Automatic Detection of Relevant Acoustic Events in Kindergarten Noisy Environments

Jens Schröder, Francois X. Nsabimana, Jan Rennies, Danilo Hollosi, Stefan Goetze
Fraunhofer IDMT, 26129 Oldenburg, Germany, E-Mail: jens.schroeder@idmt.fraunhofer.de

Introduction

In many studies conducted to monitor the health situation of kindergarten child care workers in Germany, the high noise level in the facilities has been pointed out by approximately 70% of the workers as one of the most stressful factors. One factor contributing to the stress is considerable background noise in kindergartens, many important events such as calls for help of children or colleagues might be unheard at their first utterance. This contribution presents results of a study conducted in a real kindergarten for which machine-learning approaches were tested to detect and classify acoustic events in typical background noises. For the training of applied approaches daily kindergarten noise has been recorded. The goal of this study is to develop an automatic acoustic event detection (AED) system to considerably reduce the number of unrecognized, desired events important for child-care workers. AED is increasingly used in many fields of application, e.g. for surveillance and security issues [1–3] or in the field of ambient assisted living (AAL) [4, 5]. In past AED challenges, i.e., the CLEAR'07 ("classification of events, activities and relationships") challenge [6] that was part of the CHIL project [7] and the D-CASE ("detection and classification of acoustic scenes and events") challenge [8], detecting acoustic events in a meeting room and office scenarios has been addressed. For these challenges, the proposed AED approaches were mainly based on Mel-frequency cepstral coefficient (MFCC) features in conjunction with hidden Markov models (HMMs) [9, 10].

In this contribution, two different approaches based on HMMs in combination with MFCCs are applied. The commonly used maximum likelihood approach just allows for one event detection per time section, whereas in real scenarios like kindergartens multiple events occur simultaneously and overlapping. Thus, we propose applying a set of binary classification systems by using an universal background model (UBM) that is compared to each event model resulting in binary decisions.

Classification systems

Commonly, recognition with HMMs is done by comparing the likelihoods of different event models. The model with highest likelihood is assumed to describe a time section, i.e.,

$$\hat{c} = \arg\max_{c} p\left(\boldsymbol{\chi}\,|\lambda^{c}\right), \tag{1}$$

where λ^c is the HMM of event class $c = 1 \ldots C$ with C being the total number of event classes. $\boldsymbol{\chi}$ denotes the feature vectors for a time section and $\hat{c}$ is the event with maximum likelihood. This multiclass classification system yields one detection label per time section. Since in real world scenarios, multiple events can occur simultaneously and overlapping, we propose the use of a set of binary classification systems. Therefore, an UBM is trained on all training data. Each event HMM λ^c is compared to the UBM λ^{UBM}, i.e.,

$$\hat{\boldsymbol{c}} = \left\{ c \,\middle|\, p\left(\boldsymbol{\chi}\,|\lambda^{c}\right) > p\left(\boldsymbol{\chi}\,\middle|\lambda^{\mathrm{UBM}}\right) \right\}. \tag{2}$$

Hence, the role of the UBM is to detect time sections that do not belong to a model λ^c. Thus, multiple detections $\hat{\boldsymbol{c}}$ up to C labels per time section are possible.

Experimental Setup

The evaluation is done using a database of real-world recordings from a day care center for children and a kindergarten. The data were split into five disjoint sets to perform a five-fold cross-validation . HMM recognizers were trained to model the events, silence and the UBM. The common multiclass classification system based on Eq. (1) and the binary classification system from Eq. (2) were tested. Details are presented in the following subsections.

Database

The database used for training and evaluation was recorded in a day care center for children between 0 and 3 years and in a kindergarten for children between 3 and 6 years. In both facilities, five rooms and one hallway were equipped with microphones, i.e., altogether twelve microphones were used. The data collected comprised 2840 minutes of recordings from daily activities in those facilities. The data were annotated by hand. Since the occurring events can be labeled very detailed and differently leading to a huge amount of event classes with few samples, we decided to use the labeling approach suggested in [11] to limit the number of event classes. In this report, a labeling list is proposed that was developed to label acoustic data in video clips by few, meaningful labels. These labels are supposed to present smallest components sound data can be composed of. In analogy to phonemes for words, they are called noisemes. A list is given in Table 1. The data were split into five disjoint sets to conduct a five-fold cross-validation.

