Improving acoustic event detection by localization algorithms

Moritz Brandes[1,2], Hans-Christoph Mertins[2], Jens Schröder[1], Stefan Goetze[1]

Fraunhofer IDMT, Hearing, Speech and Audio Technology[1]; Fachhochschule Münster[2]

Introduction

To increase the accuracy of acoustic event detection (AED) systems[1–3], novel approaches have been developed e.g. by Butko et al., in which the information on the acoustic localization of events (AEL) is combined with features of audio data, called feature level fusion (FLF). In this paper, the performance of FLF is compared with conventional AED for stationary and non-stationary sound-sources. It is shown that FLF leads to an improvement on the recognition performance in comparison to the AED.

Feature Fusion

In the feature level fusion, the mel frequency cepstral coefficients (MFCC, [4]) are taken as features to create a baseline detection system. Information about the source position, generated by the localization algorithm [5], is added to the MFCCs. For our experiments, only the *x*- and *y*-coordinates are used. This results in a total of 41 features per feature vector, 39 from the AED subsystem and two from the AEL, i.e. from the position estimation subsystem. They form the baseline for the training of the proposed detection system.

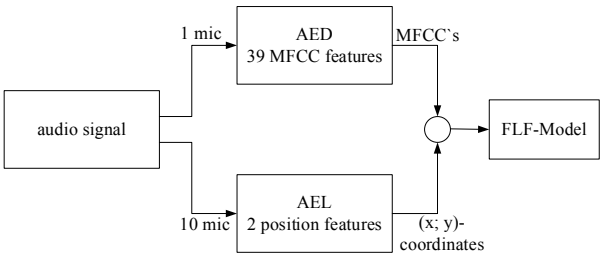

Figure 1: Block diagram of the fusion process for the feature level fusion.

Experimental Setup

All experiments within this work have been recorded in the SilverTainment-Lab of Fraunhofer IDMT [6] in Oldenburg. This was done because the developed detection system was supposed to be implemented in this room. All training-data was recorded in this room, so all acoustical information about the room, such as the typical reverberation is already contained in the recordings. The first 15 events were recorded by two-microphone arrays at different positions. One of these arrays, consisting of eight microphones is located above a television and another one is located at the wall behind the sofa (cf. Figure 2).

This work was partly funded by the EU projects "Sounds for Energy-Efficient Buildings" (S4EcoB, project no. 284628) and "Experimenting Acoustics in Real environments using Innovative Test-beds" (EAR-IT, project no. 318381)

For AED the 39 dimensional MFCC features (including Δ and ΔΔfeatures) are calculated for all recordings. The hop size is about 10 ms with a block size of 25 ms. For the acoustical classification a hidden-Markov-model (HMM [4]) was created with 5 states (3 emitting states).

Events with known positions
TV
Radio
Triangle (Kidscorner)
Drum (Kidscorner)
RC-Car-Sirene (Kidscorner)
RC-Car-Engine (Kidscorner)
SpoonCup_Jingle
Keyjingle
Table_knock
Tel_ring
Curtain
Events with unknown positions
Speech
Clap
Whistle
snap

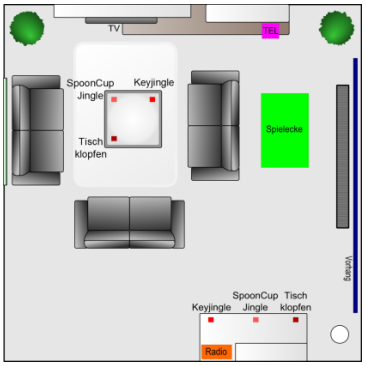

Figure 2: Scheme of the SilvertainmentLab at Fraunhofer IDMT, Oldenburg, which is the test environment for all recordings used for this work.

To determine the position of the acoustic source, the localization algorithm SRP-PHAT [5] is used. This algorithm processes the information of 10 microphones of the mentioned two arrays. The room dimensions of the SilvertainmentLab are about $6 \cdot 6$ m. Possible sound positions are localized with a precision of about 0.3 m. This results in about 400 measure points.

To develop a model which is robust in noisy environments, all models were also trained under noisy conditions (multi-condition-training). Different background noises from the NOISEX-database [7] with a signal to noise ratio (SNR) level of -5, 0, 5, 10 and 20 dB where added to the training-sets. A total number of 1.951 recordings with 18 events led to 31.216 noisy, pre-processed audio-files. Two-thirds of the data were used for the training, resulting in 20.128 recorded sequences and and 11.088 test sequences.

The most important parameter of a recognition system is the detection accuracy. For the evaluation process, 33 % of the entire audio data were used for test purposes only and can be evaluated according to the following equation (1):

$$\textbf{accuracy} = \frac{\textbf{correctly recognized events}}{\textbf{number of test data}} * \textbf{100}\% \quad [\%] \quad (1)$$

Results of Evaluation

The detection rates of all events for the AED, AEL and FLF recognizers are shown in Figure 3. It can be seen that a small improvement is achievd by the FLF compared to the AED alone of about 2 %.

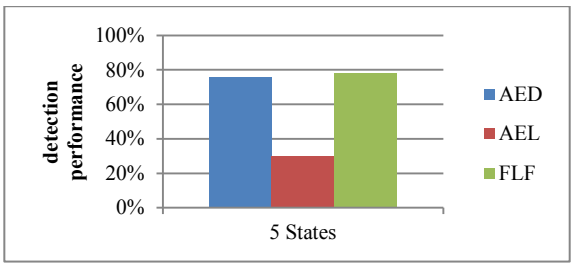

Figure 3: Comparison of the recognition performance of AED, AEL and FLF for an HMM with 5 states.

