Feasibility of automatic sound recognition in elderly care monitoring

Bernd Heim, Janto Skowronek, Richard Doornbos
Philips Research Laboratories Eindhoven, The Netherlands, Email: janto.skowronek@philips.com

Introduction

Intelligent and adaptive home environment systems can support elderly inhabitants in maintaining healthy and active lifestyles. A major part of such systems consists of intelligent monitoring of the inhabitant and its environment using a variety of sensors. Audio sensors could additionally be used to detect the type of activities of inhabitants and link them to high-level activities in daily living (ADL). For example, the sounds of a running water tap, followed by a clattering kettle, followed by the lighting of a match, and some slamming of cupboard doors could indicate, that someone is making hot water (probably for tea or coffee).

In an experimental study we wanted to investigate to what extend such a environmental sound recognition (ESR) system can provide useful information about the daily life activities in a home environment. Our main questions were: How can we deal with the scores of variety in activities and hazardous situations in daily living and what sounds can indicate these ADLs? Where are the main problems when developing such a recognition system with respect to capture and recording issues (e.g., recording quality, sound event detection), as well as classification issues (e.g., feature selection, training, robustness)?

Sounds that indicate ADL

In order to overcome the problem of the scores of variety in ADL and environmental sounds (env-sounds) respectively, we decided for a hierarchical approach: Inspired by a study from Schadow [1] dealing with hazardous situations in ADL, we defined first five scenarios of daily living: 'Coming home', 'Doing housework', 'Preparing food', 'Taking a shower' and 'Sitting/relaxing'. Then we chose a number of actions per scenario that cause sounds. Bearing our experience with env-sounds in mind [2], we described the sounds in terms of the sound producing object (e.g., vacuum cleaner) or activity and involved objects (e.g., fill water in a pot) respectively. Partly using two different objects for the same action (e.g., plastic mug or porcelain cup) we came up with 44 different sound classes such as footsteps, slamming a door, ringing the doorbell, vacuum cleaner, clattering with cutlery, running water or turning the page of a newspaper. Notice that the eventually chosen sounds do not explicitly comprise sounds of accidents. For practical reasons, in particular the trade-off between the exploratory characteristics of this study and the necessary effort of recording accidents, we decided to use only sounds of "ordinary" ADLs in first instance.

Audio recordings

We recorded all sounds manually in the entrance area, the kitchen, the staircase and the bathroom of two different home environments. The recording system comprised two kidney-characteristic mono microphones, a Mindprint two-channel microphone amplifier, a Hammerfall soundcard and a standard laptop. One microphone was placed next to the ceiling at a distance of 1.5 to 2 meters to the sound source (left channel), the other one was placed next to the wall at a distance of 0.5 to 1 meter to the sound source (right channel). We recorded every type of sound either about 40 times (transient sounds) or over a period of about one minute (continuous sounds), in order to get a sufficient amount of training material for the classification algorithm. By splitting the channels of the stereo recordings into mono files, we obtained four data sets (two home environments, two microphone positions) with about 2200 sound files.

Classification experiments

We implemented a classification algorithm with the two standard stages: feature extraction and pattern recognition. In a frame based analysis of the sounds, we computed nine standard low-level audio features [3], such as spectral centroid or zero-crossing rate, and used a standard quadratic discriminant analysis [4] for classifying the sounds. For every classification experiment the signal frames were randomly split into 80 % training and 20 % test material. Though we also looked at the confusions that occurred among individual classes, we will discuss here only the mean classification performances of six experiments that we conducted (see Table 1).

In Exp. 1a to 1d we tested whether ESR is in general feasible for detecting the activities by testing idealized situations: Test the system with the same type of actions and the same object (keep all 44 individual classes) recorded in the same environment (one home) at the same position (one microphone position) as it was trained[1]. Though we have a relatively high number of classes (44) we achieved quite good classification performance for all four tested conditions (75-78 %). Thus under these idealized conditions, ESR is indeed feasible for detecting ADL.

In Exp. 2 we merged classes whose sounds were produced by the same activity but with different objects. The mean performance increased to 90 %, meaning that same actions with different objects seem to overlap in the feature space such that a merging of those classes increases the performance.

[1] Notice that we nevertheless split the test and training frames in a training-test-ratio of 80-20.