Recognizer

As input for HMMs, MFCCs [12] are extracted from the time signal. For the back-end classifier, the Hidden Markov Toolkit (HTK) [12] is applied to build up an HMM recognition network with a task grammar. HTK provides a speech recognition network of three levels: word level, model level and HMM level. In this contribu-

Table 1: List of noisemes for labeling from [11]

Broad	Noiseme	Sounds like ...
Anim	Animal	Not identifiable animal
	Anim_...	Identified animal, e.g. anim_bird
Human_noise_s	Cry	Crying
	Human_noise	Vocal noise, e.g. cough, sneeze, throat
	Laugh	Laughter
	Scream	Screaming
	Child	Child/baby coos; animal coos
Speech_s	Mumble	Non-intelligible, single voice
	Speech	Intelligible speech English
	Speech_ne	Intelligible speech not English
Singing	Singing	Only voice; a capella
Human_m	Cheer	Intelligible speech, multiple voices
	Crowd	Non-intelligible, multiple voices
Music	Music_sing	Music with singing
	Music	Only music
Noise_pulse	Knock	Hits woods, cardboard, dry wall
	Thud	Hits floor, dirt, carpet, damped
	Clap	Hands, gun, shot-like, explosion
	Click	Quiet, mechanical click
	Bang	Hits metal, glass, tone-ish
	Beep	Very short beeps, computer
Noise_ongoing	Clatter	Bangs, knocks, pulses, irregular
	Rustle	Scratching, hiss, rustling, irregular
	Scratch	Short friction segments, regular
	Hammer	Bangs, knocks, pulses, regular
	Washboard	Fast pulses with rubbing sounds, friction, regular
	Applause	Very fast claps comb, with friction
Engine	Engine_quiet	Rattle, sewing machine, video camera
	Engine_light	High-freq. machine noise, drill-like
	Power_tool	Mid.-freq. machine noise, race car
	Engine_heavy	Low-freq. machine noise, truck, tractor
Noise_tone	Phone	Classical telephone ring, ringing
	Whistle	High-freq. tone
	Squeak	Tire squeak, friction squeak, high freq.
	Tone	Steady tone, horn, alarm
	Siren	Oscillating sound waves
Noise_backgr_nat	Water	Dubbing, splashing
	Micro_blow	Wind or breath hits microphone
	Wind	Guts, flag clatter, pulses, scratch
Noise_backgr	Radio	Radio/TV in background
	White_noise	Fuzzy signal, air cond., waterfall, hum
Other	Other_creak	Open for unseen noises

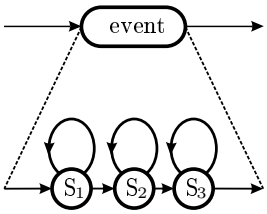

Figure 1: Schematic of a left-to-right HMM with three states that is used to model events.

tion, events are treated as words. The model level, that is used in speech recognition to represent sub-words like phonemes, is not employed here. Thus, the whole recognizer can be interpreted as a two-layer HMM. The first layer is a fully connected HMM in which each state is an event, i.e., each event can be accessed at every time. The observations of these event states are themselves HMMs that are trained independently using the extracted features. These events are modeled by left-to-right HMMs with three emitting states (cf. Figure 1) that was proposed in [13] as well. To estimate time regions in a signal in which no active event is present, an extra *silence* class is modeled. The UBM is trained using all data.

The number of Gaussian mixtures $\mathcal{M}$ for the event classes are optimized on the fifth (testing) fold.

To estimate the time regions of events in a signal, Viterbi decoding [12] is used. Since the output can be highly fragmented, i.e., several insertion and deletion errors may occur, a fixed logarithmic probability insertion penalty p is added to every event state transition [12]. Thus, the probability to remain in an event/UBM/*silence* state can be increased and a less scattered output is achieved. This parameter is also optimized on the fifth fold.

For the multiclass classification system based on Eq. (1), $\mathcal{M}$ and p are equal for each event class. For the binary classification system based on Eq. (2), this is done individually for each binary decision.

Metrics

As evaluation metrics, the F-Score and the acoustic event error rate (AEER) are used based on frame-wise, event onset (tolerance 100 ms) and event on-/offset (onset tolerance 100 ms, offset tolerance 50% of event length) measure [8]. The F-Score

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

represents the relation between the precision

$$P = \frac{H}{M} \tag{4}$$

and the recall

$$R = \frac{H}{N}, \tag{5}$$

with H denoting the number of correct hits, M the number of estimated events and N the number of reference events.

