19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

PREDICTING LISTENERS' REPORTS OF ENVIRONMENTAL SOUNDS

PACS: 4343.66.Lj

Andringa, Tjeerd; Grootel, Maarten. van
Auditory Cognition, Department of Artificial Intelligence, PO box 72, 9700 AB, Groningen, The Netherlands; tjeerd@ai.rug.nl, mgrootel@ai.rug.nl

ABSTRACT
Spontaneous verbal descriptions of environmental sounds lead to a description of the contributing sound sources and the environments in which they occur. This is a form of perception that relies crucially on the rich structure of sounds, because only rich sounds can convey detailed information about individual sources and the transmission environment. This paper uses a semantic network with connection strengths derived from listener reports to represent the content of auditory scenes. The activity of the semantic network is based on a number of source specific cues for sounds such as birds, vehicles, speech, and footsteps. These cues are not based on spectral envelope and level, but on patterns in tones, pulses and noises that capture source specific structures. Generally the system performed in a similar way as a human listener in terms of concepts activated (or named) and the choice of the acoustic environment. The robustness and performance suggest the combination of a semantic network and source specific cues can be used to design systems for sound-based ambient awareness.

INTRODUCTION
Listeners can formulate detailed reports of what they have heard[1], and they can reason about sound sources and sound producing events. Furthermore, listeners are also able to hypothesize the kind of environment that is likely to generate the sound and they can predict which events are likely to occur in the hypothesized environment. This paper investigates the requirements of an artificial system that performs these functions and that can be used as a model of Auditory Cognition.

Gaver[2] makes a distinction between musical listening and everyday listening. *Musical listening* describes sounds in terms of superficial signal properties such as pitch, corresponding to periodicity; timbre, corresponding to spectral content; and loudness related to signal energy. Musical listening is a form of listening that requires highly attentive and often trained listeners. This approach allows us to study the surface structure of sounds and its effects on the first stages of auditory processing. Much of our understanding of the auditory system stems from psychophysical experiments using musical listening with highly simplified stimuli (typically tones, pulses, and noises). However, the abstract, simplified and controlled nature of these sounds has stripped the stimuli of the richness of natural sounds. This prevents an interpretation of these sounds in terms of sound sources. In contrast, *everyday listening* relies crucially on the richness of sounds, because only rich sounds can convey detailed information about the source and the transmission environment. How to describe this richness is an open scientific question.

This paper studies everyday listening, but it extends Gaver's definition of everyday listening from mere sound event perception to include the cognitive processing required to map sounds onto a verbal description. The paper proposes a few design constraints for everyday listening and proposes some novel signal descriptors and sensory cues that might activate the concepts listeners use in their reports. Eventually this must lead to systems that are able to generate online hypotheses of the most plausible explanation of the events that caused the sounds. Systems like this can be used for automatic annotation of audio(visual) material, or they can be applied to provide camera observers or hard-hearing people with sound based situational awareness.

THEORETICAL APPROACH

Traditionally perception is viewed as a process to find the causes of input that is degraded by transmission between source and reception[3]. In this view, the task of perception is to correct the input to allow the brain to make an educated guess about what caused the input. Ecological perception includes knowledge about the perceiver's environment and his activities in the perception process. This knowledge allows the perceiver to predict specific degradations of the input. As long as the expectations develop in real-time with the changes in the environment, a relatively simple comparison between expected and received input suffices for reliable perception. Only when our knowledge-state is out of sync with the actual state of the environment, we have to fall back to perception in the traditional sense. The ecological approach to perception requires a direct influence of our knowledge during the perception process, in combination with a close correspondence between our internal environmental model and the actual environment. The closeness of this correspondence is the natural quality measure models of everyday listening, and as such it is used as objective in this paper.

