Context-based sound event recognition

Niessen, Maria Elisabeth

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2010

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
A central problem in automatic sound recognition is the mapping between signal-driven audio patterns and the semantic interpretation of a sonic environment. We propose a context model to perform this mapping. Acoustics research is predominantly devoted to mimic early stage human perceptual abilities such as audio pattern selection and grouping, which are translated into successful signal processing techniques. In contrast, not many studies are aimed at modeling knowledge and context in sound recognition, although this information is necessary to recognize a sound event in addition to segregating its components from a scene. Based on the investigation of the role of context in human sound event recognition in the previous chapter, we show that the use of knowledge in a context model can improve automatic sound event recognition by reducing the search space of the signal-driven audio patterns. Furthermore, context information dissolves ambiguities that arise from multiple interpretations of one sound event.
4.1 **INTRODUCTION**

In acoustics much research is devoted to modeling the ability of the human auditory system to segregate different events in a sonic environment based on the audio signal alone, called primitive auditory scene analysis (ASA, Bregman, 1990). Perceptual grouping based on features such as continuity of components in the audio signal and proximity in time or frequency are translated into successful models of primitive ASA (Cooke and Ellis, 2001; Godsmark and Brown, 1999; Grossberg *et al.*, 2004; Nix and Hohmann, 2007; Wang and Brown, 2006). However, primitive ASA alone will not suffice to automatically recognize sound events. We also need to model the contribution of knowledge and context to interpret the audio signal and make predictions at a higher description level (see chapter 3). Although this need has been recognized some time ago (Ellis, 1996, 1999), it has so far not resulted in models of sound event recognition that combine signal-driven (bottom-up) and context-based (top-down) methods.

In recent years there has been some research on modeling what is called context awareness in sound recognition. One group of studies focusses on estimating the context of an audio interval with varying classification techniques (Eronen *et al.*, 2006; Chu *et al.*, 2009; Aucouturier *et al.*, 2007). In these studies the context is represented by a class of sounds that can be heard in some type of environment, such as cars at a street, or people talking in a restaurant. Depending on the number of context classes that are learned, the recognition rates of these methods vary between 58% (24 classes, Eronen *et al.*, 2006) and 84% (14 classes, Chu *et al.*, 2009). Although these results are promising, the methods that are used have some attributes that make them less suitable for automatic sound event recognition.

First, no sound event segregation (see section 2.3) is applied, so the features that are used to classify an audio interval are assumed to represent information that is specific for a class. Therefore, the context class to which an audio interval belongs gives primarily information about its acoustic properties. Consequently, these methods implicitly assume that similar sound types occur in similar situations. Although this assumption is valid for some tasks in a restricted domain, like music genre determination, it cannot be guaranteed for sound events in real-world environments. The context (as defined in section 3.1) of an event in a real-world situation is not only the acoustic class it may be categorized into, but the envi-
MODELLING CONTEXT

Environment in which it can be heard as well. For example, speech can be heard in restaurant, together with music and clinking of glasses, but also in a park, together with birds and wind.

Second, tasks in multimedia applications (or a comparable setup in environmental sound classification, as in Chu et al., 2009) generally entail that a small audio interval, in the studies above typically not longer than a few seconds, is classified as a sample of one context out of a data set with a limited set of distinct contexts, which are stored as a collection of audio files. The estimation of a stationary context is easier when the intervals are longer, because the reliability of the audio features increases with time, assuming the audio signal does not change qualitatively. In contrast, in a real-world environment it cannot be assumed that the sonic composition within a context—where context refers not to the type of sound, but to the associations between events and environments (see Box 3.2 on page 35)—does not change over time. Furthermore, in continuous monitoring of a sonic environment there is no prior segmentation of interesting sound intervals. Therefore, the information in the sound used to define the context is not necessarily relevant, as it is assumed to be in the audio files in multimedia applications. To be able to determine the environmental context of a sound event, the sound event needs to be segregated and recognized, and co-occurring events that are semantically related to the same context can help to estimate the most likely context.

A second group of studies on context awareness addresses some of the above issues by retrieving semantic relatedness of sound intervals rather than the similarity of their acoustic properties (Lu and Hanjalic, 2008, 2009; Cai et al., 2006, 2008). For example, in Cai et al. (2006, 2008) the intervals are clustered based on the similarity of their audio features. Subsequently, the semantic relatedness of these intervals to each context (different tracks from shows on television, such as series or tennis games) is calculated based on their co-occurrences. However, the audio intervals that are selected give no information about the semantics of the interval itself, only about the context it belongs to. Since we want to recognize events in a real-world sonic environment, we are not interested in the context per se, but in its usage to improve recognition.

