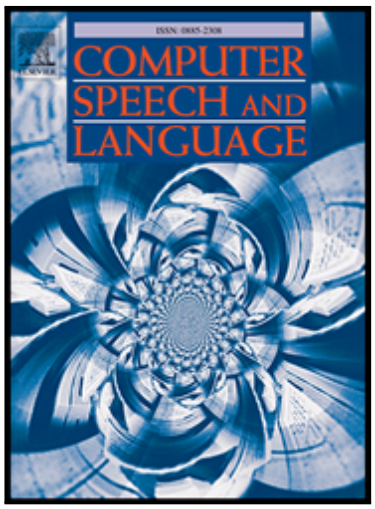


Room-Localized Spoken Command Recognition in Multi-Room, Multi-Microphone Environments

Isidoros Rodomagoulakis, Athanasios Katsamanis, Gerasimos Potamianos, Panagiotis Giannoulis, Antigoni Tsiami, Petros Maragos

PII: S0885-2308(16)30351-5
DOI: 10.1016/j.csl.2017.02.004
Reference: YCSLA 828

To appear in: *Computer Speech & Language*

Received date: 7 November 2016
Revised date: 6 February 2017
Accepted date: 10 February 2017

Please cite this article as: Isidoros Rodomagoulakis, Athanasios Katsamanis, Gerasimos Potamianos, Panagiotis Giannoulis, Antigoni Tsiami, Petros Maragos, Room-Localized Spoken Command Recognition in Multi-Room, Multi-Microphone Environments, *Computer Speech & Language* (2017), doi: 10.1016/j.csl.2017.02.004

Highlights

- Always-listening recognition pipeline for multi-room smart spaces.
- Room-localized operation based on multi-room speech activity detection.
- Channel selection and decision fusion approaches in all pipeline components.
- Robust acoustic modeling based on far-field data simulation and per-channel adaptation.
- Systematic pipeline evaluation and optimization on both simulated and real corpora.

Room-Localized Spoken Command Recognition in Multi-Room, Multi-Microphone Environments

Isidoros Rodomagoulakis[a,c,*], Athanasios Katsamanis[a,c], Gerasimos Potamianos[b,c], Panagiotis Giannoulis[a,c], Antigoni Tsiami[a,c], Petros Maragos[a,c]

[a]*School of Electrical and Computer Engineering, National Technical University of Athens, 15773, Athens, Greece.*
[b]*Department of Electrical and Computer Engineering, University of Thessaly, 38221, Volos, Greece.*
[c]*Athena Research and Innovation Center, 15125 Maroussi, Greece.*

Abstract

The paper focuses on the design of a practical system pipeline for always-listening, far-field spoken command recognition in everyday smart indoors environments that consist of multiple rooms equipped with sparsely distributed microphone arrays. Such environments, for example domestic and multi-room offices, present challenging acoustic scenes to state-of-the-art speech recognizers, especially under always-listening operation, due to low signal-to-noise ratios, frequent overlaps of target speech, acoustic events, and background noise, as well as inter-room interference and reverberation. In addition, recognition of target commands often needs to be accompanied by their spatial localization, at least at the room level, to account for users in different rooms, providing command disambiguation and room-localized feedback. To address the above requirements, the use of parallel recognition pipelines is proposed, one per room of interest. The approach is enabled by a room-dependent speech activity detection module that employs appropriate multichannel features to determine speech segments and their room of origin, feeding them to the corresponding room-dependent pipelines for further processing. These consist of the traditional cascade of far-field spoken command detection and recognition, the former based on the detection of "activating" key-phrases. Robustness to the challenging environments is pursued by a number of multichannel combination and acoustic modeling techniques, thoroughly investigated in the paper. In particular, channel selection, beamforming, and decision fusion of single-channel results are considered, with the latter performing best. Additional gains are observed, when the employed acoustic models are trained on appropriately simulated reverberant and noisy speech data, and are channel-adapted to the target environments. Further issues investigated concern the inter-dependencies of the various system components, demonstrating the superiority of joint optimization of the component tunable parameters over their separate or sequential optimization. The proposed approach is developed for the Greek language, exhibiting promising performance in real recordings in a four-room apartment, as well as a two-room office. For example, in the latter, a 76.6% command recognition accuracy is achieved on a speaker-independent test, employing a 180-sentence decoding grammar. This result represents a 46% relative improvement over conventional beamforming.

Keywords: smart homes, distant speech recognition, speech activity detection, keyword spotting, multichannel processing, decision fusion, beamforming, channel selection

☆This research was partially supported by EU project DIRHA, grant no. FP7-ICT-2011-7-288121.

*Corresponding author

Email addresses: irodoma@cs.ntua.gr (Isidoros Rodomagoulakis), nkatsam@cs.ntua.gr (Athanasios Katsamanis), gpotam@ieee.org (Gerasimos Potamianos), paniotiso@gmail.com (Panagiotis Giannoulis), antsiami@cs.ntua.gr (Antigoni Tsiami), maragos@cs.ntua.gr (Petros Maragos)

1. Introduction

Significant research effort has been devoted over the past decades to the design of Voice-enabled User Interfaces (VUIs) for natural, hands-free human-computer interaction. Such interfaces have typically been employed in interactive voice response systems at call centers and, more recently, in personal assistant applications on personal computers or smartphones (Schalkwyk et al., 2010). State-of-the-art developments in acoustic modeling for speech recognition (Hinton et al., 2012; Yu and Deng, 2015) have certainly contributed a lot to making VUIs practically usable in a variety of everyday environments; however, untethered, far-field, and always-listening operation, robust to noise, still constitutes a challenge that limits their universal applicability.

This challenge remains prominent in the very active research area of ambient assisted living inside smart homes, where, among others, VUIs are seen as crucial to the occupants' safety and well-being (Edwards and Grinter, 2001; Chan et al., 2008; Vacher et al., 2015). Indeed, domestic environments typically exhibit inter-room interference, frequent overlaps of various acoustic events and background noise with target speech, and moderate-to-high reverberation, when the acoustic scene is captured by far-field microphones, as is desired in an always-listening, untethered operation scenario. Similar conditions are present in additional everyday indoors environments, for example multi-room offices. Not surprisingly, Distant Speech Recognition (DSR) performance under such conditions lags dramatically compared to close-talking, noise-free scenarios (Kumatani et al., 2012).

A promising course for improving DSR in indoors environments is to exploit information from multiple audio channels, if such is available by distributed microphone arrays (Brandstein and Ward, 2001), located inside the smart space and providing sufficient spatio-temporal sampling of the acoustic scene. Such a solution has been investigated, for example, in the recent EU-funded project DIRHA[1]. The project focused on the design of a VUI for home automation, supporting distant speech interaction in different languages, targeting, in particular, people with kinetic disabilities. The basic use-case involved command-like voice-control of automated home equipment, for example of the room lights, temperature settings, door, window and shutter operation, etc. To enable hands-free operation, the VUI was designed to be always-listening, employing key-phrase based activation. Further, to achieve appropriate disambiguation of uttered commands, allow possible interaction with multiple users in different rooms, and provide localized feedback (VUI confirmation using room loudspeakers), room-level localization of the recognized commands was also performed. An example of the DIRHA challenging acoustic scene is depicted in Figure 1.

In this paper, we describe in detail the design of a robust multichannel distant speech processing pipeline, developed for the purposes of the aforementioned DIRHA domestic interaction scenario in the Greek language. The adopted methodology is rather general, being readily applicable to support VUIs in other everyday indoors multi-room environments equipped with multiple microphone sensors, such as smart offices, for example. The work deals with a wide range of challenging topics in the area of distant speech processing, where its contributions lie, namely addressing the following topics:

[1]DIRHA: Distant-speech Interaction for Robust Home Applications (`http://dirha.fbk.eu`).

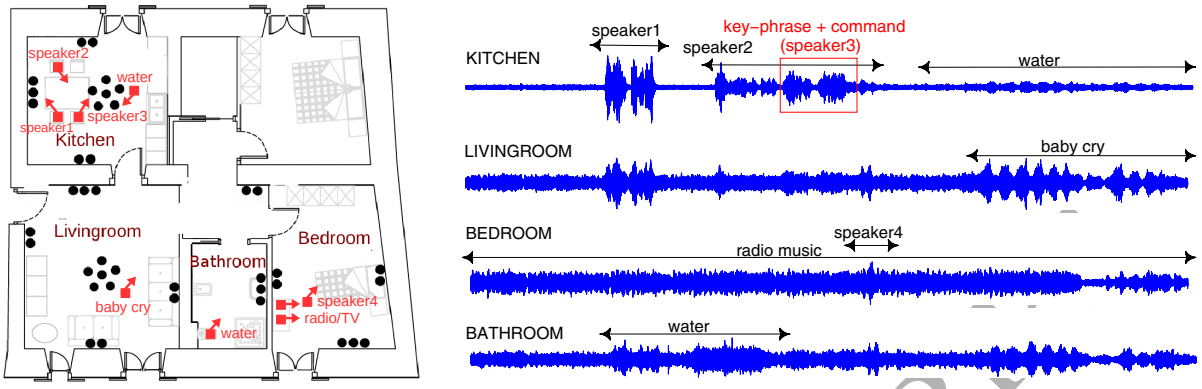


Figure 1: An example of a multi-room, multi-speaker acoustic scene considered in this paper, captured by a network of distributed microphones installed in the apartment and depicted as black dots in its floorplan (left). Four of the recorded signals are also shown (right), captured by the central microphones of the six-channel ceiling arrays (inside the Kitchen and Livingroom) and of the three-channel wall arrays (in the Bedroom and Bathroom). The goal in this example scene is to detect and recognize the command uttered by "speaker3" in the Kitchen (speech segment inside the red box), under the presence of other speech and non-speech events occurring in the various rooms (their time boundaries are annotated on the waveforms, and their source locations and directions are shown on the floorplan).

- Always-listening operation, achieved by employing Speech Activity Detection (SAD), key-phrase detection, and DSR.
- Room-localized operation, based on a multi-room SAD component used to drive separate, parallel cascades of key-phrase detection and DSR for each room of the smart space.
- Multichannel speech processing beyond beamforming, such as channel selection and decision fusion of single-channel results, considered in all pipeline components.
- Robust acoustic modeling, based on far-field data simulation and per-channel adaptation with little training data available in the target environment.
- Pipeline component optimization, studying component inter-dependencies and optimizing their tunable parameters separately, sequentially, or jointly.
- System and pipeline component evaluation on both simulated and real corpora in two multi-room, multichannel smart environments.

In more detail, to support *always-listening operation*, we build on the widely used cascade of three speech processing stages, as overviewed in Figure 2, namely: a) SAD, to separate speech from non-speech events; b) key-phrase detection, to identify a predefined system activation phrase; and c) DSR, to recognize the issued command. Combinations of some of the above components can be found in a variety of VUIs, providing partial robustness against non-speech events and increased efficiency, by processing only the speech segments of the incoming signals.

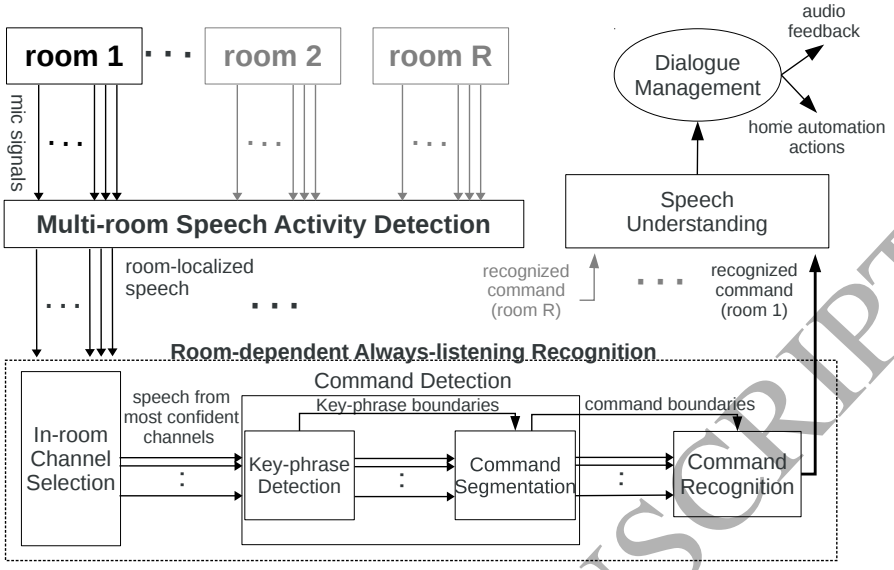


Figure 2: An overview of the proposed always-listening DSR system for smart environments that consist of R rooms equipped with multiple microphones. The system is parallelized into independent room-specific recognizers that perform command detection and recognition on room-localized speech segments produced by a multi-room SAD component. Multichannel processing is employed at all stages. The system is intended as input to a speech understanding and dialogue management component, as part of a voice interface (not addressed in this paper).

Further, to allow *room-localized operation*, we modify the aforementioned traditional cascade by designing a multi-room SAD component, instead of employing a generic, room-independent SAD. Such is able to identify speech segments in conjunction with their room of origin, robustly addressing the problem of inter-room interference. The component is used to drive separate cascades of key-phrase detection and DSR for each room of the smart space, operating in parallel. The process yields room-localized speech command recognition, as required by the VUI scenario considered in this paper.

To fulfill the needs of the detection and recognition tasks involved in the system, we elaborate and combine *multichannel speech processing* methods that have been explored in our previous preliminary studies (Giannoulis et al., 2015; Katsamanis et al., 2014; Tsiami et al., 2014a), achieving promising results and robustness in the challenging conditions considered. The implemented components make extensive use of channel selection and combination strategies to benefit from the available network of microphones inside the rooms. The advantage of these approaches is that they require no prior information regarding microphone network topology, other than mere room-microphone association. The proposed channel combination methods are based on decision fusion schemes, and they appear to outperform beamforming in most cases.

