



An Acoustic Event Detection Framework and Evaluation Metric for Surveillance in Cars

Peter Transfeld, Simon Receveur, and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig, Germany

{transfeld, receveur, fingscheidt}@ifn.ing.tu-bs.de

Abstract

The protection of cars against burglary and car theft is a major issue for automotive industry. Several techniques exist to stop the thief from driving away or to alert the owner. In this contribution we present an acoustic event detection (AED) framework for surveillance in cars, operating on a continuous audio stream which may be acquired by the existing in-car hands-free microphone. Furthermore we derive a new time-dependent accuracy measure tACC, evaluating both the accuracy and the temporal preciseness of the instant of detection. Operating on continuous audio, we examine a wide range of signal to noise ratios (SNRs) in a realistic parking scenario, showing the effectiveness of both the proposed AED framework and the new time-dependent accuracy measure.

Index Terms: Acoustic event classification and detection, sound recognition, car surveillance, hidden Markov models

1. Introduction

Within automotive industry, the protection of cars from being stolen or damaged is a major issue. Addressing this, car immobilizers have become standard in many new cars. Commonly, a radio frequency identification code is stored on a chip inside the car key [1, 2]. Only with the correct key, the car will start the engine, otherwise all car electronics is disabled. This technique prevents the car from being stolen, but does not protect any interior items such as the radio/navigation system or personal belongings of the passengers. Therefore several intruder alarm systems exist, detecting broken windows or abnormal movement inside the passenger compartment by vibration, microwave-ultrasonic [3], or in-car radar sensors [4]. More modern approaches combine mobile phones, the global positioning system (GPS) and digital cameras. Using face recognition algorithms on a camera attached to the drivers sun-shield, the identity of a person attempting to drive can be identified. If it is the owner (or any other allowed person) the car engine will start, otherwise a picture of the person and the GPS position are sent to the car owner or the police via a mobile phone [5, 6]. Even though these methods against burglary and car theft have been developed in the past years, the number of stolen cars is still rising. This may be because of the fact, that such protection systems are an extra option when buying the car as more hardware is needed.

Here, acoustic event detection (AED) or in a wider scope auditory scene analysis [7] may provide possible solutions, as they are already known in several application fields. Starting with monitoring tasks, where AED is used to monitor the health status of elderly people [8], or to detect emergencies [9]. Furthermore in the field of surveillance, AED is utilized to detect screams [10, 11] or impulsive sounds [12] to analyze the audi-

tory scene, whereas in a meeting room environment it is employed to detect special events (e. g., door sounds, phone ringing, keyboard typing) [13]. This information can then be used to analyze the social structure within a group of people or to improve the robustness of an automatic speech recognition (ASR) system. Moreover, sound analysis should be mentioned, where AED is taken to automatically classify video files based on the audio modality [14], or to analyze animal voices in bioacoustics [15]. All of these tasks face a wide range of different signal to noise ratios (SNRs).

In AED approaches, several combinations of *features*, *acoustic model*, and *classifier* have been investigated. Various kinds of *features* (temporal, spectral, and energy) are employed, stacked to feature vectors, and then reduced in their dimensionality in an optimal way [10, 11]. Simpler and even more common is the use of mel-frequency cepstral coefficients (MFCCs). In a comparative study, Cowling and Sitte have shown that MFCCs yield best results in sound recognition [16]. Accordingly, MFCCs are widely used in auditory scene analysis [8, 9, 14]. Depending on the structure of the sound and computational feasibility, the *acoustic model* and its parameters (e.g., topology) are chosen. Basic Gaussian mixture models (GMMs) [10] are as well employed as the more complex hidden Markov models (HMMs). Considering HMMs, in low-resource ASR the recognition of phonemes is commonly modeled by left-to-right HMMs. However, in (animal) sounds repetitive structures may be better modeled by cyclic HMMs [15]. As in ASR these sounds are recognized by means of the Viterbi algorithm [17] determining the most likely sequence of events. In some approaches an explicit acoustic model is even omitted and completely replaced by discriminative classification methods such as support vector machines (SVM) [13].

Evaluation of AED frameworks is done in several ways. Metrics known from ASR such as correctness and accuracy are computed for comparison of frameworks and features [16, 18]. In these cases the appearance of the correct events in the correct order is measured, without any temporal context. Based on the CLEAR evaluation [7] precision, recall, accuracy, and an event recognition score are used in an acoustic event-based manner, to measure the quality of an AED framework [13, 19, 20]. Others interpret acoustic events in a surveillance/security context. In these tasks an event can be taken as an alarm or a detection being false or correct. This interpretation involves metrics such as false detection rate and false rejection rate, either calculated based on events [11, 12, 10] or on a frame level [21].

