


RESEARCH

Open Access

Room-localized speech activity detection in multi-microphone smart homes



Panagiotis Giannoulis^{1,3*} , Gerasimos Potamianos^{2,3} and Petros Maragos^{1,3}

Abstract

Voice-enabled interaction systems in domestic environments have attracted significant interest recently, being the focus of smart home research projects and commercial voice assistant home devices. Within the multi-module pipelines of such systems, speech activity detection (SAD) constitutes a crucial component, providing input to their activation and speech recognition subsystems. In typical multi-room domestic environments, SAD may also convey spatial intelligence to the interaction, in addition to its traditional temporal segmentation output, by assigning speech activity at the room level. Such room-localized SAD can, for example, disambiguate user command referents, allow localized system feedback, and enable parallel voice interaction sessions by multiple subjects in different rooms. In this paper, we investigate a room-localized SAD system for smart homes equipped with multiple microphones distributed in multiple rooms, significantly extending our earlier work. The system employs a two-stage algorithm, incorporating a set of hand-crafted features specially designed to discriminate room-inside vs. room-outside speech at its second stage, refining SAD hypotheses obtained at its first stage by traditional statistical modeling and acoustic front-end processing. Both algorithmic stages exploit multi-microphone information, combining it at the signal, feature, or decision level. The proposed approach is extensively evaluated on both simulated and real data recorded in a multi-room, multi-microphone smart home, significantly outperforming alternative baselines. Further, it remains robust to reduced microphone setups, while also comparing favorably to deep learning-based alternatives.

Keywords: Speech activity detection, Smart homes, Active room selection, Microphone arrays, Multi-channel fusion

1 Introduction

Smart home technology has been attracting increasing interest lately, mainly in assistive scenarios for the disabled or the elderly, but also in “edutainment”, home monitoring, and automation applications, among others [1–5]. Given that interaction with users must be convenient and natural, and motivated by the fact that speech constitutes the primary means of human-to-human communication, voice-enabled interaction systems have been progressively entering the field. Indeed, multiple smart home projects have been focusing on voice-based interaction [6–13], and a number of commercial voice assistant home devices have recently been introduced in the market [14].

Such systems typically contain a sequence of modules in their architecture, with speech activity detection (SAD)

being a crucial one, as it provides input to other pipeline components, for example, speaker localization, speech enhancement, keyword spotting, and automatic speech recognition (ASR) [15–17], as well as contributing to the timing of the dialog management [18]. Further to voice-based interaction, SAD has found additional applications, such as telecommunications [19–21], variable rate speech coding [22], and voice-based speaker recognition [23, 24], among others.

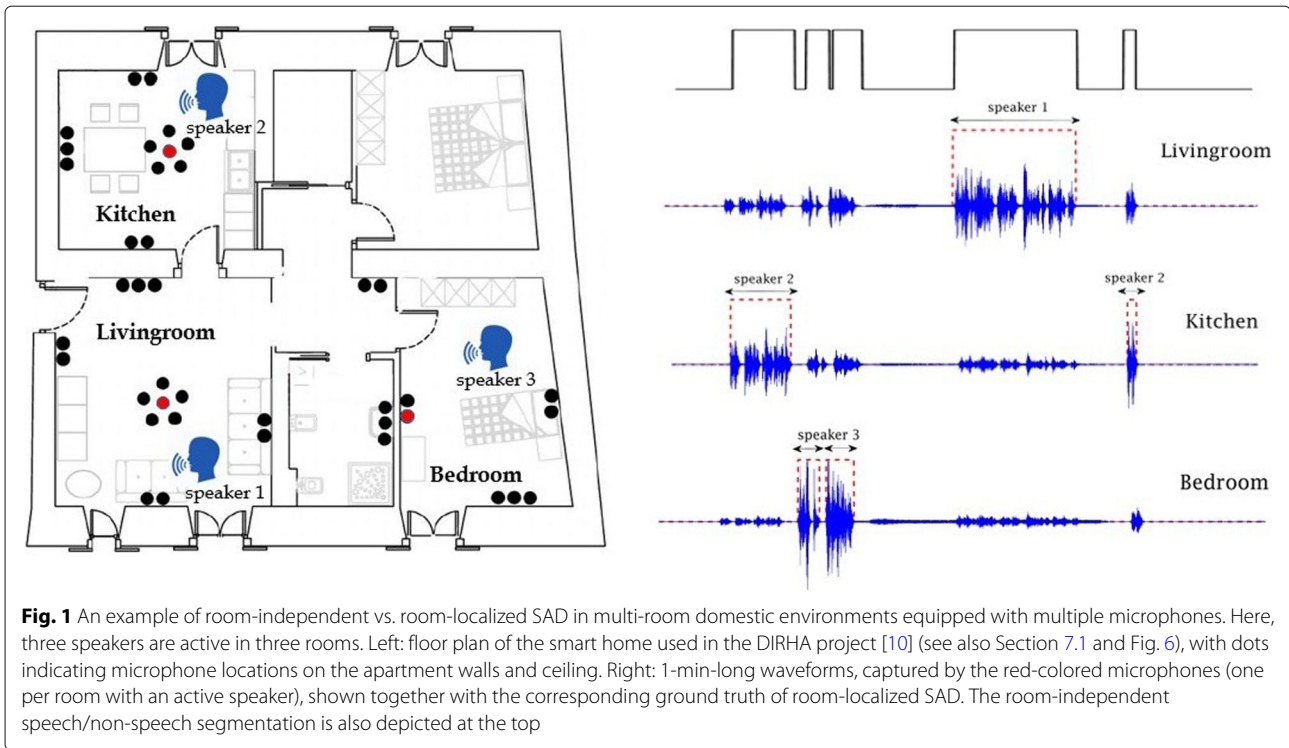
In practice, domestic environments contain multiple rooms, where one or more users may be located wishing to interact with the smart home voice interface. This scenario can be facilitated if the SAD module provides not only time boundaries of speech events (“when”), but also coarse speaker position (“where”) at the room level, i.e., assigning room “tags” to the detected speech activity, thus yielding separate speech/non-speech segmentation outputs, one per room of the smart home (see also Fig. 1). Enriching the traditional “room-independent SAD” to

*Correspondence: pangian@cs.ntua.gr

¹School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece

³Athena Research and Innovation Center, Marousi, Greece

Full list of author information is available at the end of the article



such “room-localized SAD” variant can be useful in multiple ways: It can help disambiguate user commands for voice control of devices or appliances present in multiple rooms (e.g., light switches, windows, temperature control units, television sets); enable room-localized system feedback, for example, via a loudspeaker or visual display at the room where speech activity takes place; and allow parallel voice interaction sessions by multiple subjects inside different rooms, engaging separate system pipelines, one per room [16]; finally, ASR itself can benefit significantly from room localization [25].

Designing a robust SAD system in domestic environments is a hard task due to the challenging acoustic conditions encountered. Such involve speech at low signal-to-noise ratio (SNR), presence of reverberation, and multiple background noise sources often overlapping with speech activity. In the case of room-localized SAD, these difficulties are further exacerbated due to acoustic interference between rooms. To counter these challenges, smart homes typically employ multiple microphones to capture the acoustic scene and “cover” the large multi-room interaction area. This allows exploiting multi-channel processing techniques, for example, fusion of the microphone information at the signal, feature, or decision level, in order to facilitate the analysis of the acoustic scene of interest.

Several efforts have been reported recently on room-localized SAD in multi-room environments [25–32] including our own work [33, 34]. As further overviewed in Section 2, these approaches vary in the kind of features,

classifiers, and number of microphones used per room. Depending on their design, they typically consist of one or two algorithmic stages, and may or not allow the detection of simultaneously active speakers located in different rooms.

In this paper, we present our research work on room-localized SAD for smart homes equipped with multiple microphones distributed in multiple rooms. Our approach is based on the two-stage algorithmic framework that we originally proposed in [34]. There, room-independent SAD hypotheses, obtained at the first stage by traditional statistical modeling and acoustic front-end processing, are further refined and assigned to the room level at a second stage, by means of support vector machine (SVM) classifiers operating on a set of hand-crafted features that are suitably designed to discriminate room-inside vs. room-outside speech. The aforementioned approach is further extended in this paper in multiple ways. In particular:

- Concerning the first stage of the algorithm, this is modified to already provide room-localized SAD hypotheses, various choices for the set of its statistical classes are investigated, and a number of multi-microphone decision fusion techniques are incorporated, which were originally studied in [35] for the problem of room-independent SAD only.

- Concerning the second stage of the algorithm, the set of hand-crafted features of [34] is further enriched by two additional ones: a novel spectrogram texture smoothness

descriptor, as well as a source localization feature based on the smart home floor plan. Further, various feature fusion schemes across rooms are considered, accompanied by different options for their SVM-based modeling. Among these, one remains agnostic to the number of smart home rooms. In addition, application of the second algorithmic stage is also considered on medium-sized windows sliding over the first-stage hypothesized segments, thus enabling their breakup and assignment to potentially different rooms.

- Finally, an extensive evaluation of all algorithmic components is reported, as well as of suitable alternative baselines including an extension of the seminal algorithm of [36] to the room-localized SAD problem. The experiments are conducted on a corpus of both real and simulated data in a multi-room smart home, set up for the purposes of the DIRHA project [10]. This way, insights are gained concerning the strengths, weaknesses, and design choices of the proposed system. This is demonstrated to perform well in the challenging problem of room-localized SAD without the need of a large amount of training data, being robust to the number of available microphones, while also comparing favorably to alternative deep learning approaches.

It should be noted that two of the aforementioned extensions have already been proposed in our earlier work [16], namely the room-localized operation at the first stage, as well as the sliding window mode at the second. However, since our focus there has been on the entire pipeline of the smart home spoken command recognition system, a contrastive evaluation of these enhancements for room-localized SAD has not been investigated. In particular, the room-localized operation at the first stage was also part of our system in [33], where however the focus lied on a joint SAD and speaker localization challenge over a limited area of a multi-room domestic environment [28].

The remainder of the paper is organized as follows: related work is summarized in Section 2; the overview of the proposed system is provided in Section 3, with its two algorithmic stages further detailed in Sections 4 and 5; alternative baselines are presented in Section 6; the datasets and experimental framework are discussed in Section 7; the evaluation is reported in Section 8; and finally, conclusions are drawn in Section 9.

2 Related work

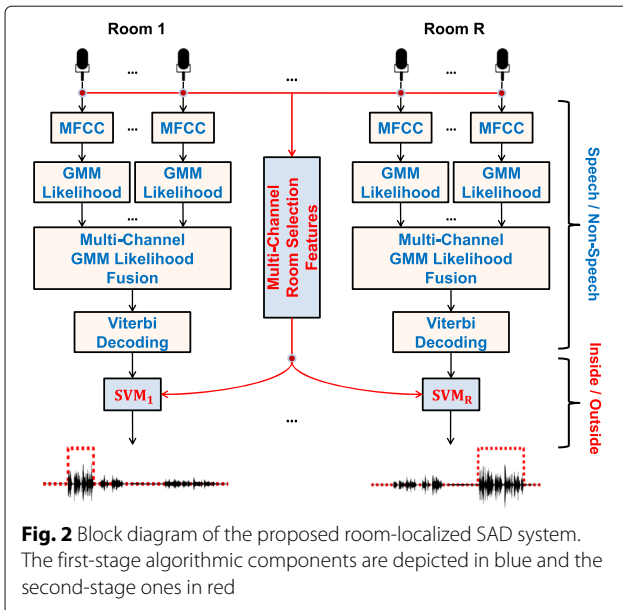
SAD has been a topic of intense research activity, with numerous algorithms proposed in the literature over more than four decades, as for example overviewed in [37]. Some of the most established methods include algorithms incorporated into standards [20, 21], the statistical model-based approach by Sohn et al. [36], and the spectral divergence proposed by Ramírez et al. [38], among others. Typically, SAD methods extract various features from

the waveform that are, for example, related to energy or zero-crossing rate [20, 21, 39, 40], harmonicity and pitch [41–43], formant structure [20, 24, 44, 45], degree of stationarity of speech and noise [46–48], modulation [49–51], or Mel-frequency cepstral coefficients (MFCCs) [24]. Feature extraction is subsequently followed by traditional statistical modeling or, more recently, by deep learning-based classifiers, for example, deep neural networks (DNNs) [52, 53], recurrent ones [54, 55], or convolutional neural networks (CNNs) [56–58], often in conjunction with autoencoders [59]. Further, end-to-end deep learning approaches applied directly to the raw signal have also been proposed [60].

Specifically for the smart home domain, several SAD systems have been developed over the last decade, following the collection of appropriate corpora in domestic environments [61–65]. For example, in [66], linear frequency cepstral coefficients are employed as features in conjunction with the Gaussian mixture model (GMM) and hidden Markov model (HMM) classifiers to detect distressed speech or acoustic events inside a smart apartment for elderly persons. In a similar task under the Sweet-Home project in [67], sound event detection is first performed by discrete wavelet transform features and an adaptive thresholding strategy, followed by speech/event classification using SVMs with GMM supervectors based on MFCCs. In [68], a simple energy-based SAD precedes the HMM-based recognition of sounds and spoken words. In [69], SAD is performed on headset microphone audio to track human behavior inside a smart home, with the proposed system employing an energy detector and a neural network trained on linear predictive coding coefficients and band-crossing features. Finally, in our earlier work within the DIRHA project [35], we investigated several fusion techniques for multi-channel SAD based on GMMs and HMMs trained on traditional MFCCs.

The aforementioned SAD systems aim to detect speech activity over the entire smart home, without however considering its typical multi-room layout. Only few recent approaches in the literature focus on the task of room-localized SAD in multi-room domestic environments that constitutes the focus of this paper, yielding a speech/non-speech segmentation for each individual room of the smart home.

The majority of such systems operate in two stages. Typically, the first stage generates speech segment hypotheses over the entire home or for each specific room, which are further examined, refined, and assigned to the proper room at the second stage. Specifically, in [25], at the first stage of the proposed algorithm, DNN-based single-channel SAD is performed in each room. Then, at the second stage, for each detected speech segment, SNR and coherence-based features are extracted from all rooms



and concatenated to feed a linear discriminant analysis classifier that yields the segment room allocation. In [26], at the first stage, statistical-based SAD is performed for each microphone, and then, majority voting over the room microphones provides the speech segments of each room. At the second stage, speaker localization output feeds a classifier (SVM or neural network) to further examine speech segments and delete those originating in other rooms. In [27], at the first stage, multi-layer perceptrons are employed for each microphone, and speech/non-speech segmentation is achieved via majority voting for each room. Then, in case of segments assigned to multiple rooms, a speech envelope distortion measure is employed to decide the correct room. In [28], three different features are investigated for room-localized SAD, namely SNR, periodicity, and the global coherence field. Speech boundaries for each room are computed by simple thresholding of these feature values and by using a heuristic rule over consecutive active frames.

