A French corpus for distant-microphone speech processing in real homes

Nancy Bertin, Ewen Camberlein, Emmanuel Vincent, Romain Lebarbenchon, Stéphane Peillon, Éric Lamandé, Sunit Sivasankaran, Frédéric Bimbot, Irina Illina, Ariane Tom, et al.

A French corpus for distant-microphone speech processing in real homes

Nancy Bertin[1]*, Ewen Camberlein*[1]*, Emmanuel Vincent*[2]*, Romain Lebarbenchon*[1]*,*
Stéphane Peillon[3]*, Éric Lamandé*[3]*, Sunit Sivasankaran*[2]*, Frédéric Bimbot*[1]*, Irina Illina*[4]*,*
Ariane Tom[5]*, Sylvain Fleury*[5]*, and Éric Jamet*[5]

[1] IRISA - CNRS UMR 6074, Rennes, France
[2] Inria, Villers-lès-Nancy, F-54600, France
[3] VoiceBox Technologies France
[4] Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
[5] CRPCC, Université Rennes 2, 35043 Rennes Cedex, France
nancy.bertin@irisa.fr

Abstract

We introduce a new corpus for distant-microphone speech processing in domestic environments. This corpus includes reverberated, noisy speech signals spoken by native French talkers in a lounge and recorded by an 8-microphone device at various angles and distances and in various noise conditions. Room impulse responses and noise-only signals recorded in various real rooms and homes and baseline speaker localization and enhancement software are also provided. This corpus stands apart from other corpora in the field by the number of rooms and homes considered and by the fact that it is publicly available at no cost. We describe the corpus specifications and annotations and the data recorded so far. We report baseline results.
Index Terms: home, distant-microphone, reverberation, noise, robustness, localization, enhancement, ASR.

1. Introduction

Distant-microphone speaker localization, speech enhancement, and speech recognition remain challenging tasks today [1–6]. The development of robust techniques able to fight reverberation and noise requires suitable corpora for development and testing. A number of real corpora are now publicly available for environments and application scenarios such as voice command for cars [7–9] and in public spaces [10], automatic transcription of lectures [11], meetings [12–14], dialogues [15, 16] and other public gatherings [17, 18], and automatic transcription of noisy or overlapped speech in broadcast media [19, 20].

More recently, distant-microphone speech processing in domestic environments has drawn much interest. This is explained not only by the financial stakes behind voice-controlled home automation and multimedia systems, human-robot communication, and speech monitoring and surveillance systems, but also by the difficult challenges raised by these environments. For instance, the reverberation time is typically higher than in, e.g., car or office environments. Talkers are located at variable distances from the microphone, from a few centimeters up to several meters. Noise backgrounds are often highly nonstationary and complex, due to the overlap of multiple noise sources such as competing talkers, TV/radio, footsteps, doors, kitchenware, electrical appliances, noise from outside...

The CHiME-1 and CHiME-2 corpora [21, 22] have contributed to popularizing research on robust speech processing in domestic environments. They feature real noise backgrounds collected in daily situations in a family home over the course of several weeks. Reverberation was generated by convolving clean speech with time-varying room impulse responses recorded in the same home using a binaural microphone setup. Reverberated speech was then scaled so as to match the intensity of normal voice at a distance of 2 m and added to randomly selected noise segments. The DIRHA Simulated corpus [23] was generated in a similar way, with more microphones across several rooms and simulated noise backgrounds obtained by summing individually recorded noises. Both corpora were released with baseline software tools [22, 24]. These corpora are realistic in several aspects and, as such, they promoted significant advances in the field. Yet, they differ from speech collected in real, ecological situations in several other aspects. For instance, in the real world, the intensity and stress level of speech depend on the amount of reverberation and noise and on the distance.

Few speech corpora have been collected in real homes so far. The DICIT corpus [25] features a constrained scenario, with talkers sitting in front of a smart TV. The DIRHA-English corpus [26] and the Sweet-Home corpus [27] relax this constraint, but are not publicly available[1]. Crucially, all recordings were made in a single home. This precludes the use of machine learning techniques (e.g., based on deep neural networks) for speech enhancement and recognition [28, 29], which require data collected in distinct homes for training, validation, and testing.

Motivated by these observations, we introduce a new corpus for distant-microphone speech processing in domestic environments, which is publicly available at no cost[2]. This corpus includes live speech from native French talkers in reverberant and noisy conditions, as well as room impulse responses and noise signals recorded in various homes with an 8-microphone device. We have already collected and annotated some data and we are currently collecting more in the scope of the voiceHome project, whose target application is the distant-microphone command of multimedia and smart home appliances via natural dialog.

In Section 2, we present the corpus specifications and annotations and the data recorded so far. In Section 3, we describe the baseline software tools and the resulting performance. We conclude in Section 4 by outlining the additional data to be recorded by the end of the voiceHome project.

