EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Context-dependent sound event detection

Toni Heittola[1*], Annamaria Mesaros[1], Antti Eronen[2] and Tuomas Virtanen[1]

Abstract

The work presented in this article studies how the context information can be used in the automatic sound event detection process, and how the detection system can benefit from such information. Humans are using context information to make more accurate predictions about the sound events and ruling out unlikely events given the context. We propose a similar utilization of context information in the automatic sound event detection process. The proposed approach is composed of two stages: automatic context recognition stage and sound event detection stage. Contexts are modeled using Gaussian mixture models and sound events are modeled using three-state left-to-right hidden Markov models. In the first stage, audio context of the tested signal is recognized. Based on the recognized context, a context-specific set of sound event classes is selected for the sound event detection stage. The event detection stage also uses context-dependent acoustic models and count-based event priors. Two alternative event detection approaches are studied. In the first one, a monophonic event sequence is outputted by detecting the most prominent sound event at each time instance using Viterbi decoding. The second approach introduces a new method for producing polyphonic event sequence by detecting multiple overlapping sound events using multiple restricted Viterbi passes. A new metric is introduced to evaluate the sound event detection performance with various level of polyphony. This combines the detection accuracy and coarse time-resolution error into one metric, making the comparison of the performance of detection algorithms simpler. The two-step approach was found to improve the results substantially compared to the context-independent baseline system. In the block-level, the detection accuracy can be almost doubled by using the proposed context-dependent event detection.

1 Introduction

Sound events are good descriptors for an auditory scene, as they help describing and understanding the human and social activities. A *sound event* is a label that people would use to describe a recognizable event in a region of the sound. Such a label usually allows people to understand the concept behind it and associate this event with other known events. Sound events can be used to represent a scene in a symbolic way, e.g., an auditory scene on a busy street contains events of passing cars, car horns, and footsteps of people rushing. Auditory scenes can be described with different level descriptors to represent the general context (street) and the characteristic sound events (car, car horn, and footsteps). As a general definition, a *context* is information that characterizes the situation of a person, place, or object [1]. In this study, the definition of context is narrowed to the location of auditory scene.

Automatic sound event detection aims at processing the continuous acoustic signal and converting it into such symbolic descriptions of the corresponding sound events present at the auditory scene. The research field studying this process is called computational auditory scene analysis [2]. Automatic sound event detection can be utilized in a variety of applications, including context-based indexing and retrieval in multimedia databases [3,4], unobtrusive monitoring in health care [5], surveillance [6], and military applications [7]. The symbolic information about the sound events can be used in other research areas, e.g., audio context recognition [8,9], automatic tagging [10], and audio segmentation [11].

Our everyday auditory scenes are usually complex in sound events, having a high degree of overlapping between the sound events. Humans can easily process this into distinct and interpreted sound events, and follow a specific sound source while ignoring or simply acknowledging the others. This process is called auditory scene analysis [12]. For example, one can follow a conversation in a busy background consisting of other people talking. Human sound perception is also robust to many

*Correspondence: toni.heittola@tut.fi
[1]Department of Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere, Finland
Full list of author information is available at the end of the article

environmental conditions influencing the audio signal. Humans can recognize the sound of footsteps, regardless of whether they hear footsteps on a pavement or on gravel, in the rain or in a hallway. In case of an unknown sound event, humans are able to hypothesize as to the source of the event. Humans use their knowledge of the context to predict which sound events they are likely to hear, and to discard interpretations that are unlikely given the context [13]. In real-world environments, sound events are related to other events inside a particular environment, providing a rich collection of contextual associations [14]. In the listening experiments, this facilitatory effect of the context to the human sound identification process has been found to partly influence the perception of the sound [15].

Automatic sound event detection systems are usually designed for specific tasks or specific environments. There are a number of challenges in extending the detection system to handle multiple environments and a large set of events. Event categories and variance within each category make the automatic sound event recognition problem difficult even with well-represented categories when having clean and undistorted signals. The overlapping sound events that constitute a natural auditory scene create an acoustic mixture signal that is more difficult to handle. Another challenge is the presence of certain sound events in multiple contexts (e.g., footsteps present in contexts like street, hallway, beach) calling for rules in modeling of the contexts. Some events are context specific (e.g., keyboard sounds present in the office context) and their variability is lower, as they always appear in similar conditions.

A possible solution to these challenges is to use the knowledge about the context in the sound event detection in the same manner as humans do [15], by reducing the search space for the sound event based on the context. We achieve this by implementing a first stage for audio context recognition and event set selection. The context information will provide rules for selecting a certain set of events. For example, it will determine excluding the footsteps class when the tested recording is from inside a car. A smaller set of event models will reduce the complexity of the event detection stage and will also limit the possible confusions and misclassifications. Further, context-dependent prior probabilities for events can be used to predict most likely events for the given context. The context information offers also possibilities for improving the acoustic sound event models used in the detection system. A context-dependent training and testing has the benefit of better fitting acoustic models for the sound event classes, by using only examples from a given context. For example, footsteps are acoustically different on a corridor (hallway context) than on the sand (beach context), and using specific models should be beneficial.

This article studies how to use context information in the sound event detection process, and how this additional information improves the detection accuracy. The proposed sound event detection system is composed of two stages: a context recognition stage and a sound event detection stage. Based on the recognized context, a context-specific set of sound events is selected for the sound event detection stage. In the detection stage, context-dependent acoustic models and count-based event priors are used. Two alternative event detection approaches are studied. In the first one, monophonic event sequence is outputted by detecting most prominent sound event at each time instance. In the second approach, a polyphonic event sequence is produced by detecting multiple overlapping sound events.

The rest of this article is organized as follows. Section 2 discusses related previous work, and Section 3 explains basic concepts of sound event detection. Section 4 presents a detailed description of the proposed context-dependent sound event detection system. Section 5 presents the audio database and metrics used in the evaluations. Section 6 contains detailed results of the evaluations and the discussions of the results. Finally, concluding remarks and future research directions are given in Section 7.

2 Previous work

Early research related to the classification of sounds for everyday life has been concentrating on problems with specific sounds. Examples include gunshots [16], vehicles [17], machines [18], and birds [19]. In addition to this, usually a low number of sound categories are involved in the studies, specifically chosen to minimize overlapping between different categories, and evaluations are carried out with one or very small set of audio contexts (kitchen [20], bathroom [21], meeting room [22], office and canteen [23]). Many of these previously presented methods are not applicable as such for the automatic sound event detection for continuous audio in real-world situations.