Exp.	Test Conditions	Num. Clas.	Mean Perf.
1a	training & test set (80/20 split): home 1, mic. position 1	44	78 %
1b	training & test set (80/20 split): home 1, mic. position 2	44	75 %
1c	training & test set (80/20 split): home 2, mic. position 1	44	78 %
1d	training & test set (80/20 split): home 2, mic. position 2	44	76 %
2	training & test set (80/20 split): home 1, mic. position 1, merged classes: same activity with different types of objects	31	90 %
3	training & test set (80/20 split): home 1 & 2 , mic. position 1, merged corresponding classes of both homes	44	71 %
4	training & test set (80/20 split): home 1 & 2, mic. position 1 kept individual classes of both homes	88	71 %
5a	home 1 training set, home 2 test set, mic. position 1	44	15 %
5b	home 2 training set, home 1 test set, mic. position 1	44	20 %
6a	home 1 training set, home 2 test set, mic. position 1, merged classes: same activity with different types of objects	31	25 %
6b	home 2 training set, home 1 test set, mic. position 1, merged classes: same activity with different types of objects	31	28 %
7a	mic. position 1 training set, mic. position 2 test set, home 1	44	38 %
7a	mic. position 2 training set, mic. position 1 test set, home 1	44	39 %

Table 1: Overview of the mean performances of the classification experiments.

Exp. 3 aimed at the influence of the different home environments. Merging the corresponding classes of the two homes led to a slight performance decrease compared to Exp. 1 (75-78 % down to 71 %). The corresponding sounds of the different home environments seem to overlap in the feature space such that the performance changes only slightly.

With an additional experiment (Exp. 4) we wanted to investigate how such a ESR system can deal with a high number of sound classes. By merging the data sets of both homes for one microphone position but by keeping the individual classes of both sets, we doubled the number of classes (88) the classification algorithm had to deal with. Again the performance decreased only slightly (about 5 %) compared to Exp. 1, meaning that under the conditions of the experimental set up, it is still possible to adequately recognize a large number of classes.

However these experiments were quite idealized. Though we considered to split test and training material, the system was always "familiar" with the type of material: Training and test material stemmed from the same environment(s) and microphone position. In order to check the system's robustness against completely unknown data, we ran further experiments (Exp. 5-7), in which we trained the system with material from one combination of home environment and microphone position, but tested with another combination of home environment and microphone position. The results of Exp. 5-7 in Table 1 show a drastic performance decrease down to 15-40 %, dependent wether we tested the robustness against two environments (Exp. 5,6) or the different microphone positions (Exp. 7).

Conclusions

In summary we saw that an ESR system as we set up (recording, features, classification method) is in principle feasible for recognizing activities of daily living. During pilot experiments we observed an influence of the quality of the recording system; the experiments discussed here showed also an influence of microphone position, home environment and the objects involved in the recorded activities. As long as the system was trained under the same circumstances as it was tested (Exp. 1-4), we observed a quite high recognition performance, even if two environments have been considered. When the system was confronted with material from completely new circumstances (Exp. 5-7), the performance broke down. Thus the tested system requires a specialized training for each particular situation (home environment, microphone position) in which it will be placed. Or a quite extensive training is necessary in order to be more robust for different situations (training has to cover various environments, microphone positions etc.)

In further work we first plan to implement a complete automatic ESR system (online sound recognition instead of off-line classification of manual recordings). Besides the technical aspects of the follow-up project, we will address the issues and problems we found here as well.

References

[1] Schadow, B., *Entwicklung einer arbeitswissentschaftlichen Methodik zur partizipativen Analyse potentieller Gefahren im Haushalt*, PhD thesis, Fakultät V - Verkehrs- und Maschinensysteme der Technischen Universität Berlin, 2004.

[2] Skowronek, J., Kappelmann, M., McKinney, M. *Automatic Classification Of Environmental Sounds*, Proceedings of DAGA 2005, 31.Jahrestagung für Akustik, pp. 475-476, Deutsche Gesellschaft für Akustik, München, 2005.

[3] McKinney, M., Breebaart, J. *Features for audio and music classification*, 4th ISMIR, Baltimore, 2003

[4] Duda, R.O., Hart, P.E., *Pattern classification and scene analysis*, Wiley-Interscience, New York 1973