The VocADom Project: Speech Interaction for Well-being and Reliance Improvement

Michel Vacher, Emmanuel Vincent, Marc-Eric Bobillier Chaumon, Thierry Joubert, François Portet, Dominique Fohr, Sybille Caffiau, Thierry Desot

The VocADom Project: Speech Interaction for Well-being and Reliance Improvement

Michel Vacher
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
Michel.Vacher@imag.fr

Emmanuel Vincent
Université de Lorraine CNRS, Inria, LORIA F-54000 Nancy, France
emmanuel.vincent@inria.fr

Marc-Eric Bobillier Chaumon
University of Lyon (Lyon 2) GREPS, F-6976 Bron, France
marc-eric.bobillier-chaumon@univ-lyon2.fr

Thierry Joubert
THEORIS
103 rue La Fayette F-75010 Paris, France
Thierry.Joubert@theoris.fr

François Portet
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
Francois.Portet@imag.fr

Dominique Fohr
Université de Lorraine CNRS, Inria, LORIA F-54000 Nancy, France
dominique.fohr@loria.fr

Sybille Caffiau
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
sybille.caffiau@univ-grenoble-alpes.fr

Thierry Desot
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
sybille.caffiau@univ-grenoble-alpes.fr

Abstract
The VocADom project aims to provide audio-based interaction technology that lets the users have full control over their home environment and at eases the social inclusion of the elderly and frail population. This paper presents an overview of the project focusing on multimodal corpus acquisition and labelling and on investigated techniques for speech enhancement and understanding.

Author Keywords
Smart Home, Voice Command, Speech Interaction, Usage Study, Assistive Technology,Context-aware Interaction

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User InterfacesNatural language; I.2.7 [Artificial Intelligence]: Natural Language ProcessingSpeech recognition and synthesis; K.4.2 [Computers and Society]: Social Issues-Assistive technologies for persons with disabilities.

Introduction
The goal of Ambient Assisted Living (AAL) is to foster the emergence of ICT-based solutions to enhance the quality of life of older and disabled people at home, at work and in the community, thus increasing their autonomy and engagement in social life, while reducing

the costs of health and social care. Smart spaces, such as smart homes, are emerging as a way to fulfill this goal [2]. Smart homes can support health maintenance by monitoring the person's health and behavior through sensors. They can also enhance security by detecting distress situations such as fall [6], and increase autonomy by enabling natural control of the people's environment through home automation [9]. Voice based solutions are far better accepted than more intrusive solutions such as video camera, and also more natural compared to remote controllers. Thus, in accordance with other user studies, audio technology may greatly ease daily living for elderly and frail persons.

Typical processing undertaken as part of smart home systems includes: the automatic location of the dwellers, speech detection and localization, activity recognition, Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), dialogue management, context-aware decision, and sound analysis. These tasks require large resources in order to build the most accurate models. However, the acquisition of datasets in smart homes is highly expensive both in terms of material and of human resources. Therefore, there is a lack of large annotated data sets recorded in real conditions. Moreover, for the available speech datasets, home automation sensor traces are note provided, limiting the amount of studies that can be performed with them (with the notable exception of [10]). Hence, it is still difficult to conduct studies related to multi-modal ASR, distant speech with free grammar and in noisy conditions using these data sets.

This paper presents the VocADom[1] project and its most important aspects: adapted user design, speech enhancement, natural language understanding and experiments.

[1] https://vocadom.imag.fr/

The VocADom Project

The VocADom project aims to define, in conjunction with end users, the features of a distant voice controlled home automation system that will adapt to the user and that can be used in real conditions (i.e., noise, presence of several people). Challenges are related to user acceptability, speech enhancement, speech understanding and experimental evaluation. To solve the challenges related to distant voice control and associated decision, robust integration of all available information modalities (audio and home sensors) is needed. Therefore, the project gathers researchers and engineers from the Laboratory of Informatics of Grenoble (specialised in speech processing, natural language understanding, smart home design and experimental evaluation), Inria (specialised in speech processing and noise cancellation), the GRePS laboratory (specialised in social psychology) and one company: THEORIS (real-time system development and integration). The used methodology takes into account experience gained during the Sweet-Home project [8] regarding scenario design and ecological validity. VocADom will dispose of one dedicated smart home in order to validate the experimental proof of concept.

