SOUND EVENT DETECTION AND CONTEXT RECOGNITION

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

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

[2]Nokia Research Center
Tampere, Finland
antti.eronen@nokia.com

1 INTRODUCTION

Humans can easily segregate and recognize one sound source from an acoustic mixture, and recognize a certain voice from a busy background which includes other people talking and music. Sound event detection and classification aims to process an acoustic signal and convert it into descriptions of the corresponding sound events present at the scene. This is useful, e.g., for automatic tagging in audio indexing, automatic sound analysis for audio segmentation or audio context classification.

An audio scene is characterized by the presence of individual sound events. In general, context can be defined as the state of the environment, the user, and the device. In our work, context means the environment, or acoustic ambiance around the recording device. We define audio context recognition is as the process of automatically determining the context using a recorded audio signal. Information about the surroundings would enable wearable devices to provide better service to users' needs, e.g., by adjusting the mode of operation accordingly. Early listening tests conducted in [1] showed that humans are able to recognize everyday auditory contexts in 70% of cases on average and confusions are mostly between contexts that have same types of prominent sound events. The study suggested that distinct sound events recognized from the auditory scene are a salient cue for human perception of audio context. However, most of the proposed context recognition systems are modeling global acoustic characteristics of the audio context rather than sound events [2, 3, 4].

Early work on sound event detection commonly has considered only a rather limited number of audio events in a small set of audio environments [5, 6, 7]. This is largely because the polyphony of the signals present a big challenge to automatic methods. Real world audio scenes contain multiple simultaneously occurring sound events, but work in sound event detection commonly considers detection of the most prominent event at each time.

This paper describes a method for sound event detection and its evaluation on a comprehensive set of event annotated audio material from everyday environments. Moreover, we present and evaluate a method for context detection using the event detection system. A system for sound event detection creates a description of the event content of a test recording, containing event labels and timestamps for each. For context recognition,

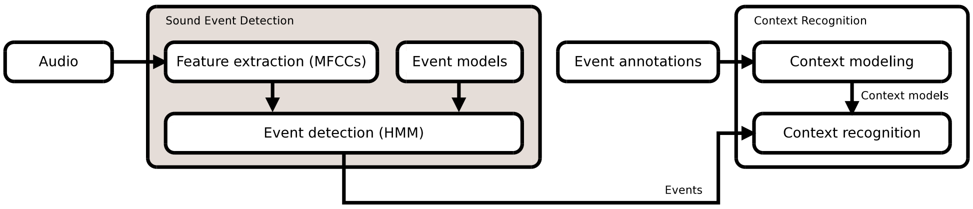

Figure 1: System overview.

this description is presented as an event histogram that will be matched with previously trained context models. The event detection system and context recognition are evaluated using recordings from ten different audio contexts that may contain the same kind of events.

2 SOUND EVENT DETECTION

The output of the sound event detection task is a sequence of the recognized event model labels and timestamps for each event. This works on the assumption that the system will indicate the most prominent event at each time in the polyphonic mixture.

The system is based on continuous density hidden Markov models (HMM). The coarse shape of spectrum is represented with short-term features. Mel-frequency cepstral coefficients (MFCC) provide a good discriminative performance with reasonable noise robustness. In addition to the static coefficients, first and second order time differentials are used to describe the dynamic properties of the cepstrum.

Manually annotated recordings with overlapping events are used for training a set of 61 event models, such as speech, laughter, applause, car door, road, dishes, door, chair, music, and footsteps. An audio segment where multiple events overlap is included in the training data of all the classes present in that segment. This means including the same observation vectors to train multiple event models. Sound event categories are modeled with three-state left-to-right HMMs. The probability density of each state is modeled using Gaussian mixture models (GMM) having 16 components. The sound event HMMs are connected into a single HMM with equal transition probabilities between the event models. Due to this, the output will be an unrestricted sequence of the 61 event models, where any event can follow any other and there is no limit for the number of events. The models are trained and tested using a database that will be described in Section 4.1.

