ACOUSTIC EVENT DETECTION IN REAL LIFE RECORDINGS

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

[1]Department of Signal Processing
Tampere University of Technology
Korkeakoulunkatu 1, 33720, Tampere, Finland
email: annamaria.mesaros@tut.fi, toni.heittola@tut.fi,
tuomas.virtanen@tut.fi

[2]Nokia Research Center
P.O.Box 100, FIN-33721 Tampere, Finland
email: antti.eronen@nokia.com

ABSTRACT

This paper presents a system for acoustic event detection in recordings from real life environments. The events are modeled using a network of hidden Markov models; their size and topology is chosen based on a study of isolated events recognition. We also studied the effect of ambient background noise on event classification performance. On real life recordings, we tested recognition of isolated sound events and event detection. For event detection, the system performs recognition and temporal positioning of a sequence of events. An accuracy of 24% was obtained in classifying isolated sound events into 61 classes. This corresponds to the accuracy of classifying between 61 events when mixed with ambient background noise at 0dB signal-to-noise ratio. In event detection, the system is capable of recognizing almost one third of the events, and the temporal positioning of the events is not correct for 84% of the time.

1. INTRODUCTION

Audio streams, such as broadcast news, meeting recordings, and personal videos contain sounds from a wide variety of sources. Examples include audio events related to human presence, such as speech, laughter, or coughing, or to sounds of animals, objects, nature, or situations. The detection of these events is useful, e.g., for automatic tagging in audio indexing, automatic sound analysis for audio segmentation or audio context classification.

An audio context or scene is characterized by the presence of individual sound events. In this respect, we may want to manage a multi-class description of our audio or video files by detecting the categories of sound events which occur in a file. For example, one may want to tag a holiday recording as being on the "beach", playing with the "children" and the "dog", right before the "storm" came. These are different level annotations, and while the beach as a context could be inferred from acoustic events like waves, wind, and water splashing, the audio events "dog barking" or "children" should be explicitly recognized, because such acoustic event may appear in other contexts, too.

The goal of this paper is to present an event detection system for a large and complex dataset. Previous related work includes audio scene recognition [1, 2, 3], analysis of video sound tracks [4, 5], and acoustic event detection [6]. Earlier work commonly considers only a rather limited number of audio events in a small set of audio environments. The work presented in this paper extends the event detection task to a comprehensive set of event-annotated audio material from everyday environments. We consider the task of recognizing and locating audio events in polyphonic long recordings. We use the term "polyphonic" for denoting recordings in which there are overlapping events, and at one instant of time there is no limitation for the number of event sound sources that can be present.

Our experiments comprise three parts. First, a study of the effect of hidden Markov model (HMM) size and topology for classification performance is performed using a database of isolated audio events. On the same database, we study the effect of the polyphony by adding environmental noise in different signal-to-noise ratios. The environmental noise is selected from a collection of appropriate ambient noises where other similar events can be present to create a realistic polyphonic fragment. Similar classification experiments are also run on real-life recordings, with the purpose of classifying the most prominent audio event in segments of various sizes. The test segments are provided by manual annotation, as it will be explained later. A final experiment is the detection of audio events in long recordings, which includes recognition and temporal positioning of a sequence of events within the recording.

The paper is organized as it follows: Section 2 presents an overview of audio scene recognition and event detection studies we find relevant to our work. Section 3 presents the tests covering isolated sound event classification. Section 4 describes the final choice for the recognition system stucture, the database of real life recordings and the experimental results in classifying and detecting audio events in the recordings. Section 5 presents discussion and conclusions and the orientation towards future work.

2. PREVIOUS WORK

Most of the previous work classifies an audio signal into one of predefined classes using standard features such as mel-frequency cepstral coefficients (MFCC) and classifiers such as hidden Markov models (HMM) or Gaussian mixture models (GMM). In [3], authors compared various features and classifiers in classifying between 24 everyday contexts, such as restaurant, car, library, and office. The system used MFCCs and their first-order time derivatives as features and HMMs with discriminative training for classification. The authors also conducted a listening test to compare the system's performance to the human abilities. The average recognition accuracy of the system was 58%, against 69% obtained in the listening tests, in recognizing between 24 everyday contexts. The accuracies in recognizing six high-level classes were 82% for the system and 88% for the humans.

