Distant Speech Processing for Smart Home Comparison of ASR approaches in distributed microphone network for voice command

Benjamin Lecouteux, Michel Vacher, François Portet

International Journal of Speech Technology manuscript No.
(will be inserted by the editor)

Distant Speech Processing for Smart Home

Comparison of ASR approaches in distributed microphone network for voice command

Benjamin Lecouteux · Michel Vacher · François Portet

the date of receipt and acceptance should be inserted later

Abstract Voice command in multi-room smart homes for assisting people in loss of autonomy in their daily activities faces several challenges, one of them being the distant condition which impacts ASR performance. This paper presents an overview of multiple techniques for fusion of multi-source audio (pre, middle, post fusion) for automatic speech recognition for in-home voice command. The robustness of the models of speech is obtained by adaptation to the environment and to the task. Experiments are based on several publicly available realistic datasets with participants enacting activities of daily life. The corpora were recorded in natural condition, meaning background noise is sporadic, so there is no extensive background noise in the data. The smart home is equipped with one or two microphones in each room, the distance between them being larger than 1 meter. An evaluation of the most suited techniques improves voice command recognition at the decoding level, by using multiple sources and model adaptation. Although Word Error Rate (WER) is between 26% and 40%, Domotic Error Rate (identical to the WER, but at the level of the voice command) is less than 5.8% for deep neural network models, the method using Feature space Maximum Likelihood Linear Regression (fMLLR) with speaker adaptation training and Subspace Gaussian Mixture Model (SGMM) exhibits comparable results.

Keywords Home automation · voice command · Smart Home · Ambient Assisted Living · Multichannel analysis

This work is supported by the Agence Nationale de la Recherche under grant ANR-09-VERS-011 in the framework of the Sweet-Home project.

Address(es) of author(s) should be given

1 Introduction

In beginning of the twenty-first century, most of the countries, whatever their gross domestic product, are undergoing a major demographic transition which will bring the large amount of baby boomers from full-time workers to full-time pensioners. This progressive ageing of most of the world population will be correlated with an increase of people with disabilities (World Health Organization, 2003). Some of these people will be incapacitated to the point at which they can no longer live independently in their own homes. However, one of the first wishes of this population is to live in their own home as cosy and safe as possible even if their autonomy decreases. Anticipating and responding to the needs of persons with loss of autonomy with Information and Communications Technology (ICT) is known as Ambient Assisted Living (AAL). In this domain, the development of smart homes is seen as a promising way of achieving in-home daily assistance (Chan et al, 2008; Peetoom et al, 2014). However, given the diverse profiles of the users (e.g., low/high technical skill, disabilities, etc.), complex interfaces should be avoided. Nowadays, one of the best interfaces, is the Voice-User Interface (VUI), whose technology is mature and provides interaction using natural language so that the user does not have to learn complex computing procedures (Portet et al, 2013; Vacher et al, 2015a). Moreover, it is well adapted to people with reduced mobility and to some emergency situations (hands-free and distant interaction).

VUI in domestic environments recently gained interest in the speech processing community as exemplified by the rising number of smart home projects that consider Automatic Speech Recognition (ASR) in their design (Charalampos and Maglogiannis, 2008; Popescu et al, 2008; Badii and Boudy, 2009; Hamill et al, 2009; Filho and Moir, 2010; Lecouteux et al, 2011; Ons et al, 2014; Christensen et al, 2013; Cristoforetti et al, 2014; Vacher et al, 2015a). However, though VUIs are frequently used in smart-phones there are still important challenges to overcome before implementing VUI at home (Vacher et al, 2011). Indeed, the task imposes several constraints to the speech technology: 1) distant speech condition, 2) cheap, 3) real-time, 4) respect of privacy[1]. Moreover, such technology must be validated in real situations (i.e. real smart homes and users). Another very important challenge is the ability to perform automatic speech processing in domestic condition with small quantity of data. Indeed, real life acoustic environment can be composed of a variety of highly dynamic background noise, interleaving speech and reverberation that is highly challenging for an ASR system. To address this challenge there has been a recent serie of effort to foster research in distant speech ASR in noisy condition as exemplified by the CHiME challenge (Vincent et al, 2013). In this challenge, a data set was synthesised from a set of read utterance mixed with a background noise (composed of real-life domestic background noise) using a binaural room impulse responses. The Word Error Rate (WER) rate

[1] Note that as any assistive technology, the intrusiveness of an ICT can be accepted if the benefit is worth it.

shows a trend from 40% in -9dB condition to 15% in 9dB condition. To move from these somewhat artificial data the challenge to more realistic conditions, CHiME 2014 (Barker et al, 2015) has recorded speech in real noisy conditions (rather than artificiality mixed) but close talking and microphones (6 microphones attached ion a tablet). In this context an impressive WER of 5.8% has been reached (Yoshioka et al, 2015). However, it is difficult to project these results in a non-read distant speech multiroom real-time ASR setting with distributed microphone signals that contain low redundancy. Moreover the baseline of these challenges are difficult to compare for non-English languages and other recording conditions.

In this context, this paper presents the results of an ASR development and evaluation for a VUI intended for elderly people and people with visual impairment in a multiroom smart home with distributed microphones (several meters apart, i.e. no array of microphones). This research supplement our early developments made in the context of the SWEET-HOME project in which state-of-the-art HMM-GMM systems with multi channel were developed and evaluated with target users in a real smart home (Vacher et al, 2015a), and whose corpora were made available (Vacher et al, 2014). In this particular paper we extend our previous research in a number of ways. First, we developed new ASR systems based on the current state-of-the-art acoustic modeling namely HMM-DNN and S-GMM. In particular, these models were compared in term of accuracy and dependency to the amount of speech material available before hand which is a real constraint for the application (e.g. the adaptation to each new home and users)[2]. Second, we present various techniques to benefit from the available set of microphones with low *a priori* knowledge (i.e. only the information about the room in which each microphone is set). This constraint is due to the fact that the installation of a home automation system must be kept as simple as possible with minimal restrictions on the material to be bought and on where it should be placed. Third, this multichannel ASR system has be evaluated using both standard measures (such as WER) and using task oriented measures (such as DER : Domotic Error Rate) and from manually and automatically segmented speech signal. This last condition is compulsory for any hands-free ASR system for real-life application since ASR performance depends on the Voice Activity Detection (VAD) accuracy. The experiments have been based on several datasets we have collected in our smart home with participants enacting activities of daily life. The corpora were recorded in realistic condition with the consequence that background noise is sporadic. For this reason the paper focuses on achieving robust real-time ASR, using multisource ASR rather than using sound source localization, and separation which would not be adapted to a distributed microphone setting where microphones are mono and their locations are subject to change.

This study is part of a system which provides voice command in a multiroom smart home for seniors and people with visual impairment. In our approach, we address the problem by using several mono-microphones set in

[2] See (Zhang et al, 2014) for an interesting study on these matters.

the ceiling, selecting the "best" sources and employing ASR decoding and voice command matching. This approach has been chosen against noise source separation which can be highly computational expensive, is sensitive to sample synchronization problem (which cannot be assumed with non professional devices) and is still not solved in real uncontrolled condition. Hands-free interaction is ensured by constant keyword detection. Indeed, the user must be able to command the environment without having to wear a specific device for physical interaction (e.g. a remote control too far from the user when needed). Though microphones in a home is a real breach of privacy, by contrast to current smart-phones, we address the problem using an in-home ASR engine rather than a cloud based one (private conversations do not go outside the home). Moreover, the limited vocabulary ensures that only relevant speech for the command of the home is correctly decoded. Finally, another strength of the approach is to have been evaluated with real users in realistic uncontrolled conditions.

The paper is organized as follow. After a short introduction to the related work in Section 2, the overall system and the different ASR strategies are described in Section 3. These strategies are experimented in Section 4, and the results of the off-line experiments are presented in Section 5. The results of the proposed methods are discussed in Section 6.

2 Related works

As stated in the introduction, several challenges are to be addressed to make distant speech recognition in Smart Homes performing well enough to provide speech based services to the dweller (Vacher et al, 2011). ASR systems obtain acceptable performances with clean close talking microphones, but the performances are significantly lower when the microphone is far from the mouth of the speaker. This deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise. Moreover, many speech controlled smart home projects are focused on AAL (Ambient Assisted Living), that adds the challenge of dealing with atypical voice. In the following, we briefly introduce some of the these challenges focusing on the literature addressing these issues by using ASR.

2.1 Diversity of speakers and situations

If ASR has reached good performance for typical users, a large field of application in speech based assistive technology in the home aims at supporting daily life of atypical users (Portet et al, 2015). Potential users are elderly and all people who may acquire a disability which affects communication. This disability can result from both motor and cognitive impairments (i.e. paralysis, hearing or visual impairment, brain injury, Alzheimer...). In speech recognition, the challenges are related to the recognition of speech uttered by elderly, dysarthric or cognitively impaired speakers.

