RECOGNITION OF VOICE COMMANDS BY MULTISOURCE ASR AND NOISE CANCELLATION IN A SMART HOME ENVIRONMENT

Michel Vacher, Benjamin Lecouteux and François Portet

Laboratoire d'Informatique de Grenoble, GETALP Team
UMR CNRS/UJF/G-INP 5217,
41 rue des Mathématiques, BP 53, 38041 Grenoble, France

ABSTRACT

In this paper, we present a multisource ASR system to detect home automation orders in various everyday listening conditions in a realistic home. The system is based on a state of the art echo cancellation stage that feeds recently introduced ASR techniques. The evaluation was conducted on a realistic noisy data set acquired in a smart home where a microphone was placed near the noise source and several other microphones were placed in different rooms. This *distant speech* corpus was recorded with 23 speakers uttering colloquial or distress sentences as well as home automation orders. Techniques acting at the decoding stage and using *a priori* knowledge gave the best results in noisy condition compared to the baseline (recall= 93.2% vs 59.2%) reaching good enough performance for a real usage although improvement still need to be made when music is used as background noise.

***Index Terms*—** Home automation, smart home, distant speech, multisource ASRs, keyword detection

1. INTRODUCTION

The demographic change and ageing in the developed countries imply challenges in the way this population will be cared for in the near future. At the same time, evolution in ICT gives many opportunities to enhance in-home quality of life and to support the elderly and disabled persons in living in their own home as autonomously as possible. One of the way to bring this everyday assistance is the development of *smart homes* which are habitations equipped with a set of sensors, actuators, automated devices and centralised software which control the increasing amount of household appliances. Various interaction methods are being developed in this setting but one of the most promising is the speech interaction. Indeed, voice interfaces are much more adapted to people who have difficulties in moving or seeing than tactile interfaces (e.g., remote control) which require physical and visual interaction [1]. Moreover, voice command is particularly suited to distress situations when a person, who cannot move after a fall but being conscious, may still have the capacity to call for assistance while a remote control may be unreachable [2]. Furthermore, given the increasing complexity of home appliances, speech interfaces seem much more natural than tactile interfaces [3].

While the speech interaction is a desirable feature of smart homes, many challenges are to be addressed before transferring this technology from the lab to the home. One of the major issues is the poor performance of Automatic Speech Recognition (ASR) in a noisy environment [4]. Indeed, in realistic conditions, the performance of ASR systems decreases significantly as soon as the microphone is *'distant'* from the speaker. This deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise such as TV, radio and devices [5]. While user's linguistic preferences, dialogues and age dependant voice interfaces have been studied during this decade [3, 2, 6], speech separation from in-home noise received attention very recently within the speech processing community [7].

In this paper, we present a system to recognise vocal home automation orders (also called home automation orders) in a noisy multiroom smart home. This work is part of the SWEET-HOME project which is introduced in Section 2 along with the evaluation dataset. The approach is based on a echo canceller stage useful in the restricted case of known noise sources like TV and radio, and a multisource ASR system that uses *a priori* knowledge to enhance the in-domain home automation orders recognition. This framework is presented in Section 3. The experiments and results are then presented in Section 4 before the conclusion.

2. CONTEXT OF THE STUDY AND EVALUATION DATASET

This study was done in the context of the SWEET-HOME project (`http://sweet-home.imag.fr`) which aims at designing a new smart home system based on audio technology to provide assistance via *natural man-machine interaction* and *security reassurance* by detecting situations of distress.

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

```
basicCmd          = key initiateCommand object |
                    key stopCommand [object] |
                    key emergencyCommand
key               = "Nestor" | "maison"
stopCommand       = "stop" | "arrête"
initiateCommand   = "ouvre" | "ferme" | "baisse" | "éteins" | "monte" |
                    "allume" | "descend" | "appelle"
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"
person            = "fille" | "fils" | "femme" | "mari" | "infirmière" |
                    "médecin" | "docteur"
organisation      = "samu" | "secours" | "pompiers" | "supérette" | "supermarché"
```

Fig. 1. Excerpt of the grammar of the voice orders (terminal symbols are in French)

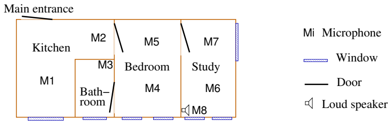

Fig. 2. Position of the microphones in the DOMUS Home

External noise	Study	Bedroom	Bathroom	Kitchen
None	30	30	30	30
Music (radio)	30	30	-	-
Speech (broadcast news)	30	30	-	-

Table 1. Sentence number as function of the room (Phase 2)

If these aims are achieved, then the person will be able to pilot, from anywhere in the house, her environment at any time in the most natural way possible.

