ATHENA: A Greek Multi-Sensory Database for Home Automation Control

Antigoni Tsiami[1,3], *Isidoros Rodomagoulakis*[1,3], *Panagiotis Giannoulis*[1,3], *Athanasios Katsamanis*[1,3], *Gerasimos Potamianos*[2,3], *Petros Maragos*[1,3]

[1] School of Electr. and Computer Eng., National Technical Univ. of Athens, 15773 Athens, Greece
[2] Department of Electr. and Computer Eng., Univ. of Thessaly, 38221 Volos, Greece
[3] Athena Research and Innovation Center, 15125 Maroussi, Greece

{antsiami,irodoma,nkatsam,maragos}@cs.ntua.gr, gpotam@ieee.org

10.21437/Interspeech.2014-382

Abstract

In this paper we present a Greek speech database with real multi-modal data in a smart home two-room environment. In total, 20 speakers were recorded in 240 one-minute long sessions. The recordings include utterances of activation keywords and commands for home automation control, but also phonetically rich sentences and conversational speech. Audio, speaker movements and gestures were captured by 20 condenser microphones installed on the walls and ceiling, 6 MEMS microphones, 2 close-talk microphones and one Kinect camera. The new publicly available database exhibits adverse noise conditions because of background noises and acoustic events performed during the recordings to better approximate a realistic everyday home scenario. Thus, it is suitable for experimentation on voice activity and event detection, source localization, speech enhancement and far-field speech recognition. We present the details of the corpus as well as baseline results on multi-channel voice activity detection and spoken command recognition.

Index Terms: smart homes, data collection, speech database

1. Introduction

Recently, much research has focused on smart homes, namely domestic environments that by using a multitude of sensors can to a certain degree understand user's state and intentions and support easy control of various automated activities [1–3]. Specifically, distant speech interaction for home applications seems particularly attractive and useful, e.g., to elderly or disabled people, because it can achieve unobtrusive and hands-free communication employing mainly microphone sensors placed in the background. However, allowing for far-field recordings in a real domestic setting requires coping with quite challening acoustic conditions, exhibiting significant noise, reverberation and background events overlapping with speech. For this purpose, the synergy of different components is required, such as source localization, event detection, voice activity detection, speech enhancement, speech recognition and understanding.

Of particular importance for research in this domain is the existence of speech databases recorded inside smart home environments in order to evaluate the various algorithms under these adverse conditions. Such databases may contain either simulated or real data. Although simulated data are easy to be produced and convenient because they allow better control of the acoustic conditions for systematic experimentation, they cannot substitute the need for real data because they present important limitations compared to the latter. One well-known single-room database that contains real speech in English collected from meetings was developed in [4], where microphones were placed on a meeting table and speech segments highly overlap. Another single-room database that contains both isolated and speech overlapping acoustic events has been presented in [5].

Concerning multiple-room environments, in [6] a multimodal corpus containing several acoustic events was collected in a smart home. A challenging database containing simulated data in four different languages (Italian, Austrian German, Greek and Portuguese) has been developed and described in [7] where the smart home consists of 4 rooms and exhibits significantly adverse conditions. In [8] speech in French was recorded in a 5-room smart home, overlapping either with noises or background events but comprising only automation commands uttered one after the other. Thus, there is no system activation or keyword spotting process.

The existence of real speech data is considered of critical importance for the development of the various components. In this paper we present ATHENA, our real speech database in Greek, recorded in Athena-RC smart home environment, consisting of two rooms. Activation keywords and home automation commands, as well as phonetically rich sentences were uttered from 20 speakers from various positions. Also, some sessions include conversation between two speakers. Most speech segments highly overlap with acoustic events and background noises, thus rendering the database quite challenging and realistic. Visual data were also recorded, both for speaker tracking and gesture capturing. The real speech database that we present is the first one to provide real audio data in a multi-room setting comprising not only speech commands but also their context. The database is publicly released [1].

