International Journal of Semantic Computing

DISAMBIGUATING SOUND THROUGH CONTEXT

MARIA E. NIESSEN
Artificial Intelligence, University of Groningen, P.O. Box 407
9700 AK Groningen, The Netherlands
m.niessen@ai.rug.nl
http://www.ai.rug.nl/~maria

LEENDERT VAN MAANEN
Artificial Intelligence, University of Groningen, P.O. Box 407
9700 AK Groningen, The Netherlands
leendert@ai.rug.nl

TJEERD C. ANDRINGA
Artificial Intelligence, University of Groningen, P.O. Box 407
9700 AK Groningen, The Netherlands
t.andringa@ai.rug.nl

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

A central problem in automatic sound recognition is the mapping between low-level audio features and the meaningful content of an auditory scene. We propose a dynamic network model to perform this mapping. In acoustics, much research is devoted to low-level perceptual abilities such as audio feature extraction and grouping, which are translated into successful signal processing techniques. However, little work is done on modeling knowledge and context in sound recognition, although this information is necessary to identify a sound event rather than to separate its components from a scene. We first investigate the role of context in human sound identification in a simple experiment. Then we show that the use of knowledge in a dynamic network model can improve automatic sound identification by reducing the search space of the low-level audio features. Furthermore, context information dissolves ambiguities that arise from multiple interpretations of one sound event.

Keywords: auditory scene analysis; dynamic network; spreading activation.

1. Introduction

When human listeners are asked to describe an auditory scene—a collection of perceptual complexes, based on acoustic evidence, in which each complex represents a single event in an acoustic environment—they will describe the different sound events in terms of the sources that caused the sounds [2, 31, 34]. They will not describe the acoustic properties of the sound signal. For instance, a passing car will

be referred to as a car, not as a noisy harmonic complex in combination with a burst of noise. The evaluation of sounds in terms of the events that produced them is often named everyday listening [15]. In everyday listening, people obviously use information that is present in the sound signal (bottom-up processing). However, they also apply knowledge of the event and the context (top-down processing). We will demonstrate the role of knowledge of the context by presenting an experiment in which participants had to identify ambiguous sounds in different contexts. The results show that sound identification depends on the context in which a sound is heard. Next, we will present a method to automatically identify ambiguous sounds with the use of context.

In acoustics much research is devoted to modeling the ability of listeners to separate different events in an auditory scene based on the sound signal alone, called primitive auditory scene analysis (ASA) [6]. Perceptual grouping based on features such as continuity of components in the sound signal and proximity in time or frequency are translated into successful models of primitive ASA (e.g., [8, 16, 17, 23, 33]). However, primitive ASA alone will not suffice to automatically identify sound events. We also need to model the contribution of knowledge and context (sometimes referred to as schema-based ASA [33]). Although this need has been recognized some time ago [13, 14], it has so far not resulted in complete models of sound event identification, which combine primitive and schema-based ASA.

While context-based identification is mostly ignored in automatic sound recognition—with the notable exception of speech, where grammatical and lexical rules are essential for automatic recognition [5, 26]—it has a long history in other research areas such as information retrieval (e.g., [7, 11, 29, 30]) and handwriting recognition (e.g., [9, 22]). Models of context-based identification assume that certain regularities exist in the contexts in which an event may occur and structure their knowledge base in such a way that these regularities are accounted for. Often this takes the form of a spreading activation semantic network [25], in which the nodes represent the states the network can be in, and the edges represent the prior probabilities that these states are encountered subsequently or together. In these models, context is incorporated by keeping nodes active over a longer period of time, thereby influencing the probabilities that certain nodes will be activated. As a consequence of the typical properties of these research areas, spreading activation networks have mostly been exploited in static and well-constrained domains. Our aim is to demonstrate that spreading activation can also be applied in a dynamic domain such as auditory scene analysis.