The targeted users are elderly people and persons with visual impairment. Vocal command is then an efficient way to provide natural man-machine interaction and to ease social inclusion [8]. Regarding the technical aspect, the project tries to make use of standardised technologies. In our case, the KNX bus system (KoNneX), a worldwide ISO standard (ISO/IEC 14543) for home and building control has been chosen as the main communication bus.

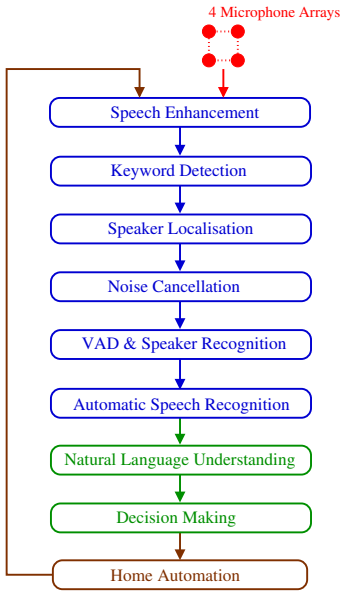

Figure 1: Sound acquisition and processing.

It is integrated through the openHAB software[2] allowing the use of multimedia standards useful for TV and radio deployment.

The architecture of the system is depicted in Figure 1. The input is composed of the information from the home automation network and information from the 16 microphones (4 arrays of 4 microphones each). Thus, information is provided directly by the persons present in the home (e.g., voice command) or via environmental sensors (e.g., temperature). Each vocal command must be differentiated from colloquial speech and therefore must begin with a keyword. The signal recorded by each microphone array is enhanced (suppression of TV noise) before keyword detection. If a keyword is detected on an array, the signal is further analysed through the following stages: speaker localisation, speech enhancement and separation and voice activity detection. An Automatic Speech Recognizer (ASR) delivers a written transcription of the best hypothesis of sentence. The Natural Language Understanding (NLU) stage is in charge of delivering the appropriate voice command if applicable and the Decision Making stage sends a command to the home automation system according to the context. the following sections describe the most important and difficult tasks of the project and first of all our study for taking the needs of potential users into account.

Home support design for elderly

In an anthropocentric (end-user centred) and ethnocentric (focused on the real activities of home actors) approach, the participatory design approach that we are developing in this research aims to prospect the use modalities as well as the "situated acceptance" conditions [1] of the VocADom device. The objective is to accompany the design and to support the final appropriation of the device within the home and with the various actors of the home. A major challenge guides the development of these methods, it consist in designing non-intrusive approaches that respect the social and psychological integrity of frail elderly people, and that also take into account the functional and cognitive limitations of these future users. This is why in this research, we mobilize a methodological triangulation approach (Flick, 1992), which articulates different methods of analysis. These must both be adapted to a public with specific needs (the elderly, dependent and visually impaired) and allow us to explore the situated acceptance of technologies [1]. In this approach, the idea is not so much that home actors have a favourable a priori representation of the technical object, as generally assessed by social acceptability models [3, 5], but that these technologies do have a favourable effect on the individual and collective practices of these people [1]. The methods we deploy therefore aim to determine these impacts, as close as possible to life situations and probable future conditions of use. They are available at three levels: 1) First, we aim the real activities, lifestyle habits and organisation/coordination modes of the different actors in the home by analysing the system of activity. The objective of this part is the knowledge of the context of life in order to determine the role, place and function that the technological artifact VocADom device should have in this ecosystem: when should it intervene, to provide which services, to support which types of activity, without harming or affecting such others? How should the tool be integrated into the activity systems of the different actors in the home in order to promote their development and articulation (without constraining or weakening them)? 2) In a second phase, it is about co-designing, through a participatory approach –with the different actors of the home and the designers of the

[2] https://www.openhab.org/

Figure 2: Instrumented kitchen in the smart home.

consortium– services, functionalities and interaction modalities (voice feedback, HMI) of the future device. 3) Finally, a final phase consists in evaluating and testing the prototype in the actual living situation, at home itself, in order to understand the difficulties of use and the various impacts, acceptable or not, of the Vocadem system on the systems of activity of the various actors in the home. The objective is to evaluate whether the VocADom system provides real efficiency and makes it possible to maintain meaning (on social, identity, relational and organisational dimensions) in the activities carried out at home.