For instance, aged voice is characterized by some specific features such as imprecise production of consonants, tremors, hesitations and slower articulation (Ryan and Burk, 1974). Some studies have shown age-related degeneration with atrophy of voice cords, calcification of laryngeal cartilages, and changes in muscles of larynx (Takeda et al, 2000)(Mueller et al, 1984). For these reasons, some authors highlight that ASR performance decreases with elderly voice. This phenomenon has been observed in the case of English, European Portuguese, Japanese and French (Vipperla et al, 2009)(Pellegrini et al, 2012)(Baba et al, 2004)(Aman et al, 2013). Vipperla et al (Vipperla et al, 2008) showed that speaker adaptation can get closer to the scores of non-aged speakers but this implies that the ASR must be adapted to each speaker.

Regarding speech impaired users, various ASR systems have been proposed in the literature. In (Potamianos and Neti, 2001; Rudzicz, 2011) speaker-independent acoustic models were adapted to speaker so that recognition of user-specific voiceisations was improved. Another way to improve ASR performance for dysarthric speakers was to train the consistency of the speakers' pronunciations using the recognition likelihood of the uttered words (Parker et al, 2006). Thus, it is the user who adapts itself to the ASR system. Other studies include the design of phoneme HMM topologies more suited to the speaker (Caballero-Morales and Trujillo-Romero, 2014) or user customizable isolated word recognition systems (Hwang et al, 2012).

Moreover, speech signal contains linguistic information but it may be influenced by the health, the social status and the emotional state (Audibert et al, 2005)(Vlasenko et al, 2011). Recent studies suggest that ASR performance decreases in case of emotional speech (Vlasenko et al, 2012), however it is still an under-researched area. In their study, Vlasenko et al (Vlasenko et al, 2012) demonstrated that acoustic models trained on read speech samples and adapted to acted emotional speech could provide better performance for spontaneous emotional speech recognition.

In voice based controlled home environments, the approach is mainly to use ASR models together with a speaker adaptation procedure to improve ASR performance for specific speakers. For instance, in the Sweet-Home (Vacher et al, 2015a) and Cirdo (Bouakaz et al, 2014), projects that aimed at providing voice based assistive technology in the home for elderly and disabled people, Maximum Likelihood Linear Regression (MLLR) and Feature space MLLR (fMMLR) speaker adaptation was used to adapt an on-line speaker-independent ASR system. In the HomeService project (Christensen et al, 2013), Maximum a posteriori (MAP) adaptation was used for speaker adaptation. Another approach was presented in the ALADIN project (Ons et al, 2014), in which a VUI model is learned from the speech and actions of the user without transcription. The speech of the user and the user's action on a device (home automation command) are two sources of information that are combined using Non-negative Matrix Factorization (NMF) so that the VUI can learn co-occurring patterns from two information sources. Although this approach requires few examples to learn, it is unclear how it can generalize to unseen situation and how the model can be reused.

Apart from acoustic modeling, language modeling is also an important issue in order to minimize the ambiguity in the decoding. For small vocabulary systems, the language model must be adapted to the speaker since even in limited vocabulary task, the user tends to deviate from the system syntax (Vacher et al, 2015a).

2.2 Reverberation

In real multi-room home it is frequent to observe the reverberation phenomenon that can alter source speech. Distorted signals can be treated in ASR either at the acoustic model level or at the input (feature) level (Wölfel and McDonough, 2009). (Deng et al, 2000) showed that feature adaptation methods provide better performances than those obtained with systems trained with data with the same distortion as the target environment (e.g. acoustic models learned with distorted data) for both stationary and non stationary noise conditions. Moreover, when the reverberation time is above 500ms, ASR performances are not significantly improved when the acoustic models are trained on distorted data (Baba et al, 2002). In our study, the home environment into consideration presents minimal reverberation . Given the small dimensions of the flat we can assume that the reverberation time stays below 500ms. Therefore, the reverberation problem will not be addressed in this paper, but this needs to be taken into account for a final system.

2.3 Background noise

The biggest obstacle to the development of distant speech based applications in domestic environment is probably the wide variety of sound events and background noise that can alter or hide the useful speech signal. For instance, to operate in real homes, an ASR system must deal with competing noise from televisions or radio, vacuum cleaners, door slamming etc., making real domestic environments characterized by highly dynamic background noise. In recent years, the research community showed an increased interest in the analysis of acoustic signals in noisy conditions and organized several challenges to deal with these extreme but realistic acoustic situations specifically for speech enhancement such as the *CHiME* challenge (Barker et al, 2013, 2015) or for acoustic events or background noise recognition such as the *D-case* challenges (Stowell et al, 2015). That means aiming at discovering, learning and detecting the hidden structure of acoustic events in these complex and seemingly unpredictable signals.

Although these challenges address very important issues to reach a generic solution, there are cases in which the state-of-the-art approaches might be satisfactory. For instance, for noise-robust ASR systems, some noise can be filtered out, or the combination of noise and speech sources can be directly modeled so as to separate them. In practice, when the noise source perturbing the signal of

interest is known, various noise removal techniques can be employed (Michaut and Bellanger, 2005). It is then possible to dedicate a microphone to record the noise source and to estimate the impulse response of the room acoustic in order to cancel the noise (Valin, 2006). This impulse response can be estimated through Least Mean Square or Recursive Least Square methods. In a previous experiment in a real smart home, these methods showed promising results when the noise was composed of speech or classical music (Vacher et al, 2012). However, in case of unknown noise sources, such as washing machine or blender noise, Blind Source Separation (BSS) techniques seem more suited. However, as showed by the ChiME challenge, noise separation in real smart home conditions remains an open challenge.

In this paper, the intended application is to provide a voice controlled multiroom smart home as an assistive technology, mainly towards the elderly population. For this reason, the paper focuses on achieving robust real-time ASR, using efficient adaptive VAD and multisource ASR, rather than using sound source localization and separation. Also, the particular acoustic background noise of the home is taken into account in the acoustic modeling by including speech uttered in the same conditions as in the test conditions. Learning or adapting acoustic models to particular acoustic environment has proven to be effective when the home is known *a priori* (Ravanelli and Omologo, 2015). The source separation stage might be added to the system once a solution would have reached adequate performance in term of accuracy, computing performance (real-time constraint), cost (array of microphone, computing poser) and resilience (in case of a broken or moved microphone).

The *CHiME-3* (Barker et al, 2015) challenge propose to use the WSJ 5k task to evaluate multi-microphone ASR in noisy settings with close microphone (40cm). This challenge has highlighted the importance of carefully engineered multi-channel enhancement. But the best systems required complex multi-pass strategies that may not be practical in real applications.

2.4 Multisource ASR

To enhance the speech signal, localize or separate sources, multisource audio processing has become the major focus of most research directions. In sound source enhancement, acoustic beamforming is usually performed to enhance signal in specific directions and to diminish it in others (Brandstein and Ward, 2001). Its limited complexity makes real time applications possible, it is still subject to many localization errors and is highly dependent on array of microphones. Another approach is to model acoustic sources using source localization. This second approach has proven to be more efficient than the first one but for a higher computing complexity (Thiemann and Vincent, 2013).

Probably, the less complex and close to real-time approaches to deal with multisource acoustic processing is to perform source selection or parallel decoding. In (Lecouteux et al, 2011), several methods for multisource ASR were compared, showing that fusion of decoding graphs from the sources with higher

SNR is more promising than late fusion (ROVER) and early fusion (beamforming). However, this study was not considering state-of-the-art techniques for ASR. In (Matos et al, 2014), a multisource technique is used by selecting the best channel in a multi-room scenario based on envelope-variance measure to reach acceptable performances.

Although most of the promising techniques to reach human-like or superior performances are based on array of microphones, many applications in real home will need to rely on distant distributed microphones with minimal *a priori* information about placement in the rooms. This is why this paper focuses on channel selection.

2.5 Finding resources

Another challenge in multi-room multi-source distant ASR is to find a relevant amount of data to train the models. Since most approaches are based on probabilistic modeling from data, finding a sufficient amount of speech material becomes essential. This is even more true since the emergence of "deep learning". Moreover, such technology must be validated in real smart homes and with potential users. At the time of writing, studies in such realistic conditions are rare (Vacher et al, 2015a), since they are very costly and time consuming. Thus, this is another reason why corpus collection is an important issue in this domain. Many approaches in the literature were tested in simulated or artificially mixed data which do not permit to evaluate the same kind of situations (controlled evaluation vs. realistic uncontrolled challenges). However, collecting real data is much more expensive than simulated one. This explains the low amount of realistic datasets in the community. In (Fleury et al, 2013), the authors report that the collection and annotation of a thirty-tree-hour corpus involving 21 participants in a smart home costs approximately 70k€.

In the last decade, some data collection efforts have been made to make this kinds of resources available. For instance, in *CHiME 2016* [3], 1600 noisy utterances from 4 speakers reading in different environments (Bus, street...) recorded on 6 channels on one close talking microphone. The DIRHA project (Ravanelli et al, 2015) made available a corpus including 24 English speakers recorded in a domestic environment equipped with a large number of microphones and microphone arrays in which speaker uttered different sets of phonetically rich sentences, newspaper articles, conversational speech, keywords, and commands.