When people produce verbal reports of auditory scenes, they associate audio cues with concepts. These concepts are interrelated like words in a dictionary are interrelated: each word is defined in terms of other words. Relational networks like this are called semantic networks[4] and are well-studied devices for long and short term memory. A semantic network represents relations between concepts. For example the concept of a busy city street is strongly correlated with the presence of cars and busses. It is also correlated with the presence of people and activities like walking and talking. The reverse is also true: walking and talking people, in combination with passing cars and busses is a strong indication that the current environment is a busy city street. While the combination of chirping birds, an occasional barking dog, cyclists, a few people, and distant cars are more indicative of a city park. Consequently, if a listener hears this combination of sound sources he/she is likely to hypothesize that the most likely recording site is a park and will not be surprised to hear a jogger passing. A screeching door however, is not predicted by this pattern of sources and warrants a reevaluation of the park as the most likely environment. The relative connection strengths in a semantic network represent its long term memory. The spreading activation in the network due to the activation of one or more nodes represents its short term memory state, which is indicative of the concepts used in the verbal reports of the auditory scene.

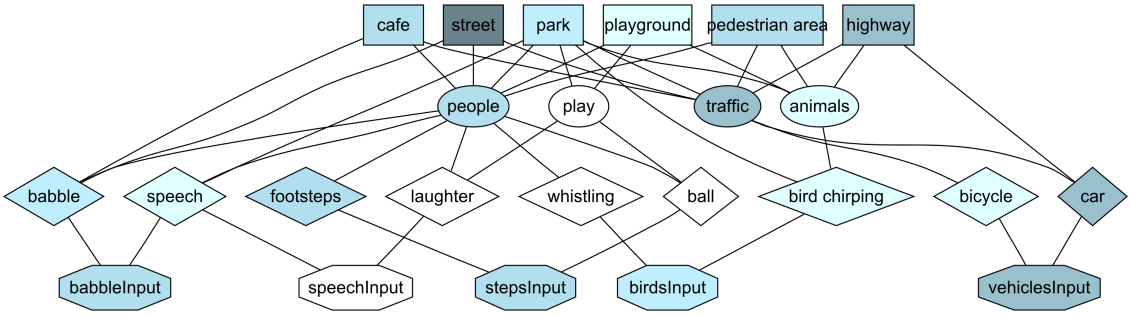

Figure 1. - A possible activation pattern due to of cues derived from a city street. The strongest cue is indicative of vehicles, but additional cues suggesting babble-sounds, footsteps, and birds are also active, which suggests the presence of people in combination with traffic. A busy city street is the most likely candidate.

If the simple semantic network of figure 1 is a sufficiently good approximation of the knowledge representation of a listener, it can be used to predict which concepts will be activated when the sound producing events at the second level are activated because cues consistent with for example babble sounds are present in the signal. This article relies on Continuity Preserving Signal Processing[5,6] (CPSP) to derive cues indicative of some of the events in the environment.

METHODS

To construct a semantic network like in figure 1, the network of descriptional concepts listeners use while they describe auditory scenes must be collected.

Sound database construction

The sounds for this experiment were recorded at several locations in Groningen (The Netherlands) on different days of the weeks, and at different times of the day. The locations used are described in table 1, showing the type of the location and the most dominant sounds. In total 8 different locations of 5 different types were visited. Recordings were made with a Sony HI-MD recorder and Core Sound low cost binaural microphones near each ear. From each file one or more fragments of 10 to 15 seconds were selected which resulted in 24 sound clips in total. Clips from the same location were chosen to vary as much as possible in content.

Table 1. - The eight recording locations and the sounds that dominate locally

Location	Type	Sounds
Park near main square	Park	Birds, conversations, laughter, distant traffic
Grote Markt	Café at busy square	Loud babble noise, motorized traffic and bicycles
Herestraat	Shopping area	Footsteps, babble noise
Noorderplantsoen (1)	Park with playground	Runners, conversations, children, bouncing ball, birds, distant traffic
Noorderplantsoen (2)	Park	Birds, distant traffic, little speech
Poeleplein	Quiet city square	Soft babble noise, footsteps, birds
Rijksweg 34	Busy highway	Heavy traffic, occasional birds
UMCG Hospital	Street	Traffic, walking people, light babble noise

Listener evaluations

In a pilotexperiment 10 of the sound clips in random order were presented via a website to 18 subjects (7 male, 11 female, average age 24,1 ± 4,27 year). The subjects were instructed to use reasonable quality headphones and were given a free association task in which they had to describe the 10 audio fragments while listenening to the sounds as often as they wanted. The subjects show a wide range of responses regarding the level of precision. Nevertheless, almost no-one used terms from the highest level depicting the environments, and most classification were on the level of sources except for the more general classes ‘people’, ‘bird’, and ‘traffic’ which were mentioned very often. On average subjects mentioned 4 categories per sound clip. The subjects reported the type of environment predominantly when they were unsure about its type. The most salient source descriptors mentioned by the subjects were ‘babble noise’ and ‘talking’, ‘speech’ and ‘conversations’, ‘bicycles’ and ‘footsteps’.

