MULTI-ROOM SPEECH ACTIVITY DETECTION USING A DISTRIBUTED MICROPHONE NETWORK IN DOMESTIC ENVIRONMENTS

Panagiotis Giannoulis [1,5], *Alessio Brutti* [2], *Marco Matassoni* [2], *Alberto Abad* [3], *Athanasios Katsamanis* [1,5], *Miguel Matos* [3], *Gerasimos Potamianos* [4,5], *Petros Maragos* [1,5]

[1] School of E.C.E., National Technical University of Athens, Athens 15773, Greece
[2] Fondazione Bruno Kessler, Trento, Italy
[3] INESC-ID/IST, Lisbon, Portugal
[4] Dept. of E.C.E., University of Thessaly, Volos 38221, Greece
[5] Athena Research and Innovation Center, Maroussi 15125, Greece

ABSTRACT

Domestic environments are particularly challenging for distant speech recognition: reverberation, background noise and interfering sources, as well as the propagation of acoustic events across adjacent rooms, critically degrade the performance of standard speech processing algorithms. In this application scenario, a crucial task is the detection and localization of speech events generated by users within the various rooms. A specific challenge of multi-room environments is the inter-room interference that negatively affects speech activity detectors. In this paper, we present and compare different solutions for the multi-room speech activity detection task. The combination of a model-based room-independent speech activity detection module with a room-dependent inside/outside classification stage, based on specific features, provides satisfactory performance. The proposed methods are evaluated on a multi-room, multi-channel corpus, where spoken commands and other typical acoustic events occur in different rooms.

Index Terms— Speech activity detection, smart homes, microphone arrays.

1. INTRODUCTION

In smart home environments, far-field speech-based interaction with devices and appliances is viewed as a key differentiator in a continuously growing market; thus, automatic speech recognition (ASR) based solutions have been progressively entering this application field. However, critical issues related to typical usage scenarios (e.g., spontaneous speech and uncontrolled acoustic conditions) present obstacles to the development of reliable voice-based interfaces.

A particularly interesting solution for improving overall robustness is the adoption of a network of distributed microphones, aiming to mitigate the impact of reverberation and background noises on spoken dialogue home automation systems. Such an approach has been the focus of the DIRHA project [1], where the targeted system selectively monitors speech activity in the household, detects and understands voice commands, and acts as the interface for appliance control. In this direction, coherent processing of multi-microphone signals to detect and properly process concurrent speech events in different rooms becomes of critical importance.

The research leading to these results has partially received funding from the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement n. 288121 - DIRHA.

In this paper, we focus on multi-room speech activity detection (SAD) for home automation applications. Since the system must operate in an always-listening mode to satisfy the hands-free concept, it becomes crucial to correctly estimate not only the time boundaries of the speech events to be recognized, but also the position of the speaker, at least at the room level, in order to correctly proceed in the dialogue flow [2, 3]. This aspect is extremely important: Since multiple concurrent interactions could be active in parallel, the dialogue manager must be able to establish which appliance has to be operated, deriving from the acoustic front-end the information about the room where the user is located. Additionally, ASR can benefit from optimal channel selection among the microphone-equipped rooms [4]. Indeed, as already pointed out in [5–9], various approaches can be successfully applied on multi-channel data, since spatial processing can effectively exploit the information available about the desired sources. Further, results in [4, 10] demonstrate the challenges presented to far-field ASR in multi-room environments, specifically showing the crucial impact of SAD in a multi-microphone spoken dialogue system. In particular, [11] focuses on the importance of room localization to ASR performance.

Our proposed solution for multi-room SAD is based on an effective two-pass strategy: First, speech segment hypotheses are obtained employing room-independent classifiers; the derived segments are then "filtered" based on a number of acoustic features, in order to identify speech events occurring in each room of the smart home. Such features are computed for each room, using its available microphones to provide signal-to-noise ratio (SNR), cross-correlation, and envelope variance measurements that are fed into an appropriate room-level classifier. We evaluate our proposed approach on the DIRHA-GRID corpus [12], a suitable multi-room, multi-channel simulated corpus. Our reported results demonstrate significant performance improvements over a number of alternatives.