Based on first experiments on acoustic event detection with noise from a car interior [22], in this paper we propose an AED framework for a car-related context including the more realistic assumption of long event-free phases in a parked car scenario. We evaluate the framework on seven event classes in-

10.21437/Interspeech.2015-452

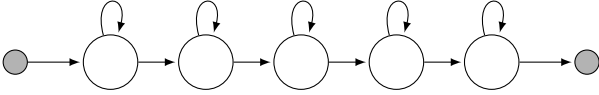


Figure 1: Event class model c_i : White nodes denote emitting states, gray nodes are non-emitting states used to connect the different models.

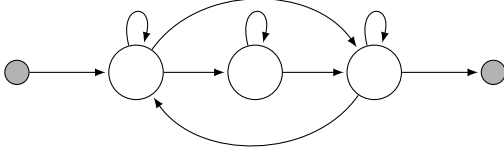


Figure 2: Silence model sil : White nodes denote emitting states, gray nodes are non-emitting states used to connect the different models.

cluding metal and glass sounds, as an abstraction of car-related critical acoustic events and place them in self-recorded in-car street noise at different SNRs. Within the evaluation, we define a new metric, combining the well-known accuracy with a temporal factor reflecting not only the correctness of the recognition process, but also its temporal exactness.

The organization of the paper is as follows: In Section 2 we present the employed acoustic event detection framework. Section 3 details the employed databases and introduces the new evaluation metric. In Section 4 the results are analyzed. Section 5 concludes the paper.

2. Acoustic Event Detection Approach

In this section the acoustic event detection framework is presented, which we propose for in-car surveillance based on continuous audio.

2.1. Feature Extraction

As features we employed 39-dimensional MFCC vectors, containing 12 cepstral coefficients, plus their zeroth component, and their first and second order derivatives. We extracted these MFCCs on a 25 ms Hamming window with 10 ms frameshift on 16 kHz sampled audio data. For feature extraction, the ETSI Advanced Frontend (AFE) [23] was used to obtain noise robust MFCCs.

2.2. Acoustic Model, Grammar, and Training

A linear left-to-right HMM with five emitting states (#2,...,#6) per modeled sound class was employed as acoustic model (see Figure 1). For modeling the state emission probability density functions, GMMs of order six with diagonal covariance matrices were used. In addition a three emitting state silence model (sil , see Figure 2) is introduced with additional transitions from the second to the fourth state and vice versa.

As the AED in a car surveillance context operates on a continuous audio stream, the employed grammar has to form a closed loop. Therefore it is written as $\langle \$word \rangle$, with $\$word$ being either one event class $c_i, i = 1, 2, \dots, 7$, or the silence class sil , and angle brackets $\langle \rangle$ allowing a sequence of unknown length with an unknown number (including zero) of events to be detected. Figure 3 gives an overview of the employed grammar network. The classes to be detected are depicted by c_i

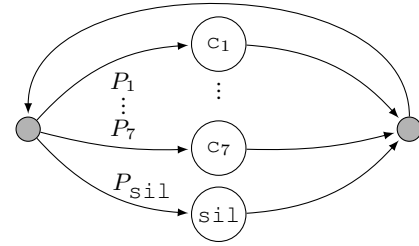


Figure 3: Grammar used for acoustic event detection: c_i denotes the employed event class models (Figure 1), sil denotes the silence model (Figure 2), gray nodes are non-emitting states used to connect the different models, and P_i and P_{sil} denote the transition probabilities for the classes c_i and the silence class, respectively.

nodes, sil indicates the silence model, and gray nodes are non-emitting nodes used to connect the different models. The transition probabilities from the first non-emitting state to the class model c_i and the silence model sil are marked as P_i and P_{sil} respectively. As each class should be treated equally, we set

$$P_i = \frac{1 - P_{sil}}{7}.$$

The employed training setup is based on the assumption that in a parked car it will neither be totally silent nor very loud. So for training of the class models and the silence model a multi-condition setup is chosen, which increases the detection performance for lower SNR. For training of the class models c_i it includes brief event files both disturbed at 10 dB SNR and clean ones. The silence model sil , which in the car context is more a background model, was trained on 10 minutes of in-car street noise. The background noise was split into files of one second and scaled to different noise levels, matching the mean background noise level within each SNR condition. HTK [24] was invoked for training, initializing the HMMs by a flat start.

