ACOUSTIC EVENT CLASSIFICATION FOR AMBIENT ASSISTED LIVING AND HEALTHCARE ENVIRONMENTS

Hussein Hussein[1], *Marc Ritter*[1], *Robert Manthey*[1], *Jan Schloßhauer*[2], *Etienne Fabian*[2] *and Manuel Heinzig*[1]

[1] *Junior Professorship Media Computing, Technische Universität Chemnitz, D-09107 Chemnitz, Germany*
[2] *Intenta GmbH, D-09125 Chemnitz, Germany*

{hussein.hussein, marc.ritter}@informatik.tu-chemnitz.de,
{robert.manthey, manuel.heinzig}@informatik.tu-chemnitz.de,
{j.schlosshauer, e.fabian}@intenta.de

Abstract: Acoustic events that are produced by people can be used to recognize activities or other critical behavior. This contribution presents our first experiments on acoustic event classification for utilization in the sector of healthcare. Ten acoustic events, including speech and non-speech events, which are usually occurred in this field are defined. The database of acoustic events is collected in a recording studio and annotated manually. A variety of features and several classifiers have been proposed for classification of acoustic events in order to detect the best feature set and classifiers for the specified acoustic events. Low-level audio features and the corresponding delta features are utilized. Statistical functionals are applied to each of the features and delta features. The best obtained classificaton results, calculated by the F-Measure, for the ten aoustic events with a feature set of 430 features is 92.50%.

1 Introduction

The number of elderly people will rise rapidly in the European Union (EU) with the demographic change. Older people often suffer from several chronic conditions and require long-term care solutions [1]. The long-term nursing home care is very expensive. Therefore, the demand for automatic systems to monitor the activities and health status is an effective solution. The human activity is reflected by acoustic events which are produced by the human body or by objects handled by humans [2]. The recognition of acoustic events that are produced by elderly or people with disabilities in Ambient Assisted Living (AAL) environments or by patients in hospital can be used to recognize Activities of Daily Living (ADL) or critical behavior. Nowadays, these places tend to be equipped with special sensors or devices that assist in daily living by collecting individual information for monitoring systems.

The recognition of acoustic events is the process of automatically determining specified acoustic events in an audio stream. Acoustic Event Classification (AEC) deals with isolated audio segments of events. Acoustic Event Detection (AED) refers to the identification of timestamps as well as types of events in continuous audio stream [2]. The detection and classification of different types of acoustic events is an important task in many fields of applications. AED is utilized in scene recognition where scenes are described in terms of elementary acoustic events, for example, outdoor, home and vehicle [3][4][5][6]. Another application in the field of AAL and healthcare environments, AED is used to monitor social activities and health status of older

people (e.g. activities occurring within a bathroom such as showering, washing-hands and brushing-teeth [7] and events showing the health status such as cough [8]) or to detect equipment alarms in a neonatal intensive care unit [9]. Furthermore, AED is utilized in smart homes to detect different types of events, for example, speech, walking steps, coffee spoon and mouse click [10] as well as in meeting room environments to detect events such as speech, paper work, chair moving and key jingle [2]. The acoustic analysis can be used in surveillance and security applications to recognize events related to car intrusion (e. g. metal scratch or breaking glass) [11]. In addition, the detection of acoustic events is utilized in applications of bioacoustics to recognize of animal sounds [12].

The process of event detection is based on feature extraction and classification. The extraction of suitable features is an important step for the detection of acoustic events. Various features and classifiers have been proposed in AED. The most popular speech perception features such as Mel Frequency Cepstral Coefficients (MFCCs) are utilized with the Hidden Markov Models (HMM) [7][2][11]. The filterbank parameters used for calculation of MFCCs are designed mainly based on the properties of the human auditory system. The spectral structure of acoustic events is different from the spectrum of speech signals. In [13], the MFCCs features yielded the best results for environmental sound recognition. However, speech features are not necessarily suitable for the detection of acoustic events as described in [14][15]. A combination of several features (time-domain and frequency-domain features) is used in [3][4][5][15]. Feature transformations are applied to reduce the space dimension of feature vectors [4][14]. The most common classification techniques used for AED are HMMs [7][2][4][11], Gaussian Mixture Models (GMM) [3][5][9], the Support Vector Machines (SVM) classifier [2][5], and the K-Nearest Neighborhood (KNN) classifier [3][4][5].