Since we want to provide a model of the role of context in sound identification, we are also interested in its role in human sound identification, which is little investigated [20]. Therefore, we will first present the results of an experiment that was designed to determine whether context facilitates one of the interpretations of an ambiguous sound. It is known that sounds are more difficult to identify when they may have multiple possible sources [3]. Context is needed to disambiguate these

sounds, as is shown in an example of the same study. In this example, participants gave a sound event a different interpretation when it was combined with another sound event and different instructions. A follow-up study [4] did not find this facilitatory effect, but did find a suppressive effect of an incongruent context. These results show that context is a complex factor. Moreover, context can manifest itself in many different ways, such as in sound and image, but also in time of day and place of occurrence. The experiment described here is designed to show one particular effect, namely the facilitatory effect, that context can have on the interpretation of an ambiguous sound. The results of the experiment will be important for the automatic sound identification in two ways. First, if context is shown to be needed in human identification of environmental sounds, it is also required in an automatic system. Second, if the experiment shows that an auditory context has a facilitatory effect on sound identification, this will be important for improving automatic sound identification.

2. Experiment

To test the facilitatory effect of context on identification in human listeners we presented homonymous sounds to participants. Homonymous sounds are characterized by having two (or more) possible causes. When these sounds are presented in isolation, the probability that they are identified as one possible cause is the same as the probability that they are identified as the other possible cause. In contrast, when homonymous sounds are preceded by a sound that predisposes the listener to one of the two causes, we would expect a biased response towards that cause.

2.1. *Method*

To create homonymous sounds we used pairs of similar sounds from high-quality commercial sound effects recordings (Hollywood Edge and Sound FX The General), which were used previously to study the similarity of sound events [18]. Sound pairs that were found maximally similar in this study were combined to form chimaeric sounds. Chimaeric sounds are composed of the fine time structure of one sound and the temporal envelope of another sound [28]. The signal properties of the sound events varied greatly because of the diversity of the environmental sounds in the database. Hence, the chimaeric sounds did not always result in homonymous sounds. For 12 selected homonymous pairs[a], listed in Table 1, we chose the combination of fine structure and envelope that sounded most natural. Most of the envelopes of sounds A were used for the chimaeric sounds, while most of the fine structures of sounds B were used. The homonymous sounds had a mean duration of 2.8 seconds. The sounds that provided context for the homonymous sounds, listed in Table 2, were obtained from additional commercial recordings (Auvidis and Dureco). All

[a]The sounds can be found on http://www.ai.rug.nl/research/acg/exp/.

sounds were sampled at 44.1 kHz. The total of 52 sound sequences (two context conditions for the homonymous sounds, and 28 filler sequences, see next paragraph) had a mean duration of 7.7 seconds. The context sounds preceded the sounds to be identified such that the sequence sounded most natural. However, the context sound always ended before the end of the target sound. For example, when the context sound was rain, it continued through the start of the sound of thunder. In contrast, when the context sound was the closing of a refrigerator door, it ended before the sound of the pouring water started.

Table 1. List of similar sound pairs used to form homonymous sounds.

Sound A	Sound B
Pouring water	Rain
Thunder	Passing airplane
Whistle	Singing bird
Footstep	Drum
Toilet flush	Pouring water
Meowing cat	Crying baby
Coughing	Barking dog
Bouncing basketball	Closing door
Ticking clock	Bouncing pingpong ball
Water bubbles	Horse running
Bowling	Thunder
Zipper	Car starting

In total 42 participants with a mean age of 24 took part in the experiment. Six participants reported a slight hearing loss, but showed no decrease in their performance on the filler sounds compared to the normal hearing participants.

The experiment comprised three conditions, one in which the context sound facilitated the interpretation of sound A, one in which the context sound facilitated the interpretation of sound B, and a control condition in which the sounds were heard in isolation. The three conditions were presented between the participants. The homonymous target sounds were interluded with 28 filler sounds taken from the same database. They were included to assess the general performance of the participants, and to make the participants unaware which sounds were the target sounds. The total of 40 sounds was presented in random order, but no target sounds were present in the first 6 exposures to get the participants familiar with the task. The identification task was a binary choice task. For the target sounds the participants could choose between the descriptions of the two original sound events, and for the filler sounds they could choose between the actual cause and some other related source description. Furthermore, the participants had to indicate on a four-point scale how confident they were of their answer. The control group of 11 listeners identified the sound events in isolation. The second group of 15 listeners first heard a sound semantically consistent with context A followed by the target chimaeric

sound. Finally, the third group of 16 listeners first heard a sound semantically consistent with context B followed by the chimaeric sound. The 28 filler sequences, the filler sounds preceded by a semantically consistent sound, were the same for the last two groups. The control group heard the filler sounds without a context sound. The experiment was conducted on line during January 2008.