The rest of the paper is organized as follows: In Section 2, we formulate the SAD problem in a multi-room scenario, and we present our proposed system and the room selection features used. In Section 3, we overview the database and experimental framework, a number of alternative systems to compare against, and present our results. Finally, in Section 4, we conclude the paper.

2. FORMULATION AND PROPOSED APPROACH

The smart home setup adopted in the DIRHA project includes 40 microphones that record data synchronously and are distributed in-

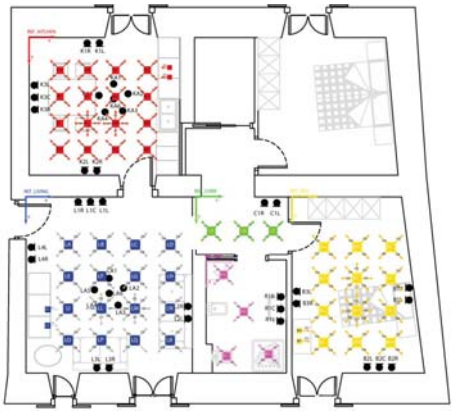

Fig. 1. The layout of the DIRHA smart home (approximately 10 × 9 × 2.7 m^3 in size) with locations of the 40 microphones in its five rooms shown (on the apartment walls and ceilings). Squares and arrows (in color) indicate the possible positions and orientations of acoustic events in the DIRHA-GRID simulated database [12].

side five rooms of an actual apartment; of these, 15 are located in the Livingroom, 13 in the Kitchen, 7 in the Bedroom, 3 in the Bathroom, and the remaining 2 in the Corridor, as depicted in Fig. 1. All microphones are placed on the room walls, with the exception of the Kitchen and Livingroom that are each also equipped with a ceiling array of 6 microphones, arranged in a star-like fashion.

2.1. Problem formulation

Let us denote[1] by $x_m^r(t)$ the signal captured by the m-th microphone ($m = 1, \ldots, M_r$) of the r-th room ($r = 1, \ldots, R$) of the smart home, and by $\mathbf{x}_t = [\, x_1^1(t), \ldots, x_{M_1}^1(t), \ldots, x_1^R(t), \ldots, x_{M_R}^R(t) \,]$ the multi-channel vector consisting of all microphone signals.

For **room-dependent** SAD, we are interested in detecting speech events occurring in a specific room, neglecting any other non-speech or speech events produced in other rooms. Defining our speech/non-speech model with 2 states, S_0 (for speech) and S_1 (for non-speech), and given the observation vector $\mathbf{X} = [\, \mathbf{x}_1, \ldots, \mathbf{x}_T \,]$, the goal is to derive state sequences $Q^r = [\, q_1^r, \ldots, q_T^r \,]$, for each room $r \in \{1, \ldots, R\}$, where $q_t^r \in \{S_0, S_1\}$, that maximize their joint conditional probability $p\,(Q^1, \ldots, Q^R | \mathbf{X})$.

In contrast, for **room-independent** SAD, the goal is to find the sequence of states $Q' = [\, q_1, \ldots, q_T \,]$ that maximizes probability $p\,(Q' | \mathbf{X})$. The resulting state sequence represents a speech/non-speech segmentation for the entire home. This of course constitutes a simpler case than room-dependent SAD, since there is no need to address possible interference among rooms or concurrent speech events, but, as also mentioned in the Introduction, may be inadequate for multi-room smart home automation applications.

2.2. System overview

Our proposed approach operates in two steps, as schematically depicted in Fig. 2. In the first step, a speech/non-speech segmentation is obtained for the entire home using a room-independent SAD module. In the second step, the resulting speech segments are further processed in order to decide whether they occurred inside or outside a given room. This step employs support vector machine (SVM) classifiers, and it relies on a set of carefully crafted room selection features. Details are discussed in the next subsections.

[1] We use such notation for the signal (and time index), denoting either signal samples or windowed signals (with corresponding subsampling of the time index), depending on the required derivations.

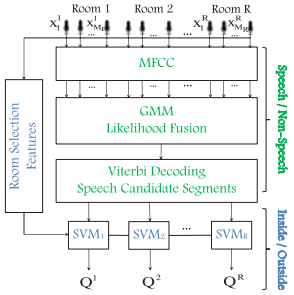

Fig. 2. Block diagram of the proposed room-dependent SAD system.