In this study, voice orders were defined using a very simple grammar as shown on Figure 1. Our previous user study showed that targeted users prefer precise short sentences over more natural long sentences [1]. Each order belongs to one of three categories: initiate command, stop command and emergency call. Except for the emergency call, all 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 key-word:

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

In this project, the targeted environment in which speech recognition must be performed is shown in Figure 2. It is a thirty square meters suite flat set up by the MULTICOM team of the Laboratory of Informatics of Grenoble, which includes a bathroom, a kitchen, a bedroom and a study, all equipped with sensors, switches and actuators. It was complemented with seven RF microphones (set in the ceiling and directed to the floor) whose audio channels are recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card [4]. This places the study in a distant-speech context where microphones may be far from the speaker and may record different noise sources. For the sake of the experiment, an eighth microphone was set in front of a loud speaker that was used to record the noise source. For practical reason, the noise signal was recorded by a microphone and not directly forwarded to the noise canceller stage. This can lead to worse performance because the noise signal can actually contains user's utterances.

Given that, to the best of our knowledge, no dataset of French utterances of voice commands in a noisy multisource home exists, we conducted an experiment to acquire a representative speech corpus composed of utterances of not only home automation orders and distress calls, but also colloquial sentences. In order to get more realistic conditions, two types of background noise were considered while the user was speaking: the broadcast news radio and a music (classical) background. These were played in the study through two speakers. Note that this configuration poses much more challenges to classical blind source separation techniques than when speech and noise sources are artificially linearly mixed.

The protocol was composed of two phases. In **Phase 1** the participant was asked to go to the study, to close the door and to read a short text of 285 words. This data was used for the adaptation of the acoustic model of the ASR. In **Phase 2**, the participant uttered 30 sentences in different rooms in different conditions. The first condition was without noise, the second one was with the radio turned on in the study and the third one was with classical music played in the study. Table 1 summarises the conditions and locations. Each sequence of 30 sentences was composed by a random selection of 21 home automation orders (9 without initiating keyword), 2 distress calls (e.g., "À l'aide"(*help*), "Appelez un docteur" (*call a doctor*)) and 7 casual sentences (e.g., "Bonjour" (*Hello*), "J'ai bien dormi" (*I slept well*)). No participant uttered the same sequences. The radio and the music were unique and pre-recorded and were started at a random time for each participant.

23 persons (including 9 women) participated to the experiment. The average age of the participants was 35 years (19-64 min-max). No instruction was given to any participant about how they should speak or in which direction. Consequently, no participant emitted sentences directing their voice to a particular microphone. The distance between the speaker and the closest microphone was about 2 meters. The total duration of the experiment was about 5 hours

At the end of the experiment, the dataset was composed, for each speaker, of a text of 285 words for **Phase 1** (36 minutes for 351 sentences in total for the 23 speakers), and of 240 short sentences for **Phase 2** (2 hours and 30 minutes per channel in total for the 23 speakers) with a total of 5520 sentences overall, 2760 of which being instances in noisy conditions (38 minutes of Radio and 37 of music). Each sentence was humanly annotated on the best Signal-to-Noise Ratio (SNR) channel. In clean condition, 1076 voice commands and 348 distress calls were uttered while they were respectively 489 and 192 in radio background noise and 412 and 205 with music. Only the microphone data were used in this study (i.e., no video data). In the following, this corpus is called the SWEET-HOME *Home Automation Speech Corpus*.

3. PROPOSED APPROACH FOR ROBUST ASR

To detect voice commands in the SWEET-HOME context, we propose a three-stage approach. The first one detects speech activities in the audio streams, the second one extracts the best utterance hypotheses using an ASR system and the last one recognises a vocal command or a distress situation from the decoded utterances. This paper describes the two last stages. For a description of the first stage, the reader is referred to [4].

