10.12871/clicit2014226

SASLODOM: Speech Activity detection and Speaker LOcalization in DOMestic environments

Alessio Brutti, Mirco Ravanelli, Maurizio Omologo
Center for Information and Communication Technology - Fondazione Bruno Kessler
via Sommarive 18, 38123, Trento
{brutti, mravanelli, omologo}@fbk.eu

Abstract

English. This paper describes the design, data and evaluation results of the speech activity detection and speaker localization task in domestic environments (SASLODOM) in the framework of the EVALITA 2014 evaluation campaign. Domestic environments are particularly challenging for distant speech recognition and audio processing in general due to reverberation, the variety of background noises, the presence of interfering sources as well as the propagation of acoustic events across rooms. In this context, a crucial goal of the front-end processing is the detection and localization of speech events generated by users within the various rooms. The SASLODOM task aims at evaluating solutions for both activity detection and source localization on corpora of multi-channel data representing realistic domestic scenes.

Italiano. *In questo articolo viene presentato il database, le metriche e i risultati della valutazione del task SASLDOM all'interno della campagna di valutazione EVALITA 2014. Gli ambienti domestici sono particolarmente sfidanti per le tecnologie di riconoscimento vocale ed elaborazione audio in genere, a causa del riverbero, della varietá di rumore di fondo, della presenza di interferenti e infine a causa della propagazione degli eventi acustico attraverso le stanze. In questo contesto un aspetto cruciale del front-end acustico è la capacità di rilevare e localizzare gli eventi acustici generati dall'utente nelle varie stanze. Il task SASLODOM mira a valutare soluzioni di rilevamento del parlato e localizzazione della sorgente su due database multicanale che rappresentano tipiche scene domestiche.*

1 Introduction

The SASLODOM challenge, within the framework of EVALITA 2014, addresses the problem of the detection in time and localization in space of speech events in domestic contexts. A considerable number of applications could benefit from natural speech interaction with distant microphones (Wölfel and McDonough, 2009). In particular, the possibility to control by voice the devices and appliances of an automated home has recently received a significantly growing interest. This scenario is being targeted by the EU project DIRHA[1] (Distant-speech Interaction for Robust Home Applications) focusing on motor-impaired users, whose life quality can considerably improve thanks to speech-driven automated home.

A desirable property of a distant-speech interaction system in domestic contexts is the capability to be "always-listening" and to always accept commands or requests from the users. This feature represents a noteworthy challenge, as the system must be able to keep as low as possible the rate of false alarms, generated by acoustic events that are not intended to convey any message addressed to the recognition system, while at the same time it must be able to detect any speech command, independently of the current environmental conditions and without introducing constraints on the user position and orientation. Hence, fundamental features of the front-end processing component are a robust Speech Activity Detection (SAD) and Source LOCalization (SLOC). A correct identification of time boundaries, room and spatial coordinates of each speech event is essential for the targeted interactive scenario. In fact, the efficiency of

[1] http://dirha.fbk.eu

a dialogue manager or of a command-and-control system, strongly depends on the performance of the ASR system in the right room: in several cases the system must be able to serve the user also on the basis of the location where the speech command has been given (i.e., the command "open the window" implies that the window to open is located in the same room.). The critical role of the SAD component both in distant-talking ASR and in acoustic event classification has been studied in (Macho et al., 2005).