2. Room setup and equipment

The Athena-RC smart home environment consists of two rooms: The main room is an office environment and the second room is part of the corridor. Figure 1 depicts the floor plan, while Fig. 2 shows a picture of the room during the recordings.

2.1. Equipment and synchronization

As depicted in Fig. 1, a total of 20 condenser microphones (Shure MX391/O, omnidirectional at 48kHz rate) are distributed on the walls and the ceiling. More specifically, 6 microphones are placed on the ceiling in a pentagon geometry with

This research was supported by the EU project DIRHA with grant FP7-ICT-2011-7-288121.

[1] http://cvsp.cs.ntua.gr/research/athenadb

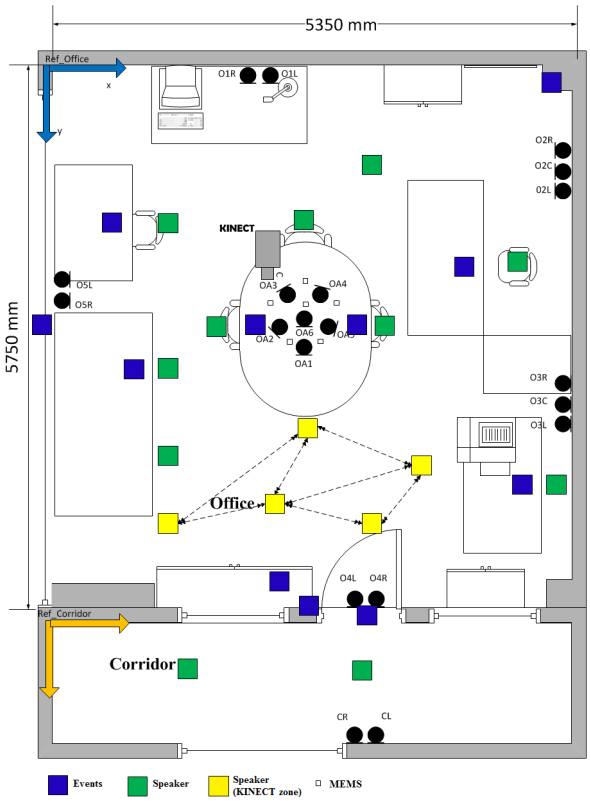

Figure 1: *Floor plan of the Athena-RC smart home environment. Sensors, speaker and event positions are depicted.*

one microphone at the center and the rest are distributed on the walls per two or three. Also, in order to keep the clean utterances, we used 2 close-talk microphones (Sennheiser ew172G3, at 48kHz rate). The 20 condenser microphones and the 2 close-talk ones were plugged in three 8-channel pre-amplifiers (Focusrite OctoPre MkII). Apart from the condenser microphones, we also used 6 MEMS microphones [9] which is a new technology of very small omnidirectional microphones (also at 48kHz rate). The latter were placed on the ceiling in a similar geometry with the condenser ones, for comparison purposes. We also used a Kinect camera (at 10 fps) in order to capture visual data. We recorded three different video streams, the RGB, Depth and Skeleton. Finally, we used a loudspeaker (Genelec 8030BPM) in order to measure the room impulse responses.

The synchronization of the 20 condenser and the 2 close-talk microphones was achieved through the connection to the synchronized pre-amplifiers. Also, the Kinect and MEMS data collection was achieved through the Robot Operating System (ROS) [10] which writes data from multiple sensors synchronized via timestamps.

2.2. Room setup and positions

A total of 16 different speaker positions with various orientations were used in the recordings. As depicted in Fig. 1, the positions are divided in two different zones, Zone A (yellow squares) and Zone B (green squares), depending on whether the Kinect camera has view of the position, or not. For some positions placed next to a desk or a table, we considered two states: one of a standing speaker and one of a seated speaker. All positions were marked in order to keep the ground truth location.