In addition to the above, single-stage approaches have also been pursued for room-localized SAD. Specifically, in [29], a DNN is employed taking as input 176-dimensional vectors composed of a variety of features, such as MFCCs, RASTA-PLPs, envelope variance, and pitch. Similar features (but 187-dimensional) and DNNs are again considered in [30], as well as alternative classifiers, including a 2D-CNN. The latter is extended to a multi-channel 3D-CNN system in [31], where log-Mel filterbank energies (40-dimensional) are employed as features, temporal context is exploited by concatenating adjacent time frames, and the resulting 2D single-microphone feature matrices are stacked across

channels. Finally, in [32], the aforementioned 3D-CNN is combined with the GCC-PHAT [70] based CNN of [71] to yield a joint SAD and speaker localization network.

As already mentioned in the introduction, we have also investigated room-localized SAD in our earlier work, following the two-stage algorithmic paradigm. Specifically, in [33], at the first stage of the developed approach, speech/non-speech segmentation was performed for each room, by means of multi-microphone decision fusion over GMMs trained on a traditional MFCC-based acoustic front-end. At the second stage, in case of speech segments simultaneously active in multiple rooms, room selection was enabled by comparing an average GMM-based log-likelihood ratio for the given segment across the different rooms. In subsequent work [34], the first stage was replaced by a multi-channel room-independent SAD module, whereas the second stage adopted the use of specific features to discriminate room-inside vs. room-outside speech by means of SVM-based classifiers. The approach was further refined in later work [16], as part of a modular pipeline of a smart home spoken command recognition system.

In the current paper, we maintain the two-stage algorithmic approach for room-localized SAD, combining the design of the first stage in [33] with that of the second stage in [34]. In the process, we introduce a number of extensions in fusion techniques, hand-crafted room discriminant features, statistical modeling, and system evaluation, as discussed earlier in Section 1. The resulting algorithm is presented in detail next.

3 Notation and system overview

Let us denote by R the number of rooms inside a given smart home that is equipped with a set of microphones \mathcal{M}_{all} . This is partitioned into subsets \mathcal{M}_r , for $r = 1, 2, \dots, R$, each containing the microphones located inside room r . Let us also denote by $\mathbf{o}_{m,t}$ the short-time acoustic feature vectors (e.g., MFCCs) extracted from the signal of microphone m , and by $\mathbf{o}_{\mathcal{M},t}$ their concatenation over microphone set $\mathcal{M} \subseteq \mathcal{M}_{\text{all}}$, with t indicating time indexing at the frame level (typically at a 10-ms resolution).

We are interested in room-localized SAD, seeking speech/non-speech segmentations for each room r , detecting speech events occurring inside it but ignoring speech originating in other rooms or any other non-speech events. As also shown in Fig. 1, this differs from room-independent SAD, where a single speech/non-speech segmentation is produced, including speech events occurring inside any of the R rooms of the smart home.

As already discussed in the previous sections and also depicted in the block diagram of Fig. 2, our proposed

system for room-localized SAD operates in two stages. The first stage, detailed in Section 4, is based on single-channel GMM classifiers, each trained on an individual room microphone, employing MFCC features and operating at the frame level. An appropriate decision fusion scheme follows, combining GMM likelihood scores across all room microphones and, by means of Viterbi decoding, providing a crude speech/non-speech segmentation for the given room. Then, at the second stage, presented in detail in Section 5, for the speech segments detected for each room, an SVM classifier is employed on a number of hand-crafted room localization features, specially designed to discriminate room-inside vs. room-outside speech. Various feature fusion schemes across rooms are considered for this purpose, accompanied by different options for their SVM-based modeling.

4 First stage: speech segment generation

We now proceed with a detailed description of the first stage of the developed room-localized SAD system. This stage generates individual speech/non-speech segmentations for every room using the specific room microphones only, thus providing initial room-localized SAD hypotheses to be refined later. To accomplish this, it employs traditional acoustic front-end processing and statistical modeling at the microphone level as discussed in Section 4.1, followed by decision fusion across microphones as detailed in Section 4.2 and appropriate decoding schemes that are presented in Section 4.3. Variations on the choices of microphones and classes considered are discussed in Section 4.4.

4.1 Single-microphone system core

At the core of the system lies the single-microphone speech/non-speech modeling. Specifically, for each microphone of the smart home, a traditional 39-dimensional MFCC-plus-derivatives acoustic front-end is employed, with features extracted over 25-ms Hamming-windowed signal frames with a 10-ms shift. Subsequently, two-class microphone-specific GMMs are trained on these features (32 Gaussian mixtures with diagonal covariance matrices are used in our implementation), with the set of classes being $\mathcal{J} = \{\text{sp}_r, \text{sil}_{\text{all}}\}$, where sp_r denotes speech originating in room r where the given microphone is located and sil_{all} indicates the lack of speech in all rooms. Alternative class choices for set \mathcal{J} are discussed in Section 4.4.

4.2 Multi-microphone decision fusion

The developed system performs multi-microphone fusion at the decision level, where the GMM log-likelihood scores of different channels are combined at the

frame level for each class of interest, potentially also incorporating channel decision confidence. In particular, the following approaches for decision fusion over microphone set $\mathcal{M} \subseteq \mathcal{M}_{\text{all}}$ are considered, which were investigated in our earlier work [35], but for room-independent SAD only:

- Log-likelihood summation, where the fused log-likelihoods (log class-conditionals) at frame t become

$$c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) = \sum_{m \in \mathcal{M}} w_{m,t} b_{m,j}(\mathbf{o}_{m,t}), \quad (1)$$

where $b_{m,j}(\mathbf{o}_{m,t})$ denotes the log-likelihoods of the GMMs for microphone m given its acoustic features $\mathbf{o}_{m,t}$ at time frame t and class $j \in \mathcal{J}$. The individual microphone scores in (1) can be uniformly weighted by setting $w_{m,t} = 1 / |\mathcal{M}|$ (where $|\bullet|$ denotes set cardinality), in which case the scheme will be referred to as *unweighted log-likelihood summation* (“u-sum”), or adaptively weighted at any given time frame t , according to channel decision confidence that is estimated as

$$w_{m,t} = \frac{|b_{m,\text{sp}_r}(\mathbf{o}_{m,t}) - b_{m,\text{sil}_{\text{all}}}(\mathbf{o}_{m,t})|}{\sum_{m' \in \mathcal{M}} |b_{m',\text{sp}_r}(\mathbf{o}_{m',t}) - b_{m',\text{sil}_{\text{all}}}(\mathbf{o}_{m',t})|}, \quad (2)$$

in which case, the method will be termed *weighted log-likelihood summation* (“w-sum”). Weighting by (2) was investigated among other schemes for room-independent SAD in [35], motivated by intuition that large log-likelihood differences between the classes imply higher classification confidence.

- Log-likelihood selection, where, at each time frame t , a microphone $\hat{m}_t \in \mathcal{M}$ is selected to provide all fused class log-likelihoods, i.e.,

$$c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) = b_{\hat{m}_t,j}(\mathbf{o}_{\hat{m}_t,t}), \quad \text{for all } j \in \mathcal{J}. \quad (3)$$

Such microphone can be chosen as the one achieving the highest frame log-likelihood over all channels and over all classes, i.e.,

$$\hat{m}_t = \arg \max_{m \in \mathcal{M}} \left\{ \max_{j \in \mathcal{J}} b_{m,j}(\mathbf{o}_{m,t}) \right\},$$

in which case the scheme will be referred to as *log-likelihood maximum selection* (“u-max”), or as the channel with the highest confidence (2), i.e.,

$$\hat{m}_t = \arg \max_{m \in \mathcal{M}} w_{m,t},$$

in which case the method will be termed *log-likelihood confidence selection* (“w-max”).

- Majority voting, where, at each time frame t , single-channel decisions, $\hat{j}_{m,t} = \arg \max_{j \in \mathcal{J}} b_{m,j}(\mathbf{o}_{m,t})$, are accumulated over microphone set \mathcal{M} , and the class with the highest decision incidence is chosen. Such accumulation can be computed uniformly over the channels, in which

case the scheme will be termed *unweighted majority voting* (“u-vote”), or scaled by means of (2), resulting in *weighted majority voting* (“w-vote”).

Among the above approaches, based on the experimental results of Section 8, the developed room-localized SAD system employs the “w-sum” scheme computed over the set of microphones inside one room at a time, i.e., $\mathcal{M} = \mathcal{M}_r$. Alternative choices for set \mathcal{M} are discussed in Section 4.4.

4.3 Speech/non-speech segmentation

Following GMM training and multi-channel fusion, two speech detection implementations are developed: The first operates on mid-sized sliding windows, thus resulting in low latency, whereas the second performs Viterbi decoding over longer sequences, providing superior accuracy (as demonstrated in Section 8), but being more suitable for off-line processing.

- GMM-based scoring over sliding window: This scheme performs sequential classification over sliding windows of fixed duration and overlap (400 ms and 200 ms, respectively, are used). Specifically, for a given time window $\mathcal{T} = [t_s, t_e]$ and microphone m , the log-likelihoods for each class $j \in \mathcal{J}$ are first computed by adding all frame scores within the window. This results in scores $b_{m,j}(\mathbf{o}_{m,\mathcal{T}}) = \sum_{t=t_s}^{t_e} b_{m,j}(\mathbf{o}_{m,t})$, where $\mathbf{o}_{m,\mathcal{T}}$ denotes all feature vectors within window \mathcal{T} . Microphone fusion is then carried out as in Section 4.2, but employing the window log-likelihoods instead.

- HMM-based Viterbi decoding over sequence: In this scheme, HMMs are built with a set of fully connected states \mathcal{J} , state transition probabilities $\{a_{jj'}\}$, for $j, j' \in \mathcal{J}$, and class-conditional observation probabilities provided by the class GMMs of Section 4.1. Then, Viterbi decoding is performed over an entire sequence of observations (in our data, such are of 1-min length, as discussed in Section 7.1), in order to provide the desired speech/non-speech segmentation. Specifically, for the single-microphone case, the well-known recursion [72]

$$\delta_{m,j}(t) = \max_j \{ \delta_{m,j'}(t-1) + \log(a_{jj'}) + b_{m,j}(\mathbf{o}_{m,t}) \},$$

is used, where $\delta_{m,j}(t)$ denotes the score of the best decoding path ending at state j and accounting for the first t frame observations of microphone m . This can be readily extended to the fusion schemes of (1) and (3) over microphone set \mathcal{M} as:

$$\delta_{\mathcal{M},j}(t) = \max_j \{ \delta_{\mathcal{M},j'}(t-1) + \log(a_{jj'}) + c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) \},$$

whereas majority voting fusion schemes “u-vote” and “w-vote” are modified to be applied over best-path scores $\delta_{m,j}(t)$ instead of log-likelihoods $b_{m,j}(\mathbf{o}_{m,t})$.

Between the two aforementioned decoding schemes, the proposed system follows the HMM-based approach

due to its superior performance, with a number of fine-tuned parameters incorporated in it. Specifically, these are the state transition penalty that tunes the flexibility of the decoder to change states, as well as the speech class prior that favors or not the selection of the speech state.

4.4 Variations in sets of classes and microphones

As already discussed, to obtain the first stage of the speech/non-speech segmentation hypothesis for room r , only the particular room microphones are considered ($\mathcal{M} = \mathcal{M}_r$). A number of variations however are possible for the set of classes \mathcal{J} , which are investigated in the experiments of Section 8.2:

- $\mathcal{J} = \{ \text{sp}_r, \text{sil}_{\text{all}} \}$, where sp_r denotes speech inside room r and sil_{all} indicates the absence of speech in all rooms of the smart home. This set is used in the proposed room-localized SAD algorithm.
- $\mathcal{J} = \{ \text{sp}_r, \text{sil}_r \}$, where sil_r indicates the absence of speech in room r . This set is used in our work in [33].
- $\mathcal{J} = \{ \text{sp}_r, \text{sp}_{\bar{r}}, \text{sil}_{\text{all}} \}$, where $\text{sp}_{\bar{r}}$ indicates the speech inside any of the other rooms, excluding room r .

In addition, in our earlier work [34], the first stage of the algorithm provides room-independent SAD output. That system uses the “w-sum” decision fusion scheme with all smart home microphones contributing to (1), i.e., $\mathcal{M} = \mathcal{M}_{\text{all}}$. Further, the set of classes employed is $\mathcal{J} = \{ \text{sp}_{\text{all}}, \text{sil}_{\text{all}} \}$, where sp_{all} denotes the speech occurring in any of the smart home rooms.

5 Second stage: room assignment

Following the generation of initial room-localized SAD hypotheses, the second stage of the developed algorithm performs the final selection of active segments for each room. For this purpose, five hand-crafted features are proposed as detailed in Section 5.1, extracted at the segment level for each room, and capable of segment discrimination as originating from inside vs. outside a given room. These features are then fused within and across rooms as presented in Section 5.2 and are fed to SVM classifiers that perform room assignment as detailed in Section 5.3, temporally operating on the given segment as discussed in Section 5.4. Various options for the above are presented.

5.1 Room discriminant features

As mentioned above, for any first-stage speech segment $\mathcal{T} = [t_s, t_e]$ starting at time-frame t_s and ending at frame t_e , segment-level features are extracted for each room. The design of these hand-crafted features is motivated by intuition concerning (a) the energy, (b) the reverberation, and (c) the arrival direction of the microphone signals. For example, microphones located inside the room where a speech segment originates are expected to yield signals

with higher energy and lower reverberation than microphones located outside it. Likewise, the room door region typically appears as the speech source for room-outside segments.

In particular, five scalar features are considered in this paper, extending our earlier work [33] with two additional novel features presented in Sections 5.1.4 and 5.1.5. The features are specially designed to provide room-inside vs. room-outside segment source discrimination, as also depicted in the histograms of Fig. 3.

It should be noted that in contrast to the acoustic front-end of the first stage of the algorithm that extracts microphone-dependent features, the features of the second stage are instead room-dependent. Indeed, their estimation typically involves all microphones located in a room (or in the entire smart home, as in Sections 5.1.1 and 5.1.3), performing in a sense fusion of their information at the signal level. Such derivation requires of course knowledge of the microphone room membership, but in the case of Section 5.1.2 also of additional information concerning which microphones lie adjacent to each other, and in the case of Section 5.1.5 further knowledge of the microphone topology and room layout. Details are provided next.

