Automatic Analysis of Speech and Acoustic Events for Ambient Assisted Living

Alexey Karpov[1,2(✉)], Alexander Ronzhin[1], and Irina Kipyatkova[1]

[1] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
{karpov, ronzhinal, kipyatkova}@iias.spb.su
http://www.spiiras.nw.ru/speech
[2] ITMO University, Saint-Petersburg, Russian Federation

Abstract. We present a prototype of an ambient assisted living (AAL) with multimodal user interaction. In our research, the AAL environment is one studio room of 60 + square meters that has several tables, chairs and a sink, as well as equipped with four stationary microphones and two omni-directional video cameras. In this paper, we focus mainly on audio signal processing techniques for monitoring the assistive smart space and recognition of speech and non-speech acoustic events for automatic analysis of human's activities and detection of possible emergency situations with the user (when an emergent help is needed). Acoustical modeling in our audio recognition system is based on single order Hidden Markov Models with Gaussian Mixture Models. The recognition vocabulary includes 12 non-speech acoustic events for different types of human activities plus 5 useful spoken commands (keywords), including a subset of alarm audio events. We have collected an audio-visual corpus containing about 1.3 h of audio data from 5 testers, who performed proposed test scenarios, and made the practical experiments with the system, results of which are reported in this paper.

Keywords: Ambient assisted living · Assistive technology · Multimodal user interfaces · Universal access · Human-Computer interaction · Automatic speech recognition · Acoustic event detection

1 Introduction

Ambient Assisted Living (www.aal-europe.eu) [1–3] is a new special area of assistive information technologies [4–6] that is focused on designing smart spaces, rooms, homes and intelligent environments to support and care some disabled and elderly people. At present AAL domain includes several International projects, for example, DOMEO, HAPPY AGEING, HOPE, SOFTCARE, Sweet Home, homeService, We-Care, etc. Arrays of microphones, video cameras and other sensors are often installed in assistive smart spaces [7].

In this study, we also analyze audio and video modalities for automatic monitoring of activities and behavior of single elderly persons and people with physical, sensory or mental disabilities. In our previous recent work [8], we presented a prototype of a

M. Antona and C. Stephanidis (Eds.): UAHCI 2015, Part II, LNCS 9176, pp. 455–463, 2015.
DOI: 10.1007/978-3-319-20681-3_43

multimodal AAL environment with main focus on video-based methods for space and user behavior monitoring by omni-directional (fish eye) cameras, mainly for detection of accidental user falls. In this article, we focus mainly on audio-based techniques for monitoring the assistive smart space and recognition of speech and non-speech acoustic events for automatic analysis of human's activities and detection of possible emergency situations (when some help is needed to the person). The use of audio-based processing additionally to video analysis makes many multimodal systems more accurate and robust [9, 10]. Acoustic events in AAL environments can be such as human's speech / commands as well as artificial sounds directly or indirectly produced by a human being (for example, cough, cry, chair movements, knocking at the door, steps, etc.). Spoken language is the most meaningful acoustic information; however other auditory events give us much information too. In scientific literature, there are some recent publications on automatic detection of individual acoustic events such as cough, sounds of human fall, cry, scream, distress calls or other events, for example in works [11–15].

In our research, the AAL environment is one room of 60 + square meters. We developed a software-hardware complex for audio-based monitoring of this AAL environment during the Summer Workshop on Multimodal Interfaces eNTERFACE in Pilsen. The room (physical model of the AAL environment) has 2 tables, 2 chairs and a sink, as well as it is equipped with 2 omni-directional video cameras and 4 stationary microphones in a grid. One Mobotix camera is placed on the ceiling and the second one is on the side wall; camera's frame resolution is 640 × 480 pixels at 8 fps rate. 4 dynamic condenser microphones Oktava MK-012 of the smart environment are connected to a multichannel external sound board M-Audio ProFire 2626. Each microphone has the cardioid diagram of direction and can capture audio signals in a wide sector below the microphone with almost equal amplification. All the microphones are placed on the ceiling (about 2.5 m above the floor) in selected locations. The scheme of the physical model of our AAL environment is shown in Fig. 1.