This paper is organized as follows: Section 2 gives an overwiew on the *localizeIT* project for object tracking using audio-visual information. The selection of acoustic events and data acquisition are described in Section 3. Section 4 reviews the acoustic features used for the recognition of acoustic events. The experimental results are shown in Section 5. Finally, conclusions and future work are presented in Section 6.

2 Audio-Visual Based Object Tracking (localizeIT)

In the *localizeIT* project, which is funded by the Federal Ministry of Education and Research in the program of Entrepreneurial Regions, passive sensors are used for object tracking. Inside the tracking area, a number of sensors is installed as shown in Figure 1. There are optical sensors (cameras with mono as well as stereo optic) and microphone arrays. They were distributed inside the area at different positions, several directions and heights to focus on different tasks and optimize theirs effectivity. The optical sensores can be hindered due to different physical properties. In this case, the acoustic sensors can be used instead. The fusion of object tracking based on audio and visual information can be used to improve the tracking task.

The audio based object tracking requires different components such as voice activity detection, acoustic source separation, acoustic event detection and source localization. In this contribution, we consider only the classification of acoustic events.

3 Database Description

This section describes the acoustic events which are selected in the experiment and the collection of related data.

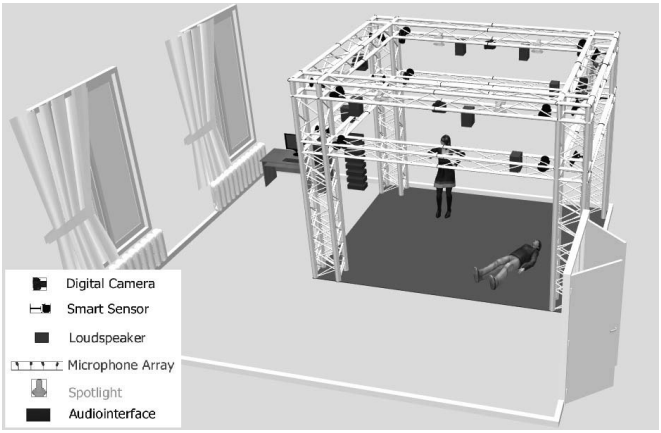

Figure 1 - Sketch of the Media Computing (MC) laboratory with depiction of its main components.

3.1 Selection of Acoustic Events

The selection of the acoustic events bases on the focus of detecting critical situations where people need help or try do something critical. There are different cases which can be consider in this field. Basic use case is an single elderly person which has fallen in the flat and is lying on the ground crying, whimpering and calling for help. Another use case is a confused person in police custody which is alone in a detention cell. Last use case is the clinical setting where mentally ill patients show critical behavior in the staying area.

So we defined ten acoustic events which are characteristic for the described use cases: help (calling of the speech signal “Hilfe” in German), scream, whimper, crying, quiet (long period of silence detected between audio segments of music, speech or background noise), strikes (strikes with an open hand on a wood plate), vandalism (destruction of furniture, strikes on a wood plate and sometimes scream), downfall of plate (downfall of a wood plate to the ground), dislocate furniture (movement of a furniture such as commode on ground), chair movement (movement of a chair on ground).

3.2 Data Acquisition

The existence of database recorded inside smart homes or healthcare environments is very important to evaluate the algorithms under real conditions. The possibility to record database in these environments is not availabe now. Hence, we recorded data from people between 25 and 82 years old in our laboratory (a subset of the data is recorded in the recording studio of the Chemnitz University of Technology and the other subset by the Intenta GmbH company) for the first research in this domain.

The audio data are recorded with an original sampling frequency of 44.1 kHz and a resolution of 16 bit (later down-sampled to 16 kHz with a resolution of 16 bit). Two measurement microphones (Behringer ECM-8000) connected to the audio interface (Focusrite Scarlett 2i2) are used for recording of audio data. The distance between microphones is about 20 cm and the distance to the acoustic source is between 20 to 30 cm. A total of 58 persons (11 female and 47 male) are participated in the recording of the acoustic events produced by human, i.e. help, scream, whimper and crying. The total amount of the acquired data from the ten acoustic events is 103 minutes. The acquired audio data is manually annotated using the FOLKER tool [16].