Audio type	#	%
Speech		
Conversation	100	20.9
Phonetically rich	140	4.0
System activation	240	2.3
Commands	240	3.6
Total	960	30.8
Long Events		
Walking steps	112	3.6
Cellphone ring	104	3.0
Keyboard	124	4.3
Glass fill	108	2.0
Coffee spoon	116	2.9
Skype call	100	3.2
Cough	96	1.9
Paper work	100	2.8
Window open/close	96	2.4
Total	956	26.1

Audio type	#	%
Short Events		
Mouse click	152	1.9
Keys	176	3.8
Knock	148	1.2
Chair moving	168	2.2
Switch on/off	156	1.2
Door open/close	144	2.1
Total	944	12.4

Background Noises		
Ambient noise	46	20.0
Fan	54	22.0
Radio music	70	29.0
Vacuum cleaner	47	20.0
Silence	21	9.0
Total	238	100

Table 1: *The audio types of the database categorized in speech, events, and background noises. Their instances (#) and their total duration as a percentage (%) over the 240 minutes of the database. Events are classified in short and long with average segment durations 1.5s and 3s respectively.*

Except for speaker positions, we also included some positions (marked with blue colour in Fig. 1) for background events occurrences. These are mainly positions on tables, workstations or next to doors and windows, so as to resemble a real smart home and the everyday activities taking place. These positions were marked as well.

3. Database description

Our database consists of 20 speakers (10 male and 10 female). Each of them participated in twelve 1-min sessions. This duration was chosen to simulate the long audio recordings typically acquired by an "always listening" speech interface. Thus, the whole database duration counts up to 240 minutes. The database contains both audio and visual data, as well as the estimated room impulse responses. Concerning audio, the recording scenarios include various speech types, acoustic events and noises, whose statistics are presented in Table 1. It should be noted that about 40% of the total speech overlaps with background events. Visual data contain gestures related with system activation and also the moving speaker's coordinates. The database is partitioned in one development and one testing set containing 10 speakers each (5 male, 5 female).

3.1. Speech

The employed speech types are keywords, commands, phonetically rich sentences and conversational speech. Regarding the first three types, the number of different available utterances were 12, 170 and 190 respectively. For conversational speech, we used several topics that could prompt the speakers to commence a spontaneous dialogue, for example cinema, music, science, etc. The phonetically rich sentences are extracts from a Greek, large vocabulary database named "Logotypografia" [11]. Keywords are short phrases containing the utterance "DIRHA" [12], and commands concern home automation control. Table 2 presents each different session.

3.2. Background events and noises

In order to simulate more realistically the conditions and the events occurring in a real smart home environment, we intro-

Type	Session	Sequence	Position Zone
Static Single Spk	1	S1: PH - DK - DC	B
	2	S1: PH - DK - DC	B
	3	S1: PH - DK - DC	B
	4	S1: PH - DK - DC	A
	5	S1: PH - DK - DC	A
Static 2 Spks	6	S1: CS - DK - DC & S2: CS	B
	7	S1: CS - DK - DC & S2: CS	B
	8	S1: CS - DK - DC & S2: CS	B
Moving Single Spk	9	S1: PH - DK - DC	A
	10	S1: PH - DK - DC	A
Moving 2 Spks	11	S1: CS - DK - DC & S2: CS	A
	12	S1: CS - DK - DC & S2: CS	A

Table 2: *Sequences per session, where DK is the keyword, DC is the command, PH is a phonetically rich sentence and CS conversational speech. S1 is the main speaker and S2 the second speaker (always static) when there is a conversation.*

duced several background events and noises occurring in every session, performed by people. Table 1 presents the various background events, categorized by duration in long and short events. In each simulation, one long and one short event takes place, with four instances each, randomly distributed into the 1-minute session. Thus, there may be either isolated events or events overlapping with speech.

Regarding noises, we employed 5 different types: a) Silence b) Radio played from a laptop c) Fan d) Ambient noise from open window e) Vacuum cleaner placed in the corridor. Noises occur during the whole 1-minute session.

