Multichannel Automatic Recognition of Voice Command in a Multi-Room Smart Home : an Experiment involving Seniors and Users with Visual Impairment

Michel Vacher[1], *Benjamin Lecouteux*[1], *François Portet*[1]

[1]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
41 rue Mathématiques, BP 53, 38041 Grenoble cedex9, France
Michel.Vacher@imag.fr, Benjamin.Lecouteux@imag.fr, Francois.Portet@imag.fr

Abstract

Voice command system in multi-room smart homes for assisting people in loss of autonomy in their daily activities faces several challenges, one of which being the distant condition which impacts the ASR system performance. This paper presents an approach to improve voice command recognition at the decoding level by using multiple sources and model adaptation. The method has been tested on data recorded with 11 elderly and visually impaired participants in a real smart home. The results show an error rate of 3.2% in off-line condition and of 13.2% in on-line condition.

Index Terms: Multi-Channel ASR, Real-Time Audio Analysis, Applications of speech technology for AAL

1. Introduction

Many developed countries are in a demographic transition which will bring the large amount of baby boomers from full-time workers to full-time pensioners. These persons are likely to live longer than the previous generation but societies have to deal with the rising budgetary costs of ageing (health and financial support as well as ensuring a good quality of life). One of the first wishes of this population is to live in their own home as comfortably and safely 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 [1]. 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 has reached a stage of maturity and that provides interaction using natural language so that the user does not have to learn complex computing procedures [2]. Moreover, it is well adapted to people with reduced mobility and to some emergency situations (handsfree and distant interaction).

Automatic Speech Recognition (ASR) in a domestic environment has recently gained interest in the speech processing community [3]. There is a rising number of smart home projects that considers speech recognition in their design [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. However, though VUIs are frequently employed in close domains (e.g., smart phone) there are still important challenges to overcome [14]. Indeed, the task imposes several constraints to the speech technology : 1) distant speech condition[1], 2) handsfree interaction, 3) affordable by people with limited financial means, 4) real-time, 5) respect of privacy[2]. Moreover, such technology must be validated in real situations (i.e., real smart home and users).

In this paper, we present an approach to provide voice command in a multi-room smart home for seniors and people with visual impairment. In our approach, we address the problem by using several mono-microphones set in the ceiling, selecting the "best" sources and employing a double ASR decoding and voice command matching. This approach has been chosen against noise source separation which can be highly computational expensive, are sensitive to sample synchronisation problem (which cannot be assumed with non professional devices) and are still not solved in real uncontrolled condition. Hand 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 speech relevant 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 condition. The paper is organised as follow. Section 2 presents the method set for multi-channel speech recognition in the home. Section 3, present the experimentation and the results. The results of the proposed methods are discussed in Section 4.

2. Method

Recall that the multi-source voice command recognition is to be performed in the context of a smart home which is equipped with microphones set into the ceiling (c.f. Figure 1). The audio processing task is to recognize predefined sentences that correspond either to a home automation command or to a distress call. The audio module should not process other private conversations. Once a command is recognized (e.g., "turn on the light"), it is sent to a intelligent controller [15] which manages the home automation system (e.g., light up the lamp the closest

[1]Another big challenge is the ability to work in noisy condition but this not the focus of this paper, see [3] for details

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

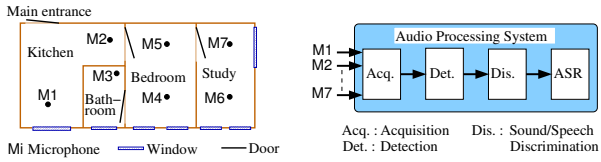

Figure 1: Microphone position in the smart home and general architecture of the PATSH audio processing system.

to the person).

The general architecture of the audio processing [16] is shown in Figure 1. Several microphones are set in the ceiling of several rooms. The microphone data are continuously acquired and sound events are detected on the fly by using a wavelet decomposition and an adaptive thresholding strategy [17]. Sound events are then classified as noise or speech and, in the latter case, sent to an ASR system. The answer of the ASR is then sent to the intelligent controller, which thanks to the context (the other information available through the home automation system) makes a context aware decision.