In many available corpora, the main focus is on typical English speakers (note though that DIRHA includes Italian, German, Greek and Portuguese) and they do not contain any atypical users such as elderly speaker. This is why we will use the Sweet-Home corpus (Vacher et al, 2014) which has been acquired in a 4-room smart home, including typical and atypical users uttering sentences for voice controlled home automation.

[3] `http://spandh.dcs.shef.ac.uk/chime_challenge/`

3 Methods

3.1 Application

This study is being done in the context of voice command for home automation or call for help by a person living alone. The smart home is fit with one or two microphones in the ceiling of each room. The distance between each microphone, its nearest neighbour and the speaker is greater than 1 meter making the speech processing in *distant speech conditions*.

3.1.1 DOMUS smart home

The DOMUS smart home, build by the LIG laboratory, has been used in the study. This 35 m^2 flat is shown Figure 1. DOMUS is fully functional and equipped with sensors, such as energy and water consumption, temperature, hygrometer. Actuators are able to control lighting, shutters, multimedia diffusion (distributed in the kitchen, the bedroom, the office and the bathroom). A independent control room permits to observe experiments in real-time (with cameras) and to collect sensors and actuator data. This flat also contains 7 radio microphones set into the ceiling that can be recorded in real-time thanks to a dedicated software that records simultaneously the audio channels. As displayed on Figure 1, two microphones are set up in each room (only one in the bathroom) and the distance between each microphone is at least 1 meter.

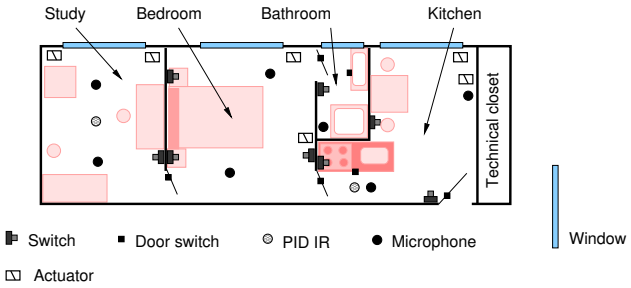

Fig. 1: DOMUS Smart Home.

3.2 Scenarios and records used for test

An experiment was conducted with users interacting with the Sweet-Home system to evaluate the accuracy of a voice command system (Chahuara et al, 2017) in realistic conditions. The possible voice commands were defined using a dedicated simple grammar described in section 4.2.1. Three categories of commands were defined: initiate command, stop command and emergency call.

Except for the emergency call, every command started with a unique keyword that permits to know whether the person is talking to the smart home or not. The grammar was built after a user study was done that showed that targeted users would prefer precise short sentences over more natural long sentences (Portet et al, 2013). Each participant had to use voice commands 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 participants to repeat the command up to 3 times in case of failure. After 3 times, a wizard of Oz technique was used to make the correct decision.

As shown on Figure 2, the SWEET-HOME system performed real-time voice command recognition anywhere in the home thanks to the PATSH software (Vacher et al, 2015a). After recognition of the command, the Intelligent Controller interpreted the available information to make decision about which command should be sent to the home automation system (Chahuara et al, 2017). The audio streams were continuously recorded during each experiment for further analysis.

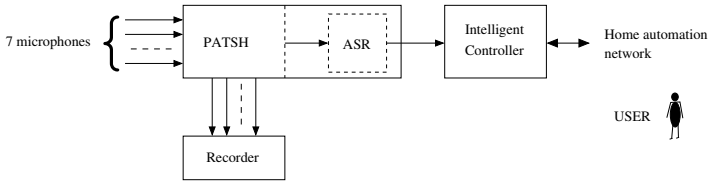

Fig. 2: SWEET-HOME system and records.

The experiments consisted in following a scenario of activities without constraint about the duration or the way of performing the activities: (1) Sleeping; (2) Resting: listening to the radio; (3) Feeding: preparing and having a meal; and (4) Communicating: having a talk with a relative thanks to the specialised communication device *e-lio* of the *Technosens*[4] company. Before the experiment, a visit was organized so that the participants find all the items necessary to perform the activities. Many decisions were to be made by the decision module such as answering commands related to giving the time or closing the blinds. Moreover, two situations related to forgetting to close a window or the front door where included in the scenario. Each time these situation were recognised, a warning message was generated thanks to a speech synthesizer. Therefore, this experiment allowed us to process realistic and representative audio events in conditions which are directly linked to usual daily living activities. Speech was transcribed manually using Transcriber (Barras et al, 2001).

[4] http://www.technosens.fr/

3.3 Experimental conditions

We assumed that during the experiment the participant was alone in the smart home, but the speech synthesizer part of the home automation system was operating and transmitted messages in case of risky situations (e.g., door no locked when the person is going to bed) or when the person asked it (about the time, etc.). The distance from the participant to the closest microphone was more than one meter and the person never spoke in the axis of the microphone because microphones were directed to the floor. Recording were on the 7 channels simultaneously, the Signal to Noise Ratio (SNR) being processed for each identified sentence on each channel.

3.4 Global architecture of the analysis system

The architecture of the audio analysis system is presented on Figure 3. Several audio event sources (in our case 7) are processed to estimate the SNR of each speech events. Depending on the applied method, the sources are either, fused at the signal level, fused at the ASR decoding level or fuse after the ASR has been run on several concurrent speech signals (*a posteriori* fusion). The last processing step is related to the identification of the voice command.

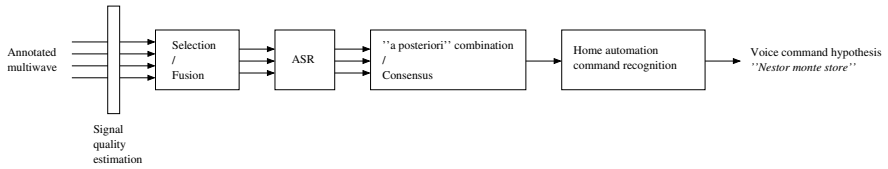

Fig. 3: Global architecture of the audio analysis system.

3.4.1 Beamforming

At the acoustic level, it may be interesting to fuse the different channels in order to enhance the signal. However, a simple sum of signals would result in a worse single channel with echoes. That is why a beamforming algorithm (Anguera et al, 2007) was used to merge all channels in a single one to feed an ASR system. Beamforming involves low computational cost and combines efficiently acoustic streams to build an enhanced acoustic signal.

The acoustic beamforming algorithm is based on the *weighted & sum microphone array* theory. Given M microphones, the signal output $y[t]$ is computed by

$$y[t] = \sum_{m=1}^{M} W_m[t]\, x_m\Big[t - D^{(m,ref)}[t]\Big] \tag{1}$$

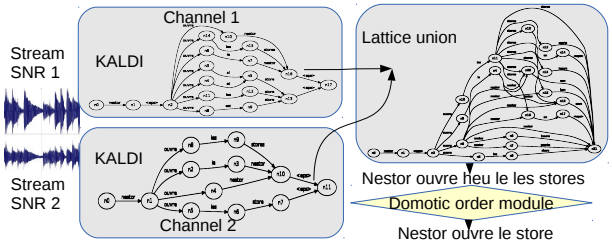

Fig. 4: **Multi-channel fusion**: voice commands are recognized from the union of the two streams lattices: "*Nestor open the blind*".

where $W_m[t]$ is the weight for microphone m at time t, $x_m[t]$ is the signal of the m^{th} channel and $D^{(m,ref)}[t]$ is the delay between the m^{th} channel and the reference channel. The weights $W_m[t]$ must satisfy $\sum_{m=1}^{M} W_m[t] = 1$. In our experiments, the reference channel was the one with the highest SNR overall and the 7 signals were entirely combined for each speaker rather than doing a sentences based combination. Once the new signal y is computed, it can feed a monosource ASR stage.

3.4.2 Fusion of lattices

Previously, we presented at the decoding level, a novel version of the Driven Decoding Algorithm allowing to guide a channel by another one (Lecouteux et al, 2013). In this work, we propose to combine channels using the FST framework. This multi-channel system is showed in Figure 4. After the decoding, the channel lattices are combined using Minimum Bayes Risk decoding as proposed in (Xu et al, 2011). The relative contribution of individual lattices is weighted according the SNR (70% for the best channel: log of the weight is subtracted from the total backward score). This method allows one to merge the information from the two streams at graph level. The applied strategy used a dynamic selection by using the two best channels for each utterance to decode (i.e. having the highest SNR).

3.4.3 Rover

ROVER can be used to combine ASR results obtained from each channel (Fiscus, 1997); it is expected to improve the recognition results by providing the best agreement between the most reliable sources. It combines systems output into a single word transition network. Then, each branching point is evaluated with a vote scheme. The word with the best score is selected (number of votes weighted by confidence measures). This approach necessitates high computational resources when several sources need to be combined and real time is needed (in our case, 7 ASR systems must operate concurrently).