Table 2. List of sounds used to facilitate context A and B.

Context A	Context B
Refrigerator door	Thunder
Rain	Airport announcement
Football cheering	Forest
Closing door	Guitar
Urinating	Refrigerator door
Barking dog	Music box
Talking	Meowing cat
Cheering audience	Footsteps
Chiming clock	Applause
Teakettle whistling	Horse neighing
People talking	Rain
Raining on tent	Car door closing

2.2. *Results*

The score of all participants in every group on the filler sounds was 100%, and they gave a mean confidence rating of 2.8 on a four-point scale ranging from 0 to 3. A two-way analysis of variance (ANOVA) was used to test the difference in the response between the participants within the homonymous sounds. The effect of context A on the mean identification score compared to the mean score in isolation was significant: $F_1(1, 11) = 8.09$, with $p < 0.017$. However, there was no effect of context B on the mean identification score compared to the mean score in isolation ($F_1(1, 11) < 1$). The results are summarized in Figure 1. The bars depict the average score on option A for all participants within a group summarized for all homonymous sounds, where option A is the sound description that is in agreement with context A. The complement, 100% minus score A, is the average score on option B.

The difference between the confidence ratings in correct responses, that is, responses for which the answer was in agreement with the context sound, compared to the confidence ratings in incorrect responses was significant in the group that heard context A: $t(101) = 3.34$, with $p < 0.002$. The confidence rating was higher when the answer was in agreement with the context. The mean confidence ratings of consistent and inconsistent identifications are depicted in Figure 2. This effect was absent in the group that heard context B ($t(159) < 1$).

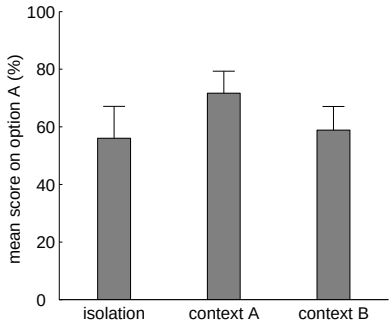

Fig. 1. Mean scores on option A in each of the three groups, with the standard error. The complements are the scores on option B.

2.3. *Discussion*

Not all chimaeric sounds appeared to be as homonymous as assumed. In particular three sounds received one interpretation exclusively in the isolated condition. When these three sounds were excluded from the ANOVA, the difference in the mean score of context A compared to the mean score in isolation had a greater F: $F_1(1, 8) = 13.28$, with $p < 0.007$. In conclusion, for the homonymous sounds we found a significant effect of one context on identification.

Although there is a significant effect of one context on the mean scores, this effect is completely absent in the other context. The explanation for the absence of the effect lies in the design of the experiment. The homonymous sounds were formed by combining the envelope of one sound and the fine structure of another sound. Most descriptions of context A predisposed the listener to the interpretation of the envelope of the homonymous sound, while the interpretation related to the

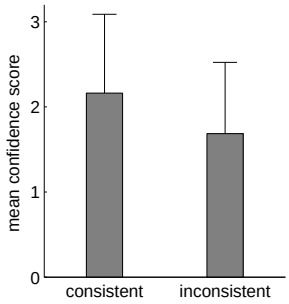

Fig. 2. Mean confidence ratings of consistent and inconsistent identifications in context A, with the standard deviation.

fine structure was most prominent in context B. Hence, the envelope seems to be a stronger cue for identification than the fine structure for this experimental design. This effect is known in speech perception [27, 28], and depends on the number of frequency bands used to create the chimaeric sound. If the number of frequency bands we used (eight) were used for the identification of chimaeric speech sounds, the fine structure would give relatively little information compared to the envelope. Hence, our results suggest this effect can be generalized to environmental sounds. As a consequence, the effect of context is canceled by the preference for the envelope in context B. This conclusion is consistent with a significant prevalence for the interpretation that coincided with the envelope of the homonymous sound (64%) compared to the fine structure (36%) when the sounds were presented in isolation ($\chi^2(1) = 9.82, p < 0.002$). Overall, the experiment demonstrates that the context in which a sound is heard determines in part its perception.

