Real-World Sound Recognition:
A Recipe

Tjeerd C. Andringa and Maria E. Niessen

Department of Artificial Intelligence, *University* of Groningen,
Gr. Kruisstr. 2/1, 9712 TS Groningen, The Netherlands
{T.Andringa,M.Niessen}@ai.rug.nl
http://www.ai.rug.nl/research/acg/

Abstract. This article addresses the problem of recognizing acoustic events present in unconstrained input. We propose a novel approach to the processing of audio data which combines bottom-up hypothesis generation with top-down expectations, which, unlike standard pattern recognition techniques, can ensure that the representation of the input sound is physically realizable. Our approach gradually enriches low-level signal descriptors, based on Continuity Preserving Signal Processing, with more and more abstract interpretations along a hierarchy of description levels. This process is guided by top-down knowledge which provides context and ensures an interpretation consistent with the knowledge of the system. This leads to a system that can detect and recognize specific events for which the evidence is present in unconstrained real-world sounds.

Key words: real-world sounds, sound recognition, event perception, ecological acoustics, multimedia access, Continuity Preserving Signal Processing, Computational Auditory Scene Analysis.

1 Introduction

Suppose you claim you have made a system that can automatically recognize the sounds of passing cars and planes, making coffee, and verbal aggression, while ignoring sounds like speech. People might react enthusiastically and ask you for a demonstration. How impressive would it be if you could start-up a program on your laptop, and the system recognizes the coffee machine the moment it starts percolating? You might then bring your laptop outside where it starts counting passing cars and it even detects the plane that takes off on a nearby airfield. In the meanwhile you tell your audience that the same system is currently alerting security personnel whenever verbal aggression occurs at public places like train stations and city centers. Your audience might be impressed, but will they be as impressed if you show a system that works with sound recorded in studio conditions but that does not work in the current environment? Or that only works in the absence of other sound sources, or only when the training data base is sufficiently similar to the current situation, which it, unfortunately, is not?

Apart from the coffee machine detector, detectors similar to those described above are actually deployed in the Netherlands by a spin-off of the University of Groningen called Sound Intelligence [16]. But although these systems reliably monitor the activities in a complex and uncontrollable acoustic environment, they require some optimization to their environment and cannot easily be extended to recognize and detect a much wider range of acoustic events or to reason about their meaning.

The one system that does have the ability to function reliably in all kinds of dynamic environments is, of course, the human and animal auditory system. The main attribute of the natural auditory system is that it deals with unconstrained input and helps us to understand the (often unexpected) events that occur in our environment. This entails that it can determine the cause of complex, uncontrollable, in part unknown, and variable input. In the sound domain we can call this *real-world sound recognition*.

A real-world sound recognition system does not put constraints on its input: it recognizes any sound source, in any environment, if and only if the source is present. As such it can be contrasted to popular sound recognition approaches, which function reliably only with input from a limited task domain in a specific acoustic environment. This leads to systems for specific tasks in specific environments. For example, a study by Defréville et al. describes the automatic recognition of urban sound sources from a database with real-world recordings [6]. However, the goals of our investigation do not allow for the implicit assumption made in Defréville's study, namely that at least one sound source on which the system is trained is present in each test sample.

Apart from being able to recognize unconstrained input, a real-world sound recognition system must also be able to explain the causes of sounds in terms of the activities in the environment. In the case of making coffee, the system must assign the sounds of filling a carafe, placing the carafe in the coffee machine, and a few minutes later the sound of hot water percolating through the machine, to the single activity of making coffee.

The systems of Sound Intelligence approach the ideal of real-world sound recognition by placing minimal constraints on the input. These systems rely on Continuity Preserving Signal Processing, a form of signal processing, developed in part by Sound Intelligence, which is designed to track the physical development of sound sources as faithfully as possible. We suspect that the disregard of a faithful rendering of physical information of, for example, Short Term Fourier Transform based methods, limits traditional sound processing systems to applications in which a priori knowledge about the input is required for reasonable recognition results.