Semantic network

To model the relations and interactions between the layers a semantic network with a strictly bottom-up, path constrained spreading activation[7] is used. This network consists of 4 layers as shown in figure 1, the lowest layer is the input layer, the second layer corresponds to sound sources, the third layer reflects more abstract classes and the highest level reflects environments. Activation of a node *j* is calculated by summing all the activations of the connecting nodes *I*, weighted by their connection strength, as seen in the following formula.

$$A_j = \sum_i A_i w_{ji} \qquad \text{(Eq. 1)}$$

The input nodes at the lowest level were activated by the file averages of the sensory cues outlined below. These were scaled to the interval (0,1) with a sigmoid function dependent of the mean and the range of the average input of that category. The weights w_{ji} were obtained from both the listener evaluations as well as knowledge about the content of the audio clips. The connection weights between the input and the sound source level were determined by “guesstimating” how important each cue was for each class. For example, the babble cue is very indicative for speech murmur, but also indicative for more discernable speech. This was a very rough estimation that can however be automated with a reliable ground truth.

The connections between the sources and the classes and environments were determined from the subjects report as the fraction of the sound source being named as a member of the more general class. The weights from class to environment were determined as the fraction the classes occurred for each environment. The environment node that receives the highest activation from the average cue pattern is the most probable category for the sound clip.

Signal analysis

CPSP is a framework for auditory scene analysis designed to track the physical development of a sound source through the identification of signal components as the smallest coherent units of physical information. Signal components are defined as physically coherent regions of the time-frequency plane delimited by qualitative changes, such as on- and offsets or discrete steps in frequency. It is often, and even usually, possible to form signal component patterns that have a very high probability to represent physically coherent information of a single source or process. This is especially true for pulse-like and sinusoidal components (such as individual harmonics) for which reliable estimation techniques have been developed[5]. An analysis of the existence and properties of both the signal components and rest of the time-frequency regions, that is the richness of the signal, might yield patterns characteristic of specific sound events recorded in realistic environments.

CPSP uses a variant of the classical spectrogram, called a cochleogram, which is continuous in time and frequency through the use of a realistic transmission-line model of the human cochlea[8]. It is possible to perform figure-ground separation with a dynamical adapting background model that assigns evidence of pulses and sinusoidal components to the foreground and noisy components to the background[9]. This method will be described here in broad lines only. The method uses cochleogram energy $E(x,t)$ to denote energy stemming from cochlea segment x at time t to compute the background energy $E_{bg}(x,t)$ according to:

$$\tau \frac{dE_{bg}(x,t)}{dt} + E_{bg}(x,t) = E'(x,t) \qquad \text{(Eq. 2)}$$

where τ of the order of 0.1 s or larger, and $E'(x,t)$ is defined as:

$$\begin{aligned} E'(x,t) &= E(x,t), \quad \text{if} \quad E(x,t) - E_{bg}(x,t) \le \mathrm{C}(x,t) \\ &E_{bg}(x,t), \ \text{if} \quad E(x,t) - E_{bg}(x,t) > \mathrm{C}(x,t) \end{aligned} \qquad \text{(Eq. 3)}$$

The function $C(x,t)$ determines the amount of deviation from a generic noise. The result is a background model that is predominantly based on relatively slowly changing noisy contributions. When the original energy $E(x,t)$ deviates sufficiently from the background model $E_{bg}(x,t)$ it can be assigned to a foreground model $E_{fg}(x,t)$ in which it is possible to track evidence of pulses and tones.