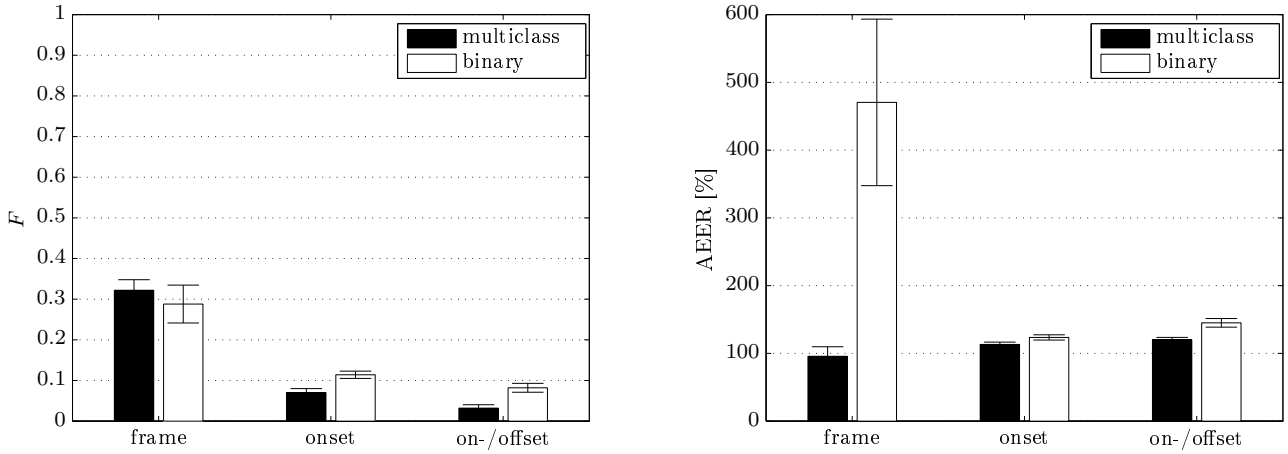

Figure 2: Mean (bars) and standard deviations (whiskers) of the F-Score F (left panel) and AEER (right panel) from the five-fold cross-validation for the multiclass (cf. Eq.(1)) (black) and the binary (cf. Eq.(2)) (white) classification system based on frame-wise, event onset and event on-/offset measure.

The AEER is the sum of insertions I, deletions D and substitutions S relative to the number of reference events N, i.e.,

$$\text{AEER} = \frac{I + D + S}{N}. \tag{6}$$

Results

The evaluation is conducted based on a five-fold cross-validation. In Figure 2, the results in terms of mean and standard deviation for the F-Score (left panel) and the AEER (right panel) are depicted. It can be seen that the F-Score based on the onset and on-/offset measure is higher for the binary classification system. For the frame based measure, the multiclass classification system is more accurate. For the AEER, the multiclass classification system leads to fewer errors than the binary system for every measured condition. Since the AEER comprises insertion errors, the AEER is not limited to 100% error.

Conclusion

In this contribution, we investigated AED for noisy kindergarten environments . For this, a database consisting of real recordings conducted in a day care center for children and a kindergarten was recorded. Since the annotation of the events within the database is complex, we proposed the use of a defined list of noisemes. Two HMM based classification systems were tested. The multiclass classification system is based on the commonly used maximum likelihood approach yielding one label per time section. The other proposed binary classification system applies a set of binary decisions between a model and a UBM. Thus, it is capable of detecting multiple events per time section.

The evaluation of these systems showed slightly better performance for the binary classification system regarding the F-Score but worse based on AEER than the multiclass classification system. However, the performances of both systems are still low. On the one hand, this results from the highly complex scenario with overlapping sounds. On the other hand, the labeling/evaluation method may not be optimal for our application scenario. For example, events like "clatter" and "knock" just differ in the number of times they occur. Hence, if the classification system outputs multiple times "knock" instead of "clatter", this leads to higher error rates. Furthermore, errors can occur from mixing up very similar classes that are not useful to discriminate against, e.g., "speech" and "mumble". All these effects can cover the actual performance difference between the two proposed classification systems and reduce the overall accuracy. Thus, labeling is crucial and of high importance. This will be further evaluated in the future considering this awareness.

Acknowledgement

This study has been conducted within the collaborative research project SmartKita funded by the German Federal Ministry of Education and Research (FK 16SV6117). We thank Felicitas Kohl, Sibylle Meyer and their colleagues for their extensive support with respect to communicating with personnel and employees at the model site.

References

[1] D. P. W. Ellis, "Detecting alarm sounds," in *Proceedings of the Recognition of real-world sounds: Workshop on consistent and reliable acoustic cues*, Aalborg, Denmark, 2001, pp. 59–62.

[2] R. A. Lutfi and I. Heo, "Automated detection of alarm sounds," *Journal of the Acoustical Society of America*, vol. 132, no. 2, Sep. 2012.

[3] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 493 – 497.

[4] J. Schröder, S. Wabnik, P. van Hengel, and S. Goetze, "Detection and classification of acoustic

events for in-home care," in *Ambient Assisted Living*. Springer, 2011, pp. 181 – 195.

[5] D. Hollosi, J. Schröder, S. Goetze, and J.-E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," in *3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, Rome, Italy, Nov. 2010.

[6] CLEAR: Classification of Events, Activities and Relationships, 2007. [Online]. Available: http://clear-evaluation.org/?CLEAR

[7] CHIL: Computers in the human interaction loop. [Online]. Available: http://chil.server.de/

[8] D-CASE: Detection and Classification of Acoustic Scenes and Events, 2013. [Online]. Available: http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/

[9] R. Stiefelhagen, R. Bowers, and J. G. Fiscus, Eds., *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 4625.

[10] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[11] S. Burger, Q. Jin, P. F. Schulam, and F. Metze, "Noisemes: Manual annotation of environmental noise in audio streams," Carnegie Mellon University (CMU), Language Technologies Institute (LTI), Tech. Report LTI-12-017, 2012, http://www.lti.cs.cmu.edu/research/ reports/2012/cmulti12017.pdf.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, 2006.

[13] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, Aug. 2010, pp. 1267–1271.