ACOUSTIC GEO-SENSING: RECOGNISING CYCLISTS' ROUTE, ROUTE DIRECTION, AND ROUTE PROGRESS FROM CELL-PHONE AUDIO

Björn Schuller[1,2], Florian Pokorny, Stefan Ladstätter[2], Maria Fellner[1], Franz Graf[1], Lucas Paletta[2]

JOANNEUM RESEARCH Forschungsgesellschaft mbH
DIGITAL – Institute for Information and Communication Technologies
[1]Research Group for Space and Acoustics
[2]Research Group for Remote Sensing and Geoinformation
Steyrergasse 17, 8010 Graz, Austria
firstname.lastname@joanneum.at

ABSTRACT

We introduce the discipline of Acoustic Geo-Sensing (AGS) that deals with the connection of acoustics and geoposition, i. e., 'local audio' – focussing on spatial rather than on temporal aspects. We motivate this field of research, and give an example by automatic determination of a cyclist's route between determined start and endpoints, the direction she advances on this route, and the progress made from cell-phone audio. The Graz Cell-phone Cycle Corpus of 16 hours audio is introduced to this end. A standardised acoustic feature set ensures reproducibility throughout extensive experimentation aiming to reveal maximal spatiotemporal resolution. In the result, principle feasibility is shown by unsupervised clustering and all presented tasks can be solved at high accuracies and correlation within Random Forest classification and Additive Regression.

Index Terms— Acoustic Geo-Sensing, Ambient Audio, Intelligent Audio Analysis, Computer Audition

1. INTRODUCTION

"When hearing a sound, our imagination often plays an important part in recognising *what* it might be" [1]. However, common experience has it that, one is also able to think *where* it might be. In this light, we want to introduce the field of Acoustic Geo-Sensing, motivate its potential, and describe the exemplary demonstration of general feasibility of three selected tasks.

1.1. Defining Acoustic Geo-Sensing

Geosensors are such devices that measure or receive environmental stimuli that can be geographically referenced. *Acoustic Geo-Sensing*, or AGS for short, is consequently related to analysing acoustic stimuli alongside their geographical reference. As such, emphasis is put on the location in the spatiotemporal continuum, i. e., we are mostly interested in finding acoustic relation to the *fixed* geolocation utmost independent of the time. Obviously, time has a significant influence if one thinks of outdoor recordings which are acoustically influenced by weather conditions or time of day and working day vs. holiday in urban environments. However, interesting applications can also include the recognition of spatiotemporal equivalence of multiple sensors, for example to determine whether two persons' phone calls originate from the same geoposition at the same time.

1.2. Application Potential

A number of applications opens up including commercially interesting ones and such that posses the potential to have an impact on society. To name a few, let us begin with applications where the geoinformation is known alongside the acoustic recording. These include acoustic monitoring for public safety, e. g., by crowdsourcing from persons willing to contribute to such a service and transmitting audio data from their cell-phones or capture devices mounted on cars, etc., e. g., to pre-filter locations of potential accidents [2], aggressive behaviour, etc. Next 'acoustic maps' can be thought of similar to connecting vision with geolocation [3], either by collecting sounds from different geolocations and allowing for acoustic playback, e. g., by 'mouseover' events. This may require acoustic thumbnailing or summarisation to identify or collect the most characteristic audio fingerprint(s) over time for a certain period. An interesting variant then will be acoustic pleasantness maps [4] that measure the agreeableness of the acoustic environment in a certain location. An automatically generated 'acoustic diary' of a journey could be another option: Such a diary could either find acoustic thumbnails along the route or find the most prominent ones and show them on a map.