The number of audio files is as follows: help (175), scream (129), whimper (176), crying (192), quiet (45), strikes (475), vandalism (78), downfall of plate (84), dislocate furniture (126), chair movement (132).

4 Feature Analysis and Feature Extraction

Audio signals can be described by a set of characteristic features. The analysis of spectral structure of acoustic events is very important to detect the suitable features in an acoustic event recognition system. Figure 2 shows the spectrograms of some acoustic events. It can be seen that the spectral structure of non-speech acoustic events is different from that of human speech (help) or produced from human (scream). Non-speech acoustic events can be considered as specific sounds which are similar to noise. Therefore, speech perception features are not necessarily suitable to produce good results by the classification of acoustic events as described in [14][15].

Feature extraction can be split into two categories according to the processing domain (more detailed descriptions can be found in [3][13][15]):

Time-Domain Features

- Zero-Crossing Rate (ZCR)
- Short-time average energy

Frequency-Domain Features

- Pitch: fundamental frequency (F_0)
- Spectral features: band energy, spectral rolloff, spectral flux, spectral centroid
- Cepstral features: MFCCs
- Linear prediction features: Linear Prediction Coefficients (LPC)

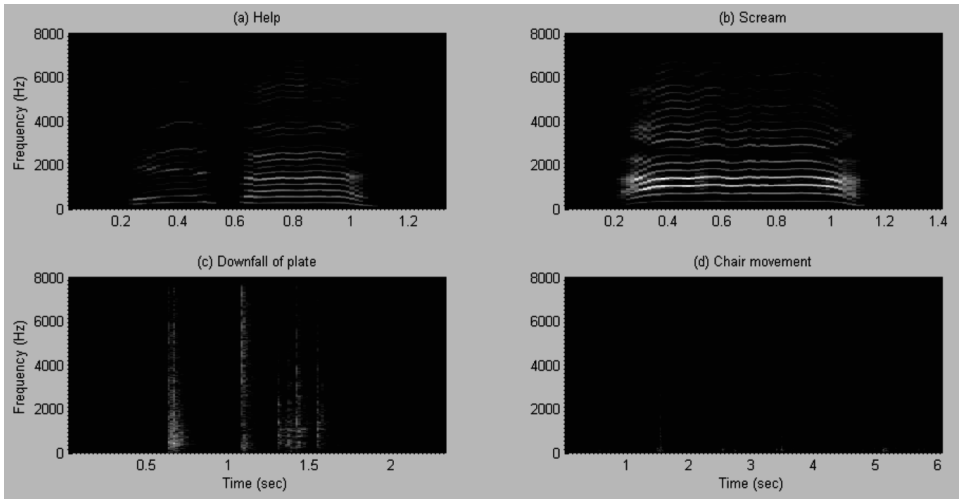

Figure 2 - Spectrograms of speech and non-speech acoustic events: help, scream, downfall of plate and chair movement.

5 Experiments and Results

The performance of a varity of audio features and several classifiers is investigated to determine the best suited features and classifiers for the classification of the isolated acoustic events.

5.1 Experimental Setup

The databse consists of 1612 recording samples. A subset of the database is used as training data (80%) and the remaining subset as test data (20%). The classifiers are trained on correctly labeled instances in order to generate models which represent the specific acoustic events. These models are used in the test phase to classify events in a specific class. The audio signals are split into frames using hamming window with a frame length of 30 msec and a step of 10 msec.