A baseline ROVER is using all available channels without *a priori* knowledge. In a second time, an *a priori* confidence measure based on SNR can be

used: for each decoded segment s_i from the i^{th} ASR system, the associated confidence score $\phi(s_i)$ is computed by

$$\phi(s_i) = 2^{R(s_i)} / \sum_{j=1}^{7} 2^{R(s_j)} \tag{2}$$

where $R()$ is the function computing the SNR of a segment and s_i is the segment generated by the i^{th} ASR system. The SNR is evaluated as:

$$R(S) = 10.log(\frac{\sum_{n \in I_{speech}} S[n]^2}{|I_{speech}|} / \frac{\sum_{n \in I_{sil}} S[n]^2}{|I_{sil}|}) \tag{3}$$

Given that channels with lowest SNR contain few and redundant information, it is possible to reach satisfactory results with reasonable computational cost thanks to a ROVER using only the three best SNR channels.

3.4.4 Voice command detection

We propose to transcribe each voice command and ASR output into a phoneme graph in which each path corresponds to a variant of pronunciation. For each phonetized ASR output T, every voice commands H is aligned to T using Levenshtein distance. The deletion, insertion and substitution costs were computed empirically while the cumulative distance $\gamma(i, j)$ between H_j and T_i is given by Equation 4.

$$\gamma(i, j) = d(T_i, H_j) + min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \tag{4}$$

The distance function $d()$ is biased according to the likelihood of phoneme confusion.

The voice command with the aligned symbols score is then selected for decision according to a detection threshold. This approach takes into account some recognition errors such as word endings or light variations. Moreover, in a lot of cases, a miss-decoded word is phonetically close to the good one (due to the close pronunciation).

4 Experiments

The experiments presented here are based on our previous study which included beamforming, ROVER and a old version of the Driven Decoding algorithm (Lecouteux et al, 2011) on a different dataset. We re-run the experiment using two new corpora: Interaction and User specific. Moreover, another major difference apart from the more adequate corpora, is that we focus the experiment on state-of-the-art and competing models: Subspace Gaussian Mixture Model (SGMM) and Deep Neural Networks (DNNs). Corpora are presented at the beginning of this section before model generation and evaluation metrics.

Subset	Duration	Number of files
Multimodal	2607	1785
Home Automation Speech	10845	5340
Cirdo set	945	414
Voix Détresse	1127	1164
All	15524 (4h 18mn 44s)	8703

Table 1: Size of the different parts of the training corpus

4.1 Corpora

The used corpora are extracted from different subsets recorded in the DOMUS smart home described in Section 3.1.1. Two corpora are mono channel, *Cirdo* and *Voix Détresse*, they were chosen because they are made of expressive speech and then more representative of speech uttered in realistic conditions in a smart home. Regarding the other one, they are made of read speech and audio records are available for all channels, they are manually annotated thanks to Transcriber software (Barras et al, 2001) and the SNR was calculated for each channel for the purpose of selecting the 2 best channels.

The recording way of each of them and their composition are described in the appendix A whereas their repartition in the training, development and test parts are described in the following sections 4.1.1 and 4.1.2.

4.1.1 Training subset

For training, we used 4h 18mn 44s of data, Table 1 resumes their principal characteristics. They are extracted from the following corpora:

1. the speech part of the *Multimodal* subset of the SWEET-HOME corpus (Vacher et al, 2014) (see Table 5a), sentences are read by the participants when they operate an Activity of Daily Living (bathing, dressing, eating and preparing a meal);
2. the non noisy part of the *Home Automation Speech* subset of the SWEET-HOME corpus (Vacher et al, 2014) (see Table 6a), sentences are read by the participant in each room, sentences are following the grammar necessary for activating the intelligent controller;
3. the *Cirdo set* corpus (Vacher et al, 2016) (see Table 5b), which is made of call for uttered by people when they fell on the carpet or when they sited on the sofa and can not go up due to a blocking hip;
4. and the *Voix Détresse* corpus (Aman et al, 2016) (see Table 6b), which is made of neutral and expressive sentences.

Unlike the other sets, the two lasts are not read and are made of expressive speech because the participants were calling for help in a distress situation. All corpora except the last one were recorded in distant speech condition.

Subset	Duration (seconde)	Number of files
Interaction	21 mn 28s	803
User Specific	17mn 48s	549

Table 2: Size of the different parts of the development/test corpus

4.1.2 Developing and testing subset

For testing, we used the *Interaction* and *User Specific* subsets of the Sweet-Home corpus (Vacher et al, 2014) recorded in realistic conditions. They are described Table 2. During the recording of these datasets, typical participants (for the first one) and elderly/visually impaired participants (for the second one) had to use voice commands to interact with the home automation system (cf. Sec. 3.2). These two corpora were recorded in distant speech condition on 7 channels. More details about these corpora are given Tables 7 and 8.

4.2 Automatic Speech Recognition System

The Kaldi speech recognition toolkit (Povey et al, 2011b) was chosen as unique ASR system. Kaldi is an open-source state-of-the-art ASR system with a high number of tools and a strong support from the community. This choice was made based on experiments we undertook with several state-of-the-art ASR systems and on the fact that DDA can be easily implemented in it.

4.2.1 Grammar and Language Models

In Sweet-Home, the actions the intelligent controller could make were the following:

- turn on/off the light, radio
- close/open the blinds, curtains
- give the temperature, time
- warn about open windows, unlocked door
- command the e-lio system to call a specific number or to send out an emergency call.

These actions constitute a subset of a larger set of possible actions resulting from a previous user study (Portet et al, 2013). Of course, this set of actions must be adapted in the future to every user and home, but this predefined list was useful for the evaluation of the system.

Possible voice commands were defined using a very simple grammar as shown on Figure 5. Each command 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 that permits to know whether the person is talking to the smart home or not. In the following, we will use *'Nestor'* as keyword:

```
basicCmd         = key initiateCommand object |
                   key stopCommand [object] |
                   key emergencyCommand
key              = "Nestor" | "maison"
stopCommand      = "stop" | "arrête"
initiateCommand  = "ouvre" | "ferme" | "baisse" | "éteins" | "monte" |
                   "allume" | "descend" | "appelle" " | "donne"
emergencyCommand = "au secours" | "à l'aide"
object           = [determiner] ( device | person | organisation)
determiner       = "mon" | "ma" | "l'" | "le" | "la" | "les" | "un" | "des" |
                    du"
device           =  "lumière" | "store" | "rideau" | "télé" | "télévision" |
                    "radio" | "heure" | "température"
person           = "fille" | "fils" | "femme" | "mari" | "infirmière" |
                    "médecin" | "docteur"
organisation     = "samu" | "secours" | "pompiers" | "supérette" | "supermarché"
```

Fig. 5: Excerpt of the grammar of the voice command (terminal symbols are in French)

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. `Nestor au secours`)

A 3-gram Language Model (LM) with a 10K words lexicon was used. It results from the interpolation of a generic LM (weight 10%) and a domain LM (weight 90%). The generic LM was estimated on about 1000M of words from the French newspapers Le Monde and Gigaword. The domain LM was trained on the sentences generated using the grammar of the application (see Figure 5). The LM combination biases the decoding towards the domain LM, but still allows decoding of out-of-domain sentences. A probabilistic model was preferred over using strictly the grammar because it makes it possible to use uncertain hypotheses in a fusion process for more robustness.

4.2.2 Acoustic Model adaptation : fMLLR+SAT baseline system

Acoustic modeling was implemented with the Kaldi framework (Povey et al, 2011b). GMMs were trained on 40 dimensional MFCC (Mel Frequency Cepstral Coefficients) feature vectors (including first and second delta components and energy). Cepstral mean and variance normalisation (CMVN) were also performed. The position-independant triphone GMM trained with Kaldi consisted in 15.000 states Hidden Markov Models (HMMs) with a total of 150.000 Gaussians and 3-state phone-silence model and the number of phones was 40.

The parameters of the acoustic model were estimated via Viterbi training by aligning the audio to the reference transcript with the most current acoustic model. The models were trained on features, spliced across 3 frames before and 3 frames after and processed with linear discriminant analysis and maximum likelihood linear transformation. Speaker adaptive training was also performed

by adapting to each specific speaker with a particular data transform. Features were then adapted with feature-space MLLR in both training and test time.

4.2.3 Acoustic Model adaptation : Subspace GMM Acoustic Modelling

The GMM and Subspace GMM (SGMM) both model emission probability of each HMM state with a Gaussian mixture model, but in the SGMM approach, the Gaussian means and the mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections.

The SGMM model (Povey et al, 2011a) is described in the following equations:

$$\begin{cases} p(\mathbf{x}|j) = \sum\limits_{m=1}^{M_j} c_{jm} \sum\limits_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i), \\ \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \\ w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}, \end{cases}$$

where $\mathbf{x}$ denotes the feature vector, $j \in \{1..J\}$ is the HMM state, i is the Gaussian index, m is the substate and c_{jm} is the substate weight. Each state j is associated to a vector $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S is the phonetic subspace dimension) which derives the means, μ_{jmi} and mixture weights, w_{jmi} and it has a shared number of Gaussians, I. The phonetic subspace $\mathbf{M}_i$, weight projections $\mathbf{w}_i^T$ and covariance matrices $\mathbf{\Sigma}_i$, i.e. the globally shared parameters $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$, are common across all states. These parameters can be shared and estimated over multiple record conditions.

A generic mixture of I gaussians, denoted as Universal Background Model (UBM), models all the speech training data for the initialisation of the SGMM.