2.3. Room-independent SAD

The multi-stream Gaussian mixture model (MS-GMM) based SAD system of [13] exploits multi-channel data to produce a room-independent speech/non-speech segmentation. Given a set of GMMs (for speech and non-speech), denoted by λ_m^r, each trained on data of a single microphone channel m in room r, the log-likelihood of room-independent speech/non-speech state q_t, given multi-channel observations $\mathbf{x}_t$, is accumulated as:

$$\mathcal{L}(\, q_t \,|\, \mathbf{x}_t \,) \;=\; \sum_{r=1}^{R} \sum_{m=1}^{M_r} \log\, p\;(\, q_t \,|\, x_m^r(t) \,;\, \lambda_m^r \,) \quad . \qquad (1)$$

Viterbi decoding is then employed to derive the most likely speech/non-speech state sequence Q'.

2.4. Room selection features

As a result of the above step, a number of speech event segments are produced, defined by their starting and ending time stamps, (t_s, t_e). In the second step of the proposed approach, three room-selection features are defined over such segments, computed for each room of the smart home. These features are chosen based on the expectation that speech signals originating from outside a room typically exhibit lower energy and higher reverberation characteristics, compared to speech events occurring inside the room.

2.4.1. K-best room SNR dominance (K-SNRD)

To compute this feature, for a given speech segment, the system first estimates

$$\mathrm{SNR}_m^r(t_s, t_e) = \frac{\displaystyle\sum_{t=t_s}^{t_s+\Delta t} x_m^r(t)^{\,2}}{\displaystyle\sum_{t=t_s-\Delta t}^{t_s} x_m^r(t)^{\,2}} \quad , \qquad (2)$$

for each microphone m in room r, where Δt is a parameter that controls the amount of speech and silence considered in the SNR estimation. Subsequently, microphone set $\mathcal{M}_K$ is determined, consisting of the K channels with the highest SNRs of (2); and, finally, the K-best room SNR dominance feature (K-SNRD) for room r is computed as the difference between the sum of SNR values for the microphones of set $\mathcal{M}_K$ that belong to room r and the sum of SNRs of microphones in $\mathcal{M}_K$ that belong to the other rooms, namely:

$$\sigma^{\,r}(t_s\,, t_e) = \sum_{m \in \mathcal{M}_K} \mathrm{SNR}_m^r(t_s\,, t_e) - \sum_{r' \neq r} \sum_{m \in \mathcal{M}_K} \mathrm{SNR}_m^{r'}(t_s\,, t_e) \quad . \qquad (3)$$

2.4.2. Room microphone cross-correlation (Coh)

The inter-microphone cross-correlation function is a good estimator of the degree of reverberation of a signal [14]. For a signal window

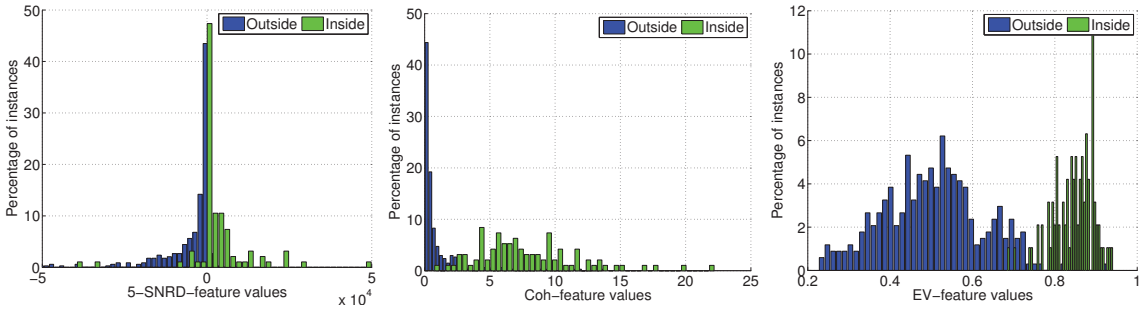

Fig. 3. Histograms of the "5-SNRD", "Coh", and "EV" features, introduced in Section 2.4, for speech occurring inside and outside the DIRHA smart home Bedroom (see also Fig. 1), for the "test1" set of the DIRHA-GRID corpus (see also Section 3.1).