In Figure 4 and Figure 5, the recognition performance is shown for each event for the AED and the FLF detection system, respectively. It can be seen that the number of false detections decreases for a few events by using the FLF recognizer. False detection are particularly problematic in detection systems since they may lead to significantly reduced acceptance of such systems if wrong actions are triggered by erroneously detected events.

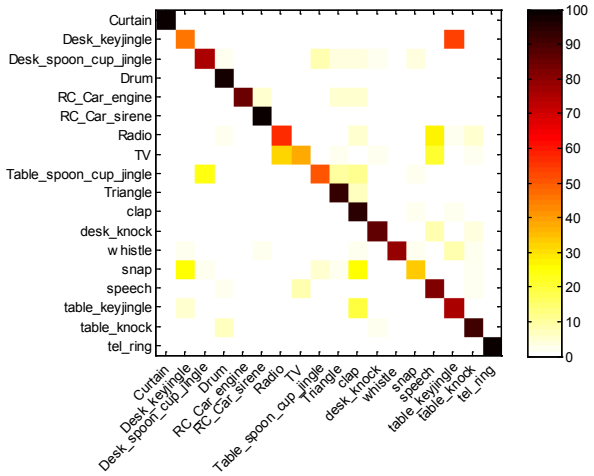

Figure 4: Confusion matrix for the AED-system.

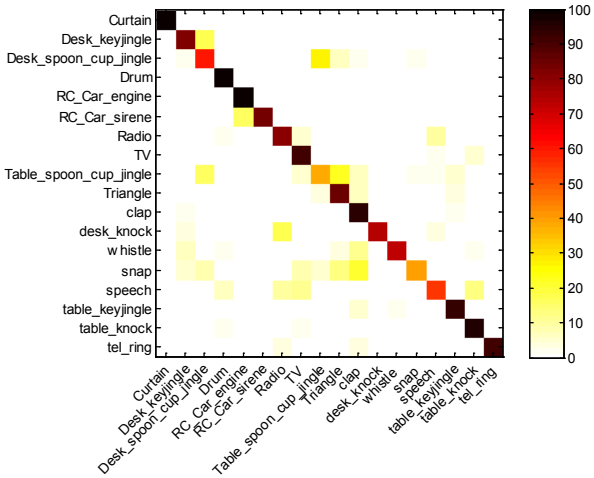

Figure 5: Confusion matrix for the FLF-system.

To investigate what happens to events for which just the positions are different but the sounds are the same, Figure 6 shows a comparism between the detection performances of the events '*TV*' and '*radio*'. By this, a test to distinguish two equal events only by their positions is conducted. It can be seen, that both events are often confused when using AED only. Above all, there is a high rate of confusion between the events *radio/TV* with the event *speech*. This could be caused by two reasons: both events are very similar since the news, played out by the loudspeaker, consist of speech. By using the fusion of AED and AEL the number of false detections can be reduced significantly.

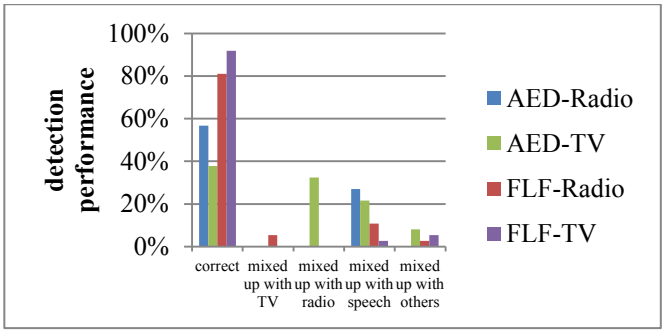

Figure 6: Comparison of the detection performance for the events *radio* and *TV*. Identical signals played at two different positions.

Conclusions

The results of the described feature level fusion for the AED and AEL systems show a significant rise of the detection performance for various stationary sound-sources. It can be seen that the average detection rate of all events shows only a slight increase in the FLF recognition performance with respect to the AED.

The findings regarding accuracy of the FLF corresponds with 78 % to the results reached in [2] with approx 76 % (cf. Table 1).

Table 1: Comparison of the detection performance of [2] with the results from this work

	Ø ACC AED	**Ø ACC FLF**
Performance in [2]	74,55 %	75,82 %
Proposed System	76,05 %	77,78 %

References

[1] C.V. Cotton, D.P.W. Ellis, Spectral vs. spectro-temporal features for acoustic event detection, in: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 69–72.

[2] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion, 2009, http://taras-butko.info/files/publication/2009-Interspeech-Butko.pdf, accessed 18 February 2014.

[3] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, M. Omologo, CLEAR Evaluation of Acoustic Event Detection and Classification Systems, in: R. Stiefelhagen, J. Garofolo (Eds.), Multimodal Technologies for Perception of Humans, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 311–322.

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK Book: for HTK Version 3.4, 2009, accessed 31 January 2013.

[5] J.H. DiBiase, A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arraysthesis, 2000, http://www.glat.info/ma/av16.3/2000-DiBiaseThesis.pdf, accessed 26 June 2013.

[6] Fraunhofer IDMT, SilverTainmentLab: Ist Ihre Lösung fit für die Generation 50+?, 2014, http://www.idmt.fraunhofer.de/de/hsa/equipment/_jcr_content/contentPar/textblockwithpics_3/linklistPar/download/file.res/SilverTainment.pdf, accessed 24 April 2014.

[7] A. Varga, H.J. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication 1993, pp. 247–251.