The purpose of our research is the development of a sound processing system that can determine the cause of a sound from unconstrained real-world input in a way that is functionally similar to natural sound processing systems. This paper starts with the scientific origins of our proposal. From this we derive a research paradigm that describes the way we want to contribute to the further development of a comprehensive theory of sound processing on the one hand,

and the development of an informative and reliable sound recognition system on the other hand. From this we spell out a proposal for a suitable architecture for such a system from an example of a coffee making process. Finally we will discuss how our approach relates to other approaches.

2   Real-World Sound Recognition

2.1   Event Perception

Ecological psychoacoustics (see for example [10, 9, 18, 3]) provides part of the basis of our approach. Instead of the traditional focus on *musical* or *analytical listening*, in which sensations such as pitch and loudness are coupled to physical properties such as frequency and amplitude in controlled experiments (for a short historical overview, see Neuhoff [13]), ecological psychoacousticians investigate *everyday listening*, introduced by Gaver [9]. Everyday or descriptive listening refers to a description of the sounds 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. William Gaver concludes:

> "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." ([9], p. 8)

Gaver links source physics to event perception, and since natural environments are, of course, always physically realizable, we insist on physical realizability within our model. This entails that we aim to limit all signal interpretations to those which might actually describe a real-world situation and as such do not violate the physical laws that shape reality.

Physical realizability, as defined above, is not an issue in current low-level descriptors [5], which focus on mathematically convenient manipulations. The solution space associated with these descriptors may contain a majority of physically impossible, and therefore certainly incorrect, signal descriptions. In fact, given these descriptors, there is no guarantee that the best interpretation is physically possible (and therefore potentially correct). For example, current speech recognition systems are designed to function well under specific conditions (such as limited background noise and adapted to one speaker). However, when these systems are exposed to non-speech data (such as instrumental music) that happen to generate an interpretation as speech with a sufficiently high likelihood, the system will produce nonsense.

2.2 Continuity Preserving Signal Processing

We try to ensure physical realizability of the signal interpretation by applying Continuity Preserving Signal Processing (CPSP, [2, 1]). Compared to other signal processing approaches CPSP has not been developed from a viewpoint of mathematical elegance or numerical convenience. Instead it is a signal processing framework 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). Although CPSP is still in development, indications are that 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 [2]. For example, estimated sinusoidal signal components (like in voiced speech) signify the presence of a sound source which is able to produce signal components with a specific temporal development of energy and frequency content. This analysis excludes a multitude of other sound sources like doors, which are not able to produce signal components with these properties; the system is one step up toward a correct interpretation.

CPSP is a form of Computational Auditory Scene Analysis (CASA). Modern CASA approaches typically aim to identify a subset of the time-frequency plane, called a mask, where a certain target sound dominates [17, 4]. The advantage of such a mask is that it can be used to identify the evidence which should be presented to a subsequent recognition phase. One major disadvantage is that it requires a hard decision of what is target and what not before the signal is recognized as a certain instance of a target class. In contrast, CPSP does not aim to form masks, but aims to identify patterns of evidence of physical processes that can be attributed to specific events or activities. CPSP is based on the assumption that sound sources are characterized by their physical properties, which in turn determine how they distribute energy in the time-frequency plane $E(f, t)$. Furthermore it assumes that the spatio-temporal continuity of the basilar membrane (BM) in the mammalian cochlea, where position corresponds to frequency, is used by the auditory system to track the development of physically coherent signal components. Typical examples of signal components are pulses, clicks, and bursts, or sinusoids, narrowband noises, wavelets, and broadband noises. A sinusoid or chirp may exist for a long time, but is limited to a certain frequency range. In contrast pulses are short, but they span a wide range of frequencies. Noises are broad in frequency, persist for some time, and show a fine structure that does not repeat itself.