5.1.1 Features

We perform feature extraction on the event-level by using the OPENSMILE feature extraction tool [17]. The feature set contains features which result from Low-Level Descriptors (LLD) with the corresponding delta coefficients (ΔLLD) and statistical functionals applied to each of the LLD and ΔLLD. The LLD used are the features described in in Section 4. The features are energy, pitch, ZCR, spectral features, MFCCs, Line Spectral Frequencies (LSP) which computed from LPC coefficients. The default values of coefficients in OPENSMILE are used (12 cepstral coefficients for MFCCs and 8 linear predictive coding coefficients for LPC). The following statistical functionals are used for every feature: min, max, range, standard deviation and mean. Many feature vector sets are used:

- A (120 features): MFCCs
- B (50 features): Energy, pitch, ZCR
- C (230 features): Energy, pitch, ZCR, spectral features
- D (350 features): Energy, pitch, ZCR, spectral features, MFCCs
- E (430 features): Energy, pitch, ZCR, spectral features, MFCCs, LSP
- F (6373 features): Baseline feature set of the Computational Paralinguistics Challenge (ComParE) in the Interspeech 2013 [18].

5.1.2 Classification

A series of classifiers that have shown promising results in performing classification tasks in the literature are selected in order to determine the best suited classifier for the evaluation on the test set. In order to conduct the classification experiments, we applied the following machine learning algorithms with default values using the WEKA data mining toolkit [19]:

- ClassificationViaRegression (CVR) using regression methods for each class.
- K-Nearest Neighborhood (KNN) using the euclidean distance and 1-nearest neighbour.
- Sequential Minimal Optimisation (SMO) for training a SVM.

5.2 Experimental Results

The performace is measured using the F-Measure which is the harmonic mean between percision and recall. Figure 3 shows the classification results by applying six kinds of feature sets and three classifiers. The Figure shows that the feature set (A) of MFCCs yielded better results than the feature set (B) of energy, pitch and ZCR features by the three classifiers. The F-Measure for MFCCs features by the KNN classifier is 89.40%. The addition of spectral features to the energy, pitch and ZCR features, i.e. feature set (C), outperforms the results based on MFCCs features (A) by CVR and SMO classifiers. In contrast, results of feature set (C) are worse than that of feature set (A) by the KNN classifier. The insertion of more features, for example, adding the MFCCs features to feature set (C) as well as adding the LSP features to feature set (D), improves the classification results by all classifiers. Increasing the number of features of more than 5900 features from feature set (E) to (F) leads to a slightly improvement in the results by the CVR classifier. But this negatively impact the performance by the KNN and SMO classifiers. This indicates that the increasing of the number of features is not always helpful. The unnecessary large number of features is justified in [15] that data points become sparser and potentially irrelevant features affect negatively the performance of classification. In addition, the selection of a smaller feature set reduces the computational cost and running time. The best performance in the experiment is achieved with the energy, pitch, ZCR, spectral, MFCCs and LSP features (feature set E) by the SMO classifier with a F-Measure of 92.50%.

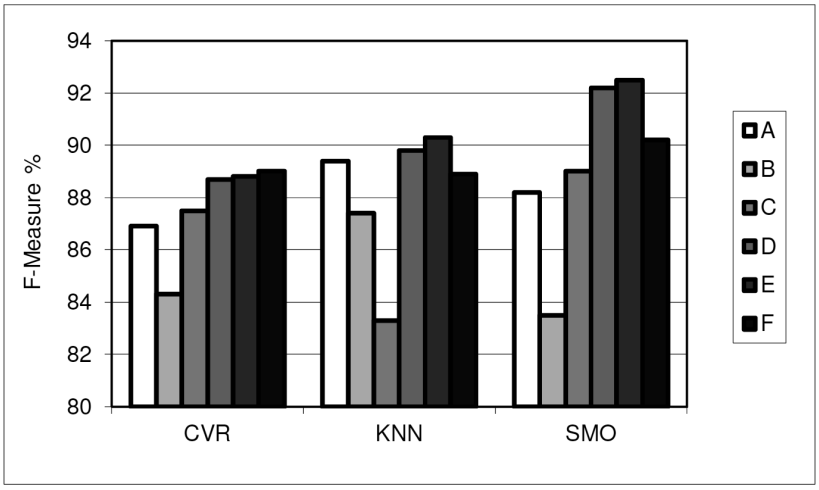

Figure 3 - Experimental results obtained on the test set by appling different feature sets and several classifiers.