In this paper, we focus on the ASR system and present different strategies to improve the recognition rate of the voice commands. For the reason emphasized in the introduction, the methods do not concentrate on enhancement of the signal but on the use of *a priori* information at the language level to generate the hypothesis the most consistent with the task as well as the most relevant available channels. The remaining of this section presents the methods employed at the acoustic and decoding level.

2.1. The acoustic modeling

The Kaldi speech recognition tool-kit [18] was chosen as 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. In the experiments, the models are context-dependent classical three-state left-right HMMs. Acoustic features are based on mel-frequency cepstral coefficients, 13 MFCC-features coefficients are first extracted and then expanded with delta and double delta features and energy (40 features). Acoustic models are composed of 11,000 context-dependent states and 150,000 Gaussians. The state tying is performed using a decision tree based on automatically phonetic question. In addition, off-line fMLLR linear transformation acoustic adaptation is performed.

The acoustic models were trained on 500 hours of transcribed French speech composed of the ESTER 1&2 (broadcast news and conversational speech recorder on the radio) and REPERE (TV news and talk-shows) challenges as well as from 7 hours of transcribed French speech from 60 speakers interacting in the Smart home [19], called SH (SWEET-HOME) in the text.

2.1.1. 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 [20] is described in the following equations:

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

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 initialization of the SGMM.

Our experiments involved obtaining SGMM shared parameters using both SWEET-HOME data (7h) and clean data (500h). In the GMM system, the two training data set are just merged in a single one. [21] shows that the model is also effective with large amounts of training data. We propose to train a classical SGMM system using all the data to train the UBM (1K gaussians). In the experiments, this SGMM model is named SGMM1. In a second way, two UBM are trained respectively on SWEET-HOME data and clean data. The two obtained UBMs contain 1K gaussians and are merged into a single one mixed down to 1K gaussian (closest Gaussians pairs are merged [22]) , this SGMM is named SGMM2. The aim is to bias specifically the acoustic model with the smart home conditions.

2.2. Spoken Keyword Spotting

The problem of recognizing voice commands with a predefined grammar but not other conversation can be seen as a spoken keyword spotting problem. Given the uncertainty of the ASR system output, spoken keyword spotting has mainly been addressed by searching instances of particular keywords in the ASR set of hypotheses or lattice obtained after processing an utterance [23]. In this work, we use the method of Can and Saraçlar [24] for Spoken Term Detection (STD) from the ASR decoding lattice of an utterance. In this approach, the lattice is transformed into a deterministic weighted Finite State Transducer (FST) called Timed Factor Transducer (TFT) embedding informations for the detection (utterance ID, start and end time and posterior score). A search of a string X in a speech utterance is then a composition of the automaton representation of X and the TFT to give the resulting transducer R which contains all the possible successful detections and their posterior probability. The interest of such approach is the fact that search complexity is linear and that performing several searches in a same utterance can be done with the same TFT. Moreover, the FST formalism makes filtering with a predefined grammar easy by using the composition operator.

2.3. Multi-channel decoding

Previously, we presented at the decoding level, a novel version of the Driven Decoding Algorithm allowing to guide a channel by an other one [25]. In this work, we propose to combine channels using the FST framework. For this, the two best

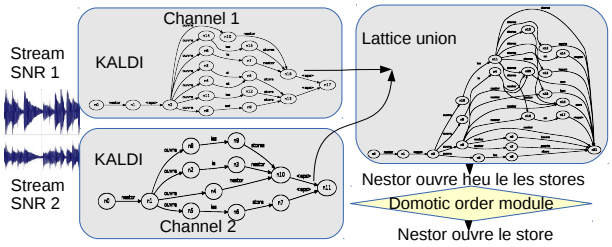

Figure 2: **Multi-channel fusion**: vocal orders are recognized from the union of the two streams lattices: "*Nestor open the blind*".

SNR channels are decoded. After the decoding the channel lattices are combined using Minimum Bayes Risk decoding as proposed in [26]. 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). The multi-channel system is showed in Figure 2.

2.4. Detection of voice commands

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 1.

