Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions

Benjamin Lecouteux, Michel Vacher, François Portet

Laboratoire d'Informatique de Grenoble, GETALP Team
UMR CNRS/UJF/G-INP 5217, Grenoble, France
Benjamin.Lecouteux@imag.fr, Michel.Vacher@imag.fr, Francois.Portet@imag.fr

Abstract

While the smart home domain has become a major field of application of ICT to improve support and wellness of people in loss of autonomy, speech technology in smart home has, comparatively to other ICTs, received limited attention. This paper presents the SWEET-HOME project whose aim is to make it possible for frail persons to control their domestic environment through voice interfaces. Several state-of-the-art and novel ASR techniques were evaluated on realistic data acquired in a multiroom smart home. This *distant speech* French corpus was recorded with 21 speakers playing scenarios including activities of daily living in a smart home equipped with several microphones. Techniques acting at the decoding stage and using *a priori* knowledge such as DDA give better results (WER=8.8%, Domotic F-measure=96.8%) than the baseline (WER=18.3%, Domotic F-measure=89.2%) and other approaches.
Index Terms: home automation, smart home, distant speech, multisource ASRs, keyword detection

1. Introduction

Since the rise of *Ubiquitous Computing*, new ways of conceiving our home environment appeared. One of these, 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 ranging from lights, automatic motorised blinds... to Hi-Fi systems, PCs, alarms systems, etc. that fit modern homes. These smart homes represent a promising solution to support the elderly and disabled persons in living in their own home as autonomously as possible. Among all the interaction and sensing technologies used in smart home, speech processing technology has a great potential to become one of the major interaction modalities in smart home. Indeed, voice interfaces are much more adapted to disabled people and the ageing population who have difficulties in moving or seeing, than tactile interfaces (e.g., remote control) which require physical and visual interaction [1, 2]. Moreover, voice command is particularly suited to distress situations. A person, who cannot move after a fall but being conscious, may have still the possibility to call for assistance while a remote control may be unreachable. Despite all this, very few smart home projects have seriously considered speech recognition in their design [1, 2]. Part of this can be attributed to the complexity of setting up this technology in a real environment and to important challenges that still need to be overcome [3].

The SWEET-HOME project has started in 2010 to address some of these challenges. One of the major issues that prevents the development of speech technology in real home setting is the poor performance of Automatic Speech Recognition (ASR) in noisy environment [3]. Indeed, ASR systems have reached correct performances with close talking microphones (e.g. head-set), but the performance decreases significantly as soon as the microphone is moved away from the mouth of the speaker. In realistic conditions, this deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise such as TV, radio and devices [4]. All these problems, related to the so called *'distant speech'* context, should be taken into account in the home context [5]. While, user linguistic preferences, dialogues and age dependant voice interfaces have been studied during this decade [1, 2, 6], distant speech in smart home received attention very recently within the speech processing community [7].

This paper presents results of state-of-the-art and novel ASR techniques evaluated on realistic data acquired in a multiroom smart home. Before presenting our experimental framework (Section 3), the proposed techniques are described in Section 2. The experiments and results are then presented in Section 4. This paper concludes with brief remarks about the results and future work.

2. SWEET-HOME project and corpus

The SWEET-HOME project (`sweet-home.imag.fr`) aims at designing a new smart home system based on audio technology focusing on three main aspects: to provide assistance via *natural man-machine interaction* (voice and tactile command), to ease *social e-inclusion* and to provide *security reassurance* by detecting situations of distress. If these aims are achieved, then the person will be able to pilot, from anywhere in the house, their environment at any time in the most natural way possible. The targeted smart environments in which speech recognition must be performed thus include multi-room homes with one or more microphones per room set near the ceiling. This places the project in a distant-speech context where microphones may be far apart from each other and may thus record similar or very different sources. The most close projects seem to have focused mainly on one-room microphone array [1] or one or unspecified number of microphones [2, 6].

To achieve the project goals, the DOMUS smart home depicted in Figure 1 was adapted to acquire a realistic corpus and to test the developed techniques. This smart home was set up by the Multicom team of the Laboratory of Informatics of Grenoble, partner of the project. It is a thirty square meters suite flat including a bathroom, a kitchen, a bedroom and a study, all equipped with sensors and effectors such as infra-red presence detectors, contact sensors, video cameras (used only for annotation purpose), etc. In addition, seven microphones were set in the ceiling.