To address the issues of the SWEET-HOME context (i.e., noise, distant-speech) and to benefit from it (i.e., multiple microphones which are continuously recording), we proposed to test the impact of some state-of-the-art and novel techniques that fuse the streams of information at three independent levels of the speech processing: *acoustic signal enhancement*, *decoding enhancement*, and *ASRs output combination* (see [8]). Despite a good improvement of voice command recognition, the method did not include a systematic treatment of the background noise. This section focuses on the implemented techniques for noise cancellation in the case of known noise sources and on the method for vocal order recognition with a specific dataset different from that of our previous work[8].

3.1. Known source noise cancellation

Listening to the radio and watching TV are very frequent everyday activities; this can seriously disturb a speech recogniser at two levels: firstly, speech emitted by the person in the flat can be altered by loudspeakers and badly recognised, and secondly, radio and TV sounds and speech will be analysed by the ASR although their information is not relevant. It is thus mandatory to cancel the radio or the TV noise. To do so we used an acoustic echo cancellation technique (AEC).

The AEC processing chain is described in Figure 3. In AEC, the sound emitted by a noise source $x(n)$ (here the Radio loudspeaker in the smart home) is altered by the room acoustics. The resulting noise $y_b(n)$ of this alteration can be expressed by a convolution product in the time domain $y_b(n) = h(n) * x(n)$, h being the impulse response of the room and n the discrete time. This noise is then mixed with the interesting signal $e(n)$ emitted in the room (here the voice order). The signal recorded by the microphone is then $y(n) = e(n) + h(n) * x(n)$. To cancel the noise, an adaptive filter that estimates the impulse response of the room $\hat{h}(n)$ is generally used to generate an estimate of the original $e(n)$ following the formula:

$\nu(n) = e(n) + y_b(n) - \hat{y}(n) = e(n) + h(n) * x(n) - \hat{h}(n) * x(n)$

The noise filter is adapted using the residual $\nu(n)$ leading to a tracking system where ν is the feedback. At the beginning of the process, some time is needed to learn $\hat{h}$ and decrease the error below a specific threshold: this is the convergence time. When no $x(n)$ signal is present, the filter adaptation tends to diverge because both y_b and $\hat{y}$ are zero. Thus AEC must be applied only when background noise is present. It is also generally required that e stays close to zero otherwise the useful signal (here the domotic order) is seen as an additive noise and the adaptation becomes unstable. This is known as double-talk (i.e., two sources emitting at the same time: here voice order and background noise). To prevent this problem, robust echo cancellers require adjustment of the learning rate to take the presence of double talk in the signal into account. Most echo cancellation algorithms attempt to explicitly detect double-talk but this approach is not very successful, especially in presence of a stationary background noise. In our case, the voice commands used in the experiment are expected to the sufficiently short and separated by silence periods to not disturb the adaptation process. We used the SPEEX library whose AEC stage is based on a multidelay block frequency (MDF) algorithm [9]. Noise cancellation was operated separately on the 7 microphone channels.

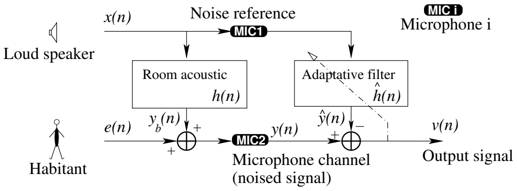

Fig. 3. Echo cancellation principle for noise cancellation

3.2. Voice order recognition

Once noise filtered, the channels feed a multisource ASR system. The ASR system under consideration is the Speeral tool-kit [10] by the the LIA (Laboratoire d'Informatique d'Avignon). Given the targeted application of SWEET-HOME, and its real-time constraints, the 1xRT Speeral configuration was used (decoding time similar to signal duration). At the decoding level, a novel version of the Driven Decoding Algorithm (DDA) was applied within Speeral. DDA aims to align and correct the *a priori* transcripts using the speech recognition engine [11]. This algorithm improves the system performance dramatically by taking advantage of the availability of the predefined transcripts (e.g., automatic speech recognition of a journalist speech using her discourse text as *a priori* information)

In the smart home context, the system knows the grammar of the voice orders and has multiple sources. Thus, in this DDA version, the *a priori* transcript is given by decoding a first channel. Then, at each new generated assumption of the ASR system, the current ASR assumption is aligned with this *a priori* transcript (from the previous decoding pass). Then, a matching score α is computed and integrated with the lan-

