33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

The SWEET-HOME Project: Audio Technology in Smart Homes to improve Well-being and Reliance

Michel Vacher, Dan Istrate, François Portet, Thierry Joubert, Thierry Chevalier, Serge Smidtas, Brigitte Meillon, Benjamin Lecouteux, Mohamed Sehili, Pedro Chahuara and Sylvain Méniard

***Abstract*— The SWEET-HOME project aims at providing audio-based interaction technology that lets the user have full control over their home environment, at detecting distress situations and at easing the social inclusion of the elderly and frail population. This paper presents an overview of the project focusing on the multimodal sound corpus acquisition and labelling and on the investigated techniques for speech and sound recognition. The user study and the recognition performances show the interest of this audio technology.**

I. INTRODUCTION

Demographic change and ageing in developed countries imply challenges for the society to continue to improve the well being of its elderly and frail inhabitants. Since the dramatic evolution of Information and Communication Technologies (ICT), one way to achieve this aim is to promote the development of smart homes. In the health domain, a *health smart home* is a habitation equipped with a set of sensors, actuators, automated devices and centralised software controllers specifically designed for daily living task support, early detection of distress situations, remote monitoring and promotion of safety and well-being [2]. The smart home domain is greatly influenced by the *Ubiquitous Computing* domain. As introduced by Weiser [15], ubiquitous computing refers to the computing technology which disappears into the background, which becomes so seamlessly integrated into our environment that we do use it naturally without noticing it. Among all the interaction and sensing technologies used in smart homes (e.g., infra-red sensors, contact doors, video cameras, RFID tags, etc.), audio processing technology has a great potential to become one of the major interaction modalities. Audio technology has not only reached a stage of maturity but has also many properties that fit the Weiser's vision. It is physically intangible and depending on the number and type of the sensors (omnidirectional microphones) that are used, it does not force the user to be physically at a particular place in order to operate. Moreover, it can provide interaction using natural language so that the user does not have to learn complex computing procedures or jargon. It can also capture sounds of everyday life which makes it even more easy to use and can be used to communicate with the user using synthetic or pre-recorded voice. More generally, voice interfaces can be much more adapted to disabled people and the ageing population who have difficulties in moving or seeing than tactile interfaces (e.g., remote control) which require physical and visual interaction. Moreover, audio processing is particularly suited to distress situations. A person, who cannot move after a fall but being conscious, has still the possibility to call for assistance while a remote control may be unreachable. Despite all this, very few smart home projects have seriously considered audio technology and notably speech recognition in their design [3], [6]. Part of this can be attributed to the complexity of setting up this technology in a real environment and to important challenges that still need to be overcome [14].

This paper presents the SWEET-HOME project which started in 2010 to address several of these issues. An introduction to the project is given in Section II. The first task of the project was to acquire a large sample of situations for training and analysis purpose. This acquisition is detailed in Section III. Then, the core audio technologies being developed, namely, every day sound detection and distant speech recognition, are presented in Section IV and V. The paper ends with an outlook of future works.

II. THE SWEET-HOME PROJECT

The SWEET-HOME project (`sweet-home.imag.fr`) is a French national supported research project aiming at designing a new smart home system based on audio technology focusing on three main aspects: to provide assistance via *natural man-machine interaction* (voice and tactile command), to ease *social inclusion* and to provide *security reassurance* by detecting situations of distress. If these aims are achieved, then the person will be able to pilot, from anywhere in the house, their environment at any time in the most natural way possible. The project is made up of researchers and engineers from the Laboratory of Informatics of Grenoble (specialised in speech processing, smart home design and evaluation), the Esigetel (specialised in audio technology) and from three companies: Theoris (real-time system development and integration), Camera Contact (diffusion of adapted services for maintenance at home) and Technosens (videophone equipment for the elderly).