There is a wide literature addressing SAD techniques. Early works on specific speech/non-speech segmentation focused on close talking interaction and were based on the use of energy thresholding and zero-crossing features (Junqua et al., 1994), in some cases exploring the use of noise reduction (Bouquin-Jeannes and Faucon, 1995). Also, well-known features among the speech recognition community, like MFCCs and PLP, have been used for audio event detection (Portelo et al., 2008; Trancoso et al., 2009). Additionally, techniques based on Spectral Variation Functions (SVF) (DeMori, 1998) or other spectro-temporal features (Pham et al., 2008) can be exploited to discriminate speech from stationary background noise, even under unfavorable SNR conditions. Various machine learning methods (Shin et al., 2010), are used to provide a final classification of the audio events such as Gaussian Mixture Models (GMMs) (Chu et al., 2004), Support Vector Machines (SVMs) (Guo and Li, 2003), Hidden Markov Models (HMMs) and Bayesian Networks (Cai et al., 2006). Recently, solutions relying on Deep Neural Networks (DNN) have been employed (Zhang and Wu, 2013). Finally, the availability of multiple acquisition channels permits the implementation of multi-channel processing (Wrigley et al., 2005; Dines et al., 2006), or the adoption of different feature sets, eventually based on the spatial coherence at two or more microphones (Armani et al., 2003). In general the reliability of the resulting system can be highly correlated to the SNR of the input, depending on the environmental noise and the distance from speaker to microphones. In (Ramirez et al., 2005), more details are given on the problem, together with a good introductory survey of the audio event detection techniques explored more recently.

Also SLOC technologies have been deeply investigated and several different approaches are available in the literature (Wölfel and McDonough, 2009; Brandstein and Ward, 2001; Huang and Benesty, 2004). In general, SLOC algorithms are based on the estimation of the Time Differences Of Arrivals (TDOA) at two or more microphones, from which the source location is inferred by applying geometrical considerations. The Generalized Cross-Correlation Phase Transform (GCC-PHAT) (Knapp and Carter, 1976), is the most common technique for estimating the TDOA at two microphones. In multi-microphone configurations SLOC techniques based on acoustic maps, like the Global Coherence Field (GCF) (DeMori, 1998) also known as SRP-PHAT (Brandstein and Ward, 2001), are particularly effective in representing the spatial distribution of sources. Under the assumption that sources are sparse in time and space short-term spatio-temporal clustering has been successfully applied to the localization of multiple sources (Di Claudio et al., 2000; Lathoud and Odobez, 2007). Sequential bayesian methods and particle filtering (Arulampalam and Maskell, 2002; Vermaak and Blake, 2001; Lehman and Johansson, 2007) have also been experimented successfully on tracking of single as well as multiple sources (Fallon, 2008; Lee et al., 2010). Beside the above-mentioned methods, more recently approaches for Blind Source Separation (BSS), relying on Independent Component Analysis (ICA) (H. Sawada et al., 2003; Loesch et al., 2009) or on sparsity-aware processing of the cross-spectrum (Araki et al., 2009; Nesta and Omologo, 2011), have been applied to the estimation of the TDOA in presence of multiple sources (Brutti and Nesta, 2013).

1.1 Motivation

One of the main issues of the multi-room scenario typical of the domestic context, is that acoustic waves propagate from one room to another (e.g. through open doors), which represents an intrinsic cause of ambiguity on the location of each sound source, especially when concurring events can occur in different rooms. Furthermore, the environmental conditions of a domestic scene (e.g., background noise, interferes, noise sources, number of users, etc...) significantly vary over time, from very quiet conditions to very noisy and challenging situations, requiring algorithmic solutions capable of coping with such variability while preserving good performance. In DIRHA,

these challenges are tackled by distributing multiple microphones in the rooms of an apartment. This approach permits the implementation of effective SLOC solutions to identify the actual location of event generation as well as the development of robust strategies for event detection and speech recognition, for instance based on channel or model selection (Wolf and Nadeu, 2013; Sehr et al., 2010). The joint use of SLOC and SAD technologies is hence required in the addressed scenario in order to realize a multi-room SLOC and SAD. Although SAD and SLOC technologies have been widely investigated over the decades and several effective solutions are available in the literature, the peculiarities of the domestic scenarios pose significant challenges for these technologies. This fact motivated the creation of the DIRHA corpora and the definition of the SASLODOM evaluation tasks.

