, B. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. doi:10.1016/j.tics.2008.02.003

KEY TERMS AND DEFINITIONS

Attention: the mental processes that allocates algorithmic and knowledge resources perception or other mental tasks.

Audition: the capacity for, or act of sound-based processing in which the existence of something or someone becomes mentally available (in the case of hearing), this availability can be used in a reasoning process to discover the consequences of what has been perceived (in the case of listening).

Bottom-Up Attention: a subprocess of attention that allows unexpected and well-trained stimuli to interrupt ongoing mental tasks in favor of a more detailed analysis.

Every Day Listening: a form of listening aimed at discovering the events, objects, and processes that caused the sound

Gist: the representation of a scene and its possible meaning that results from even a short stimulation, the gist can be refined to a reliable interpretation with subsequent analysis.

Hearing: the bottom-up, gist activation stage of audition aimed at discovering the existence of sound sources and their possible behavioral significance

Listening: the top-down, task and knowledge specific detailed analysis of sound and sound sources.

Meaning of Something for Someone: the difference in behavioral options for someone with and without the inclusion of something

Musical Listening: a form of listening aimed at listening to the properties of the sound as a physical signal

Physical Realizability: a property indicating whether or not an interpretation corresponds to a physically allowed configuration

Quasi-Stationarity: the assumption, valid for a single source, that the development of a source can be described as a set of discrete steps

Reverberation Radius: the distance around a source where the energy of the direct sound is equal to the energy of the summed indirect reflections

Sound: vibration that travel through the air or other media and can be heard when they reach a person's or animal's ear.

A/The Sound: audible sound stemming from someone or something

Top-Down Attention: a subprocess of attention that uses knowledge-based expectations to capture subsets of the input and connect these to an interpretation