The work in [7] deals with direct audio context recognition. Individual events are considered to be characteristics of the audio scene, and are not modeled themselves, but included in models of the contexts. The events and contexts are chosen such that to minimize overlapping. The authors present results for classifying 14 different contexts using MFCCs and matching pursuit features, using fixed length segments in training and testing.

In [2], the authors propose unsupervised clustering of interesting events recorded automatically in an office environment. The "interesting" events are detected by continuous monitoring of background noise and then clustered into discrete categories using unsupervised k-means. Authors of [4] propose a framework for detection of key audio effects in a continuous stream. They use 10 audio effects, distinct enough to be perceived, modeled using HMMs with parameters trained using isolated audio effects from Web, and decode the optimal sequence using the Viterbi algorithm.

Acoustic information is used also for finding interesting segments of video in video content analysis. Authors of [5] present an audio keyword generation system for sports videos based on audio. They use HMMs for classifying semantic events and a support vec-

[1]This work was financially supported by the Academy of Finland.

 ISSN 2076-1465

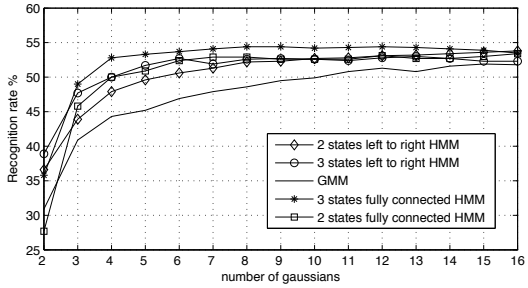

Figure 1: Isolated events classification performance for different size and type models.

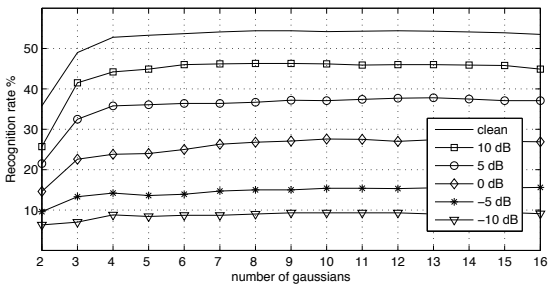

Figure 2: Isolated event classification performance under varying SNR conditions.

tor machine (SVM) classifier for finding audio keywords in soccer, basketball and tennis videos. Audio event detection can find a use also in healthcare monitoring for elderly people [8] or audio-based surveillance [1].

Efforts on acoustic events detection are presented in the CHIL project in their CLEAR evaluation [6]. The goal of the acoustic event detection task is to detect and recognize a closed set of pre-defined acoustic events. The evaluation data consisted of overlapping acoustic events occurring in the CHIL lecture and meeting corpus. Participants to the CLEAR evaluation proposed 5 systems based on HMMs and one on SVMs; the best performing system used HMMs and AdaBoost for feature selection[9]. Our proposal consisted of fully connected HMMs, using MFCCs and optimal path search decoded using the Viterbi algorithm [10].

Despite the research done so far, reliable detection and categorization of audio events from everyday audio is not mature enough for practical applications, such as automatic indexing of video sound tracks. The presented research contributes to the field by presenting a detailed evaluation of an HMM-based event detection system on a realistic and diverse set of audio material.

3. ISOLATED EVENTS CLASSIFICATION

In order to select the appropriate size and type of audio event models, we performed preliminary tests for isolated sound recognition. For this, a collection of isolated sound effects was selected from the Stockmusic online sample database [1], and organized into 61 classes. This database contains a total of 1359 samples belonging to 9 different contexts: crowd, hallway, household, human, nature, office, outdoors, shop, vehicles.

Samples from these classes were randomly selected either to the training set (70%) or to the testing set (30%). The training and testing set randomization was done five times and the average performance was calculated. Isolated event recognition was implemented for the 61 event classes, using MFCC based features and HMMs. We chose the same parametrization method as in [10]. Sixteen MFCCs were extracted from 20 ms long Hamming-windowed frames with 50% frame overlap and 40 mel-bands spanning the frequency range up to the Nyquist frequency were simulated in the frequency domain. The zeroth order coefficient was discarded. In addition to the static MFCC coefficients, we appended the first and second time derivatives. Using these features, an HMM was trained for each audio event class using the Expectation-Maximization (EM) algorithm. In the classification stage, the likelihood of each HMM producing the test observation sequence was obtained using the Viterbi algorithm, and the event was selected as the one corresponding to the HMM giving the largest likelihood.