The targeted users are elderly people who are frail but still autonomous. There are two reasons for this choice. Firstly, a

This work is a part of the Sweet-Home project founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-09—VERS-011)

M. Vacher, F. Portet, B. Lecouteux, B. Meillon, P. Chahuara and S. Méniard are with the Laboratory of Informatics of Grenoble, UMR CNRS/UJF/G-INP 5217, Grenoble, France `forname.name@imag.fr`

D. Istrate and M. Sehili are with ESIGETEL, 77210 Avon-Fontainebleau, France `dan.istrate@esigetel.fr`

T. Joubert is with THEORIS, 103 rue La Fayette - 75010 Paris, France `thierry.joubert@theoris.fr`

T. Chevalier is with TECHNOSENS, 15 rue d'Alsace, 69150 Décines, France `thierry.chevalier@technosens.fr`

S. Smidtas is with VISAGE, 75000, Paris, France `visage@camera-contact.com`

domotic system is costly and would be much more profitable if it is use in a life-long way rather than only when the need for assistance appears. Secondly, in the case of a loss of autonomy, the person would continue to use their own system with some adaptations needed by the new situation (e.g., wheelchair) rather than having to cope simultaneously with their loss of autonomy and a new way of life imposed by the smart home. To assess the acceptance of this new technology, a qualitative user evaluation was performed [10]. 8 healthy persons between 71 and 88 years old, 7 relatives (child, grand-child or friend) and 3 professional carers were questioned in co-discovery in a fully equipped smart home alternating between interview and wizard of Oz periods. The four important aspects of the project have been assessed: voice command, communication with the outside world, domotic system interrupting a person's activity, and electronic agenda. In each case, the voice based solution was far better accepted than more intrusive solutions (e.g., video camera). Thus, in accordance with other user studies [6], audio technology appears to have a great potential to ease daily living for elderly and frail persons. To respect privacy, it must be emphasized that the adopted solution analyses the audio information on the fly and does not store the raw audio signal. Moreover, the speech recognizer is made to recognize only a limited set of predefined sentences which prevents recognition of intimate conversations.

Regarding the technical aspect, the project tries to make use of already standardised technologies and applications rather than building communication buses and purpose designed material from scratch. As emphasized in [7], standards ensure compatibility between devices and ease the maintenance as well as orient the smart home design toward cheaper solutions. In our case, the KNX bus system (KoNneX), a worldwide ISO standard (ISO/IEC 14543) for home and building control has been chosen as main communication bus. Another example of this strategy is that SWEET-HOME includes already designed systems such as the e-lio (`www.technosens.fr`) or Visage (`camera-contact.com`) systems which will make it easier for the user to connect with their relative, grocer or caregiver. We believe this strategy is the most realistic one given the large spectrum of skills that are required to build a complete smart home system. Moreover, in order for the user to be in full control of the system and also in order to adapt to the users' preferences, three ways of commanding the system are possible: voice order, PDA or classical tactile interface (e.g. switch).

The architecture of the system is depicted Figure 1. The input is composed of the information from the domotic network and information from the 8 microphones (M1, M2...M8) transmitted through radio frequency channels. Thus, information can be provided directly by the user (e.g., voice order) or via environmental sensors (e.g., temperature). The information coming from the domotic system is transmitted on-line to the intelligent controller (through several preprocessing steps). This controller captures all streams of data, interprets them and executes the required actions. The knowledge of the controller is defined using two semantic layers: the *low-level* ontology, devoted to the representation

