AUDIO CONTEXT RECOGNITION USING AUDIO EVENT HISTOGRAMS

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

[1]Department of Signal Processing
Tampere University of Technology
Korkeakoulunkatu 1, 33720, Tampere, Finland
email: toni.heittola@tut.fi, annamaria.mesaros@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 method for audio context recognition, meaning classification between everyday environments. The method is based on representing each audio context using a histogram of audio events which are detected using a supervised classifier. In the training stage, each context is modeled with a histogram estimated from annotated training data. In the testing stage, individual sound events are detected in the unknown recording and a histogram of the sound event occurrences is built. Context recognition is performed by computing the cosine distance between this histogram and event histograms of each context from the training database. Term frequency–inverse document frequency weighting is studied for controlling the importance of different events in the histogram distance calculation. An average classification accuracy of 89% is obtained in the recognition between ten everyday contexts. Combining the event based context recognition system with more conventional audio based recognition increases the recognition rate to 92%.

1. INTRODUCTION

Context recognition is defined as the process of automatically determining the context around a device. Information about the surroundings would enable wearable devices to provide better service to users' needs, e.g., by adjusting the mode of operation accordingly. Compared to image or video sensing, audio has certain distinctive characteristics. Audio captures information from all directions and is relatively robust to sensor position and orientation, which allows sensing without troubling the user. Audio can provide a rich set of information which can relate to location, activity, people, or what is being spoken. The acoustic ambiance and background noise characterizes a physical location, such as inside a car, restaurant, or office.

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].

In this paper, we propose a context recognition system based on detection of individual acoustic events. Our approach assumes that different contexts, such as a street or a restaurant, are characterized by the occurrence of certain sound events. Contexts are modeled with event histograms collected from annotated recordings. The proposed system is divided into two stages, sound event detection and context recognition. A sound event detection system is used to detect sound events present in the tested context and the event histogram constructed from the recognition result is matched with context models. The system is evaluated with ten contexts that may contain the same events. The overall system scheme is presented in Figure 1.

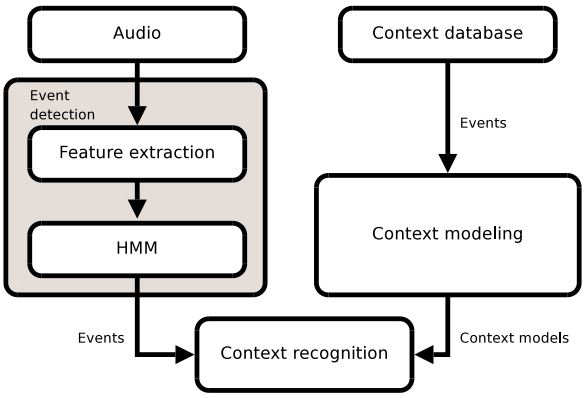

Figure 1: System overview.

The rest of this paper is organized as follows. Section 2 briefs the related work. Section 3 presents the event detection system, and Section 4 describes how detected events are used in the context recognition. Section 5 explains the context database used in the evaluation and the evaluation itself. Section 6 provides conclusions and suggestions for further study.

2. RELATED WORK

Automatic recognition of the context or environment based on audio information is known from many earlier works. However, most of the work on context recognition has been done by directly recognizing the context from the acoustic information, without explicitly detecting the individual sound events in the auditory scene. Eronen *et al.* [2] presented an approach to recognize 24 everyday context with mel-frequency cepstral coefficients (MFCC) and hidden Markov models (HMM). They reached a 58% recognition accuracy against 69% obtained in a human listening tests using the same material. The study in [3] presented an HMM-based environmental noise classification system and reported over 91% accuracy in classifying 10 contexts using three second test segments. The authors also performed a listening test on the same data. The listeners' performance for the three

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

seconds segments was significantly worse than the system performance. More recently, Chu *et al.* [4] proposed an approach using matching pursuit to select a small set of time-frequency features to represent each context. They achieved a 84% performance for 14 contexts for four second segments using these features jointly with MFCC. The used contexts were chosen to be as different as possible to minimize overlapping.