Our experiments aims at obtaining SGMM shared parameters using both Sweet-Home data (7h) and clean data (ESTER+REPERE 500h). Regarding the GMM part, the three training datasets were merged in a single one. Povey et al (2011a) showed that the model is also effective with large amounts of training data. Therefore, two UBMs were trained respectively on Sweet-Home data and clean data. These two UBMs contained 1K gaussians and were merged into a single one mixed down to 1K gaussians (closest Gaussians pairs were merged (Zouari and Chollet, 2006)). The aim was to bias specifically the acoustic model with the smart home and expressive speech conditions.

4.2.4 Acoustic Model adaptation : DNN

In a DNN-HMM hybrid system, the Deep Neural Network (DNN) is trained to provide posterior probability estimates for the HMM states. For an observation corresponding to time t in utterance, the output of the DNN for the HMM state is obtained using the softmax activation function. The networks are trained to optimize a given training objective function using the standard error back-propagation procedure (Rumelhart et al, 1986). In our experiments cross-entropy is used as the objective and the optimization is done through

stochastic gradient descent. For any given objective, the important quantity to calculate is its gradient with respect to the activations at the output layer. The gradients for all the parameters of the network can be derived from this one quantity based on the propagation procedure.

DNN for ASR is a feed-forward neural network with hidden layers. Optimizing hidden layers can be done by pretraining using Restricted Boltzmann Machines (RBM). The generative pretraining strategy builds stacks of RBMs corresponding to the number of desired hidden layers and provides better starting point (weights) for DNN fine-tuning through backpropagation algorithm. Pretraining a DNN can be carried out in a unsupervised manner because it does not involve specific knowledge. Only the softmax layer is sensitive to the target data. It is added on top of the hidden layers during fine-tuning and its output corresponds to the HMM states. Finally, we built specific DNN for acoustic environment by fine-tuning the hidden layers from clean data on sweethome training data. We use a DNN system in order to adapt speaker features from the GMM system, (after a first pass of GMM decoding and adaptation). The 40-dimensional features from GMM are spliced across 4 frames of context before and 4 frames of context after and used as input to the DNN. The DNNs are trained on the same LDA+FMLLR features as the GMM-HMM baselines, except that the features are globally normalized to have zero mean and unit variance. The fMLLR transforms are the same as those estimated for the GMM-HMM system during training and testing. We use a p-norm DNN (Zhang et al, 2014) with 4 hidden layers and p-norm (input, output) dimensions of (4000, 400) respectively. We use 8000 sub-classes, and the number of parameters is 15.5 million. It is trained for 12 epochs with learning rate varying from 0.02 to 0.004 (the optimization terminates when the frame accuracy increases by less than 0.1%). The frames are presented in a randomized order while training both of these networks using Stochastic Gradient Descent (SGD) to minimize the cross-entropy between the labels and network output. We use minibatches of 128 frames.

4.3 Evaluation metrics

In our application framework, ASR performances can't be the unique criterion for system evaluation. Correct recognition of voice commands and short processing time are very important too, therefore these 3 metrics are considered.

4.3.1 Word Error Rate

Performance of the ASR system is evaluated through the Word Error Rate (WER), which is a common evaluation metric and analyzed in McCowan et al (2005):

$$WER = \frac{S + D + I}{S + D + C} \tag{5}$$

where S is the number of substitutions, D the number of deletions, I the number of insertions and C the number of the corrects.

4.3.2 voice command recognition and distress detection

It is important to recognize voice commands or distress calls and to not miss any of them. We define the DER (Domotic Error Rate i.e. home automation error rate) as:

$$\text{DER} = \frac{\text{Missed} + \text{False Alarms}}{\text{Voice Commands}_{\text{syntactically correct}}} \tag{6}$$

For the DER, the ground truth is the number of uttered voice commands respecting the grammar; i.e. the utterances where the person's intention was to utter a command but was not following the voice command syntax were not considered as true voice commands. The "Missed" correspond to the true voice commands not recognized and the "False Alarms" to sound events incorrectly classified as voice commands. This metric is inspired by information retrieval metrics as presented in McCowan et al (2005).

4.3.3 Decoding time

The Decoding Time Rate (DTR) is simply evaluated as the ratio of the decoding time of the entire corpus to the corpus duration:

$$DTR = \frac{\text{Decoding time}}{\text{Corpus duration}} \tag{7}$$

5 Results

Results on manually annotated data are given Table **??**. The most important performance measures are the WER of the overall decoded speech and those of the specific voice commands as well as the DER. Most of the improvement is due to fMLLR and adapted data to the acoustic environment. However, adaptation techniques based on SGMM and DNN significantly improve the WER.

For the WER measure SGMMs generate better results than DNN. SGMM has a relatively small amount of parameters tied to the acoustic state, with many of the model parameters being globally shared. This make it possible to train models on less in-domain data (in our case, SWEET-HOME data quantity is very low) than would otherwise be necessary for heavy data consuming approaches such as DNN. By contrast, if DER is observed, the DNN models are slightly better.

All proposed SNR-based approaches benefited from the multiple available microphones. Beamforming shows a little improvement. The lattice fusion method showed the best improvement by using the SNR with a very high stability. Finally, the SNR-based ROVER obtains results similar to the Beamforming approach.

Regardless of the models used, the fusion of two channels improves the initial results. This means that the information provided by the different channels

	GMM-HMM + fMLLR+SAT		SGMM-HMM		DNN-HMM	
WER/DER	DEV	TEST	DEV	TEST	DEV	TEST
WER Interaction SNR1	35.00	30.95	30.65	**27.86**	29.78	29.11
WER Interaction SNR2	30.65	34.21	26.52	31.75	27.83	**31.31**
WER Interaction Beamforming	29.60	30.10	26.20	29.12	27.83	**28.49**
WER Interaction ROVER	28.70	30.10	26.33	**28.22**	27.83	28.49
WER Interaction SNR1&2	27.61	29.83	26.04	**27.22**	26.83	27.49
WER User specific SNR1	28.16	39.79	29.99	**34.88**	28.55	37.95
WER User specific SNR2	36.77	39.96	38.98	**38.34**	35.98	39.01
WER User specific Beamiforming	27.95	38.63	30.10	**35.42**	26.01	37.31
WER User specific ROVER	27.57	38.57	28.90	**35.00**	28.21	37.51
WER User specific SNR1&2	27.47	37.56	28.94	**34.38**	27.81	37.11
DER Interaction SNR1	6.98	4.75	5.43	**3.86**	6.20	5.93
DER Interaction SNR2	6.20	5.93	3.10	5.79	5.43	**5.49**
DER Interaction Beamforming	6.20	5.00	3.10	**4.56**	5.43	5.64
DER Interaction ROVER	5.43	5.00	3.30	**3.86**	5.43	5.64
DER Interaction SNR1&2	3.88	4.15	2.65	**3.86**	3.88	5.03
DER User specific SNR1	2.19	2.19	1.09	**1.91**	1.09	**1.91**
DER User specific SNR2	4.92	4.10	3.83	3.28	3.28	**2.46**
DER User specific Beamforming	2.00	2.10	1.29	1.91	2.00	**1.78**
DER User specific ROVER	1.8	2.19	1.09	1.91	1.75	**1.78**
DER User specific SNR1&2	1.64	1.37	1.09	1.73	1.09	**1.37**

Table 3: DER and WER results on the Interaction and User Specific corpora using different channel combinations and acoustic models. We report the results for "interaction" and "user specific" corpora using three main acoustic models : feature MLLR with speaker adaptation training (fMLLR+SAT), SGMM and(DNN). We compare 5 decoding methods: SNR1 means using only the best SNR channel, SNR2 is using only the second best SNR channel, beamforming is based on 7 channels, ROVER is based on 7 channels and SNR1&2 is the combination at the graph level between SNR1 and SNR2.

Table 4: DTR for the Interaction corpus

GMM-HMM + fMLLR+SAT	SGMM-HMM	DNN-HMM
1.86	4.1	2.8

are complementary. Finally, using DNN and channel fusion, a DER of 1.37% for User specific and 5% for interaction corpora is obtained. The WER is high on the speech portions that do not correspond to home automation commands but almost perfect for home automation orders.

Results regarding Decoding Time Rate are given in Table 4 for the Interaction corpus. Although the best decoding time is reasonable, it is still too long for use in a real application in home automation. For example, if duration is 1s, the system could not operate before 1.86s after the end of the order in the case of fMLLR+SAT, that may not be acceptable by the user.

6 Conclusion

In this paper, the multichannel ASR part of a voice command system in a smart home is presented. Since voice based smart homes is perturbed by the distant speech condition, an overview of multiple techniques for fusion of multi-source audio signal for automatic speech recognition is presented and evaluated on two corpora collected in a real smart home with typical, senior and visually impaired participants enacting activities of daily life. The corpora were recorded in realistic conditions, meaning background noise is sporadic so there is not an extensive background noise in the data. The smart home is equipped with two microphones in each room, the distance between each of them and with the user is larger than 1 meter.

