SOUND EVENT DETECTION FOR OFFICE LIVE AND OFFICE SYNTHETIC AASP CHALLENGE

Aleksandr Diment, Toni Heittola and Tuomas Virtanen

Tampere University of Technology
Korkeakoulunkatu 10,
FI-33720 Tampere, Finland
[aleksandr.diment, toni.heittola, tuomas.virtanen]@tut.fi

ABSTRACT

We present a sound event detection system based on hidden Markov models. The system is evaluated with development material provided in the AASP Challenge on Detection and Classification of Acoustic Scenes and Events. Two approaches using the same basic detection scheme are presented. First one, developed for acoustic scenes with non-overlapping sound events is evaluated with Office Live development dataset. Second one, developed for acoustic scenes with some degree of overlapping sound events is evaluated with Office Synthetic development dataset.

Index Terms— Sound event detection, hidden Markov model, multiple Viterbi passes

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 in an office contains events of door opening, keyboard typing, mouse clicking and various human sounds (speaking, laughing, and coughing).

Automatic sound event detection aims at processing a continuous acoustic signal and converting it into a sequence of event labels with associated start times and end times. The sound event detection can be utilized in a variety of application areas, including context-based indexing and retrieval in multimedia databases [1], unobtrusive monitoring in health care [2], and audio-based surveillance [3]. Furthermore, the detected events can be used as mid-level-representation in other research areas, e.g. audio context recognition [4, 5], automatic tagging [6], and audio segmentation [7].

Most of the research on sound event detection has been concentrated on detecting only one sound event at a time, considerably simplifying the detection problem [8, 9, 10]. However, everyday auditory scenes are usually complex in sound events, having multiple overlapping sound events active at the same time. If an algorithm that detects only a single event at a time is applied to material consisting of overlapping events, the majority of detection errors will be caused by temporally overlapping sound events. In order to detect all sound events, a way to deal with overlapping events is needed. Recently, this problem has been addressed at various levels of the detection process. At the signal level, sound source separation techniques can be used to minimize the acoustical interference of overlapping sound sources [11]. In the event detection stage, overlapping events can be detected with multiple iterative detection passes and by excluding already detected events from the following detection iterations until the desired amount of overlapping events have been reached [12].

In this work, we utilize a sound event detection system based on the system described in [12]. The system is slightly simplified by bypassing the context-detection stage, since the event detection task at AASP Challenge [13] is restricted only to office environment. For the Office Live task we use the prominent event detection scheme, and for the Office Synthetic task we extend the prominent detection to multiple detection passes.

2. SYSTEM OVERVIEW

This section explains the sound event detection approach used in the proposed system. The system recognizes and temporally locates sound events in the test recordings.

2.1. Event Models

The coarse shape of the power spectrum of the recording from the auditory scene is represented with 20 mel-frequency cepstral coefficients (MFCCs). In order to describe the dynamic properties of the cepstrum, first and second time derivatives of the static coefficients are also utilized. Features are calculated in 20 ms frames with 50 % overlap.

Sound-event-conditional feature distributions are modeled using continuous-density hidden Markov models. Left-to-right model topology having three states was chosen to represent sound events having a beginning, a sustained part, and an end part. The probability density functions of observations in each state is modeled using a mixture of multivariate Gaussian density functions (8 Gaussians).

In addition to the targeted sound events, the test recordings contain background noise at various levels. Background noise is modeled by a one-state HMM (32 Gaussians) trained using segments of the development dataset which contain no annotated sound events.

In the detection stage, the sound event models are connected into a network with transitions from each model to any other. An example of model network is shown in Figure 1.

This document is licensed under the Creative Commons
Attribution 3.0 License (CC BY 3.0).
http://creativecommons.org/licenses/by/3.0/
© 2013 The Authors.

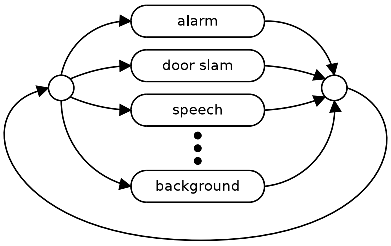

Figure 1: Fully-connected sound event model network.