The problem of sound event detection in real environments having a large set of overlapping events was addressed in the acoustic event detection task (AED) of the Classification of Events, Activities and Relationship (CLEAR) evaluation campaign [24]. The goal of the AED task was to detect non-speech events in the meeting room environment. The metric used in the evaluation was designed for the detection system outputting a monophonic sequence of sound events. The best performing system submitted to the evaluation achieved a 30% detection accuracy by using AdaBoost-based feature selection and a Hidden Markov Model (HMM) classifier [25]. Later this study was extended into a two-stage system having a tandem connectionionist-HMM-based classification stage and a re-scoring stage [26]. The authors

achieved a 45% detection accuracy on the CLEAR evaluation database. Sound event detection for a wider set of real-world audio contexts was studied in [27]. A system based on Mel-frequency cepstral coefficients (MFCC) features and an HMM classifier achieved on average a 30% detection accuracy over ten real-world audio contexts.

In addition to the acoustic features and classification schemes, different methods have been studied to include prior knowledge of the events to the detection process. Acoustically homogeneous segments for the environment classification can be defined using frame level n-grams, where n-grams are used to model the prior probabilities of frames based on previously observed ones [28]. In a complex acoustic environment with many overlapping events, the number of possible combinations is too high to be able to define such acoustically homogeneous segments and for modeling transitions between them. In [3], a hierarchical probabilistic model was proposed for detecting key sound effects and audio scene categories. The sound effects were modeled with HMMs, and a higher-level model was used to connect individual sound effect models through a grammar network similar to language models in speech recognition. A method of modeling overlapping event priors has been addressed in [29], by using probabilistic latent semantic analysis to calculate priors and learn associations between sound events. The context-recognition stage proposed in this article will solve the associations of the sound events by splitting the event set into subsets according to the context. Furthermore, the count-based priors estimated from training material can be used to provide probability distributions for the sound events inside each context.

In order to be able to do context-dependent sound event detection, we introduce a context recognition step. In recent years, there has been some research on modeling what is called *context awareness* in sound recognition. One group of studies focuses on estimating the context of an audio segment with varying classification techniques [8,30,31]. 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, [30]) and 84% (14 classes, [8]). Although these results are promising, the methods that are used have some attributes that make them less suitable for automatic sound event detection. Features that are used to classify an audio interval are assumed to represent information that is specific for a class, and therefore, the context class to which an audio interval belongs gives primarily information about its acoustic properties. Tasks in multimedia applications (or a comparable setup in environmental sound classification, as in [8]) generally entail that a small audio interval, typically not longer than a few seconds, is classified as a sample of one context out of a dataset with a limited set of distinct contexts, which are stored as a collection of audio files. 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 [32,33]. For example, in [32] the intervals are clustered based on the similarity. Our approach for event detection will include a step of context recognition by classifying short intervals, before the main step of event detection.

3 Event detection

This section explains the sound event detection approach used in the proposed method, which recognizes and temporally locates sound events in recordings. In Section 4, this approach is extended to use context-dependent information.

3.1 Event models

The coarse shape of the power spectrum of the recording from the auditory scene is represented with MFCCs. They provide a good discriminative performance with reasonable noise robustness. In addition to the static coefficients, their first and second time derivatives are used to describe the dynamic properties of the cepstrum.

Sound-event-conditional feature distributions are modeled using continuous-density HMMs. Left-to-right model topology having three states was chosen to represent sound events having a beginning, a sustained part, and an end part. A mixture of multivariate Gaussian density functions is used in modeling the probability density functions of observations in each state. The acoustic models are trained using audio signals where the start and end times of events as well as their classes have manually been annotated. The traditional approach would be to use non-overlapping sound events to train the acoustic event models. However, realistic auditory scenes are usually too complex to provide enough such material for reliable training. Thus, each event instance annotated represents one training sample for the model of the event class regardless whether there were overlapping events present or not. The regions of the sound that contain overlapping events are used as training instances of both event classes when training the models. The assumption behind this procedure is that in the model training stage the variability caused by overlapping sound events classes will average out and the models will learn a reliable representation of the target sound events. The procedure of assigning training material to the event classes is illustrated in Figure 1. The models for sound events are trained with these samples using the Baum–Welch algorithm [34].

In the testing stage, the sound event models are connected into a network with transitions from each model to any other. A model network is shown in Figure 2.

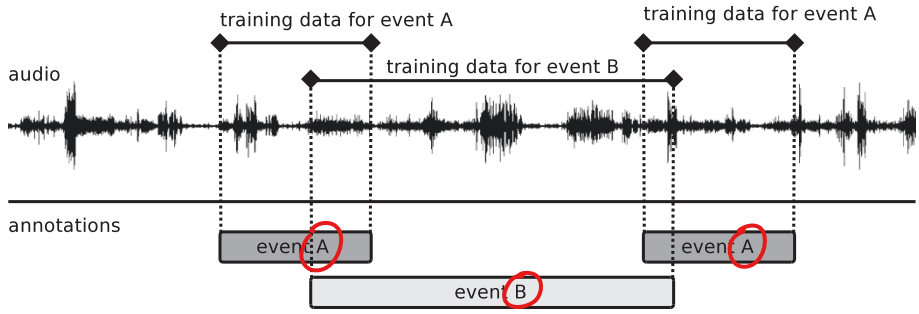

Figure 1 Training material containing overlapping sound events is used to train both sound event models.

Since it is possible that a test recording will contain some sound events which were not present in the training set, the system has to be able deal with such situations. A universal background model (UBM) is often used in speaker recognition to capture general properties of the signal [35]. We are using a UBM to capture events which are unknown to the system. A one-state HMM is trained with all available recordings for this purpose.

3.2 Count-based priors

Equally probable events can be represented by a network with equal inter-model transition probabilities. In this case, the output will be an unrestricted sequence of relevant labels, in which any event can follow any other.

In reality, the sound events are not uniformly distributed. Some events are more likely than others, e.g., speech is more common than a car alarm sound. If we regard each event as a separate entity and model the event counts, the histogram of the event counts inside certain context will provide us event priors. The event priors can be used to control event transitions inside the sound event model network shown in Figure 2. The count-based event priors are estimated from the annotated training material.

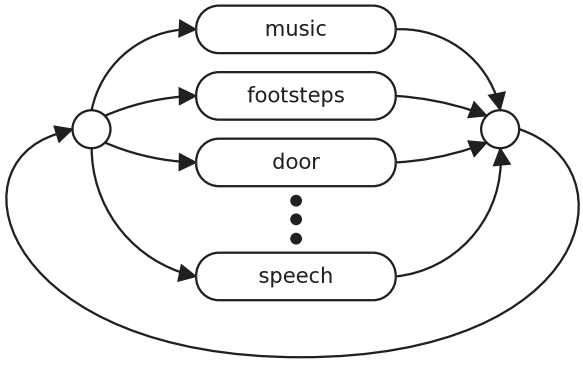

Figure 2 Fully connected sound event model network.