Furthermore, sound event recognition is different from segmentation, where an audio signal is divided into different intervals based on features, such as the zero crossing rate in the time domain and the spectral centroid, that represent prop-
Chapter 4

Properties of the signal (Tzanetakis and Cook, 1999). Figure 4.1 displays a schematic overview of the segmentation process. The segmented intervals do not necessarily correspond to events. Sound events in a real-world environment are often heard simultaneously, or no (interesting) events are heard at all. In other words, we do not want to know the acoustic class of each interval, but which events are present in a continuous stream of audio. Context can help to limit the search space of possible sound events.

While context-based recognition has received little attention in automatic sound recognition—with the notable exception of speech, where grammatical and lexical rules are considered for automatic recognition (Barker et al., 2005; Scharenborg, 2007)—it has a long history in other research areas such as information retrieval (Cohen and Kjeldsen, 1987; Crestani, 1997; Van Maanen, 2007; Van Maanen et al., 2010) and handwriting recognition (Côté et al., 1998; McClelland and Rumelhart, 1981). Models of context-based recognition 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 (Quillian, 1968; Collins and Loftus, 1975), in which the nodes represent the states the network can be in, and the vertices 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. 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 sound event recognition.

4.2 Context Model

Based on existing models of spreading activation and the findings of the context facilitation experiment in section 3.3 we introduce a model for context-based recognition that can be used with dynamic real-world audio input. This model allows automatic recognition of events in a complex and changing sonic environment. In complex real-world environments a sound event may have different interpretations, depending on the situation in which it occurs. Therefore, the model needs knowledge about the context to interpret the audio features, similar to humans.
Figure 4.1: Schematic overview of a segmentation process. Acoustic features, such as the zero crossing rate and the spectral centroid, are extracted from the audio signal. Based on the feature values for the different time frames, the audio signal can be segmented into different intervals that have corresponding feature values, indicated by gray scales in the bottom panel.
Figure 4.2: Schematic overview of a network initiated by a pattern segregated from an audio signal. The pattern is connected to event hypotheses that are possible interpretations of the pattern. The learned associations to contexts can help to infer the most likely interpretation for a certain pattern. All possible connections are depicted as lines. The strength of these connections is indicated with a symbol $w$.

The model dynamically builds a network that generates semantic hypotheses of sound events based on signal-driven audio patterns and knowledge of the events. Moreover, context information is used to compute the support for competing hypotheses, and consequently a most likely hypothesis for all segregated patterns can be assessed. Figure 4.2 shows a schematic representation of a network.

With our model we want to qualitatively improve automatic sound event recognition. Our approach starts with signal-driven techniques for the selection and grouping of audio components (the methods for segregation are explained in section 5.2.1). Every segregated pattern represents a possible sound event. The ability of people to use their knowledge of the context to disambiguate sounds, which we demonstrated in the experiment in chapter 3, should also be present in the model. Therefore, a context model evaluates the signal-driven input with knowledge of
Table 4.1: Algorithm for updating the dynamic network configuration at times when new signal-driven information is presented to the network.

For all patterns (grouped components) at time $t$:

1. Segregate the audio signal into patterns $P(t) = \{p_{t,i}\}$
2. For each $p_{t,i} \in P(t)$ add possible event hypotheses $\{e_{t,j}\} \in E(t)$ and connect them with strength $w_{i,j}$
3. For each new event $e_{t,j} \in E(t)$ add appropriate contexts $c_k$ not yet present in the set of active contexts $C$
4. Connect each new event $e_{t,j}$ with the appropriate context $c_k$ with strength $w_{j,k}$
5. Spread signal-driven activation
6. Spread context-based activation
7. Evaluate activation values

the event and its context (Andringa and Niessen, 2006). We will illustrate the behavior of the model through an example of a sound event 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. In the following sections we will describe how the model dissolves this ambiguity through the use of context knowledge in a dynamic network. The algorithm for the construction and updating of the dynamic network is summarized in Table 4.1.

4.2.1 Dynamics of context model

Because we want to combine a signal-driven (bottom-up) and context-based (top-down) approach to sound recognition, the model maps hypotheses of the sound event based onto segregated audio patterns to expectations that are formed by knowledge of the relations between the events and the context. This mapping process will lead to a best hypothesis about the event that causes the sound in this context, at every description level in the network (apart from the lowest, which are the segregated audio patterns). 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 interpretations for one event, the hypothesis with the highest support will win. For example, in Figure 4.3 the reverberant impact sound could be either a closing door or a basketball bouncing,
based on the audio pattern 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 input. In the following paragraphs 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 signal-driven information is presented, and spreads its activation when the network is stable, that is, when the available knowledge about the signal-driven 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

Figure 4.3: Network configuration for the identification of a reverberant impact sound preceded by the sound of cheering. The best hypotheses at the highest two levels (the gray nodes) correspond to the best interpretation for the signal-driven evidence at that description level.
MODELING CONTEXT

signal, and the highest level to a (provisional) interpretation of the environment. The intermediate levels 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 segregated patterns, one for the event hypotheses that are inferred from the patterns, and one for the context of the environment, which can raise particular expectations about future events (Figure 4.2).

