Evaluation of a Real-Time Voice Order Recognition System from Multiple Audio Channels in a Home

Michel Vacher[1], *Benjamin Lecouteux*[1], *Dan Istrate*[2],
Thierry Joubert[3], *François Portet*[1], *Mohamed Sehili*[2], *Pedro Chahuara*[1]

[1]LIG, UJF/CNRS/Grenoble-INP/UPMF, UMR5217, 38041 Grenoble, France
[2]ESIGETEL, 77210 Avon, France
[3]THEORIS, 75000 Paris, France

{Michel.Vacher, Benjamin.Lecouteux, Pedro.Chahuara, Francois.Portet}@imag.fr
dan.istrate@esigetel.fr, mohamed.sehili@esigetel.fr, thierry.joubert@theoris.fr

Abstract

The SWEET-HOME[1] project aims at providing audio-based interaction technology that lets the user have full control over their home environment and at detecting distress situations. This paper presents the audio analysis system PATSH developed for this project and a user experiment in a smart home that evaluates the performances of the first revision of the system regarding vocal order recognition with realistic scenarios.
Index Terms: Real-time audio analysis, experimental in-situ evaluation, Smart Home, Home Automation, AAL

1. Introduction

The development of smart homes and intelligent companions is seen as a promising way of achieving in-home daily assistance [1]. Despite, some smart home projects have considered speech recognition in their design [2, 3, 4, 5, 6, 7, 8], there are still important challenges to be overcome to set up this technology in a real environment [9].

Designing and applying speech interface in Smart Home to provide *security reassurance* and *natural man-machine interaction* is the aim of the SWEET-HOME[2] project [10]. The project addresses the important issues of distant voice command recognition and sound source identification to improve the robustness of voice-based home automation control. In this paper, we introduce the PATSH system which performs the real-time identification of the voice command anywhere in the home and sends the result to an intelligent controller in charge of making decision based on these [11].

2. The Audio Analysis System

The global architecture of PATSH is illustrated in Figure 1. The PATSH framework is developed with the .Net cross-plateform technology. The main data structure is the **Sound object**, which contains a segment of the multidimensional audio signal whose interpretation is continuously refined along the processing pipeline. PATSH deals with the distribution of the data among the several plugins that perform the processing to interpret the audio events. The execution can be done, in parallel, synchronously or asynchronously, depending on the settings

[1]SWEET-HOME is founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-09-VERS-011)

[2]http://sweet-home.imag.fr

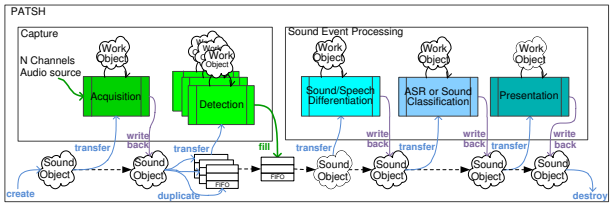

Figure 1: The PATSH architecture.

stored in a simple configuration file. For a complete description of the system the reader is referred to [12].

In the SWEET-HOME configuration, PATSH runs plugins that performs the following tasks: (1) multichannel data Acquisition through the NI-DAQ6220E card (16bits, 16kHz, 7channels), (2) sound detection, (3) sound/speech discrimination, (4) sound classification, (5) automatic speech recognition (ASR) and extracting vocal orders, and (6) presentation, communicating the sound event to the Intelligent Controller. If a vocal order is detected and according to the context (activity and localisation of the user in the flat), an home automation command is generated to make the light up, close the curtains or emit a warning message thanks to a speech synthesizer.

3. Experiments in real conditions

Experiments were run in the DOMUS smart home at the Laboratoire d'Informatique de Grenoble [13]. The flat has been equipped with 7 radio microphones set in the ceiling for audio analysis. A specialized communication device, *e-lio*, from the Technosens company is used to initiate a communication between a senior and a relative [10]. The possible voice orders were defined using a very simple grammar :

set an actuator on:	(e.g. `Nestor ferme fenêtre`) `key initiateCommand object`
stop an actuator:	(e.g. `Nestor arrête`) `key stopCommand [object]`
emergency call:	(e.g. `au secours`)

Our previous user study showed that targeted users prefer precise short sentences over more natural long sentences [14]. Each order belongs to one of three categories: initiate command, stop command and emergency call. Except for the emergency call, every command starts with a unique key-word (here : *'Nestor'*) that permits to know whether the person is talking to the smart home or not.