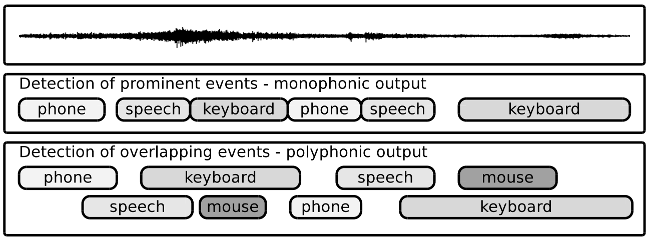
Detection of prominent events - monophonic output
phone
speech
keyboard
phone
speech
keyboard
Detection of overlapping events - polyphonic output
phone
keyboard
speech
mouse
speech
mouse
phone
keyboard

Figure 2: Example of sound event detection output for the two approaches: monophonic system output and polyphonic system output.

2.2. Event Detection

The detection of the most prominent event at each time instance is applied for the Office Live-subtask of the challenge to produce monophonic event sequence as an output. The audio material in this subtask is strongly sequential having only one sound event active at the time, so this type of approach suits well the problem. This approach is later referred as *monophonic* detection.

The detection of a predefined number of overlapping events is applied to the Office Synthetic-subtask of the challenge to produce a polyphonic event sequence as an output. The audio material in this subtask has varying degree of overlapping events with varying SNR ratios. This approach is later referred as *polyphonic* detection. Examples of the outputs of these two approaches are shown in Figure 2.

2.2.1. Monophonic detection

Segmentation of a recording into regions containing the most prominent event at a time will be obtained by doing Viterbi decoding inside the network of sound event models. The number of events in the output event sequence is controlled by a cost for inter-event transitions. This parameter is called insertion penalty, and the value was chosen experimentally using the development dataset.

2.2.2. Polyphonic detection

To detect overlapping events in the Office Synthetic-task, the monophonic detection approach is extended by using consecutive passes of the Viterbi algorithm as proposed in [14] for the detection of overlapping musical 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 background model is allowed in each iteration. This method will provide iterative decoding of the next-best sequence of events that are at each time different than in the previously decoded one. The main idea of this approach is illustrated in Figure 3. The number of iterations depends on the expected polyphony of the tested acoustic material and is fixed based on preliminary evaluations on the development dataset.

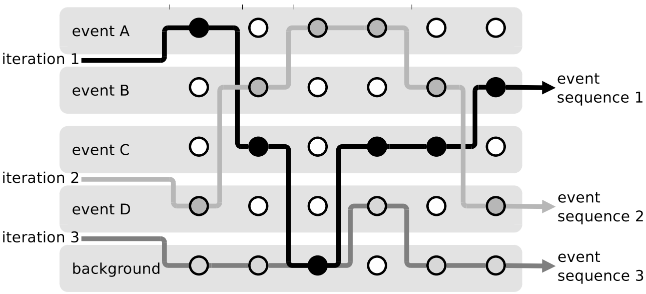

Figure 3: Concept of multiple path decoding using three consecutive passes of Viterbi algorithm.

3. EXPERIMENTS ON DEVELOPMENT DATASET

The datasets used in the challenge are restricted to 16 sound event classes: alarm, clearing throat, cough, door slam, drawer, keyboard, keys, door knock, laughter, mouse, page turning, pen drop, phone, printer, speech and switch.

3.1. Metrics

The evaluation used in the AASP challenge highlight different aspects of the event detection by using three metrics [13]: frame-based, event-based and class-wise event-based. The frame-based F-score metric is evaluated in a 10-ms frames by checking active events within the frame. The event-based metric considers either only onsets or both onsets and offsets. For this metric, the onset of an event is correct if the onset of the output event is within 100 ms tolerance window compared to the ground truth. The offset is considered correct if it is within a tolerance window ranging from mid point of the ground truth event until the end of it. The class-wise event-based metric is used in order to avoid effect of repetitive events in the test dataset.