3.3. Gesture

For purposes of multi-modal processing and interaction, as well as to further aid a system activation and keyword spotting process, we introduced a gesture while the speaker was uttering a keyword in Kinect sessions. The gesture type is a raised hand in fist, in order for the Kinect to be able to track the gesture independently of the speaker's orientation. An example can be found in Fig. 2, where the three captured Kinect streams, RGB, Depth, Skeleton are being depicted along with the MEMS microphones outputs for a keyword instance.

3.4. Impulse response estimation

Apart from the collection of real data, we also measured the room impulse responses (IRs) from each source position and orientation to the microphones. The IRs measurements were based on a professional studio monitor (Genelec 8030A) able to excite the target environment with long sequences of Exponential Sine Sweep (ESS) signal [13]. As pointed out in [14], ESS method ensures IRs measurements with high SNR and robustness against harmonic distortions.

4. Real data collection and annotation

In order to achieve a fair distribution of source positions, utterances, events positions and time boundaries over all sessions, we randomized all the above parameters, ensuring that each utterance should appear in at least one session of one speaker.

The Athena-RC team used ELAN annotation tool [15] to guide the speakers during the recording process to follow the recording script. The tool indicated the speaker positions, the sentences he/she should pronounce and the background events in time slots, as can be seen in Fig. 3.

The speaker guidance was achieved through synchronized

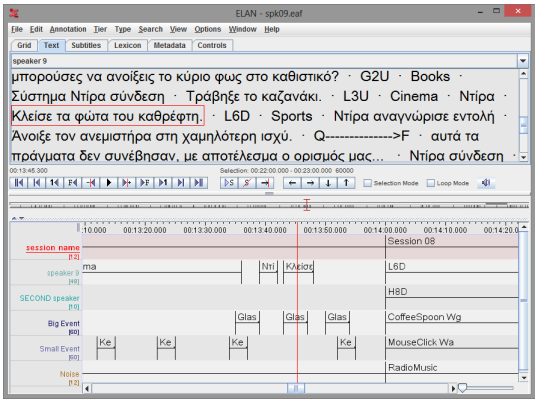

Figure 3: *ELAN annotation tool: The different tiers represent speech, position, events etc. The red contour indicates the sentence to be uttered.*

monitors distributed in the two rooms displaying the annotation information. Before the session beginning, the speaker was prompted to stand on the position marker indicated by the annotation tool, looking towards the direction indicated by a number next to the source position information.

After the recordings, we re-annotated the data, in order to correct the time boundaries and also annotate some external events that were not included in the recording scenarios. Such events may concern babble noise coming from neighbouring offices, doors opening and closing or walking steps.

5. Baseline experiments

We conducted some preliminary experiments on the ATHENA database, concerning voice activity detection and far-field command recognition using at this point the 20 condenser microphones. The methodologies and the corresponding results are being described in the following sections.

5.1. Voice activity detection (VAD)

Both single and multi-channel VAD approaches have been proposed in the bibliography [16–18]. The proposed VAD system implements a multichannel approach to determining the temporal boundaries of speech activity in the smart home. The sequence of audio events, namely speech or non-speech, is estimated by means of the Viterbi algorithm [19] on combined scores coming from single-channel event models for the entire observation sequence. These combined scores are estimated as averages of scores based on channel-specific event models ("sum of log-likelihoods").

The probability distributions for each event at each channel are modelled as Gaussian mixture models. The features used are baseline MFCCs with Δ's and $\Delta\Delta$'s. The Viterbi algorithm allows the identification of the optimal sequence of audio events in the smart home for a given recording session. The incorporation of the multichannel score essentially leads to a decision that is informed by all the microphones in the home. The details of the proposed VAD are presented in [20].