3.3 Detection

We will present two alternative approaches for the sound event detection: in the first one, we find the most prominent event at each time instance, and in the second one we find a predefined number of overlapping events. The detection of the most prominent event will produce a monophonic event sequence as an output. This approach is later referred as *monophonic* detection. The detection of overlapping events will produce a polyphonic event sequence as an output. This approach is later referred as *polyphonic* detection. Examples of the outputs of these two approaches are shown in Figure 3.

3.3.1 *Monophonic detection*

Segmentation of a recording into regions containing the most prominent event at a time will be obtained by doing Viterbi decoding [36] inside the network of sound event models. Transitions between models in this network are controlled by event prior probabilities. The balance between the event priors and the acoustic model is adjusted using a weight in combining the two likelihoods when calculating the path cost through the model network. A second parameter, insertion penalty, controls the number of events in the event sequence by controlling the cost of inter-event transition. These parameters are experimentally chosen using a development set.

3.3.2 *Polyphonic detection*

As discussed in Section 2, the previous studies related to sound event detection consider audio scenes with overlapping events that are explicitly annotated, but the detection results are presented as a sequence that is assumed to contain only the most prominent event at each time. In this respect, the systems output only one event at each time, and the evaluation considers the output correct if the detected event is one of the annotated ones. The performance of such systems is very limited in the case of rich multisource environments.

In order to detect overlapping events, we propose to use consecutive passes of the Viterbi algorithm as proposed in [37] for the detection of overlapping musical

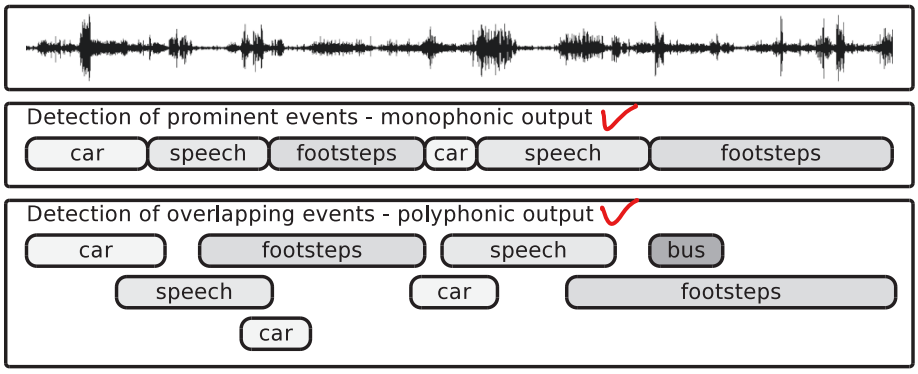

Figure 3 Example of sound event detection output for two approaches: monophonic system output and polyphonic system output.

notes. After one iteration, the decoded path through the model network is marked and the next iteration is prohibited from entering any states belonging to the sound event decoded at that frame in the previous iteration. The UBM is allowed in each iteration. This method will provide iterative decoding of the next-best path containing events that are at each time different than in the previously decoded one. This is difficult to achieve with conventional N-best decoding, which provides too many paths that have only minor state changes between them. These state changes do not produce the desired outcome. The proposed approach is illustrated in Figure 4. The number of iterations is chosen depending on the expected polyphony of the acoustic material.

4 Context-dependent event detection

Many sound events are acoustically dissimilar across contexts, and in these cases usage of context-specific acoustic models should provide better modelling accuracy. Sound events also have context-dependent prior probabilities, and using more accurate prior probabilities should also increase detection accuracy. Thus, we propose a sound event detection system utilizing the context information. The proposed system has two stages. In the first stage, the recording is tested for audio context classification. The second stage is the event detection. Based on the recognized context label, a specific set of sound event models is selected and acoustics models trained with the context-dependent material are selected to be used in the detection stage. In addition to this, context-dependent event priors are applied in the event detection. The system overview is presented in Figure 5. The details of each stage will be presented in the following sections.

4.1 Context recognition

As discussed in Section 2, an audio context can be recognized robustly among a small and restricted set of context classes. For our system, we chose a simple state-of-the-art context recognition approach [30] based on MFCCs and Gaussian mixture models (GMMs).

In the recognition stage, the audio is segmented into 4-second segments which are classified individually using the context models. Log-likelihoods are accumulated over all the segments and the model with the highest total

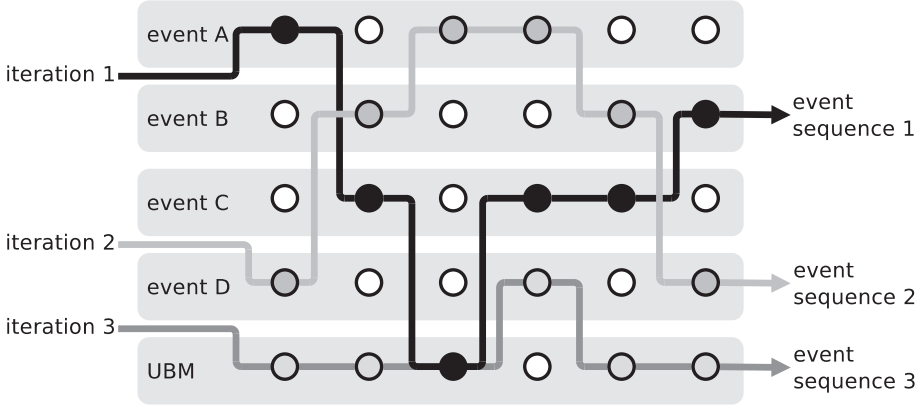

Figure 4 Concept of multiple path decoding using three consecutive passes of Viterbi algorithm.

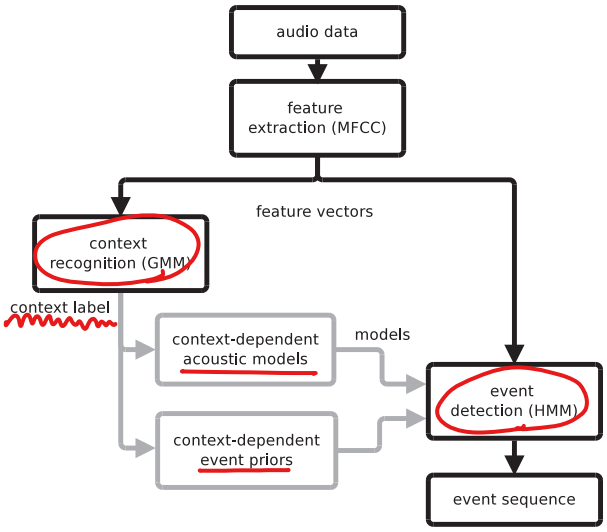

Figure 5 System overview.

likelihood is given as the label for the recording. The performance of context recognition will influence the performance of the sound event detection, as incorrectly recognized context will lead to choosing a wrong set of events for the event detection stage. Results for the context recognition are presented in Section 6.1.