An experiment was conducted to acquire a representative speech corpus composed of utterances of domotic order, distress call and casual sentence. This corpus is called the SWEET-

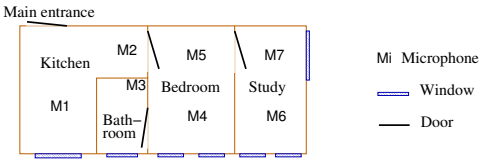

Figure 1: Position of the 7 microphones in the DOMUS Home

HOME *speech corpus*. 21 persons (including 7 women) participated to a 2-phase experiment to record, among other data, speech corpus in a daily living context. The average age of the participants was 38.5 ± 13 years (22-63, min-max). To ensure that the audio data acquired would be as close as possible to real daily living sounds, the participants were asked to perform several daily living activities in the smart home. A visit, before the experiment, was organized to make sure that the participants will find all the needed items to perform the activities. No instruction was given to any participant about either 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 is about 2 meters. Sound data were recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card [3].

The first phase (**Phase 1**) consisted in following a scenario of activities without condition on the time spent and the manner of achieving them (having a breakfast, simulate a shower, get some sleep, clean up the flat using the vacuum, etc.). During this first phase, participants uttered 40 predefined casual sentences on the phone (e.g., "Allo" (*Hello*), "J'ai eu du mal à dormir" (*I slept badly*)) but were also free to utter any sentence they wanted (some did speak to themselves aloud). The second phase (**Phase 2**) consisted in reading aloud a list of 44 sentences whose 9 were distress sentences (e.g., "À l'aide"(*help*), "Appelez un docteur" (*call a doctor*)) and 3 were domotic orders (e.g., "Allumez la lumière" (*turn on the light*)). In this paper, experiments are performed without device noise (TV, radio, vacuum, ..).

Finally, the French SWEET-HOME speech corpus is made of 862 sentences (38 minutes 46s per channel in total) for **Phase 1**, and 917 sentences (40 minutes 27s per channel in total) for **Phase 2** all from 21 speakers. Each sentence is available for each channel and has been humanly annotated on the best Signal-to-Noise Ratio (SNR) channel. The average SNR for the considered sentences of the best SNR channel is 20.3 dB (SNR is typically around 55 db in studio record). It must be clear that the data from the 7 microphones was the only data source used in this study.

3. Proposed approaches for robust ASR

To detect domotic commands in the SWEET-HOME context, we propose a three-stage approach. The first one detects audio activity and classifies it as speech or other sound, the second one extracts the utterances using an ASR system and the last one recognizes a vocal command or a distress situation from the decoded utterances. This paper describes the two last stages, for the first stage the reader is referred to [4].

To address the issues of the SWEET-HOME context (noise, distant-speech) and to benefit from it (multiple microphones which are continuously recording) we propose 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*. The remaining of this section presents the implemented techniques for robust ASR and the chosen method for vocal order recognition.

3.1. 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 beam-forming algorithm [8] was used to merge all channels in a single one which fed 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:

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

where $W_m[t]$ is the weight for microphone m at time t, knowing that $\sum_{m=1}^{M} W_m[t] = 1$, the signal of the m^{th} channel is $x_m[t]$ and $D^{(m,ref)}[t]$ is the delay between the m^{th} channel and the reference channel. In our experiments, the reference channel was the one with the highest SNR overall in **Phase 2** and the 7 signals were entirely combined for each speaker rather than doing a sentences based combination (the tested algorithm failed with too short sentences). Once the new signal y is computed, it can feed a monosource ASR stage.

3.2. Driven Decoding Algorithm

At the decoding level, a novel version of the Driven Decoding Algorithm (DDA) was applied. DDA aims to align and correct auxiliary transcripts by using a speech recognition engine [9, 10]. This algorithm improves system performance dramatically by taking advantage of the availability of the auxiliary transcripts.

DDA acts at each new generated assumption of the ASR system. The current ASR assumption is aligned with the auxiliary transcript (from a previous decoding pass). Then a matching score α is computed and integrated with the language model [9]:

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

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.

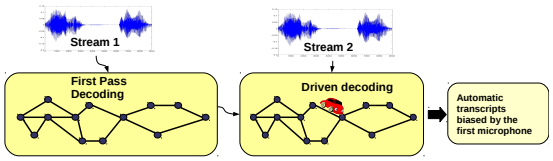

Figure 2: **DDA** used with two streams: the first stream allows one to drive the second stream

We propose to use a variant of the Driven Decoding Algorithm where the output of the first microphone is used to drive the output of the second one (cf. Figure 2). This approach presents two main benefits:

- The second ASR system speed is boosted by the approximated transcript (only 0.1xRT),

- DDA merges truly and easily the information from the two streams while voting strategies (such as ROVER) do not merge ASR systems outputs.

The applied strategy is dynamic and used, for each utterance to decode, the best channel for the first pass and the second best channel for the last pass.