Experiment in a Smart Home

To provide data to test and train the different processing stages of the VocADom system, experiments were run in the instrumented apartment Amiqual4Home [4]. The 87m^2 smart home is equipped with home automation systems, multimedia, water and electricity meters, and means for observing human activity. The kitchen and the living room are on the ground floor, the bedroom and the bathroom on the first floor. This Smart Home is fully functional and equipped with sensors, such as energy and water consumption, level of hygrometry, temperature, and actuators able to control devices, lighting, shutters, multimedia diffusion, distributed in all rooms. Home automation actuators were activated through an experimenter operating as a Wizard of Oz.

The keyword is followed by a sentence specifying the voice command. The voice commands are grouped into two main categories:

- checking-question : KEYWORD Est-ce que la porte est fermée ? ("Is the door closed?")
- command : KEYWORD Allume la télévision ("Turn on the TV")

Before starting the dataset collection (with the participant playing domestic activities), an experimenter explained the objectives and the participants role (with a consent form to sign 4), the participant visited the apartment (alone) and chose a keyword to control the home. To collect a multi-source controlled and spontaneous corpus of voice base interaction in the home, three recording phases were defined as:

- Phase 1: Graphical based instruction to elicite spontaneous voice commands. In this phase, the participant is mostly alone.
- Phase 2: Two-inhabitant scenario enacting a visit by a friend. Activities were planned as well as the kind of voice command to utter. However, there was complete freedom in the way the activities and the sentence were performed.
- Phase 3: Noisy voice commands. In that phase, the participant is alone except during the reading by a third person of a text on the ground floor. There were 11 positions at which the participant had to be (6 on the ground floor living room / kitchen, 2 in the bathroom and 3 in the room). There were, in each case, 5 sentences to read in each position and each noise context.

Eleven people between 23 and 25 years old participated to the experiments. Corpus duration is 12h 28mn 45s. Keywords chose by participants were vocadom, hé cirrus, ulysse, téraphim, allo cirrus, ichefix, minouche and hestia.

Speech Enhancement

Speech enhancement occurs at two different stages. In the first stage, we remove noises for which a reference signal is available (e.g., TV, radio) from the microphone array recordings using an acoustic echo cancellation

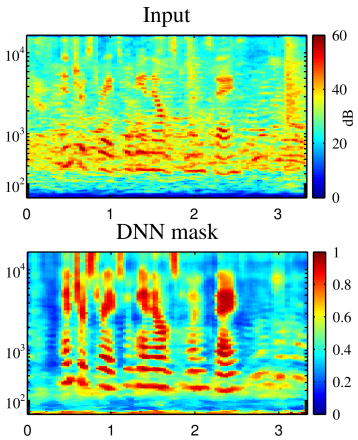

Figure 3: Time-frequency mask estimated by a DNN.

technique. In the second stage, we remove other noises and separate the target speaker's voice signal from other concurrent speakers. State-of-the-art methods are designed for a single speaker in noise. They are based on beamformers (a.k.a. spatial filters) derived from time-frequency representations of speech and noise estimated by a deep neural network (DNN) [11]. We are currently extending this approach in order to handle overlapping speakers by localising the target speaker, i.e., estimating his/her spatial direction with respect to the microphone array, and then using this information to estimate a time-frequency mask and derive a beamformer in a way similar to [7]. We developed a DNN-based keyword-dependent speaker localisation method that exploits knowledge of the keyword uttered by the target speaker in order to discriminate him/her from other speakers. To the best of our knowledge this had not been considered in the literature before. We are now seeking to integrate this method with speech enhancement in a way that is robust to localisation errors.