The sound event detection stage represents the first block in the overall system presented in Figure 1. The features are extracted for the entire audio clip. Based on the input audio and the event models, the event detection stage uses Viterbi algorithm to output the most likely event sequence and their timestamps. A detailed explanation of this system can be found in [8]. A histogram representation of this result will be used for recognizing the context where the audio was recorded.

3 CONTEXT RECOGNITION

Our context recognition approach assumes that each context is characterized by the presence of certain sound events. Models of contexts are constructed by summing up event histograms of individual recordings. We represent a recording as an event occurrence histogram by counting all the annotated events in it. To prevent bias related to the length of the recording, the event counts in the histogram are divided by the total number of events present in the recording. The context model histogram is normalized so that the bins sum up to one.

For recognizing the context of a recording, a similar representation of the recording is constructed. The histograms are calculated based on the output of the detection system that uses Viterbi or based on the classifier output for segments of the recording. The context recognition is based on comparing this histogram with histogram models of contexts. The cosine distances are calculated between context models and the test recording and the closest match is selected as recognition result. A more detailed description of the context recognition system can be found in [9].

4 EVALUATIONS

The proposed sound event detection system and context recognition system are evaluated with an audio database collected from real-life environments. The database was split into non-overlapping training and testing sets such that in five folds all the material gets tested.

4.1 Database

The database consists of 103 recordings, each of which is 10 to 30 minutes long. The recordings are collected from ten audio contexts: basketball game, beach, inside a bus, inside a car, hallway, office, restaurant, grocery shop, street and stadium with track and field events. The audio is recorded using binaural microphones placed inside the human ears. 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 24 bit resolution. In this work, we are using monophonic versions of the recordings, i.e., two channels are averaged to one channel.

The sound events in the recordings were manually annotated indicating the start and end times of all clearly audible sound events. Repetitive sounds were annotated as long events (e.g. footsteps), while long events like conversation are annotated as multiple successive speech events if there is perceivable pause in the conversation. The annotations comprise a number of 61 event classes. Within each context there are from 9 to 16 annotated event classes. Some events classes appear in multiple contexts, others are context specific.

Table 1: Sound event detection accuracy, grouped by context.

Context	ACC	Context	ACC
basketball	49 %	office	47 %
beach	23 %	restaurant	20 %
bus	24 %	shop	29 %
car	25 %	street	16 %
hallway	27 %	track & field	38 %

4.2 Sound event detection

The performance of the sound event detection system is evaluated using the accuracy metric from the CLEAR 2007 evaluation [7]. This metric is used to score detection of relevant sound events without taking into account how exact is the temporal coincidence of the annotated and system output events. Accuracy is defined as the balanced F-score between precision and recall:

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

where precision is defined as number of correct system outputs divided by number of all system outputs and recall is defined as number of correctly detected reference events divided by number of all reference events.

Table 1 presents event detection results, grouped by audio contexts. The average accuracy of the event detection is 30 %. The best results are obtained in contexts with very specific sound events, like basketball (referee whistle, crowd cheering, announcer) and office (chair squeaks, typing, mouse clicks). Other contexts like beach, bus, restaurant and shop contain lots of common events related to human presence, especially speech. The worst results are obtained for the noisiest context, street.

4.3 Context recognition

Context recognition was performed by comparing the event count histogram of the test recording with histogram models of the contexts.

We tested a simple method for obtaining event counts for a recording, instead of Viterbi segmentation [9]. In this case, isolated event classification is performed over four second segments of the tested audio. The classifier provides the most likely event label for each segment. The events detected in the segments within the tested recording are collected to form an event histogram.

The second method for obtaining event counts is based on the detection system presented previously. In this case, the Viterbi algorithm provides the most likely event sequence for the entire recording and this sequence will be used to construct the histogram.

In addition to the event based context recognition, we evaluated a system based on acoustic information of contexts. We constructed a baseline system where each of the

Table 2: Context-wise average recognition accuracy.