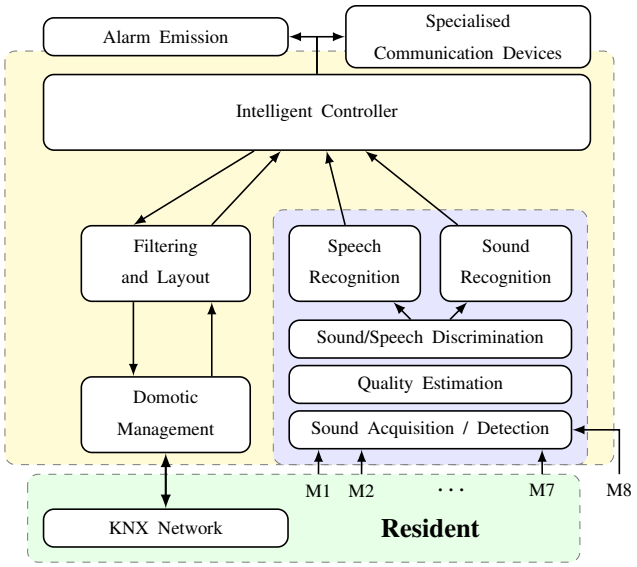

Fig. 1. Block diagram of the SWEET-HOME system architecture.

of raw data and network information description (e.g., state of switches and actuators); and the *high-level* ontology which represents concepts being used at the reasoning level. This separation between low and high levels makes it possible a high re-usability of the reasoning layer when the sensor network and the home have to be adapted [4]. Once the current situation/order has been determined, the controller either reacts to the order or acts pro-actively by modifying the environment without an order (e.g. turns off the light when nobody is in the room). Outputs of the system thus include domotic commands but also interactions with the user when a vocal order was not understood for example, or in case of alert messages (e.g. turn off the gas, remind the person of an appointment).

While the domotic system provides symbolic information, raw audio signals must be processed to extract information from speech and sound. This extraction will be based on our experience in developing the AUDITHIS system [13], a real time multi-threaded audio processing system for ubiquitous environments. The audio information is composed of 8 channels, seven are devoted to the house and another one is free for specific application (e.g., hand microphone, radio). Succinctly, *Acquisition / Detection* stage simultaneously detects in real-time the audio events on the 8 input channels of a sound card at a 16 kHz sampling rate by an adaptive threshold algorithm based on the discrete wavelet transform. Data of each channel is buffered and processed sequentially and separately. At the *Quality* stage, noise level (and other parameters) is evaluated to assess the Signal-to-Noise Ratio (SNR) of each acquired sound. The events are then discriminated between speech and everyday life sound using Gaussian Mixture Model (GMM) and then undergo further processing to recognize either the uttered sentence or the type of sound (e.g., object fall, door slam). These two stages are further described in sections IV and V.

Fig. 2. Images captured by the DOMUS video cameras during the experiment

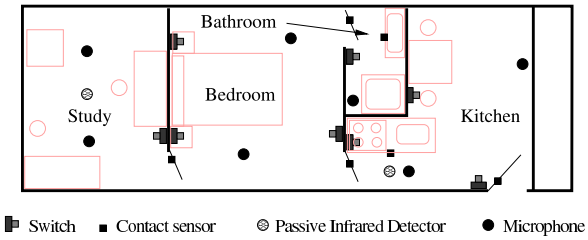

Fig. 3. Layout of the DOMUS Smart Home and position of the sensors.

III. SMART HOME CORPUS ACQUISITION

To provide data to test and train the different processing stages of the SWEET-HOME system, experiments were run in the DOMUS smart home that was designed by the Multicom team of the Laboratory of Informatics of Grenoble to observe users' activities interacting with the ambient intelligence of the environment. Figure 3 shows the details of the flat. It is a thirty square meters suite flat including a bathroom, a kitchen, a bedroom and a study, all equipped with sensors and effectors so that it is possible to act on the sensory ambiance, depending on the context and the user's habits. The flat is fully usable and can accommodate a dweller for several days. The technical architecture of DOMUS is based on KNX. More than 150 sensors, actuators and information providers are managed in the flat (e.g., lighting, shutters, security systems, energy management, heating, etc.). A residential gateway architecture has been designed, supported by a virtual KNX layer seen as an OSGI service (Open Services Gateway Initiative) to guarantee the interoperability of the data coming and to allow the communication with virtual applications, such as activity tracking. The flat has also been equipped with 7 radio microphones set into the ceiling that can be recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card [13].