We gain additional benefits by employing *robust modeling*, in order to reduce mismatch between training and test conditions. In particular, we generate artificial training data simulating the test conditions, and, furthermore, we employ statistical model adaptation for each microphone channel, using few data from the target environment, if available.

Further, we consider *optimization* of a number of tunable system component parameters, while taking into consideration their inter-dependencies. Specifically, we observe that their joint optimization, rather than separate or sequential optimization, leads to improved command recognition accuracy.

Finally, we conduct *extensive experimentation* on both simulated and real datasets, where the adopted system architecture is evaluated systematically. For this purpose, we employ three separate databases: (a) DIRHA-sim, a corpus of simulated long audio recordings inside a real multi-room apartment (Cristoforetti et al., 2014); (b) ATHENA-real, a set of real recordings in a two-room office environment (Tsiami et al., 2014b); and (c) DIRHA-real, a corpus of real recordings captured inside the multi-room apartment also used for the first set. The first two consist of both development and test subsets, allowing for model adaptation and system optimization, while the third one is employed for testing the proposed pipeline on real data, unseen during its training. Reported results vary due to different characteristics and challenges of each dataset, reaching 76.6% in command recognition accuracy on the ATHENA-real corpus.

The rest of the paper is organized as follows: Section 2 overviews related work in the literature. Section 3 presents the proposed system, describes its components in detail, and reviews the adopted robust modeling and multichannel processing methods. Section 4 describes the databases used for the development and evaluation of the system pipeline. Section 5 introduces the adopted experimental framework and presents results of both the isolated components and the integrated system. Details on system optimization, final pipeline evaluation, and an error analysis are also included. Finally, Section 6 concludes the paper with a brief discussion.

2. Related work

Several projects and challenges have been launched over the last decade targeting intelligent interfaces for indoors smart environments and addressing DSR via multiple distributed microphones. Initially, the community focused on single-room setups for the analysis of lectures and meetings. Research projects like CHIL (Chu et al., 2006) and AMI (Hain et al., 2008) produced a wide range of results under the framework of the NIST Rich Transcription evaluation campaigns (Fiscus et al., 2008). Although meeting rooms are more controlled environments with fewer background noises compared to multi-room domestic environments, the largest portion of the corresponding acoustic scenes consisted of conversations between multiple speakers, and thus speech often overlapped. The focus there has mainly been on large-vocabulary continuous speech recognition in English, based on language models with dictionaries of approximately 50k words. A representative recognition result reported by Hain et al. (2012) on the AMI corpus was 33.2% Word Error Rate (WER) on non-overlapped speech, employing beamforming, speaker adaptation, and lattice rescoring methods. Further improvements were achieved in more recent works by Liu et al. (2014) and Renals and Swietojanski (2014), where beamforming was replaced by multichannel processing based on convolutional neural networks for training acoustic models on supervectors of concatenated single-channel features.

Moving from single-room to multi-room environments with more complex acoustic conditions, a hierarchical sound analysis system (Sehili et al., 2012) has been developed within the SWEET-HOME project (Vacher et al., 2011; 2015) for command recognition in French and de-

tection of distressed situations in apartments with elderly or impaired occupants. Its authors conducted a user evaluation on a small set of real recordings of regular and impaired speakers inside a four-room apartment with Signal-to-Noise Ratios (SNRs) within 15-25 dB. To recognize every speech instance, they used task-dependent language models of about 10k words, combining results from different rooms, but employing one microphone per room for cost efficiency. In most of the reported WERs (35% – 120%) there was a significant amount of insertions, caused by false detection of speech. In more recent work, though, Vacher et al. (2014b) presented improved recognition results using robust training and environmental adaptation, reducing command recognition error rate to 13%. This score was obtained after correcting misrecognized words at the syllable level, in order to match words of predefined commands. A similar system for voice command recognition and emergency detection inside smart homes was also proposed in the recent work of Principi et al. (2015), however its evaluation was conducted on single-room multichannel data (in both Italian and English). In other work, Morales-Cordovilla et al. (2014) employed beamforming to recognize room-localized commands of the DIRHA-GRID corpus (Matassoni et al., 2014), consisting of six-word English sequences simulated in a five-room environment with distributed microphone arrays. Although the acoustic scenes were simpler, without overlapped commands and background speech, the obtained WER of 39% showcased the degradation caused by factors such as background noise, interference across rooms, and reverberation.

In the context of robust speech technologies, the REVERB (Kinoshita et al., 2013), CHIME (Vincent et al., 2013), and ASpIRE (Harper, 2015) Challenges have been recently launched to provide a common evaluation framework concerning datasets, tasks, and evaluation metrics for a wide range of problems related to DSR in single-room noisy and reverberant environments with mismatch between training and testing conditions. In general, the approaches reported in the aforementioned campaigns can be grouped into two categories, namely a) robust modeling and b) multichannel processing (Delcroix et al., 2015). The former refer to data contamination and environmental adaptation methods. More specifically, in the absence of training data, the mismatch between complex acoustic environments and generic speech models can be reduced by artificially distorting training data (Matassoni et al., 2002; Ravanelli et al., 2012), and/or adapting to an available development set (Matassoni et al., 2002; Lecouteux et al., 2011). On the other hand, methods for channel selection and combination constitute multichannel processing approaches. Channel selection is based on channel confidence measures, mainly signal-based, such as SNR (Wölfel et al., 2006), or decoder-based (Wolf and Nadeu, 2014). Channel combination may be realized at the signal-level, e.g., by beamforming (Wölfel et al., 2006; Lecouteux et al., 2011), or at the decision-level employing techniques such as ROVER (Chu et al., 2006), SNR-weighted confusion-network based fusion (Wölfel et al., 2006), or the driven decoding algorithm of Lecouteux et al. (2011).

Research, development and evaluation of the involved multichannel processing modules in the recognition chain of an always-listening distant VUI depend on the existence of databases, either simulated or recorded in smart environments. Due to the complexity of the targeted acoustic scenes, collecting data in a realistic setup is demanding in terms of design, resources, and data annotation. A possible solution is the production of simulated data by convolving clean pre-recorded signals with estimated room impulse responses, and then mixing the signals to form sequences with overlaps and noise (Cristoforetti et al., 2014). Although simulated data are easier to produce for more controllable acoustic scenes, experimentation on real data is essential in order to evalu-

ate the system in real conditions. Regarding the number of rooms, most of the publicly available databases (Le Roux and Vincent, 2014) have been acquired in a single-room multi-microphone setup for meeting analysis (Janin et al., 2003; Mostefa et al., 2007; Carletta et al., 2006), acoustic event detection (Temko et al., 2007) and DSR (Bertin et al., 2016). A limited only number of corpora have been released for the case of home automation in multi-room setups. For instance, Vacher et al. (2014a) recorded speech in French by regular and impaired participants performing activities of daily living while interacting with a VUI through commands in a health smart home with four rooms equipped with two microphones. Another database was acquired in a similar health apartment by Fleury et al. (2013), giving emphasis on the task of distress situation detection via voice or other related acoustic events.

To our knowledge, concerning the language targeted in this paper, there exist few only works that address Greek for VUIs in smart environments. For example, Giannakopoulos et al. (2005) report preliminary results on distant command recognition for the control of home appliances using a similar pipeline to the one described in this work. The authors mainly focus on implementation issues of source localization and beamforming techniques in order to locate and enhance speech in a reverberant room, where the user walks and utters commands while engaged in conversation with other speakers in the same room. Although the task is challenging due to user motion and speech overlap, the reported task completion rates are above 80%. However, the experiments are restricted to three conversation scenarios, designed for a minimal set of 20 commands, in which the employed linear microphone array is steered to a specific area where the conversation is taking place, while the room is mainly quiet. Generally speaking, Automatic Speech Recognition (ASR) of Greek remains challenging due to the rich morphology of the language and the limited resources available for acoustic and language modeling (Gavrilidou et al., 2012). As a matter of fact, few only works in the literature address large-vocabulary Greek ASR. Indicative results are reported in the works of Digalakis et al. (2003) and Rodomagoulakis et al. (2013) for read newspaper articles, achieving WERs within the 11.5% – 21% range, and by Riedler and Katsikas (2007) and Dimitriadis et al. (2009) for the transcription of news broadcasts, with WERs close to 38%. Finally, multi-lingual acoustic modeling approaches are examined for Greek (together with other under-resourced languages) by Imseng et al. (2012).

3. Proposed multichannel, always-listening, distant speech recognition pipeline

As already outlined, the proposed speech processing pipeline aims at recognizing spoken commands for home and office automation. The user is potentially able to address the system from any position in the multi-room space. This is achieved by designing it to operate in parallelized room-dependent speech processing cascades, consisting of a) microphone selection, b) command detection, and c) command recognition, all driven by multi-room SAD that provides candidate speech segments for each room (see also Figure 2). Details of the system modules follow.

3.1. Multi-room speech activity detection

Detection of room-localized distant speech in multi-room environments presents several challenges, compared to traditional SAD approaches as applied to single-channel, single-space recordings, with interference across rooms causing additional significant difficulties. To solve this prob-

lem, the multi-room SAD approach of Giannoulis et al. (2015) is employed with slight modifications. There, two steps are followed: a) First, speech/non-speech segmentation is performed for the entire multi-room space using multi-stream speech/non-speech Gaussian Mixture Models (GMMs). b) Subsequently, the resulting speech segments are further processed to decide whether they occurred inside or outside a given room, by utilizing room-dependent Support Vector Machine (SVM) classifiers, trained on carefully crafted acoustic features that capture reverberation and attenuation effects in the microphone signals.

First step of multi-room SAD. In this paper, the first step of the aforementioned approach is modified to perform speech/non-speech segmentation for each room independently, using only the microphones located inside it. As a result, detected speech is more room-localized, facilitating effective inside/outside speech classification at the second step. In more detail, channel-dependent two-class (speech/non-speech) GMMs, consisting of 32 mixtures with diagonal covariances, are trained on the development set data of each channel (see Section 4 and Table 1 for details). A traditional acoustic front-end is used, based on 13-dimensional Mel-frequency Cepstral Coefficients (MFCCs) appended by their first- and second-order temporal derivatives, extracted every 10 *ms* over Hamming-windowed signal frames of 25 *ms* duration. A multichannel score for both speech and non-speech classes at a given frame and room is subsequently obtained, by summing the single-channel GMM log-likelihoods of all M_r microphones located inside room r (room index $r \in \{1, \ldots, R\}$, where R denotes the number of available rooms). These scores can be viewed as observation probabilities of a simple Hidden Markov Model (HMM) having a speech and a non-speech state. Viterbi decoding can then be applied to determine the most likely sequence of such states, yielding a speech/non-speech segmentation for each room.

During the decoding process, a Speech Prior Log Probability (SPP) can be added to the speech scores, in order to promote the occurrence of speech against non-speech, while transitions from speech to non-speech states and vice-versa can be reduced by a Speech/non-speech Insertion Penalty (SIP), enforcing temporal smoothness of the detected segments. Both parameters are tunable and can affect detection performance in terms of precision and recall. For example, if SPP increases, detection promotes speech classification, leading to higher recall performance.

Second step of multi-room SAD. At the second step of the approach, and given the detected speech segments derived from its first step, inside/outside-room classification decisions are made for each room, based on appropriately designed room-dependent SVMs. This step is also modified, compared to the earlier work of Giannoulis et al. (2015), to yield decisions every 100 *ms*, instead of the entire segment, thus allowing its breakup across rooms. The classification is performed using all available microphone data over longer windows of 600 *ms* in duration, shifted by the desired decision step of 100 *ms* at a time. Results are further refined by simple majority voting over all consecutive windows that partially overlap with the current 100 *ms* decision frame, with speech prevailing in case of a tie. In addition, post-processing is applied to the results, by merging nearby speech segments (lying closer than 0.7 seconds) and subsequently discarding any segments smaller than 0.2 seconds in duration.

The employed SVMs are trained on development set data (like the GMMs in the first step of this module), and they operate on multichannel features that are indicative of whether the candidate segment source lies inside or outside the room of interest. Feature design is based on the

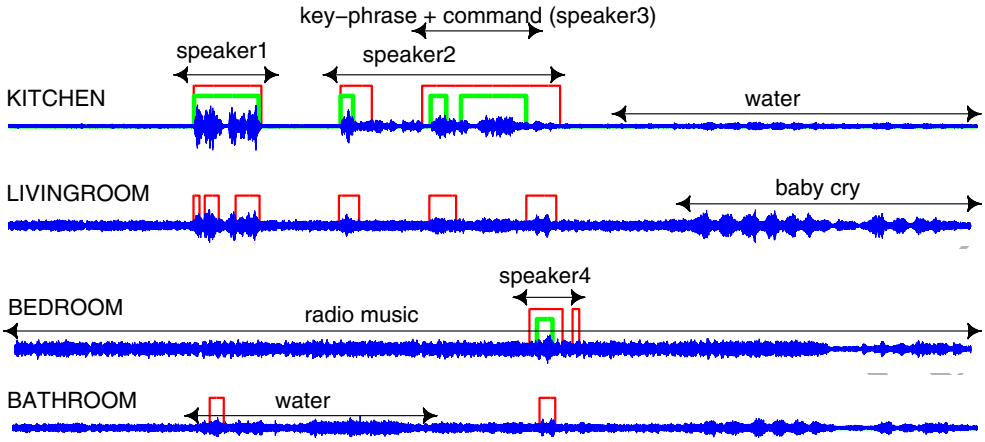


Figure 3: Example output of the two-step multi-room SAD algorithm detailed in Section 3.1, applied to the recordings of Figure 1. Speech segments resulting from the first step are depicted with red, thin rectangles. These are refined at the second step, as shown by the green, thick rectangles.

expectation that a speech signal recorded by the microphones of a given room exhibits lower energy and higher reverberation when produced outside the room compared to inside it. The feature set comprises of three measurements, as discussed next. Note that these are extracted for each of the R rooms, thus forming $3R$-dimensional feature vectors, over which the room-dependent SVMs operate.