3. Dynamic network Model

Based on existing models of spreading activation and the findings of the experiment we introduce a model for context-based identification that can be used with dynamic sound input. This model allows automatic identification of events in a complex and changing auditory scene of a real-world environment. In complex real-world environments a sound signal may have different causes, depending on the situation in which it occurs. Furthermore, the bottom-up estimated audio features are meaningless by themselves and require interpretation. Therefore, we need knowledge to give meaning to the low-level audio features, and context to restrict the possible causes that they represent, similar to human listeners. The model dynamically builds a network that generates meaningful hypotheses of sound events based on low-level audio features and knowledge of these events. Moreover, context information is used to compute the support for competing hypotheses, and consequently a most likely hypothesis for all input events can be assessed.

3.1. *Network construction*

With our model we want to qualitatively improve automatic sound recognition. Our approach starts with data-driven techniques for the extraction and grouping of low-level audio features. The ability of human listeners to use context to disambiguate sounds, which we demonstrated in the experiment, should also be present in the model. Therefore, a dynamic network is added that uses knowledge of the event and the context to limit the search space of the bottom-up input [1]. We will describe the network's behavior through one of the sound events that was also used in the experiment, the mix of a bouncing basketball and a closing door, which can be identified as both in the absence of context information. Without any context information, similar to the control condition in the experiment, the actual cause of the sound is indefinite, as illustrated in Figure 3. In the following paragraphs we will describe how the model dissolves this ambiguity through the use of context.

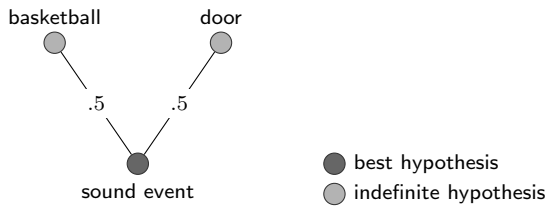

Fig. 3. Network configuration for the identification of a reverberant impact sound without context.

As described in the previous paragraph, we want to combine a bottom-up and top-down approach to sound recognition. In other words, hypotheses of the sound event based on the low-level audio features are matched to expectations that are formed by knowledge of the relations between the events and the context. This matching process will lead to a best hypothesis about the event or source causing the sound in this context, at every description level in the network. All hypotheses hold a confidence value reflecting their support from relations to other events and the context in which the hypothesized event is occurring. In case of conflicting explanations for one event, the hypothesis with the highest support will win. In Figure 4 for example, the sound event could be either a closing door or a basketball bouncing, based on the low-level audio features alone. However, knowledge about the context actuated by a previous sound event, cheering, will increase the support for the hypothesis that the second sound event is a basketball bouncing. Furthermore, the confidence value of the first hypothesis, cheering, is increased, because the context of a sports game, and hence the cheering, is more likely considering the new information. In the following we will describe the process of how the network is dynamically built, and how the confidence of all hypotheses is established through spreading activation.

The network is updated if and only if new bottom-up information is presented, and spreads its activation when the network is stable, that is, when all available knowledge about the bottom-up information is processed. The hierarchy in the network is captured by the interdependent relations of all the hypotheses. The lowest description level in the network corresponds to the physics of the signal, and the highest level to the semantics of the scene. The levels in between represent hypotheses of increasing generality. The number of levels depends on the complexity of the domain, but usually three levels will suffice: one for the audio features, one for the event that is inferred from the features, and one for the context of the scene, which can raise particular expectations. In the first step, audio features are extracted from the time-frequency plane of the sound. Figure 5 shows the spectrogram of a sports game scene with annotations of the audio features—the current version of the model operates on annotations of low-level audio features instead of automatically extracted audio feature descriptions. Every sound can be represented as a specific

pattern of these audio features. For example, the cheering is a noisy collection of distorted harmonic complexes. Each pattern comes with a basis activation based on the confidence given by the low-level grouping algorithms. For example, a confidence value may reflect how well a pattern fits a particular mask. However, since these algorithms are not coupled to the model yet, the basis activation of all patterns is set to 1.