2 The DIRHA corpora

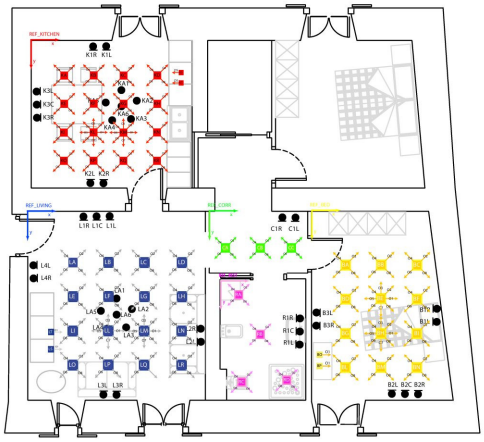

Figure 1: Layout of the apartment used for the collection of the DIRHA corpora. Circles indicate the microphone positions. Squares and arrows indicate the possible positions and orientations of acoustic events in the simulated corpus.

The general scenario addressed in the DIRHA project refers to a real automated apartment consisting of 5 rooms. In each room a set of microphones is deployed on the walls and the ceiling, as shown in Figure 1. 15 microphones are located in the Livingroom (bottom-left), 13 in the Kitchen (top-left), 7 in the Bedroom (bottom-right), 3 in the Bathroom (bottom-middle) and 2 in the Corridor (central). A star-shaped 6-microphone array is mounted on the ceiling of the Livingroom and of the Kitchen, where the majority of the speech events is expected to occur in every-day interactions. Overall 40 microphones monitor the house. For this target scenario, both simulated and real corpora of multi-channel multi-lingual acoustic data were created, in order to reproduce a variety of typical domestic scenes for experimental purposes (Cristoforetti et al., 2014). For each of the 40 microphones a 48 kHz/16 bit WAV audio file is available, fully synchronized and aligned at sample level with the other channels. Detailed annotations in terms of acoustic events, source positions and other information are also available. The corpora are publicly available upon request to the DIRHA consortium. The next sections provide a brief description of the two corpora. Table 1 summarizes the main differences between the simulated and real data collections.

	Real	Simulations
source	human	loudspeaker
movement	moving	static
system feedback	yes	no
background	quiet	various
noise source rate	low	high
overlapping events	no	yes

Table 1: Main differences between the real and simulated scenes.

2.1 The DIRHA SimCorpus

First of all, for a set of predefined positions and orientations (represented by squares and arrows in Figure 1) Room Impulse Responses (RIR) were measured for the 40 microphones by exciting the environment with long Exponential Sine Sweep (ESS) signals (Farina, 2000) reproduced by a loudspeaker. This procedure ensures high SNR and remarkable robustness against harmonic distortions (Ravanelli et al., 2012).

Speech events including sentences uttered by 120 speakers in 4 languages (Greek, German, Italian and Portuguese) were recorded using high-quality close-talking microphones and ensuring very high SNR and absence of artifacts. These sentences are typical commands for the domestic system, phonetically rich sentences and conversational speech. For what concerns "non-speech" events, they were selected from Logic Pro and from the Freesound[2] high-quality database, con-

[2]http://www.freesound.org/

sidering those sounds typical of domestic environments. Moreover, a selection of copyright-free radio shows, music and movies were used to simulate radio and television sounds. To increase the realism of the acoustic sequences, 21 common home-noise sources (shower, washing machine, oven, vacuum cleaner, etc.) were directly recorded by the distributed microphone network of the apartment.

Given the ingredients described above, the DIRHA SimCorpus (Cristoforetti et al., 2014) was created as a collection of acoustic scenes with a duration of 60 seconds. Each scene consists of real background noise, with random dynamics, to which a variety of localized acoustic and speech events are superimposed. Events occur randomly in time and in space, constrained on the grid of the predefined positions and orientations for which RIR measurements are available. The acoustic wave propagation from the sound source to each single microphone is simulated by convolving dry signals with the respective RIR.

Data set	Development	Test
Simul	40 scenes 40 min. 23.4% speech	40 scenes 40 min. 23.7% speech
Real	12 scenes 11 min. 9% speech	10 scenes 10 min. 30 sec. 17% speech

Table 2: Development and test material used in the SASLODOM task.