Three state-of-the-art methods were implemented to fuse speech events from different channels. The proposed approaches were acting at the three main levels of the ASR task: acoustic, decoding and hypothesis selection. They were: beamforming (earl fusion), ASR lattice fusion (middle fusion) and ROVER (late fusion). Regarding the ASR, three acoustic models with model adaptation were used: classical HMM-GMM with fMLLR+SAT, Subspace GMM (SGMM) and DNN.

The results of the fusion techniques do not reveal a definite superiority for any of them. Beamforming and ROVER are competitive but when ASR lattice fusion is the best (which is frequent) it is by a larger gap than the two others. Beamforming improved the WER, however its performance was very close to the baseline one. ROVER also improved the WER but never beat the ASR lattice fusion. This may be due to the fact that the seven microphones are too far apart from each other to contain enough redundancy for an enhanced acoustic signal or ASR hypotheses. The lattice fusion gave the best performance with only two channels, while ROVER (using 7 ASR systems) perform similar results. Moreover, since the ASR lattice fusion only uses 2 channels against 7 for the two other methods, it can be concluded that ASR lattice fusion is the most adequate method for the task.

Regarding the acoustic models of the ASR, their robustness is achieved by adaptation to the environment and the task. This adaptation is performed at the learning level by including corpora recorded in the same conditions as the evaluation corpora as well as using model adaptation techniques during the decoding. Although the overall WER is between 26% and 40%, DER is always less than 6%. This confirms the interest of using of the Levenshtein distance at the phonetic level. The fMLLR+SAT model never gives the best WER and DER. The SGMM models gives the best WER for almost all the conditions and the best DER for the Interaction corpus. The DNN model has the best DER for the User Specific corpus. Thus, although DNN models have brought a substantial performance improvement in speech processing and other fields, SGMM are still competitive in case of a low amount of training data. However, DNN models were far quicker in processing data than the SGMM ones.

These results obtained in realistic conditions give a fairly accurate idea of the performances that can be achieved with state-of-the-art ASR systems.

As stated above, obtained results are not sufficient to allow the system to be uses in real conditions and we plan to focus on three challenges to address. Firstly, the processing time of the ASR system must be improved. We plan to work on concurrent speech decoding and on-line decoding. On-line decoding consists in processing speech frames as they arrive to the ASR system rather when the entire signal is acquired. This make it possible to process the speech signal as soon as it is detected. Secondly, distant speech recognition should be able to be performed in noisy conditions (television, sound of water) thus future work include the use of speech enhancement. We are currently working with source separation specialists in the framework of the ANR VocADom project[5] supported by the French national government. One of the issues for this challenge is to be able to process noisy speech without impacting the processing time. On-line solutions, such as noise cancellation, can be applied when the noise source is clearly identified (Vacher et al, 2012). Thirdly, the voice command system should be able to work with multiple users, hence it is important to include a speaker recognition stage to manage command privileges. Previous work has shown the feasibility of the approach but also emphasized the challenge of speaker recognition with short signal (Vacher et al, 2015b).

Acknowledgements The authors would like to thank the participants who accepted to perform the experiments.

References

Aman F, Vacher M, Rossato S, Portet F (2013) Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences. In: The 7th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013, Cluj-Napoca, Romania, pp 177–184

Aman F, Aubergé V, Vacher M (2016) Influence of expressive speech on ASR performances: application to elderly assistance in smart home. In: Sojka P, Horak A, Kopecek I, Pala K (eds) Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Springer International Publishing, pp 522–530, DOI 10.1007/978-3-319-45510-5_60

Anguera X, Wooters C, Hernando J (2007) Acoustic beamforming for speaker diarization of meetings. IEEE Transactions on Audio, Speech, and Language Processing 15(7):2011–2022, DOI 10.1109/TASL.2007.902460

Audibert N, Aubergé V, Rilliard A (2005) The prosodic dimensions of emotion in speech: the relative weights of parameters. 9th european conference on speech communication and technology. In: Interspeech 2005, Lisbon, Portugal, pp 525–528

Baba A, Lee A, Saruwatari H, Shikano K (2002) Speech recognition by reverberation adapted acoustic model. In: ASJ General Meeting, pp 27–28

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

Baba A, Yoshizawa S, Yamada M, Lee A, Shikano K (2004) Acoustic models of the elderly for large-vocabulary continuous speech recognition. Electronics and Communications in Japan, Part 2, Vol 87, No 7, 2004 87(2):49–57

Badii A, Boudy J (2009) 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, pp 18–20

Barker J, Vincent E, Ma N, Christensen H, Green PD (2013) The PASCAL chime speech separation and recognition challenge. Computer Speech & Language 27(3):621–633

Barker J, Marxer R, Vincent E, Watanabe S (2015) The third 'chime' speech separation and recognition challenge: Dataset, task and baselines. In: Workshop on Automatic Speech Recognition and Understanding (ASRU), pp 504–511

Barras C, Geoffrois E, Wu Z, Liberman M (2001) Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication 33(1-2):5–22

Bouakaz S, Vacher M, Bobillier-Chaumon ME, Aman F, Bekkadja S, Portet F, Guillou E, Rossato S, Desserée E, Traineau P, Vimon JP, Chevalier T (2014) CIRDO: Smart companion for helping elderly to live at home for longer. IRBM 35(2):101–108

Brandstein M, Ward D (eds) (2001) Microphone Arrays : Signal Processing Techniques and Applications. Springer-Verlag Berlin Heidelberg

Caballero-Morales SO, Trujillo-Romero F (2014) Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition. Expert Systems with Applications 41(3):841–852

Chahuara P, Portet F, Vacher M (2017) Context-aware decision making under uncertainty for voice-based control of smart home. Expert Systems with Applications 75:63–79, DOI 10.1016/j.eswa.2017.01.014

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

Charalampos D, Maglogiannis I (2008) 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, pp 1:1–1:8

Christensen H, Casanuevo I, Cunningham S, Green P, Hain T (2013) homeservice: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition. In: SLPAT, pp 29–34

Cristoforetti L, Ravanelli M, Omologo M, Sosi A, Abad A, Hagmueller M, Maragos P (2014) The DIRHA simulated corpus. In: The 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, pp 2629–2634

Deng L, Acero A, Plumpe M, Huang X (2000) Large-vocabulary speech recognition under adverse acoustic environments. In: ICSLP-2000, ISCA, Beijing, China, vol 3, pp 806–809

Filho G, Moir T (2010) From science fiction to science fact: a smart-house interface using speech technology and a photorealistic avatar. International Journal of Computer Applications in Technology 39(8):32–39

Fiscus JG (1997) A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In: Proc. IEEE Workshop ASRU, pp 347–354, DOI 10.1109/ASRU.1997.659110

Fleury A, Vacher M, Portet F, Chahuara P, Noury N (2013) A French corpus of audio and multimodal interactions in a health smart home. Journal on Multimodal User Interfaces 7(1):93–109

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

Hwang Y, Shin D, Yang CY, Lee SY, Kim J, Kong B, Chung J, Kim S, Chung M (2012) Developing a voice user interface with improved usability for people with dysarthria. In: 13th International Conference on Computers Helping People with Special Needs, ICCHP'12, pp 117–124

Lecouteux B, Vacher M, Portet F (2011) Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions. In: Proc. InterSpeech, pp 2273–2276

Lecouteux B, Linares G, Estève Y, Gravier G (2013) Dynamic combination of automatic speech recognition systems by driven decoding. IEEE Transactions on Audio, Speech & Language Processing 21(6):1251–1260

Matos M, Abad A, Astudillo R, Trancoso I (2014) In: IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, pp 178–188

McCowan I, Moore D, Dines J, Gatica-Perez D, Flynn M, Wellner P, Bourlard H (2005) On the use of information retrieval measures for speech recognition evaluation. Tech. rep., Idiap

Michaut F, Bellanger M (2005) Filtrage adaptatif : théorie et algorithmes. Hermes Science Publication, Lavoisier

Mueller P, Sweeney R, Baribeau L (1984) Acoustic and morphologic study of the senescent voice. Ear, Nose, and Throat Journal 63:71–75

Ons B, Gemmeke JF, hamme HV (2014) The self-taught vocal interface. EURASIP Journal on Audio, Speech, and Music Processing 43

Parker M, Cunningham S, Enderby P, Hawley M, Green P (2006) Automatic speech recognition and training for severely dysarthric users of assistive technology: The stardust project. Clinical linguistics & phonetics 20(2-3):149–156

Peetoom KKB, Lexis MAS, Joore M, Dirksen CD, De Witte LP (2014) Literature review on monitoring technologies and their outcomes in independently living elderly people. Disability and Rehabilitation: Assistive Technology pp 1–24

Pellegrini T, Trancoso I, Hämäläinen A, Calado A, Dias MS, Braga D (2012) Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese. In: Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings, pp 139–147

Popescu M, Li Y, Skubic M, Rantz M (2008) 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, pp 4628–4631

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

Portet F, Christensen H, Rudzicz F, Alexandersson J (2015) Perspectives on Speech and Language Interaction for Daily Assistive Technology: Overall Introduction to the Special Issue Part 3. ACM - Transactions on Speech and Language Processing 7(2)