5.1.1 Energy-based feature

Originally proposed in [33], this feature is motivated by intuition that microphones inside the room where speech activity occurs will exhibit, on average, higher SNRs compared to ones outside it. For its computation, given detected speech segment $\mathcal{T} = [t_s, t_e]$, the energy

ratio (ER) of speech over non-speech is first computed for all smart home microphones. For this purpose, the initial part of the speech segment, as well as the trailing part of non-speech preceding it, both of length $\Delta \tau$, is utilized to yield

$$\text{ER}_{m,\mathcal{T}} = \left(\sum_{\tau=Lt_s}^{Lt_s+\Delta\tau-1} x_m(\tau)^2 \right) / \left(\sum_{\tau=Lt_s-\Delta\tau}^{Lt_s-1} x_m(\tau)^2 \right), \quad (4)$$

for all microphones $m \in \mathcal{M}_{\text{all}}$. In (4), $x_m(\tau)$ denotes the signal captured by microphone m , with τ indicating indexing at the sample level. The latter is related to frame-level indexing by $\tau = Lt$, where L is the number of signal samples over the short-time window shift. Following computations (4), the ERs are sorted across all smart home microphones, and the microphone set with the K largest values is derived, denoted by $\mathcal{M}^{(K)}$. Finally, the desired energy-based feature for room r is extracted as the difference between the sum of the ERs of the microphones in set $\mathcal{M}^{(K)}$ that are located inside room r and the ER sum of the ones in $\mathcal{M}^{(K)}$ but located in other rooms, namely

$$f_{r,\mathcal{T}}^{(\text{en})} = \sum_{m \in \mathcal{M}^{(K)} \cap \mathcal{M}_r} \text{ER}_{m,\mathcal{T}} - \sum_{m \in \mathcal{M}^{(K)} \setminus \mathcal{M}_r} \text{ER}_{m,\mathcal{T}},$$

for all rooms $r = 1, 2, \dots, R$. In our implementation, $K = 5$ and, in (4), $\Delta \tau$ corresponds to a 0.5-s interval.

5.1.2 Coherence feature

Originally proposed in [25] and re-used in [33], this feature is motivated by intuition that signals captured by

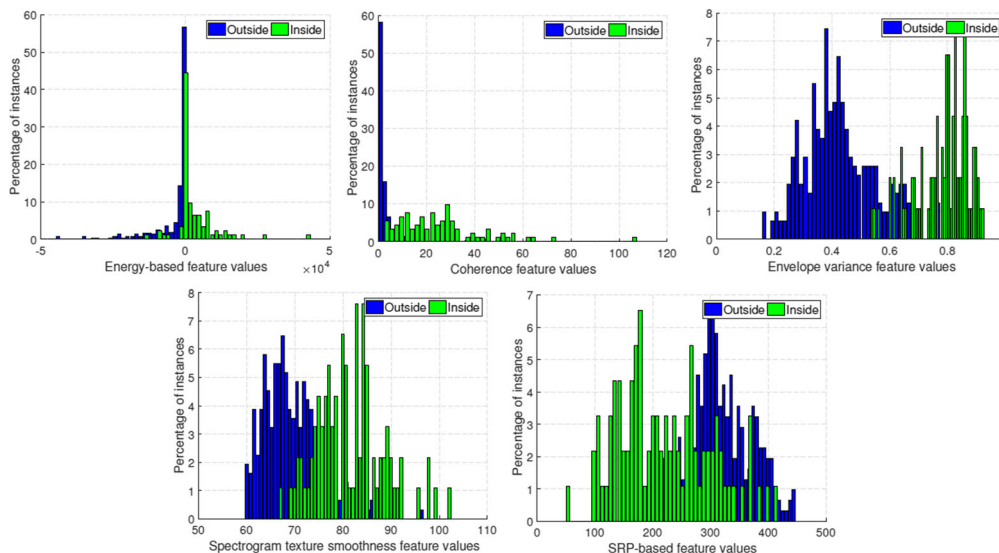


Fig. 3 Histograms of the five hand-crafted scalar features of Section 5.1, demonstrating their ability to discriminate room-inside vs. room-outside speech. Histograms are computed over the development set of the simulated dataset of Section 7.1, for the case of the smart home bedroom (see also Fig. 1). Upper row, left-to-right: energy-based feature (Section 5.1.1), coherence feature (Section 5.1.2), and envelope variance one (Section 5.1.3). Lower row, left-to-right: spectrogram texture smoothness feature (Section 5.1.4) and SRP-based one (Section 5.1.5)

pairs of adjacent microphones located outside a speech-active room will exhibit higher reverberation and thus lower cross-correlation than pairs inside it. To compute the coherence feature for room r , the set of adjacent pairs of microphones inside the room is first determined, denoted by $\{\mathcal{M}_r \times \mathcal{M}_r\}_{\text{adj}}$. Such pairs typically consist of neighboring microphones in larger arrays (see also Section 7.1). Then, for every time frame t within detected speech segment \mathcal{T} , the maximum cross-correlation of the signal frames of adjacent microphone pair (m, m') is computed, denoted by $C_{m,m'}(t)$. This is repeated for all pairs $(m, m') \in \{\mathcal{M}_r \times \mathcal{M}_r\}_{\text{adj}}$ and the maximum retained. Finally, the result is averaged over the entire segment \mathcal{T} , yielding the coherence feature for room r , as:

$$f_{r,\mathcal{T}}^{(\text{coh})} = \text{avg}_{t \in \mathcal{T}} \left\{ \max_{\substack{(m,m') \\ \in \{\mathcal{M}_r \times \mathcal{M}_r\}_{\text{adj}}}} C_{m,m'}(t) \right\}.$$

Note that this feature employs the un-normalized cross-correlation function in order to also “capture” signal attenuation. In our implementation, signal cross-correlation is computed over fixed size sliding windows of 100 ms in length and a 25-ms shift.

5.1.3 Envelope variance feature

Originally proposed in [73] for ASR channel selection and used in [27, 29, 33] for room-localized SAD, this feature is motivated by intuition that higher reverberation (indicative of room-outside speech) results in smoother short-time speech energy, also observed as reduced dynamic range of the corresponding envelope. To compute the envelope variance feature, we follow the derivations in [73]. Briefly, for each microphone m , the short-time filterbank energy, denoted by $X_m(n, t)$, is obtained for time frames $t \in \mathcal{T}$, where, as above, \mathcal{T} is the detected speech segment and n denotes the sub-band (20 linear filters are used here). Then, the n^{th} sub-band envelope of microphone m is computed as:

$$\hat{X}_m(n, t) = \exp \left\{ \log [X_m(n, t)] - \text{avg}_{t \in \mathcal{T}'} \left\{ \log [X_m(n, t)] \right\} \right\},$$

where \mathcal{T}' denotes medium-sized windows sliding over segment \mathcal{T} , the time progression of which will be indexed by t' (600-ms-long windows with a 50-ms shift are used). Then, the variance of each sub-band envelope is computed (following cube root compression) as:

$$V_m(n, t') = \text{var}_{t \in \mathcal{T}'} \left\{ \hat{X}_m(n, t)^{1/3} \right\},$$

subsequently normalized over all smart home microphones, and its average over all sub-bands obtained:

$$\text{EV}_m(t') = \text{avg}_n \left\{ \frac{V_m(n, t')}{\max_{m' \in \mathcal{M}_{\text{all}}} V_{m'}(n, t')} \right\}. \quad (5)$$

In this work, we define the envelope variance feature of segment \mathcal{T} for room r as the average over all mid-sized shifting windows within \mathcal{T} of the maximum value of (5) over the set of all room microphones \mathcal{M}_r , i.e.,

$$f_{r,\mathcal{T}}^{(\text{ev})} = \text{avg}_{t' \in \mathcal{T}} \left\{ \max_{m \in \mathcal{M}_r} \text{EV}_m(t') \right\}. \quad (6)$$

5.1.4 Spectrogram texture smoothness feature

For measuring the degree of reverberation, an additional feature is proposed in this paper, based on the “smearing” effect that reverberant conditions cause to the speech signal spectrogram. An example is shown in Fig. 4: There, for a speech occurrence inside the bedroom of the smart home of Fig. 1, the spectrograms of two signals captured by a microphone located in the bedroom and one in the kitchen are depicted, showing that the latter (located outside the speech-active room) is much smoother (smeared).

To measure this effect, the proposed feature considers the signal spectrogram as a 2D image, and attempts to quantify its texture smoothness by applying to it the 2D discrete Teager energy operator of [74], yielding

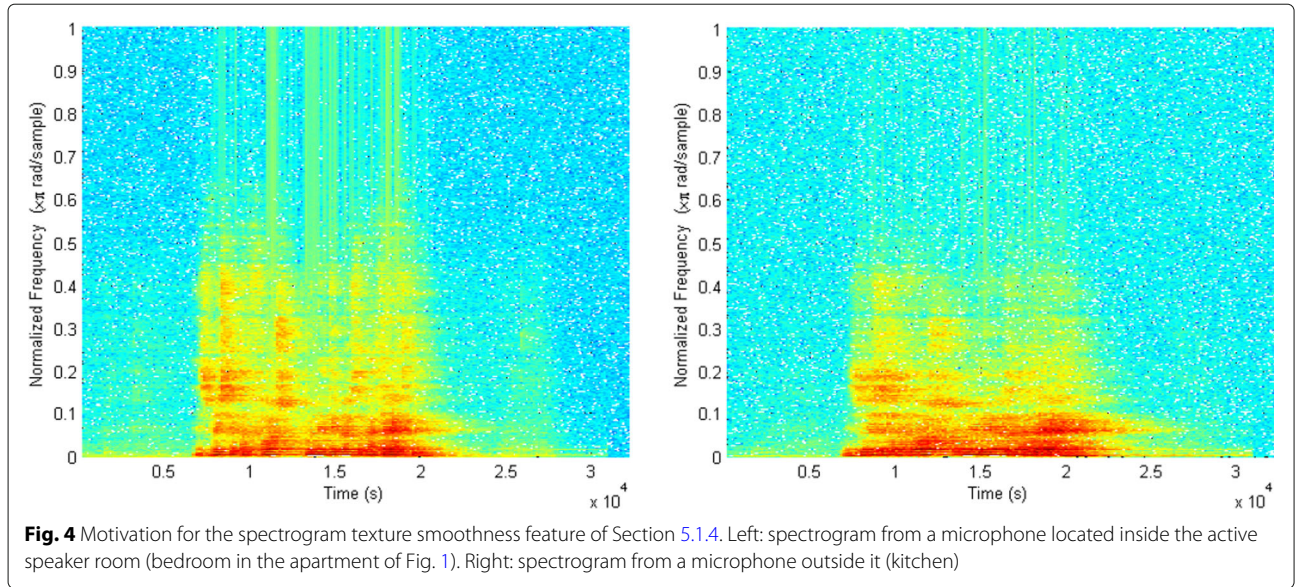
$$\Phi_m(n, t) = 2 (S_m(n, t))^2 - S_m(n, t-1) S_m(n, t+1) - S_m(n-1, t) S_m(n+1, t),$$

where $S_m(n, t)$ denotes the signal spectrogram of microphone m at short-time frame $t \in \mathcal{T}$, and n is the frequency index (40-ms-long Hamming windows with a 20-ms shift and 960 frequency bins are used here). Then, as in Section 5.1.3, medium-sized windows \mathcal{T}' sliding over segment \mathcal{T} are considered, the time progression of which is indexed by t' (600-ms-long windows with a 50-ms shift are used). The values of $\Phi_m(n, t)$ are then averaged over a part of the resulting 960×30 -sized spectrogram image as:

$$\Phi_m(t') = \text{avg}_{n=1, \dots, 200} \text{avg}_{t \in \mathcal{T}'} \{ \Phi_m(n, t) \},$$

where the frequency domain averaging is carried out over the 200 lower frequency bins that correspond to the 0–5 kHz frequency range of the 48-kHz sampled signal, focusing on speech content. Finally, the spectrogram texture smoothness feature for room r and segment \mathcal{T} is obtained by maximizing over all room microphones and averaging the result over all medium-sized windows, namely

$$f_{r,\mathcal{T}}^{(\text{ts})} = \text{avg}_{t' \in \mathcal{T}} \left\{ \max_{m \in \mathcal{M}_r} \Phi_m(t') \right\}. \quad (7)$$



5.1.5 SRP-based feature

The final feature considered for room assignment of detected speech segments is based on the steered response power (SRP-PHAT) approach of [75], and it is proposed for the first time in this paper for room-localized SAD. Employing SRP allows the creation of an acoustic map, by computing the signal power when steering microphone arrays in the direction of a specific location. The position of the sound source corresponds to that with the maximum SRP value over all possible locations. In the case of multi-room smart homes, one expects that speech originating from outside a given room will likely exhibit high SRP values at the door region that connects that room to the rest of the apartment. In contrast, for room-inside speech, the actual source location should yield the highest SRP instead. An example for this motivation is depicted in Fig. 5.

To compute the SRP-based feature for room r , a 3D region is first defined, denoted by \mathcal{A}_r that corresponds to cylindrically shaped volume(s) covering the room door(s). Specifically, on the floor plane, this lies inside room r , delineated by a 0.7-m radius semicircle around the door center, while also containing all points above it. Using a 10-cm spatial resolution for each dimension, and depending on the number of doors of the room, this scheme yields approximately between $2k$ and $4.3k$ points, denoted as $\vec{y} \in \mathcal{A}_r$, expressed in the 3D room coordinate system (see also Fig. 5).

Then, for all points $\vec{y} \in \mathcal{A}_r$, the corresponding SRP-PHAT values for time frame $t \in \mathcal{T}$ are computed (200-ms-long frames with a 100-ms shift are used), by summing the generalized cross-correlations over all pairs

of adjacent microphones in room r , as:

$$P_r(t, \vec{y}) = \sum_{\substack{(m, m') \\ \in \{\mathcal{M}_r \times \mathcal{M}_r\}_{\text{adj}}}} \int_0^{2\pi} \frac{X_m(\omega, t) X_{m'}^*(\omega, t)}{|X_m(\omega, t) X_{m'}^*(\omega, t)|} e^{j\omega \tau_{mm'}(\vec{y})} d\omega,$$

where $X_m(\omega, t)$ denotes the DTFT of the m th microphone signal frame and $\tau_{mm'}(\vec{y})$ is the time difference of arrival at point \vec{y} between the signals of adjacent microphones m and m' . Finally, the SRP-based feature is computed by summing all above values and averaging them over all windows $t \in \mathcal{T}$, i.e.,

$$f_{r, \mathcal{T}}^{(\text{srp})} = \text{avg}_{t \in \mathcal{T}} \left\{ \sum_{\vec{y} \in \mathcal{A}_r} P_r(t, \vec{y}) \right\}. \quad (8)$$

Clearly, the computation of this feature requires knowledge of the microphone topology and room layout.