2.2 Real corpus

Besides the simulated scenes, a real data set was derived from excerpts of a Wizard-of-Oz data collection, resulting in 22 scenes, each one approximately 60 second long. Each real scene includes a human speaker uttering typical commands while moving within the Livingroom and the Kitchen. The background is rather quiet (in particular if compared to the simulated scenes), and the main noise of interference is the system output reproduced by the Wizard through a loudspeaker installed on the ceiling of the Livingroom or of the Kitchen (e.g., the replies of the system to the user commands). The reference signal of the system output is also made available.

2.3 Data used in the SASLODOM task

For the SASLODOM task a subset of the simulated data, consisting in 80 scenes in Italian, was considered. The scenes are selected in such a way that different degrees of complexity are covered. Notice that the language is probably not relevant for the addressed technologies.For what regards the real data, the full data set is used since it is relatively small and in Italian.

The data are evenly split in two sets for development and tests. Table 2 summarizes the amount of data used in the evaluation and the ratio between the total length of speech events over the full datasets duration.

3 The Task

Given the multi-room domestic scenario addressed in the DIRHA project, the goal of the SASLODOM task is, for each speech event, to:

- provide the corresponding time boundaries,
- determine the room where it was generated,
- derive the spatial coordinates of the speaker.

When considering a specific room, speech events occurring in other rooms must be discarded. Similarly, any other noise event must be neglected. In case a speech event occurring in a given room is associated by the system to another room, this will result in a false alarm and a deletion. Although speech and noise events may occur anywhere in the apartment, the evaluation considers only speech events generated in the Livingroom and Kitchen (i.e., speech events in other rooms must be discarded). This choice is motivated by the fact that a small number of microphones is available in the other rooms.

To allow the participation of laboratories without effective solution for SLOC, a subtask is defined where the localization stage does not require the estimation of the speaker coordinates but just the identification of the room where the event occurred (localization is implicit in the SAD component). This subtask is referred to as SAD.

4 System Evaluation

Reference speaker positions and speech activities are reported every 50 ms in a reference file, together with the annotation of other acoustic events occurring in the 5 rooms. The system under evaluation delivers, for each room and each scene, a

similar hypothesis file with a time resolution of at least 50 ms. If the time resolution of the hypotheses is higher, the evaluation tool averages the estimated coordinates.

In the evaluation step, the hypothesis sequence and the reference file are compared one each other. For each reference line, the closest (in time) hypothesis line is selected and one of the four events below is generated:

- Deletion: no hypothesis available for a given reference line (SAD);
- False Alarm: an hypothesis is produced when there is no speech activity in the targeted room (SAD);
- Fine error: the distance between the estimated source position and the reference is smaller than 50 cm;
- Gross error: the distance between the estimated source position and the reference is larger than 50 cm.

4.1 Metrics

Given the classifications listed above, a series of metrics is computed to characterize the performance of the system under evaluation:

- Time boundaries accuracy:
  - **Deletion Rate**: number of missing hypotheses over all speech frames.
  - **False Alarm Rate**: number of false alarms over all non-speech frames.
- Event-based Detection performance:
  - **Precision** of the SAD component.
  - **Recall** of the SAD component.
  - **F score**.

Systems are ranked according to the **Overall SAD Detection error**, defined as:

$$\text{SAD} = \frac{N_{\text{fa}} + \beta N_{\text{del}}}{N_{\text{nsp}} + \beta N_{\text{sp}}},$$

where N_{del}, N_{fa} are the total numbers of deletion and false alarms respectively, N_{sp} is the total number of speech frames, N_{nsp} is the total number of non-speech frames while $\beta = \frac{N_{\text{nsp}}}{N_{\text{sp}}}$ weights the contributions of false alarm and deletions. This weighting is necessary to avoid that results are biased due to the unbalanced distribution of speech and non-speech frames in the data (see Table 2). The SAD metric is equivalent to the Equal Error Rate in most of the cases. For a deeper understanding of the evaluation results, wherever possible the scores are reported in a disaggregated fashion, differentiating among cases in which there are noises in the targeted room, interferes (noise or speech) in another room, background noises.