One of the approaches to use sound events in the context recognition was presented in [5]. The authors propose a framework for detection of key audio effects in a continuous stream. The optimal key effect sequence is determined using Viterbi decoding, controlled by a two-loop network defining possible transitions between sound effects. They use 10 audio effects, distinct enough to be perceived, with models trained using isolated audio effects from Web. The different audio effects are modeled using HMMs with 5 to 11 states per model, trained with various features. The authors treat overlapping events by using the label of the dominant one for that region. The detected audio effects are used to recognize the scene as one of 5 possible (non-overlapping) categories - humor, pursuit, etc. More recently, the authors proposed an unsupervised co-clustering approach for the same task [6]. Authors of [7] propose an audio keywords generation system for sports videos. Low-level features are extracted from audio and after off-line feature selection hierarchical SVM is used find audio keywords. Hidden Markov models are used to detect the semantic events in sports videos. The system was tested with soccer, basketball, and tennis videos.

Sound event detection from audio signals can be performed in an unsupervised or supervised manner. In the unsupervised approach, the categories of sound events are not specified beforehand but distinct portions of the audio signal are detected as potential events, e.g. via clustering [8]. In the supervised approach, predefined sound event classes are used to segment and classify sound events. In [9], we presented a sound event detection system for the meeting room environment using MFCC based features and a HMM classifier.

3. EVENT DETECTION

The sound event detection in the proposed context recognition system is based on continuous density HMMs and the audio signal power spectrum is represented with MFCCs. These short-term features represent the coarse shape of the spectrum and provide a good discriminative performance with reasonable noise robustness. The system uses 16 MFCCs calculated from the outputs of a 40-channel filterbank. In addition to the static coefficients, their first and second order time differentials are used to describe the dynamic properties of the cepstrum. Features are extracted in 20 ms frames with a 50% frame shift.

We train 61 HMMs to represent 61 sound event categories. Three-state left-to-right HMMs are trained with the standard Baum-Welch training procedure using a training database that will be described in Section 5.1. 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.

Manually annotated recordings with overlapping events were used for training the event models. 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. In the detection stage, features are extracted for the entire audio clip, and the event detection is organized in two ways. Event detection over the entire recording is done using the Viterbi algorithm to obtain the most likely event sequence. However, the order of the sound events will not be used in the context recognition. In addition to this, we use isolated event recognition over four second segments by finding the event HMM that has most likely produced the observation sequence of each segment. In this case, the system is used to recognize the most prominent event in each segment. A more detailed explanation of the event detection system can be found in [10].

4. CONTEXT RECOGNITION

We assume that each context is characterized by the presence of certain sound events. The event histogram for a recording is constructed by collecting all the sound events into an event occurrence histogram. In order to prevent a bias towards longer recordings, the event counts in the histogram are divided by the number of events present in the recording. The models for the contexts are constructed by summing up these event histograms. The context model histogram is normalized so that the bins sum up to one.

In the recognition stage, an event histogram is collected from the events that are detected in the tested recording. Histograms are calculated either from the output of the Viterbi segmentation or by accumulating the events recognized in the four second segments. The context recognition is based on comparing this histogram with the context histogram.

The event histograms are compared by calculating a distance between them. In the preliminary studies, we tested three distance metrics for the task: the cosine distance, the correlation distance and one based on the Kullback-Leiber divergence. Since they provided rather similar performance, in the final system we chose to use only one of them, the cosine distance. The cosine distance is defined as the cosine of the angle between an event histogram for context C and an event histogram for tested recording Q:

$$Dist_{cos}(Q,C) = \frac{\sum_{i=1}^{T} q_i c_i}{\sqrt{\sum_{i=1}^{T} q_i^2 \sum_{i=1}^{T} c_i^2}}, \qquad (1)$$

where q_i is the normalized event count of event i in the tested recording, c_i is the normalized event count of event i in the context and T is number of events in the vector. The context corresponding to the closest distance is selected as the recognition result.

In order to better model the within context variation in the distribution of events, k-nearest neighbor (k-NN) classification is also used. With k-NN, all the recordings in the training database can be used to represent the context they belong to. In this case each context is represented by several event histograms, each calculated from a single recording in the training database. Distances to each recording are calculated and the context recognition is done by majority voting among classes corresponding to the k nearest context instances.