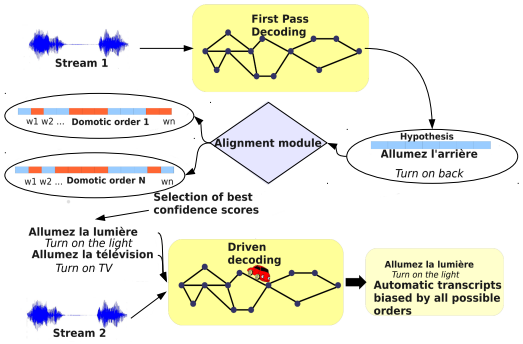

Figure 3: **DDA 2-level**: vocal orders are recognized from the first decoded stream which are then used to drive the decoding of the second stream

This approach was extended to take into account *a priori* knowledge about the expected utterances. The ASR system is driven by vocal orders recognized during the first pass. This method is called **DDA 2-level**: speech segments of the first pass are projected into the $3 - best$ vocal orders by using an edit distance (cf. 3.4) and injected via DDA into the ASR system for the fast second pass as presented in Figure 3.

3.3. ROVER

At the ASR combination level, a ROVER [11] was applied. ROVER 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 was tested using all available channels without *a priori* knowledge. In a second time, an *a priori* confidence measure based on the SNR was used: for each decoded segment s_i from the i^{th} ASR system, the associated confidence score $\phi(s_i)$ was computed by $\phi(s_i) = 2^{R(s_i)}/\sum_{j=1}^{7} 2^{R(s_j)}$ where $R()$ is the function computing the SNR of a segment and s_i is the segment generated by the i^{th} ASR system. For each annotated sentence a silence period I_{sil} at the beginning and the end is taken around the speech signal period I_{speech}. The SNR is thus evaluated as:

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

Finally, a ROVER using only the two best channels overall was tested in order to check whether other channels contain redundant information and whether good results can be reach with reasonable computational cost.

3.4. Detection of domotic orders and distress sentences with proposed approaches

We propose to transcript each domotic order and distress sentences in a phoneme graph in which each path corresponds to a variant of pronunciation. Then the number of sentences to detect is 12 (3 domotic orders + 9 distress sentences). Automatic transcripts are transcribed in the same way.

In order to locate domotic orders into automatic transcripts T of size m, each sentence of size n from domotic orders H are aligned to T by using a Dynamic Time Warping (DTW [12]) algorithm at the phonetic level. The deletion, insertion and substitution costs were computed empirically. The cumulative distance $\gamma(i,j)$ between H_j and T_i is computed as: $\gamma(i,j) = d(T_i, H_j) + min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\}$

Each domotic order is aligned and associated with an alignment score: the percentage of well aligned symbols. The domotic order with the best score is then selected for decision according a detection threshold. This approach takes into account some recognition errors such as word endings or slight variations. Moreover, in many cases, a miss-decoded word is phonetically close from the good one (due to the close pronunciation).

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.

4.1. The Speeral ASR system

The LIA (Laboratoire d'Informatique d'Avignon) speech recognition tool-kit Speeral [13] was chosen as unique ASR system. This choice was made based on experiments we undertook with several state-of-the-art ASR systems and on the fact that DDA is only implemented in Speeral. Speeral relies on an A^* decoder with HMM-based context-dependent acoustic models and trigram language models. HMMs are classical three-state left-right models and state tying is achieved by using decision trees. Acoustic vectors are composed of 12 PLP (Perceptual Linear Predictive) coefficients, the energy, and the first and second order derivatives of these 13 parameters.

In the study, the acoustic models were trained on about 80 hours of annotated speech. Given the targeted application of SWEET-HOME the computation time should not be a breach of real-time use. Thus, the 1xRT Speeral configuration was used. In this case, the time required by the system to decode one hour of speech signal is real-time (noted 1xRT). The 1xRT system uses a strict pruning scheme. Furthermore, acoustic models were adapted for each of the 21 speaker by using the Maximum Likelihood Linear Regression (MLLR) and the annotated **Phase 1** corpus. MLLR adaptation is a good compromise while only a small amount of annotated data is available

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 specialized LM (weight 90%). The *generic* LM was estimated on about 1000M of words from the French newspapers *Le Monde* and *Gigaword*. The *specialized* LM was estimated from the sentences (about 200 words) that the 21 participants had to utter during the experiment (domotic orders, casual phrases, etc.).

4.2. Results

Results of the approaches are presented table 1. The ASR stage was evaluated using the Word Error Rate (WER) whereas the vocal order recognition (classification) stage was evaluated using recall/precision/F-measure triplet: the number of domotic orders is about 10. Domotic orders were manually specified by annotating all sentences. During the detection, if a marked do-

motic order is well detected, it is considered as detected. In all other cases, a detected order is considered as false detection. For each approach, the presented results are the average over the 21 speakers (plus standard deviation for the WER). For the sake of comparison, results of a baseline and an oracle baseline systems are provided. The baseline system outputs the best decoding amongst 7 ASR systems according to the highest SNR. The oracle baseline is computed by selecting the best WER for each speaker.