K-best room SNR dominance: This feature is based on the assumption that microphone recordings inside the source room for a particular speech event will generally have higher SNRs than microphones in other rooms. Consequently, the K-best signal-to-noise energy ratios are computed over the current speech segment window (no logarithm is used in this calculation). The feature is then estimated as the sum of these quantities for microphones located inside the room of interest, minus the ones of microphones outside it.

Room microphone cross-correlation: This stems from the expectation that microphone recordings inside the source room will be less reverberant than those in other rooms, thus exhibiting higher pairwise cross-correlation on average (Morales-Cordovilla et al., 2014). To compute this feature, given a candidate speech segment and a room of interest, the maximum cross-correlation among all pairs of adjacent microphones inside the room is estimated, computed over 100 *ms* signal lengths. For improved robustness, consecutive such estimates are averaged with a 25 *ms* window shift within the examined 600 *ms* window.

Room envelope variance: Similarly to the above, this is based on the expectation that short-time signal energy will vary more (be less smooth) inside the source room compared to microphone recordings outside it, due to less reverberation of the former. Such effect is captured by the signal Envelop Variance (EV), further discussed in Section 3.2 (Wolf and Nadeu, 2014). To compute this feature, EVs are first estimated for each microphone channel located inside the room of interest, over the entire 600 *ms* window examined. The desired feature is then obtained as the maximum of the resulting EVs.

An example of the multi-room SAD module output, applied to the acoustic scene of Figure 1, is provided in Figure 3. It can be readily observed that, at its first step, the algorithm successfully overlooks non-speech acoustic activity such as water, baby cry, and radio music. Further, at the second step, it manages to exclude speech by "speaker4", located in the Bedroom, from the speech segments localized in the Kitchen.

3.2. In-room channel selection

The tasks of distant key-phrase detection and ASR are strongly affected by noise and reverberation. In scenarios where speech is captured by a network of distributed microphones, the degree of distortion may differ significantly among them. Channel/microphone selection aims at identifying a subset of them, considered as more reliable for further processing. The advantage of channel selection versus signal fusion/enhancement approaches based on beamforming lies on the fact that a good trade-off between recognition accuracy, latency, and computational cost can be accomplished, avoiding source localization or time-difference-of-arrival (TDOA) estimation that can be error-prone in challenging acoustic scenes.

Channel selection in the proposed pipeline is based on the EV measure, advocated by Wolf and Nadeu (2014), who showed its superiority in DSR over other signal-, statistical-, or model-based selection criteria. As also mentioned in Section 3.1, EV indicates how reverberant or, in general, distorted a channel is, by capturing the smoothness of short-time speech energy. It is estimated as the average of the variances of properly normalized and cube-root compressed energies, which are computed on 24 mel-spaced sub-bands, frame-by-frame (exactly as in MFCC feature extraction), and subjected to log-domain mean subtraction (over the segment) to remove short-term channel effects. To obtain reliable variance estimates, EV is typically calculated over longer windows (here, 400 *ms* in duration). Further, EV over a longer duration segment is obtained by shifting the window by 50 *ms* at a time and computing the average of the resulting EV sequence.

Based on the above, for a given speech segment detected inside room r by the multi-room SAD module, $\hat{M}_r$ channels are selected among the M_r available room microphones as the ones with the highest speech segment EVs. After experimenting with $\hat{M}_r$ values within the range $[2, \ldots, 6]$, based on the resulting performance of system modules that follow, the choices of $\hat{M}_r = 4$ for command detection (Section 3.3) and $\hat{M}_r = 3$ for command recognition (Section 3.4) are made. Of course, for rooms with fewer microphones, $\hat{M}_r = M_r$. Generally speaking, the choice of $\hat{M}_r$ depends on the microphone setup, e.g., larger values of $\hat{M}_r$ may add outlier microphones in the subsequent channel combination when the microphones arrays are placed sparsely in a room and their recognition results are expected to be quite different due to the localized interfering noises.

3.3. Command detection

The role of the proposed command detection component, following multi-room SAD and in-room channel selection, is two-fold: a) It first detects whether a key-phrase has been uttered within a given room-localized speech segment, and b) it specifies the temporal boundaries of the command that follows. Key-phrases are typically followed by commands, but the pause duration between them may vary. Given that other speech events can occur simultaneously, finding the exact

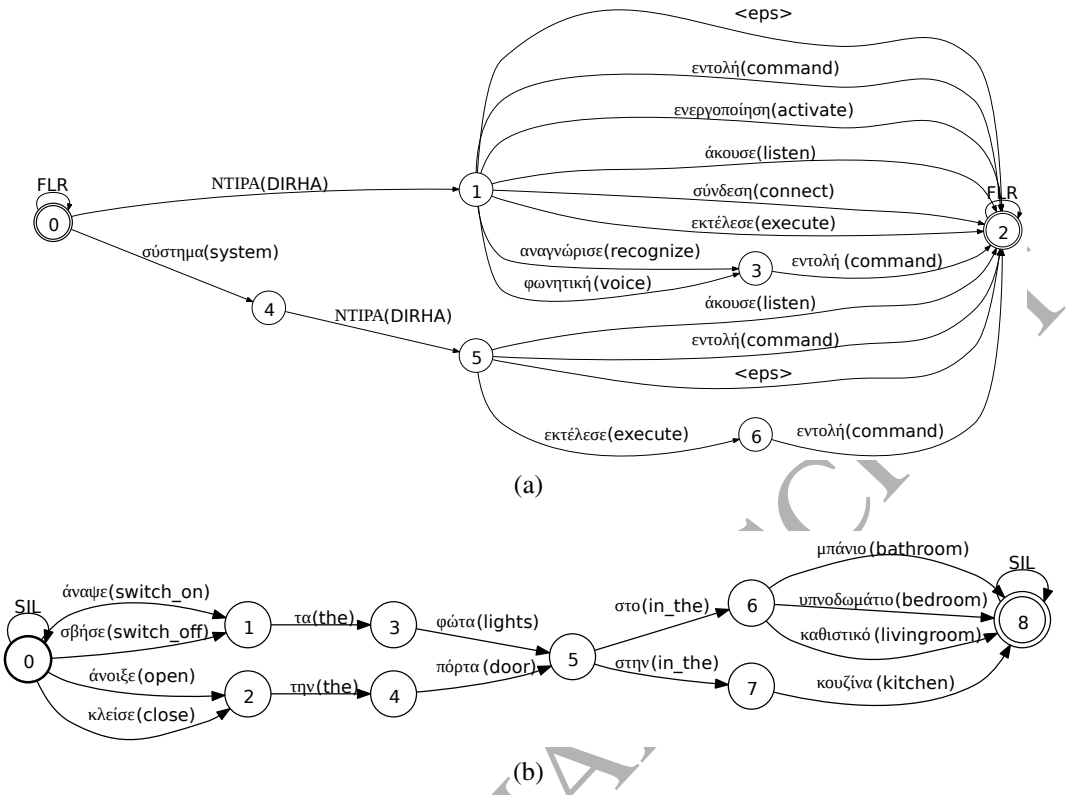


Figure 4: (a) Finite-State-Automaton (FSA) representation of the finite state grammar used for key-phrase detection of 12 possible Greek phrases, as discussed in the first part of Section 3.3. (b) FSA for parts of the finite state grammar used for command recognition, as discussed in Section 3.4, with 16 out of the 180 possible home automation commands depicted. English translations are also provided. The filler model is denoted by FLR and silence/non-speech as SIL. Double circles indicate final states, bold circles initial ones, and <eps> implies an empty transition.

start- and end-time of an uttered command can be challenging. To address this issue, a multichannel key-phrase detection scheme is introduced, followed by a rule-based command segmentation module. Details are provided next.

Key-phrase detection. The adopted methodology is based on the classical keyword-filler approach (Wilpon et al., 1990; Katsamanis et al., 2014). It employs whole-word HMMs for the words in the key-phrases and a separate HMM for general/irrelevant speech, known as the filler model. Following experimentation with the filler HMM topology, optimal detection is achieved when 24 states are used with left-to-right state transitions and observation probabilities consisting of 32-mixture GMMs with diagonal covariances, based on a standard MFCC-plus-derivatives front-end. In the absence of domain-specific training data, HMMs for the key-phrase words are constructed by concatenating sub-word models (tri-phones), pooled from the large-vocabulary continuous speech recognition system in Greek, built by Rodomagoulakis et al. (2013) on the "Logotypografia" corpus (Digalakis et al., 2003), as further discussed in Sections 3.4 and 4.4. A subset of the same database, consisting of 10 hours of speech, is also employed for filler HMM training. Additional model training and per-channel adaptation are also performed to better match the far-field condi-

Input : t_0 - detected key-phrase end-time,
t_1 - current SAD segment end-time,
t_2, t_3 - next SAD segment time boundaries
Parameters: d_{min}, d_{max} - minimum and maximum keyword-command distance,
l_{min}, l_{max} - minimum and maximum command duration
Output : command start- and end-times

if $t_0 + d_{min} + l_{min} \leq t_1$ **then**
  /* command expected in the same SAD segment as key-phrase */
  command-start = $t_0 + d_{min}$
  command-end = $\min(t_0 + d_{min} + l_{max},\ t_1)$
else if $t_2 - t_1 \leq d_{max}$ **and** $t_3 - t_2 \geq l_{min}$ **then**
  /* command is expected in the following SAD segment */
  command-start = t_2
  command-end = $\min(t_3,\ t_2 + l_{max})$
else
  no command found
end

Algorithm 1: Command segmentation algorithm.

tions, following the robust modeling steps detailed in Section 3.5. The employed keyword-filler approach is designed to detect a predetermined set of 12 short key-phrases in Greek for system activation, for example translating to "DIRHA activate", "DIRHA execute", and "DIRHA listen", among others. This is accomplished by grammar-based ASR employing the finite state grammar depicted in Figure 4(a). Viterbi decoding is further controlled by a tunable Filler Word Insertion Penalty (FWIP) that penalizes transitions between models.

Key-phrase recognition is implemented to generate results every 2 seconds, for a 3-second long sliding window inside a speech segment (as detected by SAD). A hypothesis test is performed over each such window, in which the log-likelihood score of the optimal model sequence is found, assuming that a key-phrase is present, and compared to the filler model score, assuming that no key-phrase is uttered. A key-phrase is detected for a particular channel if the resulting Log-Likelihood Difference (LLD) exceeds a threshold T, which is tunable to allow the desired balance between recall and precision. In case a key-phrase is detected in multiple windows within a given segment, the one scored with the maximum LLD value is kept.

The above algorithm is applied separately to each of the $\hat{M}_r$ channels that are selected based on the microphone EVs of the room-localized candidate segment, as discussed in Section 3.2. The resulting binary decisions (key-phrase presence or absense) can be easily combined via majority voting with equal weights among them, thus exploiting the available multichannel information. In this scheme, if at least half of the microphones agree on key-phrase presence inside the examined segment, command detection prevails. In such case, the detected key-phrase (content and temporal boundaries) of the most confident channel (that with the highest LLD) are kept.

Command segmentation. Given SAD output and the end-point of a detected activation key-phrase, the next module in the pipeline determines the accompanying command temporal boundaries as accurately as possible, providing input to the DSR component for command recognition. This is achieved using heuristics on keyword-command distance and expected command duration, assuming that commands are short speech segments appearing shortly after key-phrases. Command segments are thus expected inside the nearest speech segment for the particular room, following the key-phrase segment, or within the same segment where the key-phrase lies, if it is sufficiently long. Command duration must also not exceed a maximum. This rule-based approach depends on a set of four tunable parameters that correspond to the expected minimum and maximum values of the keyword-command distance (denoted by d_{min} and d_{max}, respectively) and command duration (denoted by l_{min} and l_{max}). The method is more formally described in Algorithm 1.

3.4. Command recognition

A multichannel speech recognition module is introduced as the last component of the proposed pipeline, designed to perform robust DSR, separately for each room. The module uses the $\hat{M}_r$ most confident channels of a given room r, as provided by the channel selection algorithm of Section 3.2. In particular, it first employs the most confident channel to generate a list of possible command hypotheses, and subsequently exploits the remaining $\hat{M}_r - 1$ channels to rescore this list and yield the recognition result. More formally, the algorithm consists of the following steps:

1. The $\hat{M}_r$ most confident microphones in terms of envelope variance (EV) are selected for a given room r into a sorted list $\{m_1, m_2, \ldots, m_{\hat{M}_r}\}$, where m_1 denotes the microphone with the highest EV over the speech segment where the command is detected.
2. A variation of the Viterbi algorithm (Chow and Schwartz, 1989) is applied on the recording of microphone m_1, returning an N-best list of hypotheses, denoted as $\{\mathcal{H}_j,\ j = 1, 2, \ldots, N\}$.
3. Each hypothesis $\mathcal{H}_j$ is rescored for each selected microphone. This is achieved by forced-alignment of $\mathcal{H}_j$ using the Viterbi algorithm on the corresponding microphone recording and employing microphone-specific acoustic models. Thus, best-path log-likelihood scores $\{c_{i,j}\}$ are obtained for each hypothesis $\mathcal{H}_j$ and microphone m_i, where $i = 1, \ldots, \hat{M}_r$, $j = 1, \ldots, N$.
4. The recognition result is hypothesis $\mathcal{H}_{\hat{j}}$, where

$$\hat{j} = \underset{j \in \{1, \ldots, N\}}{\arg\max} \sum_{i=1}^{\hat{M}_r} c_{i,j}\ , \tag{1}$$

namely the hypothesis with the highest combined score.