If the location needs to be inferred from the audio, e. g., from phone calls, a further range of applications opens up. Advertisement placement suiting the environmental condition by 'any sensor' of a remote device (thus including audio capture) has recently been patented [5]. In forensics, identification and verification of location by audio can be of interest: Imagine a phone call recording without knowledge of the geoposition, e. g., by tracking. Then, either a location identification or verification can be based on the acoustic recording. Further, emergency hotlines can infer the position of callers in case they are not able to give coordinates by themselves, for example, because they dial the number secretly while being kept hostage or similar. Such hotlines could also prioritise new callers from new positions in case of mass catastrophes: Given an incident that involves several hundred or thousands of people, emergency hotlines will likely receive hundreds of calls providing similar information giving little to no chance to callers in need outside of this event. Given a system that identifies geographic similarity, it could prefer calls

The authors acknowledge funding from the Advanced Audio Processing project by the Austrian Research Promotion Agency and the European Community's Seventh Framework Programme (FP7/2007 – 2013) under grant agreement No. 288587 (MASELTOV). The responsibility lies with the authors. The authors express their gratitude for permission to use map material from Digitaler Atlas der Steiermark, Abteilung 7 – Landes und Gemeindeentwicklung, Referat Statistik und Geoinformation, GIS-Steiermark, Austria.

Fig. 1. From left to right: four-channel high quality audio recorder with wind shield and helmet mounting in 90° angle of inclination, smart phone with windshield and positioning in the backpack. The additional GPS tracker was positioned inside the backpack ensuring sufficient satellite visibility.

from new destinations to switch to human operators first.

1.3. Example

In this work, we want to exemplify feasibility of three research questions in the field of Acoustic Geo-Sensing: Can it be inferred from the audio recording (I) which route between two endpoints was taken by a cyclist – the *route recognition*, (II) in which direction the cyclist proceeds along the route – the *route direction recognition*, and (III) which progress was made along this route– the *route progress recognition*. In fact, this is what we sometimes do ourselves, e. g., when talking to family members over the phone which are on their way home or similar: In such a case, one can often estimate from the ambient audio and acoustics where on their way they roughly are.

A particular concern will then be the spatiotemporal resolution, i. e., which amount of time is needed for a sufficiently reliable conclusion and which local resolution results from it. Obviously, a limit similar to Heisenberg's uncertainty relation exists: At some point the window in time for chunking will be too small to lead to a perfect spatial determination, in particular given the variability of acoustics over time.

1.4. Relation to prior work

There have been several recordings of environmental sounds for Acoustic Event Detection or Classification [6] and Computational Auditory Scene Analysis [7] or more general Computer Audition purposes [8]. Recognition of sound events is increasingly pursued, e. g., in [9] – also in the urban environment [10], yet, not related to (*exact*) geographic location. This is similarly true for classifying acoustic ambience. The field of Acoustic / Sound (Source) Localisation [11] infers local information from sound, however, usually without relation to the audio *content* and again without relation to a *fixed* geographical location. The field of Acoustic Remote Sensing mostly deals with tomography [12]. As such, acoustic waves are used for imaging by sections or sectioning. However, this name was already set into relation with connecting acoustics and position, e. g., for ecological area analysis by acoustics of wildlife [13]. Most closely to the introduced Acoustic Geo-Sensing may be works within the MediaEval evaluation campaign's placing task for location determination in Flickr videos – contributions focussing on audio were made in this context, e. g., for the recognition of the city of recording out of 18 [14]. We are not aware of any prior work on acoustic route monitoring – let alone in the specific case of cycling.

The remainder of this paper is structured as follows: In Section 2 we introduce the Graz Cell-phone Cycle Corpus recorded for experimentation, in Section 3 we report experimental results for the above named tasks, before concluding in Section 4. There, a number of concrete steps for future development are given.