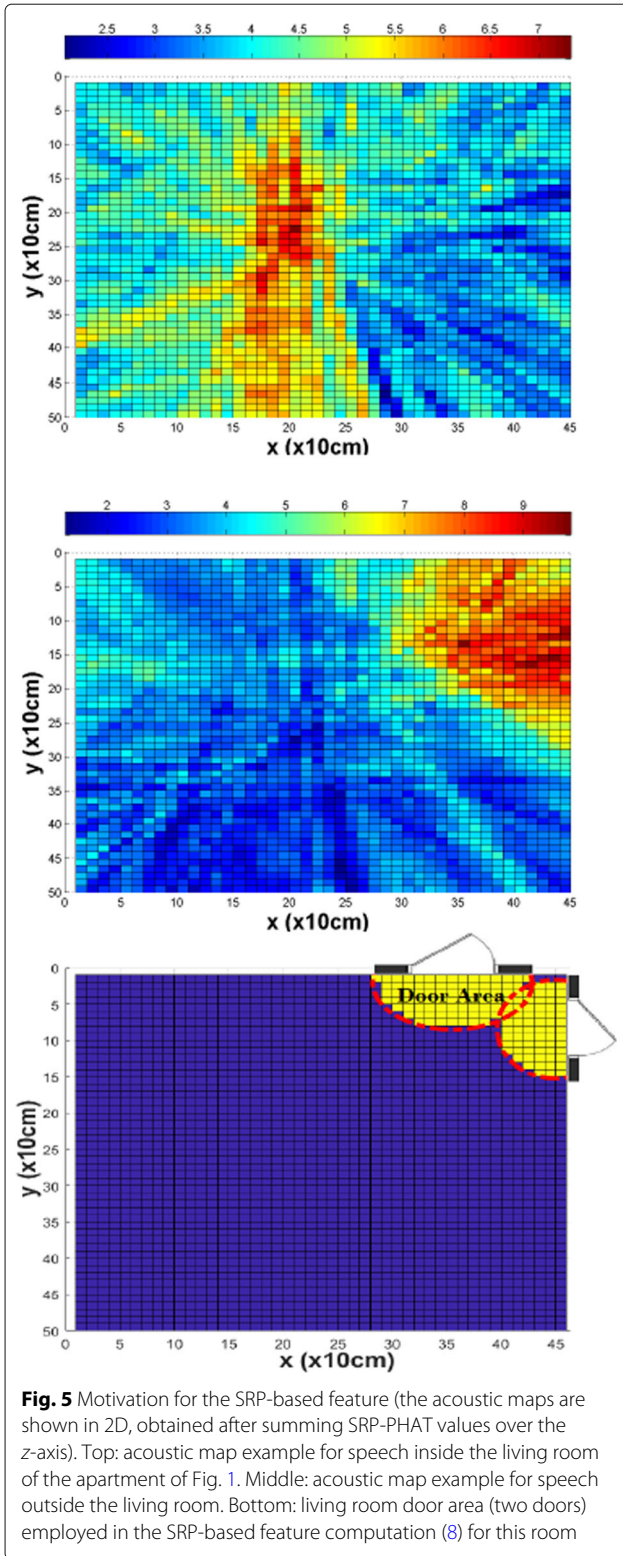
5.2 Intra- and inter-room feature fusion

Using the above framework in the proposed system, for each candidate speech segment \mathcal{T} , five features are extracted for each room r . The features are then combined by intra-room feature fusion (plain concatenation), resulting in five-dimensional feature vectors

$$f_{r, \mathcal{T}}^{(\text{all})} = \left[f_{r, \mathcal{T}}^{(\text{en})}, f_{r, \mathcal{T}}^{(\text{coh})}, f_{r, \mathcal{T}}^{(\text{ev})}, f_{r, \mathcal{T}}^{(\text{ts})}, f_{r, \mathcal{T}}^{(\text{srp})} \right], \quad (9)$$

for each room $r = 1, 2, \dots, R$.

In addition, inter-room feature fusion can be beneficial to room-inside vs. room-outside speech discrimination. Two schemes are considered for this purpose:



- Inter-room feature concatenation, where vectors from all R rooms are concatenated, resulting in a single $5R$ -dimensional feature vector for segment \mathcal{T} ,

$$f_{\text{home},\mathcal{T}}^{(\text{all})} = \left[f_{1,\mathcal{T}}^{(\text{all})}, f_{2,\mathcal{T}}^{(\text{all})}, \dots, f_{R,\mathcal{T}}^{(\text{all})} \right]. \quad (10)$$

- Inter-room feature averaging, where vectors from each room are augmented by the feature average across the remaining $R - 1$ rooms, resulting in ten-dimensional representations of segment \mathcal{T} ,

$$f_{r,\text{avg},\mathcal{T}}^{(\text{all})} = \left[f_{r,\mathcal{T}}^{(\text{all})}, \text{avg}_{r' \neq r} \left\{ f_{r',\mathcal{T}}^{(\text{all})} \right\} \right], \quad (11)$$

for each room $r = 1, 2, \dots, R$. This way, feature vector dimensionality is no longer a function of R . Alternatives to (11) can also be designed, for example, employing feature extrema instead of averages.

5.3 SVM classification

The fused feature vectors are fed to appropriately designed classifiers, in order to determine the room of origin for a given segment. In this paper, linear SVMs are employed for this purpose, due to the two-class nature of the problem (room-inside vs. room-outside segment classification), as well as the relatively small corpus size (see also Section 7.1).¹ Specifically, two SVM modeling approaches are considered, resulting to a total of five different models, as discussed next:

- Room-specific SVM models, where a separate classifier is built for each smart home room. Each training segment thus provides data to a total of R SVMs as a room-inside or -outside class sample, while during testing, a candidate segment is fed to the SVM of the room in which it was detected by the first stage. The SVMs can be built on any of the three feature vectors of Section 5.2, given by (9), (10), or (11), thus resulting in three different systems.

- Global SVM models, where a single SVM is developed being applicable to all rooms, thus removing dependence of the number of SVM models on R . Each training segment provides its data to the global SVM a total of R times (once as a room-inside sample and $R - 1$ times as a room-outside one). During testing, candidate segments are fed to this global SVM. In both cases, room-dependent features are used, provided by (9) or (11), yielding two different systems. Features (10) are not used, as they would have re-introduced dependency on R .

Among the above modeling options, the proposed system employs room-specific SVMs on inter-room concatenated features (10). Note also that since each room

¹All SVMs are trained in Matlab[®], by its `svmtrain.m` function. By default, the regularization parameters are set taking into account the unbalanced nature of the two classes of interest. For this purpose, different penalties are set for misclassifying each class samples, with their ratio being equal to the inverse ratio of the two class sample sizes.

decides for its own final segments, it is possible that a segment gets assigned to multiple rooms or to no rooms at all.

5.4 Temporal operation and post-processing

In practice, the SVM classification of speech segments can be performed at two different temporal scales:

- Over the entire segment, where a single scalar feature is extracted for the segment for each of the five categories of Section 5.1, providing a single sample for SVM training or testing. Thus, assignment to a given room is made for the whole segment.

- Over segment sliding windows, where features are extracted on medium-sized windows sliding over the given segment. As a result, each segment provides multiple data points for SVM training or testing (per window). The scheme allows segment breakup and selective assignment of its parts into the room that it was detected in by the first stage of the algorithm.

The proposed room-localized SAD system employs the sliding window approach, using windows of 600 ms in size advancing by a 100-ms shift. This necessitates minor modifications to the feature extraction methodology of Section 5.1. In particular, there is no longer the need of averaging in (6) and (7), since the medium-sized window sizes coincide, thus trivially allowing for one window only. Further, in (4), the non-speech energy is computed over the 0.5-s interval preceding the first window of the segment.

As a final step, post-processing is also applied to the results. Specifically, speech segments with less than 0.7-s distance between them are unified, whereas speech segments of less than 0.4-s duration are deleted.

6 Baseline approaches

Two additional, simpler systems are presented in this section, both following a two-stage architecture, to serve as baselines against the developed room-localized SAD system. The first method employs MFCC features and GMM classifiers in both its algorithmic stages, while the second extends the well-known statistical model-based approach of Sohn et al. [36] to room-localized SAD, by incorporating a simple SNR-based room assignment criterion. Details follow.

6.1 MFCC/GMM-based system

This baseline follows our earlier work [33], and it is mainly considered in order to evaluate a system based entirely on a standard acoustic front-end (MFCC features), aiming also to demonstrate the value of the room discriminant features of Section 5.1.

Its first stage is identical to that of the proposed system. Namely, for every smart home room, it performs weighted log-likelihood summation of MFCC/GMM-based scores

by means of (1) and (2) over all room microphones ($\mathcal{M} = \mathcal{M}_r$) for classes $\mathcal{J} = \{\text{sp}_r, \text{sil}_{\text{all}}\}$ (see also Section 4.4).

At the second stage, segments generated by the first stage are further examined and classified as room-inside or room-outside speech. For this purpose, room-specific GMMs are trained for each class $\mathcal{J} = \{\text{sp}_r, \text{sp}_{\bar{r}}\}$, and unweighted log-likelihood summation of MFCC/GMM-based scores is performed over all room microphones ($\mathcal{M} = \mathcal{M}_r$), followed by averaging over all short-time frames in the segment. Segments classified as room-outside speech are then deleted from the SAD output of the given room.

6.2 Sohn's algorithm with SNR criterion

The first stage of this baseline employs the well-known and effective SAD algorithm of Sohn et al. As they detail in [36], the method is based on a likelihood ratio test between speech and noise models, considered as Gaussians in the frequency domain under an i.i.d. assumption in frequency and that of additive uncorrelated noise. Following noise model estimation using observed noise and of the necessary SNRs by a decision-directed approach, the likelihood ratio test is performed, and decision results are smoothed by means of an HMM-based hangover scheme [36].

In the designed baseline, Sohn's SAD is employed for each smart home room r , using a single ad hoc selected room microphone $m \in \mathcal{M}_r$. Then, at the second stage, for a first-stage generated segment in room r , the SNR of microphone m is compared to a global threshold; if below it, the particular segment is deleted from the room's SAD output. This baseline thus presents a well-established and relatively simple to implement approach for room-localized SAD.

7 Databases and experimental framework

We now proceed to describe the databases where the proposed system, its variations, and baselines are evaluated, as well as to discuss the adopted experimental framework and evaluation metrics used. In particular, the presentation refers to the experiments of Sections 8.1–8.4. An additional dataset and a slightly modified evaluation framework, necessary to allow comparisons with recent deep learning-based works, are detailed in the corresponding Section 8.5.

7.1 The DIRHA corpora

The experiments in Sections 8.1–8.4 are conducted on two databases: the Greek-language part of DIRHA-simcorpora II [61], hereafter referred to as "DIRHA-sim", and the "DIRHA-real" Greek corpus [16].² The datasets are either simulated or recorded inside a smart

²DIRHA-sim is found at <https://dirha.fbk.eu/simcorpora>, whereas DIRHA-real is available on request to the authors.

home apartment (with an average reverberation time of 0.72 s), developed for the purposes of the DIRHA research project [10]. Its floor plan is depicted in Figs. 1 and 6, showing that five of its rooms (living room, kitchen, bathroom, corridor, and bedroom) are equipped with a total of 40 microphones grouped in 14 arrays. Most arrays consist of two or three microphones (with linear topology) located on the room walls, while, for each of the living room and kitchen, a six-element pentagon-shaped array is also located at the ceiling. As a result, concerning the set of adjacent microphone pairs used in Sections 5.1.2 and 5.1.5, the two-element arrays provide one such pair, the three-element arrays two, and the pentagon-shaped arrays five, with all latter pairs containing the central array microphone. The corridor thus yields the least pairs (one), while the living room the most (ten).

As indicated by its name, the DIRHA-sim dataset contains audio simulations, produced as detailed in [61]. Briefly, first, about 9k room impulse responses are measured at each of 40 smart home microphones from 57 possible source locations uniformly distributed in the rooms of interest and with up to 8 source orientations for each (as shown in Fig. 6). These are then used to convolve high-quality, close-talk speech by 20 subjects (recorded at a 48-kHz sampling rate and an SNR average of 50 dB), while real, long-duration background noises and shorter acoustic events are also added to the resulting simulations. In total, 150 1-min simulation sequences containing speech and noise are available. In contrast, the DIRHA-real set

contains actual recordings of 5 subjects acquired by the 40 microphones inside the smart home under realistic noise conditions [16]. In total, 60 1-min recorded sequences of speech and noise are available. Statistics of the two sets are summarized in Table 1.

Apart from the main difference concerning the nature of the two sets (simulated vs. real), there exist two additional variations, as can be also observed in the waveform examples of Fig. 7. First, DIRHA-sim is characterized by more adverse noise conditions, containing more background noises and acoustic events besides speech. Further, in DIRHA-sim, speech often overlaps with other acoustic events or speech in different rooms of the smart home. Indeed, as listed in Table 1, speech overlap there reaches 47% (22 out of 47 min). These facts deem DIRHA-sim much more challenging for room-localized SAD than DIRHA-real.

7.2 Experimental framework and metrics

In the experiments of Sections 8.1–8.4, the DIRHA-sim dataset of 150 simulations is partitioned into a training set containing 75 of them and a test set with the remaining 75. Optimization of the first-stage algorithmic parameters of Section 4.3 (i.e., the transition penalty and constant prior added to the speech-class log-likelihood), as well as of the global threshold used in conjunction with Sohn's baseline, are performed on the training set. In the case of DIRHA-real, all 60 recordings are used for testing systems developed on the DIRHA-sim training data. This framework allows to also gauge the usefulness of simulated databases for training models and developing features and systems that can perform well in real-case scenarios, even when differences between the sets are significant.