CPSP makes use of signal components because they have several characteristics which are useful for automatic sound recognition: The first characteristic is the low probability that the signal component consists of qualitatively different signal contributions (for instance periodic versus aperiodic), the second is the low probability that the signal component can be extended to include a

larger region of the time-frequency plane without violating the first condition, and the third is the low probability that the whole region stems from two or more uncorrelated processes. Together these properties help to ensure a safe and correct application of quasi-stationarity and as such a proper approximation of the corresponding physical development and the associated physical realizability.

Signal components can be estimated from a model of the mammalian cochlea [7]. This model can be interpreted as a bank of coupled (and therefore overlapping) bandpass-filters with a roughly logarithmic relation between place and center-frequency. Unlike the Short Term Fourier Transform, the cochlea has no preference for specific frequencies or intervals: A smooth development of a source will result in a smooth development on the cochlea. A quadratic, energy-domain rendering of the distribution of spectro-temporal energy, like a spectrogram, results from leaky integration of the squared BM excitation with a suitable time constant. This representation is called a cochleogram and it provides a rendering of the time-frequency plane $E(f, t)$ without biases toward special frequencies or intervals ([2], chapter 2). Although not a defining characteristic, signal components correspond often to the smallest perceptual units: it is usually possible to hear-out individual signal components, and in the case of complex patterns, such as speech, they stand out when they do not comply with the pattern.

2.3 An Example: Making Coffee

Figure 1 shows a number of cochleograms derived from a recording of the coffee making process in a cafeteria[1]. The upper three cochleograms correspond to filling the machine with water, positioning the carafe in the machine, and the percolation process, respectively. The lower row shows enlargement of the last two processes. The positioning of the carafe in the machine (a) results in a number of contacts between metal and glass. The resulting pulses have significant internal structure, do not repeat themselves, and reduce gradually in intensity. Some strongly dampened resonances occur at different frequencies, but do not show a discernible harmonic pattern. The whole pattern is consistent with hard (stiff) objects that hit each other a few times in a physical setting with strong damping. This description conveys evidence of physical processes which helps to reduce the number of possible, physically meaningful, interpretations of the signal energy. Note that signal components function as interpretation hypotheses for subsets of the signal. As hypotheses, they may vary in reliability. A context in which signal components predict each other, for instance through a repetition of similar components, enhances the reliability of interpretation hypotheses of individual signal components. In this case the presence of one pulse predicts the presence of other more or less similar pulses.

The lower right cochleogram (b) corresponds to a time 6 minutes after the start where the water in the coffee machine has heated to the boiling point and where it starts to percolate through the coffee. This phase is characterized

---

[1] The sounds can be found on our website, `http://www.ai.rug.nl/research/acg/research.html`
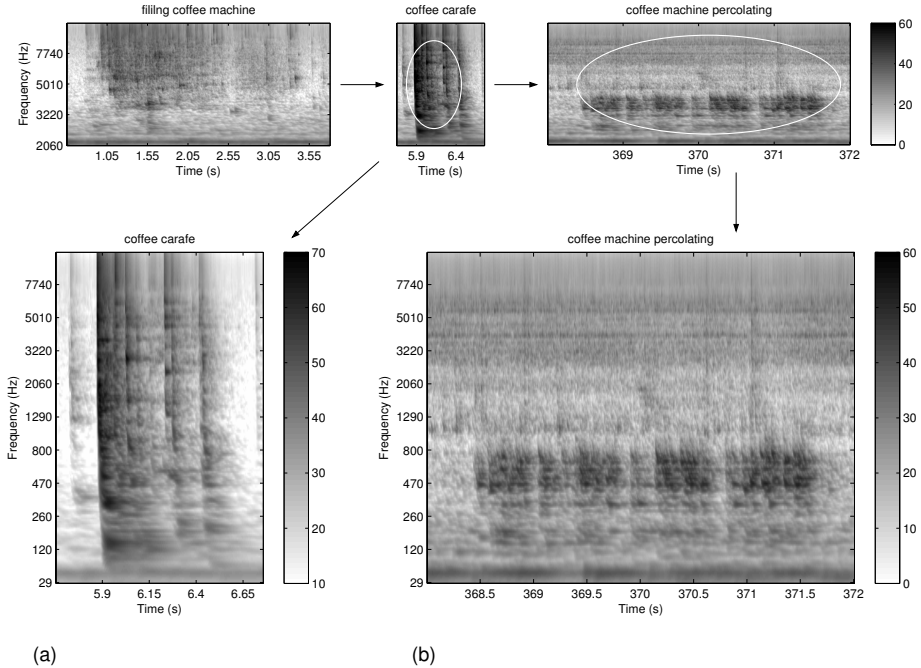
Fig. 1. Cochleograms of different sound events involved in making coffee. The top three cochleograms are of the sound of, from left to right, the filling of the water compartment of a coffee machine, the placing the coffee carafe, and the coffee machine percolating, respectively. These three cochleograms have the same energy range, as shown by the right color bar. The bottom two cochleograms are enlargements of the two top right cochleograms. (Note that the energy ranges and time scales of the bottom two cochleograms do not correspond to each other.)