A pattern of audio features may have multiple causes. Hence, all possible causes of a pattern will be initialized as hypotheses, after the extraction of patterns of audio features. Knowledge of the hypothesized events will then initiate more hypotheses,

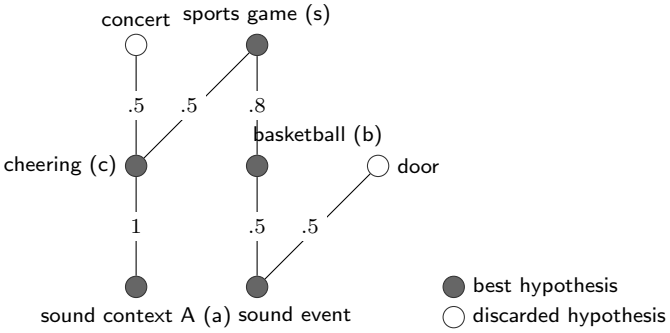

Fig. 4. Network configuration for the identification of a reverberant impact sound in context A. The best hypotheses at each level corresponds to a best explanation for the bottom-up evidence at that description level.

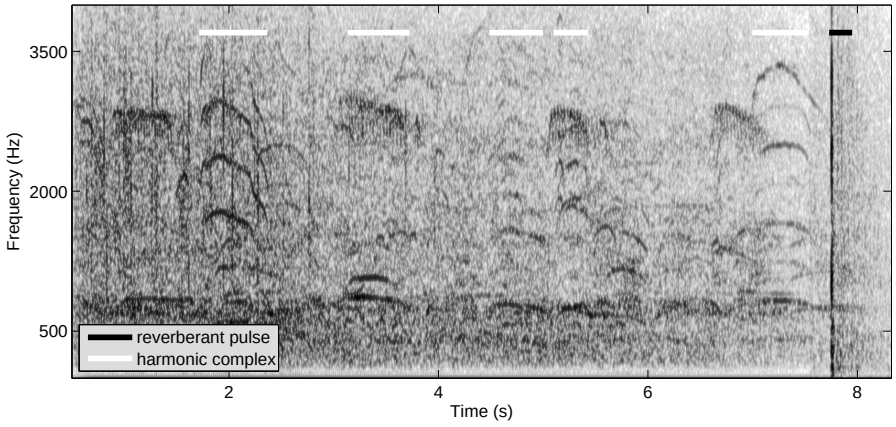

Fig. 5. Spectrogram of cheering followed by a chimaeric basketball/door sound, including the audio feature annotation episodes (harmonic complexes and a reverberant pulse).

for example about an event sequence or a context. In Figure 4, the cheering could mean a pop concert or a sports game. The higher level hypotheses create expectations about sound events that will follow, like a basketball in a sports game. If the expected event is matched with bottom-up evidence, it will receive extra support when its hypothesis is created. When all knowledge is processed and the network is stable, the activation of the low-level audio features spreads through the network.

3.2. *Spreading activation*

When the network configuration is stable after updating, the activation first spreads upward to the highest level in the network, and then downward to other connected events in the past, if they exist. The spreading can only go up once and down once through every path that denotes a past event, after which it terminates. The activation of the individual hypotheses is a time-dependent weighted sum that decays exponentially over time. The activation of each hypothesis is limited to a maximum. As a consequence, hypotheses that are highly active over a longer period of time are not repeatedly reinforced by new input, because the effect of the input decreases when the activation of a hypothesis reaches its maximum. Furthermore, the activation represents the reliability of a hypothesis when it is modulated. The computation of the spreading activation is similar to the method used in McClelland and Rumelhart's model of letter perception [22]. However, we only incorporate excitatory and no inhibitory connections, since the inhibitory effect is accounted for by missing connections, in which case only the decay function has an effect on the activation value. Furthermore, the decay function applied in our model is a continuous function of time instead of a constant value that is applied to discrete time steps.

The activation of hypothesis i at time $t + \Delta t$ is

$$A_i(t + \Delta t) = \mathrm{e}^{-\frac{\Delta t}{C}} A_i(t) + E_i(t), \tag{1}$$

where C is a constant parameter controlling the speed of decay, whose value depends on the application domain, and Δt is the elapsed time since the hypothesis i is last updated. The effect of connected events is represented by

$$E_i(t) = n_i(t)(M - \mathrm{e}^{-\frac{\Delta t}{C}} A_i(t)), \tag{2}$$

where M is the maximum activation level, usually set to 1, and $n_i(t)$ is the input of the event:

$$n_i(t) = \sum_j \alpha_{ij} a_j(t), \tag{3}$$

where j is a connected hypothesis at different updating times t, $a_j(t)$ is its activation, and α_{ij} is the weight of the relation between hypotheses i and j.