For evaluation, the recall, precision, and F -score metrics are used, all computed at the frame level with a 10-ms time resolution and reported in percentage. Evaluation of room-localized SAD differs somewhat to the traditional room-independent case, as can be easily inferred from

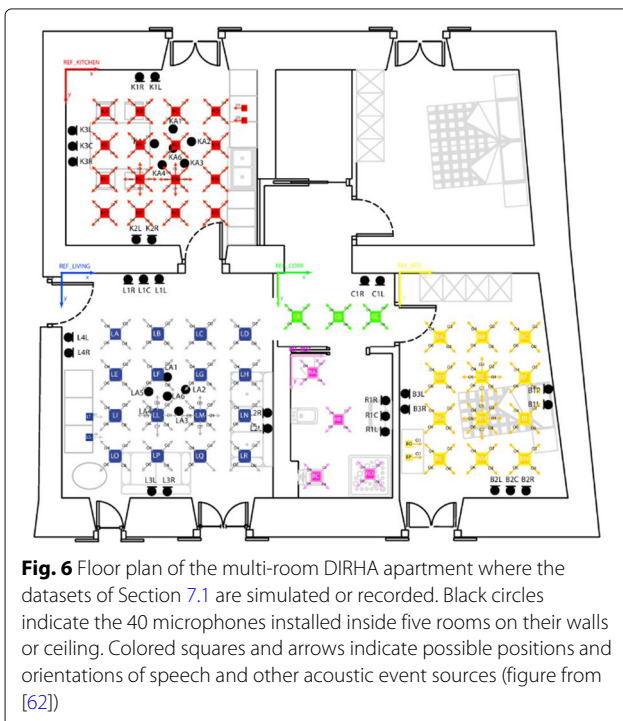


Fig. 6 Floor plan of the multi-room DIRHA apartment where the datasets of Section 7.1 are simulated or recorded. Black circles indicate the 40 microphones installed inside five rooms on their walls or ceiling. Colored squares and arrows indicate possible positions and orientations of speech and other acoustic event sources (figure from [62])

Table 1 Characteristics and statistics of the DIRHA-sim and DIRHA-real corpora, used in the experiments of this paper

Data characteristics	Databases	
	DIRHA-sim	DIRHA-real
Speech source	Loudspeaker	Human
1-min-long sequences (#)	150	60
Total speech (min)	47	19
Overlapped speech (min)	22	0
Non-speech events (#)	72	Untranscribed
Background noises (#)	10	Untranscribed
Subjects (#)	20	5
Average SNR (dB)	13	15

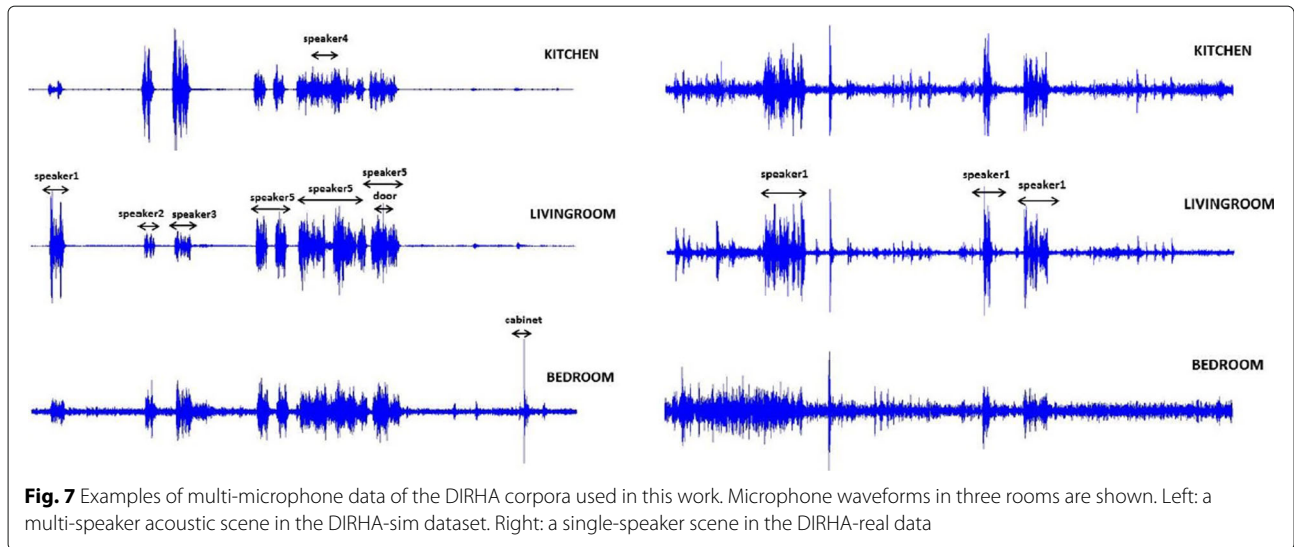


Fig. 7 Examples of multi-microphone data of the DIRHA corpora used in this work. Microphone waveforms in three rooms are shown. Left: a multi-speaker acoustic scene in the DIRHA-sim dataset. Right: a single-speaker scene in the DIRHA-real data

Fig. 1. In traditional SAD, the aim is to detect speech anywhere in the smart home, and as a result, each test-set sequence is evaluated only once (75 sequences for DIRHA-sim and 60 for DIRHA-real). In contrast, in the room-localized case, for each sequence, a total of $R = 5$ SAD outputs are evaluated (one for each room), with ground truth each time considering only speech occurring inside the given room. Thus, $75 \times 5 = 375$ and $60 \times 5 = 300$ SAD outputs in total are evaluated for the DIRHA-sim and DIRHA-real test sets, respectively. This affects the evaluation metrics: for example, recall for room-localized SAD is computed as the ratio between the number of correctly detected room-inside speech frames and the total number of such frames in the ground truth. In total, the test set contains 447 room-inside and 1788 room-outside speech segments in the DIRHA-sim case, and 232 and 928 segments, respectively, in DIRHA-real.

8 Experimental results

Next, we report our experiments. We first focus on room-independent SAD results, subsequently covering the room-localized case extensively. We also provide an error analysis of the proposed system, as well as a study on its robustness to the number of available microphones. We conclude the section with a comparison to recent deep learning-based approaches.

8.1 Room-independent SAD

Room-independent SAD is evaluated first, primarily to showcase its easier nature compared to the room-localized task, as well as to benchmark differences between the various techniques of Sections 4 and 6 and simple channel selection schemes. Results are reported in Table 2 for both DIRHA-sim and DIRHA-real sets in terms of recall, precision, and F -score.

Specifically, in the lower part of Table 2, both the GMM- and HMM-based decoding schemes of Section 4.3 are presented in conjunction with the six fusion techniques of Section 4.2, but for the room-independent SAD system variant discussed at the end of Section 4.4 that uses all 40 smart home microphones ($\mathcal{M} = \mathcal{M}_{\text{all}}$ and $\mathcal{J} = \{\text{sp}_{\text{all}}, \text{sil}_{\text{all}}\}$). These results are compared to two single-channel systems where microphone selection is driven by the best SNR per test-set sequence (actual based on ground-truth segmentation, or estimated), as well as the oracle-best channel result (that with the maximum F -score per sequence) and the average of all channel results. Finally, Sohn's algorithm is also considered, applied for each room (a single room microphone is used for each room), with the union of the results across rooms obtained.

For DIRHA-sim (left side of Table 2), we immediately observe the superiority of HMM-based Viterbi decoding over frame-based GMM segmentation. The best result is obtained by multi-channel fusion using log-likelihood summation scheme “w-sum”, achieving an F -score of 91.80%. This is significantly higher than Sohn's method (68.29%), and it represents a 53.5% relative error reduction in F -score compared to the best estimated SNR single-channel system (91.80% vs. 82.38%). Note that the latter performs similarly to the average of all channel results (82.69%), while it lags the ideal actual SNR case (90.10%) where channel SNR computations employ ground-truth information. These comparisons confirm that the challenging nature of DIRHA-sim adversely affects SNR estimation. Note finally that the best multi-channel system still lags the oracle-best channel one (95.73%), showing potential for further improvements.

Similar observations hold in the DIRHA-real case (right side of Table 2). The best system again employs log-

Table 2 Room-independent SAD results on the DIRHA-sim (left) and DIRHA-real (right) test sets, further discussed in Section 8.1

Method	DIRHA-sim						DIRHA-real						
	Recall		Precision		<i>F</i> -score		Recall		Precision		<i>F</i> -score		
	GMM	HMM	GMM	HMM	GMM	HMM	GMM	HMM	GMM	HMM	GMM	HMM	
Oracle-best	96.94	94.67	94.01	96.82	95.45	95.73	93.01	95.49	95.91	96.46	94.44	95.97	
Channel avg.	87.86	82.26	76.64	83.13	81.82	82.69	65.56	71.57	89.47	87.42	75.37	78.34	
Best act.-SNR	94.56	92.36	83.85	87.95	88.88	90.10	88.77	90.33	88.95	86.87	88.86	88.57	
Best est.-SNR	96.60	93.63	66.56	73.54	78.81	82.38	92.43	93.41	74.38	74.02	82.43	82.59	
Sohn's	81.22		58.91		68.29		78.05		61.51		68.80		
Decision fusion	"u-sum"	94.39	91.08	83.60	90.97	88.67	91.01	74.76	89.11	96.54	91.70	84.26	90.39
	"w-sum"	95.00	91.78	83.57	91.82	88.92	91.80	76.87	87.37	96.58	93.37	85.67	90.27
	"u-max"	74.17	82.51	75.28	73.69	74.72	77.85	45.66	68.40	97.21	95.01	62.14	79.54
	"w-max"	95.44	95.53	82.34	87.16	88.41	91.15	79.76	89.66	95.77	88.70	87.03	89.18
	"u-vote"	92.55	88.92	84.18	92.24	88.16	90.55	69.12	83.39	96.61	95.02	80.58	88.82
	"w-vote"	91.37	91.83	87.39	90.40	89.34	91.11	74.76	85.03	96.54	94.82	84.26	89.66

likelihood summation, with scheme "u-sum" reaching an *F*-score of 90.39%. This corresponds to a 44.8% relative *F*-score error reduction compared to the best estimated SNR single-channel system (90.39% vs. 82.59%). The latter performs now better than the average of all channel results (78.34%), and it lies somewhat closer to the best actual SNR system (88.57%) than in the DIRHA-sim case, due to the less adverse DIRHA-real environment. Note finally that, as above, the best multi-channel system lags the oracle-best channel result (95.97%).

8.2 Room-localized SAD results

We now switch focus to the room-localized SAD task. Our experiments are organized as follows: First, we evaluate the several possible choices of the system's first stage discussed in Section 4.4. Next, we investigate its second stage and the performance of the room discriminant features of Section 5.1. Finally, we present comparative results between our proposed system and the alternative baselines of Section 6.

The first experiment, reported in Table 3, compares the various design choices concerning the possible classes

Table 3 Effect of the various choices in the design of the system's first stage (discussed in Section 4.4) to the room-localized SAD performance on the DIRHA-sim test set

Oper	\mathcal{M}	Classes \mathcal{J}	Recall	Precision	<i>F</i> -score
RI	\mathcal{M}_{all}	{sp _{all} , sil _{all} }	72.30	56.63	63.51
RL	\mathcal{M}_r	{sp _r , sil _{all} }	72.07	61.08	66.12
		{sp _r , sil _r }	71.20	60.39	65.35
		{sp _r , sp _r , sil _{all} }	71.00	62.40	66.43

For consistency, the first stage is always followed by the second stage of the MFCC/GMM baseline of Section 6.1. RI denotes room-independent operation ("oper") of the first stage and RL room-localized one

and microphones used in the first stage of the room-localized SAD system, as summarized in Section 4.4. In all cases, decision fusion by means of log-likelihood summation scheme "u-sum" is employed across microphones. For consistency in the comparisons, the various first stages considered are always followed by an identical second stage, namely that of the MFCC/GMM baseline of Section 6.1.

It is clear from Table 3 that the room-independent scheme leads to the worst performance, trailing all room-localized variants. The basic reason is that in the latter schemes, the first stage can achieve high recall for room-inside speech and produces less room-outside segments compared to the room-independent case; thus, the second stage has an easier task. The second line of the table corresponds to the classes and microphone set options chosen in the proposed system. These yield the highest recall (72.07%) among the room-localized SAD variants, with an *F*-score second, but very close, to the three-class modeling approach of the last line (66.12% vs. 66.43%).

The second experiment, reported in Table 4, concentrates on the proposed room discriminant features of Section 5.1, as well as their feature fusion schemes of Section 5.2 and the SVM modeling approaches of Section 5.3 operating over entire segments. The evaluation is conducted for the room-inside vs. room-outside speech classification task of the second stage of the developed algorithm. For this purpose, the ground-truth speech boundaries are used, thus decoupling the comparisons from the first stage. Further, results include four rooms of the smart home, excluding the corridor ($R = 4$). Importantly, in addition to single features and their intra-room fusion (9), various feature subsets are also considered. Specifically, in Table 4, the best two, three,

Table 4 Performance of the room discriminant features of Section 5.1 and their combinations, in conjunction with inter-room fusion (Section 5.2) and SVM modeling (Section 5.3) for the room-inside vs. room-outside speech classification task of the second stage of the proposed algorithm

Set	SVM models	Feature (●)	Recall			Precision			F-score		
			$f_{r,\mathcal{T}}^{(\bullet)}$	$f_{r,avg,\mathcal{T}}^{(\bullet)}$	$f_{home,\mathcal{T}}^{(\bullet)}$	$f_{r,\mathcal{T}}^{(\bullet)}$	$f_{r,avg,\mathcal{T}}^{(\bullet)}$	$f_{home,\mathcal{T}}^{(\bullet)}$	$f_{r,\mathcal{T}}^{(\bullet)}$	$f_{r,avg,\mathcal{T}}^{(\bullet)}$	$f_{home,\mathcal{T}}^{(\bullet)}$
DIRHA-sim		(en)	63.97	37.93	40.06	50.51	86.03	86.92	56.45	52.65	54.84
		(coh)	47.46	87.41	88.66	67.90	77.01	76.05	55.87	81.88	81.87
		(ev)	82.89	90.81	90.38	78.01	74.85	76.28	80.37	82.06	82.74
		(ts)	71.91	86.00	89.35	52.21	74.46	79.28	60.50	79.82	84.01
	Room-specific	(srp)	76.76	79.85	79.25	53.94	56.44	60.94	63.36	66.13	68.90
		(ts,srp)	80.67	89.33	90.58	66.72	79.37	82.97	73.03	84.05	86.61
		(ts,srp,ev)	91.74	90.74	91.86	85.20	83.26	85.27	88.35	86.84	88.44
		(ts,srp,ev,coh)	90.62	90.42	92.27	83.65	84.96	85.80	86.99	87.61	88.92
		(en,coh,ev)[34]	89.48	87.65	90.37	78.90	81.16	81.69	83.86	84.28	85.81
		(all)	91.14	89.65	91.40	83.93	85.30	85.40	87.39	87.42	88.30
Global		91.12	92.21	n/a	78.49	79.63	n/a	84.34	85.46	n/a	
DIRHA-real		(en)	63.65	24.39	27.68	55.30	100.00	100.00	59.18	39.22	43.36
		(coh)	5.61	71.35	78.99	100.00	61.67	57.22	10.62	66.16	66.71
		(ev)	99.02	99.73	99.73	97.40	98.07	98.21	98.21	98.89	98.96
		(ts)	68.94	97.44	97.94	81.42	95.25	93.41	74.67	96.33	95.62
	Room-specific	(srp)	85.36	87.91	80.75	75.50	77.98	75.29	80.13	82.65	77.93
		(ts,srp)	90.28	94.52	97.33	91.58	95.32	86.76	90.92	94.92	91.74
		(ts,srp,ev)	99.90	98.82	97.81	99.82	97.87	97.24	99.86	98.34	97.53
		(ts,srp,ev,coh)	98.52	98.99	98.11	99.94	98.37	87.09	99.23	98.68	92.27
		(en,coh,ev)[34]	98.25	99.73	99.50	99.60	98.64	90.84	98.92	99.18	94.98
		(all)	98.89	98.85	95.68	99.94	98.46	80.21	99.42	98.66	87.26
Global		99.33	100.00	n/a	100.00	99.84	n/a	99.66	99.92	n/a	

Results are reported on $R = 4$ rooms of the DIRHA smart home (excluding the corridor) on the DIRHA-sim (top) and DIRHA-real (bottom) test sets using ground-truth speech segment boundaries. All SVMs operate over entire segments

and four feature combinations are listed, as selected by wrapper-based sequential forward feature selection [76, ch. 5.7.2] that is conducted on DIRHA-sim (based on the corresponding proposed system F -scores). In addition, the three-feature subset of our previous work [34] is evaluated. Notice that the notation in (10) and (11) is slightly extended to allow inter-room fusion of single features and subsets.