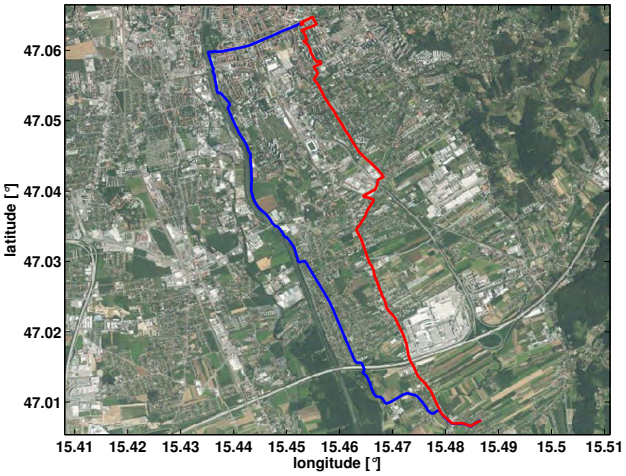

Fig. 2. Two alternative routes (left/blue: RIVER route, right/red: CITY route) recorded in the Graz area/Austria from Bucherlweg/Grambach to Steyrergasse/Graz. The starting point is found at the bottom – note that, for better visibility only the red colour is used from both routes' start until their branch. The endpoint is at the top with only the final parking lot as overlap.

	#	#	[m]	T [min:sec]			
Route	fw	bw	L	min	mean	sdev	max
CITY	6	3	8431.9	22:44	25:44	1:53	28:12
RIVER	17	7	9546.1	27:02	30:29	2:48	39:36

Table 1. *Statistics of the 16 h GC*^{3} *route takes: number per forward (fw) and backward (bw) direction, length (L), and duration (T).*

2. THE GRAZ CELL-PHONE CYCLE CORPUS

2.1. Recording

As cycle, a 26" Scott Elite Racing mountain bike was used. To record audio data during bicycle rides without particularly demanding hardware conditions, an Android smart phone of the brand Samsung Galaxy Nexus was used as main device. This device samples spatially equidistant and was operated at 0.2 m^{-1}. The phone was loosely located in a side pocket of a backpack worn on the back of the cyclist. As only additional measure, a wind shield from an ordinary microphone was put on top of the device (cf. Figure 1). The standard media recording APIs of Android use an AMR codec with poor quality. Therefore, the audiostream was accessed directly and saved uncompressed as PCM at 44 kHz, 16 bit. The recording component was implemented as a service and thus could run in the background with the phone locked and the screen turned off. For GPS measurements, the phone utilises the SIRFStar IV GSD4t chipset, which enables maintaining position locks also in challenging environments such as urban canyons or dense forests. In addition to this, limited motion sensing is available through an InvenSense MPU-3050 accelerometer unit. This sensor contains a MEMS accelerometer and a gyroscope. Linear and angular accelerations can be captured at a sampling rate of up to 100 Hz. Since the audio recording thread already puts the CPU under considerable load, the actual achievable sampling rate was in the range of 10 – 20 Hz, though. The MARIA application [15] for public transport guidance was adjusted in a way to log the audio alongside GPS coordinates and acceleration data from the motion

Fig. 3. Left to right: Routes from start at home to office – CITY route (top) and RIVER route (bottom). Pictures above each other are taken at same distance from start and evenly distributed along the pathes.

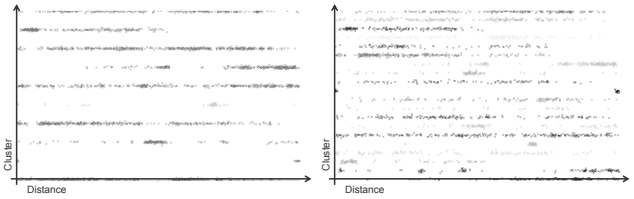

Fig. 4. Qualitative unsupervised kMeans clustering of the 17 takes of the RIVER route in forward direction: 10 (left) and 20 (right) clusters. Y-axis: cluster number (each cluster is shown in one horizontal line – a cluster's presence is marked by 'x' symbols along this line. X-axis: distance from start (left) to end (right). Grey shading is only used for better visibility.