The evaluation protocol includes also a set of metrics for the source localization tasks. Since none of the participants provided results on this problem they are not fully described here. They comprises: the average (bias) and RMS errors for fine and gross errors respectively as well as the ratio between the two categories (percentage of correct localization estimates).

It is worth mentioning that in an ASR perspective false alarms are less problematic than deletion as the rejection model offers an effective and practical way to deal with them. Therefore, it could make sense to give Deletions a higher weight in the overall SAD error rate computation. However, in the addressed context false alarms include also correct event associated to wrong rooms: this case would be detrimental for ASR and dialogue engines. This is the reason why the two rates are equally weighted.

4.2 Participants

As reported in Table 3, two laboratories participated in the evaluation, focusing on event detection and room selection only, and both participants submitted more than one system. The Spoken Language Systems Laboratory of the Instituto de Engenharia de Sistemas e Computadores Investigao e Desenvolvimento in Lisbon (INESC-ID L^2F) submitted three systems based on Multi-Layer Perceptron (MLP) and Major Voting Fusion (MVF) of the multiple channels. The three systems differ in the way the room selection is performed: MVF-MLP-NRS does not select the room while MVF-MLP-MRS and MVF-MLP-RRS adopt two slightly different procedures. The Multimedia Assistive Technology Laboratory - Dipartimento di Ingegneria dell'Informazione of the Universitá Politecnica delle Marche (MATeLab-DII) presented two approaches based on Deep Belief Networks (DBN) and Bidirection Long Short-Term Mem-

ory Recurrent Neural Networks (BLSTM) respectively. It must be mentioned that, although no SASLODOM specific data were used for system tuning, neither simulated nor real, the MLP models used by INESC-ID L^{2}F have been adapted on a rather large set of in-domain DIRHA data, not available to the other participant, which could give a significant improvement in the performance.

4.3 Results

Table 4 reports the evaluation results on the simulated corpus. Besides the official metrics the table reports the results also in terms of event-based metrics. The best performing system is "MVF-MLP-NRS" from INESC-ID L^{2}F which achieves a 7.7% error rate at frame level. However, this is obtained allowing events to occur in more than one room, which results in a considerable increase of false alarms and a significant reduction in the event-based metrics. In particular, the false alarm rate doubles in presence of events outside the target room. The reason why "MVF-MLP-NRS" performs better than the other two systems could be that the room selection scheme fails in several cases, in particular when noises outside the room occur. This fact confirms that the room selection problem is not a trivial task at all. In general all system submitted by INESC-ID L^{2}F handles properly the background noise, while a performance degradation is observed when events occurs outside the room. Note that the second best approach, which achieves a 9.5% overall error rate, has a very low precision despite acceptable false alarm and deletion rates: the reason could be in the generation of several short events. For both MATeLab-DII solutions background noise determines an increase of deletions (features are not observable) while noise events outside the rooms results in a higher false alarm rate (events are detected in the wrong room). It must be kept in mind that DNN solutions are penalized by the limited amount of training material.

4.4 Real Data

Table 5 reports the results on the real data. As expected the performance of the best systems is much higher than on the simulated data, thanks to the reduced amount of background noise and the absence of interfering sources. Furthermore, in the real data set events never overlap in time. In this case the best approaches are "MVF-MLP-MRS" and "MVF-MLP-RRS" of INESC-ID L^{2}F which outperform the solution without room selection. Given the easier conditions the room selection behaves properly and this provides a significant improvement to the performance. The methods proposed by MATeLab-DII performs considerably worse than on the simulated data, probably due to the limited amount of training material available.

5 Conclusions

The SASLODOM task at EVALITA 2014 addressed the problem of detecting and localizing speech event in a multi-room domestic scenario. The evaluation, based on real and simulated acoustic corpora collected within the EU DIRHA project, attracted two participants who focused on the SAD subtask. The submitted systems implement state of the art MLP and DNN solutions for the speech/non-speech classification task. The results confirm that the domestic scenario is extremely challenging and specific solutions based on multi-channel processing and room selection/localization are crucial to obtain satisfactory performance. In terms of absolute numbers, a very good accuracy is achieved on the real data.