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

The voice command with the aligned symbols score is then selected for decision according a detection threshold. This approach takes into account some recognition errors like 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). From this the DER (Domotic Error Rate i.e., home automation error rate) is defined as:

$$\text{DER} = \frac{\text{Missed} + \text{False Alarms}}{\text{Voice Commands}_{\text{syntactically correct}}} \quad (2)$$

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 an order but was not following the voice command syntax were not considered as a 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.

3. Experimentation and results

3.1. Live Experiment

An experiment was run in the DOMUS smart home which is part of the experimentation platform of the LIG laboratory. This is a four-room flat (see Figure 1) equipped with home automation system and with 7 microphones set in the ceiling for audio analysis. A communication device was also present to allow video-conferencing. The SWEET-HOME system consisted in the PATSH software presented in Section 2 which was continuously analysing the audio streams to detect voice commands [27] and an intelligent controller [15] in charge of executing the correct action (e.g., lighting the lamp, or giving the temperature using TTS) based on these voice commands.

Each participant, had to follow 4 successive scenarios whose topic was: 1) 'finishing breakfast and going out', 2) 'coming back from shopping and cleaning', 3) 'communicating with a relative', and 4) 'waiting for friends'. Each of these scenarios was designed to last between 5 to 10 minutes but there was no constraint on the execution time. Scenario 1 and 2 were designed to make the user performing daily activities while uttering voice commands. The participant was provided with a list of actions to perform and voice commands to utter. Each participant had to use vocal orders to turn the light on or off, open or close blinds, ask about temperature and ask to call his or her relative.

Six seniors and five people with visual impairment were recruited. The seniors (81.2 years old (SD=5.8)) were women living alone in an independent non-hospitalised accommodation. The focus of study was to target seniors who where in the edge of losing some autonomy not seniors who have lost complete autonomy. In other words, we sought seniors who were still able to make a choice regarding how the technology could help them in case of any physical degradation. The visual impaired category (62.2 (SD=6.9) years old, 3 were women) was composed of adult people living either single or as couple and whose handicap was acquired after their childhood. No upper age limit was given. They were not blind and can see but with very low visual acuity.

3.1.1. Voice orders

Possible voice orders were defined using a very simple grammar which was built after a study revealing that targeted users prefer precise short sentences over more natural long sentences [2], more details about are given in [27]. As shown below, each order belongs to either initiate command or emergency call. Every command starts with an optional key-word (e.g. *'Nestor'*) to make clear whether the person is talking to the smart home or not. Some basic commands are then `'Nestor allume lumière'` *'turn on the light'* or `'Nestor quelle est la température'` *'what is the temperature'*:

set an actuator on/off:	`key initiateCommand object` (e.g., `Nestor ferme fenêtre`) (*e.g., Nestor close the window*)
emergency call:	`key emergencyCommand` (e.g., `Nestor au secours`) (*e.g., Nestor help*)

3.1.2. Acquired data

During the experiment, audio data were recorded in two ways. Firstly, the 7-channel raw audio stream was stored for each participant for further analysis. Secondly, audio events were automatically extracted by the PATSH software on the fly. Some of the events were missed or discarded and some of the detected speech events were misclassified as everyday life sound, and some noise were misclassified as speech (bell ring, music, motor). In the later case, these non-speech events were sent to the ASR. These two speech data sets (manual vs. PATSH segmentation) were transcribed using transcriber [28]. For the PATSH data set, there are 617 uttered sentences. 291 were home automation orders (46%), 66 (10%) were actually gener-

Table 1: Recorded audio data

ID	Category	Age	Sex	Scenario duration	Speech utterances	Voice commands		SNR (dB)
						uttered	missed	
S01	Aged	91	F	24mn	59	37	30	16
S02	Visually	66	F	17mn 49s	67	26	5	14
S03	Visually	49	M	21mn 55s	53	26	9	20
S04	Aged	82	F	29mn 46s	74	27	12	13
S05	Visually	66	M	30mn 37s	47	25	10	19
S06	Aged	83	F	22mn 41s	65	31	20	25
S07	Aged	74	F	35mn 39s	55	25	13	14
S08	Visually	64	F	18mn 20s	35	22	8	21
S09	Aged	77	F	23mn 5s	46	23	17	17
S10	Visually	64	M	24mn 48s	49	20	7	18
S11	Aged	80	F	30mn 19s	79	26	18	23
All	-	-	-	4h 39mn	629	291	149	-