sensor. This allowed to foster synchrony between audio recording and GPS and additional sensor data. However, depending on the phone hardware, deviation may occurr – a maximum of 1 s was measured for a 42 min take. This deviation was considered as tolerable. In addition, a Zoom H2 four-channel recording device recording at 48 kHz, 16 bit was mounted to the helmet of the cyclist for high quality audio (cf. Figure 1). Finally, a secondary GPS sensor was used for verification purposes: The NAVIN Mini Homer GPS tracker was operated at a sample rate of 0.2 Hz. The recordings were made by the second author on his way to and from work at various times throughout the day. No weekend or night takes are contained, which renders the current analyses optimistic as this may impact traffic load, however, without too strong limitation of application range. Two different routes were chosen on purpose, see Figure 2. Both routes are common alternatives for the recording cyclist on his way from/to work and were followed strictly throughout repeated takes. As can be seen in the figure, one route is characterised by going alongside the Mur river for roughly half of the route. The river is, however, hardly audible in the recordings. Pictures along the route are visible in Figure 3 for illustration. Table 1 shows statistics of the routes.

2.2. Annotation, Partitioning, and Release

In principle, the audio is annotated by the GPS track without further labelling effort. However, a range of additional information was noted alongside. The following protocol was established for metadata transcription: date and time of start, type of movement (cycling, walking, stationary), from/to (addresses), position of recorders, weather condition (cloudy / clear / partly cloudy / heavy cloud, dry / drizzle / rain / heavy rain / thunderstorm, temperature in °C). Further, the GPS coordinate tuples were transformed to distance values by cumulative Euclidean distance between two GPS sample points. This allows for spatially equidistant chunking of the audio by non-overlapping windows. Alternatively, temporally equidistant chunking will be considered for comparison. While one can expect spatial equidistance to be more precise, an application may often not be able to chunk in such a way, as it first does not have the spatial information available.

LLD ($16 \cdot 2$)	**Functionals** (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel. position, range
(Δ) MFCC 1–12	linear regression: offset, slope, MSE

Table 2. *Extracted audio features: low-level descriptors (LLDs) and functionals. MSE: mean square error.*

To ensure independent testing and development, the recorded routes were divided into three partitions. Train and development data are united after the optimisation phase for testing. This partitioning was carried out once in chronological order taking the first third of the CITY and RIVER routes, each, for training, the next third for development, and the last for independent testing. The data can be obtained freely per request. In the ongoing, we refer to the data set by the Graz Cell-phone Cycle Corpus or GC^3 for short reference.

3. EXPERIMENTS

In this work, all experiments are carried out on the cell-phone audio takes to demonstrate feasibility even in low audio quality condition without extra mounting effort. Further, we exploit only GPS information for training and testing from the additional sensor data recordings and no metadata for the moment. To foster reproducibility of findings, we use the openSMILE feature extractor [16] with a standardised feature set, and the Weka toolkit for classification [17]. We decided for the set of the INTERSPEECH 2009 Challenge event [18]. This set was preferred over other standards such as given by the MPEG 7 low-level descriptors (LLDs) [19], as also the implementation of features is well defined and accessible by the open source openSMILE extractor [16]. 16 LLDs are contained. To each of these, the delta coefficients are additionally computed. Next, 12 functionals are applied on a chunk basis as depicted in Table 2. Thus, the total feature vector per chunk contains $16 \cdot 2 \cdot 12 = 384$ attributes.

Classification is performed by Random Forests. A number of 30 unpruned trees were found a good choice on development data across tasks. In a similar fashion, Additive Regression with 50 iterations of Simple Regression as regression learner was found well suited for regression in the case of continuous route progress prediction. All results are shown in Table 3 – in the following, details of the experiments are given. The results are reported by means of unweighted accuracy (UA: recall of all classes added and divided by number of classes to respect imbalances) and 'normal' weighted accuracy (WA) for classification, and by correlation coefficient (CC) for regression. UA chance level in case of binary or ternary decision resembles 50 % and 33 %, respectively. For training and testing the same chunk size is used, each. Note that, by principle different chunk sizes thus lead to different numbers of testing instances – results can thus not be directly compared across different chunk sizes.