Concerning DIRHA-sim (Table 4, top), in the case of room-specific SVMs, we observe that for most individual features of Section 5.1 (with the exception of the energy-based one), performance improves by inter-room fusion. The best feature is the proposed spectrogram texture smoothness, achieving an F -score of 84.01% after fusion by (10). In contrast, the energy-based feature performs the worst at a 52.65% F -score after fusion by (11). For the entire feature vector (“all”) obtained by intra-room fusion (9), small differences are observed between no room combination and inter-room fusion by (10) or

(11), with the best F -score reaching 88.30%. Global SVM modeling performs slightly worse (85.46% F -score with fusion (11)).

Regarding feature subsets, the best two-feature set consists of the spectrogram texture smoothness and the SRP-based feature; envelope variance is then added to yield the best three-member set; and subsequently, the coherence-based one is chosen. All subsets demonstrate better performance than individual features, when fused by (10) or (11). Also, we can observe that energy does not boost performance further, as the best four-feature set slightly outperforms the “all” set, achieving an 88.92% vs. 88.30% F -score with fusion (10). Finally, compared to our previous work [34], the “all” set achieves a 17.5% relative error reduction in F -score (88.30% vs. 85.81% with (10)).

In the less challenging DIRHA-real set (Table 4, bottom), the coherence, envelope variance, and spectrogram texture smoothness features take advantage of inter-room combination, whereas the energy- (as also on DIRHA-

sim) and SRP-based ones fail to do so. The highest performing feature is the envelope variance with an F -score of 98.96% after fusion by (10), closely followed by spectrogram texture smoothness at 96.33% after fusion by (11). For the entire feature vector (“all”) obtained by intra-room fusion (9), small differences are observed between no room combination and inter-room fusion by (11), regardless of the SVM models used. However, concatenation across all rooms by (10) fails to improve matters (an F -score of only 87.26% is attained). This is probably due to the high dimensionality of the resulting vector and the use of multiple SVMs, in conjunction with the mismatch between the DIRHA-sim trained models and DIRHA-real test conditions. This seems also supported by the fact that inter-room fusion by means of (11) in most cases outperforms (10). Nevertheless, the best “all” feature system reaches an almost perfect F -score of 99.92%, obtained by global SVMs and fusion (11). Note also that this is very close to the 99.86% F -score of the spectrogram texture smoothness-SRP-envelope variance combination with no inter-room fusion.

As a complement to this experiment and to further gain insights into the room discriminant features, their correlation is investigated. For this purpose, the Pearson correlation coefficient is computed among all features over the speech segments of the DIRHA-sim test set, resulting in the matrix of Fig. 8. As expected, the envelope variance, spectrogram texture smoothness, and coherence-based features demonstrate high correlation between them, as they are all related to reverberation. On the contrary, the energy- and SRP-based ones exhibit low correlation with all features.

In the third experiment, reported in Table 5, once again ground-truth segments are considered as input to the

second stage. The aim here is threefold: first, to showcase the superiority of the proposed room discriminant feature approach over the baselines of Section 6; second, to highlight performance differences among the various smart home rooms; and third, to further compare the fusion schemes of Section 5.2. Specifically, the MFCC/GMM-based second stage of the baseline of Section 6.1 is listed in the first line of Table 5, followed by the SNR-based room assignment scheme of Section 6.2, as well as room-specific SVM modeling on (9), (11), and (10) operating over entire segments. F -scores are reported for each room separately (no corridor F -score is shown for DIRHA-real, as there are no ground-truth room-inside segments there), as well as for all four (excluding the corridor) or five rooms.

It is clear from Table 5 that the proposed approach dramatically outperforms the baselines, e.g., for $R = 5$, on DIRHA-sim, the best result (84.26%) represents a 46.7% and 73.2% relative error reduction over the baselines of Sections 6.1 and 6.2, respectively, while on DIRHA-real, the corresponding reductions of the best result (93.34%) stand at 78.6% and 87.8% relative. It is also clear that the corridor is a challenging room, as seen by its low DIRHA-sim F -scores and the performance drop from the $R = 4$ to the $R = 5$ case. This is primarily due to its central location in the smart home floor plan (see also Fig. 6) exposing it to sounds coming from all other rooms, as well as the small number of microphones in it (only two). Regarding the multi-room results of the feature fusion schemes of Section 5.2, inter-room feature concatenation (10) performs best on DIRHA-sim, followed by (11). This can be expected as (10) captures more detailed information (albeit at higher dimensionality). Similarly, fusion (11) is superior to the lack of inter-room combination in (9). On DIRHA-real, however, the above are reversed, as features (9) outperform (11) and, in turn, fusion by (10). This is primarily due to the mismatch of the DIRHA-sim trained SVMs to the DIRHA-real conditions, thus favoring lower-dimensional representations that generalize better, as also observed in Table 4.

Finally, Table 6 reports on the full task of room-localized SAD. Its upper part covers single-stage methods, namely the best room-independent approach (“best RI”), as well as the first stages of the MFCC/GMM baseline of Section 6.1 (recall that this is identical to the proposed system’s first stage) and Sohn’s algorithm (Section 6.2). The complete two-stage baselines are evaluated next, followed by the proposed algorithm employing room-specific SVMs on features (10) operating over the entire segments (“seg”) or over sliding windows (“win”), where results both with and without the corridor are reported.

As shown in Table 6, the proposed system operating over sliding windows reaches satisfactory performance,

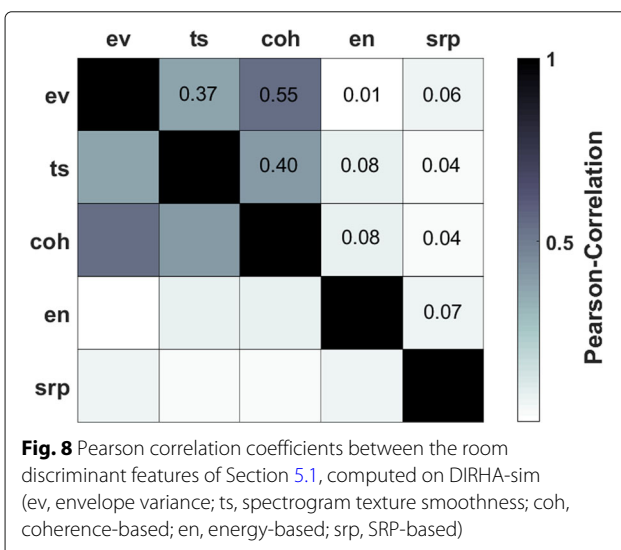


Table 5 Comparison of the two baselines of Section 6 (upper part) and the room discriminant feature-based approach (lower part) for the room-inside vs. room-outside speech classification task

Features	DIRHA-sim					DIRHA-real								
	Single room					Multi-room			Single room				Multi-room	
	Liv.	Kitch.	Bath.	Bed.	Corr.	$R = 4$	$R = 5$	Liv.	Kitch.	Bath.	Bed.	$R = 4$	$R = 5$	
MFCCs	70.96	72.52	61.25	76.94	39.03	72.32	70.49	46.93	71.50	80.98	58.11	68.91	68.91	
SNR	55.59	57.57	17.38	50.75	8.80	48.59	41.22	41.47	72.99	53.42	31.63	53.74	45.54	
$f_{r,\mathcal{T}}^{(all)}$	83.74	92.83	83.33	86.76	34.28	87.39	80.63	97.17	100.00	99.89	100.00	99.42	93.34	
$f_{r,avg,\mathcal{T}}^{(all)}$	84.80	92.83	84.48	85.05	38.11	87.42	81.46	99.50	97.51	99.89	99.16	98.66	89.94	
$f_{home,\mathcal{T}}^{(all)}$	86.29	89.96	91.67	88.25	39.66	88.30	84.26	97.88	79.23	95.00	93.63	87.26	79.19	

F -scores are reported for each room, as well as over $R = 4$ rooms (excluding the corridor) and all $R = 5$ rooms of the DIRHA smart home, on both DIRHA-sim (left) and DIRHA-real (right) test sets using ground-truth speech segment boundaries. Room-specific SVMs are employed, operating over entire segments

namely a 80.98% F -score on DIRHA-sim and 87.68% on DIRHA-real, which are further improved if the corridor is excluded. The algorithm clearly outperforms the two-stage baselines dramatically, resulting to relative error reductions of 43.9% and 73.2% on DIRHA-sim compared to the methods of Sections 6.1 and 6.2, respectively. The corresponding improvements stand at 44.3% and 82.4% on DIRHA-real. The single-stage systems considered perform even worse. Not surprisingly, the addition of the second stage helps both baselines, especially the MFCC/GMM system.

Concerning the operation of the second stage over entire segments vs. sliding windows, it can be observed in Table 6 that the latter scheme fares slightly better on the more challenging DIRHA-sim dataset. An example of its superiority is provided in Fig. 9 (same as in Fig. 7 (left)). There, the kitchen SAD results are shown for a case of two overlapping speakers located inside different rooms (“speaker 5” in the living room and “speaker 4” in the kitchen). The first stage of the system returns a segment containing both. Then, at the second stage, the segment-operating scheme classifies it entirely as kitchen-inside speech, whereas the sliding window one

allows to only keep the part belonging to “speaker 4”. Further, both schemes delete three erroneous first-stage segments, but fail to do so for two that originate in the living room. However, on the less challenging DIRHA-real set, a slightly worse performance for the sliding-window scheme is observed in Table 6. This can be attributed to the lack of overlapping speech segments originating in different rooms, in conjunction with the obvious fact that window-based decisions rely on less data than entire segments.

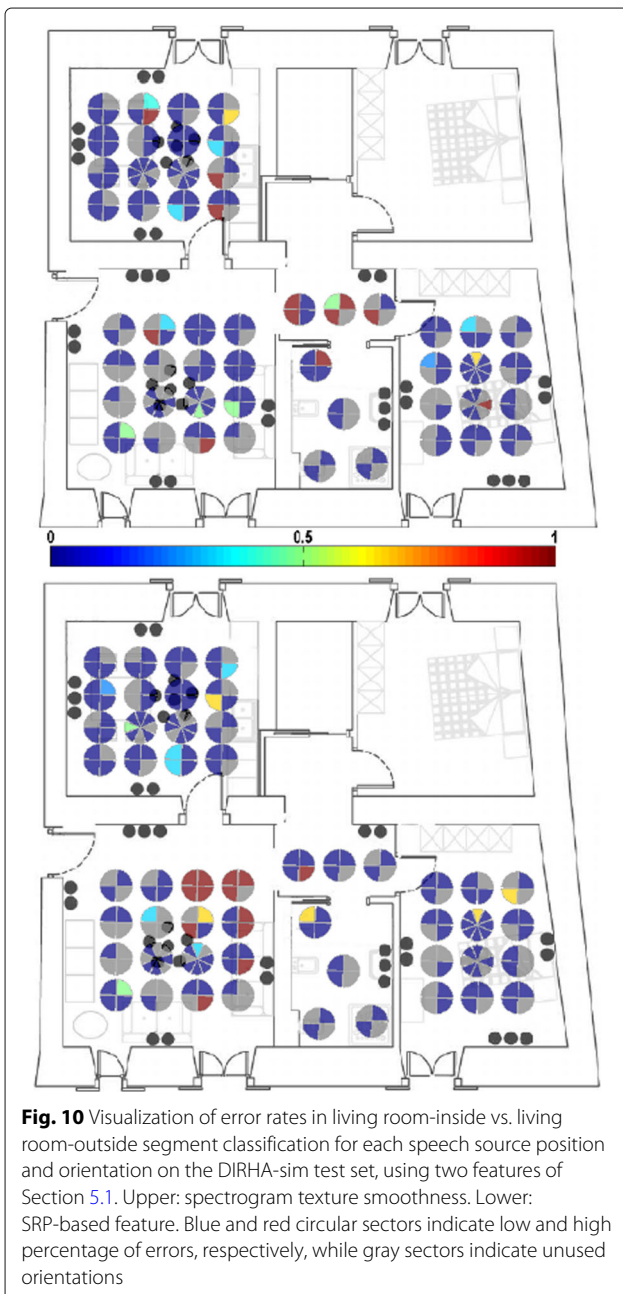
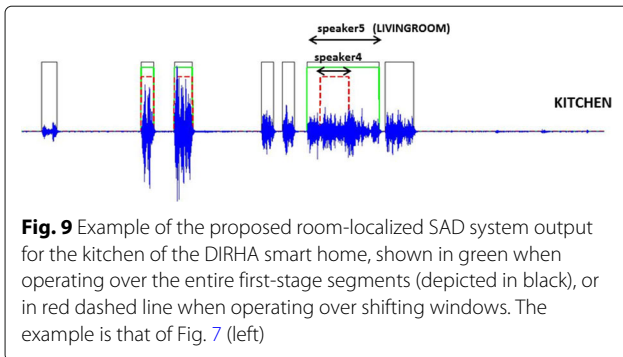
8.3 Error analysis

This section attempts to provide additional insights into the performance of the various room discriminant features of Section 5.1. In particular, the focus lies on how such is affected by the speech source location and the amount of overlap in the detected segment.