21 persons (including 7 women) participated to an experiment to record multimodal data in a daily living context. To make sure that the data acquired would be as close as possible to real daily living data, the participants were asked to perform several daily living activities in the smart home. The average age of the participants was 38.5 ± 13 years (22-63, min-max). The experiment consisted in following a scenario of activities without condition on the time spent and the manner of achieving them (e.g., having a breakfast, simulate a shower, get some sleep, clean up the flat using the vacuum, etc.). Figure 2 shows participants performing activities in the different rooms of the smart home. A visit, before the experiment, was organized to make sure that the participants will find all the items necessary to perform the activities. During the experiment, event traces from the domotic network, audio and video sensors were captured. Video data were only captured for manual marking up and are not intended to be used for automatic processing. In total, more than 26 hours of data have been acquired.

To use the multimodal corpus for training and testing, data of interest must be annotated. Participants' activities, location, etc. were annotated using a standard software. However, sound annotation requires an heavier procedure given its transient nature. A detection algorithm [11] was applied to the seven channels to detect intervals of sounds of interest. Then, for human annotation purpose, a unique signal resulting of the combination of the seven channels using a weighted sum with coefficients varying with the signals energy was created. Moreover, sound intervals were fused by making the union of overlapping intervals of the seven channels. This signal, the merged intervals, and the videos were then used for the annotation. The resulting annotation file contains the start, the end and the class of the sound or speech event.

IV. SOUND CLASSIFICATION

In a home automation context, many daily living sounds are interesting to detect in order to obtain direct information about the resident status (e.g., cry, snoring) or to help the home automation system disambiguate situations (e.g., textile handling sound during a wearing activity inference). A first step of the sound detection task was to identify the most useful sound classes to recognize according to the project objectives and users' feedback. Sounds were grouped based on spectral proximity in generic classes such as human sounds (cough, yawn, screams, cry, snoring), water type, electrical engine (hair dryer, store engine, vacuum cleaner), object falls, etc.

Two approaches were developed to classify sounds of everyday living. A GMM (Gaussian Mixture Models) based classifier with an optimum number of Gaussians determined using the *ANASON* software developed at ESIGETEL and an SVM (Support Vector Machine) based approach. In a previous experiment [12], the GMM was trained on a corpus with sounds described by 16 MFCC (Mel-Frequency Cepstral Coefficient) acoustical parameters while the SVM was trained on the same corpus using a Gaussian RBF kernel and a one-versus-one approach. 14 sound classes were used (breathlessness, door-slamming, electrical shaver, female cry, female scream, male scream, hair-dryer, paper, glass breaking, laugh, sneeze, cough, dishes and yawn).

As the Sweet-Home corpus is currently being annotated, the preliminary tests have been done using a database composed of local recordings and some downloaded files. The GMM led to higher results than the SVM one (92% of correct classifications vs. 87%). But, that can be due to the naive SVM aggregation implemented method which simply sums up the score of all frames (about 16ms) of a sound to get the global score. However, the corpus need to be extended and a hierarchical classification approach will be implemented in order to increase classification robustness.

V. SPEECH RECOGNITION

Automatic Speech Recognition systems (ASR) have reached good performances with close talking microphones (e.g. head-set), but the performance decreases significantly as soon as the microphone is moved away from the mouth of the speaker (e.g., when the microphone is set in the ceiling). This deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise such as TV, radio and devices. All these problems should be taken into account in the home context and have become hot topics in the speech processing community [1].

In the SWEET-HOME project, all the sentences uttered in the flat are not to be recognized, only vocal orders or some distress sentences need to be detected. Term detection has been extensively studied in the last decades in the two different contexts of spoken term detection: large speech databases and keyword spotting in continuous speech streams. The first topic recently faced a growing interest, stemming from the critical need of content-based structuring of audio-visual collections. Performances reported in the literature are quite good in clean conditions, especially with broadcast news data. However, performances of state-of-the-art approach are unknown in noisy situation such as the one considered in SWEET-HOME. This section summarises experiments that were run to test to which extend standard and research ASR systems can be used in this context.