4.2.2 Algorithm of context model

In the first step (see Table 4.1), audio patterns that are likely to be caused by sound events are segregated from the time-frequency plane of the sound. Figure 4.4 shows the time-frequency plane of a sports game scene with annotated audio patterns. Every hypothesis of a sound event corresponds to a specific pattern of audio components. For example, the cheering is a noisy collection of distorted harmonic complexes. Each segregated event comes with a base-level activation based on the confidence given by the segregation algorithms. For example, a confidence value may reflect how well a pattern, such as a harmonic complex, fits a particular mask, such as a calculated harmonic complex. For illustrative purposes we set the base-level activations of all patterns to 1 in the example.

A segregated pattern of audio components may have multiple interpretations. Hence, all proposed interpretations of a pattern will be initialized as sound event hypotheses (step 2 in Table 4.1). These event hypotheses are connected to the patterns that initiated them, with some strength, denoted by weight \( w \). Subsequently, knowledge about the hypothesized events will initiate hypotheses about the context in which an event can occur, for example about an event sequence or an environmental setting (step 3), and connected to the event hypotheses (step 4). In Figure 4.3, the cheering could point to a pop concert or a sports game. These higher level context hypotheses create expectations about sound events that will follow, like a basketball in a sports game. If the expected event is matched with signal-driven evidence, it will receive extra support when its hypothesis is created.

---

1 In the application of the model we use grouped audio components that are automatically extracted from the audio signal instead of annotations (see chapter 5).

2 The knowledge about the context is learned in a training phase. The method used for training is discussed for each application, that is, section 5.2.2 and 6.2.2.


**Figure 4.4:** The cochleogram of the sound of cheering followed by a chimaeric basketball/door sound. The gray bars indicate annotations of harmonic complexes and the black bar the annotation of a reverberant pulse.

When the knowledge is processed and the network is stable, the activation of the audio pattern spreads through the network (step 5 and 6).

The connections in the dynamic network are symmetrical, and only between hypotheses at different levels. For instance, the event hypotheses are connected to the segregated patterns that initiated them, and to hypotheses of their possible contexts, but not to each other (Figure 4.2). Connections between hypotheses at the same level would be redundant, since they can reinforce each other through shared parent hypotheses. Furthermore, the hierarchy of the network is now captured by the connections. Therefore, it is not necessary to store a global representation of the complete network. Instead, each hypothesis contains information of its relative position in the network, that is, it stores its direct connections. The only information that is globally available is which hypotheses are active.
4.3 Activation Spreading

When the network configuration is stable after updating, the activation first spreads upward to the highest level in the network (signal-driven activation spreading, step 5), and then downward to other connected events in the past, if they exist (context-based spreading, step 6). 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 with time. The rate of decay is determined by a time constant $\tau$. The activation of each hypothesis is limited to a maximum value. 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. Because the connection strength between the hypotheses are learned on training data (section 5.2.2 and 6.2.2), the activation represents a pseudo-probability (confidence) that the hypothesis is true.

The computation of the spreading activation is similar to the method used in the model of letter perception by McClelland and Rumelhart (1981). However, we only incorporate excitatory and no inhibitory connections. Furthermore, the decay function applied in our model is a continuous function of time instead of a constant value that is applied at discrete time steps.

The input activation $n_i(t)$ of the individual hypotheses is the weighted sum of all connected hypotheses, either from the level below, for signal-driven activation spreading (step 5), or from the level above, in case of context-based spreading (step 6):

$$n_i(t) = \sum_j w_{ji} A_j(t), \quad (4.1)$$

where $j$ is a hypothesis connected to $i$, $A_j(t)$ is its activation, and $w_{ji}$ is the connection strength between hypotheses $j$ and $i$, retrieved from the stored knowledge.