by small droplets of boiling water and steam emerging from the heating system. This produces a broadband signal between 400 and 800 Hz with irregular amplitude modulation and, superimposed, many not-repeating wavelet-like components that probably correspond to individual drops of steam driven boiling water that emerge from the machine. Again it is possible to link signal details to a high level description of a complex event like percolating water.

3   Research Paradigm

We propose a novel approach to real-world sound recognition which combines bottom-up audio processing, based on CPSP, with top-down knowledge. The top-down knowledge provides context for the sound event and guides interpretations of the hypotheses. Succeeding description levels in a hierarchy gradually include more semantic content. Low levels in the hierarchy represent much of the details of the signal and are minimally interpreted. High levels represent minimal signal

detail, but correspond to a specific semantic interpretation of the input signal which is consistent with the systems knowledge of its environment and its current input. Before we present the model for real-world sound recognition, we first propose a paradigm, comprising of seven focus points, to guide the development of the model, inspired by considerations from the previous section:

Real-world sound recognition The model should recognize unconstrained input, which means there is no a priori knowledge about the input signal such as the environment it stems from. The system will use knowledge, but this knowledge is typically of the way we expect events to develop through time and frequency, used a posteriori to form hypotheses. Besides this the knowledge can also be about the context, for example an interpretation of the past which may need updating, or about the environment, which may have changed as well. Matching top-down knowledge with bottom-up processing helps to generate plausible interpretation hypotheses. However, top-down knowledge does not pose restrictions on the bottom-up processing and the generation of hypotheses; it will only influence activation and selection of hypotheses.

Domain independent Techniques that are task or environment specific are to be avoided, because every new task or environment requires the development of a new recognizer. In particular, optimizing on closed tasks or closed and constrained environments should be avoided. This excludes many standard pattern recognition techniques such as neural networks, and Hidden Markov Models (HMM's), based on Bayesian decision theory (see for example Knill and Young [11]). Note that specifying a general approach toward a specific task is allowed, but including a task specific solution in the general system is not.

Start from the natural exemplar Initially the model should remain close to approaches known to work in human and animal auditory systems. The problem in this respect is the multitude of fragmented and often incompatible knowledge about perception and cognition, which (by scientific necessity) is also based on closed domain research in the form of controlled experiments. Nevertheless domains such as psycholinguistics have reached many conclusions we might use as inspiration.