starting at a given time t, we denote the maximum value of inter-microphone cross-correlation as:

$$c^{\,r}_{m\,m'}(t) = \max_{\tau} R_{x^{\,r}_{m}\,x^{\,r}_{m'}}(\tau, t) \quad , \tag{4}$$

where m, m' are two adjacent microphones in the same room r, and $R_{x^{\,r}_{m}\,x^{\,r}_{m'}}(\tau, t)$ is the cross-correlation between windowed signals $x^{\,r}_{m}(t)$ and $x^{\,r}_{m'}(t)$, with τ denoting the time lag. The quantity in (4) is expected to reach high values when an acoustic event occurs inside room r, where microphones m and m' are located, compared to events occurring outside it, in which case, direct source to room microphone propagation paths do not exist [11].

For a given speech segment, our system computes (4) between pairs of adjacent microphones over fixed-size sliding windows inside the segment. Then, the final cross-correlation feature $C^{\,r}(t_s, t_e)$ for room r and segment $(t_s\,, t_e)$ is obtained by averaging over the segment the maximum cross-correlation among all pairs of adjacent microphones $m\,, m'$ in room r (denoted below as $m\,, m' \in r$):

$$C^{\,r}(t_s\,, t_e) \;=\; \underset{t\,\in (t_s, t_e)}{\text{avg}} \left\{ \max_{m, m' \in\, r} c^{\,r}_{m\,m'}(t) \right\} \quad . \tag{5}$$

Note that this feature uses the simple cross-correlation function instead of its generalized versions [15], in order to also take advantage of the signal energy attenuation.

2.4.3. Room envelope variance (EV)

The distortion measure proposed in [16], named envelope variance (EV), is based on the idea that reverberation smooths the short-time speech energy, so its effect may be observed as a reduction in the dynamic range of the corresponding envelope. This measure has been shown to be a good indicator of the amount of reverberation in each microphone for automatic channel selection purposes [16]. To obtain the EV, first the k-th sub-band envelope of channel m at time t is computed as:

$$\hat{X}^{\,r}_{m}(k, t) = e^{\log[X^{\,r}_{m}(k,t)] - \mu_{\log[X^{\,r}_{m}(k)]}} \;,$$

where $X^{r}_{m}(k, t)$ is the short-time filter-bank energy (FBE), and quantity $\mu_{log[X^{r}_{m}(k)]}$ denotes the mean of the log-FBE of each sub-band, computed over time. Such normalization allows for constant channel effects removal. Then, the variance of each sub-band envelope is computed after cube root compression:

$$V^{\,r}_{m}(k) \;=\; \text{Var}\,[\,\hat{X}^{\,r}_{m}(k, t)^{1/3}\,] \;,$$

and the weighted average variance over all sub-bands as:

$$EV^{\,r}_{m} \;=\; \sum_{k} w^{\,r}_{m}(k) \; \frac{V^{\,r}_{m}(k)}{\underset{\substack{r' \in \{1,\ldots,R\} \\ m' \in \{1,\ldots,M_{r'}\}}}{\max} \left\{\, V^{\,r'}_{m'}(k) \,\right\}} \quad , \tag{6}$$

where $w^{\,r}_{m}(k)$ are the sub-band weights that are usually set to $1/n$, with n being the number of sub-bands.

In this work, we define the room envelope variance as the maximum value of the envelope variance (6) over all microphones in room r (denoted below as $m \in r$). In practice, we compute it in fixed-size sliding windows (extending notation of (6) to $EV^{\,r}_{m}(t)$), and averaging the result; namely, for a given segment $(t_s\,, t_e)$, the "room EV" is computed as:

$$EV^{\,r}(t_s\,, t_e) \;=\; \underset{t\,\in (t_s, t_e)}{\text{avg}} \left\{ \max_{m \in r} EV^{\,r}_{m}(t) \right\} \quad . \tag{7}$$

2.5. Feature combination for room selection

The features introduced above provide significant room discrimination information as shown in Fig. 3, which depicts histograms of the feature values obtained for speech segments inside and outside a particular room of the DIRHA-GRID corpus (see Section 3.1). Therefore, in the second step of our proposed system, we proceed to exploit them for room localization.