Acknowledgements

This work has partially received funding from the European Union's 7th Framework Programme (FP7/2007-2013), grant agreement n. 288121-DIRHA.

Site ID	Full Name	Task	Runs
INESC-ID L^{2}F	Spoken Language Systems Laboratory Instituto de Engenharia de Sistemas e Computadores Investigao e Desenvolvimento Lisboa, Portugal	SAD	3
MATeLab-DII	Multimedia Assistive Technology Laboratory Dipartimento di Ingegneria dell'Informazione Universitá Politecnica delle Marche Ancona, Italy	SAD	2

Table 3: The participants of the SASLODOM task.

Lab	System	**SAD**	FA	Del	P	R	Fscore
INESC-ID L^{2}F	MVF-MLP-MRS	14.4	3.6	25.2	82.3	75.1	78.5
	MVF-MLP-RRS Sys2	11.8	5.4	18.2	73.4	79.2	76.2
	MVF-MLP-NRS Sys3	7.7	12.0	3.4	53.5	95.9	68.7
MATeLab-DII	BLSTM	12.1	11.9	12.3	30.6	98.6	46.5
	DBN	9.5	8.7	10.3	25.3	99.5	40.4

Table 4: Evaluation results on the simulated data.

Lab	System	**SAD**	FA	Del	P	R	Fscore
INESC-ID L^{2}F	MVF-MLP-MRS1	2.0	2.7	1.3	100	96.2	98.1
	MVF-MLP-RRS	2.0	2.7	1.3	100	96.2	98.1
	MVF-MLP-NRS	13.7	26.1	1.3	49.2	96.2	65.1
MATeLab-DII	BLSTM	19.7	33.7	5.6	22.5	98.7	36.7
	DBN	12.2	9.7	14.7	28.5	98.7	44.2

Table 5: Evaluation results on the real data.

References

Shoko Araki, Tomohiro Nakatani, Hiroshi Sawada, and Shoji Makino. 2009. Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem. In *Proc. of the International Conference on Independent Component Analysis and Signal Separation*.

L. Armani, M. Matassoni, M. Omologo, and P. Svaizer. 2003. Use of a CSP-based voice activity detector for distant-talking ASR. In *EUROSPEECH*.

M. Arulampalam and S. Maskell. 2002. A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), February.

R.L. Bouquin-Jeannes and G. Faucon. 1995. Study of a voice activity detector and its influence on a noise reduction system. *Speech Communication*, 16.

M. Brandstein and D. Ward. 2001. *Microphone Arrays*. Springer-Verlag.

A. Brutti and F. Nesta. 2013. Tracking of multidimensional tdoa for multiple sources with distributed microphone pairs. *Computer Speech And Language*, 27(3).

R Cai, L. Lu, A. Hanjalic, H. Zhang, and L. Cai. 2006. A flexible framework for key audio effects detection and auditory context inference. *IEEE Trans. on Audio, Speech and Language Processing*, 14(3).

W. Chu, W. Cheng, J. Wu, and J. Hsu. 2004. A study of semantic context detection by using SVM and GMM approache. In *Proc. of IEEE International Conference on Multimedia and Expo*.

L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos. 2014. The DIRHA simulated corpus. In *LREC*.

R. DeMori. 1998. *Spoken Dialogues with Computers*. Academic Press, London. Chapter 2.

E. Di Claudio, R. Parisi, and G. Orlandi. 2000. Multisource localization in reverberant environments by root-music and clustering. In *Proc. of IEEE conference on Acoustics, Speech, and Signal Processing*.

J. Dines, J. Vepa, and T. Hain. 2006. The segmentation of multichannel meeting recordings for automatic speech recognition. In *Proc. Int. Conf. on Speech Communication and Technology*.