The context models used in the context recognition stage are essentially identical to the context-dependent UBMs later used in the event detection stage. This simplifies the training process of the whole system and speeds up the event detection process allowing the calculated observation probabilities to be shared between stages.

4.2 Context-dependent modeling

In order to have more accurate modeling, the acoustic models for sound events are trained within each available context. Context-dependent count-based priors for the sound events are collected from the annotations of training material.

In the testing stage, the set of possible sound events is determined by the context label provided by the context recognition stage. The sound event models belonging to the recognized context will be selected and connected into a network with transitions from each model to any other (see Figure 2). The transitions between events are controlled with count-based event priors estimated for the recognized context.

5 Evaluation setup

The sound event detection system was trained and tested using an audio database collected from various contexts. The system was evaluated using an established evaluation metric [38] and a new metric introduced for a better understanding of the overlapping event detection results.

5.1 Database description

A comprehensive audio database is essential for training context and sound event models and for estimating count-based event priors. To the best of the authors' knowledge, there are only two publicly available audio databases for sound event detection from auditory scene. The database used in CLEAR 2007 evaluation [38] contains only material from meeting rooms. The DARES-G1 database [39] published in 2009 offers a more diverse set of audio recordings from many audio contexts. Event annotations for this database have been implemented using free-from event labels. The annotations would require label grouping in order to make the database usable for the event detection. At the time of this study, there was not any multi-context database publicly available that could be used for the evaluation without additional processing, and we recorded and annotated our own audio database. Our aim was to record material from common everyday contexts and to have as representative collection of audio scenes as possible.

The recordings for the database were collected from ten audio contexts: basketball game, beach, inside a bus, inside a car, hallways, inside an office facility, restaurant, grocery shop, street, and stadium with track and field events. Hallways and office facility contexts were selected to represent typical office work environments. The street, bus, and car contexts represent typical transportation scenarios. The grocery shop and restaurant contexts represent typical public space scenarios, whereas the beach, basketball game, and track and fields event contexts represent examples of leisure time scenarios.

The database consists of 103 recordings, each of which is 10–30-min long. The total duration of recordings is 1133 min. Each context is represented by 8 to 14 recordings. The material for the database was gathered using a binaural audio recording setup, where a person is wearing the microphones in his/her ears during the recording. The recording equipment consists of a Soundman OKM II Klassik/Studio A3 electret microphone and a Roland Edirol R-09 digital recorder. Recordings were done using 44.1 kHz sampling rate and 24-bit resolution. In this study, we are using monophonic versions of the recordings, i.e., two channels are averaged to one channel.

The recordings are manually annotated indicating the start and end times of all clearly audible sound events in the auditory scene. Annotations were done by the same person responsible of the recordings; this ensured as detailed as possible annotations since the annotator had prior knowledge of the auditory scene. In order to help the annotation of complex contexts, like street, also a low-quality video was captured during the recording of audio to help the annotator recall the auditory scene while doing annotation. The annotator had the freedom to choose descriptor labels for the sound events. The event

labels used in the annotations were manually grouped into 61 distinct event classes. Grouping was done by combining labels describing essentially the same sound event, e.g., "cheer" and "cheering", or labels describing acoustically very similar event, e.g., a "barcode reader beep" and a "card reader beep". Event classes were formed from events appearing at least ten times within the database. More rare events were included in a single class labeled as "unknown".

Figure 6 illustrates the event classes and their frequencies of occurrence for different contexts in the database. Each context contains 9 to 16 event classes and many event classes appear in multiple contexts (e.g., speech). There are also event classes which are highly context specific (e.g., dishes, or referee whistle). As expected in a natural auditory scenes, the event classes are not well balanced. It can been seen that some events are context specific (e.g., pressure release noise in the bus context), while others are very common across different contexts (e.g., speech). The number of events annotated per context is presented in Table 1.

5.2 Performance evaluation

In order to provide comparable metrics to the previous studies [25-27], in the performance evaluations we are using two metrics also used in the CLEAR 2007 evaluation [38]. The CLEAR evaluation defines the calculation of the precision and recall for the event detection, and the balanced F-score is calculated based on these. This accuracy metric is later denoted by ACC. The CLEAR evaluation also defines a temporal resolution error to represent the erroneously attributed time. This metric is later denoted by ER. Exact definition of these metrics can be found in the evaluation guidelines [38].

For evaluating a system output with overlapping events, the recall calculated in this way is limited by the number of events the system can output, compared to the number of events that are annotated. As a consequence, even if the output contains only correct events, the accuracy for the event detection is limited by the used metric. The temporal resolution error represents all the erroneously attributed time, including events wrongly recognized and events missed altogether by the lack of sufficient polyphony in the detection. The two metrics are therefore complementary, and tied to the polyphony of the annotation. This complicates the optimization of the event detection system into finding a good balance between the two.

In order to tackle this problem and to have a single understandable metric for sound event detection, we propose a block-wise detection accuracy metric. The metric combines the correctness of the event detection with a coarse temporal resolution determined by the length of the block used in the evaluation.

The proposed block-wise metric will evaluate how well the events detected in non-overlapping time blocks coincide with the annotations. The detected events are regarded only at the block level. In the evaluations, we are using two block lengths: 1 (later denoted by A1) and 30 s (later denoted by A30). This metric is designed