Figure 1 presents the recognition rates for different size and type of HMMs and number of gaussians per state. At a sufficiently high number of gaussians per state, the system attains its maximum possible performance for the task, which in our case is 54% for 61 events. We also tried adjusting the number of states according to the average length of the audio events; this did not result in higher performance. Most of the fully-connected models became diagonalized during the training. Based on the simulations, it appears that a three-state left-to-right HMM with 4 to 16 mixture densities per state is a good choice for modeling audio events.

We conducted an additional study of how the environment richness influences the recognition of events. To simulate a natural polyphonic environment, we studied the effect of different signal-to-noise ratios, the signal being the event to be recognized and "noise" being selected from a database of ambient noises [2]. Ambient noise samples were chosen from the same 9 context classes as the sound effects. The background samples were randomly selected for each sound effect from the same context to which the event belongs, and the same background sample was used for the different SNR-cases. The results of sound effects classification under varying SNR conditions is presented in Figure 2 for a three-state HMM as a function of the number of gaussians per state. It can be observed how the performance decreases considerably with the introduced polyphony. This happens also in everyday life; when the acoustic power of the environmental noise is too high compared to individual events, we simply do not hear or recognize them anymore.

4. EVENT DETECTION IN REAL LIFE RECORDINGS

In the event detection in real life recordings, two tasks are evaluated: classification of isolated events in polyphonic recordings and detection of events in continuous sequences. For classification of isolated events, the test data provided to the recognizer consists of a short segment of audio containing one specific event, but the segment can have a rich content meaning that other events may also be present on the duration of the target event to be recognized. This task is similar to the SNR experiments from Section 3. In the acoustic event detection, the system also needs to temporally position the events. The test data consists of an entire track, and the system performs segmentation and classification simultaneously.

4.1 System description

The system for event detection consists of 61 event class models represented by three-state left-to-right HMMs with 16 gaussians per state. The set of features used for constructing the models are the MFCCs. The parameterization was the same as in Section 3.

For event classification, the class corresponding to the model resulting in the largest likelihood for the test observation sequence is chosen as recognition result. For event detection, the 61 models are connected into a network HMM, having equal transition probabilities from one event model to another. The detection task output is an unrestricted sequence of the 61 models, where any model can follow any other and there is no limit for the number of events. The optimal sequence of events is decoded using the Viterbi algorithm. The output of the system contains the timestamps for the recognized

[1] http://stockmusic.com/

[2] http://www.sound-ideas.com/

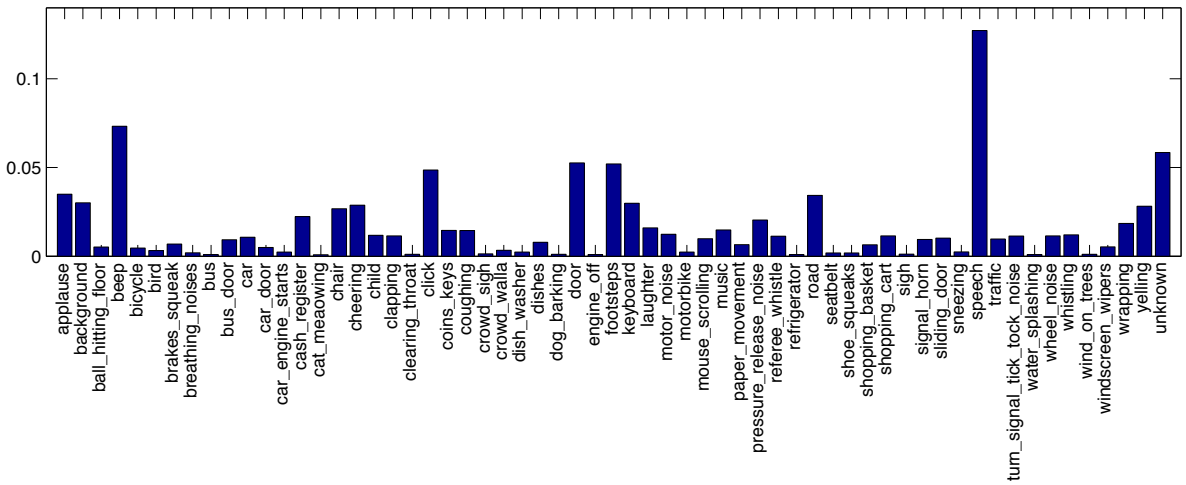

Figure 3: Count-based probabilities for the event classes calculated for the entire database. The histogram is dominated by "speech", as it is the most frequently annotated event, appearing in all the recorded contexts.

events, assuming that the system will indicate the most prominent event at a given polyphonic segment.