The optimal value for parameter N was searched over the range of $[2, \ldots, 6]$, and it was found that $N = 3$ performed best. A version of this algorithm was originally proposed for fusing heterogeneous speech recognition engines by Ostendorf et al. (1991) and then for the first time applied in the context of multichannel DSR in our previous work (Katsamanis et al., 2014). For the individual task of DSR, it was shown there to provide additional performance improvements compared to single-channel DSR based on just the most confident microphone. In the current work, we further investigate how channel combination behaves in the proposed integrated setup.

The employed speech recognition engine is grammar-based, with the grammar designed to include a pre-defined set of commands that cover a wide range of home automation tasks, for example, door/window/shutter opening/closing, light switching on/off, etc. The commands are 180 in total, possibly including two or three different wordings for the same task, and also specifying the room of interest, e.g., "in the Livingroom". An excerpt of the corresponding command grammar is depicted in Figure 4(b).

Regarding acoustic modeling, GMM-HMM cross-word tri-phone models are used, based on a standard MFCC-plus-derivatives front-end. The tri-phones have tied states and are approximately 8k in total, with 16 diagonal-covariance Gaussians per state. These are trained on 22.6 hours of clean recordings that are part of a subset of the "Logotypografia" corpus consisting of high-quality utterances recorded by a close-talk microphone. Similarly to key-phrase detection, additional model training and per-channel adaptation are performed to better match the far-field environment, as detailed in the robust acoustic modeling steps of Section 3.5. Finally, the Viterbi decoding stage of the recognizer is fine-tuned by properly adjusting the Word Insertion Penalty (WIP) parameter.

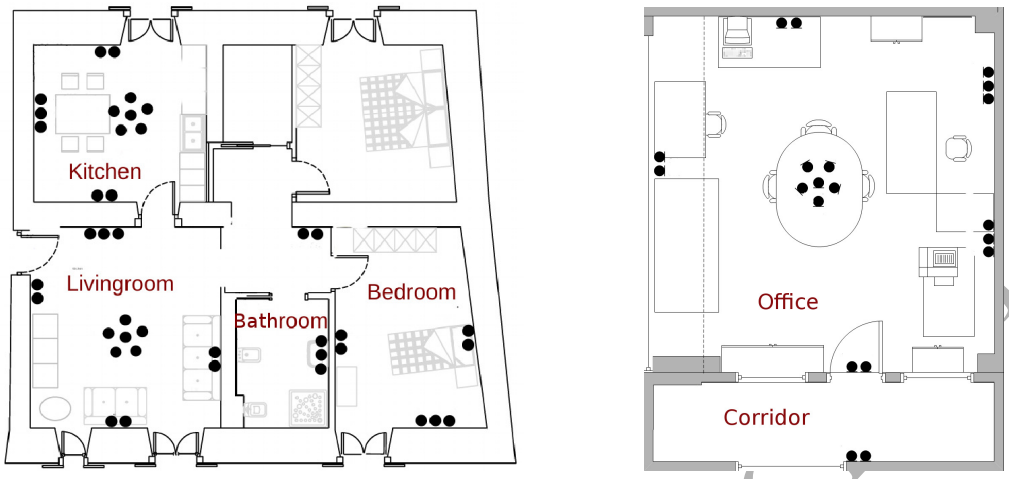
3.5. Robust acoustic modeling

To increase robustness and reduce mismatch with the acoustic conditions in the targeted multi-room environments, in addition to the clean acoustic models for key-phrase detection and command recognition, discussed in Sections 3.3 and 3.4, respectively, further model training strategies are pursued. In particular, HMMs for these two modules are also trained on artificially distorted data, following the same recipes as in the aforementioned sections, and they will be referred to in this paper as the "reverbed" acoustic models. For this purpose, data contamination is performed on available clean training data (discussed in Section 4.4), following the paradigm of Matassoni et al. (2002). The distortion/simulation process involves convolution of all utterances of the clean speech training corpus with Room Impulse Responses (RIRs) and addition of white Gaussian noise. The employed RIRs were measured in real environments (here, for the domestic DIRHA one) using the exponential sine sweep technique (Farina, 2000; Ravanelli et al., 2012). The exact number of RIRs employed is known to not affect ASR performance significantly (Ravanelli and Omologo, 2014).

The reverbed acoustic models are further adapted to the environment conditions employing Maximum Likelihood Linear Regression (MLLR) for additional performance gains. In our work, we only consider supervised adaptation using the development data of the available corpora (see Table 1), according to which few speakers (different to the test set ones) utter pre-defined commands or other phrases inside the multi-room space, to be used for offline transformation of the acoustic models. We apply MLLR separately for each microphone channel, ending up with channel-specific, environment-adapted (but not speaker-adapted) acoustic models. For comparison purposes, adaptation of the clean acoustic models is also considered in our experiments.

4. Simulated and real corpora for indoor automation

Three challenging multichannel datasets are employed for the development and evaluation of the proposed system: a) The DIRHA simulated corpus (DIRHA-sim), b) the DIRHA real corpus (DIRHA-real), and c) the ATHENA real database (ATHENA-real). All sets have been acquired in



(a) (b)

Figure 5: Floorplans of the two multi-room spaces considered in the paper: (a) The ITEA apartment is a multi-room home with 40 microphones and a surface area of approximately 50 m^2, used in the creation of the DIRHA-sim and DIRHA-real corpora. (b) The ATHENA office is a two-room space with 20 microphones and a surface area of about 35 m^2, used in the collection of the ATHENA-real corpus. Black dots in the plans represent microphones installed on the walls in pairs or triplets, or arranged in pentagon-shaped arrays on the ceiling. Inter-microphone distance is 30 cm for pairs and 15 cm for triplets, whereas, in the ceiling array, the peripheral-central microphone distance is 30 cm.

multi-room smart environments and include one-minute long recordings of a variety of commands and activation phrases in Greek, as well as non-speech events and background noises, deeming the recordings very realistic for always-listening, distant command recognition for home automation. The next paragraphs describe the corpora in more detail, and Table 1 summarizes them.

4.1. DIRHA-sim corpus

The DIRHA simulated corpus[2] (Cristoforetti et al., 2014) for Greek comprises simulated recordings of speech in the ITEA apartment that has been set up within the context of the DIRHA project at the Fondazione Bruno Kessler (FBK) in Trento, Italy. The apartment, as shown in Figure 5(a), is equipped with 40 microphones distributed in five rooms, either in linear arrays of 2-3 microphones, or in pentagon-shaped arrays of six microphones placed on the ceilings of the Kitchen and Livingroom that are considered as the most active rooms. Note that the apartment Corridor, although equipped with a pair of wall microphones, is not considered as an independent room. The corresponding recordings are therefore excluded in our experiments.

To create the simulations, high-quality speech (48 kHz, 16 bits PCM format, 50 dB SNR on average) were first captured in a sound-proof studio using a professional close-talk microphone. Twenty speakers (10 male, 10 female) were recorded, resulting to 1703 utterances containing approximately 140 minutes of various speech types, including phonetically rich sentences, read and spontaneous commands, system activation key-phrases, and conversational speech.

[2]The DIRHA simulated corpus is publicly available at `http://dirha.fbk.eu/simcorpora`

data characteristics	databases		
	DIRHA-sim	DIRHA-real	ATHENA-real
one-minute sessions (#)	150	60	240
rooms (#)	4	4	2
microphones (#)	40	40	20
subjects (#)	20	5	20
ages	25 – 50	25 – 55	18 – 55
total speech (min)	37	18	72
unique commands (#)	99	59	172
activation phrases (#)	12	12	12
background noises (#classes)	10	not transcribed	4
non-speech events (#classes)	73	not transcribed	15
avg SNR (dB)	13	15	9
avg T_{60} (sec)	0.72	0.72	0.50
overlapped speech (%)	47%	not transcribed	40%
close-talk mic available	no	no	yes
split into	dev, test	test-only	dev, test

Table 1: Overview of the corpora employed in this work. DIRHA-sim and ATHENA-real are used in the development and evaluation of the proposed system, whereas DIRHA-real for its evaluation only. Reported SNRs are average estimates over all speech segments from all the available microphones while reverberation times (T_{60}) are averaged over a number of RIRs in each environment, estimated by the method of Farina (2000). Speech is overlapped by speech and non-speech events while background noises are present constantly in most of the 350 one-minute sessions.

Subsequently, acoustic simulations were realized by convolving this material with more than 9k RIRs, estimated for each of the 40 microphones from 57 source locations, uniformly distributed inside the apartment and having 4-8 orientations each. Real, long-duration background noises and shorter acoustic events were also added in the simulations, for example music, various appliance sounds, drilling noises, water pouring, door knocking, etc., originating from randomly selected locations, or uniformly distributed in the apartment rooms, possibly concurrently. As a result, 150 one-minute long simulated recordings of speech and noise were created. For our experiments, half of these data, involving half of the speakers, are held out as a development set and the rest form the test set.

4.2. DIRHA-real corpus

The DIRHA-real database, presented in this paper for the first time, is a smaller set of real, instead of simulated, recordings acquired in the ITEA apartment. The environment, as well as the microphone configuration, is exactly the same as in the DIRHA-sim corpus. The data include five speakers, each recorded in 12 one-minute sessions, uttering phonetically rich sentences, commands preceded by system activation key-phrases, and in free conversation with a second speaker. Speaker positions are static, uniformly distributed across sessions inside four rooms of the apartment (4 sessions take place in the Livingroom, 4 in the Kitchen, 2 in the Bedroom, and 2 in the Bathroom). Various background noises and non-speech events also occur during the recordings, such as music, appliance sounds, and other typical home environment sounds. We use this corpus in our experiments as previously unseen, test data only.

4.3. ATHENA-real database

The ATHENA-real database[3] (Tsiami et al., 2014b) is a multimodal[4] database for home automation. It consists of 240 one-minute long sessions recorded in the two-room office environment depicted in Figure 5(b), where 20 microphones are installed, either in linear arrays of 2-3 microphones on the walls, or in a pentagon-shaped array of six microphones placed on the ceiling of the main office room. Additionally, head-mounted close-talk microphones are worn by the speakers to provide clean speech as reference for transcription and experimentation. Overall, the corpus contains data by 20 speakers, recorded while still or moving inside the two rooms, uttering phonetically rich sentences, system activation phrases followed by home automation commands, as well as in conversation with another speaker in some sessions. Most speech segments highly overlap with one of 15 acoustic events (e.g., opening/closing doors and windows) and four types of background noise, i.e., ambient office noise, vacuum cleaner, radio music, fan noise, thus rendering the database quite challenging and realistic. Similarly to the DIRHA-sim corpus, the data are split into a development and a test set in our experiments.

4.4. Additional speech material: Greek large vocabulary close-talk speech for acoustic modeling

In the absence of in-domain Greek speech material for acoustic modeling, we utilize the clean recordings of the "Logotypografia" database (Digalakis et al., 2003), collected for the development of Greek ASR. The corpus is akin to the Wall Street Journal task (Paul and Baker, 1991), consisting of 72 hours of large-vocabulary continuous speech of read newspaper text with 50k unique words, and containing a total of 125 speakers (55 male, 75 female) recorded in three environments (studio, office, and a quiet room) using two microphones (a head-mounted one and a desktop). A subset of this material, 22.6 hours in duration, recorded with the head-mounted microphone, is used to build clean acoustic models for key-phrase detection and command recognition, as described in Sections 3.3 and 3.4, respectively.

Furthermore, the high quality ($>$ 50 dB) utterances within this subset were contaminated to provide material for training reverbed acoustic models (see Section 3.5). In particular, distorted data were generated by convolving the available utterances with one of ten randomly selected source-microphone impulse responses, measured in the ITEA apartment environment, and also adding white Gaussian noise at a randomly chosen level among three possible ones.

5. Experimental framework and system evaluation

The design of the experimental framework for the development and evaluation of the presented system pipeline is complex due to the inter-dependency of the connected modules. To account for the behavior of each component individually and relatively to the others, we group experimental tasks into three categories, discussing details in the following subsections, mainly:

[3]The ATHENA-real database is available upon request at `http://cvsp.cs.ntua.gr/research/athenadb`.

[4]Kinect RGB-D data are also available, depicting the user activating the system by performing a gesture (raised hand in fist) in addition to the spoken key-phrase.

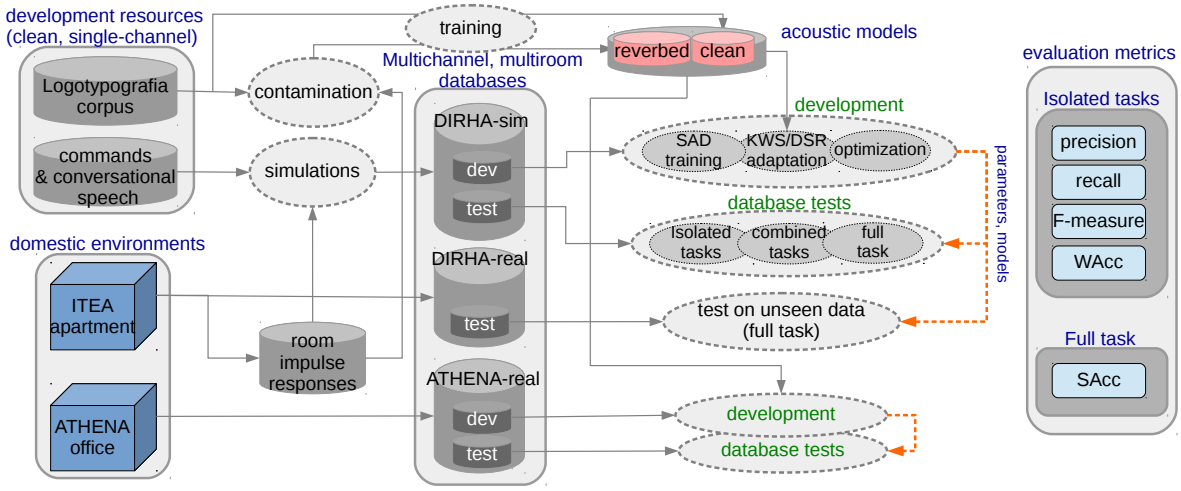


Figure 6: The experimental framework that is followed to develop and evaluate the proposed pipeline in simulated and real data that correspond to two smart environments: the ITEA apartment and the ATHENA office. Simulations of far-field speech are realized by convolving clean recordings with RIRs in order to produce simulated acoustic scenes (DIRHA-sim) for experimentation, and to contaminate large vocabulary clean speech (e.g., Logotypografia) for robust training, i.e., for training reverbed acoustic models. The individual components are trained/adapted on the "dev" sets of the DIRHA-sim and ATHENA-real databases where they are also optimized separately or jointly. Evaluation is performed on the corresponding "test" sets. The DIRHA-real corpus is used as an unseen set for the final evaluation of the system in terms of sentence accuracy (SAcc).