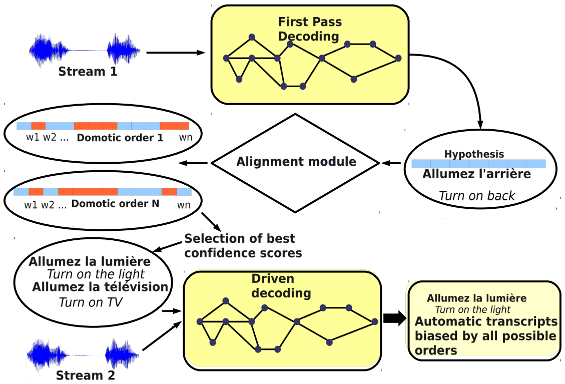

Fig. 4. **DDA 2-level**: vocal orders are recognised from the first decoded stream, the result is then used to drive the decoding of the second stream

guage model [11]:

$$\tilde{P}(w_i|w_{i-1},w_{i-2}) = P^{1-\alpha}(w_i|w_{i-1},w_{i-2})$$

where $\tilde{P}(w_i|w_{i-1},w_{i-2})$ is the updated trigram probability of the word w_i knowing the history w_{i-2}, w_{i-3}, and $P(w_i|w_{i-1},w_{i-2})$ is the initial probability of the trigram.

The applied strategy is dynamic and uses, for each utterance to decode, the best channel for the first pass and the second best channel for the last pass. This approach was extended to take into account *a priori* knowledge about the expected utterances. The ASR system is driven by vocal orders recognised during the first pass : speech segments of the first microphone are projected into the $3 - best$ vocal orders by using an edit distance and injected via DDA into the ASR system for the fast second pass as presented in Figure 4.

By using DDA, the output of the first microphone drives the output of the second one (cf. Figure 4). This approach presents several benefits: - the second ASR system speed is boosted by the approximated transcript (only 0.1xReal-Time), - DDA merges truly and easily the information from the two streams while other strategies (such as ROVER) do not merge ASR systems outputs, - while a strict usage of the grammar may bias toward voice orders, the projection of the voice orders does not prevent the ASR recognising colloquial sentences.

4. EXPERIMENTS AND RESULTS

In all experiments, the **Phase 1** corpus was used for development and training whereas the **Phase 2** corpus served for the evaluation. This section presents the ASR tuning and the experimental results of the proposed approaches.

In the study, the acoustic models were trained on about 80 hours of annotated speech. Furthermore, acoustic models were adapted to each of the 23 speaker by using the Maximum Likelihood Linear Regression (MLLR) on the **Phase 1** data.

Method	Home automation recall	Distress recall	Distress precision	Neutral recall
Without noise	62.1(±16.9)	84.2(±29.2)	88.8(±18.5)	97.5(±5.2)
Without noise+DDA	**92.7(±10.1)**	**87.2(±27.3)**	**89.0(±18.1)**	**97.9(±5.2)**
BN	29.3 (±23.5)	74.3(±22)	73.7(±19.8)	94.5(±4.8)
BN + DDA	**57.2(±30.8)**	**75.2(±22.1)**	**74.7(±19.9)**	**94.6(±5)**
BN+denoising	42.6 (±21.1)	79.4(±19.4)	87.5(±17.6)	97.2(±3.8)
BN+DDA+denoising	**83.5(±16.1)**	**81.2(±19.1)**	**88.0(±18.2)**	**97.8(±3.9)**
Music	59.0(±21)	81.6(±27.6)	87.3(±16.2)	96.8(±4)
Music+DDA	**90.6(±15)**	**82.5(±26.1)**	**87.6(±16.1)**	**97.1(±3.9)**
Music+denoising	46.9(±23.8)	64.5(±36.4)	79.7(±27.1)	94.8(±5.3)
Music+DDA+denoising	**79.2(±16.5)**	**66.5(±34.3)**	**80.7(±27.2)**	**95.1(±4.8)**

Table 2. Home automation and distress detection in three cases: music, Broadcast News (BN) and noiseless

For the decoding, a 3-gram Language Model (LM) with a 10K 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. 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 its possible to use uncertain hypotheses in a fusion process for more robustness.

Results of the approaches are presented in Table 2. In this study we focus on the voice order recognition (classification) stage of each speech event into one of three classes: home automation orders, distress calls and neutral sentences (i.e., sentences that are neither home automation orders nor distress). The recognition is evaluated using recall/precision/F-measure triplet. During the detection, a genuine voice order/distress call is considered as detected only if it completely matches the grammar/distress call sentences. Neutral sentences are considered detected as long as they are not recognized as distress call or voice orders. All other cases are considered as incorrect classification. For each approach, the presented results are the average over the 23 speakers. For the sake of comparison, results of a baseline system (without DDA nor denoised stream) are provided.