4.2 Database of real life recordings

For the modeling and recognition of acoustic events we collected long recordings (10 to 30 min each) from ten different acoustic environments (see the list in Table 1). All the recordings are made using a binaural setup, where a person is wearing the microphones in his ears during the recording. The recording equipment consists in a Soundman OKM II Klassik/studio A3 electret microphone and Roland Edirol R-09 wave recorder using 44.1 kHz sampling rate and 24bit resolution.

The events in the recordings were manually annotated by specifying the name and exact location (start and end time) of each audible event within the files. For each context there are 8 to 14 recordings, with a total of 103 recordings in the database. Within each context there are from 9 to 16 annotated event classes, totalling to 61 event classes, and there are many event classes appearing in multiple contexts. We formed distinct classes for events appearing at least 10 times, while more rare events are included in a class labeled as "unknown". Figure 3 illustrates the event classes and their frequencies of occurrence within the database. The classes are not balanced, some events are very frequent, while other are very common, as it is expected in a natural environment.

The data was split into non-overlapping training and testing sets such that in five folds all the material gets tested. Individual event instances as annotated are used for training. The features for one event instance were calculated directly from the polyphonic mixture, in the region of each track that was annotated as having that event present. In the case when more events appear simultaneously, the same part of the track (therefore the same observation vectors) was assigned to all the event classes present in that segment. The observations for individual events were used to construct models for each class. Table 1 presents information about the number of event instances extracted from each context.

4.3 Event classification

In this experiment we are interested in recognizing one event per presented test segment, considering that the system will identify the most prominent event in that segment. The experiments were performed in the described five fold setup. In this case, the test data is segmented into chunks containing one event, according to the annotated start and end times for each event instance. These segments

Table 1: Number of events extracted for each context of the recordings

basketball	990	beach	738
bus	1729	car	582
office	1220	hallway	822
restaurant	780	shop	1797
street	827	tracknfield	793

Table 2: Acoustic event classification evaluated using using one, two and three-best list

	one best	2-best	3-best
accuracy	23.8 %	35.4 %	44.1 %

are similar to the data used for training the event classes. In this respect, the task is isolated event classification, but with polyphonic audio, where other events may also be present on the duration of the target event to be recognized.

The average recognition accuracy is 23.8%, and some event classes have zero recognition rate. The confusion matrix is presented in Figure 4, and the recognition rates for individual classes are presented in Figure 5. There are cases when one event class is not present both in training and testing, thus we expect it to be wrongly classified, while in other cases there may be acoustic events that are more prominent for a given segment than the target one – for example water splashing is often recognized as wind on trees, which is a concurrent event in the beach recordings. To take into account the possibility of recognizing multiple superimposed events, we chose from one to three best scoring models for each tested file. The results of the experiments are presented in Table 2. The evaluation considers an event to be correctly recognized if its model is among one to three most likely models.

In the SNR experiments from Section 3, the recognition rates drop with approximately 10% every 5 dB. At the 0dB level, the concurrent background ambient noise has the same level as the acoustic event to be classified. At that value, the recognition rate is comparable with the results obtained for the real life recordings. This suggests that the level at which our annotator could still clearly hear

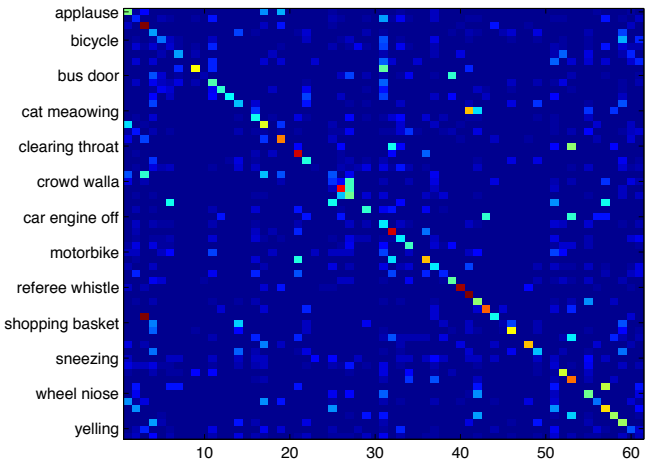

Figure 4: Confusion matrix for event classification. The labels presented in the figure represent every fifth event class in alphabetical order.

and annotate a distinct sound event is when the acoustic power of the power is approximately the same as the power of the event itself.