1. **Individual**: every module of the pipeline is tested separately in terms of standard evaluation metrics such as precision, recall, and F-measure for the detection tasks, and word accuracy for recognition, by assuming ground-truth inputs, e.g., the module of key-phrase detection is evaluated for segmented speech based on the annotated boundaries.
2. **Combined**: a pair of connected modules is tested together in order to assess their dependencies and to explore possible strategies for their joint optimization, while assuming ground-truth inputs from the preceding modules in the pipeline. For example, the performance of command recognition is examined in conjunction with that of command detection given ground-truth speech boundaries.
3. **Full**: the whole pipeline is tested when all its components are fully functional and no ground-truth information is provided.

An overview of the adopted experimental framework is summarized in Figure 6. Apart from the described processes of simulating and contaminating data for development and testing, another issue that is depicted in the diagram concerns the ability of the implemented pipeline to generalize and perform well on new data. For this purpose, we consider the DIRHA-real corpus as an unseen test set for evaluating the system that is developed on the DIRHA-sim development set. Although the two databases correspond to the same environment of the ITEA apartment, the cross-database experimentation shed light on the effectiveness of training on simulated data and then testing on real data. The proposed experimental setup targets the maximization of sentence accuracy (SAcc), which is the percentage of correctly recognized sentences (commands) penalized by the insertion

	parameters		
module	**symbol**	**description**	**operation ranges**
SAD	SPP	speech prior log probability	$[-3, -2.5, .., 3]$
	SIP	speech/non-speech insertion penalty	$[0, 10, .., 110]$
key-phrase detection	FWIP	filler/word insertion penalty	$[-300, -250, .., -100]$
	T	filler/word log-likelihood difference threshold	$[-3, -2.5, .., 3]$
command segmentation	l_{min}	min command duration	$[0.5, 1, .., 2.5]$
	l_{max}	max command duration	$[0.5, 1, .., 2.5]$
	d_{min}	min distance between key-phrase and command	$[0.5, 1, .., 2]$
	d_{max}	max distance between key-phrase and command	$[2, 2.5, .., 8]$
command recognition	WIP	word insertion penalty	$[0, 10, .., 50]$

Table 2: The proposed system is tunable by a set of nine parameters optimized for maximum back-to-back performance of the four modules in the pipeline.

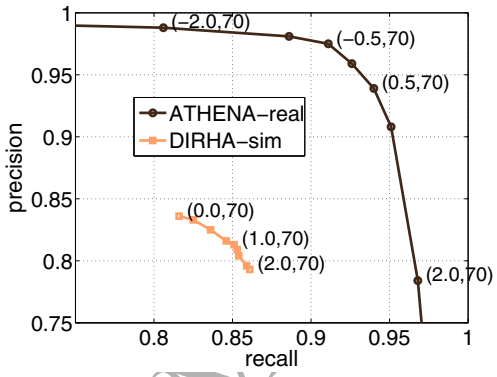


Figure 7: ROC curves showing the trade-off between precision and recall of the multi-room SAD as tuned by its parameters (SPP, SIP), whose values are indicatively shown next to the corresponding operating points. The operation is mostly affected by SPP which causes improved recall when increased, while SIP, which regulates the smoothness of the detection results, is less critical and thus kept constant across the depicted operating points.

rate of falsely detected commands, taking into consideration that an excessive number of false alarms will render the system unusable in practice. This measure reflects well the efficiency of the system in terms of user experience and is expected to provide an indication of the system behavior related to speech understanding and dialogue management (not considered in this paper). Optimization of the various tunable system parameters (summarized in Table 2) is discussed in detail in the following sections.

5.1. Evaluation of individual modules

5.1.1. Evaluation of multi-room speech activity detection

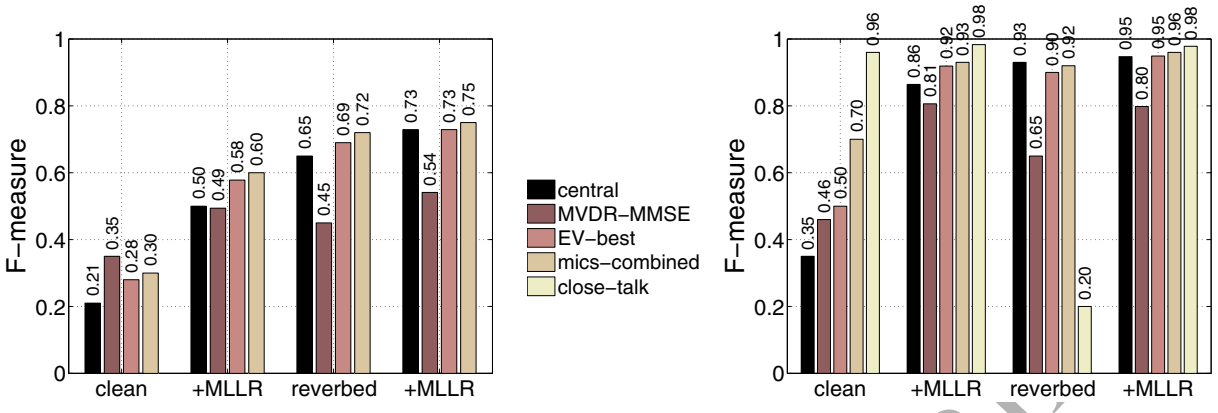
The multi-room SAD component is trained on the development sets and evaluated on the test sets of the DIRHA-sim and ATHENA-real databases in terms of precision, recall, and F-measure. Detection is evaluated for each room independently by framing the detected segments in non-overlapping frames of 10 *ms* and comparing them with the corresponding frames of the room localized speech/non-speech annotations. Average scores are taken over all the rooms of the examined multi-room environments. The Receiver Operating Curves (ROCs) shown in Figure 7 are

obtained by manipulating the SPP and SIP parameters as described in Section 3.1. The best F-measures on the DIRHA-sim and ATHENA-real databases are 0.83 and 0.95, respectively, indicating that the performance is almost excellent in the ATHENA office, but SAD remains challenging in the ITEA apartment, where many inter- and intra-room speech overlaps occur.

5.1.2. Evaluation of command detection and recognition

The individual tasks of command detection and recognition are evaluated assuming ground-truth room-localized time boundaries of the speech intervals and command sub-segments, respectively. The next paragraphs present results and comparisons showing the effectiveness of the adopted methods for robust modeling and multichannel processing. Command detection and recognition are realized using either the "EV-best" microphone (selected for each speech segment), the proposed channel combination aproach ("mics-combined"), or a state-of-the-art beamforming. For the latter, Minimum Variance Distortionless Response (MVDR) beamforming is employed, based on the work of Lefkimmiatis and Maragos (2007), where a single-channel Wiener post-filter is applied with weights estimated using Minimum Mean Square Error (MMSE). The necessary alignment of the beamformed channels is performed by employing TDOAs estimated by the speaker localization method described in the work of Tsiami et al. (2014a). Comparisons are conducted in the subset of sessions in which the user is located in rooms where pentagon-shaped arrays are installed approximately at the center of their ceilings and used for beamforming. Additionally, the performance of the central microphone of these arrays is presented ("central"), along with the performance of the close-talk microphone, which is available only in the ATHENA-real database. In order to demonstrate the effectiveness of robust training on contaminated data followed by environmental adaptation, experimentation is conducted using both clean and reverbed acoustic models, as well as their adapted versions. Note that the parameters of command detection (T, FWIP, d_{min}, d_{max}, l_{min}, l_{max}) and command recognition (WIP), summarized in Table 2, are optimized for each set of acoustic models and each database in terms of F-measure and word accuracy, respectively. Optimization has been held on the development sets of the two databases by using ground-truth speech boundaries, in a subset of sessions, in which the user was located in rooms with ceiling arrays, from which the central microphone was used for detection and recognition, respectively. More details follow.

Command detection results. As described in Section 3.3, command detection involves key-phrase detection followed by command segmentation. Starting with the evaluation of key-phrase detection, the corresponding F-measures are shown in Figure 8. It is evident that the reverbed models outperform the clean ones significantly, and that performance increases further when they are adapted to the actual environment. For example, in the case of the central microphone, the absolute improvement from the original clean to the adapted reverbed models is dramatic, averaged to 56% across the databases (from 0.21 and 0.35 to 0.73 and 0.96 in the DIRHA-sim and ATHENA-real databases, respectively). Moving from single- to multi-channel detection, the EV-best microphone yields further improvements compared to the central microphone, mainly in the case of using original clean models, in which the performance is absolutely increased by 7% and 15% in the two databases. The EV-best microphone is outpeformed by the proposed channel combination via majority voting by 2.3% and 6% on average, in the two databases, respectively. The proposed ap-



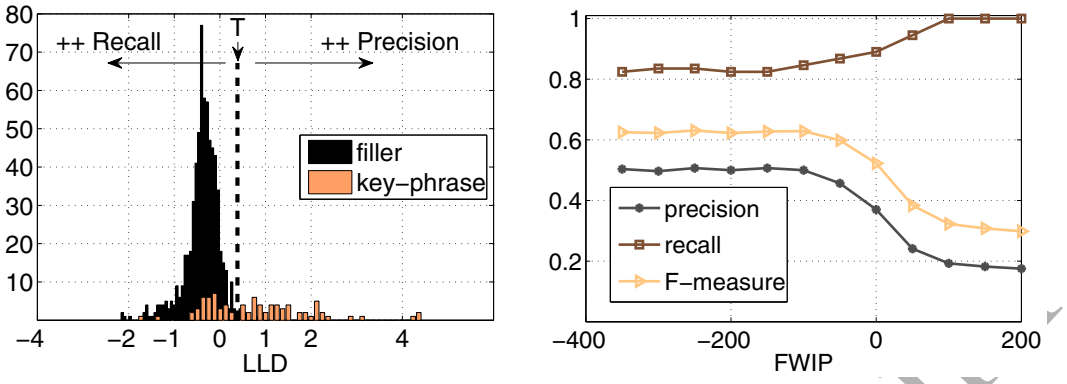
(a) DIRHA-sim (b) ATHENA-real

Figure 8: Key-phrase detection F-measures in the test speech segments using their ground-truth boundaries. Channel selection and the proposed channel combination are compared to MVDR-MMSE beamforming in sessions where the user is located in rooms with ceiling arrays. The performance of the central microphones in the corresponding rooms and the close-talk microphone (in the ATHENA-real database) are also reported for completeness.

proach achieves the best F-measures in the DIRHA-sim (0.75) and ATHENA-real (0.96) corpora with the latter being close to the F-measure achieved when using the close-talk microphone (0.98).

Regarding beamforming, the combination of MVDR-MMSE with the original clean models yields significant relative improvement of 67% (from 0.21 to 0.35) and 31% (from 0.35 to 0.46), against the central microphone in the two databases, respectively. However, beamforming results drop significantly when using reverbed models trained on contaminated signals, which are mismatched with the denoised beamforming signals. Although adaptation improves beamforming performance, overall, the best F-measure results obtained in the two databases (0.54 and 0.81) are significantly lower than the ones (0.75 and 0.96) corresponding to the proposed multichannel methods combined with robust modeling. The inferior performance of beamforming can be explained by several reasons. First, source localization errors caused by speaker movements and reverberation effects may affect the signal alignment stage. For example, an average F-measure increase of 8% (7.33% and 8.67% for adapted clean and reverbed models) is observed in the DIRHA-sim database when we use ground-truth instead of estimated locations. Secondly, post filtering appears to be beneficial only when using clean models. When using unadapted reverbed models, the F-measure is improved by 6% after removing the post filtering stage. Finally, note that the employed acoustic models are adapted to perfectly aligned beamformed signals based on the available ground-truth source locations. The performance may increase further if source localization errors are accounted in the adaptation process.

Figure 9 shows an example of how the T and FWIP parameters of key-phrase detection were optimized on the development set of the DIRHA-sim database. The parameters were swept over a range of values in order to maximize the F-measure. As the histogram of Figure 9(a) indicates, the selected value for threshold T is 0.15. Increasing or decreasing T favors precision or recall respectively. Accordingly, based on the curves of Figure 9(b), the F-measure is maximized over a wide range of FWIP values between -300 and -100 in the log-likelihood domain (the value of -250 is used).



(a) (b)

Figure 9: Optimizing the parameters of key-phrase detection in a subset of sessions in which the user was located in the Kitchen of the ITEA apartment. Detection was performed using the central microphone and the reverbed models for the recordings in the development set of the DIRHA-sim database. (a) Histograms of LLD values for key-phrase and filler segments demonstrating their discrimination by an appropriate T threshold. (b) Manipulating the filler/word insertion penalty (FWIP) included in Viterbi decoding for the estimation of the corresponding likelihood probabilities.