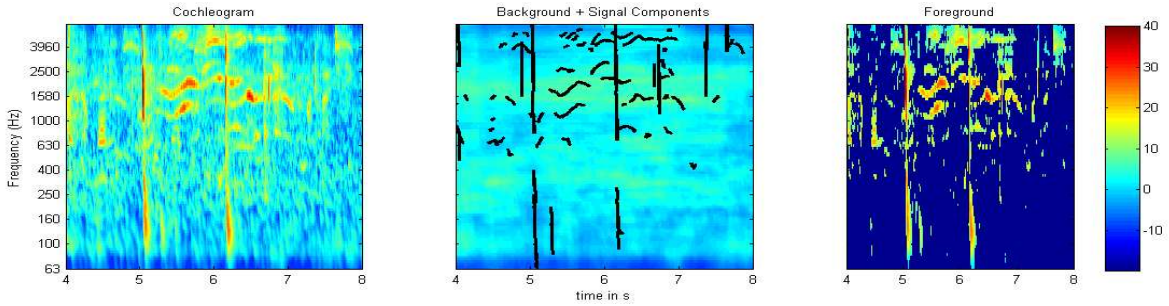

Figure 2. - Separation into fore- and background and the resulting signal components. The left panel shows a 4 second cochleogram (60 dB dynamic range) with two balls bouncing and a yelling child. The middle panel shows the background model computed according to Eq. 1 and 2 with signal components estimated from the foreground signal shown in the right panel.

It is possible to use the background model $E_{bg}(x,t)$ as input for one or more additional foreground separation passes by increasing τ and adapting $C(x,t)$ to match the criteria for acceptance in the next level background. For this paper a second level background model has been computed with a time-constant of 4 seconds that separates relatively fast events like a passing car from a more stationary background.

Sensory cues

A sensory cue is a statistic or signal that can be extracted from sensory input and which indicates the value of some property or state of the world. The cues are used to activate the sound source level in Figure 1. Note that the dynamics of the semantic network, i.e. the interaction between signal and knowledge, must eventually determine which sound sources are present. The input of the lower level must therefore not be the result of a full-fledged sound source detector, but a set of simple cues that are indicative, but generally inconclusive, of the presence of sound producing events. The computed cues are based on a wide range of signal properties that can be used for many different sound classes. Specific combinations of these signal properties form the cues for vehicles, babble sound, speech, footsteps, and birds. Note that none of the usual spectral-envelope (e.g. MFCC, LPC), or sound-level features is used.

The cue for motorized vehicles is a superposition of low frequency tonal components; mean frequency below 200 Hz and at least 1 s long in combination with the content of the second level foreground, which predominantly contains evidence with a temporal development between 0.1 and about 4 seconds. Vehicles at a distance tend to be characterized by the low frequency tonal components. Vehicles that pass at a short distance are characterized by a prominent noise-cloud visible in the level two foreground. The motorized vehicle cue is a scaled superposition of the two types of contributions.

The cue for babble is derived from the fact that babble noise consists of the superposition of a large number of speakers of which the strongest formants are most likely to dominate a coherent region of the time-frequency plane. Compared to white noise, these regions are more periodic and tend to show hints of tonal components. It is possible to compute how well the fine structure matches the shape of an ideal sinusoid around this position. A comparison between white nose and babble noise shows that babble noise has indeed more tonal content than white noise. The cue for babble-noise is based on the difference with the expectation for white noise; but only for those regions not assigned to the second level foreground which is indicative for the vehicle.

The cue for discernable speech is based on the number of tonal signal components between 400 and 2500 Hz, which last between 30 ms and 0.5 s. This set is further constrained by demands that the average rate of change, normalized to one second, is more than 1 and less than 6 times the instantaneous frequency of the tonal component. This accounts for the characteristic variation and covers about 95% of most spontaneous speech. A similar, but less important restriction is placed on the average energy fluctuation. The speech cue is based on the sum of the local foreground to background ratio for those intervals where two or more tonal components co-occur.

The cue for footsteps aims at the detection of repeated pulses and is based on the sum of the foreground to background ratio assigned to pulse-like signal components between 200 and 4000 Hz. This time signal is Fourier transformed and squared to form a spectrum. The cue for steps is based on a comparison of a smoothed spectrum with the original spectrum for the range of 1 to 3 Hz. Footsteps are detected whenever the unsmoothed spectrum contains a frequency sufficiently far (in this case a factor 2) above the smoothed spectrum. The cue-value, for the clip, is the fraction above this threshold. Informal listening suggests that this algorithm can detect barely discernable footsteps.