3. Data Preparation and Evaluation Setup

In this section we describe our database setup and the chosen evaluation methodology. Furthermore we introduce a new evaluation metric, combining the well-known accuracy with a time penalty.

3.1. Database Description

As event database we used the Real World Computing Partnership Sound Scene Database in Real Acoustical Environments (RWCP-SSD-RAE) [25]. The database was recorded under high-quality acoustic conditions in an anechoic chamber without any acoustic disturbance. Each file contains only one event wrapped in almost silence. For this paper we selected seven sound classes from the database, including three metal and four glass sounds. These sound classes serve as an abstraction of car-intrusion related sounds (e. g., metal scratch or breaking glass). The data of each class contains 100 files, yielding a number of 700 single events being processed in this paper.

To investigate the AED framework in a car-related context, we employed street noise as the origin of disturbance and for training of the silence model. Note that in our scenario the sound is acquired *inside* the car, assuming the hands-free microphone as the source of the continuous audio stream. Therefore, we recorded realistic in-car street noise from a

parked Volkswagen Touran at a four-lane public street. A G.R.A.S. microphone type 40AE with a G.R.A.S. preamplifier type 26CA were mounted at the passenger's sun shield and a HEAD acoustics labHMS was used for data acquisition. The in-car street noise was recorded at 48 kHz and subsequently downsampled to 16 kHz to fit the system requirements. The chosen recording conditions ensure a realistic scenario of an unguarded parked car.

3.2. Data Preprocessing

For evaluation we split the $N=700$ files of all event classes into three sets: training (500 files), development (100 files), and test (100 files), such that each class is about equally represented within the three sets. The experiments were repeated over seven different assignments of training, development, and test. As background noise is involved in the training of the silence and the event class models both in training and test, it was also split into three disjoint sections for training, development, and test.

The test scenario simulates a realistic use case of a parked car, which basically means a low frequency of (critical) events surrounded by long intervals of just background noise. Reflecting this scenario, *one* event is randomly placed and added to a file of five minutes of background noise with a given SNR. Also the background noises scaled prior to addition at a given SNR are included as separate event-free files into the test process. This doubles the files for test and also includes the case of *no* event during the five minutes, providing an even more realistic test scenario.

Within the simulations, several SNR levels are investigated. For scaling and adding noise at a specific SNR, typically ITU-T P.56 [26] is used. Unfortunately the voice activity detection included in ITU-T P.56 is unable to detect the time instants of the acoustic events, even though the files of the RWCP-SSD-RAE database are of high quality and almost free of acoustic background or sensor noise. However, due to the professional recording of the database we were able to reliably identify the time instant of an event in a file of N samples as the set of indices

$$\mathcal{N}_{\text{event}} = \{\min\{\mathcal{I}_\theta\}, \min\{\mathcal{I}_\theta\} + 1, \dots, \max\{\mathcal{I}_\theta\}\}, \quad (1)$$

$$\text{with } \mathcal{I}_\theta = \{n \mid n \in \{0, 1, \dots, N-1\} \wedge |s(n)| \geq \theta\},$$

with $s(n)$, $n \in \{0, 1, \dots, N-1\}$, denoting the signal in the entire file. In this notation the event itself is given by $s(n)$, $n \in \mathcal{N}_{\text{event}}$, and $\theta = 50$ being our chosen threshold. When preparing and scaling signal and background noise for addition at a given SNR, we employed a root mean square (rms) measurement of $s(n)$, $n \in \mathcal{N}_{\text{event}}$, for the event signal, and of the background noise $d(n)$, with $n \in \{0, 1, \dots, N-1\}$, following ITU-T P.56, respectively.

3.3. New Evaluation Methodology

Given the grammar definition from Section 2.2 the recognizer output is produced based on the HTK Viterbi algorithm [24]. Beforehand to evaluation, consecutive occurrences of identical classes in the Viterbi output were merged to a single instance, as we want to count them as one. Detected silence occurrences were removed as they do not count as an event.