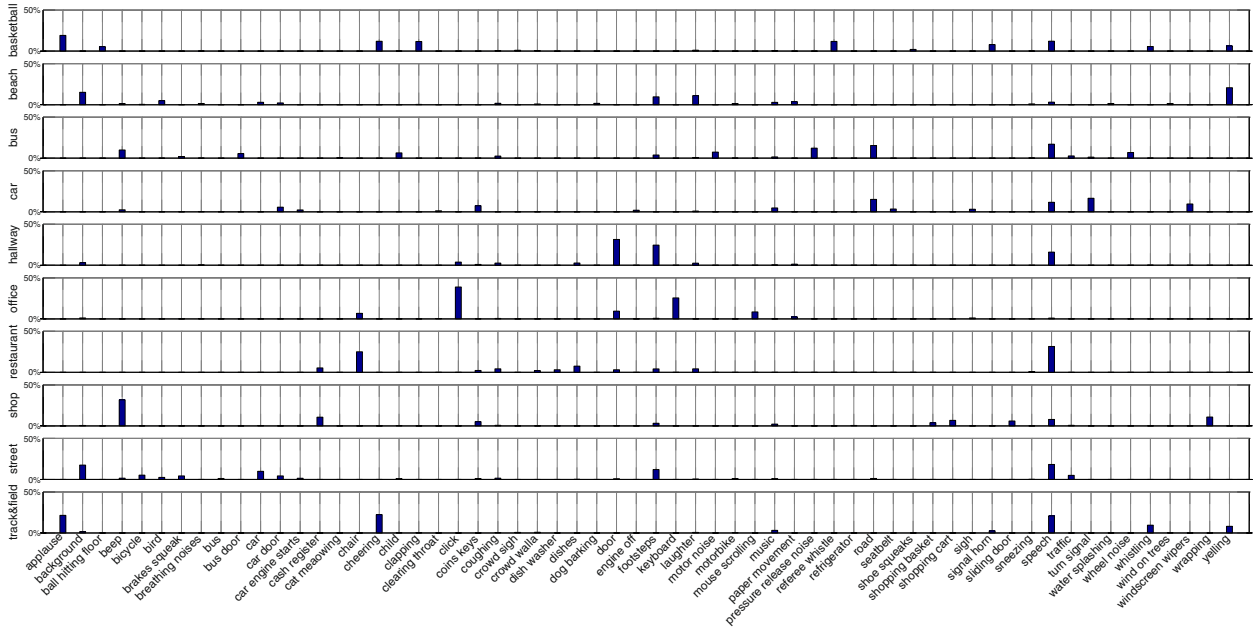

Figure 6 Percentage of sound event classes annotated per audio context in the database.

Table 1 Number of events annotated per context and total length of recordings (in minutes)

	Number of events	Length
Basketball game	990	80
Beach	738	197
Inside a bus	1729	146
Inside a car	582	111
Office facility	1220	105
Hallway	822	100
Restaurant	780	96
Grocery shop	1797	88
Street	827	102
Track & field stadium	793	108

for applications requiring a fairly coarse time resolution, placing more importance into finding the correct events within the block than finding their exact location. Inside the blocks, we calculate precision and recall. Precision is defined as the number of correctly detected sound event classes divided by the total number of event classes detected within the block. Recall is defined as the number of correctly detected sound event classes divided by the number of all annotated event classes within the block. An event is regarded as correctly detected if it has been detected somewhere within the block and the same event label also appears in the annotations within the same block. The accuracy represented by the F-score is calculated based on the precision and recall by the formula:

$$\text{Block accuracy} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

where

$$\text{Precision} = \frac{\text{Number of correct events}}{\text{Number of detected events}} \quad (2)$$

and

$$\text{Recall} = \frac{\text{Number of correctly detected events}}{\text{Number of all annotated events}}. \quad (3)$$

An illustration of how this metric works can be seen in Figure 7. In block 1, the annotated events are A, B, C, and D. The monophonic system output for the block 1 contains events A, C, and E. The events A and C are correctly detected by the system. For this block, the precision is 2 out of 3 (2/3) and recall 2 is out of 4 (2/4). The calculated block-wise accuracy for this block is 57.1% and the average block-wise accuracy for the entire example is 57.3%. For comparison, the CLEAR metrics are calculated on the level of entire output. The detection accuracy (ACC) is 76.2% having a precision 8/12 and recall 8/9. The time resolution error (ER) is calculated by counting the units that are wrongly labeled or missed altogether

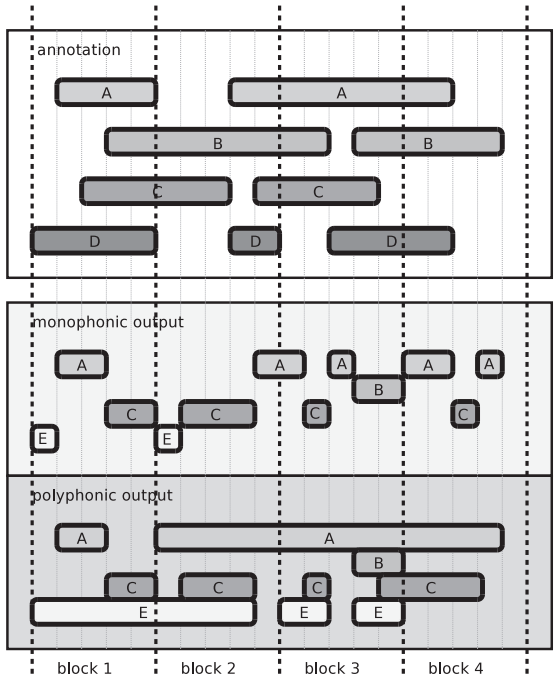

Figure 7 Block-wise accuracy for sound event detection.

(42) and dividing it with the total number of units (51) covered by the annotated events. This results in a 82.4% time resolution error.

For the polyphonic system output, the block-wise accuracy for the first block is 57.1% and the average accuracy for the entire example 58.3%. This is easily comparable with the same metric for the monophonic output. The CLEAR metric for the detection accuracy (ACC) is 63.2% (precision 6/10 and recall 6/9). The time resolution error (ER) is 109.8%, having 56 wrongly labeled or missed time units, compared to 51 in the annotation. This makes it hard to compare the monophonic and polyphonic outputs. In addition to this, an error value over 100% does not have proper interpretation. The proposed block-wise metric is comparable among monophonic and polyphonic outputs, with similar accuracy in the two illustrated cases. Therefore, the metric is equally valid for a system outputting only one event at time (monophonic output) as for a system outputting overlapping events (polyphonic output).

6 Experimental results

The database was split randomly into five equal-sized file sets, with one set being used as test data and other four for training the system. The split was done five times for a fivefold cross-validation setup. One fold was used in the development stage for determining parameters in the decoding. The evaluation results are presented as the average of the other four folds.

Table 2 Context recognition results

	4 s	20 s	40 s	Whole signal
Overall	**70.0**	**80.7**	**85.0**	**91.0**
Context-wise results				
Basketball	91.0	99.0	100.0	100.0
Beach	57.0	69.0	71.0	81.0
Bus	41.0	52.0	58.0	67.0
Car	84.0	93.0	95.0	100.0
Hallway	55.0	60.0	67.0	75.0
Office	85.0	87.0	88.0	88.0
Restaurant	77.0	89.0	95.0	100.0
Shop	72.0	87.0	94.0	100.0
Street	52.0	76.0	83.0	100.0
Track&Field	86.0	96.0	98.0	100.0

Percentage of correctly recognized segments.