We first carried out a qualitative pre-study by looking at unsupervised kMeans clustering over all takes of the RIVER route in forward direction, as it possesses the highest number of takes (cf. Table 1). In this study, spatially equidistant chunking is more meaningful, and we chose a resolution of 50 m for visualisation in Figure 4 – we had further investigated 100 m and 500 m with similar effects. In this figure, one notices several acoustic clusters appear subsequently along the

	Route			Direction			Progress					
L_{chunk}	# test	UA	WA	# test	UA	WA	# test	UA2	WA2	UA3	WA3	CC
equitemporal chunking [sec]												
5	2 995	62.0	65.5	3 080	73.3	76.7	2 358	68.0	68.4	56.5	57.2	.440
10	1 501	65.6	68.1	1 543	76.1	78.8	1 181	68.0	68.2	60.1	60.5	.446
30	505	65.4	70.5	519	73.2	78.0	397	75.3	75.6	60.8	61.5	.568
60	255	78.4	78.4	261	76.3	78.5	200	67.8	68.5	69.2	60.9	.601
equispatial chunking [m]												
50	1 488	67.0	72.5	1 533	76.9	79.1	1 151	67.8	67.8	55.9	55.9	.446
100	746	69.0	73.7	769	81.2	81.9	577	68.8	68.8	61.2	61.2	.369
500	153	72.3	76.5	159	77.9	81.8	119	84.0	84.0	68.9	68.7	.752

Table 3. *Route and direction recognition in dependency of chunk length L_{chunk} in sec or m: unweighted (UA) and (weighted, WA) accuracy in %, route progress estimation: either for binary (UA2/WA2) or ternary (UA3/WA3) quantisation or by correlation coefficient (CC) for continuous modelling, and number of respective resulting test instances after chunking.*

distance axis – this can be interpreted as strong indication that indeed, even over repeated recordings, certain geographic areas tend to be marked by specific types of acoustics.

For *route recognition*, we chose to use only 'forward' direction instances (home to work), as more exist from these for both routes and we wanted to keep the question of direction and route separate. Training and development data are united for final testing. Due to imbalance among the two classes, we use random downsampling without replacement to 60 % and uniform class distribution for the overall learning data. Table 3 shows results as high as 78.4 % UA given one minute of audio.

For *route direction recognition*, we chose to use only RIVER instances, as more exist from this route and – as stated above – we wanted to keep the question of direction and route separate. Training and development data are united for final testing. Due to imbalance among the two classes, we use random downsampling without replacement to 40 % and uniform class distribution for the overall learning data. As can be seen in Table 3, the recognition reaches 78.8 % UA from as little as 10 sec of audio.

In the final question of *route progress recognition*, we analyse the progress along the RIVER route in forward direction. This is a consequent choice not only because most examples exist for this route, but, by that we so far first looked at deciding if or not we are dealing with the river route, then in which direction it is being progressed – forward or backward – and now finally decide on the progress in the forward case. We consider regression by comparison with the distance in m. Upsampling of training material is not necessary in this case, as the tracks always contain each distance from beginning to end. Here, 75.2 % UA can be reached for binary (beginning/end) decisions from 30 sec of 'observation' audio. For ternary (beginning/middle/end) decisions, remarkable 69.2 % are obtained using a 1 min audio chunk length. Finally, fully continuous regression determining the distance in metres leads to a CC around .6 as of 30 sec of audio. Obviously, equispatial chunking is in particular suited in this task, and .752 CC are reached for 500 m chunking.

4. CONCLUSION

We introduced the field of Acoustic Geo-Sensing and exemplified it by three tasks related to inferring geoposition by acoustic information. Unsupervised chunking visually demonstrates that 'similar acoustics' tend to appear in local neighbourhood even over time. Then, we showed the feasibility of route, route direction, and route progress determination from audio with results highly significantly above chance levels (p value $< 10^{-3}$ for all results in Table 3 in one-sided z-testing respecting varying number of test instances).