For evaluation we will compute two performance metrics. At first as a standard metric we use the accuracy

$$\text{ACC} = \frac{N - D - S - I}{N}, \quad (2)$$

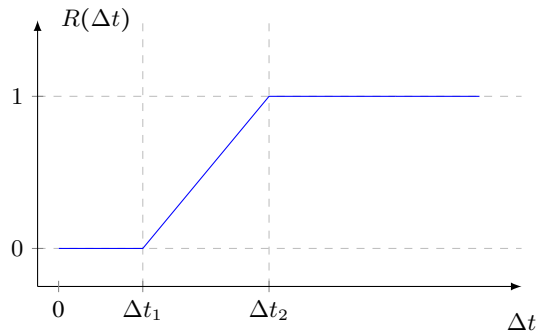


Figure 4: Cost function $R(\Delta t)$

where N , D , S and I denote the number of events, deletions, substitutions, and insertions, respectively.

At this point accuracy gives a measure of how good the correct event is detected within the described detection framework. But in this special task, and in every other surveillance use case, it is even more important that the event is detected at the right time. To combine this requirement with accuracy, we secondly introduce a new metric taking into account the time instants of each event as provided by the Viterbi algorithm, leading to a *time-dependent accuracy* value. We define the *detection time error* of an event by

$$\Delta t = |t_e - \hat{t}_e|,$$

where t_e marks the center of the event, while \hat{t}_e denotes its estimate from the Viterbi algorithm. Depending on Δt , a cost function (see Figure 4) is defined as

$$R(\Delta t) = \begin{cases} 0 & \text{if } \Delta t \leq \Delta t_1 \\ 1 - \frac{\Delta t_2 - \Delta t}{\Delta t_2 - \Delta t_1} & \text{if } \Delta t_1 < \Delta t < \Delta t_2 \\ 1 & \text{else,} \end{cases} \quad (3)$$

with Δt_1 being the maximum detection time error an event is accepted as correctly detected in time, and Δt_2 being the detection time error beyond a correctly recognized event nevertheless counts as an error $R(\Delta t) = 1$. For detection time errors $\Delta t_1 < \Delta t < \Delta t_2$ a correctly recognized event counts at least as a partial error. Finally a time-dependent accuracy is defined by

$$\text{tACC} = \frac{N - R - D - S - I}{N} = \text{ACC} - \frac{R}{N} \leq \text{ACC}, \quad (4)$$

with the *time penalty*

$$R = \sum R(\Delta t) \quad (5)$$

being the summation over costs (3) of all events correctly detected (at their particular time offset). The time-dependent accuracy tACC in (4) now measures both the accuracy *and* the temporal exactness of the AED framework.

4. Experiments and Discussion

4.1. Evaluation Strategies and Optimization

Employing SNR levels of 0 dB to 40 dB in 10 dB steps, and clean data, and using the two metrics (2) and (4), we evaluate the AED framework in two manners. First, we follow the classic evaluation strategy, counting only those events as correct where the correct class out of seven is detected. We call this a *multi-class* detection. Second, we make a binary decision as it would

	SNR					
	clean	40 dB	30 dB	20 dB	10 dB	0 dB
S/N	38.4	37.3	38.4	38.3	36.6	36.7
D/N	1.4	1.6	3.4	7.7	10.7	21.3
I/N	3.3	1.0	1.6	1.0	2.9	5.9
R/N	2.0	4.5	5.4	5.6	3.4	2.1
ACC	56.9	60.1	56.6	53.0	49.9	36.1
tACC	54.9	55.7	51.2	47.4	46.4	34.0

Table 1: Multi-class detection: ACC, tACC, ratio of substitutions S/N, deletions D/N, insertions I/N, and time penalties R/N for different SNRs, in %, respectively; N=100 events.

be appropriate for car surveillance. In this use case, it is not as important to know exactly which critical event was detected, but to know if anything is detected what is seen as a possible threat. We call this *single-class* detection. Note that for single-class detection we always have $S = 0$, whereas both insertions and deletions may happen.

For the Viterbi algorithm, the word insertion penalty parameter (WIP) is used to optimize the recognizer output. In the HTK implementation [24] (and so in our AED framework) at every word boundary the WIP is added to the logarithmic probability, to bias the amount of insertions in the recognition. A negative WIP punishes insertions, whereas a positive one even rewards them. On the other hand a negative WIP rewards shorter sequences of events which means that deletions occur more often. For evaluation we monitored both ACC and tACC, exploring the WIP in the range $[-350, 0]$.