[1] The authors of [26] plan to distribute it via the LDC for a fee.

[2] http://voice-home.gforge.inria.fr/voiceHome_corpus.html

2. Specifications, recording, and annotation

2.1. Specifications

The voiceHome corpus contains audio recordings, annotations and transcriptions of speech utterances from several speakers in diverse domestic noises, and room impulse responses that were recorded in various realistic, reverberant environments (RT_{60} varies from 395 to 585 ms).

Prompts are generated from two distinct grammars, one for home automation applications and one for multimedia applications, which result from industry specifications and user study and span the basic functionalities expected in a smart home controlled in natural language. All utterances starts with a keyword, "OK Vesta", to allow for future wake-up-word technology deployment. The keyword is followed either by:

- a question: "est-ce que la climatisation fonctionne?" *("is the air conditioning turned on?")*,
- a wish: "j'aimerais que tu m'enregistres le prochain Desperate Housewives" *("I'd like you to record the next Desperate Housewives for me")*,
- a command: "ouvre la porte du garage" *("open the garage door")*,

with possible adjuncts of time, space, or other adjuncts specifying the query. The vocabulary includes about 450 words, including named entities. The dataset does not contain true negatives until now but some will be added in future releases.

2.2. Recording protocol

With the exception of near-field speech, all audio data have been recorded as 16-bit, 16 kHz, 8-channel WAV files recorded by means of 8 MEMS microphones[3] plugged on the faces of a 10 cm cube (see Fig. 1). A USB interface allows direct digital recording from the array to a computer.

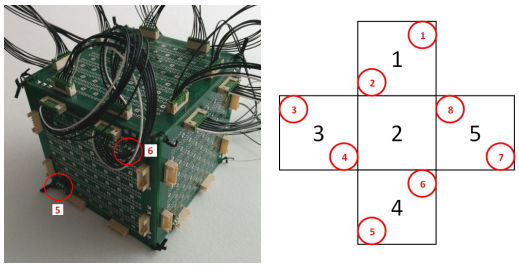

Figure 1: Microphone array (left) and a possible placement of microphones on the cube when unfolded (right).

All speech data were recorded during a unique session in two different rooms of a smart room experimental facility (`home1`), furnished and equipped to mimic a real home[4]. The position of the microphones was fixed during the whole session. Each of the 3 speakers (2 males, 1 female) was assigned a list of 20 sentences and asked to read them : *i)* All at once in the first room (noiseless, near-field, single-channel recording[5] with low reverberation); *ii)* 5 by 5, each group of 5 sentences corresponding to one noise condition (noiseless, vacuum cleaner, television broadcasting a talk show, someone washing dishes in the sink). Noise conditions are pre-determined for each sequence and manually produced or triggered by another experimenter. This sequence is repeated at 5 different predefined positions in the second room, represented in Fig. 2. The fact that the speaker doesn't move during one utterance is realistic, given the short duration of the utterances. Each recording contains 12 to 15 s of noise-only signal before and after each utterance.

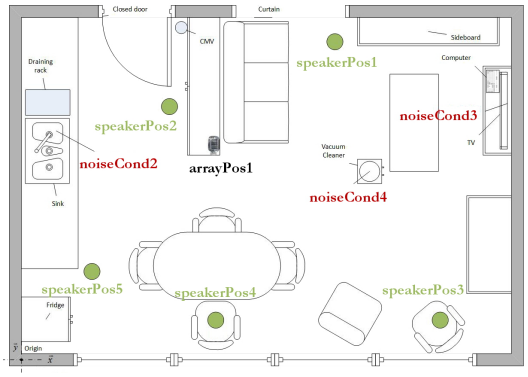

Figure 2: Schematic map of `home1/room1`.

Impulse responses were obtained by processing recordings of a 6 second chirp from 0 to 8 kHz, played by a loudspeaker[6], in 12 different rooms of 3 real homes (4 rooms per home: living room, kitchen, bedroom, bathroom). In each room, recordings were performed with 2 different positions of the microphone array and 7 to 9 different positions of the loudspeaker. These positions span a range of angles and are distributed logarithmically across distance. The recordings were then convolved with the inverse chirp to obtain the estimated room impulse responses, stored as 24-bit WAV files.

In addition, in each of the 12 rooms, 5 complex, everyday noise scenes relevant to the function of the room (speech, television, footsteps, meal preparation, shutters opening or closing, water flowing...) were recorded, at the same two array positions as above. Noise sources are different in each home.