Figure 10 concentrates on the two novel features proposed, namely the spectrogram texture smoothness (upper part) and the SRP-based feature (lower figure). There, in the case of segments with ground-truth boundaries and no overlap, performance of the features for living room-inside vs. living room-outside classification on DIRHA-sim data is visualized by an appropriate

Table 6 Performance of various approaches for the full task of room-localized SAD on DIRHA-sim (left) and DIRHA-real (right)

Method		DIRHA-sim			DIRHA-real		
		Recall	Precision	F -score	Recall	Precision	F -score
Single-stage	Best RI	92.22	19.49	32.18	92.71	16.25	27.66
	MFCC/GMM	89.87	41.60	56.87	88.02	57.06	69.24
	Sohn’s	73.17	17.33	28.02	73.40	17.71	28.53
Two-stage	MFCC/GMM	72.07	61.08	66.12	78.94	76.87	77.89
baselines	Sohn’s	43.14	21.96	29.11	46.39	22.26	30.08
Proposed	Seg ($R = 5$)	82.16	77.35	79.68	88.27	89.30	88.78
	Win ($R = 5$)	83.09	78.96	80.98	86.51	88.87	87.68
	Win ($R = 4$)	84.65	86.10	85.37	86.51	94.03	90.11



coloring scheme within circular sectors that correspond to the various speech source positions and orientations in the smart home (blue indicates low misclassification rates, while red high ones). It can be observed that errors mostly occur around the living room boundaries, but differ across features. For example, the spectrogram texture smoothness misclassifies mainly segments of adjacent rooms with orientation towards the living room doors, as they reach its microphones with less reverberation. In contrast, the SRP feature classifies such correctly, as they produce high acoustic energy at the living room doors. However, it misclassifies room-inside segments near these doors.

Finally, Fig. 11 aims to quantify the effects of overlap to the room discrimination performance of the various features. For this purpose, two cases are considered: “low overlap” concerning speech segments with less than 30% of overlap with acoustic events of other rooms, and “high overlap” with more than 30%. Performance is measured in frame-based F -score, using ground-truth first-stage (room-independent) speech boundaries. In all single-feature sets, five-dimensional vectors are produced (one feature per room). Clearly, most sets exhibit low performance in the high overlap condition, with some (spectrogram texture smoothness, energy-based, and fused features) affected more.

8.4 Robustness to reduced microphone setups

The proposed room-localized SAD algorithm relies on the availability of multiple microphones in the multi-room DIRHA apartment. As this installation includes 40 microphones, the question naturally arises as to how dependent the system is on such an expensive setup.

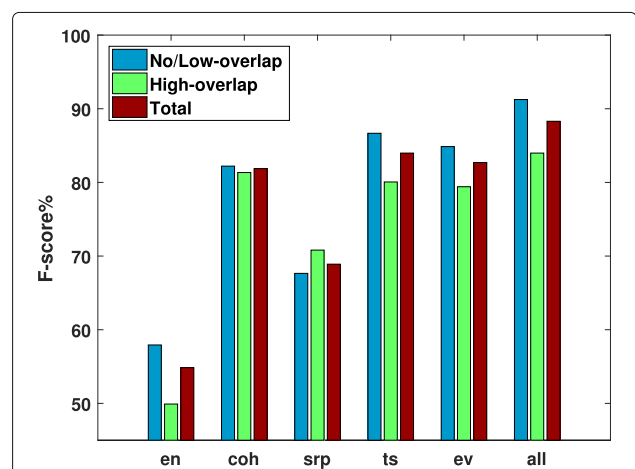


Fig. 11 Performance of the room discriminant features of Section 5.1 in classifying speech segments exhibiting low or high overlap with audio events in other rooms for the DIRHA-sim test set (en, energy-based; coh, coherence-based; srp, SRP-based; ts, spectrogram texture smoothness; ev, envelope variance; all, intra-room fusion (9)) In all cases, inter-room fusion (10) and room-specific SVMs are used

To investigate this, four reduced “nested” setups are considered, gradually decreasing the number of smart home microphones $|\mathcal{M}_{\text{all}}|$ from 40 down to 5, specifically to $|\mathcal{M}_{\text{all}}| = 25, 16, 10,$ and 5 , as depicted in Fig. 12 (compare to the original configuration of Figs. 1 and 6). Note that the $|\mathcal{M}_{\text{all}}| = 10$ setup includes one microphone pair in each room, while the $|\mathcal{M}_{\text{all}}| = 5$ configuration only one microphone per room. For the latter, coherence- and SRP-based features cannot be computed due to the absence of microphone pairs, thus reducing the set of available features to three per room (see also Fig. 13).

The first experiment, summarized in Table 7, quantifies the effects of reduced microphone setups to the GMM-HMM based SAD module. Specifically, room-independent SAD performance on the DIRHA-sim test set is reported (see also Table 2), employing HMM-based Viterbi decoding and “w-sum” decision fusion over the microphones of the various setups. To further reduce system complexity, a simplified modeling approach is also evaluated, where only a single GMM is trained on data of a specific microphone, in place of microphone-specific models. In particular, the living room ceiling central microphone, available in all configurations, is used for this purpose. In that case, (1) and (2) are slightly modified by setting $b_{m,j}(\mathbf{o}_{m,t}) \leftarrow b_{M_j}(\mathbf{o}_{m,t})$, for all $m \in \mathcal{M}$, where M denotes the specific GMM-training microphone.

Concerning SAD performance, it is evident from Table 7 that it remains robust to the number of available microphones. In particular, the F -score degrades gracefully and monotonically as the installation becomes leaner: In the microphone-specific modeling case, the full-setup F -score of 91.80% reduces to 89.60% for $|\mathcal{M}_{\text{all}}| = 5$, exhibiting an absolute drop of only 2.2%. A similar trend is also observed in the single-GMM case. Further, comparing the two modeling approaches, the single-GMM one yields small only F -score absolute degradations within the 1.7 to 2.6% range (depending on the setup). Thus, in the lack of multi-channel training data, a single-microphone model constitutes a viable approach leading to satisfactory results.

In the second experiment, depicted in Fig. 13, the performance of the room discriminant features of Section 5.1 is examined as a function of the number of available microphones. For this purpose, the room-inside vs. room-outside classification task (second stage of the algorithm) is considered with ground-truth segmentation on the DIRHA-sim test set. It can be readily noted that reduced setups have a noticeable, albeit not dramatic, effect on the performance of the intra-room fused features (“all”), degrading the full-setup F -score of 88.30% ($|\mathcal{M}_{\text{all}}| = 40$) to 85.10% for $|\mathcal{M}_{\text{all}}| = 16$ and 80.86% for $|\mathcal{M}_{\text{all}}| = 5$. Thus, the multi-channel second stage can benefit from larger microphone numbers, but can also perform satisfactorily with fewer microphones. Regarding individual feature performance in reduced setups, the envelope variance seems the most robust, while the SRP-based feature the least.

8.5 Comparison to deep learning approaches

As overviewed in Section 2, a number of works on room-localized SAD have appeared recently, proposing single-stage algorithms based on deep learning methods [29–32]. In this section, a performance comparison to our developed system is provided.

To enable such comparison, the experimental framework of these works is followed, deviating from that of Section 7. In particular, the corpus used is the Italian-language part of the DIRHA simcorpora (DIRHA-sim-evalita), first introduced as part of the SASLODOM evaluation campaign at the EVALITA’14 workshop [77]. This contains 80 1-min simulations, generated in the DIRHA apartment as discussed in Section 7.1. Experiments are conducted by tenfold cross-validation to reduce performance variance, with each test fold containing eight simulations. Results are reported in terms of the “overall SAD detection error” metric, as defined in [77], considering only two rooms ($R = 2$) of the DIRHA apartment, i.e., living room and kitchen.

Comparative results between the best deep learning results of [29–32] and our proposed algorithm are presented in Table 8. In particular, the best system of

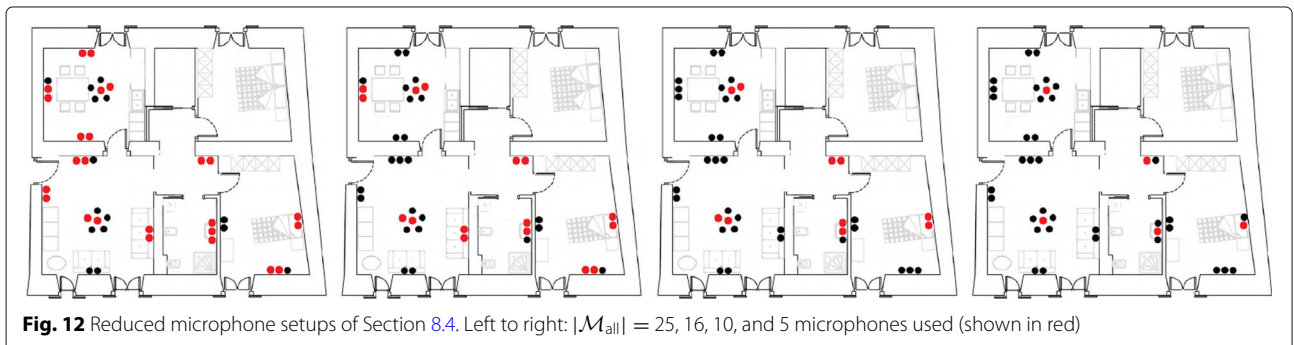


Fig. 12 Reduced microphone setups of Section 8.4. Left to right: $|\mathcal{M}_{\text{all}}| = 25, 16, 10,$ and 5 microphones used (shown in red)

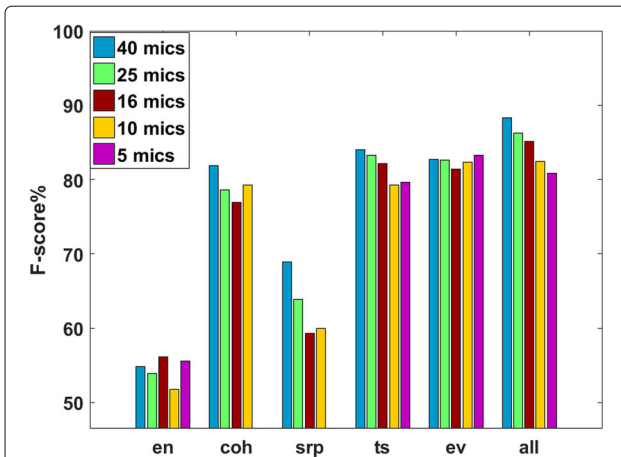


Fig. 13 Performance of the room discriminant features of Section 5.1 for the speech-inside vs. speech-outside classification task with ground-truth segmentation on the DIRHA-sim test set, using various numbers of microphones (en, energy-based; coh, coherence-based; srp, SRP-based; ts, spectrogram texture smoothness; ev, envelope variance; all, intra-room fusion (9)) Inter-room fusion (10) and room-specific SVMs are used

[29, 30], employing a DNN over 187-dimensional features of various types that are extracted from the best microphone per room, yields a SAD error of 5.8%. Further, the 3D-CNN system of [31], operating on 40-dimensional log-Mel filterbank energies after temporal splicing and combining information from the three best microphones per room, exhibits a 7.0% SAD error. This improves to 5.2%, when employing all microphones available in the two rooms [32]. Finally, a 3.5% SAD error is reported in [32], when the aforementioned 3D-CNN is extended incorporating 51-dimensional GCC-PHAT patterns [70] to jointly provide SAD and speaker location estimates (marked as “SAD+SLOC” in the table). However, it should be noted that this system employs additional information during its training, in the form of ground-truth speaker positions (in 2D room coordinates).

Table 7 Room-independent SAD results on the DIRHA-sim test set, employing all available microphones ($|\mathcal{M}_{\text{all}}| = 40$) or the reduced setups of Fig. 12

$ \mathcal{M}_{\text{all}} $	Microphone-specific GMMs			Single-microphone GMM		
	Recall	Precision	<i>F</i> -score	Recall	Precision	<i>F</i> -score
40	91.78	91.82	91.80	91.21	87.25	89.19
25	91.45	91.41	91.43	90.66	87.58	89.09
16	91.50	90.89	91.19	87.51	90.69	89.07
10	90.84	89.39	90.11	90.33	86.42	88.33
5	88.22	91.02	89.60	89.21	86.61	87.89

In all cases, HMM-based Viterbi decoding and “w-sum” decision fusion are used, where the combined log-likelihoods result from microphone-specific GMMs (left) or a GMM trained on a single microphone (right)

Table 8 Performance (in overall SAD detection error [77]) of deep learning-based approaches vs. the proposed algorithm for room-localized SAD on the DIRHA-sim-evalita corpus

Method		SAD error(%)
Deep learning	DNN [30]	5.8
	3D-CNN [31]	7.0
	3D-CNN [32]	5.2
	3D-CNN (SAD+SLOC) [32]	3.5
Proposed	Seg	5.7
	Win	4.7

In comparison, our proposed algorithm exhibits SAD errors of 5.7% and 4.7%, when operating over entire segments (“seg”) or sliding windows (“win”), respectively. The latter represents a 19% relative SAD error reduction over the DNN of [30] and 10% over the 3D-CNN of [32], proving better than segment-based operation in the challenging and noisy DIRHA-sim-evalita data (as also observed in Table 6 for DIRHA-sim). These comparisons highlight the competitiveness of our two-stage system and the suitability of the five room discriminant features of its second stage. Of course, it is possible that the deep learning methods could have gained advantage if more training data had been available in the DIRHA corpora.

9 Conclusions

We have presented an efficient multi-channel, two-stage approach to address speech activity detection in multi-room smart home environments, equipped with multiple microphone arrays distributed inside them. In the general scenario, possibly, concurrent speech activity in different rooms needs to be detected and the effect of cross-room interference suppressed. For this purpose, the proposed room-localized SAD system first employs a multi-channel speech/non-speech segmentation module per room, and it subsequently determines whether detected speech activity occurs inside or outside each room by utilizing a novel set of room discriminant features. Experiments on a suitable multi-room, multi-channel dataset demonstrate satisfactory performance on both simulated and real data, reaching *F*-scores of 81.0% and 87.7%, respectively, while significantly outperforming alternatives that combine well-known baselines and features (MFCCs, Sohn’s SAD, SNR), as well as comparing favorably to deep learning-based approaches (DNNs, CNNs). The evaluation results verify the robustness of the two-stage system and the suitability of the devised hand-crafted features, while also highlighting the realistic design and value of the current simulated database for developing algorithms that generalize well to real recorded data.