The baseline without noise presents a home automation recall of 62% and distress recall of 84%. This better detection can be explained by the lower number of possibilities of distress call expressions than voice orders which reaches about 400. When DDA is used, voice order detection rise to 92.7% and distress call detection is slightly improved (87.2%). Its impact is better for voice command detection because it introduces directly the grammar in the ASR and in the case of distress recall, acts as a combination between two microphones.

In the case of broadcast news, home automation recall falls to 29.3% while distress recall decreases to 74.3%. The introduction of DDA double the home automation detection (57.2%) but has no effect on the distress calls. Using the AEC system both the home automation detection (42.6%) and the distress detection (79.4%) are increased. Finally, the best configuration is obtained by combining the two approaches: the home automation detection is then dramatically increased to 83.5% and distress detection to 81.2%.

In the music context, results are surprising in two points. The music does not seem to impact strongly the ASR as the results with music are just slightly below the results without noise. When the AEC system is used, the performances are overall worst than without it. Thus AEC does not seem adapted to this kind of noise. It must be emphasized that there is only one participant for which the AEC improved the results, and in this case the music was set at a very loud level. Regarding DDA, as in the other cases, results are improved for all classes excepted when it is used in conjunction with the AEC stage.

In all configurations, accuracy of the home automation orders recognition is improved by using DDA: the recognition rate is above 80%. The AEC approach makes a significant difference in case of broadcast news, but does not seem adapted to other background noise.

5. CONCLUSION

This paper describes an approach for voice command recognition in multi-room smart-homes where audio information is captured by several microphones in a distant speech context. This approach is designed to perform in various known background noise. Given that a fully equipped smart home is likely to have a media network (such as UPnP), we performed experiments considering two noise sources that can be known: broadcast news and music (in contrast with other independent source such as a drilling). The approach acts at two levels of the ASR system by applying an AEC and DDA.

Very good results are obtained. Using the DDA, more than 80% of well detected voice commands and distress calls were obtained in noisy and clean conditions. While the DDA is the most conclusive improvement, the denoising method is effective only in the case of radio broadcasts as it leads to very poor performance in music condition. This might be due to the fact that the AEC system introduces noise and non-linear distortion. We will make some further experiments in order to analyse this aspect. It must be emphasized that the baseline ASR system and the DDA one are far less perturbed by music than by voice. Indeed, when a speaker speaks on radio, her words are decoded by the ASR system, thus introducing many errors. But in the case of music, some of the spectrum is filtered by the acoustic models. The AEC is thus highly relevant for this specific application when voice background noise is present, because, in the context of home automation, it is plausible that the system knows what radio or TV broadcasts the person is listening. The future step of our work is to study the use of classical denoising methods like BSS in our context with real time constraints when interference source in unknown (i.e. vacuum).

6. REFERENCES

[1] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects," *Personal and Ubiquitous Computing*, in press.

[2] 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.

[3] A. Vovos, B. Kladis, and N. Fakotakis, "Speech operated smart-home control system for users with special needs," in *Proc. InterSpeech 2005*, 2005, pp. 193–196.

[4] 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.

[5] Matthias Wölfel and John McDonough, *Distant Speech Recognition*, Published by Wiley, 2009.

[6] R. C. Vipperla, M. Wolters, K. Georgila, and S. Renals, "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*, 2009.

[7] J. Barker, H. Christensen, N. Ma, P. Green, and E. Vincent, "The PASCAL 'CHiME' Speech Separation and Recognition Challenge," in *InterSpeech 2011*, 2011.

[8] B. Lecouteux, M. Vacher, and F. Portet, "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions," in *Proc. InterSpeech*, 2011, pp. 2273–2276.

[9] J.-M. Valin and I. B. Collings, "A new robust frequency domain echo canceller with closed-loop learning rate adaptation," in *Proc. ICASSP'07*, 2007, vol. 1, p. 93–96.

[10] G. Linarès, P. Nocéra, D. Massonié, and D. Matrouf, "The LIA speech recognition system: from 10xRT to 1xRT," in *Proc. TSD'07*, 2007, pp. 302–308.

[11] B. Lecouteux, G. Linarès, Y. Estève, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Proc. IEEE ICASSP 2008*, 2008, pp. 1549–1552.