Method	WER$^{\pm SD}$	Domotic recall	Domotic precision	F -measure
Baseline	$18.3^{\pm 12.1}$	88.0	90.5	89.2
Oracle Baseline	$17.7^{\pm 10.3}$	88.5	91.3	89.9
Beam Forming	$16.8^{\pm 8.3}$	89.0	92.6	90.8
DDA +SNR	$11.4^{\pm 5.6}$	93.3	97.3	95.3
DDA 2 lev.+SNR	$\mathbf{8.8}^{\pm 3.7}$	**95.6**	**98.1**	**96.8**
ROVER	$20.6^{\pm 8.5}$	85.0	90.0	87.4
ROVER 2c+SNR	$13.0^{\pm 6.6}$	91.3	95.3	93.3
ROVER +SNR	$\mathbf{12.2}^{\pm 6.1}$	**92.7**	**97.4**	**95.0**
ROVER Oracle	$7.8^{\pm 2.7}$	99.4	98.9	99.1

Table 1: WER, Domotic orders detection

The baseline system achieved a 18.3 % WER (best SNR channel). All proposed SNR-based approaches benefited from the multiple available microphones. Beamforming led to a 8.1% relative WER improvement. This result shows that combining all channels increases the ASR task robustness. The DDA method showed a 37.8% relative improvement by using the SNR. The 2 level DDA presented a 52 % relative improvement with a very high stability (SD=3.7): this gain is easily explained as the second decoding pass was perfused with *a priori* knowledge (i.e., domotic orders) triggered by the first pass. Finally, the SNR-based ROVER led to a 33.4% relative improvement.

In all configurations, accuracy of the vocal orders recognition was good: the baseline recognition gave a 89.2% F-measure. It can be observed that in other configurations the spotting task correlated with the WER. Thereby, ROVER and the two DDA configurations led to a significant F-measure improvement over the baseline of about 7% absolute. Beamforming gain was not significant. ROVER performed detections similar to the DDA approaches, but required to decode all channels. Finally, the best configuration was based on the 2 level DDA leading to a 96.8% F-measure.

5. Conclusion

Several approaches were presented to perform accurate vocal order recognition in multi-room smart-homes where audio information is captured by several microphones in a distant speech context. The proposed approaches were acting at the three main levels of the ASR task: acoustic, decoding and hypothesis selection. Some of them included *a priori* knowledge either dynamically computed such as the SNR or acquired off line such as the predefined domotic orders.

Results confirmed that the use of the seven microphones improved the ASR accuracy. Beamforming improved the WER (16.8%), however its performance were very close to the baseline one (18.3%). This may be due to the fact that the seven microphones are far apart from each other and might not contain enough redundancy to obtain a really enhanced acoustic signal. The Driven Decoding Algorithm gave the best performance with a 11.4% WER and 95.3% F-measure for vocal order classification. DDA results were only slightly better than the ROVER results, however DDA needs only two channels while ROVER necessitates 7 ASR systems performing concurrently to approach DDA performances. The DDA computational cost is thus very low compared to the ROVER one. Moreover, the 2-level DDA approach makes it possible to include *a priori* knowledge to increase performances to 8.8% WER and 96.8% F-Measure with much better stability than the baseline (3.7% WER standard deviation vs. 10.3%). However, this amelioration will be achieved only if test data contains domotic orders. This study shows that good recognition rate can be obtained by adapting classical ASR systems mixing multisource and domain knowledge. We plan to adapt these approaches to noisy conditions notably by applying source separation techniques to real daily living records composed of uncontrolled noise.

6. Acknowledgements

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)

7. References

[1] 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.

[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] 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.

[4] M. Vacher, A. Fleury, J.-F. Serignat, N. Noury, and H. Glasson, "Preliminary Evaluation of Speech/Sound Recognition for Telemedicine Application in a Real Environment," in *Proc. InterSpeech 2008*, 2008, pp. 496–499.

[5] M. Wölfel and J. 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 International: 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, (to appear).

[8] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[9] B. Lecouteux, G. Linarès, J. Bonastre, and P. Nocéra, "Imperfect transcript driven speech recognition," in *Proc. InterSpeech'06*, 2006, pp. 1626–1629.

[10] 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.

[11] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop ASRU*, 1997, pp. 347–354.

[12] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Workshop on Knowledge Discovery in Databases (KDD'94)*, 1994, pp. 359–370.

[13] 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.