The results for the VAD task are depicted in Table 3. Performance of the various approaches is reported in terms of two metrics; "Success rate" which practically corresponds to frame classification and "F-score" which corresponds to the harmonic mean of precision and recall. The baseline in our experiments corresponds to the "best-SNR channel" output per session. For

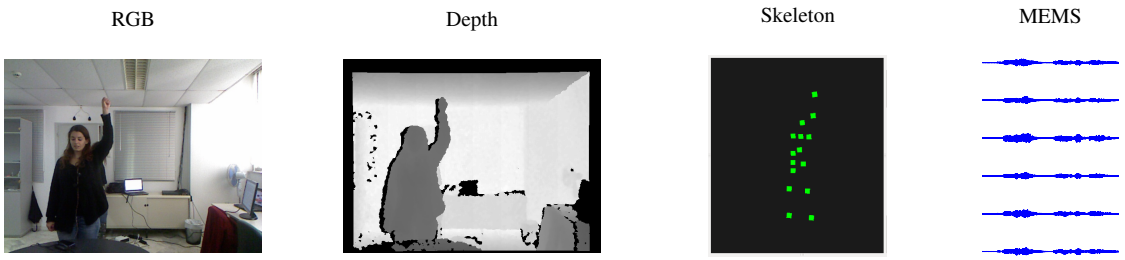

Figure 2: *RGB, Depth, Skeleton and MEMS outputs for a spoken keyword performed along with the gesture.*

VAD methods	Success rate (%)	F-score
average over channels	95.85	0.793
best-SNR channel	96.25	0.820
sum of log-likelihoods	96.62	0.832

Table 3: *Results for the VAD task.*

a given speech segment detected by a particular channel, the local SNR is computed as the ratio of energies between the speech segment and the preceding non-speech segment of 0.5-sec duration. An average SNR across segments is computed for each session and channel. The "average over channels" field denotes the average scores across channels for the single-channel experiment. Note that VAD output was evaluated only on non-conversational speech, because conversational speech has not yet been fully transcribed. As noted, the multi-channel approach outperforms the single-channel ones, yielding satisfying results. Also, the "best-SNR channel" selection achieves a better performance than the "average over channels", as expected.

5.2. Far-field command recognition

This section demonstrates a baseline system for the recognition of the 170 commands contained in the 1-minute long sessions of ATHENA database. These first experiments are designed to evaluate the task of far-field command recognition focusing on the challenges that emerge due to the real conditions of the database as described in the above sections. Thus, in this work, the task is limited to the recognition of commands that are segmented using the speech boundaries provided by the VAD output for the segment following the keyword. The keyword's location is retrieved using the ground truth transcriptions. Moreover, as the focus is on the acoustic conditions, the language model of the recognizer is factored out by using a finite state grammar for the set of commands to be recognized.

The system described here is based on the baseline part of our previous work [21] on far-field command recognition for simulated data. Methods such as environmental adaptation of the acoustic models and channel selection, which led to improvements, are also applied here for the case of real data. The employed recognizer is our HTK [22] based system for large vocabulary continuous speech recognition in Greek [23], which consists of tied state triphones trained on MFCCs extracted from clean speech of "Logotypografia" database [11]. The models are adapted on the development data of each microphone using Maximum Likelihood Linear Regression (MLLR) and channel selection is based on the SNR of the speech segment to be recognized. More details on adaptation and channel selection can be found in [21].

Table 4 presents recognition performance for the 120 sessions of the testing set. The "min", "median" and "best" microphones depict how the performance over all sessions can vary among the 20 microphones. The wide ranges of the distributions which are approximately 35% and 40% for the original and adapted clean models respectively indicate that the performance of each microphone is strongly correlated to the source positions and the background events and noises. The performance of the "SNR-best" microphone per session is better by almost 7% compared to the "best" microphone over all sessions. Regarding adaptation, microphone dependent MLLR adaptation leads to an improvement of median accuracy close to 18% and when adaptation is combined with channel selection the performance reaches 85.67% which is the best performance of this baseline system.