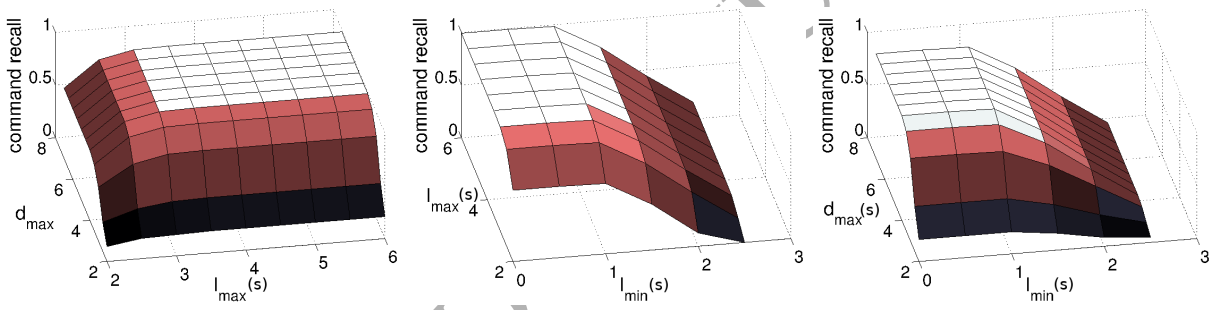
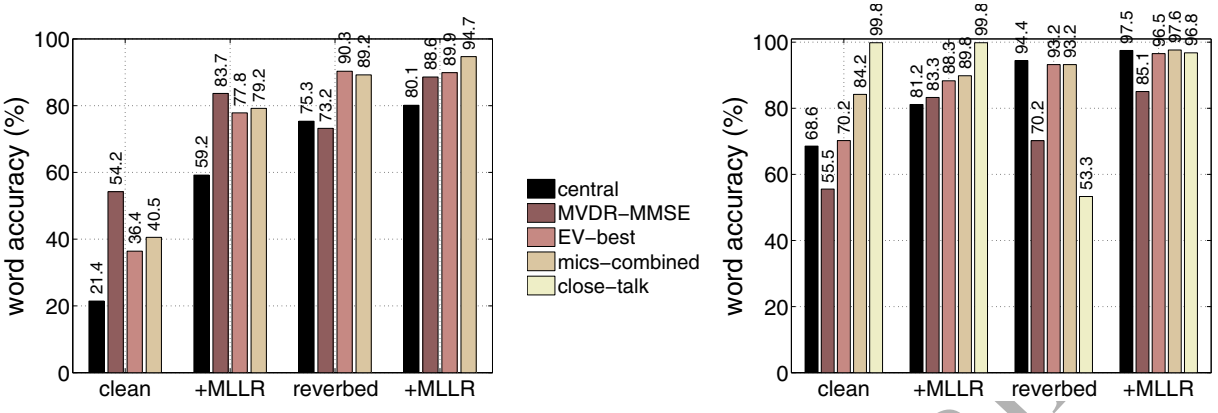


Figure 10: Optimizing the command segmentation parameters (d_{min}, d_{max}, l_{min}, l_{max}) of Algorithm 1 in the development set of the DIRHA-sim corpus in order to maximize command recall. For visualization purposes, the 4-D parameter space is projected onto 2-D ones, where the projected parameters are set to their optimal values. The presented pairs of parameters were found to affect performance the most. Each parameter search has a resolution of 0.5s.

The command segmentation stage is evaluated separately by assuming ground-truth inputs namely, room-localized speech boundaries and key-phrase end-points. We experiment with a variety of values in order to tune the four temporal parameters (d_{min}, d_{max}, l_{min}, l_{max}) of the algorithm. The detection criterion is segment-based: a command segment is considered as correctly detected if the estimated one covers 90% of its duration and the total distance between their boundaries is lower than 100 ms. The command recall on the DIRHA-sim corpus is 0.94 for the following parameter values (in seconds): $(d_{min}, d_{max}, l_{min}, l_{max}) = (0.5, 2.0, 2.0, 4.5)$. Accordingly, the obtained command recall on the ATHENA-real corpus is perfect for values $(0.5, 5.0, 2.0, 6.0)$. The optimal combination of parameter values is found by applying a greedy search in the four-dimensional parameter space, as depicted in the examples of Figure 10.

Command recognition results. The evaluation of command recognition is shown in Figure 11. Similarly to key-phrase detection, robust modeling appears to boost performance significantly



(a) DIRHA-sim (b) ATHENA-real

Figure 11: DSR word accuracy (%) on the test-set command segments of the DIRHA-sim and ATHENA-real databases assuming ground-truth boundaries. Channel selection and the proposed channel combination are compared to MVDR-MMSE beamforming for simulations in which the user is located in rooms with ceiling arrays. The performance of the central microphones in the corresponding rooms and the close-talk microphone (available on ATHENA-real data only) are also reported.

compared to the original clean models. For example, in the case of the central microphone, the absolute improvement from the original clean to the adapted reverbed models is also dramatic, averaged to 43.8% across the databases (from 21.4% and 68.6% to 94.7% and 97.6% in the DIRHA-sim and ATHENA-real databases, respectively). Further improvements are obtained by using channel selection and channel combination. Compared to the central microphone, the EV based channel selection and the proposed channel combination via N-best hypothesis rescoring improve the average word accuracy in most cases. For example, when using adapted reverbed models, the EV-best and mics-combined approaches outperform the central microphone by 9.8% and 14.6%, respectively, in the DIRHA-sim database. The corresponding improvements are moderate in the ATHENA-real database, where further analysis showed that the central microphone is very often chosen by the employed channel selection method as the most reliable, yielding similar results with the EV-best and mc-combined recognition approaches. Overall, similarly to the key-phrase detection results of Figure 8, channel combination combined with the MLLR-adapted reverbed acoustic models lead to the best recognition in the DIRHA-sim (94.7%) and ATHENA-real (97.6%) corpora, with the latter being close to the performance of the close-talk microphone (99.9%). Moreover, it is interesting to mention that the reverbed models exhibit cross-environment robustness. Although they have been trained on data produced in the ITEA apartment, they perform well on the ATHENA office data as well. It seems that the contamination process increases data variability and thus modeling becomes robust to mismatched conditions. Finally, we observe that recognition results in the DIRHA-sim corpus, using MVDR-MMSE beamforming with original and adapted clean models, show an absolute increase of 14% and 4% compared to the proposed channel combination approach. In this case, beamforming appears to be an effective solution for recognition if no contaminated data are available for training reverbed models. However, its maximum performance on both databases is significantly lower by 6% and 12.5%, compared to the proposed method. Note that the optimum values of the WIP parameter in the aforementioned

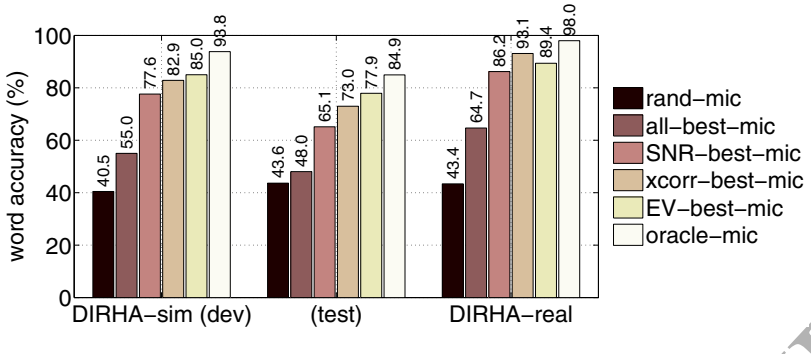


Figure 12: Comparison of signal-based channel selection methods for command recognition in all ITEA-apartment datasets, assuming ground-truth boundaries and using adapted clean models. Channel selection is conducted over all the available microphones in the apartment. Based on the EV, xcorr and SNR measures, the most confident microphones (EV-best-mic, xcorr-best-mic, SNR-best-mic) are selected from the 40 microphones in the entire apartment targeting command recognition independently of which room the source is located in. Random and oracle selection (the best microphone per segment in terms of recognition accuracy) are also reported for completeness, along with the microphone with the best overall performance (all-best-mic) selected a-posteriori for all sessions.

experiments were among [10, 20, 30].

Next, we compare various signal-based channel selection measures reported in the literature. We conduct the comparisons in all available datasets acquired in the environment of the ITEA apartment, where the variability across the microphones of the entire apartment is expected to provide insights regarding the examined channel selection methods. These datasets are the development and test sets of the DIRHA-sim database as well as the DIRHA-real dataset (see Table 1). The employed EV measure is compared to SNR and cross-correlation (xcorr) measures. The latter is an average of the maximum values of the cross-correlations between all possible pairs of signals recorded by neighboring microphones within an array. This measure gives an indication of how reverberant the acoustic signal that reaches the microphones of an array is, and may be used for array selection. Based on SNR and EV, selection is realized for every speech segment in order to obtain the EV-best and SNR-best microphones. Accordingly, based on xcorr, the central microphone of the most confident array is selected. The results of Figure 12 show that the EV-best microphone results in better recognition compared to the SNR-best and xcorr-best microphones. The absolute improvement has been, on average, measured to 7.7% and 1.2%, respectively, over the three employed test sets. Additional comparisons show that the EV-best microphone is by far better than a randomly selected microphone, although there is room for improvement in order to reach the "oracle" selection that results in the best possible microphone per segment in terms of recognition accuracy. Nevertheless, all the presented selection strategies achieve better performance in comparison to the best microphone ("best-mic"), selected a-posteriori and remaining the same for all sessions.

Additionally, to better understand the behavior of the various microphones in the ITEA apartment in relation to speaker location, we visualize the recognition results for the test-set simulations of the DIRHA-sim corpus, as shown in Figure 13. Each cell corresponds to the result of a specific microphone, for a specific simulation, using ground-truth command boundaries. Simulations are grouped by source location, e.g., simulations where the user is in the bathroom (BA) are repre-

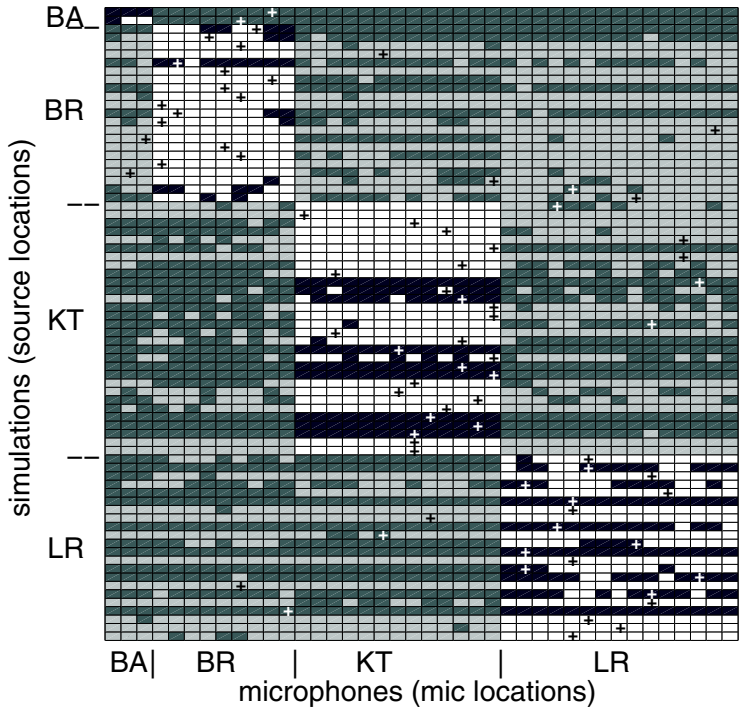


Figure 13: Performance analysis and channel selection over microphones located inside and outside the room where a command was uttered. Recognition results of commands with ground-truth boundaries are grouped in rooms and correspond to the test-set simulations of the DIRHA-sim corpus. The adapted reverbed acoustic models are used for recognition. Rows correspond to simulations and columns to microphones. From left to right, the microphones in the Bathroom (BA), Bedroom (BR), Kitchen (KT) and Livingroom (LR) are indexed. From top to bottom, the simulations where the system user is in the corresponding rooms are indexed. Black or dark cells reflect command recognition errors inside or outside the room, respectively. The EV-best microphone is also depicted for each simulation (once per row) by a cross ("+") sign, black when recognition is correct, white otherwise.

sented by the first two rows. Correct recognition of a command is depicted in the lightest colors (white/light blue) for microphones inside/outside the room where the user is located. Accordingly, black/dark blue cells indicate erroneous recognition. Overall, as expected, the microphones in the same room with the source perform significantly better. The probability of a microphone to recognize correctly a command uttered in the same room is measured to 0.7 compared to the probability of 0.55 corresponding to correct recognition by microphones outside the room. It also appears that the EV-best microphone is located in the same room with the source in approximately 75% of the cases. Both facts are evident by observing that the cells in the diagonal blocks are mainly white with more crosses than the cells on the off-diagonal blocks.