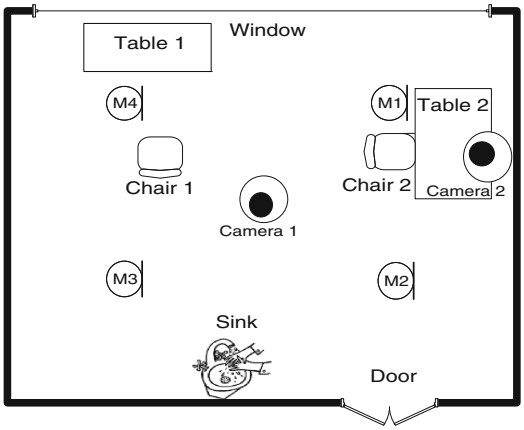

Fig. 1. Scheme of the physical model of the Ambient Assisted Living environment

The paper below describes the architecture and implementation issues of the automatic recognition system in Sect. 2, as well as presents some experimental results and analysis of its evaluation in Sect. 3.

2 Architecture of the Automatic Recognition System

The recognition vocabulary includes 12 non-speech acoustic events for different types of human activities plus 5 possible spoken commands. We defined also a set of alarm audio events X = {"Cough", "Cry", "Fall" (a human being), "Key drop" (a metal object), "Help", "Problem" (commands)}, which can serve as a signal on an emergency situation with the user inside the AAL environment. Figure 2 presents a tree classification of audio signals in the AAL environment including speech commands and acoustic events, which are modelled in the automatic system.

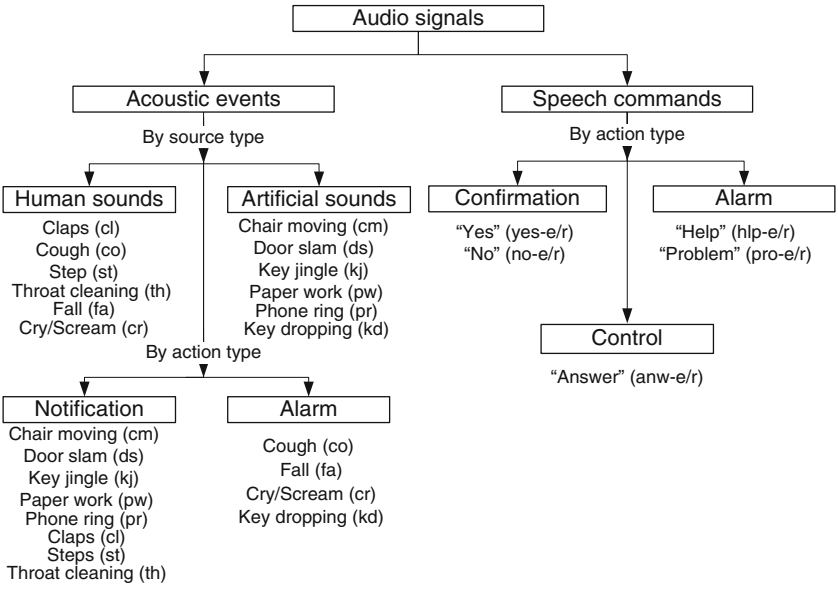

Fig. 2. A tree classification of audio signals of the AAL environment model

Figure 3 shows the software-hardware architecture of the automatic system for recognition of speech and non-speech acoustic events. Acoustic modeling in the system is based on single order Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) as many modern automatic speech recognition systems [16]. Our system extracts feature vectors consisting of 13 Mel-frequency cepstral coefficients (MFCC) with deltas and double deltas from the multichannel audio signals.

The system uses HMM-based Viterbi method and software algorithm for finding an optimal model for input audio signal (Fig. 4). HMMs having a unique topology with

different number of states (from 1 to 6 states per one model) represent all acoustic events and phonemes of speech commands depending on duration.

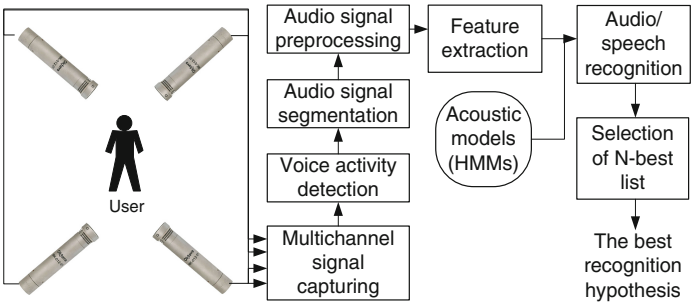

Fig. 3. Architecture of audio event/speech recognition system

The recognition process is made in the on-line mode (the speed factor is less than 0.1 real-time) and a recognition hypothesis is issued almost immediately after energy-based voice/audio activity detection. We apply a speech recognizer dependent to its speaker/user [17] because of the system aim and usability issues. All speech commands, audio events, as well as garbage (any unknown acoustic events) and silence models are presented by a grammar that allows the system to output only one the recognition hypothesis at the same time. In this grammar, there are also some restrictions; for example, two or more "Fall" events cannot follow each other opposite to the "Step" events. The developed ASR system is bilingual and able to recognize and interpret speech commands both in English and in Russian.