When an audio pattern holds a low confidence value, the activation spreading from the higher levels to the event hypotheses is more important than the activation spreading upward from the pattern, and vice versa. In other words, the lower the saliency of the signal, the more influential the context is. As a consequence, the dynamic network is more robust to unreliable input than models that rely only on signal-driven techniques.
4.3.1 Activation evaluation

After the activation has spread through the network, the activation of each hypothesis is evaluated (step 7). The activation evaluation is an accumulation of current input and the previous activation corrected with a decay. The decay ensures that items in short-term memory are forgotten without reinforcement (new signal-driven evidence) in contrast to information in long-term memory (Quillian, 1968). The activations of all hypotheses decay exponentially with time toward a default situation. Therefore, the decay function is dependent on the a priori activation of a hypothesis:

\[ f_i(\Delta t) = e^{-\frac{\Delta t}{\tau}}(1 - \hat{A}_i) + \tilde{A}_i, \]  

(4.2)

where \( \hat{A}_i \) is the default activation of hypothesis \( i \), which is non-zero for a closed set of contexts. For example, when the context is represented by one of \( N \) locations, the default activation can be determined from their incidences in the training data. In such a situation, the sum of the default activations of all context hypotheses is 1. For all other hypotheses \( \hat{A}_i = 0 \). Furthermore, \( \tau \) is a time constant controlling the rate of decay, and \( \Delta t \) is the elapsed time since hypothesis \( i \) is evaluated last. As a result, hypotheses deactivate when they do not receive input activation from other hypotheses. When the activation value decreases below a minimum value \( (\theta_A) \), the hypothesis is no longer evaluated, and removed from the dynamic network. A new hypothesis will be initiated when new evidence is found for the same type of event.

The activation value of the hypotheses is normalized to the maximum input activation, so that it is scaled between 0 and 1:

\[ A_i(t) = f_i(\Delta t)A_i(t - \Delta t) + n_i(t)(M - f_i(\Delta t)A_i(t - \Delta t)), \]  

(4.3)

where \( M = 1 \) is the maximum activation level, and \( A_i(t - \Delta t) \) is the activation of hypothesis \( i \) when the network was last updated, multiplied with a decay \( f_i(\Delta t) \), computed according to equation 4.2. Furthermore, \( n_i(t) \) is the input activation as calculated in equation 4.1. It should be noted that the activation of the audio patterns will always decay, because they do not get any more input activation \( (n_i(t) = 0) \) after being initiated. In contrast, event and context hypotheses can get reinforced by new evidence from subsequent audio patterns, and thus can stay active for a longer period of time. The result of the activation evaluation of a hypothesis is treated as the pseudo-probability that the hypothesis is true.
Going back to the example of Figure 4.3, the activation of the sports game hypothesis is summed over the two time steps when new signal-driven information is presented to the network. The value of the time constant $\tau$ is arbitrarily set to 100 to demonstrate its effect in the calculation of the activation value. However, in different application domains the value of $\tau$ can be estimated based on training data. At the first moment, 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 4.3):

$$A_s(1.7) = n_s(1.7) = w_{cs}A_c(1.7) = 0.5 \times 1 = 0.5$$ (4.4)

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

$$A_s(7.7) = e^{-\frac{7.7-1.7}{100}}A_s(1.7) + w_{bs}A_b(7.7)(1 - e^{-\frac{7.7-1.7}{100}}A_s(1.7))$$
$$= 0.94 \times 0.5 + 0.8 \times 0.7 \times (1 - 0.94 \times 0.5) = 0.77$$ (4.5)

The activation of the cheering hypothesis is not included in equation 4.5, 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 signal-driven evidence:

$$A_c(1.7) = n_c(1.7) = w_{hc}A_h(1) = 1 \times 1 = 1$$ (4.6)

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

$$A_c(7.7) = e^{-\frac{7.7-1.7}{100}}A_c(1.7) + w_{sc}A_s(7.7)(1 - e^{-\frac{7.7-1.7}{100}}A_c(1.7))$$
$$= 0.94 \times 1 + 0.5 \times 0.77 \times (1 - 0.94 \times 1) = 0.96$$ (4.7)

The activation values of all the higher level hypotheses are shown in Figure 4.5.
Figure 4.5: Activation of all higher-level hypotheses in the example of Figure 4.3. At the two moments when signal-driven patterns are presented to the network, the activation values of the hypotheses that are connected to these patterns increase.

4.4 CONCLUSION

The network described in the example is rather simple, while in a real-world environment there will be many more events, mostly of unreliable sound quality. The complexity of a real-world environment will be captured by knowledge about the relations between the real-world events. Furthermore, the expansion of the network must be controlled. This is partly achieved by keeping track of which hypotheses are active, and which hypotheses are of finished or discarded events. These last two classes are excluded from 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. People use much more contextual information in the recognition of sound events, such as time of day, environmental setting and ecological frequency (Ballas, 1993). This information can also be included in our model in the form of nodes in the network that help support or discard hypotheses.

In the next part we will show applications of the model in real-world situations,
where the input patterns of the model are supplied by automatic audio signal segregation algorithms. Furthermore, we will show how the knowledge of the model is acquired in the training phase of an application. Although the model is being developed for audio input, its general implementation allows for other signal-driven input, such as image descriptions, as long as they represent a single event or object (chapter 6). If different types of descriptions can serve as input to the model, they may be combined in one model for use in multimedia applications.