In the example of Figure 4, the activation of the sports game hypothesis is summed over the two time steps when new bottom-up information is presented to the network. The value of the decay parameter C is arbitrarily set to 100 to

demonstrate its effect in the calculation of the activation value. However, in different application domains the value of C can be estimated based on training data. In the first step, the activation of the sports game hypothesis consists of the input it gets from the cheering hypothesis, which starts at time $t = 1.7$ (the subscript letters are in parentheses in Figure 4):

$$A_{\mathrm{s}}(1.7) = \alpha_{\mathrm{cs}} a_{\mathrm{c}}(1.7) = 0.5 * 1 = 0.5 \tag{4}$$

A few seconds later, at time $t = 7.7$, the input is delivered by the basketball hypothesis:

$$\begin{aligned} A_{\mathrm{s}}(7.7) &= \mathrm{e}^{-\frac{7.7-1.7}{100}} A_{\mathrm{s}}(1.7) + \alpha_{\mathrm{bs}} a_{\mathrm{b}}(7.7)(1 - \mathrm{e}^{-\frac{7.7-1.7}{100}} A_s(1.7)) \\ &= 0.94 * 0.5 + 0.8 * 0.7 * (1 - 0.94 * 0.5) = 0.77 \end{aligned} \tag{5}$$

The activation of the cheering hypothesis is not included in the second step, because at every update only the active connected hypotheses can deliver input to the sports game hypothesis. As a consequence of the two-way spreading, the cheering hypothesis will receive an increased support from the basketball bouncing, through the sports game hypothesis. In the first step the hypothesis receives activation from the bottom-up evidence:

$$A_{\mathrm{c}}(1.7) = \alpha_{\mathrm{ac}} a_{\mathrm{a}}(1) = 1 * 1 = 1 \tag{6}$$

In the second step the sports game hypothesis contributes to the activation of the cheering hypothesis:

$$\begin{aligned} A_{\mathrm{c}}(7.7) &= \mathrm{e}^{-\frac{7.7-1.7}{100}} A_{\mathrm{c}}(1.7) + \alpha_{\mathrm{sc}} a_{\mathrm{s}}(7.7)(1 - \mathrm{e}^{-\frac{7.7-1.7}{100}} A_c(1.7)) \\ &= 0.94 * 1 + 0.5 * 0.78 * (1 - 0.94 * 1) = 0.96 \end{aligned} \tag{7}$$

The activation values of all the higher level hypotheses after spreading at the two time steps are depicted in Figure 6. Note that at time $t = 1.7$ the basketball and door hypothesis are not generated yet, because at that time the associated sound features have not been presented to the network.

3.3. *Discussion*

The network described in the example is rather simple, while in a real-world environment there will be many more events, mostly of deteriorated sound quality. The complexity of a real-world environment will have to be captured by the knowledge of the relations that exist between the real-world events. Furthermore, the expansion of the network will have to be controlled. This is partly done by keeping track of which hypotheses are active, and which hypotheses are finished or discarded. These last two classes are not included in the search space of connected hypotheses when new information is presented to the network. As a consequence, the search space at any time is limited to the hypotheses that are active at that time. An advantage of a complex environment is its supply of information. Human listeners use much more contextual information in the identification of sounds, such as time, location

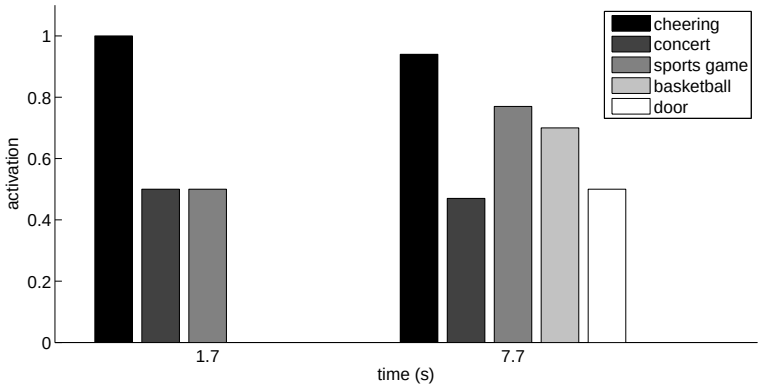

Fig. 6. Activation of higher-level hypotheses in the network at the two times when bottom-up audio features are presented.

and ecological frequency [2]. This information can also be included in our model as nodes in the network that help support or discard hypotheses.