M. Fallon. 2008. Multi target acoustic source tracking with an unknown and time varying number of targets. In *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, May.

A Farina. 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *110th AES Convention*, February.

G. Guo and S. Li. 2003. Content-based audio classification and retrieval by support vector machines. *IEEE Trans. on Neural Networks*, 14(1).

H. H. Sawada, R. Mukai, and S. Makino. 2003. Direction of arrival estimation for multiple source signals using independent component analysis. In *Proceedings of ISSPA*.

Y. Huang and J. Benesty. 2004. *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Kluwer Academic Publishers.

J.C. Junqua, B. Mak, and B. Reaves. 1994. A robust algorithm for word boundary detection in the presence of noise. *IEEE Trans. on Speech and Audio Processing*, 2(3).

C. H. Knapp and G. C. Carter. 1976. The generalized correlation method for estimation of time delay. In *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, volume 24, pages 320–327.

G. Lathoud and J.M. Odobez. 2007. Short-term spatio-temporal clustering applied to multiple moving speakers. *IEEE Trans. on Audio, Speech and Language Processing*, 15(5), July.

Y. Lee, T.S. Wada, and Biing-Hwang Juang. 2010. Multiple acoustic source localization based on multiple hypotheses testing using particle approach. In *IEEE International Conference on Acoustics Speech and Signal Processing*.

E Lehman and A. Johansson. 2007. Particle filter with integrated voice activity detection for acoustic source tracking. *EURASIP Journal on Applied Signal Processing*.

B. Loesch, S. Uhlich, and B. Yang. 2009. Multidimensional localization of multiple sound sources using frequency domain ICA and an extended state coherence transform. *Proceedings of IEEE Workshop on Statistical Signal Processing*.

D. Macho, J. Padrell, A. Adad, J. McDonough, M. Wolfel, A. Brutti, M. Omologo, G. Potamianos, S. Chu, U. Klee, P. Svaizer, C. Nadeu, and J. Hernando. 2005. Automatic speech activity detection, source localization and speech recognition on the chil seminar corpus. In *Proc. of IEEE International Conference on Multimedia and Expo*.

F. Nesta and M. Omologo. 2011. Generalized State Coherence Transform for multidimensional TDOA estimation of multiple sources. *Audio, Speech, and Language Processing, IEEE Transactions on*.

T.V. Pham, M. Stadtschnitzer, Pernkopf F., and Kubin G. 2008. Voice activity detection algorithms using subband power distance feature for noisy environments. In *Proc. of Interspeech*.

J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro. 2008. Non-speech audio event detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*.

J. Ramirez, J.C. Segura, C. Benitez, A. De la Torre, and A. Rubio. 2005. An effective subband osf-based vad with noise reduction for robust speech recognition. *IEEE Trans. on Speech and Audio Processing*, 13(6), Nov.

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

A. Sehr, R. Maas, and W. Kellermann. 2010. Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. *IEEE Trans. on Audio, Speech, and Language Processing*, 18(7):1676–1691.

J. W. Shin, J.H. Chang, and N. S. Kim. 2010. Voice activity detection based on statistical models and machine learning approaches. *Computer Speech and Language*, page 515=530.

I. Trancoso, J. Portelo, M. Bugalho, J. da Silva Neto, and A. Serralheiro. 2009. Training audio events detectors with a sound effects corpus. In *Proc. of Interspeech.*

J Vermaak and A. Blake. 2001. Nonlinear ltering for speaker tracking in noisy and reverberant environments. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*.

M. Wolf and C. Nadeu. 2013. Channel selection measures for multi-microphone speech recognition. *Speech Communication*.

M. Wölfel and J. McDonough. 2009. *Distant speech recognition*. Wiley.

S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals. 2005. Speech and crosstalk detection in multichannel audio. *IEEE Trans. on Speech and Audio Processing*, 13(1):84–91, Jan.

X.L. Zhang and J. Wu. 2013. Deep belief networks based voice activity detection. *IEEE Trans. on Audio, Speech, and Language Processing*, 21(4):679–710, April.