4.4 Event detection

As mentioned, for the event detection task, the optimal sequence of events is decoded using the Viterbi algorithm within the system HMM network, assuming that the system will indicate the most prominent event at a given time. The output contains the start and end times for the recognized events, marked as the points when the search path goes from one event model to another.

Prior knowledge of the events frequency of occurrence can be used in the detection. This information is presented as a normalized histogram of the event counts, as illustrated in Figure 3. These are prior probabilities for the event classes. The likelihoods of the event classes during recognition will be multiplied by their prior probabilities in order to determine a posterior probability that will then be used in the Viterbi search.

As a performance evaluation measure for the events detection we use the accuracy evaluation metric from the CLEAR 2007 evaluation. This metric is used to score detection of relevant acoustic events (AE). It does not take into account temporal coincidence of the annotated and system output timestamps. It is defined as the F-score (the harmonic mean between precision and recall). In the evaluation, the balanced F-score was used:

$$ACC = 2 * \frac{Precision * Recall}{Precision + Recall},$$

where

$$Precision = \frac{\text{number of correct system output AEs}}{\text{number of all system output AEs}}$$

and

$$Recall = \frac{\text{number of correctly detected reference AEs}}{\text{number of all reference AEs}}$$

The system output is considered correct if there exists at least one annotated sound event whose temporal centre is situated between the timestamps of the system output, and the annotated label and system output are similar, or if the temporal centre of the system output lies between the timestamps of at least one annotated event and the annotated label and system output are similar. The annotated sound event is considered correctly detected if there exists at least one system output whose temporal centre is situated between the timestamps of annotated sound event and the labels are similar, or if the temporal centre of the annotated sound event lies between the timestamps of at least one system output and the labels are similar. The results are presented in Table 3.

Table 3: Acoustic event detection evaluation results

system	Precision	Recall	Accuracy
no priors	38.9%	24.5%	30.1%
using priors	39.6%	24.2%	30.0%

Table 4: Acoustic event detection error

system	missed events	false alarms	substitutions	overall error
no priors	60.6%	1.4%	22.1%	84.1%
using priors	60.7%	1.4%	21.8%	84.0%

The temporal resolution of the detected acoustic events is scored using the metric for Speaker Diarization, adapted to the task of audio event detection in the CLEAR evaluation. A one-to-one mapping of the reference acoustic events to the acoustic events output by the system is computed, and the measure is the aggregation over all reference acoustic events of the time that is jointly attributed to both the reference and the corresponding system output acoustic event to which that reference events are mapped. This is computed over all audio segments, including regions of overlapping.

The overall error score ER will be computed as the fraction of the time that is not attributed correctly to an acoustic event:

$$ER = \frac{\sum_{seg}\{dur(seg) * max(N_{ref}, N_{sys}) - N_{correct})\}}{\sum_{seg}\{dur(seg) * N_{ref}\}}$$

where the audio data is divided into adjacent segments whose border coincide with the points where either a reference or a system output acoustic event starts or stops, so that for the given segment, the number of current reference AEs and the number of system output AEs do not change. For each segment seg, dur is the duration of the seg, N_{ref} is the number of reference AEs in seg, N_{sys} is the number of system output AEs in seg and $N_{correct}$ is the number of reference AEs in seg which have a corresponding mapped system output AEs in seg .

The overall detection error of the system and some details about the errors are presented in Table 4. The total amount of scored time is 920 min; this represents the added duration of all annotated events, being 2.5 times more than the actual time covered by overlapping events. The overall acoustic event detection error of the presented system for the 61 event classes is 84.1% of the total scored time.

Using the prior information based on overall events counts did not improve the results for event detection. Such direct count may not reflect the true probability of events in different contexts; because of averaging over all the contexts, the histogram in Figure 3 is dominated by "speech". Indeed, speech is present in all the contexts and it overlaps practically all other events, and also gets a lot of confusions in the classification.