Both the context recognition stage and the event detection stage used MFCC features and shared the same parameter set. MFCCs were calculated in 20-ms windows with a 50% overlap from the outputs of a 40-channel filterbank which occupied the frequencies from 30 Hz to half the sampling rate. In addition to the 16 static coefficients, the first and second time derivatives were also used.

In the event detection stage, the parameters controlling the balance between the event priors, the acoustic model, and the sequence length were experimentally chosen using a development set by finding parameter values which resulted in an output comprising approximately the same total amount of sound events that was manually annotated for the recording.

6.1 Context recognition

Context recognition was performed using the method presented in Section 4.1. The number of Gaussian distributions in the GMM model was fixed to 32 for each context class. This amount of Gaussian distributions was found to give a good compromise between computational complexity and recognition performance in the preliminary studies conducted with the development set.

The performance of the context recognition is presented in Table 2 as a fourfold average performance for the evaluation sets for four different segment lengths: 4 s, 20 s, 40 s, and the whole signal. Figure 8 shows the context recognition performance as a function of segment length used in the recognition. It can be seen that already after 2–3 min the system achieves a good recognition accuracy. A decision about the context could be taken already after the first minutes, in order to minimize the complexity of the context recognition stage and avoid processing the whole recording. However, we use the decision obtained after processing the whole signal to maximize the recognition accuracy. When using the whole length of the recording for the decision, six out of ten contexts have perfect 100% recognition rate, and rest of the contexts have also reasonable good, around 80% recognition rate.

The performance could positively be affected by the fact that recordings for the same context were done around the same geographical location, e.g., along the same street. Thus, the training and testing sets might contain recordings around the same area having quite a similar auditory scene, leading to over-optimistic performance.

6.2 Monophonic event detection

First we study the accuracy of the proposed system to find the most prominent event at each time instance. Since the performance of the context recognition stage affects on the selected event set for the event detection, the system is first tested when provided with the ground-truth context label. This will provide us the maximum attainable performance of the monophonic event detection. Later the system is evaluated in conjunction with the context recognition stage to provide a realistic performance evaluation. The system is evaluated using either uniform event priors or count-based event priors.

The number of Gaussian distributions per state in the sound event models was fixed to 16 for each event class.

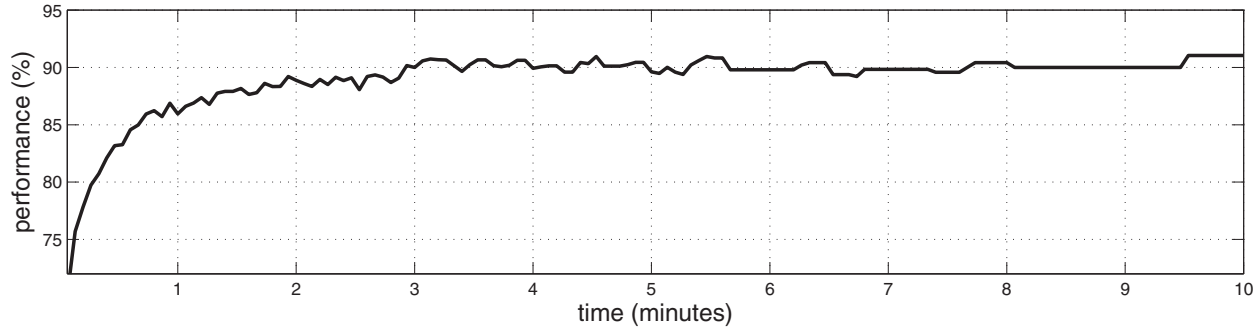

Figure 8 Context recognition performance as a function of used segment length.

This was found to give a high enough accuracy in the preliminary studies using the development set.

All the results are calculated as an average of the four test sets. The results are evaluated first with the CLEAR metrics (ACC and ER) in order to provide a way to compare the results to those of previously published systems [27,29]. In addition to this, block-wise accuracy is presented for two block lengths: 1 (denoted by A1) and 30 s (denoted by A30).

6.2.1 Event detection with the ground-truth context

The system is evaluated first using global acoustic models and then context-dependent acoustic models. At the same time also count-based event priors are evaluated. Event detection results using given ground-truth context labels are presented in Table 3.

The context-dependent acoustic models provide better fitting modeling and this is shown by the consistent increase in the results. Using the count-based event priors increases the system performance in the event detection for most of the contexts in both metrics. The overall accuracy increases from 34.7 to 41.1 while the time resolution error decreases from 86.9 to 83.4. The performance increase is reflected in the block-wise metric with an increase from 10.9 to 14.8 in 1-s block accuracy and 27.0 to 31.2 for 30-s block accuracy.

6.2.2 Event detection with recognized context

The true performance of the system is evaluated using the two steps: context recognition is performed on the test recording and then a set of event models and event priors are chosen according to the recognized context. Event detection results using the proposed two-step system are presented in Table 4. For comparison, the results of a context-independent baseline system [27] is also presented.

The results of the two-step system are slightly lower than the ones presented in Table 3 with the ground-truth context label. This is due to the 9% error in the context recognition step. A wrongly recognized context will lead to choosing the wrong model set and event priors. Even so, the different contexts do contain some common events and some of those events are correctly detected.

Table 3 Monophonic event detection performance based on ground-truth context

	ACC	ER	A1	A30
Global acoustic models				
Uniform event priors	32.3	85.2	10.0	21.9
Count-based event priors	36.6	84.7	12.0	25.8
Context-dependent acoustic models				
Uniform event priors	34.7	86.9	10.9	27.0
Count-based event priors	41.1	83.4	14.8	30.2

Table 4 Monophonic event detection performance comparison with context-independent baseline system and context-dependent system using context recognition

	ACC	ER	A1	A30
Context-independent detection				
No priors, baseline system	28.3	87.0	8.4	17.8
Context-dependent detection				
Uniform event priors	33.8	87.8	10.9	27.0
Count-based event priors	40.1	84.2	14.6	29.8

6.3 Polyphonic event detection

Overlapping events are detected using consecutive passes of the Viterbi algorithm as explained in Section 3.3.2. The average polyphony of the recorded material was estimated based on the annotations, and based on this the number of Viterbi passes was fixed to four.

The system is evaluated first with the ground-truth context label to get the maximum attainable performance of the polyphonic event detection. Later the full system having the context recognition stage is evaluated in order to get the realistic performance evaluation. As discussed in Section 5.2, the CLEAR evaluation metrics are not sensible to be used for polyphonic system output, and only block-wise accuracies are presented. Results for overlapping event detection with ground-truth context labels and recognized context labels are presented in Table 5.