Physical optimality Because the auditory systems of different organisms are instances of a more general approach to sound recognition, we may try to generalize these instances toward physical optimality or physical convenience. We might for example ignore physiological non-linearities which may be necessary only to squeeze the signal into the limited dynamic range of the natural system. Physical optimality rules out the application of standard frame-based methods in which quasi-stationarity, with a period equaling the frame size, is applied on an yet unknown mixture of sound sources. Physically, the approximation of a sound source as a number of different discrete steps is only justified with a suitable (inertia dependent) time-constant. For closed domains with clean speech this might be guaranteed, but for open domains quasi-stationarity can only be applied on signal evidence which,

firstly, is likely to stem from a single process, and secondly, is known to develop slowly enough for a certain discrete resampling. Signal components comply with these requirements, while in general the result of frame blocking does not. ([2], chapter 1)

Physical realizability The model should at all times only consider those solutions which are physically plausible. This is important because we want to link the physics of the source of the sound to the signal, which only makes sense if we actually consider only plausible sound events as hypotheses. An extra advantage of this rule is the reduction of the solution space, since solutions which are not physically plausible will not be generated. Again this is different from standard pattern recognition techniques, which consider *mathematically* possible (that is, most probable) solutions.

Limited local complexity To ensure realizability, the different steps in the model should neither be too complex, nor too large. This can be ensured through a hierarchy of structure which is not imposed by the designer, but dedicated by the predictive structures in the environment.

Testing The model should not remain in the theoretical domain, but should be implemented and confronted with the complexities of unconstrained input. This also means the model should not be optimized for a target domain, but confronted with input from many other domains as well.

4 Model of Real-World Sound Recognition

Figure 2 captures the research paradigm of the previous section. The system processes low-level data to interpretation hypotheses, and it explicates semantically rich, top-down queries to specific signal component expectations. The top-down input results from task specific user queries, like "Alert me when the coffee is ready" or "How many times is the coffee machine used in the cafeteria every day?", and contextual knowledge, like the setting, for example a cafeteria, and the time of day, for example 8:30 am. However, the system recognizes sounds and updates its context model even without top-down input: The more the system learns about an environment (that is, the longer it is processing data in an environment), the better its context model is able to generate hypotheses about what to expect and the better and faster it will deal with bottom-up ambiguity. The bottom-up input is continuous sound input, which can be assigned to signal components with a delay of less then a few hundred milliseconds (in part dependent on the properties of the signal components). Subsequent levels generate interpretation hypotheses of each signal component combination which complies with predictions about the way relevant sound events, such as speech, combine signal components.[2] Several levels match top-down expectations with bottom-up hypotheses. This ensures consistency between the query and the sound (set

---

[2] The bottom-up hypotheses generation from signal component patterns will be hard-coded in the first implementations of the system, but eventually we want to use machine learning techniques so that the system learns to classify the signal component patterns. A similar approach is chosen for top-down queries and contextual
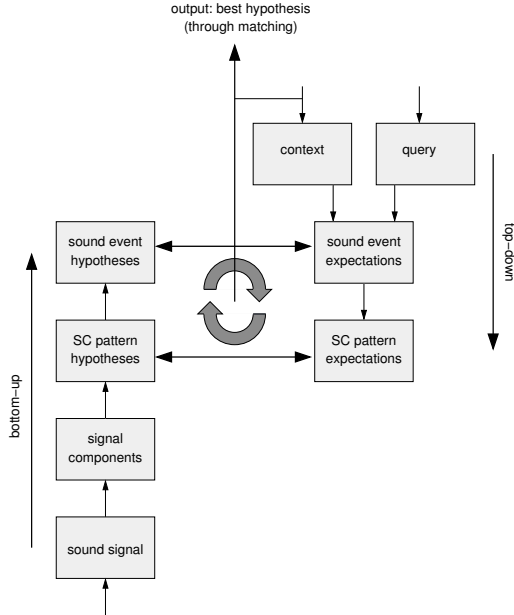
Fig. 2. The sound is analyzed in terms of signal components. Signal component patterns lead to sound event hypotheses. These hypotheses are continuously matched with top-down expectation based on the current knowledge state. The combination of bottom-up hypotheses and top-down expectations results in a best interpretation of the input sound.

of signal components) which is assigned to it. When top-down knowledge is sufficiently consistent with the structure of reality (that is, potentially correct) the resulting description of the sound is physically realizable.