In particular, using the features of (3), (5), and (7), for each room r, and for each segment $(\,t_s\,, t_e\,)$ returned by the room-independent SAD, we form the 3-dimensional feature vector

$$\theta^{\,r}(t_s\,, t_e) \;=\; [\,\sigma^{\,r}(t_s\,, t_e)\,,\, C^{\,r}(t_s\,, t_e)\,,\, EV^{\,r}(t_s\,, t_e)\,] \quad . \tag{8}$$

Further, in order to jointly process the observations from all rooms, we perform early fusion by concatenating all room-specific feature vectors (8) into a single one. This yields a $3R$-dimensional vector for the segment of interest,

$$\theta(t_s\,, t_e) \;=\; [\,\theta^{\,1}(t_s\,, t_e)\,, \ldots, \theta^{R}(t_s\,, t_e)\,] \quad . \tag{9}$$

Based on feature vector (9), a linear SVM classifier is trained for each room to discriminate between speech segments inside it vs. outside. The outputs of these classifiers determine the final room-dependent segmentation Q^{r}, defined in Section 2.1. Note that the adoption of R parallel classifiers allows for the detection of multiple overlapping events in different rooms.

3. EXPERIMENTS AND RESULTS

3.1. Experimental framework: Database and metrics

For evaluating our proposed approach, we conduct experiments on the DIRHA-GRID corpus [12], a multi-microphone, multi-room simulated database, developed as part of the DIRHA project activities. This database contains a set of acoustic scenes of one-minute each in duration, "observed" by the 40 microphone channels distributed in the DIRHA apartment, following appropriate mixing of

Table 1. Evaluation of room-independent SAD modules on DIRHA-GRID test data. All results (F-score, precision, and recall) are in %.

System (classifier)	F-score	Prec.	Recall
Baseline (GMM)	84.49	79.09	90.67
Contrastive 2 (MLP)	83.07	75.00	93.09
Proposed (MS-GMM)	90.79	95.33	86.68

source material with measured room impulse responses. Each acoustic scene is composed of both speech, i.e., short English commands from the GRID database [17], and non-speech acoustic events, i.e., typical home noises. Both speech and non-speech events can take place in any of the microphone-equipped rooms, occurring randomly in time and space, within a specified set of possible source locations and orientations (see also Fig. 1). In particular, a variable number of short commands (ranging from 4 to 7) appear in each acoustic scene. Overlap in time between speech and non-speech sources is possible, while overlap between speech sources is not allowed. The corpus is divided into 3 chunks ("dev1", "test1", and "test2") containing 75 acoustic sequences each, with 12 different speakers (6 male and 6 female) appearing in each. In our experiments, we use the "dev1" set for training speech/non-speech models and for feature extraction parameter tuning. We employ the remaining two sets ("test1" and "test2", totaling 150 sequences) for evaluation.

For reporting results, we adopt the recall, precision, and F-score metrics, all computed at the frame level (every 10 ms). In the room-independent case, we evaluate one SAD output, i.e., sequence of time-stamped speech events, over the entire apartment for all 150 acoustic test sequences. For room-dependent SAD, we typically evaluate five SAD outputs (one per room) for all 150 acoustic scenes.

3.2. Evaluated systems

In addition to our proposed two-step room-dependent SAD system, for comparison, we also evaluate an alternative baseline for room-independent SAD and two contrastive room-dependent systems.

3.2.1. Implementation details of the two-step proposed system

The 40 GMMs (one for each microphone) of the first step of our system (see Section 2.3) operate on the typical acoustic front-end of 13-dimensional MFCCs with their Δ's and $\Delta\Delta$'s appended, extracted over 25 ms frames with a 10 ms shift. Each class model (speech/non-speech) employs 32 Gaussians with diagonal covariances.

Regarding room selection features (Section 2.4), in (2), Δt is set to 0.5 s. Further, in (5), 100 ms windows with a 25 ms shift are used, and, in (7), 600 ms windows with a 50 ms shift. Note that the dimensionality of feature vector (9), used in SVM-based room selection, is 15, when all five smart home rooms are considered, or 12, in case the Corridor is excluded (see also Section 3.3).