Natural Language Understanding for Voice Command

Previous studies showed that smart home users don't agree with the use of a formal grammar for voice commands [9]. Understanding of the intent included in the vocal command is then necessary.

Despite growing interest in smart-homes, semantically-annotated voice command corpora are scarce, especially for languages other than English. Yet, large corpora are necessary to develop advanced Natural Language Understanding (NLU) tools. Therefore, we are developping an approach to generating customizable synthetic corpora of semantically-annotated smart-home commands. We use it to develop an artificial French corpus for the smart-home. We evaluated this corpus using the smaller corpus of real users interacting with Amiqual4Home presented in previous section. Two state-of-the-art NLU models a triangular CRF and an attention-based RNN were trained on the synthetic dataset (and on Port-Media) and tested on the real smart-home corpus. Results demonstrate that the current synthetic corpus permits accurate prediction of user intents, but accurate slot-filling requires modification to state- of-the-art models. We release the artificial and real datasets as the first (to our knowledge) publicly available, French, semantically-annotated smart-home corpus.

Conclusion and Future Works

This paper presents an overview of the VocADom project. Several steps have been completed to provide real-time detection and speech recognition in the house. This technology can benefit both the disabled and the elderly population who have difficulties in moving or seeing and want security reassurance.

Next steps in the project include the improvement of the current audio processing algorithms and further developments of the intelligent controller. Integration of the different modules in a real-time system is also an important aspect of the project. The resulting system is planned to be tested in different homes (from fully equipped with home automation devices to poorly equipped) to validate the reliability of the approach and with elderly or visually impaired people to get feedback from this targeted population.

Acknowledgements

This work is part of the VocADom project founded by the French National Agency (Agence Nationale de la Recherche / ANR-16-CE33-0006).The authors would like

to thank the participants who took part to the different experiments. Thanks are also extended to N. Bonnefond ans S. Humblot for their support during the experiments inside the smart home.

References

[1] Bobillier Chaumon, M., B., C., Bekkadja, S., and Cros, F. Detecting Falls at Home: User-Centered Design of a Pervasive Technology. *Human & Technology : An Interdisciplinary Journal on Humans in ICT Environments 12(2)* (2016), 165–192.

[2] Chan, M., Estve, D., Escriba, C., and Campo, E. A review of smart homes- present state and future challenges. *Computer Methods and Programs in Biomedicine 91*, 1 (2008), 55–81.

[3] Davis, F. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly 13(3)* (1989), 329–340.

[4] Lago, P., Lang, F., Roncancio, C., Jiménez-Guarín, C., Mateescu, R., and Bonnefond, N. The ContextAct@A4H real-life dataset of daily-living activities Activity recognition using model checking. In *CONTEXT*, vol. 10257 of *LNCS* (2017), 175–188.

[5] Peek, S. T. M., Wouters, E. J. M., van Hoof, J., Luijkx, K. G., B., R., H., and Vrijhoef, H. J. M. Factors influencing acceptance of technology for aging in place: A systematic review. *International Journal of Medical Informatics 83(4)* (2014), 235248.

[6] Peetoom, K. K. B., Lexis, M. A. S., Joore, M., Dirksen, C. D., and Witte, L. P. D. Literature review on monitoring technologies and their outcomes in independently living elderly people. *Disability and Rehabilitation: Assistive Technology 10*, 4 (2015), 271–294.

[7] Perotin, L., Serizel, R., Vincent, E., and Gurin, A. Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (2018), 1–5.

[8] Portet, F., Vacher, M., Golanski, C., Roux, C., and Meillon, B. Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects. *Personal and Ubiquitous Computing 17*, 1 (2013), 127–144.

[9] Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., and Chahuara, P. Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing 7*, issue 2 (2015), 5:1–5:36.

[10] Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. The Sweet-Home speech and multimodal corpus for home automation interaction. In *Proceedings of 9th edition of the Language Resources and Evaluation Conference (LREC)* (2014), 4499–4506.

[11] Wang, Z., Vincent, E., Serizel, R., and Yan, Y. Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments. *Computer Speech and Language 49* (2018), 37–51.