	4 second segments	Viterbi segmentation
Baseline system	88.5 %	-
Event based system	88.5 %	84.5 %
Combined system	91.4 %	92.4 %

ten contexts is modeled with a GMM, using MFCCs as features. For this method, the test recordings are split into four second segments which are classified individually. A final decision for the entire recording is taken by accumulating context model likelihoods over the entire recording and choosing the higher scoring one.

The baseline system provides context information based on the global acoustic characteristics of the audio context. This is complementary information to the sound events, which represent details of the audio context, and a combination could lead to improved results. To combine the two systems, we map the distances from the event based method into probabilities using an inverted sigmoid-function. The obtained probabilities are then multiplied with the context likelihoods produced by the baseline system.

The results for the context recognition are presented in Table 2. The two methods of obtaining event histograms are marked as “4 second segments” and “Viterbi segmentation”. The baseline system (global characteristics) and the classification based histogram (event-based, in 4 second segments) perform equally well, while the detection based histogram obtains slightly lower results. The combination of the global and detailed characteristics in modeling improves the performance with few percent units.

The performance of the event based context recognition system is not superior to the baseline system. The event based context recognition system is more complex and requires long test segments to work properly. However, it gives complementary information (sound event labels) compared to a single context label assigned to the recording. The baseline system performs nicely with contexts which are acoustically distinguishable. Combining the event based system with the baseline system provides slightly better accuracy and robustness with acoustically similar contexts.

5 CONCLUSION

This paper presented an evaluation of an HMM-based sound event detection system using recordings of ten different natural environments. Detected events were further used to recognize the contexts of the tested recordings. The sound event detection achieved an accuracy of 30 %. This was found to be sufficient for the event based context recognition, allowing performance comparable to the approach using global acoustic characteristics of the recording for the context recognition. When combining the event based context recognition with the approach using global acoustic characteristics of the recording the overall performance was found to increase with few percent units, implying that these two context recognition approaches provide complementary information about the context.

6 ACKNOWLEDGEMENTS

This research has been funded by TEKES, Nokia Research Center and the Academy of Finland.

REFERENCES

[1] PELTONEN V, ERONEN A, PARVIAINEN M, & KLAPURI A, Recognition of everyday auditory scenes: Potentials, latencies and cues, in *In Proc. 110th Audio Eng. Soc. Convention*, Hall, 2001.

[2] ERONEN A, PELTONEN V, TUOMI J, KLAPURI A, FAGERLUND S, SORSA T, LORHO G, & HUOPANIEMI J, Audio-based context recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(2006) 1, 321–329, ISSN 1558-7916.

[3] MA L, MILNER B, & SMITH D, Acoustic environment classification, *ACM Trans. Speech Lang. Process.*, **3**(2006) 2, 1–22, ISSN 1550-4875.

[4] CHU S, NARAYANAN S, & KUO C C J, Environmental sound recognition with time-frequency audio features, *IEEE Trans. on Audio, Speech and Language Process.*, **17**(2009) 6, 1142–1158, ISSN 1063-6676.

[5] CAI R, LU L, HANJALIC A, ZHANG H J, & CAI L H, A flexible framework for key audio effects detection and auditory context inference, *IEEE Transactions on Audio, Speech and Language Processing*, **14**(2006) 3, 1026–1039.

[6] XU M, XU C, DUAN L, JIN J S, & LUO S, Audio keywords generation for sports video analysis, *ACM Trans. Multimedia Comput. Commun. Appl.*, **4**(2008) 2, 1–23, ISSN 1551-6857.

[7] STIEFELHAGEN R, BOWERS R, & FISCUS J, editors, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Springer-Verlag, Berlin, Heidelberg, 2008.

[8] MESAROS A, HEITTOLA T, ERONEN A, & VIRTANEN T, Acoustic event detection in real-life recordings, in *18th European Signal Processing Conference*, 2010.

[9] HEITTOLA T, MESAROS A, ERONEN A, & VIRTANEN T, Audio context recognition using audio event histograms, in *18th European Signal Processing Conference*, pages 1272–1276, Aalborg, Denmark, 2010.