ated by the speech synthesizer, 10 (2%) were noise occurrences wrongly classified as speech and 250 were other spontaneous speech (42%, mostly during the video-conferencing with a relative). Only 29 speech utterances were missed (4%), but 85 of the detected ones were rejected (14%) either because their SNR was below 0dB or because they were out of the acceptable duration range (2.2 seconds). Therefore, 18% of the utterances were not treated by the system. The recorded audio data are summarized Table 1.

3.2. Off line experiments

The methods presented in Section 2 were run on the data set presented Table 1. Regarding S11, PATSH crashed in the middle of the experiment and due to time constraints S11 data was not considered in the study. Therefore 550 sentences (2559 words) including 250 orders, questions and distress calls (937 words) were used. Two acoustic models were used: AM (500h) and AM (500h+ SH, were SH = SWEET-HOME data), speaker adaptation was provided by fMLLR using the text read by each speaker. The two versions of the Subspace Gaussian Mixture Models (SGMM1 and 2 cf. Section 2.1.1) were also applied to all different combinations. For the decoding, a 2-gram language model (LM) with a 10K lexicon was used. It results from the interpolation of a generic LM and a specialized LM. The *generic* LM was estimated on about 1000M of words from the French newspapers *Le Monde* and *Gigaword*, and the broadcast news manual transcripts provided during the ESTER campaign. The *specialized* LM was estimated from the grammar and from the transcript of the 60 speakers, containing voice commands and casual speech. All the methods were run on the transcribed data (manually annotated) and on the PATSH data (automatically segmented). We present the keyword spotting approach in order to show that a conventional approach is limited because of language variations introduced by the protagonists (i.e. home automation commands are rarely pronounced correctly).

Results on manually annotated data are given Table 2. The most important performance measures are the Word Error Rate (WER) of the overall decoded speech and those of the specific voice commands as well as the Domotic Error Rate (DER: c.f. equation 2).

It can be noticed that most of the inprovement is brought by the use of fMLLR and the use of data adapted to the acoustic environment (the SH dataset). The WER obtained from the overall speech goes from 59.8 to 35.7. But most of this reduction is driven by the dramatic decrease of error in the voice command decoding. It starts from 32.9% for the baseline, and is reduced to 22.8% with the use of an acoustic model adapted to the smart home and to 14.0% thanks to speaker adaptation. Best results, WER=10.1%, DER=3.2%, are obtained by using SGMM applied to the 2 best channels (i.e., those with utterances with the highest SNR). However given the rather low number of test items results in a $\pm 1.2\%$ uncertainty interval. The improvement brought by the multi-channel decoding is below the significance level.

Table 2: WER and DER from the manually segmented data (1: best channel, 1+2: 2 best channels)

Chan.	Method	WER all (%)	WER voice commands (%)	DER (%)
1	Keyword Spotting, AM (500h+SH), SAT + fMLLR, SGMM2	-	-	57.6
1	AM (500h)	59.8	32.9	20.8
1	AM (500h), SAT + fMLLR	46.0	15.9	5.6
1	AM (500h+SH)	51.9	22.8	14.4
1	AM (500h+SH), SAT + fMLLR	39.0	14.0	4.4
1	AM (500h+SH), SAT + fMLLR, SGMM1	38.1	11.4	3.6
1	AM (500h+SH), SAT + fMLLR, SGMM2	36.1	10.9	3.2
1+2	**AM (500h+SH), SAT + fMLLR, SGMM2**	**35.7**	**10.1**	3.2

Regarding the data set extracted by PATSH. The original ASR performance with a decoding on only one channel [27] was WER=43.2%, DER=41% while the ASR using AM (500h+SH), SAT fMLLR, SGMM2 gave WER=49%, DER=13.6% on an only channel and WER=49.0%, DER=13.2% on 2 channels. The most important contribution to the DER is due to missed speech utterances at the detection or speech/sound discrimination level. Therefore this is a very significant improvement from the experimental condition.