Top-down combinations of query and context lead to expectations of which instances of sound events might occur and what the current target event is. Although important, the way queries lead to expectations is not yet addressed in this coffee-making detection system. The event expectations are translated into probability distributions of (properties of) the signal component combinations to expect. For example, knowledge about the type of the coffee machine, for instance filter or espresso, leads to a reduction of the signal components to expect. The derivation to top-down expectations through a top-down process ensures that the expectations are always consistent with our knowledge.

The continuous matching process between the bottom-up instances and the top-down probability distributions ensures that bottom-up hypotheses which are inconsistent with top-down expectations will be deactivated (or flagged irrelevant for the current task, in which case they might still influence the context model).

input; we first hard-code the knowledge between sound events and signal component patterns, but eventually we plan to use machine learning techniques here as well.
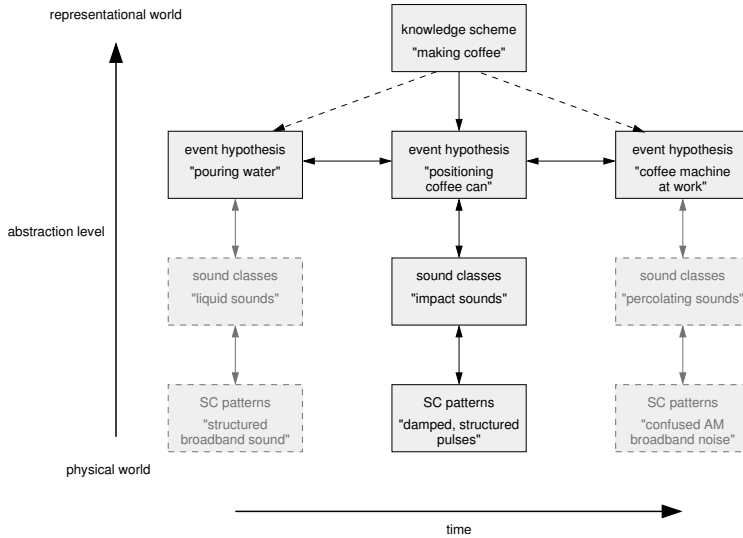
Fig. 3. Different stages in the process of recognizing coffee making. The top level is the highest abstraction level, representing most semantic content, but minimal signal detail. The bottom level is the lowest abstraction level, representing very little or no interpretation, but instead reflecting much of the details of the signal. From left to right the succeeding best hypotheses are shown, generated by the low-level signal analysis, and matched to high-level expectations, which follow from the knowledge scheme.

Vice versa, signal component combinations which are consistent with top-down expectations are given priority during bottom-up processing. The consistent hypothesis that fits the query and the context best is called the *best hypothesis*, and is selected as output of the system. The activation of active hypotheses changes with new information: Either new top-down knowledge or additional bottom-up input changes the best interpretation hypothesis whenever one hypothesis becomes stronger than the current best hypothesis. Similar models for hypothesis selection can be found in psycholinguistics [12, 14]. For example, the Shortlist model of Norris [14] generates an initial candidate-list on the basis of bottom-up acoustic input; a best candidate is selected after competition within this set. The list of candidates is updated whenever there is new information.

Let us turn back to the example, the process of making coffee. Figure 3 depicts the coffee making process at three levels: a sequence of activities, a sequence of sound events to expect, and a series of signal component combinations to expect. These three levels correspond to the top-down levels of the system. The lowest level description must be coupled to the signal components and the cochleogram (see figure 1) from which they are estimated. We focus on the second event hypothesis, the sound of a coffee carafe being placed on the hot plate (a).