To validate the system in realistic conditions, we built scenarios in which every participant was asked to perform the following activities: (1) Sleeping; (2) Resting: listening to the radio; (3) Feeding: realizing and having a meal; and (4) Communicating: going out of the flat to do shopping and having a talk with a relative thanks to *e-lio*. Therefore, this experiment allowed us to process realistic and representative audio events in conditions which are directly linked to usual daily living activities. The participant had to use vocal orders to make the light on or off, open or close blinds, ask about temperature and ask to call his or her relative. The instruction was given to the participant to repeat the order up to 3 times in case of failure. A wizard of Oz was used in case of persistent problem.

Sixteen healthy participants (including 7 women) were asked to perform the scenarios without condition on the duration. Before the experiment, the participant was asked to read a text of 25 short sentences in order to adapt the acoustic models of the ASR for future experiments. The average age was 38 years (19-62, min-max) and the experiment lasted between 23min and 48min. While the scenario included 15 vocal orders more sentences were uttered because of repetitions.

During the experiment, audio data were recorded and saved in two ways. Firstly, the 7-channel raw audio signal was stored for each participant to make subsequent analysis possible. In total, 8h 52min 36s of data was recorded for the 16 participant. Secondly, the individual sound events automatically detected by PATSH were recorded to study the performances of this framework. Overall, 4595 audio events were detected whose 993 were speech utterances.

In this study, we are only interested in recognizing vocal orders or distress sentences. All other spontaneous sentences and system messages were discarded. Therefore, only the syntactically correct vocal orders were annotated. The average SNR and duration are 15.8dB and 1s with is far from the recording studio condition. Thanks to this annotation, an oracle corpus corpus was extracted. The comparison between experimental real-time results with thus obtained with the same ASR on the oracle corpus will allow to analyse the performance degradation introduced by the PATSH system.

4. Results

4.1. Discrimination between speech and sounds

All the results presented take into account the performances of the detection because the signals are extracted automatically by the system. The sound/speech discrimination badly classified 108 sound and 232 speech occurrences which gives a total error rate of about 7.4% which is correlated with our previous results [12]. 23.4% of speech occurrences were missed speech and 3% of the detected speech was actually sounds. Theses poor performances are explained by the fact that PATSH was not successful in selecting the best audio event among the set of simultaneous event and thus the event with low SNR introduced errors and were not properly segmented. Theses performances are explained by the fact that the multichannel information was not used; each channel was processed separately. The channel with low SNR introduced errors. For the sounds, the number of missed sounds is about 6.6% and of speech identified as sound is about 3%. Sounds such as dishes, water flow or electric motor were often confused with speech. For instance, when certain persons stirred the coffee and chocked the spoon on the cup or when they chocked plates and cutlery, the emitted sounds had resonant frequencies very to the speech one. This is emphasizing the difficulty of the task and Models must be improved to handle these problematic samples [9].

Table 1: *Home automation order error rate (DER)*

Spk. ID	Nb.	Expe. (%)	Oracle (%)	Spk. ID	Nb.	Expe. (%)	Oracle (%)
S01	20	35	20	S02	32	12.5	6.2
S03	22	22.7	22.7	S04	26	23	7.7
S05	26	15	3.8	S06	24	21	8.3
S07	19	79	52.6	S08	33	30	33.3
S09	40	40	22.5	S10	40	67	47.5
S11	37	46	27	S12	26	21	7.7
S13	21	43	19	S14	27	48	29.6
S15	28	71	55.5	S16	22	18	13.6
Average	28	38%	23.9%				

4.2. Home automation order recognition

The global performance of the system is directly related to vocal order recognition. The DER (Domotic Error Rate) is shown in Table 1, the 3^{nd} and 7^{th} columns "Expe." indicate the results for the real-time experiment, the columns "Nb." indicate the number of vocal orders for each speaker. This error rate is evaluated after filtering at the input of the intelligent controller and includes the global effects of all stages: detection, discrimination between speech and sound, ASR. When the uttered voice orders were not respecting the grammar, for example when a sentence such as “Nestor heure” is uttered instead of the command “Nestor donne l’heure”, these utterrances were discarded. The error rate is 38% on average and 23.9% for the oracle.