Tables 1 to 2 state accuracy (2) and the new time-dependent accuracy (4) for the two different evaluation setups, with the values of $\Delta t_1 = 0.05s$, $\Delta t_2 = 0.25s$, and $P_{s11} = 0.99$ as it appears to be reasonable for a car surveillance application. For the multi-class and the single-class detection we maximized the sum over both accuracies (ACC + tACC) for each SNR condition individually, as we assume the SNR to be estimated inside the car. As the accuracy and the time-dependent accuracy have shown similar behavior over the monitored SNR range it was possible to jointly optimize them. All optimization steps were made on the development set.

4.2. Results

Having a rough look at the multi-class and the single-class detection, the employed training setup clearly becomes visible, both, in terms off ACC and tACC. Using a multi-condition training for both, the event classes and the silence model, yields to a better detection performance of the framework for low SNR, whereas the detection performance is decreasing for high SNR and clean condition. For both detection types, the detection performance is best at 40 dB SNR in terms of accuracy.

Now have a detailed look, at first at the results of the multi-class detection in Table 1. In clean condition, the accuracy is 56.9 %, increasing to 60.1 % at 40 dB SNR, an then decreasing to 36.1 % at 0 dB SNR. Having a look at the ratios of substitutions, deletions, and insertions, it is clearly visible that under all SNR conditions the accuracy is mainly affected by substitutions, about 37 % slightly differing by about 1 % over all SNR. The ratio of deletions and insertions overall increases with higher SNR.

The time-dependent accuracy is as well increasing from 54.9 % in clean condition to 55.7 % at 40 dB SNR, and then

	SNR					
	clean	40 dB	30 dB	20 dB	10 dB	0 dB
S/N	0.0	0.0	0.0	0.0	0.0	0.0
D/N	1.4	1.1	4.1	6.1	7.7	18.9
I/N	3.3	1.3	0.9	2.9	5.7	8.0
R/N	5.2	8.2	10.9	11.8	10.0	11.9
ACC	95.3	97.6	95.0	91.0	86.6	73.1
tACC	90.1	89.3	84.1	79.2	76.6	61.2

Table 2: Single-class detection: ACC, tACC, ratio of substitutions S/N, deletions D/N, insertions I/N, and time penalties R/N for different SNRs, in %, respectively; N=100 events.

decreasing to 34.0 % at 0 dB SNR. As the time-dependent accuracy is always lower than the accuracy (see formulas (2) and (4)), this trend is not quite surprising. But the gap between ACC and tACC of 2 % up to 5.6 % shows, that there is indeed a temporal misalignment of the recognition results and the ground truth. Having a look at the ratio of R , which is the sum over all weights $R(\Delta_t)$, up to 5.6 % (at 20 dB) of all correctly recognized events are temporally misaligned by more than 0.25 seconds.

Now have a look at the results of the single-class detection in Table 2. As already stated before, no substitutions happen as any event counts as an critical event and in this way can not be substituted by another. The accuracy increases from 96.3 % under clean conditions to 97.6 % at 40 dB SNR and then decreases to 73.1 % at 0 dB SNR. Having a look on the ratio of insertions and deletions, overall the same effect as in the multi-class detection can be seen. But for almost all SNR the ratio of insertions is slightly higher, whereas the ratio of deletions is slightly lower, compared to the multi-class detection.

For the single-class detection the time-dependent accuracy (in contrast to the multi-class detection) decreases from 90.1 % at clean to 61.1 % at 0 dB SNR. Furthermore the gap between ACC and tACC is much bigger, as in the multi-class detection. The gap differs from 5.2 % (clean) up to 11.9 %. Looking again at the ratio of R this means that at 0 dB SNR 11.9 % of all correctly recognized events are temporally misaligned by more than 0.25 seconds. Through the single-class detection the amount of events that count as correctly recognized rises, and so the temporal misalignment does as well.

5. Conclusions

In this contribution we propose an acoustic event detection (AED) framework and a new evaluation metric for surveillance in cars based on continuous audio from the handsfree microphone. We have evaluated the framework on critical acoustic events embedded in in-car street noise of a parked car. In order to take into account the correct timing of the correctly recognized events, we proposed a new time-dependent accuracy measure tACC that in general is smaller than the conventional accuracy ACC. The new evaluation metric shows that depending on the evaluation setup and SNR, conventional accuracy measures such as ACC may be too optimistic, since an alarm being (correctly) detected but way too late is not penalized witch ACC. Within surveillance applications this is important, however, as not only the correct event has to be detected, but also at the right time.