6.3.1 Event detection with the ground-truth context

The consecutive passes of the Viterbi algorithm increase the event detection performance especially when measured on 1-s block-level. On longer 30 s block-level the performance difference is smaller between monophonic output and polyphonic output. The monophonic output can capture small segments of the overlapping events as they become more prominent than other events within

Table 5 Polyphonic event detection results and comparison with monophonic event detection system performance

	Ground-truth context		Recognized context	
	A1	A30	A1	A30
Monophonic system output				
Uniform event priors	10.9	27.0	10.9	27.0
Count-based event priors	14.8	30.2	14.6	29.8
Polyphonic system output				
Uniform event priors	19.8	28.9	18.9	28.2
Count-based event priors	20.4	30.0	19.5	29.4

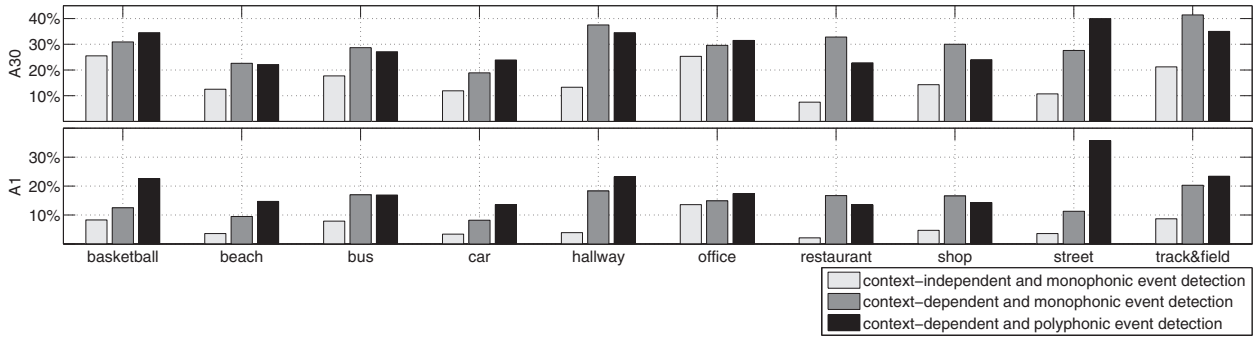

Figure 9 Context-wise results for context-independent baseline system and proposed context-dependent sound event detection systems.

the block. This way the monophonic system can detect many of the overlapping sound events on longer blocks.

6.3.2 Event detection with recognized context

The true performance of the system is evaluated using the context recognizer to get the context label for the test recording. The differences in the performance between the monophonic and polyphonic detection are quite similar to the detection where the true context was given. A slight overall performance decrease is due to the contexts which are not recognized 100% correctly (see Table 2).

6.4 Discussion

The context-dependent sound event detection substantially improves the performance compared to the context-independent detection approach. The improvement is partly due to the context-dependent event selection, and partly due to more accurate sound event modeling within the context. The event selection simplifies the detection task by reducing the number of sound events involved in the process. A context-dependent acoustic model represents particular characteristics of the sound event specific to the context, and provides more accurate results. The two-step classification scheme allows the proposed system to be extended easily with additional contexts later. The training process has to be applied only for the new context to get the context model for the context classification and to get the sound event models for the event detection.

Analysis of the individual contexts reveals interesting performance differences between contexts. Selected context-wise results are presented in Figure 9. Results are presented for three different system configurations: the context-independent baseline system, context-dependent monophonic event detection system using count-based event priors, and context-dependent polyphonic event detection system using count-based event priors. The context-dependent sound event detection approach increases the accuracy on all the studied contexts, especially on the rather complex contexts like street and restaurant. On the other hand, some contexts, like basketball, beach, and office, do not benefit as much.

The proposed overlapping event detection approach provides equal or better performance than prominent event detection approach for most of the contexts. The multiple Viterbi passes increases the detection accuracy in the shorter 1-s blocks relatively more than in 30-s blocks. This property can be exploited when a more responsive detection is required. An impressive improvement of 23% units is achieved in the 1-s block-wise accuracy for the street context, which is probably the noisiest context. On the other hand, the contexts also having a complex auditory scene, the restaurant, and the shop have a slight decrease in the accuracy. Varying complexity per context, i.e., having a different amount of overlapping events present at different times, may require also a different amount of Viterbi passes to overcome this. Examples of the audio recordings used in the evaluations along with their manual annotations and automatically detected sound events are available at arg.cs.tut.fi/demo/CASAbrowser.

7 Conclusion

The benefits of using the context-dependent information in the sound event detection were studied in this article. The proposed approach utilizing the context information comprised a context recognition stage and a sound event detection stage using the information of the recognized context. The evaluation results show that the knowledge of context can be used to substantially increase the acoustic event detection accuracy compared to the context-independent baseline approach. The context information is incorporated in multiple ways into the system. The detection task is simplified by using context-dependent event selection and the acoustic models of the sound events are made more accurate within each context by using context-dependent acoustic modeling. The context-dependent event priors are used to model

event probabilities within the context. For example, the detection accuracy in the block-metrics is almost doubled compared to the baseline system. Furthermore, the proposed approach for detecting overlapping sound events increases the responsiveness of the sound event detection by providing better detection accuracy on the shorter 1-s blocks.

Auditory scenes are naturally complex, having usually many overlapping sound events active at the same time. Hence, the detection of overlapping sound events is an important aspect for more robust and realistic sound event detection system. Recent developments in the sound source separation provide interesting possibilities to tackle this problem. In the early studies, sound source separation has already proven to substantially increase the accuracy of the event detection [40]. Further, the event priors for the overlapping sound events are difficult to model because of high number of possible combinations and transitions between them. Latent semantic analysis has emerged as a interesting solution to learn associations between overlapping events [29], but the area requires more studying to apply it efficiently to the overlapping event detection.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere, Finland. [2]Nokia Research Center, Visiokatu 3, Tampere, Finland.