Potamianos G, Neti C (2001) Automatic speechreading of impaired speech. In: AVSP 2001-International Conference on Auditory-Visual Speech Processing

Povey D, Burget L, Agarwal M, Akyazi P, Kai F, Ghoshal A, Glembek O, Goel N, Karafiát M, Rastrow A, Rose RC, Schwarz P, Thomas S (2011a) The subspace Gaussian mixture model – A structured model for speech recognition . Computer Speech & Language 25(2):404 – 439

Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K (2011b) The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, iEEE Catalog No.: CFP11SRW-USB

Ravanelli M, Omologo M (2015) Contaminated speech training methods for robust DNN-HMM distant speech recognition. In: INTERSPEECH 2015, Dresden, Germany, pp 756–760

Ravanelli M, Cristoforetti L, Gretter R, Pellin M, Sosi A, Omologo M (2015) The dirha-english corpus and related tasks for distant-speech recognition in domestic environments. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp 275–282

Rudzicz F (2011) Acoustic transformations to improve the intelligibility of dysarthric speech. In: Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, pp 11–21

Rumelhart D, Hinton G, Williams R (1986) Learning representations by back-propagating errors. Nature vol 323 pp 533-536

Ryan W, Burk K (1974) Perceptual and acoustic correlates in the speech of males. Journal of Communication Disorders 7:181–192

Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD (2015) Detection and classification of audio scenes and events. IEEE Transactions on Multimedia 17(10):1733–1746

Takeda N, Thomas G, Ludlow C (2000) Aging effects on motor units in the human thyroarytenoid muscle. Laryngoscope 110:1018–1025

Thiemann J, Vincent E (2013) An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement. In: MLSP - 23rd IEEE International Workshop on Machine Learning for Signal Processing - 2013, Southampton, United Kingdom

Vacher M, Serignat J, Chaillol S, Istrate D, Popescu V (2006) Speech and sound use in a remote monitoring system for health care. In: P Sojka

KP I Kopecek (ed) Text Speech and Dialogue, LNCS 4188/2006, Springer Berlin/Heidelberg, vol 4188/2006, pp 711–718

Vacher M, Portet F, Fleury A, Noury N (2011) Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges. International Journal of E-Health and Medical Communications 2(1):35–54

Vacher M, Lecouteux B, Portet F (2012) Recognition of Voice Commands by Multisource ASR and Noise Cancellation in a Smart Home Environment. In: EUSIPCO (European Signal Processing Conference), Bucarest, Romania, pp 1663–1667, URL `https://hal.inria.fr/hal-00953511`

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

Vacher M, Caffiau S, Portet F, Meillon B, Roux C, Elias E, Lecouteux B, Chahuara P (2015a) 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):5:1–5:36

Vacher M, Lecouteux B, Serrano-Romero J, Ajili M, Portet F, Rossato S (2015b) Speech and Speaker Recognition for Home Automation: Preliminary Results. In: 8th International Conference Speech Technology and Human-Computer Dialogue "SpeD 2015", IEEE, Bucarest, Romania, Proceedings of the 8th International Conference Speech Technology and Human-Computer Dialogue, pp 181–190

Vacher M, Bouakaz S, Bobillier Chaumon ME, Aman F, Khan RA, Bekkadja S, Portet F, Guillou E, Rossato S, Lecouteux B (2016) The CIRDO corpus: comprehensive audio/video database of domestic falls of elderly people. In: 10th International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Portoroz, Slovenia, pp 1389–1396

Valin JM (2006) Speex: a free codec for free speech. In: Australian National Linux Conference, Dunedin, New Zealand

Vincent E, Barker J, Watanabe S, Le Roux J, Nesta F, Matassoni M (2013) The Second 'CHiME' Speech Separation and Recognition Challenge: An overview of challenge systems and outcomes. In: 2013 IEEE Automatic Speech Recognition and Understanding Workshop, Olomouc, Czech Republic, pp 162–167

Vipperla R, Renals S, Frankel J (2008) Longitudinal study of ASR performance on ageing voices. In: 9th International Conference on Speech Science and Speech Technology (InterSpeech 2008), Brisbane, Australia, pp 2550–2553

Vipperla RC, Wolters M, Georgila K, Renals S (2009) Speech input from older users in smart environments: Challenges and perspectives. In: HCI Internat.: Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments

Vlasenko B, Prylipko D, Philippou-Hübner D, Wendemuth A (2011) Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In: Proceedings of Interspeech 2011, pp 1577–1580

Vlasenko B, Prylipko D, Wendemuth A (2012) Towards robust spontaneous speech recognition with emotional speech adapted acoustic models. Proc of the KI 2012

Wölfel M, McDonough J (2009) Distant Speech Recognition. Published by Wiley

World Health Organization (2003) What are the main risk factors for disability in old age and how can disability be prevented? Available from: `http://www.euro.who.int/document/E82970.pdf`

Xu H, Povey D, Mangu L, Zhu J (2011) Minimum bayes risk decoding and system combination based on a recursion for edit distance. Computer Speech & Language 25(4):802 – 828, DOI http://dx.doi.org/10.1016/j.csl.2011.03.001, URL `http://www.sciencedirect.com/science/article/pii/S0885230811000192`

Yoshioka T, Ito N, Delcroix M, Ogawa A, Kinoshita K, Yu MFC, Fabian WJ, Espi M, Higuchi T, Araki S, Nakatani T (2015) The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In: IEEE Automatic Speech Recognition and Understanding Workshop

Zhang X, Trmal J, Povey D, Khudanpur S (2014) Improving deep neural network acoustic models using generalized maxout networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, pp 215–219

Zouari L, Chollet G (2006) Efficient gaussian mixture for speech recognition. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol 4, pp 294–297, DOI 10.1109/ICPR.2006.475

A Composition of the different corpora

All corpora were recorded in distant speech conditions with the exception of Voix Détresse. Corpora used for training are detailed in Section A.1 and those used for development and test are in section A.2. Training corpora are made of in distant speech and multichannel conditions with microphones each at a distance of about 1 to 2 meters of the nearest one, and of expressive speech. Development and testing corpora were recorded in distant speech and multichannel conditions by persons interacting with the voice command system Sweet-Home.

Each sentence was manually annotated on the best Signal-to-Noise Ratio (SNR) channel using Transcriber. Moreover, regarding the User Specific set, an automatic transcription is available, that was obtained using the Patsh software operating line during the experiments while participants interacted with the Sweet-Home system. This set is important because it would be possible, using it, to determine the performances that can be achieved using a fully automatic system in a smart home application.

A.1 Training corpora

The detailed composition of each corpus is presented in Section A.1.1 (Table 5a) for the *Multimodal* subset, in Section A.1.2 (Table 5b) for the *Cirdo Set* corpus, in section A.1.3 (Table 6a) for the *Home Automation* corpus and in Section A.1.4 (Table 6b for the *Voix Détresse* corpus).

Speaker ID	Age (year)	Sex	Nb. of files	Size (s)
M01	32	M	83	129.9
M02	22	M	73	120.8
M03	56	F	84	126.9
M04	51	M	86	139.2
M05	25	F	82	149.4
M06	23	M	84	123.8
M07	50	F	91	128.0
M08	27	F	81	121.6
M09	36	M	84	137.3
M10	24	M	82	99.6
M11	38	F	89	113.9
M12	42	M	84	107.2
M13	41	M	86	115.5
M14	23	F	84	114.4
M15	62	M	85	108.8
M16	38	M	84	110.6
M17	28	M	85	136.7
M18	46	M	106	157.8
M19	63	M	86	136.2
M20	33	M	86	114.2
M21	48	F	80	115.7
All	-	-	1785	2607.4 (43mn 27s)

(a) Composition of the MULTIMODAL subset of the SWEET-HOME corpus (7 channels, read sentences), number of files and size are related to each channel

Speaker ID	Age (year)	Sex	Nb. of files	Size (s)
C01	30	M	22	37.6
C03	24	F	16	27.5
C04	83	F	66	92.0
C05	29	M	24	35.6
C06	64	F	23	31.2
C07	61	M	23	26.0
C08	44	M	25	44.0
C09	16	M	32	38.2
C10	16	M	19	348.3
C11	52	M	12	18.8
C12	28	M	15	23.6
C13	66	M	24	50.4
C14	52	F	23	39.9
C15	23	M	20	29.8
C16	40	F	29	44.0
C17	40	F	24	33.5
C18	25	F	17	25.3
All	-	-	414	945.6 (15mn 45s)

(b) Composition of the CIRDOSET corpus (1 channel, participants calling for help when they fall on the carpet or when they can't go up from the sofa)

Table 5: *Multimodal* and *Cirdo* corpora

A.1.1 Multimodal

The *Multimodal* subset of the SWEET-HOME corpus (Vacher et al, 2014) was recorded by 21 participants (7 females and 14 males) to train models for automatic human activity recognition and location. These two types of information are crucial for context aware decision making in smart home. For instance, a voice command such as "allume la lumière" (turn

on the light) cannot be handled properly without the knowledge of the user's location. 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 talk on the phone, having a breakfast, simulating a shower, getting some sleep, cleaning up the flat using the vacuum, etc.). During the experiment, even tracks from the home automation network, audio and video sensors were captured. Speech was recorded using 7 microphones set up in the ceiling directed towards ground in the DOMUS smart home (see Figure 1).