In the audio events detection of the CLEAR evaluation, the best system score was 36.3% accuracy and 99.5% detection error. In comparison, our system has a lower detection error for a much higher number of classes, but the accuracy of recognition is lower.

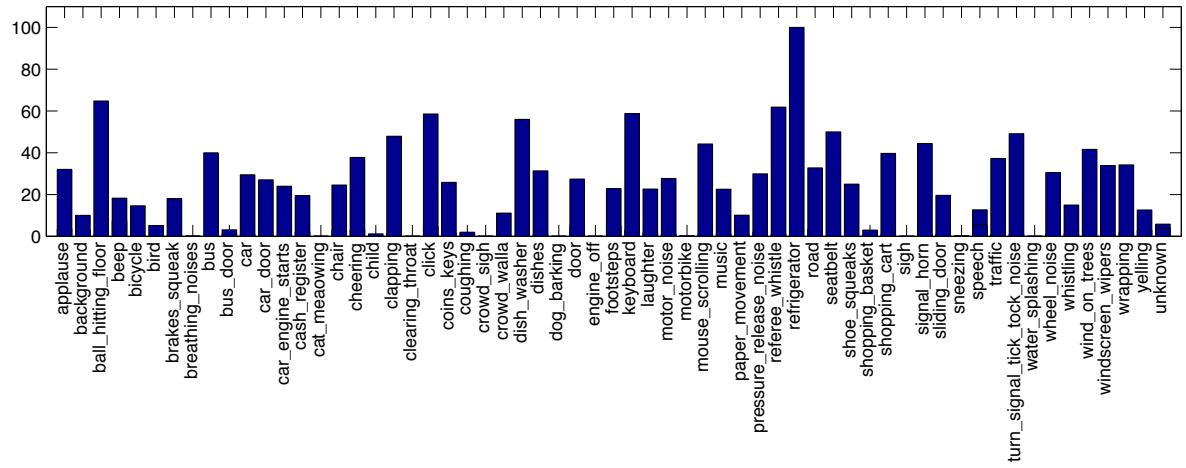

Figure 5: Event classification performance of individual classes

.

5. CONCLUSIONS

This paper presented a detailed evaluation of an HMM-based event detection and classification system using recordings of ten different natural environments. Three different tests were performed. A study of the topology and size of the selected models was performed on a database containing isolated audio events, obtaining a maximum performance of 54% for the three-state left-to right and fully-connected HMMs. Based on these results, we selected a three-state left-to-right model for the subsequent experiments. We performed a similar event classification task on the real-life recordings, obtaining a recognition performance of 24%. Similar performance was obtained in isolated events recognition with with background noise mixed at 0 db SNR, suggesting that this is the level where humans can clearly hear and annotate an audio event in a natural context. For detecting successive events in a long recording, the proposed system has an accuracy of 30% for 61 classes and a detection error of 84.1%. Using prior information based on overall event count did not bring any improvement. We think this is due to adding up all the events from different environments, which averages out the differences in count between events specific to certain environments. Our future work will consider e.g. using missing feature techniques for improving the event detection robustness in polyphonic mixtures. The current event detection system is used in an audio context recognition system based on acoustic events.

REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*, 2005, pp. 1306–1309.

[2] A. Härmä, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference onMultimedia and Expo*, Los Alamitos, CA, USA, 2005, vol. 0, p. 4 pp., IEEE Computer Society.

[3] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[4] R. Cai, L. Lu, A. Hanjalic, H-J. Zhang, and L-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1026–1039, 2006.

[5] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[6] Rainer Stiefelhagen, Rachel Bowers, and Jonathan Fiscus, Eds., *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Springer-Verlag, Berlin, Heidelberg, 2008.

[7] S. Chu, S. Narayanan, and C-C. Jay Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Transactions on Speech, Audio, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[8] Ya-Ti Peng, Ching-Yung Lin, Ming-Ting Sun, and Kun-Cheng Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *ICME'09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, Piscataway, NJ, USA, 2009, pp. 1218–1221, IEEE Press.

[9] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "Hmm-based acoustic event detection with adaboost feature selection," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Berlin, Heidelberg, 2008, pp. 345–353, Springer-Verlag.

[10] T. Heittola and A. Klapuri, "TUT acoustic event detection system 2007," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Berlin, Heidelberg, 2008, pp. 364–370, Springer-Verlag.