4.1 Weighted event histograms

A weighing scheme for the events can be developed in a similar manner to the term frequency–inverse document frequency (TF-IDF) used for document indexing [11, 12]. In our case, the indexing term is the sound event and the document is a recording from a specific context or the entire context depending on the evaluation setup. The main idea of TF-IDF is that a term is an important indexing term for document d if it occurs frequently in it. This is denoted as term frequency (TF). On the other hand, terms which occur in many documents are rated less important for indexing due to their widely common nature. This is denoted as inverse document frequency (IDF) and it is defined as follows:

$$IDF(term) = log\left(\frac{|D|}{DF(term)}\right) \quad (2)$$

where $|D|$ is the total number of recordings and $DF(term)$ is the number of documents in which the term occurs at least once. The inverse document frequency of a term is low if it occurs in many documents and is highest if the term occurs in only one. The weight w_i of a term i in document d is calculated as

$$W_i = TF(term_i, d) \bullet IDF(term_i), \quad (3)$$

where $TF(term_i, d)$ is the term frequency, i.e., the number of times $term_i$ occurs in the document d.

In the training stage, IDF is collected from the training data and event histograms (TF) for contexts are weighted. In the testing stage, event histogram (TF) is collected from the test data and IDF calculated from the training data is used in the weighting of the event histogram.

5. EVALUATION

The proposed context recognition system is evaluated with an audio database collected from real-life environments. The database is used to train the event detection system and the context recognition system. Two different methods for obtaining the events are evaluated. In the first method, event recognition is done by splitting each recording into four second segments and classifying each segment as corresponding to the most likely event. The events detected in the segments within the tested recording are collected to form an event histogram. The second method uses the Viterbi algorithm to obtain the most likely event sequence for the entire recording and this sequence will be used to construct the histogram. In addition to this, two different methods for modeling each context are evaluated. The first method is to characterize each context by one histogram constructed from all the events. In the second method each recording belonging to a context is used as an example of that context and k-NN classification is used. We also study the effect of the test segment length on the recognition accuracy in detail.

5.1 Database

The material for the database was gathered by recording 10 to 30 minute long recordings in ten real-life environments or contexts. The selected audio contexts were basketball game, beach, inside a bus, inside a car, hallway, office, restaurant, grocery shop, street and stadium with track and field events. For each context, 8 to 14 recordings were made with binaural microphones placed inside the human ears. In total, 103 stereophonic recording was included in the database. In this paper, we are using monophonic versions of the recordings, i.e., two channels are averaged to one channel.

Table 1: Event statistics from the database.

Context	Number of present event classes	Total number of events	Average events per 1 min.
basketball	14	990	11.3
beach	16	738	3.7
bus	14	1729	12.0
car	12	582	5.3
hallway	9	822	7.4
office	12	1220	12.3
restaurant	13	780	7.8
shop	14	1797	20.4
street	15	827	7.6
track & field	11	793	6.9

Table 2: Context-wise average recognition performances.

	4 sec. segments	Viterbi segmentation
Cosine	88.5	84.5
TF-IDF	61.1	59.3

The recordings were manually annotated indicating the start and end times of all clearly audible sound events in the auditory scene. The repetitive sound events are usually annotated as long events, e.g. ball hitting the floor in the basketball game, while long events like conversation are annotated as multiple successive speech events if there is perceivable pause in the conversation. Annotated sound events present in the recordings were grouped into 61 event classes. The event classes include e.g. speech, laughter, applause, car door, road, dishes, door, chair, music, and footsteps. Each context contains events from 9 to 16 event classes and many event classes appear in multiple contexts. There are also event classes which are context specific. Event statistics from the recording database are presented in Table 1. Figure 2 shows the event histograms collected from the database.

The database was organized in a five-fold manner into training and testing sets, to test all the available recordings. The audio of the training set is used to train the event detection system and histograms of annotated event class occurrences are used to train the context recognition system.

5.2 Event based recognition

The results for event based context recognition are presented in Table 2. “Cosine” denotes a system were the distance between the estimated event histograms and the context histograms is calculated with the cosine distance. “TF-IDF” denotes a system were the event histograms are TF-IDF weighted before calculating the cosine distance. Two methods of collecting events are used in this evaluation. The method where event recognition is done with four second segments is denoted as “4 sec. segments” and the method using Viterbi decoding is denoted as “Viterbi segmentation” in the table.