6. References

- [1] B. Davis and R. DeLong, "Combined Remote Key Control and Immobilization System for Vehicle Security," in *Power Electronics in Transportation, 1996, IEEE*, Oct 1996, pp. 125–132.
- [2] K. Khangura, N. V. Middleton, and M. Ollivier, "Vehicle Anti-Theft System uses Radio Frequency Identification," in *Proc. of IEE Colloquium on Vehicle Security Systems*, London, England, Oct. 1993, pp. 4/1–4/7.
- [3] H. Ruser and V. Magori, "Highly Sensitive Motion Detection with a Combined Microwave-Ultrasonic Sensor," *Sensors and Actuators A: Physical*, vol. 67, pp. 125–132, 1998.
- [4] S. W. Redfern, "A Radar Based Mass Movement Sensor for Automotive Security Applications," in *Proc. of IEE Colloquium on Vehicle Security Systems*, London, England, Oct. 1993, pp. 5/1–5/3.
- [5] Z. Liu and G. He, "A Vehicle Anti-Theft and Alarm System Based on Computer Vision," in *Proc. of IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, Xian, China, Oct. 2005, pp. 326–330.
- [6] J. Xiao and H. Feng, "A Low-Cost Extendable Framework for Embedded Smart Car Security System," in *Proc. of International Conference on Networking, Sensing and Control (ICNSC)*, Okayama, Japan, Mar. 2009, pp. 829–833.
- [7] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification Systems," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds. Springer Berlin Heidelberg, 2007, vol. 4122, pp. 311–322.
- [8] J. Chen, A. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom Activity Monitoring Based on Sound," in *Pervasive Computing*, ser. Lecture Notes in Computer Science, H.-W. Gellersen, R. Want, and A. Schmidt, Eds. Springer Berlin Heidelberg, 2005, vol. 3468, pp. 47–61.
- [9] W. Huang, T.-K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream Detection for Home Applications," in *Proc. of 5th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Taichung, Taiwan, Jun. 2010, pp. 2115–2120.
- [10] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems," in *Proc. of 15th European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, Sep. 2007, pp. 21–26.
- [11] C. Clavel, T. Ehrette, and G. Richard, "Events Detection for an Audio-Based Surveillance System," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, Amsterdam, The Netherlands, Jul. 2005, pp. 1306–1309.
- [12] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini, "Automatic Sound Detection and Recognition for Noisy Environment," in *Proc. of 10th European Signal Processing Conference (EUSIPCO)*, Tampere, Finland, Sep. 2000, pp. 1033–1036.
- [13] A. Temko and C. Nadeu, "Acoustic Event Detection in Meeting-Room Environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, Oct. 2009.
- [14] L.-H. Cai, L. Lu, A. Hanjalic, and H.-J. Zhang, "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, May 2006.
- [15] F. Weninger and B. Schuller, "Audio Recognition in the Wild: Static and Dynamic Classification on a Real-World Database of Animal Vocalizations," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 337–340.
- [16] M. Cowling and R. Sitte, "Comparison of Techniques for Environmental Sound Recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895–2907, Nov. 2003.
- [17] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
- [18] J. Dennis, H. D. Tran, and E.-S. Chng, "Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 367–377, Oct. 2013.
- [19] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-Dependent Sound Event Detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–13, Jan. 2013.
- [20] A. Mesaros, T. Heittola, and A. Klapuri, "Latent Semantic Analysis in Sound Event Detection," in *Proc. of 19th European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug.-Sep. 2011, pp. 1307–1311.
- [21] Q. Huang and S. Cox, "Hierarchical Language Modeling for Audio Events Detection in a Sports Game," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 2286–2289.
- [22] P. Transfeld and T. Fingscheidt, "Towards Acoustic Event Detection for Surveillance in Cars," in *Proc. of 11th ITG Conference on Speech Communication*, Erlangen, Germany, Sep. 2014, pp. 113–116.
- [23] ETSI, *ETSI ES 202 050 V1.1.5 Advanced Front-End Feature Extraction Algorithm*, European Telecommunication Standards Institute, Jan. 2007.
- [24] S. Young, *The HTK Book*, Cambridge University Engineering Department, 1995.
- [25] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound Scene Data Collection in Real Acoustical Environments," *The Journal of the Acoustic Society of Japan*, vol. 20, no. 3, pp. 225–231, May 1999.
- [26] ITU-T, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Dec. 2011.