In future work, we aim to compare unsupervised clustering with semantically meaningful tags such as 'park', 'crossing', etc. Next, the data may be added by further cities and routes, and other forms of transportation such as car riding or walking. Further, we will continue our first additional recordings with a mobile electrodermal activity and skin temperature sensor (Affectiva's Q Sensor) [20] to correlate indication of human arousal (or valence) with the current acoustic scenery, such as 'alongside river' or 'downtown city'. This can also be set in relation to prediction of sound emotion analysis [21], such as the arousal regression presented in [22, 23]. In a similar fashion, higher level features that determine sound event types [24, 6] can provide additional information over LLD-type feature information. In particular, 'bag of audio words' [25] seems a promising approach as data-trained feature type. These could be based on variable length audio words, for example by using Bayesian Information Criterion (BIC) or similar to detect audio 'word boundaries'. Also, N-Grams of audio words can be thought of. As for pre-processing, the audio could be (blindly) separated into multiple sources, such as by non-negative matrix factorisation (NMF) [26, 27]. In fact, semi-supervised or supervised NMF-type activation features would allow for another type of higher level audio features such as degree of presence of car events or similar [28]. Further, modelling of temporal context seems promising, such as by long short-term memory recurrent architectures [29] – in particular when routes are analysed. These can be combined in an elegant manner with bottleneck topologies for self-learning feature space compression [30]. In addition, dynamic modelling approaches such as by dynamic time warping in case of few references or hidden Markov models (HMMs) shall be evaluated – potentially in tandem operation. HMM forward and reverse path models could have the same states, but in reversed orders. Next, variable chunk lengths such as by BIC and 'multi-condition' style training with different lengths can be thought of. In case of dropout of GPS, e. g., in the event of tunnels or partial indoor activity, interpolation will need to be considered – this was not required along the routes recorded in our experiments. In fact, AGS may even be of help to keep track in such a dropout event in navigation systems, and the combination of audio and GPS data could be used to ensure plausibility (lately, cases of GPS hijacking have been reported, e. g., of drones, by intentional emission of erroneous GPS information). Also, obviously, further multimodal combination such as audiovisual geo-sensing can be thought of in combination with video information [31]. Finally, the recorded movement sensor data would allow for research questions such as inferring movement from the audio.

5. REFERENCES

[1] M.A. Forrester, "Auditory perception and sound as event: theorising sound imagery in psychology," *Sound Journal*, 2000, no pagination.

[2] M. Fellner, F. Graf, H. Rainer, and B. Rettenbacher, "Intelligent Acoustic Solutions in Road Traffic Telematics," *ÖGAI Journal*, vol. 26, no. 4, pp. 2–9, 2012.

[3] L. Vande Velde, P. Luley, A. Almer, C. Seifert, and L. Paletta, "Intelligent Maps for Vision Enhanced Mobile Interfaces in Urban Scenarios," in *Proc. 12th World Congress on Intelligent Transportation Systems and Services*, Toronto, Canada, 2006, pp. 566–572.

[4] T.H. Park, B. Miller, A. Shrestha, S. Lee, J. Turner, and A. Marse, "Citygram One: Visualizing Urban Acoustic Ecology," in *Proc. Digital Humanities*, Hamburg, Germany, 2012, ADHO.

[5] T.B. Heath, "Advertising based on environmental conditions," US Patent 8,138,930, 2012.

[6] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. EUSIPCO*, Aalborg, Denmark, 2010.

[7] D.L. Wang and G.J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, IEEE Press, 2006.

[8] B. Gygi and V. Shafiro, "Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, Article ID: 654914, 12 pages, 2010.

[9] H.D. Tran and H. Li, "Sound Event Recognition With Probabilistic Distance SVMs," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, pp. 1556–1568, 2011.

[10] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Automatic recognition of urban environmental sound events," in *Proc. CIP*. 2008, pp. 110–113, EURASIP.

[11] G. Doblinger, "Localization and Tracking of Acoustical Sources," in *Topics in Acoustic Echo and Noise Control*, pp. 91–121. Springer, Berlin – Heidelberg, 2006.

[12] W. Munk and C. Wunsch, "Ocean acoustic tomography: a scheme for large scale monitoring," *Deep-Sea Research*, vol. 26A, pp. 123–161, 1979.