The full confusion matrix for the system ’Cosine’ is shown in Table 3. Some of the confusions are understandable when looking at the sound events present in the contexts. For

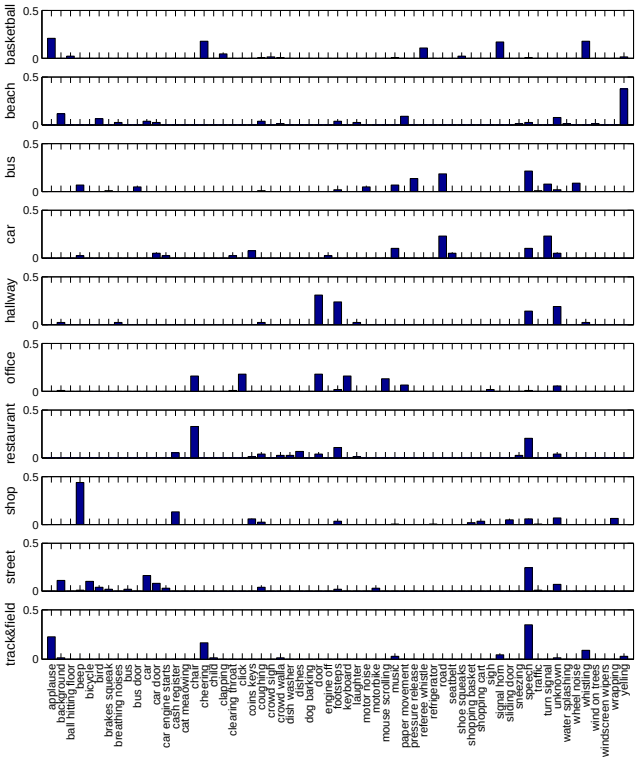

Figure 2: Normalized event histograms for contexts.

Table 3: Confusion matrix for context recognition using event histograms. Rows in the matrix correspond to presented context and columns to the recognition result.

		1	2	3	4	5	6	7	8	9	10
basketball	1	100									
beach	2		64							36	
bus	3		9	91							
car	4				100						
hallway	5					60	20	10		10	
office	6			10			90				
restaurant	7			10				90			
shop	8								100		
street	9					10				90	
track & field	10										100

example, in the hallway there are footsteps and ventilation noise present while footsteps are also present in the street context and similar ventilation noise in the office context.

Recognition results using k-NN approach with varying values for k are presented in Table 4. In this case, TD-IDF weighting helps the context recognition and provides a better performance than when using unweighted histograms. Since the idea of TF-IDF is to weigh rare events more than the common ones, collecting all the events from the database to form only one context model for each context will average out the rare events within each context and the recognition will only become more difficult.

5.3 Combining event and direct acoustic information

In addition to the event based context recognition, a system based on acoustic information of contexts was evaluated. More specifically, we constructed a baseline system where each of the ten contexts is modeled with a GMM (16 Gaussians) and using MFCCs (static, first and second order time derivatives). The test recordings for this system are cut into four second segments which are then classified individually. This system is later referred as the baseline system.

Table 4: Multiple context instances and kNN based recognition.

	$k=1$	$k=3$	$k=5$	$k=7$	$k=9$
4 second segments					
Cosine	87.3	84.6	85.8	84.8	83.8
TF-IDF	89.3	85.6	84.6	85.5	86.6
Viterbi segmentation					
Cosine	86.4	84.6	84.6	82.6	81.5
TF-IDF	89.3	87.5	87.5	89.4	89.4

Table 5: Context-wise average recognition performances.

	4 sec. segments	Viterbi segmentation
Baseline	88.5	
Cosine + Baseline	91.4	92.4
TF-IDF+ Baseline	90.5	90.4

Since the baseline system models global acoustic characteristics of the audio context instead of sound events, it may provide complementary information compared to the proposed event based system. Combining these two may thus lead to improved performance. To combine these two systems, the distance between the test event histogram and the context histograms are mapped into probabilities using an inverted sigmoid-function. The mapped probabilities are then multiplied with the context likelihood produced by the baseline system.