Previously, the system found some broadband irregularly amplitude modulated signal components with local, not repeating, wavelet-like components su-

perimposed. Particular sound events with these features are liquid sounds [9]. However, the system did not yet have enough evidence to classify the liquid sound as an indication of the coffee machine being filled; the same evidence might also correspond to filling a glass. But since the signal component pattern does certainly not comply with speech, doors, moving chairs, or other events, the number of possible interpretations of this part of the acoustic energy is limited, and each interpretation corresponds, via the knowledge of the system, to expectations about future events. If there are more signal components detected consistent with the coffee making process, the hypothesis will be better supported. And if the hypothesis is better supported, it will be more successful in predicting subsequent events which are part of the process. In this case additional information can be estimated after 6 minutes where the water in the coffee machine has heated to the boiling point and starts percolating through the coffee (b).

5   Discussion

This article started with a description of a sound recognition system that works always and everywhere. In the previous sections we have sketched an architecture for real-world sound recognition which is capable of exactly that. Our general approach is reminiscent to the prediction-driven approach of Ellis [8]. However, the insistence on physical realizability is an essential novelty, associated with the use of CPSP.

The physical realizability in our approach ensures that the state of the context model in combination with the evidence is at least consistent with physical laws (insofar reflected by the sounds). In the example of the coffee making detection system only a minimal fraction of the sounds will match the knowledge of the system and most signal component patterns will be ignored from the moment they are flagged as inconsistent with the knowledge of the system. This makes the system both insensitive to irrelevant sounds as well as specific for the target event, which helps to ensure its independence of the acoustic environment. A richer system with more target events functions identically, but more patterns will match the demands of one or more of the target events. An efficient tree-like representation of sound source properties can be implemented to prevent a computational explosion.

Our approach is an intelligent agent approach [15] to CASA. The intelligent agent approach is a common within the fields of artificial intelligence and cognitive science, and generally involves the use of explicit knowledge (often at different levels of description). The intelligent agent based approach can be contrasted to the more traditional machine learning based approaches, such as reviewed in Cowling [5], and other CASA approaches [4, 8, 17]. An intelligent agent is typically assumed to exist in a complex situation in which it is exposed to an abundance of data, most of which irrelevant — unlike, for example, the database in Defréville's study [6], where most data is relevant. Some of the data may be informative in the sense that it changes the knowledge state of

the agent, which helps to improve the selection of actions. Each executed action changes the situation and ought to be in some way beneficial to the agent (for example because it helps the user). An intelligent agent requires an elaborate and up-to-date model of the current situation to determine the relevance of the input.

More traditional machine learning and CASA approaches avoid or trivialize the need for an up-to-date model of the current situation and the selection of relevant input by assuming that the input stems from an environment with particular acoustic properties. If the domain is chosen conveniently and if sufficient training data is available, sophisticated statistical pattern classification techniques (HMM, Neural Networks, etcetera) can deal with features like Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficients (LPC), which represent relevant and irrelevant signal energy equally well. These features actually make it more difficult to separate relevant from irrelevant input, but a proper domain choice helps to prevent (or reduce) the adverse effects of this blurring during classification. CASA approaches [4, 17] rely on the estimation of spectro-temporal masks that are intended to represent relevant sound sources, only after information in the masks is presented to similar statistical classification systems. These approaches require an evaluation of the input in terms of relevant/irrelevant *before* the signal is classified. This entails that class-specific knowledge cannot be used to optimize the selection of evidence. This limits these approaches to acoustic domains in which class-specific knowledge is not required to form masks that contain sufficiently reliable evidence of the target sources.