5.2. Evaluation of combined modules

Evaluation of the individual components shows that the employed methods may accomplish satisfactory levels of performance, when assuming ground-truth inputs. A question that arises is how these components can be fine-tuned in order to function effectively in the pipeline where errors are expected to propagate. To address this issue, we focus on testing pairs of successive

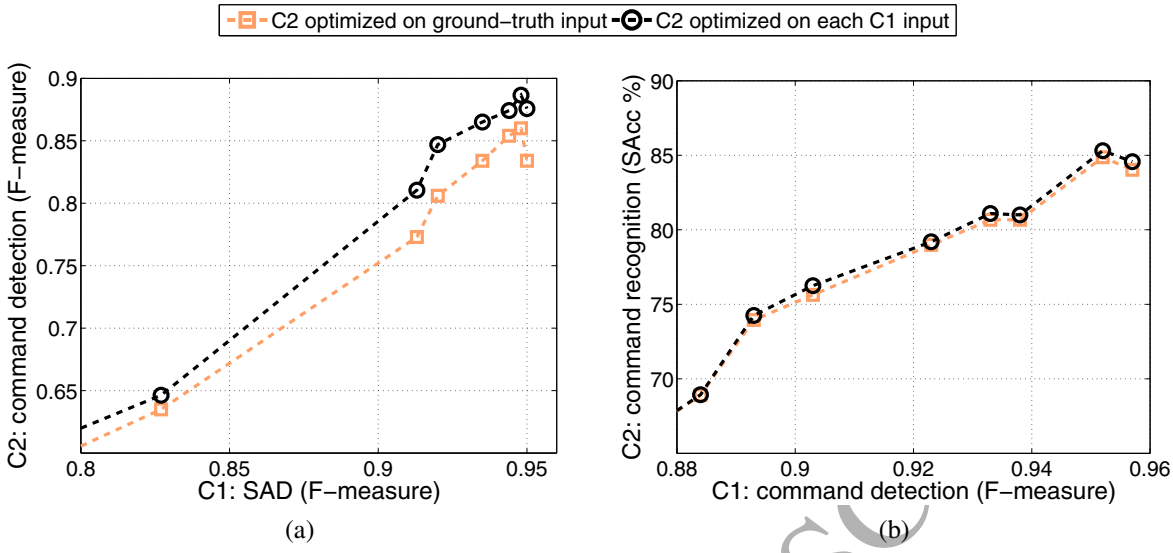


Figure 14: Investigating the inter-dependency of pairs of successive components of the proposed system pipeline: (a) command detection following SAD and (b) command recognition following command detection. In each plot, the tunable parameters of the C_2 component are either fixed to optimal values assuming ground-truth input (as in each isolated evaluation) or optimized to the input provided by the C_1 component with its parameters varying over a range. The depicted C_1 and C_2 component performance are reported on the test set of the ATHENA-real corpus, using the EV-best microphone and the adapted reverbed acoustic models (for command detection and recognition).

pipeline components operating back-to-back and given ground-truth input from their preceding components, if such exist. The goal is to find the best configuration of parameters for each examined pair in order to maximize the performance of the combined modules. At this stage of partial integration, two such pairs are considered, with results in Figure 14: (a) command detection, getting input from SAD and (b) command recognition, getting input from command detection, which is for the sake of this experiment, preceded by ground-truth SAD. For convenience, we denote the first component of each pair as C_1 and the second one as C_2. For each pair, the tunable parameters of the C_1 component are varied over a range of operating points. Subsequently, the parameters of the C_2 component are optimized based on ground-truth input or alternatively on the input provided by the C_1 component. The output of each component is evaluated using an appropriate metric, identical to the one defined previously in the evaluation of the isolated tasks. It is interesting to note that the maximum C_2 performance does not correspond to the best C_1 performance in terms of the employed metrics. For instance, in Figure 14(a), the maximum F-measures (0.86 and 0.89) of command detection (C_2), for the two considered optimization schemes, achieved with a SAD operating point yielding a lower F-measure (0.86) compared to its maximum (0.95). However, by optimizing the C_2 parameters based on real inputs from the C_1 components, the maximum F-measure of command detection is increased by 4% (from 0.86% to 0.89%) for the SAD-command detection pair, while the maximum SAcc in command recognition is increased by 1.5% (from 84.5% to 86.0%) for the command detection-command recognition pair. Further investigation of optimization strategies is presented in the next section, where the full integrated pipeline is evaluated.

5.3. Evaluation of full system pipeline

The final stage of the presented bottom-up experimental framework involves the evaluation of the full pipeline functioning without any ground-truth knowledge. First, we compare baseline systems in the test sets of the DIRHA-sim and ATHENA-real databases, and, subsequently, we present the performance of the proposed system compared to a baseline system. The final results are reported on all available databases, including also the unseen data of the DIRHA-real dataset. Note that all considered acoustic models in these experiments, referred to in the following text, are MLLR-adapted.

The first set of results is presented in Figure 15(a) in order to show the effectiveness of the EV-best microphone against MVDR-MMSE beamforming when employed at the stages of key-phrase spotting and command recognition, in the easier-to-implement approach of simply using clean acoustic models. Clean models are expected to be more matched with the denoised beamformed signals. The comparison takes place in the subset of sessions where the target command is localized in rooms with pentagon-shaped ceiling arrays (ITEA Livingroom and Kitchen, ATHENA office), where beamforming is expected to be more beneficial. Combined with adapted clean acoustic models, the EV-best microphone outperforms beamforming by 5.9% and 19.3% on the DIRHA-sim and ATHENA-real corpora. The performance is also better by 5.8% and 12.6% respectively, compared to the central microphone of the ceiling arrays. Due to its superiority against beamforming, which suffers by source localization errors and unwanted distortions caused in the post-filtering stage, we consider this system (EV-best microphone with adapted clean acoustic models) as the baseline. Additionally, there are practical reasons that strengthen this choice and make it an interesting alternative, compared to the proposed approach. Training clean acoustic models and adapting them in the target environment is easier than producing simulated data needed for the training of reverbed acoustic models. Additionally, channel selection is practically less time-consuming than the proposed majority voting and rescoring approaches for channel combination.

In the two baseline systems presented above, the parameters of each component have been optimized separately, based on component-specific metrics and given ground-truth inputs. This will be referred to as the S1 optimization scheme, and it corresponds to the most straightforward approach in which the modules are optimized individually before their combination, while no other fine-tuning is performed afterwards. As this may lead to suboptimal configurations due to the inter-component dependencies that are not taken into account, a second scheme, named by S2, is also considered, where each component is optimized given inputs from sequentially optimized preceding components. For example, the parameters of command recognition are optimized based on input from an optimized command detector, already optimized based on input from optimized SAD. Further, a third optimization scheme is considered, referred to as S3, that involves parameter tuning based on joint optimization of all components to maximize SAcc. The grid of parameters that is searched belongs to a nine-dimensional space (see Table 2), containing 1152 points produced by selecting at least two values for each of the nine system parameters based on the obtained results of the individual modules, as described in Section 5.1. We apply a brute-force parameter search by testing the grid points one by one in order to find the global maximum in terms of SAcc. Figure 15(b) shows how the performance of the baseline system improves when employing the reverbed models instead of the clean ones and then optimizing the system parameters using

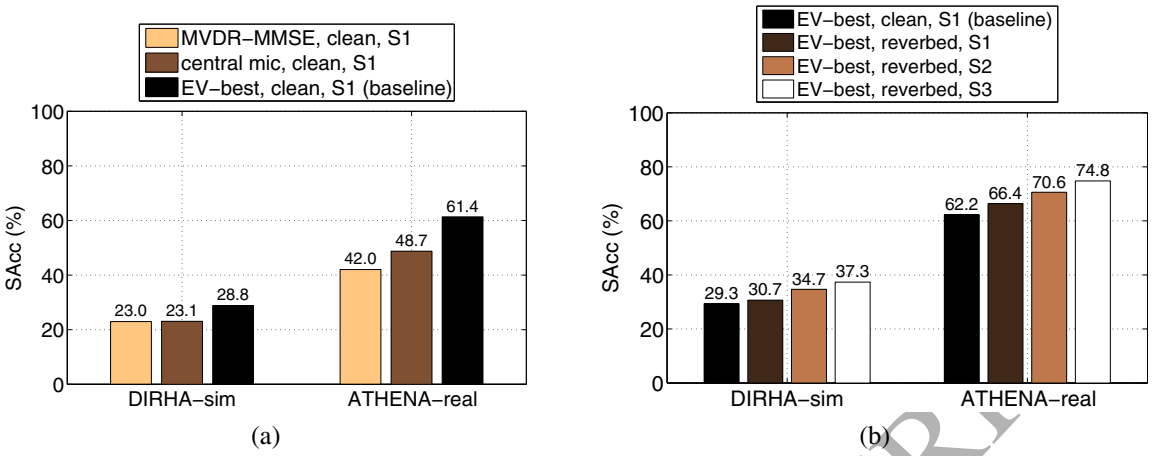


Figure 15: Baseline system and optimization results. (a) Comparison of the EV-best microphone (baseline) against MVDR-MMSE beamforming using the adapted clean models and individual optimization (S1) of the components. The comparison is conducted on a subset of sessions where the user is located in rooms with pentagon arrays installed at the center of their ceilings (ITEA Livingroom and Kitchen, ATHENA office). The performance of the central microphone of these arrays is reported as a single-channel recognition scenario. (b) Improving the baseline system by using the adapted reverbed models and two more optimization scenarios in which the pipeline components are optimized sequentially (S2) or jointly (S3). The results correspond to all sessions of the employed test sets.

schemes S2 and S3. The accomplished relative improvements in SAcc using S2 and S3 instead of S1 are 14% and 22% on the DIRHA-sim database, while on the ATHENA-real database are 6% and 13%. Note that the employed optimization schemes have been conducted on the development sets of the corresponding databases.

Figure 16(a) shows the performance of the proposed pipeline, including the S3 optimization scenario for all datasets. When comparing channel-combination vs. just using the EV-best setup, we obtain an absolute improvement of 1.3% and 2% for DIRHA-sim and ATHENA-real data, respectively. On both databases, the system has been optimized on their corresponding development subsets. On the other hand, the DIRHA-real dataset is treated as an unseen test dataset for which the DIRHA-sim optimized parameterization is applied. As mentioned before, although the former consists of real recordings in contrast with the latter which is simulated, the two databases have been acquired in the same apartment. Consequently, by sharing the models and the optimized parameters from simulated to real data, we are able to test the effectiveness of the simulation process for training and optimization. The obtained SAcc is 60% on the real unseen data of the DIRHA-real dataset, while the corresponding results on the DIRHA-sim and ATHENA-real databases are 38.7% and 76.6%, respectively. The proposed system outperforms the baseline by 15%, 11.2%, and 14.4% in the three databases, respectively, yielding a significant absolute improvement of 14% on average.

Additionally, based on the results of Figure 16(b) that correspond to a hypothetical scenario of a system operating with oracle components that give perfect results, the most significant degradation on the DIRHA-real dataset appears to occur at the command detection stage. The poor performance on the DIRHA-sim corpus can be explained by the fact that the simulated conditions are extremely challenging, presenting a variety of noises and speech overlaps that occur in the

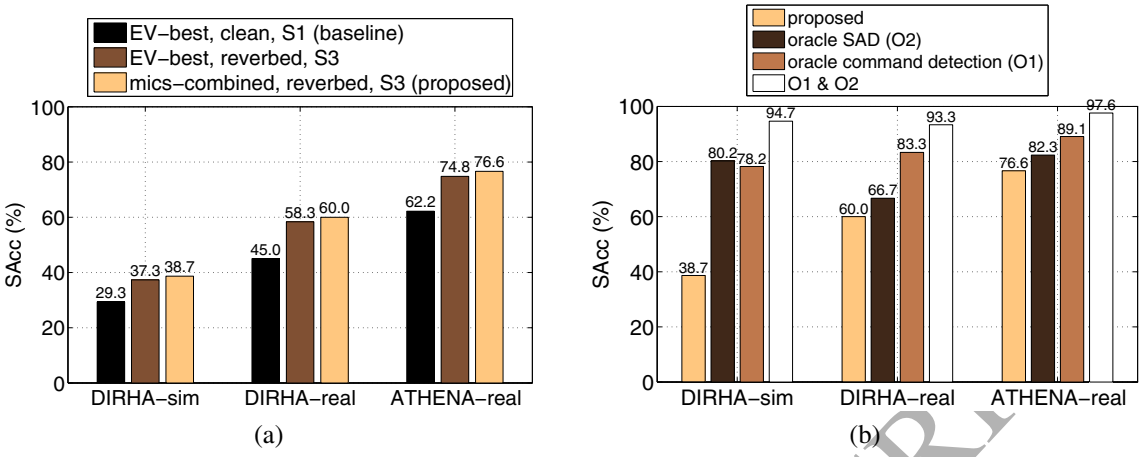


Figure 16: Proposed and oracle system results. (a) The proposed system, in which channel combination is employed at the stages of command detection and recognition, is compared to the baseline, in which the EV-best microphone is used, combined with clean and reverbed models. (b) Comparison with a hypothetical system in which certain components (SAD and/or command detection) of the proposed system are replaced with oracle ones that process ideally the inputs given by their preceding components.

same room where the user is located and cannot be fully resolved by the current pipeline design. Indicatively, a significant absolute improvement of 42% would be achieved if the current SAD module was substituted by an oracle providing ground-truth key-phrase plus command segment boundaries. On the other hand, the obtained SAcc of 76.6% on the ATHENA-real database is closer to the one yielded by systems $O1$ and $O2$ (82.3% and 89%), with oracle SAD and command detection, respectively. Compared to the results on DIRHA-sim data, the performance on ATHENA-real data is better mainly due to the absence of strong overlaps in speech segments.

Further error analysis of the obtained SAcc (60%) on the DIRHA-real dataset shows that: a) the percentage of correct sentences is 63.33%, b) the insertion rate of falsely detected commands is relatively small (3.33%) and c) the recognition word error rate of correctly detected commands is 4.33%. It is worth noting that a large portion of misrecognized words is due to confusion between synonyms. For example, recognition may incorrectly produce “shut the door” instead of “close the door”. In such cases, the meaning of the uttered and recognized commands is the same, and such errors will not be detrimental for speech understanding and dialogue management.