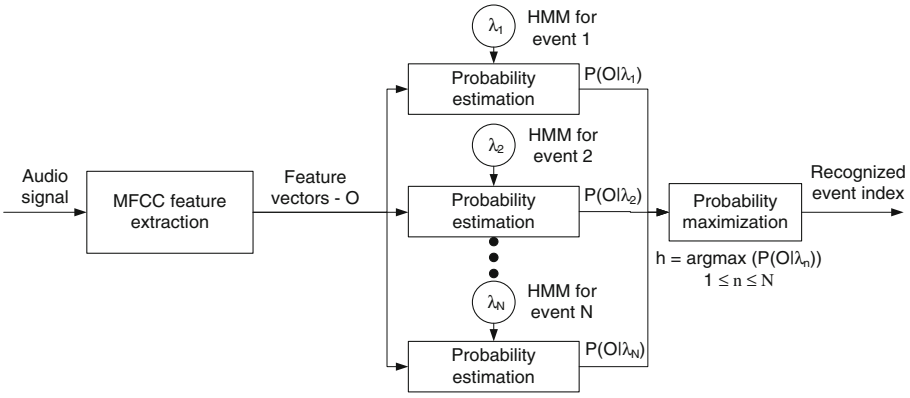

Fig. 4. HMM-based method for automatic recognition of audio signals

In order to train probabilistic HMM-based acoustic models of the recognizer, a new audio corpus has been collected in normal room conditions with an acceptable level of

background noise (SNR > 20 dB). In total, we have recorded over 1.3 h of audio data from several potential users performing certain scenarios. Above 2 K audio files were recorded; almost half of which are non-speech audio events. Approximately 2/3 audio data of each subject we used for the training and development purpose and the rest data were employed during the evaluation.

3 Results of the Experiments with the AAL Model

We have developed some scenarios for modeling basic actions performed by people in a living room (studio). The first scenario involves one person and simulates an emergency situation (drop of a metal object and fall of a human on the floor in the end). The second scenario involves up to 3 subjects, who can interact each other, and subjects may occlude each other at some frames (it is used for testing the video-based user monitoring sub-system) [8]. The main scenario involving audio-visual data supposes the following actions of a tester:

(1) Enter the room from the door side (open & close the door).
(2) Walk to the table 1.
(3) Take a glass with water from the table 1.
(4) Walk to the chair 1.
(5) Sit on the chair 1.
(6) Drink water.
(7) Long cough after drinking.
(8) Stand up.
(9) Walk to the table 1.
(10) Release the glass (it drops).
(11) Walk to the sink.
(12) Wash hands in the sink.
(13) Exit the room (open & close the door).
(14) Enter the room again.
(15) Walk to the chair 2.
(16) Sit on the chair 2.
(17) Telephone rings on the table 2.
(18) Say “Answer phone”.
(19) Talk by the telephone.
(20) Say “Hello”.
(21) Say “I’m fine”.
(22) Say “Don’t worry”.
(23) Say “Good bye”.
(24) Stand up.
(25) Walk to the table 1.
(26) Take a metallic cup on the Table
(27) Free walk (several steps).
(28) Drop the cup on the floor.
(29) Make a step.

(30) Fall on the floor.
(31) Cry.
(32) Ask for “Help”

During the multimodal database collection we have recorded audio-visual samples of the presented first Scenario from 5 different subjects (potential users). They were free to perform the fall (on the hard floor that produces a sound) as it was comfortable for them. Also a training part of the audio database has been recorded in the same room, where a factor of entering new people into the room during the recording session was avoided that allows removing major external noises. 5 check points have been defined in the room for collecting the training audio data; 4 of them were located on the floor under each microphone and the last one was in the centre of the room. Each of 5 testers performed the following sequence of user’s actions:

(1) Come to a checkpoint.
(2) Give a speech command or simulate a non-speech audio event.
(3) Move to the following checkpoint (1).