Abbreviations

ASR: Automatic speech recognition; CNN: Convolutional neural network; DTFT: Discrete-time Fourier transform; DNN: Deep neural network; GMM: Gaussian mixture model; HMM: Hidden Markov model; MFCCs: Mel-frequency cepstral coefficients; RI: Room-independent; RL: Room-localized; SAD: Speech activity detection; SNR: Signal-to-noise ratio; SRP: Steered response power; SVM: Support vector machine

Authors' contributions

All authors have contributed equally to this work. All authors have read and approved the final manuscript.

Funding

P. Maragos and G. Potamianos' work was partially supported by the European Regional Development Fund of the EU and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call 'Research - Create - Innovate' (project T1EDK-02890, "e-Prevention").

Availability of data and materials

DIRHA-sim is found at <https://dirha.fbku.eu/simcorpora>, whereas DIRHA-real is available on request to the authors.

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece. ²Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece. ³Athena Research and Innovation Center, Marousi, Greece.

Received: 17 October 2018 Accepted: 15 July 2019

Published online: 27 August 2019

References

1. M. Chan, E. Campo, D. Estève, J.-Y. Fourniols, Smart homes – current features and future perspectives. *Maturitas*. **64**(2), 90–97 (2009)
2. M. P. Poland, C. D. Nugent, H. Wang, L. Chen, Smart home research: projects and issues. *Int. J. Ambient Comput. Intell.* **1**(4), 32–45 (2009)
3. D. Ding, R. A. Cooper, P. F. Pasquina, L. Fici-Psquina, Sensor technology for smart homes. *Maturitas*. **69**(2), 131–136 (2011)
4. M. R. Alam, M. B. I. Reaz, M. Mohd Ali, A review of smart homes – past, present, and future. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **42**(6), 1190–1203 (2012)
5. M. Amirbesheli, A. Benmansour, A. Bouchachia, A review of smart homes in healthcare. *J. Ambient Intell. Humanized Comput.* **6**(4), 495–517 (2015)
6. M. Matassoni, M. Omologo, R. Manione, T. Sowa, R. Balchandran, M. E. Epstein, L. Seredi, in *Proc. International Conference on Intelligent Information Systems (IIS)*. The DICIT project: an example of distant-talking based spoken dialogue interactive system, (2008), pp. 527–533
7. A. Badii, J. Boudy, in *Proc. Congrès Société Française des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG)*. CompanionAble - integrated cognitive assistive and domestic companion robotic systems for ability and security, (2009), pp. 18–20
8. G. L. Filho, T. J. Moir, From science fiction to science fact: a smart-house interface using speech technology and a photo-realistic avatar. *Int. J. Comput. Appl. Technol.* **39**(1/2/3), 32–39 (2010)
9. M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuaara, S. Méniard, in *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. The SWEET-HOME project: audio technology in smart homes to improve well-being and reliance, (2011), pp. 5291–5294
10. DIRHA: Distant-speech interaction for robust home applications. <http://dirha.fbku.eu>. Accessed 22 Apr 2019
11. J. F. Gemmeke, B. Ons, N. Tessema, H. Van hamme, J. van de Loo, G. De Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. Van Den Broeck, P. Karsmakers, B. Vanrumste, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Self-taught assistive vocal interfaces: an overview of the ALADIN project, (2013), pp. 2039–2043
12. M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, P. Chahuaara, Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation. *ACM Trans. Accessible Comput.* **7**(2:5), 1–36 (2015)
13. M. Malavasi, E. Turri, J. J. Atria, H. Christensen, R. Marxer, L. Desideri, A. Coy, F. Tamburini, P. Green, An innovative speech-based user interface for smart homes and IoT solutions to help people with speech and motor disabilities. *Stud. Health Technol. Inform.* **242**, 306–313 (2017)
14. V. Kēpuska, G. Bohouta, in *Proc. IEEE Annual Computing and Communication Workshop and Conference (CCWC)*. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home), (2018), pp. 99–103
15. E. Principi, S. Squartini, F. Piazza, D. Fuselli, M. Bonifazi, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. A distributed system for recognizing home automation commands and distress calls in the Italian language, (2013), pp. 2049–2053
16. I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, P. Maragos, Room-localized spoken command recognition in multi-room, multi-microphone environments. *Comput. Speech Lang.* **46**, 419–443 (2017)
17. E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, F. Piazza, An integrated system for voice command recognition and emergency detection based on audio signals. *Expert Syst. Appl.* **42**(13), 5668–5683 (2015)
18. R. C. Rose, H. K. Kim, in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems, (2003), pp. 198–203
19. D. K. Freeman, G. Cosier, C. B. Southcott, I. Boyd, in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. The voice activity detector for the Pan-European digital cellular mobile telephone service, (1989), pp. 369–372
20. A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, J.-P. Petit, ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Commun. Mag.* **35**(9), 64–73 (1997)
21. ETSI EN 301 708 V7.1.1: Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels (GSM 06.94 version 7.1.1 Release 1998) (ETSI, France, 1999)
22. D. Enqing, Z. Heming, L. Yongli, in *Proc. IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCOM)*, vol. 1. Low bit and variable rate speech coding using local cosine transform, (2002), pp. 423–426
23. D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* **10**(1), 19–41 (2000)
24. T. Kinnunen, P. Rajan, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data, (2013), pp. 7229–7233
25. J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, G. Kubin, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Room localization for distant speech recognition, (2014), pp. 2450–2453
26. Y. Tachioka, T. Narita, S. Watanabe, J. Le Roux, in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. Ensemble integration of calibrated speaker localization and statistical speech detection in domestic environments, (2014), pp. 162–166
27. A. Abad, M. Matos, H. Meinedo, R. F. Astudillo, I. Trancoso, in *Proc. Italian Conference on Computational Linguistics (CLIC-it) and International Workshop EVALITA*. The L2F system for the EVALITA-2014 speech activity detection challenge in domestic environments, (2014), pp. 147–152
28. A. Brutti, M. Ravanelli, P. Svaizer, M. Omologo, in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. A speech event detection and localization task for multiroom environments, (2014), pp. 157–161
29. G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, F. Piazza, in *Proc. International Joint Conference on Neural Networks (IJCNN)*. A deep neural network approach for voice activity detection in multi-room domestic scenarios, (2015), pp. 1–8

30. F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, F. Piazza, in *Proc. International Joint Conference on Neural Networks (IJCNN)*. Deep neural networks for multi-room voice activity detection: advancements and comparative evaluation, (2016), pp. 3391–3398
31. P. Vecchiotti, F. Vesperini, E. Principi, S. Squartini, F. Piazza, in *Multidisciplinary Approaches to Neural Computing, vol. SIST-69*, ed. by A. Esposito, M. Faudez-Zanuy, F. C. Morabito, and E. Pasero. Convolutional neural networks with 3-D kernels for voice activity detection in a multiroom environment (Springer, Cham, 2018), pp. 161–170
32. P. Vecchiotti, E. Principi, S. Squartini, F. Piazza, in *Proc. European Signal Processing Conference (EUSIPCO)*. Deep neural networks for joint voice activity detection and speaker localization, (2018), pp. 1567–1571
33. P. Giannoulis, A. Tsiami, I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Maragos, in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. The Athena-RC system for speech activity detection and speaker localization in the DIRHA smart home, (2014), pp. 167–171
34. P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, P. Maragos, in *Proc. European Signal Processing Conference (EUSIPCO)*. Multi-room speech activity detection using a distributed microphone network in domestic environments, (2015), pp. 1271–1275
35. P. Giannoulis, G. Potamianos, A. Katsamanis, P. Maragos, in *Proc. European Signal Processing Conference (EUSIPCO)*. Multi-microphone fusion for detection of speech and acoustic events in smart spaces, (2014), pp. 2375–2379
36. J. Sohn, N. S. Kim, W. Sung, A statistical model-based voice activity detection. *IEEE Signal Process. Lett.* **6**(1), 1–3 (1999)
37. S. Graf, T. Herbig, M. Buck, G. Schmidt, Features for voice activity detection: a comparative analysis. *EURASIP J. Adv. Signal Process.* **2015**(91), 1–15 (2015)
38. J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information. *Speech Comm.* **42**(3–4), 271–287 (2004)
39. B. Kotnik, Z. Kacic, B. Horvat, in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*. A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm, (2001), pp. 197–200
40. C. Shahnaz, W.-P. Zhu, M. O. Ahmad, in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*. A multifeature voiced/unvoiced decision algorithm for noisy speech, (2006), pp. 2525–2528
41. R. Tucker, Voice activity detection using a periodicity measure. *IEEE Proc. I Commun. Speech Vis.* **139**(4), 377–380 (1992)
42. T. Kristjansson, S. Deligne, P. Olsen, in *Proc. Conference of the International Speech Communication Association (Interspeech)*. Voicing features for robust speech detection, (2005), pp. 369–372
43. S. O. Sadjadi, J. H. L. Hansen, Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Process. Lett.* **20**(3), 197–200 (2013)
44. L. R. Rabiner, M. R. Sambur, Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Trans. Acoust. Speech Signal Proc.* **25**(4), 338–343 (1977)
45. J. A. Haigh, J. S. Mason, in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*. A voice activity detector based on cepstral analysis, (1993), pp. 1103–1106
46. P. K. Ghosh, A. Tsiartas, S. Narayanan, Robust voice activity detection using long-term signal variability. *IEEE Trans. Audio Speech Lang. Process.* **19**(3), 600–613 (2011)
47. Y. Ma, A. Nishihara, Efficient voice activity detection algorithm using long-term spectral flatness measure. *EURASIP J. Audio Speech Music Process.* **2013**(21), 1–18 (2013)
48. A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, S. S. Narayanan, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Multi-band long-term signal variability features for robust voice activity detection, (2013), pp. 718–722
49. N. Mesgarani, M. Slaney, S. A. Shamma, Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Trans. Audio Speech Lang. Process.* **14**(3), 920–930 (2006)
50. G. Evangelopoulos, P. Maragos, Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. Audio Speech Lang. Process.* **14**(6), 2024–2038 (2006)
51. J.-H. Bach, B. Kollmeier, J. Anemüller, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Modulation-based detection of speech in real background noise: generalization to novel background classes, (2010), pp. 41–44
52. X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection. *IEEE Trans. Audio Speech Lang. Process.* **21**(4), 697–710 (2013)
53. X.-L. Zhang, D. Wang, Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(2), 252–264 (2016)
54. T. Hughes, K. Mierle, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Recurrent neural networks for voice activity detection, (2013), pp. 7378–7382
55. F. Eyben, F. Wenginger, S. Squartini, B. Schuller, in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies, (2013), pp. 483–487
56. S. Thomas, S. Ganapathy, G. Saon, H. Soltau, in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions, (2014), pp. 2519–2523
57. I. McLoughlin, Y. Song, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Low frequency ultrasonic voice activity detection using convolutional neural networks, (2015), pp. 2400–2404
58. S.-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, O. Vinyals, in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Temporal modeling using dilated convolution and gating for voice-activity-detection, (2018), pp. 5549–5553
59. Y. Jung, Y. Kim, Y. Choi, H. Kim, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Joint learning using denoising variational autoencoders for voice activity detection, (2018), pp. 1210–1214
60. R. Zazo, T. N. Sainath, G. Simko, C. Parada, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Feature learning with raw-waveform CLDNNs for voice activity detection, (2016), pp. 3668–3672
61. L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, P. Maragos, in *Proc. International Conference on Language Resources and Evaluation (LREC)*. The DIRHA simulated corpus, (2014), pp. 2629–2634
62. M. Matassoni, R. F. Astudillo, A. Katsamanis, M. Ravanelli, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones, (2014), pp. 1613–1617
63. M. Vacher, B. Lecouteux, P. Chahua, F. Portet, B. Meillon, N. Bonnefond, in *Proc. International Conference on Language Resources and Evaluation (LREC)*. The Sweet-Home speech and multimodal corpus for home automation interaction, (2014), pp. 4499–4506
64. N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, É. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury, É. Jamet, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. A French corpus for distant-microphone speech processing in real homes, (2016), pp. 2781–2785
65. N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina, F. Bimbot, VoiceHome-2, an extended corpus for multichannel speech processing in real homes. *Speech Comm.* **106**, 68–78 (2019)
66. A. Fleury, N. Noury, M. Vacher, H. Glasson, J.-F. Seri, in *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Sound and speech detection and classification in a health smart home, (2008), pp. 4644–4647
67. M. A. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, J. Boudy, in *Ambient Intelligence: Third International Joint Conference, Aml 2012 Proceedings, vol. LNCS-7683*, ed. by F. Paternò, B. de Ruyter, P. Markopoulos, C. Santoro, E. van Loenen, and K. Luyten. Sound environment analysis in smart home (Springer, Berlin, Heidelberg, 2012), pp. 208–223
68. A. Karpov, L. Akarun, H. Yalçın, A. Ronzhin, B. E. Demiröz, A. Çoban, M. Železný, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. Audio-visual signal processing in a multimodal assisted living environment, (2014), pp. 1023–1027
69. O. Brdiczka, M. Langet, J. Maisonnasse, J. L. Crowley, Detecting human behavior models from multimodal observation in a smart home. *IEEE Trans. Autom. Sci. Eng.* **6**(4), 588–597 (2009)

70. X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, H. Li, in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. A learning-based approach to direction of arrival estimation in noisy and reverberant environments, (2015), pp. 2814–2818
71. F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, F. Piazza, Localizing speakers in multiple rooms by using deep neural networks. *Comput. Speech Lang.* **49**, 83–106 (2018)
72. L. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition*. (Prentice Hall, Englewood Cliffs, 1993)
73. M. Wolf, C. Nadeu, in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. On the potential of channel selection for recognition of reverberated speech with multiple microphones, (2010), pp. 574–577
74. P. Maragos, A. C. Bovik, Image demodulation using multidimensional energy separation. *J. Opt. Soc. Am. A.* **12**(9), 1867–1876 (1995)
75. J. H. DiBiase, H. F. Silverman, M. S. Brandstein, in *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M. Brandstein, D. Ward. Robust localization in reverberant rooms (Springer, Berlin, Heidelberg, 2001), pp. 157–180
76. S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, 4th edn. (Academic Press, Burlington, 2009)
77. A. Brutti, M. Ravanelli, M. Omologo, in *Proc. Italian Conference on Computational Linguistics (CLiC-it) and International Workshop EVALITA*. SASLODOM: Speech Activity detection and Speaker LOcalization in DOMestic environments, (2014), pp. 139–146

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