6. Conclusions, discussion and future work

In this work, we detail the design, optimization and systematic evaluation of a speech processing and recognition pipeline for an always-listening voice enabled user interface in Greek. The pipeline aims at robust far-field spoken command recognition in challenging multi-room smart environments as homes and offices equipped with sparsely distributed microphone arrays. The proposed system architecture is based on the synergy between multichannel speech activity detection, key-phrase detection, and automatic speech recognition building on a channel selection and decision fusion scheme to benefit from a distributed network of microphones inside the rooms.

The systematic evaluation of the developed system is based on a bottom-up experimental

framework, from the individual components to the complete integration using both simulated and real data, offering valuable insight regarding the behavior of the integrated components and their dependencies. The results show that overall, the proposed design constitutes a robust solution for always-listening distant speech recognition. The applied channel selection approach to the tasks of command detection and recognition on the ATHENA-real database yields 46% relative improvement in sentence accuracy compared to a conventional solution of beamforming, while the proposed channel combination approaches further increase the absolute system performance by 1.8%. Regarding acoustic modeling, data contamination in simulated conditions similar to those of the testing environments, leads to a relative improvement up to 36% compared to clean models. Finally, it is found that sequential and joint optimization of the pipeline components yields up to 14% and 22% relative improvement in sentence accuracy over isolated component optimization.

The proposed system achieves promising command recognition results in the two corpora with real recordings, i.e., the two-room ATHENA-real and the multi-room DIRHA-real corpora, reaching sentence accuracy scores of 76.6% and 60% respectively with the latter being representative of the system performance in real unseen data. On the other hand, the moderate performance of 38.7% on the simulated corpus DIRHA-sim can be explained mainly due to the high simulated noise contaminating the recordings, but also due to the appearance of speech overlaps occurring either across rooms or even in the same room. Inter-room overlaps may be resolved by using room selection but intra-room speech overlaps may be unsolved based on the current design of the pipeline. As a result, such overlaps may affect both the envelope variance estimation in channel selection and also the rule-based command detection that depends on the speech boundaries that speech activity detection provides. It is worth noting that when command detection and recognition are fed with exact speech segments for each speaker, the performance is significantly improved to 80% for the simulated DIRHA dataset.

Several directions of improvement may be followed based on the presented insightful results of this work. To name a few, speech activity detection may be benefited by incorporating speaker diarization and sourse separation methods in order to cope with intra-room speech overlaps, resulting to finer speech segmentation. Additional gains are feasible in speaker localization, as well as in command detection and recognition tasks, by capturing and processing multimodal information from the targeted noisy scenes using cameras and other sensors. For example, detection of audio-gestural activation key-phrases (as included in the ATHENA-real database) may be helpful when speech is noisy and overlapped. Last but not least, the exploitation and integration of the promising DNN-based approaches for speech still constitute an open field for research which may boost significantly the performance of such systems. Besides, the opportunities are increased due to the upcoming, growing market of commercial interfaces for home automation such as the Amazon Alexa and Google Home products, which establish speech technologies for everyday living and calls for further research. To conclude, a simplified version of the presented system has been implemented, performing online, always-listening command recognition in real time. Details, demos, and code are provided by Tsiami et al. (2016).

References

Bertin, N., Camberlein, E., Vincent, E., Lebarbenchon, R., Peillon, S., Lamandé, É., Sivasankaran, S., Bimbot, F., Illina, I., Tom, A., et al., 2016. A french corpus for distant-microphone speech processing in real homes. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech).

Brandstein, M., Ward, D., 2001. Microphone arrays: signal processing techniques and applications. Springer.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al., 2006. The AMI meeting corpus: A pre-announcement. In: Machine Learning for Multimodal Interaction. Vol. LNCS-3869. Springer, pp. 28–39.

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

Chow, Y. L., Schwartz, R., 1989. The N-best algorithm: An efficient procedure for finding top N sentence hypotheses. In: Proc. of the ACM Workshop on Speech and Natural Language. pp. 199–202.

Chu, S., Marcheret, E., Potamianos, G., 2006. Automatic speech recognition and speech activity detection in the CHIL smart room. In: Machine Learning for Multimodal Interaction. Vol. LNCS-3869. Springer, pp. 332–343.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M., Maragos, P., 2014. The DIRHA simulated corpus. In: Proc. Int. Conf. on Language Resources and Evaluation (LREC). pp. 2629–2634.

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., Nakatani, T., 2015. Strategies for distant speech recognition in reverberant environments. EURASIP Journal on Advances in Signal Processing 2015 (1).

Digalakis, V., Oikonomidis, D., Pratsolis, D., Tsourakis, N., Vosnidis, C., Chatzichrisafis, N., Diakoloukas, V., 2003. Large vocabulary continuous speech recognition in Greek: Corpus and an automatic dictation system. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech). pp. 1565–1568.

Dimitriadis, D., Metallinou, A., Konstantinou, I., Goumas, G., Maragos, P., Koziris, N., 2009. GridNews: A distributed automatic Greek broadcast transcription system. In: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP). pp. 1917–1920.

Edwards, W. K., Grinter, R. E., 2001. At home with ubiquitous computing: seven challenges. In: Ubicomp 2001: Ubiquitous Computing. Vol. LNCS-2201. Springer, pp. 256–272.

Farina, A., 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: Audio Engineering Society Convention 108.

Fiscus, J. G., Ajot, J., Garofolo, J. S., 2008. The Rich Transcription 2007 meeting recognition evaluation. In: Multimodal Technologies for Perception of Humans. Vol. LNCS-4625. Springer, pp. 373–389.

Fleury, A., Vacher, M., Portet, F., Chahuara, P., Noury, N., 2013. A French corpus of audio and multimodal interactions in a health smart home. Journal on Multimodal User Interfaces 7 (1-2), 93–109.

Gavrilidou, M., Koutsombogera, M., Patrikakos, A., Piperidis, S., 2012. The Greek language in the digital age. In: Rehm, G., Uszkoreit, H. (Eds.), Meta-Net, White Paper Series. Springer.

Giannakopoulos, T., Tatlas, N., Ganchev, T., Potamitis, I., 2005. A practical, real-time speech-driven home automation front-end. IEEE Trans. on Consumer Electronics 51 (2), 514–523.

Giannoulis, P., Brutti, A., Matassoni, M., Abad, A., Katsamanis, A., Matos, M., Potamianos, G., Maragos, P., 2015. Multi-room speech activity detection using a distributed microphone network in domestic environments. In: Proc. European Signal Processing Conf. (EUSIPCO). pp. 1271–1275.

Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., van Leeuwen, D., Lincoln, M., Wan, V., 2008. The 2007 AMI(DA) system for meeting transcription. In: Multimodal Technologies for Perception of Humans. Vol. LNCS-4625. Springer, pp. 414–428.

Hain, T., Burget, L., Dines, J., Garner, P. N., Grezl, F., Hannani, A. E., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2012. Transcribing meetings with the AMIDA systems. IEEE Trans. on Audio, Speech, and Language Processing 20 (2), 486–498.

Harper, M., 2015. The automatic speech recognition in reverberant environments (ASpIRE) challenge. In: Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU). pp. 547–554.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.

Imseng, D., Bourlard, H., Garner, P. N., 2012. Using KL-divergence and multilingual information to improve ASR for under-resourced languages. In: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP). pp. 4869–4872.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C., 2003. The ICSI meeting corpus. In: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP). pp. 364–367.

Katsamanis, A., Rodomagoulakis, I., Potamianos, G., Maragos, P., Tsiami, A., 2014. Robust far-field spoken command recognition for home automation combining adaptation and multichannel processing. In: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP). pp. 5547–5551.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., 2013. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 1–4.

Kumatani, K., McDonough, J., Raj, B., 2012. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. IEEE Signal Process. Mag. 29 (6), 127–140.

Le Roux, J., Vincent, E., 2014. A categorization of robust speech processing datasets. Tech. rep., Mitsubishi Electric Research Labs.

Lecouteux, B., Vacher, M., Portet, F., 2011. Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech). pp. 2273–2276.

Lefkimmiatis, S., Maragos, P., 2007. A generalized estimation approach for linear and nonlinear microphone array post-filters. Speech Communication 49 (7-8), 657–666.

Liu, Y., Zhang, P., Hain, T., 2014. Using neural network front-ends on far-field multiple microphones based speech recognition. In: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP). pp. 5542–5546.

Matassoni, M., Astudillo, R. F., Katsamanis, A., Ravanelli, M., 2014. The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech). pp. 1613–1617.

Matassoni, M., Omologo, M., Giuliani, D., Svaizer, P., 2002. Hidden Markov model training with contaminated speech material for distant-talking speech recognition. Computer Speech and Language 16 (2), 205–223.

Morales-Cordovilla, J. A., Pessentheiner, H., Hagmüller, M., Kubin, G., 2014. Room localization for distant speech recognition. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech). pp. 2450–2453.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., et al., 2007. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. Journal of Language Resources and Evaluation 41 (3-4), 389–407.

Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R. M., Rohlicek, J. R., 1991. Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses. In: Proc. Human Language Technology Conf. (HLT). pp. 83–87.

Paul, D. B., Baker, J. M., 1991. The design of the Wall Street Journal-based CSR corpus. In: Proc. Work. on Speech and Natural Language (HLT). pp. 357–362.

Principi, E., Squartini, S., Bonfigli, R., Ferroni, G., Piazza, F., 2015. An integrated system for voice command recognition and emergency detection based on audio signals. Journal of Expert Systems with Applications 42 (13), 5668–5683.

Ravanelli, M., Omologo, M., 2014. On the selection of the impulse responses for distant-speech recognition based on contaminated speech training. In: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP). pp. 1028–1032.

Ravanelli, M., Sosi, A., Svaizer, P., Omologo, M., 2012. Impulse response estimation for robust speech recognition in a reverberant environment. In: Proc. European Signal Processing Conf. (EUSIPCO). pp. 1668–1672.

Renals, S., Swietojanski, P., 2014. Neural networks for distant speech recognition. In: Proc. Hands-free Speech Communication and Microphone Arrays (HSCMA). pp. 172–176.

Riedler, J., Katsikas, S., 2007. Development of a modern Greek Broadcast-News corpus and speech recognition system. In: Proc. Nordic Conf. Computational Linguistics (NODALIDA). pp. 380–383.

Rodomagoulakis, I., Potamianos, G., Maragos, P., 2013. Advances in large vocabulary continuous speech recognition

in Greek: Modeling and nonlinear features. In: Proc. European Signal Processing Conf. (EUSIPCO). pp. 1–5.

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M., Strope, B., 2010. "Your word is my command": Google search by voice: A case study. In: Advances in Speech Recognition. Springer, pp. 61–90.

Sehili, M., Lecouteux, B., Vacher, M., Portet, F., Istrate, D., Dorizzi, B., Boudy, J., 2012. Sound environment analysis in smart home. In: Ambient Intelligence. Vol. LNCS-7683. Springer, pp. 208–223.

Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., Omologo, M., 2007. CLEAR evaluation of acoustic event detection and classification systems. In: Multimodal Technologies for Perception of Humans. Vol. LNCS-4112. Springer, pp. 311–322.

Tsiami, A., Katsamanis, A., Maragos, P., Potamianos, G., 2014a. Experiments in acoustic source localization using sparse arrays in adverse indoors environments. In: Proc. European Signal Processing Conf. (EUSIPCO). pp. 2390–2394.

Tsiami, A., Katsamanis, A., Rodomagoulakis, I., Potamianos, G., Maragos, P., 2016. Home sweet home... listen! : A distant speech recognition system for home automation commands. In: IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP) Show and Tell Demonstrations.
URL `http://cvsp.cs.ntua.gr/research/dirha`

Tsiami, A., Rodomagoulakis, I., Giannoulis, P., Katsamanis, A., Potamianos, G., Maragos, P., 2014b. ATHENA: A Greek multi-sensory database for home automation control. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech). pp. 1608–1612.

Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., Chahuara, P., 2015. Evaluation of a context-aware voice interface for ambient assisted living: Qualitative user study vs. quantitative system evaluation. ACM Trans. Access. Comput. 7 (2), 1–36.

Vacher, M., Istrate, D., Portet, F., Joubert, T., Chevalier, T., Smidtas, S., Meillon, B., Lecouteux, B., Sehili, M., Chahuara, P., et al., 2011. The SWEET-HOME project: Audio technology in smart homes to improve well-being and reliance. In: Proc. of Annual Int. Conf. of Engineering in Medicine and Biology Society (EMBC). pp. 5291–5294.

Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., Bonnefond, N., 2014a. The SWEET-HOME speech and multimodal corpus for home automation interaction. In: Proc. Int. Conf. on Language Resources and Evaluation (LREC). pp. 4499–4506.

Vacher, M., Lecouteux, B., Portet, F., 2014b. Multichannel automatic recognition of voice command in a multi-room smart home: An experiment involving seniors and users with visual impairment. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech). pp. 1008–1012.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., 2013. The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes. In: Proc. Automatic Speech Recognition and Understanding Work. (ASRU). pp. 162–167.

Wilpon, J., Rabiner, L. R., Lee, C.-H., Goldman, E. R., 1990. Automatic recognition of keywords in unconstrained speech using hidden Markov models. IEEE Trans. on Acoust., Speech, Signal Process. 38 (11), 1870–1878.

Wolf, M., Nadeu, C., 2014. Channel selection measures for multi-microphone speech recognition. Speech Communication 57, 170–180.

Wölfel, M., Fügen, C., Ikbal, S., McDonough, J. W., 2006. Multi-source far-distance microphone selection and combination for automatic transcription of lectures. In: Proc. Int. Conf. on Speech Communication and Technology (Interspeech). pp. 361–364.

Yu, D., Deng, L., 2015. Automatic Speech Recognition: A Deep Learning Approach. Springer-Verlag London.