All the speech commands and acoustic events were simulated many times by different testers. In total, we have recorded above 2800 audio files (in the PCM WAV format); 44 % of them contain non-speech events and the rest – speech commands. 70 % recordings for each subject were used for system’s training and the rest of the data were used in the experiments. This corpus (SARGAS DB) has been registered in the RosPatent (№ 2013613086 on 25/03/2013).

Table 1. Confusion matrix for speech command recognition (accuracy, %)

Speech command	Num.	ans -e	hlp -e	no -e	pro -e	yes -e	yes -r	no -r	ans -r	hlp -r	pro -r
ans-e	388	100	0	0	0	0	0	0	0	0	0
hlp-e	100	0	85	0	0	0	0	0	0	0	15
no-e	120	0	0	100	0	0	0	0	0	0	0
pro-e	249	0	0	0	94	0	0	0	0	0	6
yes-e	104	0	0	0	0	100	0	0	0	0	0
yes-r	100	0	0	0	0	0	100	0	0	0	0
no-r	100	0	0	0	0	0	0	100	0	0	0
ans-r	118	0	0	0	0	0	0	0	100	0	0
hlp-r	155	0	0	0	2	0	0	0	0	95	3
pro-r	151	0	0	0	9	0	0	0	0	0	91

The automatic system for recognition of speech commands and audio events was evaluated using audio recordings including the audio corpus part containing data of the first Scenario made in the ambient assisted living environment. Table 1 shows the confusion matrix with the accuracy rates (in %) for recognized speech command.

Table 2. Confusion matrix for acoustic event recognition (accuracy, %)

Acoustic event	Num.	cl	cm	co	cr	ds	fa	kd	kj	pw	pr	st	th
cl	112	100	0	0	0	0	0	0	0	0	0	0	0
cm	152	0	100	0	0	0	0	0	0	0	0	0	0
co	72	0	0	100	0	0	0	0	0	0	0	0	0
cr	108	0	0	0	100	0	0	0	0	0	0	0	0
ds	111	0	0	0	2	98	0	0	0	0	0	0	0
fa	108	0	0	1	0	0	63	0	0	0	0	36	0
kd	129	15	0	0	0	0	0	75	0	10	0	0	0
kj	44	0	0	0	0	0	0	0	100	0	0	0	0
pw	68	4	0	0	0	0	0	0	0	96	0	0	0
pr	32	0	0	0	0	0	0	0	0	0	100	0	0
st	166	0	0	6	0	0	0	0	0	0	0	94	0
th	124	0	0	0	0	0	0	0	0	0	0	0	100

Presented results show that the most of speech commands were recognized with a high accuracy over 90 %, however still there were some recognition errors during test scenarios.

Table 2 shows another confusion matrix with the accuracy rates (in %) for recognized acoustic events in the AAL environment model. The presented results show that the lowest accuracy was observed for the non-speech audio event “Fall”. In the third of such cases this acoustic event was recognized as “Step”.

In average, the recognition accuracy for acoustic events was 93.8 % and 96.5 % – for speech commands.

4 Conclusion

We presented a software-hardware complex for audio-based monitoring of the AAL environment model. Audio signals in AAL environments can be human’s commands/speech and artificial sounds produced by a human directly or indirectly (for example, chair movements, cough, cry, knocking at the door, steps, etc.). The system uses Viterbi-based algorithm for finding an optimal model for input audio signal. HMMs having a unique topology with different number of states (1-6) model each acoustic event and command depending on its duration. The recognition is performed in on-line mode (speed factor < 0.1 real-time) and a hypothesis is issued almost immediately after energy-based voice/audio activity detection. We apply the speaker-dependent recognizer because of the system aim and usability issues. The vocabulary includes 12 non-speech acoustic events for different types of human activities plus 5 user’s spoken commands including a set of some alarm events, which can serve as a signal about an emergency situation inside the AAL environment. To train acoustic models, we have collected an audio corpus in quiet room conditions with

a low level of background noise (SNR > 20 dB). Above two thousand audio files were recorded; almost half of which are non-speech audio events. Approximately 2/3 audio data of each subject we used for the training and development purpose and the rest data were employed during the evaluation. In the experiments, the best recognition accuracy for the acoustic events was 93.8 % in average and 96.5 % – for speech commands.