6 Conclusions

Dealing with real-world sounds requires methods which do not rely on a conveniently chosen acoustic domain as traditional CASA methods do. Furthermore, these methods do not make use of explicit class and even instance specific knowledge to guide recognition, but rely on statistical techniques, which, by their very nature, blur (that is, average away) differences between events. Our intelligent agent approach can be contrasted to traditional CASA approaches, since it assumes the agent (that is, the recognition system) is situated in a complex and changing environment. The combination of bottom-up and top-down processing in our approach allows a connection with the real world through event physics. The insistence on physical realizability ensures that each event specific representation is both supported by the input and consistent with the systems knowledge. The reliability of a number of CPSP-based commercial products for verbal aggression and vehicle detection, which function in real-live and therefore unconstrained acoustic environments, is promising with respect to a successful implementation of our approach. We have a recipe, and a large number of ingredients. Let us try to proof the pudding.

Acknowledgments

References

1. Andringa, T.C. et. al. (1999), "Method and Apparatuses for Signal Processing", International Patent Application WO 01/33547.
2. Andringa, T.C. (2002), "Continuity Preserving Signal Processing", PhD dissertation, *University of Groningen*, see `http://irs.ub.rug.nl/ppn/237110156`.
3. Ballas, J.A. (1993), "Common Factors in the Identification of an Assortment of Brief Everyday Sounds", *Journal of Experimental Psychology: Human Perception and Performance 19(2)*, pp 250-267.
4. Cooke, M., Ellis, D.P.W. (2001), "The Auditory Organization of Speech and Other Sources in Listeners and Computational Models", *Speech Communication 35*, pp 141-177.
5. Cowling, M., Sitte, R. (2003), "Comparison of Techniques for Environmental Sound Recognition", *Pattern Recognition Letters 24*, pp 2895-2907.
6. Defréville, B., Roy, P., Rosin, C., Pachet, F. (2006), "Automatic recognition of urban sound sources", *Audio Engineering Society 120th Convention*.
7. Duifhuis, H., Hoogstraten, H.W., Netten, S.M., Diependaal, R.J., Bialek, W. (1985), "Modeling the Cochlear Partition with Coupled Van der Pol Oscillators", in J.W. Wilson and D.T Kemp (Ed.), *Cochlear Mechanisms: Structure, Function and Models*, pp 395-404. Plenum, New York.
8. Ellis, D.P.W. (1999), "Using Knowledge to Organize Sound: The Prediction-Driven Approach to Computational Auditory Scene Analysis and its Application to Speech/Nonspeech Mixtures", *Speech Communication 27(3-4)*, pp 281-298.
9. Gaver, W.W. (1993), "What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception", *Ecological Psychology 5(1)*, pp 1-29.
10. Gibson, J.J. (1979), *The Ecological Approach to Visual Perception*, Boston, MA: Houghton Mifflin.
11. Knill, K., Young, S. (1997), "Hidden Markov Models in Speech and Language Processing", in S. Young and G. Bloothooft (Ed.), *Corpus-Based Methods in Language and Speech Processing*, pp 27-68. Kluwer Academic, Dordrecht.
12. McClelland, J.L., Elman, J.L. (1986), "The TRACE Model of Speech Perception", *Cognitive Psychology 18*, pp 1-86.
13. Neuhoff, J.G. (2004), "Ecological Psychoacoustics: Introduction and History", in J.G. Neuhoff (Ed.), *Ecological Psychoacoustics*, pp 1-13. Elsevier Academic Press.
14. Norris, D. (1994), "Shortlist: A Connectionist Model of Continuous Speech Recognition", *Cognition 52*, pp 189-234.
15. Russell, S.J., Norvig, P. (1995), *Artificial Intelligence: A Modern Approach*. Prentice-Hall.
16. Sound Intelligence, `http://www.soundintel.com/`
17. Wang, D. (2005), "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis", in P. Divenyi (Ed.), *Speech Separation by Humans and Machines*, pp 181-197. Kluwer Academic, Norwell MA.
18. Warren, W.H., Verbrugge, R.R. (1984) "Auditory Perception of Breaking and Bouncing Events: A Case Study in Ecological Acoustics", *Journal of Experimental Psychology: Human Perception and Performance 10(5)*, pp 704-712.