		models	
microphones		clean	clean+MLLR
far-field	min	27.49	40.20
	median	55.41	73.25
	best	62.13	80.12
	SNR-best	69.74	85.67
	oracle	82.02	92.69
close-talk		95.47	95.47

Table 4: *Single-microphone command recognition: word accuracy across all condenser microphones. The results are with MLLR adaptation and SNR based channel selection. The "SNR-best" microphone per session is the one selected based on the highest SNR. Performances corresponding to the "oracle" microphone per session and the "close-talk" microphone are also presented as upper limits for microphone selection and single-channel recognition respectively.*

6. Conclusions

We have presented ATHENA database, a new real speech database in Greek for smart home applications. Various background events and noises overlap with speech data uttered from different positions, thus approximating a realistic domestic scenario. The employment of multiple audio and video sensors permits both multi-channel and multi-modal processing, rendering our database suitable for developing and evaluating algorithms for source localization, speech enhancement, acoustic event detection, voice activity detection and far-field speech recognition. For the two latter problems, the baseline results presented also indicate that there is space for improvement and future research.

7. Acknowledgements

The authors would like to thank the NTUA CVSP lab and Athena-RC members who participated in the recordings and especially Valia Sfika for the data re-annotation.

8. References

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

[2] M. P. Poland, C. D. Nugent, H. Wang, and L. Chen, "Smart home research: projects and issues," in *Ubiquitous Developments in Ambient Computing and Intelligence: Human-Centered Applications*, K. Curran, Ed. Information Science Reference, 2011, pp. 259–272.

[3] 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, pp. 127–144, 2013.

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, 2003, pp. 364–367.

[5] A. Temko, D. Macho, C. Nadeu, and C. Segura, "UPC-TALP database of isolated acoustic events," *Internal UPC report*, 2005.

[6] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A multimodal corpus recorded in a health smart home," in *Proc. LREC*, 2010.

[7] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. LREC*, 2014.

[8] 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 *Proc. LREC*, 2014.

[9] "MEMS audio sensor omnidirectional digital microphone, MP34DT01," STMicroelectronics, 2013, [Online] Available at: http://www.st.com.

[10] "ROS (Robot Operating System)," [Online] Available at: http://www.ros.org/.

[11] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in greek: Corpus and an automatic dictation system," in *Proc. Interspeech*, 2003.

[12] "The DIRHA (Distance-speech Interaction for Robust Home Applications) EU project," [Online] Availble at: http://dirha.fbk.eu/.

[13] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. AES Convention*, 2000.

[14] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *Proc. EUSIPCO*, 2012.

[15] "ELAN (EUDICO Linguistic Annotator)," [Online] Available at: http://tla.mpi.nl/tools/tla-tools/elan/.

[16] Q. Li, J. Zheng, A. Tsai, and Q. Zhou "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. Speech and Audio Process.*, vol. 10, no. 3, pp. 146–157, 2002.

[17] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using MFCC features and Support Vector Machine," in *Proc. SPECOM*, 2007.

[18] J. E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto, "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," in *Proc. ICASSP*, 2007.

[19] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

[20] P. Giannoulis, A. Tsiami, I. Rodomagoulakis, A. Katsamanis, G. Potamianos, and P. Maragos, "The Athena-RC system for speech activity detection and speaker localization in the DIRHA smart home," in *Proc. HSCMA*, 2014.

[21] A. Katsamanis, I. Rodomagoulakis, G. Potamianos, P. Maragos, and A. Tsiami, "Robust far-field spoken command recognition for home automation combining adaptation and multichannel processing," in *Proc. ICASSP*, 2014.

[22] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, United Kingdom, 2002.

[23] I. Rodomagoulakis, G. Potamianos, and P. Maragos, "Advances in large vocabulary continuous speech recognition in Greek: Modeling and nonlinear features," in *Proc. EUSIPCO*, 2013.