2.3. Annotations and transcriptions

The experimental settings were documented in a series of annotation files describing: the global position of the microphone array and its orientation (in the room coordinate system), the positions of the 8 microphones (in the array coordinate system), the speaker position (free text description, such as "sitting on the couch", coordinates and orientation of the mouth), the dimensions and type of the room (at this stage, rough drawings of walls and furniture with handwritten annotations), and the noise condition (noise type, approximate noise position when fixed and known).

For speech recordings, transcriptions include the start and end times of each utterance (silence / noise only segments being labeled as [$NO_SPEECH]), the sentence that the speaker was asked to read (prompt), and an accurate transcription of what he/she actually uttered. Transcriptions were manually reviewed and corrected. All text files have been encoded in UTF-8.

2.4. Filenaming conventions

The dataset has the following directory structure:

[3] MP34DT01 Digital MEMS by ST Microelectronics.

[4] Pictures and more information on this platform can be consulted at `http://www.loustic.net`

[5] AKG CK91 microphone with pop filter and AKG SE300B preamplifier.

[6] KEF IQ3 120W 8Ω loudspeaker.

```
voiceHome_corpus/
  └annotations/
    └arrays/              (microphone positions)
    └rooms/          (room geometry with positions)
  └audio/
    └clean/                    (clean speech)
    └noises/                   (noise signals)
    └noisy/             (reverberated noisy speech)
    └rirs/               (room impulse responses)
  └transcriptions/
    └clean/           (clean speech transcriptions)
    └noisy/           (noisy speech transcriptions)
    └prompts/                       (prompts)
```

Each file of the dataset, irrespective of its nature, follows a naming convention describing its contents. It is built from the general pattern

```
home<>_room<>_arrayGeo<>
  _arrayPos<>_speaker<><>_speakerPos<>
              _noiseCond<>_uttNum<>.<ext>
```

where

- `home<>` and `room<>` encode the home index and the room index in that home, respectively;
- `arrayGeo<>` and `arrayPos<>` identify the position of the microphones and of the array;
- `speaker<><>` describes the speaker identity with two keys: a letter (Female/Male/Child) and a number;
- `speakerPos<>` is the speaker position identifier;
- `noiseCond<>` encodes the noise condition;
- `uttNum<>` is the utterance number;

and brackets `<>` are filled with integer indexes, allowing us to match each audio file with the corresponding annotations and transcriptions. Depending on the contents of the file, irrelevant fields are removed from the pattern to build the full filename. For instance, the file `home3_room2_arrayPos1.txt` contains annotations of the array position 1 in room 2 of home 3, and the file `uttNum26.txt` is the 26th prompt. Filetypes depend on the contents (.wav for audio files, .txt and .pdf for annotations). Exhaustive information about each field and full name by type of file can be found in the dataset documentation.

2.5. Summary

At the current stage, the voiceHome corpus contains:

- 60 clean utterances, 75 reverberated utterances (including quiet background noise) and 225 noisy (reverberated) utterances from 3 different speakers, for a total duration of about 2.5 hours (including background noise before and after the actual utterance);
- 188 8-channel impulse responses from 12 different rooms of 3 real homes;
- 120 minutes of various noises in the same rooms as for the impulse responses;
- full annotations for all data.

In particular, it includes challenging situations (high reverberation, low SNR, nonstationary noises with variable position or diffuse spatial distribution, impulse responses without direct path), which make it particularly suited for the development and test of next-generation speech processing techniques. Reverberated noisy speech is mainly intended for the testing of source localization, speech enhancement and speech recognition, while clean speech, impulse responses and noise-only signals are intended for training, in particular for generating data for acoustic model adaptation.

3. Baseline tools and results

In order to assess the realism and the level of difficulty of the voiceHome corpus as currently defined, we provide here some baseline tools and results of source localization, speech enhancement, and speech recognition.

3.1. Multichannel source localization

The location of the speaker or its Direction-of-Arrival (DOA) with respect to the microphone array is a valuable information for subsequent signal processing, in particular for speech enhancement and recognition. As a baseline, we used our own implementation of the state-of-the-art SRP-PHAT algorithm [30]. We made this implementation freely available, together with 7 other angular spectrum-based localization techniques [31], in a Matlab toolbox named Multichannel BSS Locate[7].

We investigated the capability of the source localization algorithm to return the correct DOA of the speaker, depending on the number of DOAs returned for each different noise condition of the `home1` subset. Fig. 3 displays the obtained results, as the cumulative recall of correct speaker DOA (with a tolerance of ±15°) with respect to the number of DOAs returned. Localization is performed with all 8 microphones on the portion of signal corresponding to the wake-up word.

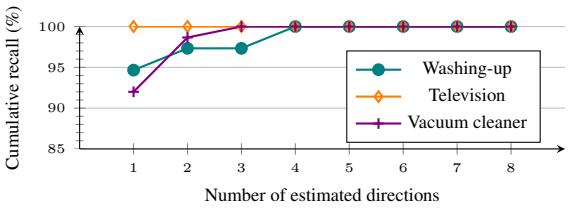

Figure 3: Baseline speaker localization performance.