The current implementation of the model receives annotations of grouped low-level audio features as bottom-up input, instead of automatically generated features. However, several techniques for low-level sound event descriptions are being developed [21, 32], which will supply the model with these grouped features. Furthermore, although the model is being developed for audio input, its general implementation allows for other low-level input, such as image descriptions, as long as they represent a single event or object. If different types of descriptions can serve as input to the model, they may be combined in one model for use in multimedia applications (cf. [35]). These issues will be subject to further investigation.

4. General discussion

As mentioned in the description of the model, we want to qualitatively improve automatic sound recognition. The main difference with existing models of environmental sound recognition (e.g., [10, 12]) lies in the explicit use of knowledge and the focus on identification rather than classification of sound events. Classification techniques assume a limited set of classes to which a sound signal may belong. Hence, such a system will assign all bottom-up input to a class, irrespective of how small the evidence is. In contrast, the model described here will only create hypotheses of events based on grouped bottom-up audio features, such as a harmonic complex, that relate directly to the source of a sound event. When this bottom-up evidence is absent, that is, when the audio input cannot be mapped onto any hypothesis, no identification will be made. Furthermore, we apply knowledge explicitly in the identification process, instead of implicitly by training the system on data that are

similar to the input the system will receive. Therefore, if our model operates in a different scene, or needs to identify different events, the information about the events and the context will be different, but the implementation of the model needs no adjustments. In contrast, most existing models of environmental sounds are developed for specific sounds or scenes, and cannot be generalized—an exception is the model used in the study of Defréville et al. [12], which is based on audio features that are automatically selected for a specific problem [24].

Although the model is generally applicable, the problem of the acquisition of scene and sound specific knowledge remains. Without knowledge the model cannot create hypotheses. Therefore, the relations between events and contexts need to be learned, as in classification models. However, instead of training the classifier, we want to explicitly learn these relations beforehand from examples, and continue to update them while the system is running. The current implementation of the model works with static databases, but we will incorporate machine learning techniques to be able to dynamically update the knowledge of the model.

In our model we use one benefit that context has for the identification of environmental sounds, namely to disambiguate a sound that may have multiple causes. However, the few studies that have been done on context-based sound identification in humans show several effects that context may have. While Ballas and Howard [3] concluded that context has a facilitatory effect on a particular interpretation of a homonymous sound, Gygi and Shafiro [19] found the opposite effect in a study where sounds were mixed with either congruent or incongruent scenes. People showed an increased identification performance when sounds were mixed with an incongruent scene, which indicates that context can also have a habituating effect on the identification of environmental sounds. However, we are interested in the improvement of automatic identification of environmental sounds. We have shown that context can have a facilitatory effect in identification, and used this asset to automatically disambiguate sounds that may have multiple causes.

Another extensive study on context-based sound identification [4] found only a suppressive effect of context. That is, an incongruent context decreased the correct identification score compared to an isolated condition, but a consistent context did not increase the performance compared to the isolated condition. The lack of facilitatory priming may be due to the experimental design, as the authors concluded. Furthermore, the use of contextual information might be highly dependent on the quality of the bottom-up evidence. More challenging acoustic environments are likely to be more influenced by top-down expectations. In these experiments the sounds were offered clean, and consequently required less need for context to be identified. Automatic environmental sound identification decreases in challenging environments, because the number of possible causes increases when the bottom-up evidence is unreliable. In these situations, context is even more important to select the most likely cause.

In summary, in the sound identification experiment we have demonstrated that context in part determines the perception of sound events, although the effect of

context is not straightforward. Furthermore, we introduced a computational model for the analysis of dynamic auditory scenes, in which we used the facilitatory effect of context to dissolve ambiguities. To show the validity of the model, we plan to test it on databases of real events. Furthermore, testing the model on real events will allow us to compare its performance to other computational models.

Acknowledgements

This work is supported by SenterNovem (Dutch Companion project grant no. IS053013) and NWO (ToKeN/I^{2}RP project grant no. 634.000.002). We are grateful to Brian Gygi for making the database used in his study [18] available to us. Ronald van Elburg, Dirkjan Krijnders, Renante Violande, Hedde van de Vooren, and Elske van der Vaart are acknowledged for valuable comments on earlier versions of the paper.