The SWEET-HOME speech corpus was built from the multimodal SWEET-HOME corpus. It is made of 1779 sentences uttered by 21 persons and each channel lasts about 1 hour 19 minutes 13 s. The average SNR (Signal-to-Noise Ratio) for the considered sentences is 20.3 dB.

In a first step, two ASR systems were used: Sphinx [9] and Speeral [8] working with different graph exploration algorithms to build a baseline system. The best baseline system was composed of interpolated language models between a large vocabulary one and a specialized one (composed of predefined domotic sentences); and used speaker acoustic adaptation. The best baseline obtained a 14.5% Word Error Rate and a 16.9% Classification Error Rate.

We plan to develop a robust original approach to benefit from the multiple microphones of the smart home and from a priori knowledge about the sentences being uttered. This approach will be based on the Driven Decoding Algorithm (DDA) which permits to drive a audio stream being decoded by the results of the decoding on another one [5].

VI. CONCLUSIONS AND FUTURE WORKS

This paper presents an overview of the SWEET-HOME project which aims at designing a new smart home system based on audio technology. Several steps have been completed to provide real-time detection of sound of everyday living and speech recognition in the house. This technology can benefit both the disabled and the elderly population that 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 domotic devices to poorly equipped) to test the reliance of the approach and with elderly people to get essential feedback from this targeted population. Finally, social inclusion will undergo separated tests using specialised devices to find out which information can be usefully exchanged between them in order to improve both the user's environment and her communication capacity.

VII. ACKNOWLEDGMENTS

The authors would like to thank the participants who took part to the different experiments. Thanks are also extended to N. Bonnefond and S. Pons for their support during the experiments inside the smart home.

REFERENCES

[1] J. Barker, H. Christensen, N. Ma, P. Green, and E. Vincent, "The pascal 'chime' speech separation and recognition challenge," in *Interspeech 2011*, 2011.

[2] M. Chan, E. Campo, D. Estève, and J.-Y. Fourniols, "Smart homes — current features and future perspectives," *Maturitas*, vol. 64, no. 2, pp. 90–97, 2009.

[3] 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.

[4] M. Klein, A. Schmidt, and R. Lauer, "Ontology-centred design of an ambient middleware for assisted living: The case of soprano," in *30th Annual German Conference on Artificial Intelligence (KI 2007)*, 2007.

[5] B. Lecouteux, G. Linarès, Y. Estève, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 1549–1552.

[6] R. López-Cózar and Z. Callejas, "Multimodal dialogue for ambient intelligence and smart environments," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. C. Augusto, Eds. Springer US, 2010, pp. 559–579.

[7] F. Mäyrä, A. Soronen, J. Vanhala, J. Mikkonen, M. Zakrzewski, I. Koskinen, and K. Kuusela, "Probing a proactive home: Challenges in researching and designing everyday smart environments," *Human Technology*, vol. 2, pp. 158–186, 2006.

[8] P. Nocera, G. Linares, and D. Massonié, "Principes et performances du décodeur parole continue speeral," in *JEP 2002*, 2002.

[9] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer, "The 1996 hub-4 sphinx-3 system," in *Proc. of the 1997 ARPA Speech Recognition Workshop*, 1997, pp. 85–89.

[10] 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*, in press.

[11] J. Rougui, D. Istrate, and W. Souidene, "Audio sound event identification for distress situations and context awareness," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, Minneapolis, USA, 2009, pp. 3501–3504.

[12] M. A. Sehili, D. Istrate, and J. Boudy, "Primary investigations of sound recognition for a domotic application using support vector," *Annals of the University of Craiova, Series: Automation, Computers, Electronics and Mechatronics*, vol. 7(34), no. 2, pp. 61–65, 2010.

[13] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, and N. Noury, *Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living*. Intech Book, 2010, pp. 645 – 673.

[14] 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.

[15] M. Weiser, "The computer for the 21st century," *Scientific American*, vol. 265, no. 3, pp. 66–75, 1991.