3.2. Office Live

The training dataset containing 20-24 examples of each sound event class recorded in the office environment is used to train acoustic models for sound events. In addition this, three live recordings of scripted non-overlapping event-sequences from an office environment are provided as development dataset. The recording length varies between one to three minutes. Two sets of annotations by different annotators are provided. For simplicity, we are using only one of them for reporting the results on the development dataset.

The results on the development dataset, averaged over the three recordings are shown in Table 1.

3.3. Office Synthetic

The same training dataset as in Office Live is used to train acoustic models. The development dataset consists of nine artificial samples created by concatenating overlapping sound events from an office

Table 1: Average detection results on development set of Office Live task.

	F-score
Frame-based	61.6
Event-based	
Onset-only	55.4
Onset-offset	46.7
Class-wise event-based	
Onset-only	41.1
Onset-offset	35.0

Table 2: Average F-scores of the evaluation metrics on development dataset of Office Synthetic task.

Number of iterations	1	2	3	4
Frame-based	43.2	**43.9**	39.8	35.6
Event-based				
Onset-only	24.5	**26.5**	23.8	20.7
Onset-offset	18.4	**20.4**	18.8	16.5
Class-wise event-based				
Onset-only	22.6	**31.0**	31.0	30.3
Onset-offset	16.3	**24.7**	25.8	25.1

environment. The dataset contains signals with various SNRs (-6dB, 0, 6dB) in respect to the background noise and different level of density of sound events (low, medium, high). The acoustic model for the background class is trained using segments extracted from these samples.

The results on the development dataset, averaged over the nine synthetic audio samples are shown in Table 2. The best results are obtained using two iterations. The average event-based onset-offset F-scores over different SNR conditions using two iterations are: 14.7% (-6dB) , 17.7% (0dB), 28.8% (6dB). The average event-based onset-offset F-scores over different event densities using two iterations are: 12.7% (high), 32.5% (medium), 15.9% (low).

Two iterations provides a good compromise for all event density levels used in the evaluation. The number of iterations was fixed to two in the submission.

4. CONCLUSION

We presented sound event detection system based on the system described in [12]. The system uses HMMs and Viterbi decoding to find the most probable event sequence. For the Office Live task, we used the prominent event detection scheme to produce monophonic event sequence. For the Office Synthetic task, we extended the prominent detection scheme with multiple detection passes to produce polyphonic event sequence. In the preliminary evaluations with the development dataset, the system is showing promising results for both tasks.

5. REFERENCES

[1] 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.

[2] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. 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: IEEE Press, 2009, pp. 1218–1221.

[3] A. Härmä, M. F. McKinney, and J. Skowronek, “Automatic surveillance of the acoustic activity in our living environment.” Los Alamitos, CA, USA: IEEE Computer Society, 2005, pp. 634–637.

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

[5] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Audio context recognition using audio event histograms,” in *18th European Signal Processing Conference*, Aalborg, Denmark, 2010, pp. 1272–1276.

[6] M. Shah, B. Mears, C. Chakrabarti, and A. Spanias, “Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices,” in *Emerging Signal Processing Applications (ESPA), 2012 IEEE International Conference on*, Jan. 2012, pp. 99–102.

[7] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, “Segmentation, indexing, and retrieval for environmental and natural sounds,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 688–707, Mar. 2010.

[8] 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: Springer-Verlag, 2008, pp. 345–353.

[9] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, “Real-world acoustic event detection,” *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010, pattern Recognition of Non-Speech Audio.

[10] A. Mesaros, T. Heittola, T. Virtanen, and A. Eronen, “Acoustic events detection in real life recordings,” in *Proc. of the 2010 European Signal Processing Conference (EUSIPCO-2010)*, 2010, pp. 1267–1271.

[11] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, “Sound event detection in multisource environments using source separation,” in *Workshop on Machine Listening in Multisource Environments, CHiME2011*, Florence, Italy, 2011.

[12] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” *EURASIP Journal on Audio, Speech, and Music Processing*, 2013.

[13] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, L. M., and M. Plumbley, “An ieee aasp challenge: Detection and classification of acoustic scenes and events,” Tech. Rep.

[14] M. Ryynänen and A. Klapuri, “Polyphonic music transcription using note event modeling,” in *In Proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 319–322.