Acknowledgements. This research is partially supported by the Council for Grants of the President of Russia (Projects No. MD-3035.2015.8 and MK-5209.2015.8), by the Russian Foundation for Basic Research (Projects No. 15-07-04415 and 15-07-04322), and by the Government of the Russian Federation (Grant No. 074-U01).

References

1. Burzagli, L., Di Fonzo, L., Emiliani, P.L.: Services and applications in an ambient assisted living (aal) environment. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2014, Part III. LNCS, vol. 8515, pp. 475–482. Springer, Heidelberg (2014)
2. Sacco, M., Caldarola, E.G., Modoni, G., Terkaj, W.: Supporting the design of AAL through a SW integration framework: the D4All project. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2014, Part I. LNCS, vol. 8513, pp. 75–84. Springer, Heidelberg (2014)
3. Mora, N., Bianchi, V., De Munari, I., Ciampolini, P.: A BCI platform supporting AAL applications. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2014, Part I. LNCS, vol. 8513, pp. 515–526. Springer, Heidelberg (2014)
4. Karpov, A., Ronzhin, A.: A Universal assistive technology with multimodal input and multimedia output interfaces. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2014, Part I. LNCS, vol. 8513, pp. 369–378. Springer, Heidelberg (2014)
5. Argyropoulos, S., Moustakas, K., Karpov, A., Aran, O., Tzovaras, D., Tsakiris, T., Varni, G., Kwon, B.: A Multimodal framework for the communication of the disabled. J. Multimodal User Interfaces **2**(2), 105–116 (2008). Springer
6. Karpov, A., Ronzhin, A., Kipyatkova, I.: An assistive bi-modal user interface integrating multi-channel speech recognition and computer vision. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part II, HCII 2011. LNCS, vol. 6762, pp. 454–463. Springer, Heidelberg (2011)
7. Portet, F., Vacher, M., Golanski, C., Roux, C., Meillon, B.: Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Pers. Ubiquit. Comput. **32**(1), 1–18 (2011)
8. Karpov A., Akarun L., Yalçın H., Ronzhin Al., Demiröz B., Çoban A., Zelezny M.: Audio-visual signal processing in a multimodal assisted living environment. In: Proceedings of the 15th International Conference, INTERSPEECH-2014, Singapore, pp. 1023–1027 (2014)
9. Karpov, A.: An automatic multimodal speech recognition system with audio and video information. Autom. Remote Control **75**(12), 2190–2200 (2014). Springer
10. Karpov, A., Ronzhin, A.: Information Enquiry Kiosk with Multimodal User Interface. Pattern Recogn. Image Anal. **19**(3), 546–558 (2009). Springer
11. Drugman T., Urbain J., Dutoit T. Assessment of audıo features for automatıc cough detectıon. In: Proceedings of the 19th European Signal Processing Conference, EUSIPCO-2011, Barcelona, Spain, pp. 1289–1293 (2011)

12. Zigel, Y., Litvak, D., Gannot, I.: A method for automatic fall detection of elderly people using floor vibrations and sound - proof of concept on human mimicking doll falls. IEEE Trans. Biomed. Eng. **56**(12), 2858–2867 (2009)
13. Miao, Yu., Naqvi, S.M., Rhuma, A., Chambers J.: Fall detection in a smart room by using a fuzzy one class support vector machine and imperfect training data. In: Proceedings of the 36th International Conference, ICASSP-2011, Prague, Czech Republic, pp. 1833–1836 (2011)
14. Huynh, T.H., Tran, V.A., Tran, H.D.: Semi-supervised tree support vector machine for online cough recognition, In: Proceedings of the 12th International Conference, INTERSPEECH-2011, Florence, Italy, pp. 1637–1640 (2011)
15. Aman, F., Vacher, M., Rossato S., Portet, F.: In-Home Detection of Distress Calls: The Case of Aged Users. In: Proceedings of the 14th International Conference, INTERSPEECH-2013, Lyon, France, pp. 2065–2067 (2013)
16. Levin, K. et al.: Automated Closed Captioning for Russian Live Broadcasting. In: Proceedings of the 15th International Conference, INTERSPEECH-2014, Singapore, pp. 1438–1442 (2014)
17. Matveev, Y.: The Problem of voice template aging in speaker recognition systems. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 345–353. Springer, Heidelberg (2013)