5. Conclusion

This paper presents the PATSH system which performs real-time identification of voice commands in the home. The architecture is based on a pipeline of sound event detection, speech/noise discrimination and ASR or sound classification modules. Each identified sound event is then sent to an intelligent controller for final decision about the action to make on the house.

Our application of this technology within a realistic smart home, showed that one of the most sensible tasks is the speech/noise discrimination [9]. According to the SNR level, the performance can be quite poor, which has side effects on both the ASR and the sound classification (and then on the decision making). Another issue is linked to the lack of handling of simultaneous sound event records. These fill the sound object queue, which is the system bottleneck, and thus slow down the processing while real-time performances are required. To increase the performance and free this bottleneck, we had implemented a filtering strategy to remove low SNR audio events as well as too delayed events. The preliminary results showed a significant increase in performance. In a second step, PATSH will be modified to allow in real-time a multisource ASR thanks to the Driven Decoding Algorithm [15].

Although the 16 participants had to repeat, sometimes up to three time, the voice commands, they did not complained about the reaction time nor about the syntax. They were overall very excited about commanding their own home by voice. The next experiment, will involve aged participants and a more controlled evaluation of their feeling and reaction about the system in order to compare with our previous user study [14] and conclude about the potential of the state of the art speech technologies for Ambient Assisted Living.

6. References

[1] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes- present state and future challenges," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[2] D. Istrate, M. Vacher, and J.-F. Serignat, "Embedded implementation of distress situation identification through sound analysis," *The Journal on Information Technology in Healthcare*, vol. 6, pp. 204–211, 2008.

[3] D. Charalampos and I. Maglogiannis, "Enabling human status awareness in assistive environments based on advanced sound and motion data classification," in *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, 2008, pp. 1:1–1:8.

[4] M. Popescu, Y. Li, M. Skubic, and M. Rantz, "An acoustic fall detector system that uses sound height information to reduce the false alarm rate," in *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, 20–25 Aug. 2008, pp. 4628–4631.

[5] A. Badii and J. Boudy, "CompanionAble - integrated cognitive assistive & domotic companion robotic systems for ability & security," in *1st Congres of the Société Française des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG'09)*, Troyes, 2009, pp. 18–20.

[6] M. Hamill, V. Young, J. Boger, and A. Mihailidis, "Development of an automated speech recognition interface for personal emergency response systems," *Journal of NeuroEngineering and Rehabilitation*, vol. 6, 2009.

[7] G. Filho and T. Moir, "From science fiction to science fact: a smart-house interface using speech technology and a photorealistic avatar," *International Journal of Computer Applications in Technology*, vol. 39, no. 8, pp. 32–39, 2010.

[8] B. Lecouteux, M. Vacher, and F. Portet, "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions," in *Interspeech 2011*, Florence, Italy, aug 2011, p. 4p. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00642306/fr/

[9] M. Vacher, F. Portet, A. Fleury, and N. Noury, "Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges," *International Journal of E-Health and Medical Communications*, vol. 2, no. 1, pp. 35–54, 2011.

[10] M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuara, and S. Méniard, "The SWEET-HOME Project: Audio technology in Smart Homes to improve well-being and reliance," in *33th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'11)*, Boston, USA, 2011, pp. 5291–5294.

[11] P. Chahuara, A. Fleury, F. Portet, and M. Vacher, "Using Markov Logic Network for On-line Activity Recognition from Non-Visual Home Automation Sensors," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science, F. Paternò, B. de Ruyter, P. Markopoulos, C. Santoro, E. van Loenen, and K. Luyten, Eds., vol. 7683. Pisa, Italy: Springer, nov 2012, pp. 177–192.

[12] M. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, and J. Boudy, "Sound Environnment Analysis in Smart Home," in *Ambient Intelligence*, Pisa, Italy, 2012, pp. 208–223.

[13] M. Gallissot, J. Caelen, F. Jambon, and B. Meillon, "Une plateforme usage pour l'intégration de l'informatique ambiante dans l'habitat : Domus," *Technique et Science Informatiques (TSI)*, vol. 32, 2013.

[14] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[15] M. Vacher, B. Lecouteux, and F. Portet, "Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment," in *EUSIPCO (European Signal Processing Conference)*, Bucarest, Romania, August 27-31 2012, pp. 1663–1667.