The cue for birds is a 0.5 s moving average of the number of tonal signal components between 3.5 and 6 kHz, which last between 30 ms and 1 s, and which have energies at least 6 dB above the background model.

RESULTS AND DISCUSSION

Of the 24 sound clips, only 6 were incorrectly classified compared with the ground truth based on the actual recording site. Three of these clips were also incorrectly classified by many subjects, indicating that some environments are not trivial to classify. The network had a tendency to activate ‘people’, which led to a bias for ‘café’ and ‘park’. The network is robust against minor changes in the connection strengths, as small changes, in the order of 10-20%, didn’t influence the classification much. With considerably bigger changes the number of correct responses remained often constant. A similar insensitivity to the parameters that determine the

sensory cue values was also observed. This inherent insensitivity is important since it suggests the stability and robustness required of a system that can deal with unconstrained environmental input. The suboptimal classifications corresponded usually to confusions based on the true signal content. For example, one rather crowded park sample was classified as a pedestrian area. Only one classification confused a highway with some bird song with a park. This was the consequence of the startup-time required for the second level foreground (in de order of several times τ= 4 s). Furthermore the activation of the sound sources and the more general classes appeared to be consistent with the verbal reports.

Simple machine learning techniques applied to the weights in the semantic network are likely to lead to a perfect score, but since the parameter space is large enough to fit a very complex classification task; this would not be convincing. The fact that the system has not been optimized in any way is more suggestive of its future potential. The cues are very simple and not even conclusive for the target classes. But because they are not based on traditional level and spectral shape cues used in sound and speech recognition[10] they provide evidence that sound source recognition can be performed with other cues than tradition suggests. The use of these cues, and especially their environment and source dependence will be the focus of further research.

Although this is a knowledge based approach (as opposed to standard statics-based approaches such as HMM's and neural networks) it is in principle possible to automate much of the development of a system for the description of auditory scenes. More cues can be formulated from first principles or detected from structures in annotated data. Furthermore each cue can be improved with machine learning techniques. The weights of the semantic network are based on a combination of a ground truth (i.e. the knowledge of where the recording is made) and counts of the number of descriptive words the subjects used per situation. This can also be optimized with an improved experimental design and with machine learning techniques on annotated data.

Further research must investigate how to reach and maintain a high level of correspondence between the properties of the actual and the predicted input. The current work does not use these expectations, but future developments in the recognition and the description of environments may. This might allow the development of artificial perceptive systems that not merely react to a signal, but are actively prepared to process and explain its content in terms of physical causes.

References:

1. Dubois, D., Guastavino, C. & Raimbault, M. A cognitive approach to soundscape: Using verbal data to access everyday life auditory categories. *Acta Acustica united with Acustica* **92**, 865-874 (2006).
2. Gaver, W. W. What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception. *Ecological Psychology* **5**, 1-29 (1993).
3. Turvey, M. T. & Carello, C. Cognition: The view from ecological realism. *Cognition* **10**, 313-321 (1981).
4. Steyvers, M. & Tenenbaum, J. B. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science: A Multidisciplinary Journal* **29**, 41-78 (2005).
5. Andringa, T. C. Continuity Preserving Signal Processing. *PhD thesis, Univerity of Groningen* (2002).
6. Andringa, T. C., Hengel, P. W. J., Duifhuis, H. & et.al. Method and Apparatuses for Signal Processing. (International Patent Application WO 01/33547). 1999.
7. Crestani, F. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review* **11**, 453-482 (1997).
8. Duifhuis, H., Hoogstraten, H. W., van Netten, S. M., Diependaal, R. J. & Bialek, W. Modelling the cochlear partition with coupled Van der Pol oscilators. *Cochlear Mechanisms: Structure, Function and Models* 395-404 (1985).
9. Hengel, P. W. J. & Andringa, T. C. Verbal aggression detection in complex social environments. IEEE International Conference on Advanced Video and Signal based Surveillance . 2007.
10 Cowling, M. & Sitte, R. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters* **24**, 2895-2907 (2003).