4. Discussion

Efficient on-line recognition of voice command is mandatory for the dissemination of in-home VUI. This task must address many challenges such as distant speech recognition and respect for privacy. Moreover, such technology must be evaluated in realistic conditions. In this paper, we showed that a careful selection of the best channel as well as good adaptation to the environmental and acoustic characteristics increase dramatically the voice command classification performance. In the manual segmentation, SGMM acoustic model learned from data previously acquired in the smart home as well as fMLLR diminish the DER from 20.8% to 3.2% surpassing more standard methods such as keyword spotting. In the recognition task based on the PATSH detection and discrimination, the best technique (2-channel SGMM, fMLLR) shows a rise of DER to 13.2%. This can be explained by the imperfect detection, segmentation and classification of the system. Indeed, some sentences were missing or split in two parts (e.g., 'Nestor light the lamp' $\rightarrow$ 'Nestor' then 'light the lamp'). Hesitations in real speech are natural but it is still unclear whether they are going to be frequent in real use or due to the experimental condition (people must learn the grammar).

Despite these progresses, there are still improvements to be performed. Indeed, in the experiment, people regularly deviated from the grammar (e.g., adding politeness terms or reformulation) and did not like the predefined chosen keyword. An interesting research direction would be to adapt the language model to the words a user 'naturally' utters in different situations, hence learning the grammar from the data rather than imposing a grammar. An other direction would be to exploit the smart-home capacity of sensing the environment to provide context-aware ASR. Finally, we are in the process of releasing the data used in this experiment to support the development of in-home assistive speech technology [19].

5. Acknowledgements

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

6. References

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

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

[3] *The second CHiME Speech Separation and Recognition Challenge: Datasets, tasks and baselines*, Vancouver, Canada, 2013.

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

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

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

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

[8] 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.

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

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

[11] J. F. Gemmeke, B. Ons, N. Tessema, H. V. hamme, J. van de Loo, G. D. Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. V. D. Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces: an overview of the aladin project," in *Interspeech 2013*, 2013, pp. 2039–2043.

[12] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "homeservice: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition," in *4th Workshop on Speech and Language Processing for Assistive Technologies*, 2014.

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

[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] P. Chahuara, F. Portet, and M. Vacher, "Making Context Aware Decision from Uncertain Information in a Smart Home: A Markov Logic Network Approach," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science, vol. 8309. Dublin, Ireland: Springer, 2013, pp. 78–93. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00953262

[16] M. E. A. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, and J. Boudy, "Sound environment analysis in smart home," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science, vol. 7683. Pisa, Italy: Springer, 2012, pp. 208–223. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00840918

[17] M. Vacher, D. Istrate, and J. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees," in *Proc. 12th European Signal Processing Conference*, S. LTD, Ed., Vienna, Austria, sep. 2004, pp. 1171–1174.

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

[19] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "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, 2014, pp. 4499–4506. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00953006

[20] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model—a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011, language and speech issues in the engineering of companionable dialogue systems.

[21] ——, "The subspace gaussian mixture model—a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011, language and speech issues in the engineering of companionable dialogue systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088523081000063X

[22] L. Zouari and G. Chollet, "Efficient gaussian mixture for speech recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4, 2006, pp. 294–297.

[23] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, M. Fapso, and J. Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Interspeech'05*, Lisboa, Portugal, 2005, pp. 633–636.

[24] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2338–2347, Nov 2011.

[25] B. Lecouteux, G. Linares, Y. Estève, and G. Gravier, "Dynamic combination of automatic speech recognition systems by driven decoding," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 6, pp. 1251–1260, 2013.

[26] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802 – 828, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230811000192

[27] M. Vacher, B. Lecouteux, D. Istrate, T. Joubert, F. Portet, M. Sehili, and P. Chahuara, "Experimental Evaluation of Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home," in *4th Workshop on Speech and Language Processing for Assistive Technologies*, Grenoble, France, 2013, pp. 99–105.

[28] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, 2001.