3.2.2. Room-independent SAD baseline

In order to demonstrate the effectiveness of channel fusion in our proposed approach (see (1) in Section 2.3), a simple alternative for room-independent SAD is considered. This baseline employs R room-specific speech/non-speech GMMs, each trained on a single representative microphone of the corresponding room, using an identical front-end and model complexity, as discussed in Section 3.2.1. The system produces separate room-dependent speech/non-speech state sequences using the Viterbi algorithm (thus, 5 sequences are obtained), with the final, room-independent segmentation resulting as the union of the speech event hypotheses generated for each room.

Table 2. Room-dependent SAD performance (in F-score, %) on the DIRHA-GRID test data, when employing the various room selection features of Section 2.4 in the second step of the proposed method, assuming ground-truth room-independent SAD boundaries. Results are depicted for each of the five rooms of the smart home of Fig. 1.

Features	Liv.	Kitch.	Bath.	Bed.	Cor.
5-SNRD	62.03	72.55	29.32	79.57	12.50
Coh	84.78	89.38	91.96	89.88	22.34
EV	85.56	90.55	79.50	88.70	23.76
all	88.55	93.26	95.02	91.90	25.86

3.2.3. Contrastive 1: Single-step room-dependent SAD with GMMs

As an alternative to the proposed, we consider a single-step system that employs three-state GMMs (3s-GMM), one for each room. In particular, one state is used to model "speech inside" the room, another "speech outside" it, and a third "non-speech". Similarly to Section 3.2.2, a single microphone for each room is considered, and room-dependent audio stream segmentations are obtained by Viterbi decoding (considering the three possible states). For the final room-dependent SAD, events labeled as "speech outside" are set to "non-speech" for that particular room.

3.2.4. Contrastive 2: Two-step room-dependent SAD with MLPs

We also consider the system of [18], which, instead of GMMs, employs multi-layer perceptron (MLP) models, followed by a finite state machine decoder, for speech/non-speech segmentation. In more detail, this constitutes a two-step system, where, first, segmentations for each room are obtained based on MLPs and channel combination via majority voting. Then, as a second step, simple strategies using EV measures are exploited to "filter" speech segments as occurring inside or outside each room, thus yielding the desired room-dependent SAD. Note that the output of the first stage can also provide room-independent SAD, based on the union of the initial MLP segmentations per room.

3.3. Experimental results

We first evaluate room-independent SAD performance of the systems discussed above (typically, their first steps), presenting results in Table 1. One can readily observe that in terms of F-score, the MS-GMM based proposed method achieves approximately a 40% relative error reduction compared to the baseline GMM approach of Section 3.2.2, improving F-score from 84.49% of the latter to 90.79%. Clearly, exploiting all 40 microphones (instead of one per room) and fusing likelihood scores by (1), as opposed to fusion of single-microphone SAD outputs, improves performance significantly.

Regarding room-dependent SAD, we first investigate the capacity of the proposed room selection features of Section 2.4 to discriminate between speech occurring inside or outside a room, given the room-independent segmentation. Table 2 reports the SVM-based classification performance (in terms of F-score) of different feature sets for each room of the DIRHA smart home, using ground-truth room-independent SAD boundaries. In all single feature cases, 5-dimensional vectors are produced (one feature per room), yielding 15-dimensional vectors when all 3 feature sets are considered (see also (9)). Clearly, the combined feature vector outperforms any single feature set, suggesting that different features carry complementary information for room selection. In addition, we can observe a large performance drop in the case of the Corridor, compared to the

Table 3. Evaluation of room-dependent SAD systems on DIRHA-GRID test data. All results (F-score, precision, and recall) are in %. In all cases, all five rooms are considered, with the exception of the last row, where the Corridor is excluded ($R = 4$).

System	F-score	Prec.	Recall
Contrastive 2 (MLP), 1st step only	40.92	26.29	92.31
Proposed (MS-GMM), 1st step only	49.27	35.32	81.47
Contrastive 1 (3s-GMM)	60.23	52.69	70.30
Contrastive 2 (MLP), both steps	57.61	48.22	71.56
Proposed (MS-GMM), both steps	74.46	68.50	81.58
MS-GMM / MLP, and 2nd step	75.01	69.49	81.50
MS-GMM / MLP, and 2nd step (R=4)	84.27	86.37	82.28

other rooms. This is not surprising, as this is not a typical room, but more like a hallway in the center of the apartment, exposed to sounds coming from all other rooms. Consequently, it is subject to higher false alarm rates.