The evaluation results are presented in Table 5. "Baseline" denotes the system based on acoustic information of contexts and "Cosine+Baseline" denotes the system where the output of the baseline system is combined with the event based context recognition system without TF-IDF weighting. "Baseline+TF-IDF" denotes a system were the weighting of the event histograms is used. The proposed context recognition system provides comparable recognition accuracy with the baseline system (see Tables 2 and 4). The recognition accuracy is slightly improved when the proposed system is combined with the baseline system.

The full confusion matrix for the baseline system is shown in Table 6. The full confusion matrix for the system where the output of the baseline system is combined with the proposed event based system without TF-IDF weighting (see Table 3) is presented in Table 7. By comparing the confusions in Tables 6 and 7, one can see that the event based system increased the performance on the bus and hallway contexts. Confusions of the bus context are now made with the street context which is understandable since they share some sound events.

5.4 Test segment length

The effect of different test segment lengths on the recognition accuracy was evaluated. Evaluation was done by constructing the event histogram from the classification results of different number of four second segments. Using the baseline system, the likelihoods of successive four second segments are accumulated over time. The recognition results based on the test segment length are shown in Figure 3 for the baseline system and the system using k-NN approach.

Table 6: Confusion matrix for context recognition using the baseline system.

		1	2	3	4	5	6	7	8	9	10
basketball	1	100									
beach	2		73			9				18	
bus	3			73		9		18			
car	4				100						
hallway	5					50	30			20	
office	6					10	90				
restaurant	7							100			
shop	8								100		
street	9									100	
track & field	10										100

Table 7: Confusion matrix for context recognition using the "Cosine+Baseline" system with Viterbi segmentation.

		1	2	3	4	5	6	7	8	9	10
basketball	1	100									
beach	2		73			9				18	
bus	3			91						9	
car	4				100						
hallway	5					80	20				
office	6					10	90				
restaurant	7							100			
shop	8								100		
street	9								10	90	
track & field	10										100

5.5 Discussion

TF-IDF weighting was found to help recognition only when using multiple examples of one context, represented by the recordings in the training database. This is due to the fact that TF-IDF weights rare events more than the common ones and having only one model for the complex contexts will smooth out the rare events. Furthermore, this weighting has problems with short segments having small amount of events which are all common events, and thus will be weighted to zero.

The performance of the event based system is not superior to the baseline system. The 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.

6. CONCLUSIONS

In this paper, event histograms were used for context recognition. Recognition was evaluated on a database consisting of 103 recordings from ten different contexts. The best recognition result, 89.4% correct, for the event based recognition was obtained using multiple context instances from the training database and a *k*-NN classification approach. When combining the event based context recognition with a baseline context recognition system, the performance was increased to 92.4%.

In the future, other classification methods than distance metrics and *k*-NN will be studied. For example, training support vector machines with the event histograms might provide better recognition results.

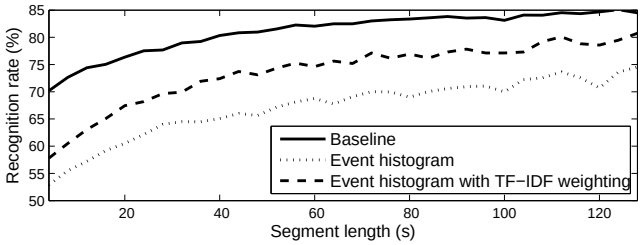

Figure 3: Context recognition accuracy as function of test segment length.

REFERENCES

[1] V. T. K. Peltonen, A. J. Eronen, M. P. Parviainen, and A. P. Klapuri, "Recognition of everyday auditory scenes: Potentials, latencies and cues," in *In Proc. 110th Audio Eng. Soc. Convention*, Hall, 2001.

[2] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 14, pp. 321–329, Jan. 2006.

[3] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–22, 2006.

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

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

[6] R. Cai, L. L., and H. A., "Co-clustering for auditory scene categorization," *IEEE Trans. on Multimedia*, vol. 10, no. 4, pp. 596–606, 2008.

[7] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 2, pp. 1–23, 2008.

[8] A. Härmä, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *IEEE Int. Conf. on Multimedia and Expo*, pp. 634–637, 2005.

[9] 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*, pp. 364–370, Springer-Verlag, 2008.

[10] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference*, 2010.

[11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[12] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.