Received: 26 June 2012 Accepted: 4 December 2012
Published: 9 January 2013

References

1. AK Dey, Understanding and using context. Person. Ubiquit Comput. **5**, 4–7 (2001)
2. D Wang, GJ Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. (Wiley-IEEE Press, New York, 2006)
3. R Cai, L Lu, A Hanjalic, H Zhang, LH Cai, A flexible framework for key audio effects detection and auditory context inference. IEEE Trans. Audio Speech Lang. Process. **14**(3), 1026–1039 (2006)
4. M Xu, C Xu, L Duan, JS Jin, S Luo, Audio keywords generation for sports video analysis. ACM Trans. Multimed. Comput. Commun. Appl. **4**(2), 1–23 (2008)
5. Y Peng, C Lin, M Sun, K Tsai, in *IEEE International Conference on Multimedia and Expo, 2009. ICME 2009*. Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models (IEEE Computer Society, New York, NY, USA, 2009), pp. 1218–1221
6. A Härmä, MF McKinney, J Skowronek, in *IEEE International Conference on Multimedia and Expo*. Automatic surveillance of the acoustic activity in our living environment (IEEE Computer Society, Amsterdam Netherlands, 2005), pp. 634–637
7. S Ntalampiras, I Potamitis, N Fakotakis, in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '09*. On acoustic surveillance of hazardous situations (IEEE Computer Society, Washington, DC, USA, 2009), pp. 165–168
8. S Chu, S Narayanan, CCJ Kuo, Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009)
9. T Heittola, A Mesaros, A Eronen, T Virtanen, in *18th European Signal Processing Conference*. Audio context recognition using audio event histograms (Aalborg, Denmark, 2010), pp. 1272–1276
10. M Shah, B Mears, C Chakrabarti, A Spanias, in *2012 IEEE International Conference on Emerging Signal Processing Applications (ESPA)*. Lifelogging: archival and retrieval of continuously recorded audio using wearable devices (IEEE Computer Society, Las Vegas, NV, USA, 2012), pp. 99–102
11. G Wichern, J Xue, H Thornburg, B Mechtley, A Spanias, Segmentation, indexing, and retrieval for environmental and natural sounds. IEEE Trans. Audio Speech Lang. Process. **18**(3), 688–707 (2010)
12. AS Bregman, *Auditory Scene Analysis*. (MIT Press, Cambridge MA, 1990)
13. M Bar, The proactive brain: using analogies and associations to generate predictions. Trends Cogn. Sci. **11**(7), 280–289 (2007)
14. A Oliva, A Torralba, The role of context in object recognition. Trends Cogn. Sci. **11**(12), 520–527 (2007)
15. M Niessen, L van Maanen, T Andringa, in *IEEE International Conference on Semantic Computing*. Disambiguating sounds through context (IEEE Computer Society, Santa Clara, CA, USA, 2008), pp. 88–95
16. C Clavel, T Ehrette, G Richard, in *IEEE International Conference on Multimedia and Expo*. Events detection for an audio-based surveillance system (IEEE Computer Society, Los Alamitos, CA, USA, 2005), pp. 1306–1309
17. H Wu, J Mendel, Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers. IEEE Trans. Fuzzy Syst. **15**, 56–72 (2007)
18. L Atlas, G Bernard, S Narayanan, Applications of time-frequency analysis to signals from manufacturing and machine monitoring sensors. Proc. IEEE. **84**(9), 1319–1329 (1996)
19. S Fagerlund, Bird species recognition using support vector machines. EURASIP J. Appl. Signal Process. **2007**, 64–64 (2007)
20. F Kraft, R Malkin, T Schaaf, A Waibel, in *Proceedings of Interspeech*. Temporal ICA for classification of acoustic events in a kitchen environment (International Speech Communication Association, Lisboa, Portugal, 2005), pp. 2689–2692
21. J Chen, AH Kam, J Zhang, N Liu, L Shue, in *Pervasive Computing*. Bathroom activity monitoring based on sound (Springer, Berlin, 2005), pp. 47–61
22. A Temko, C Nadeu, Classification of acoustic events using SVM-based clustering schemes. Pattern Recognit. **39**(4), 682–694 (2006)
23. TH Dat, H Li, in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Probabilistic distance SVM with Hellinger-exponential kernel for sound event classification (IEEE Computer Society, Prague, Czech Republic, 2011), pp. 2272–2275
24. R Stiefelhagen, R Bowers, J(eds) Fiscus, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*. (Springer, Berlin Germany, 2008)
25. X Zhou, X Zhuang, M Liu, H Tang, M Hasegawa-Johnson, T Huang, in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*. HMM-based acoustic event detection with AdaBoost feature selection (Springer, Berlin, Germany, 2008), pp. 345–353
26. X Zhuang, X Zhou, MA Hasegawa-Johnson, TS Huang, Real-world acoustic event detection. Pattern Recognit. Lett. (Pattern Recognition of Non-Speech Audio). **31**(12), 1543–1551 (2010)
27. A Mesaros, T Heittola, A Eronen, T Virtanen, in *18th European Signal Processing Conference*. Acoustic event detection in real-life recordings (Aalborg, Denmark, 2010), pp. 1267–1271
28. M Akbacak, JHL Hansen, Environmental sniffing: noise knowledge estimation for robust speech systems. IEEE Trans. Audio Speech Lang. Process. **15**(2), 465–477 (2007)
29. A Mesaros, H Heittola, A Klapuri, in *19th European Signal Processing Conference*. Latent semantic analysis in sound event detection (Barcelona, Spain, 2011), pp. 1307-1311
30. A Eronen, V Peltonen, J Tuomi, A Klapuri, S Fagerlund, T Sorsa, G Lorho, J Huopaniemi, Audio-based context recognition. IEEE Trans. Audio Speech Lang. Process. **14**, 321–329 (2006)
31. JJ Aucouturier, B Defréville, F Pacher, The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. J. Acoust. Soc. Am. **122**(2), 881–891 (2007)
32. R Cai, L Lu, A Hanjalic, Co-clustering for auditory scene categorization. IEEE Trans. Multimed. **10**(4), 596–606 (2008)
33. L Lie, A Hanjalic, Text-like segmentation of general audio for content-based retrieval. IEEE Trans. Multimed. **11**(4), 658–669 (2009)
34. LR Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. **77**(2), 257–286 (1989)

35. D Reynolds, R Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**, 72–83 (1995)
36. GD Forney, The Viterbi algorithm. Proc. IEEE. **61**(3), 268–278 (1973)
37. M Ryynänen, A Klapuri, in *Proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Polyphonic music transcription using note event modeling (IEEE Computer Society, New York, NY, USA, 2005), pp. 319–322
38. A Temko, C Nadeu, D Macho, R Malkin, C Zieger, M Omologo, in *Computers in the Human Interaction Loop*, ed. by AH Waibel, R Stiefelhagen. Acoustic event detection and classification (Springer, New York, 2009), pp. 61–73
39. M Grootel, T Andringa, J Krijnders, in *Proceedings of the NAG/DAGA Meeting 2009*. DARES-G1: database of annotated real-world everyday sounds (Rotterdam, Netherlands, 2009), pp. 996–999
40. T Heittola, A Mesaros, T Virtanen, A Eronen, in *Workshop on Machine Listening in Multisource Environments, CHiME2011*. Sound event detection in multisource environments using source separation (Florence, Italy, 2011), pp. 36–40

doi:10.1186/1687-4722-2013-1
Cite this article as: Heittola *et al.*: Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:1.