In total, more than 26 hours of data have been acquired (audio, home automation sensors and videos). The speech part is made of a telephonic conversation at the office, that represents 1785 sentences and 43 minutes and 27 seconds of speech signal. No instruction was given to the participants about how they should speak or in which direction.

A.1.2 Cirdo

The *Cirdo* corpus (Vacher et al, 2016) was recorded by 17 participants (9 men and 8 women) with average age of 40 years old (SD 19.5). This corpus was recorded in the framework of a project aiming at the development of a system able to recognize calls for help in the home of seniors in order to provide reassurance and assistance. Among them 13 people were under 60 and worn a simulator which hampered their mobility and reduced their vision and hearing to simulate aged physical conditions. The persons of the aged group (4 participants) were between 61 and 83 years old (mean 68.5). The persons of the young group were between 16 and 52 years old.

The participants simulated five options chosen from the 28 risky situations identified (1 slip, stumble 1, 2 falls in a stationary position and a position of hip blocked on the sofa). These situations were selected because they were representative falling downs at home and because they could safely be simulated by the participants. During the scenario, the participant uttered calls for help, some of them were part of the scenario but others were spontaneous speech. In these 414 calls or sentences were isolated, this represents 15 minutes and 45 seconds of speech. Due to the recording conditions, this corpus is made of expressive speech because the participants are in disturbing situations, i.e. when they fall down on the carpet. An unique microphone was used.

The interest in use of such a corpus for training is that participants spoke in a spontaneous way with affects in the voice, even if sentences were learnt at the beginning of the experiment. Therefore it is like real condition at home compared to usual corpora.

A.1.3 Home Automation

The *Home Automation Speech* subset of the SWEET-HOME corpus (Vacher et al, 2014) was recorded by 23 speakers (9 females and 14 males) to develop robust automatic recognition of voice commands in a smart home in distant conditions. The audio channels were recorded to acquire a representative speech corpus composed of utterances of not only home automation commands and distress calls, but also colloquial sentences in clean or noisy conditions. No instruction was given to the participants about how they should speak or in which direction. Speech was recorded using 7 microphones set up in the ceiling directed towards ground in the DOMUS smart home (see Figure 1).

The home automation commands follow a more simplified grammar than one defined for the test in section 4.2.1. The non noisy part is composed, for each speaker, of a text of 285 words for acoustic adaptation (36 minutes for 351 sentences in total for the 23 speakers), and of 240 short sentences (2 hours and 30 minutes per channel in total for the 23 speakers). In clean condition, 1076 voice commands and 348 distress calls were uttered. With a total of 5340 sentences overall, this corpus is made of 3 hours and 45 seconds of speech signal.

A.1.4 Voix Détresse

The *Voix Détresse* French corpus was recorded in the DOMUS smart home in order to determine if ASR performances can be affected by expressive speech (Aman et al, 2016). Firstly,

Speaker ID	Age (year)	Sex	Nb. of files	Size (s)
P01	56	M	234	585.7
P02	33	M	239	576.2
P03	38	F	229	591.7
P04	26	M	229	571.2
P05	23	M	226	440.0
P06	28	F	224	491.2
P07	30	M	224	405.5
P08	61	M	229	597.0
P10	19	F	239	373.9
P11	64	M	247	405.9
P12	57	M	237	449.7
P13	46	F	227	402.7
P14	26	M	232	425.9
P15	45	M	241	410.6
P16	23	F	233	361.1
P17	26	M	224	380.6
P18	39	F	227	395.0
P19	26	F	236	453.9
P20	57	M	223	541.4
P21	29	M	233	565.2
P22	23	M	228	608.7
P23	22	F	245	434.0
P24	25	F	234	378.2
All	-	-	5340	10845.4 (3h 0mn 45s)

(a) Composition of the *Home Automation* subset (7 channels, participants reading a short text and voice commands in each room of the appartment), number of files and size are related to each channel

Speaker	Age	Sex	Nb. of files	Size (s)
A01	84	F	80	103.9
A02	85	F	60	62.1
A03	83	F	60	60.3
A04	67	F	60	62.8
A05	73	F	60	85
J01	31	M	69	58.0
J02	30	M	83	80.5
J03	60	F	60	56.6
J05	26	F	63	65.8
J07	26	M	84	66.1
J08	32	M	64	59.6
J09	23	F	76	68.1
J10	25	F	74	61.2
J13	29	F	97	98.7
J14	24	F	85	71.3
J15	25	M	89	67.1
All	–	–	1164	1126.8 (18mn 46s)

(b) Composition of the *Voix Détresse* corpus (1 channel, expressive and acted speech)

Table 6: *Home Automation* and *Voix Détresse* corpora

speakers had to read 20 distress sentences in a neutral manner, these sentences were extracted from the *AD80* corpus(Vacher et al, 2006). Then, elicited emotions were recorded: a photography showing a person in a distress situation was associated to each sentence, the participants were asked to stand in that individuals shoes and to utter in an expressive manner. Desired emotions were mainly negative emotions like fear, anger, sadness.

Speaker ID	Age (year)	Sex	Nb. of files	Size (s)	SNR (dB)
S01	24	M	31	40.5	17
S02	27	M	52	93.3	17
S03	19	M	46	96.5	19
S04	31	M	66	93.9	18
S05	33	M	39	69.3	12
S06	62	M	37	61.7	15
S07	58	F	62	102.9	25
S08	41	F	62	90.8	12
S09	29	F	74	109.1	20
S10	27	F	49	72.1	11
S11	46	M	45	78.8	14
S12	32	M	35	53.2	17
S13	27	M	34	44.5	14
S14	52	F	42	68.4	12
S15	55	F	94	158.1	14
S16	50	F	35	55.7	14
All	–	–	803	1288.6 (21mn 28s)	–

Table 7: Composition of the INTERACTION subset (7 channels, interaction with the home automation system through voice commands, generic participants, manual transcription), number of files and size are related to each channel

Speaker	Category	Age	Sex	Number of files	Size (second)
S01	Aged	91	F	59	153.7
S02	Visually	66	F	71	114.6
S03	Visually	49	M	53	72.2
S04	Aged	82	F	72	116.9
S05	Visually	66	M	45	111.6
S06	Aged	83	F	67	134.8
S07	Aged	74	F	58	103.2
S08	Visually	64	F	35	75.7
S09	Aged	77	F	45	81.0
S10	Visually	64	M	44	104.4
All	–	–	–	549	1068 (17mn 48s)

Table 8: Composition of the USER SPECIFIC subset (7 channels, interaction with the home automation system through voice commands, elderly and visually impaired participants, manual transcription), number of files and size are related to each channel

This corpus was recorded using a microphone by 5 elderly speakers and 11 younger speakers. It is made of 1164 neutral and expressive sentences, its duration is 18mn 45s. The interest in use of such a corpus for training is that it is made of neutral and expressive sentences. Therefore it is nearest to real record condition at home than usual corpora.

A.2 Development and testing corpora

The *Interaction* and *User Specific* subset of the SWEET-HOME corpus was recorded in realistic conditions according to the conditions described in Section 3.2, thanks to the participation

Speaker	Age	Sex	Number of files	Size (second)	RSB (dB)
S01	91	F	54	161.5	14
S02	66	F	61	134.4	14
S03	49	M	48	119.2	20
S04	82	F	68	137.6	13
S05	66	M	45	124.7	19
S06	83	F	63	184.2	25
S07	74	F	54	126.0	14
S08	64	F	35	97.2	21
S09	77	F	44	97.8	17
S10	64	M	46	165.4	18
All	–	–	518	1348.0 (22mn 28s)	–

Table 9: Automatic transcription of the User Specific subset by Patsh (2 best channels)

of 16 people (7 female and 9 male, minimal age 19 years, maximal age 62 years) for the first one. The experiment duration was 8h 52mn, and 993 sentences were recorded and annotated in the same conditions that for the Training subset. The participants were in realistic life conditions and must retrieve themselves the voice command appropriate to the situation, so they don't respect perfectly the grammar: particularly, the keyword was frequently omitted or uttered a long time before the command itself. The second one implied elderly people (5 women, minimal age 74 years, maximal age 91 years) and visually impaired people (2 women and 3 men, minimal age 49 years, maximal age 66 years), for this reason, scenarios were a little simplified.

Development and testing corpora manually transcribed are detailed in Table 7 for the Interaction corpus and in Table 8 for the User Specific one. The number of voice commands is different for each speaker because if a voice command was not correctly recognized, the requested action was not directed by the intelligent controller (light on or off, curtains up or down...) and thus the speaker often uttered the command two or three times.

Moreover, regarding the User Specific set, an automatic transcription is available, it was obtained using the Patsh software operating online during the experiments while participants interacted with the Sweet-Home system; this set is described in Table 9 but was not used in the framework of this paper.

In a nutshell, two corpora recorded in realistic conditions are available for test and development, that is 21mn 28s and 17 mn 48s of manually transcribed data and 22mn 28s of automatically transcribed data.