Finally, in Table 3, we present results for room-dependent SAD. The upper part of the table shows performance of single-step methods. Note that the "Contrastive 1" system that employs 3-state GMMs performs significantly better, since it is designed to yield room-localization within its single step, by introducing the third GMM state. In contrast, the first step of the proposed and "Contrastive 2" systems are more prone to introduce false alarms (low precision), as rooms are processed independently. Next, in the remaining parts of the table, we show results for various two-step methods. The proposed method exhibits a remarkable segmentation performance, attaining an approximately 35% relative error reduction in terms of F-score, compared to the best single-step system ("Contrastive 1"). Such is mostly because of a dramatic increase in precision, due to the second step. Further, the proposed approach also outperforms the "Contrastive 2" system. Nevertheless, since the two methods employ different classifiers, it is of interest to consider their combination within the first step, aiming to obtain more accurate room-independent speech segments, before proceeding to their processing at the second step of the proposed method. Such combination is performed by a simple speech label intersection of their first-step segmentation outputs, and, as observed in the table, provides an additional slight improvement in precision and F-score. Results can of course further improve, if the Corridor is excluded in the evaluation, as observed in the last row of Table 3.

4. CONCLUSIONS

We have introduced a multi-channel, two-step approach to address speech activity detection in multi-room domestic environments, where (possibly concurrent) speech events in separate rooms need to be detected, robustly addressing cross-room interference. The presented system produces first a multi-channel, room-independent speech/non-speech segmentation that is subsequently localized in the smart home room(s), based on a novel set of room selection features. Experiments on a multi-room, multi-channel corpus of simulated audio recordings demonstrate the superiority of the proposed approach against alternatives, yielding absolute F-score improvements of at least 15%.

REFERENCES

[1] "DIRHA: Distant-speech Interaction for Robust Home Applications," [Online] Available: http://dirha.fbk.eu/.

[2] P. Giannoulis, A. Tsiami, I. Rodomagoulakis, A. Katsamanis, G. Potamianos, and P. Maragos, "The Athena-RC system for speech activity detection and speaker localization in the DIRHA smart home," in *Proc. HSCMA*, 2014, pp. 167–171.

[3] Y. Tachioka, T. Narita, S. Watanabe, and J. Le Roux, "Ensemble integration of calibrated speaker localization and statistical speech detection in domestic environments," in *Proc. HSCMA*, 2014, pp. 162–166.

[4] M. Matos, A. Abad, R. Astudillo, and I. Trancoso, "Recognition of distant voice commands for home applications in Portuguese," in *Proc. IberSpeech*, 2014, pp. 178–188.

[5] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Comm.*, 57:170–180, 2014.

[6] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Comp. Speech Lang.*, 27(3):621–633, 2013.

[7] M. Delcroix, K. Kinoshita *et al.*, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Comp. Speech Lang.*, 27(3):851–873, 2013.

[8] F. Nesta and M. Matassoni, "Blind source extraction for robust speech recognition in multisource noisy environments," *Comp. Speech Lang.*, 27(3):703–725, 2013.

[9] I. Rodomagoulakis, P. Giannoulis, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes," in *Proc. DSP*, 2013, pp. 1–6.

[10] M. Vacher, B. Lecouteux, and F. Portet, "Multichannel automatic recognition of voice command in a multi-room smart home: an experiment involving seniors and users with visual impairment," in *Proc. Interspeech*, 2014, pp. 1008–1012.

[11] J.A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, and G. Kubin, "Room localization for distant speech recognition," in *Proc. Interspeech*, 2014, pp. 2450–2453.

[12] M. Matassoni, R.F. Astudillo, A. Katsamanis, and M. Ravanelli, "The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones," in *Proc. Interspeech*, 2014, pp. 1613–1617.

[13] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *Proc. EUSIPCO*, 2014.

[14] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, 4(2):148–152, 1996.

[15] C. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, 24(4):320–327, 1976.

[16] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Proc. Interspeech*, 2010, pp. 574–577.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, 120(5):2421–2424, 2006.

[18] A. Abad, M. Matos, H. Meinedo, R.F. Astudillo, and I. Trancoso, "The L^2F system for the EVALITA-2014 speech activity detection challenge in domestic environments," in *Proc. CLiC-it/EVALITA*, 2014, vol. II, pp. 147–152.