[13] G. Sanchez, R.C. Maher, and S. Gage, "Ecological and environmental acoustic remote sensor (EcoEARS) application for long-term monitoring and assessment of wildlife," in *Proc. U.S. Department of Defense Threatened, Endangered and at-Risk Species Research Symposium and Workshop*, Baltimore, MD, USA, 2005.

[14] H. Lei, J. Choi, and G. Friedland, "City-Identification on Flickr Videos Using Acoustic Features," Tech. Rep. TR-11-001, ICSI, Berkley, USA, 2011.

[15] L. Paletta, R. Sefelin, J. Ortner, J. Manninger, R. Wallner, M. Hammani-birnstingl, V. Radoczky, P. Luley, P. Scheitz, O. Rath, M. Tscheligi, B. Moser, K. Amlacher, and A. Almer, "MARIA - Mobile Assistance for Barrier-Free Mobility in Public Transportation," in *Proc. CORP*, Vienna, Austria, 2010, pp. 1151–1155.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. 9th ACM Multimedia*, Florence, Italy, 2010, pp. 1459–1462, ACM.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[18] B. Schuller, "The Computational Paralinguistics Challenge," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, July 2012.

[19] H.-G. Kim, J. J. Burred, and T. Sikora, "How efficient is MPEG-7 for general sound recognition?," in *Proc. of AES 25th Int. Conf.*, London, UK, June 2004.

[20] M.Z. Poh, N.C. Swenson, and R.W. Picard, "A Wearable Sensor for Unobtrusive, Longterm Assessment of Electrodermal Activity," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1243–1252, May 2010.

[21] M. Fellner, H. Rainer, G. Neubauer, M. Dvorzak, M. Kropf, R. Krainz, M. Cvetkovic, and H. Sulzbacher, "Predicting the attractivity of a sound by its features - A tool for sound design in casino game industry," in *Proc. 1st EAA-EuroRegio*, Ljubljana, Slovenia, 2010, European Acoustics Association, no pagination.

[22] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic Recognition of Emotion Evoked by General Sound Events," in *Proc. 37th ICASSP*, Kyoto, Japan, 2012, pp. 341–344, IEEE.

[23] S. Sundaram and R. Schleicher, "Towards evaluation of example-based audio retrieval system using affective dimensions," in *Proc. ICME*, Singapore, 2010, pp. 573–577, IEEE.

[24] Z. Zhang and B. Schuller, "Semi-supervised Learning Helps in Sound Event Classification," in *Proc. 37th ICASSP*, Kyoto, Japan, 2012, pp. 333–336, IEEE.

[25] M. Riley, E. Heinen, and J. Ghosh, "A Text Retrieval Approach to Content-based Audio Retrieval," in *Proc. ISMIR*, Philadelphia, PA, USA, 2008, pp. 295–300.

[26] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[27] F. Weninger and B. Schuller, "Optimization and Parallelization of Monaural Source Separation Algorithms in the openBliSSART Toolkit," *Journal of Signal Processing Systems*, vol. 69, no. 3, pp. 267–277, 2012.

[28] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, "Non-Negative Matrix Factorization as Noise-Robust Feature Extractor for Speech Recognition," in *Proc. 35th ICASSP*, Dallas, TX, USA, 2010, pp. 4562–4565, IEEE.

[29] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. ICANN*, Warsaw, Poland, 2005, pp. 602–610.

[30] M. Wöllmer, B. Schuller, and G. Rigoll, "A Novel Bottleneck-BLSTM Front-End for Feature-Level Context Modeling in Conversational Speech Recognition," in *Proc. 12th Biannual IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2011*, Big Island, HY, 2011, pp. 36–41, IEEE.

[31] L. Paletta, G. Fritz, C. Seifert, P. Luley, and Almer A., "A Mobile Vision System for Multimedia Tourist Applications in Urban Environment," in *Proc. IEEE Intelligent Transportation System Conf., ITSC*, Toronto, Canada, 2006, pp. 566–572.