References

[1] T. C. Andringa and M. E. Niessen. Real-world sound recognition: A recipe. In *Proceedings of the 1st Workshop on Learning Semantics in Audio Signals (LSAS 2006)*, pages 106–118, 2006.

[2] J. A. Ballas. Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2):250–267, 1993.

[3] J. A. Ballas and J. H. Howard. Interpreting the language of environmental sounds. *Environment and Behavior*, 19(1):91–114, 1987.

[4] J. A. Ballas and T. Mullins. Effects of context on the identification of everyday sounds. *Human Performance*, 4(3):199–219, 1991.

[5] J. P. Barker, M. P. Cooke, and D. P. W. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45(1):5–25, 2005.

[6] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[7] P. R. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, 23(4):255–268, 1987.

[8] M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35(3-4):141–177, 2001.

[9] M. Côté, E. Lecolinet, M. Cheriet, and C. Y. Suen. Automatic reading of cursive scripts using a reading model and perceptual concepts. *International Journal on Document Analysis and Recognition*, 1(1):3–17, 1998.

[10] M. Cowling and R. Sitte. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895–2907, 2003.

[11] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997.

[12] B. Defréville, P. Roy, C. Rosin, and F. Pachet. Automatic recognition of urban sound sources. In *Proceedings of the 120th Audio Engineering Society Convention (AES 2006)*, 2006.

[13] D. P. W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996.

[14] D. P. W. Ellis. Using knowledge to organize sound: The prediction-driven approach

to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication*, 27:281–298, 1999.

[15] W. W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[16] D. J. Godsmark and G. J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366, 1999.

[17] S. Grossberg, K. K. Govindarajan, L. L. Wyse, and M. A. Cohen. ARTSTREAM: a neural network model of auditory scene analysis and source segregation. *Neural Networks*, 17(4):511–536, 2004.

[18] B. Gygi, G. R. Kidd, and C. S. Watson. Similarity and categorization of environmental sounds. *Perception & Psychophysics*, 69(6):839–855, 2007.

[19] B. Gygi and V. Shafiro. Effect of context on identification of environmental sounds. *The Journal of the Acoustical Society of America*, 119(5):3334, 2006.

[20] B. Gygi and V. Shafiro. General functions and specific applications of environmental sound research. *Frontiers in Bioscience*, 12:3152–3166, 2007.

[21] J. D. Krijnders, M. E. Niessen, and T. C. Andringa. Robust harmonic complex estimation in noise. In *Proceedings of the 19th International Congress on Acoustics (ICA 2007)*, 2007.

[22] J. L. McClelland and D. E. Rumelhart. An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, 88(5):375–407, 1981.

[23] J. Nix and V. Hohmann. Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE Transactions on Audio Speech and Language Processing*, 15(3):995–1008, 2007.

[24] F. Pachet and P. Roy. Exploring billions of audio features. In *Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing (CBMI 2007)*, pages 227–235, 2007.

[25] M. Ross Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 216–270. MIT Press, Cambridge, MA, 1968.

[26] O. Scharenborg. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347, 2007.

[27] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

[28] Z. M. Smith, B. Delgutte, and A. J. Oxenham. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416:87–90, 2002.

[29] L. van Maanen. Mediating expert knowledge and visitor interest in art work recommendation. In *Proceedings of the Workshop Lernen-Wissen-Adaption (LWA 2007)*, pages 367–372, 2007.

[30] L. van Maanen, H. van Rijn, M. van Grootel, S. Kemna, M. Klomp, and E. Scholtens. Personal publication assistant: Abstract recommendation by a cognitive model. In press.

[31] N. J. Vanderveer. *Ecological acoustics: Human perception of environmental sounds.* PhD thesis, Cornell University, 1979.

[32] R. Violanda, H. van de Vooren, and T. C. Andringa. Signal component estimation using phase and energy information. In preparation.

[33] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis.* John Wiley and Sons, Holoken, NJ, 2006.

[34] W. A. Yost. Auditory image perception and analysis: The basis for hearing. *Hearing Research*, 56(1-2):8–18, 1991.

[35] W. Zajdel, J. D. Krijnders, T. C. Andringa, and D. M. Gavrila. CASSANDRA: audio-

video sensor fusion for aggression detection. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007)*, pages 200–205, 2007.