6 Conclusion and Future Work

We presented our first experiments on classification of acoustic events within the *localizeIT* project for utilization in the field of smart homes and healthcare. Ten acoustic events, including speech and non-speech events, which are usually occur in this field are defined. The databse is collected in our laboratory and manually annotated. In order to detect the optimal feature set for the classification of the defined acoustic events, different kinds of audio features are

investigated with several classifiers. The F-Measure of the best results by using a subset of features, including energy, pitch, spectral, MFCCs and LSP features (430 features), and SMO classifier is 92.50%. It is verified that the increasing of the number of features is not always helpful to improve the classification results.

In the next steps, the effect of frame length and step size on the classification results can be investigated with different values of window and shift. The aim of the project is to implement a real-time acoustic event detection system and to test it under realistic environments. In addition, the object localization and tracking based on audio and visual information must be implemented.

7 Acknowledgements

This work is funded by the program of Entrepreneurial Regions InnoProfile-Transfer in the project group *localizeIT* (funding code 03IP608).

References

[1] European Commission Staff. Europe's demographic future: Facts and figures on challenges and opportunities, October 2007.

[2] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. Acoustic Event Detection and Classification in Smart-Room Environment: Evaluation of CHIL Project Systems. In *Proc. of IV Jornadas en Tecnología del Habla - The IV Biennial Workshop on Speech Technology*, Zaragoza, Spain, November 2006.

[3] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa. Computational Auditory Scene Recognition. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florida, USA, May 2002.

[4] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-Based Context Recognition. *Proc. of IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):321 – 329, January 2006.

[5] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric. Where am I? Scene Recognition for Mobile Robots using Audio Features. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 885–888, Toronto, Canada, Juli 2006.

[6] M. Ritter, M. Rickert, L. J. Chenchu, S. Kahl, R. Herms, H. Hussein, M. Heinzig, R. Manthey, D. Richter, G. S. Bahr, and M. Eibl. Technische Universität Chemnitz at TRECVID Instance Search 2015. In *Proceedings of TRECVID Workshop*, Gaithersburg, MD, USA, November 2015.

[7] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue. Bathroom Activity Monitoring Based on Sound. In *Proceedings of the Third International Conference on Pervasive Computing*, PERVASIVE'05, pages 47–61, Munich, Germany, 2005. Springer-Verlag.

[8] J. Schröder, S. Wabnik, P.W.J. van Hengel, and S. Goetze. *Detection and Classification of Acoustic Events for In-Home Care*, pages 181–195. Springer, 2011.

[9] G. Raboshchuk, P. Jančovič, C. Nadeu, A. P. Lilja, M. Köküer, B. M. Mahamud, and A. R. Veciana. Automatic Detection of Equipment Alarms in a Neonatal Intensive Care Unit Environment: A Knowledge-Based Approach. In *Proc. of Sixteenth Annual Conference of*

the International Speech Communication Association (Interspeech 2015), Dresden, Germany, September 2015.

[10] A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos, and P. Maragos. ATHENA: a Greek Multi-Sensory Database for Home Automation Control. In *Proc. of 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, pages 1608–1612, Singapore, September 2014.

[11] P. Transfeld, S. Receveur, and T. Fingscheidt. An Acoustic Event Detection Framework and Evaluation Metric for Surveillance in Cars. In *Proc. of 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, Dresden, Germany, September 2015.

[12] H. Klinck and M. Ritter. Automated Identification of Blue and Fin Whale Vocalizations using an Ensemble Based Classification System. In *International DCLDE [Detection, Classification, Localization, and Density Estimation] Workshop*, La Jolla, CA, USA, July 2015.

[13] M. Cowling and R. Sitte. Comparison of Techniques for Environmental Sound Recognition. *Pattern Recognition Letters*, 24(15):2895–2907, November 2003.

[14] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson. Feature Analysis and Selection for Acoustic Event Detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, pages 17–20, Caesars Palace, Las Vegas, Nevada, USA, 2008.

[15] S. Chu, S. Narayanan, and C.-C. J. Kuo. Environmental Sound Recognition with Time-Frequency Audio Features. *Proc. of IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, August 2009.

[16] T. Schmidt and W. Schütte. FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2091 – 2096, Valletta, Malta, May 2010.

[17] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, Barcelona, Spain, 2013.

[18] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proceedings of Interspeech*, pages 148–152, Lyon, France, 2013.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18, 2009.