98 % recall is achieved with this short analysis duration and only two estimated DOAs, which seems to be a sufficient performance for later enhancement (*e.g.*, by beamforming). In addition, the correct returned DOAs are accurate, with an average error of 0.96° in azimuth and 2.16° in elevation. For other enhancement approaches, *e.g.* source separation, an estimation of speaker and noise DOAs would be necessary. As seen in Fig. 4, this task is more difficult and requires returning a larger number of DOAs estimated on a longer time interval.

We stress here the diversity of noise conditions used in the corpus, which span different levels of difficulty or complexity of the scene. In particular, the recorded sounds include cases where no direct path was observed between the noise source and the microphones, and cases where the "ground truth" annotation can only be approximate (sources with diffuse spatial distribution or complex noise with several DOAs).

[7] `http://bass-db.gforge.inria.fr/bss_locate/#mbss_locate`

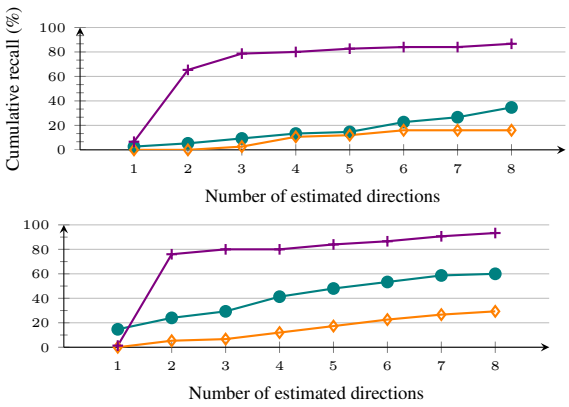

Figure 4: Baseline noise localization performance: short analysis duration (512 ms, top) vs. long analysis duration (whole utterance, 2.8 s average duration, bottom).

3.2. Enhancement by audio source separation

Source separation has proved to be a valuable enhancement strategy for subsequent speech recognition, as evaluated for instance in the CHiME challenges [21, 22]. We replicated the strategy deployed in [32] and adapted it to perform source separation on the voiceHome corpus using the FASST toolbox [33][8]. FASST is based on local Gaussian modeling in the time-frequency domain. The multichannel covariance of each source in each time-frequency bin is expressed as the product of a spatial term (spatial covariance matrix) and a spectral term (short-term power spectrum), which is itself factored into the product of template spectra and time activation coefficients by means of multichannel Nonnegative Matrix Factorization (NMF).

We consider that there are two sources: speech and noise. First, the single-channel near-field data (`audio/clean`) are used to train speaker-independent template spectra by 32-component NMF. This is done by a first call to the FASST toolbox, where the template spectra are initialized by vector quantization of the input magnitude spectrogram. Second, a spatial and spectral model of the noise is trained from the segments of noise preceding and following the utterance to be enhanced. This is achieved by a second call to FASST, where the spatial covariance matrices are initialized via a rank-1 model computed from the ground truth position of the noise source, and the template spectra are learned by 16-component NMF. Eventually, separation is performed by a last call to FASST, using the previous trained models and the ground truth speaker position to initialize its spatial covariance matrix. The template spectra are now kept fixed and only the time activation coefficients and the spatial covariance matrices are adapted to the test signal. 50 Expectation-Maximization (EM) iterations are performed within each call to FASST.

In the absence of ground truth noise and speech signals in the test subset (`home1/room1/`), we can only roughly evaluate the enhancement performance by computing the average SNR before and after separation. The annotation is used to compute the average noise power P_n (from the segments of noise preceding and following the utterance) and the average noise+speech power $(P_n + P_s)$ (from the utterance itself), and the SNR is estimated as P_s/P_n. The resulting SNRs are shown in Table 1.

[8] `http://bass-db.gforge.inria.fr/fasst/`

Noise condition	Input SNR (dB)	Output SNR (dB)
Washing up	5.7	12.4
Television	14.3	15.3
Vacuum cleaner	8.0	13.1

Table 1: Baseline speech separation performance. The average input SNR in the "noiseless" condition is 20.6 dB.

3.3. Speech recognition

We conducted a first speech recognition experiment to serve as a baseline for further use of the corpus. Training is performed on the ESTER corpus [34], which contains approximately 200 hours of broadcast speech, and test on the `home1/room1` subset of the voiceHome corpus. Speech recognition is conducted using a DNN-HMM acoustic model implemented using the Kaldi speech recognition toolkit [35]. We use 40 logmel features with 5 left and right context frames, that is an input feature dimension of $40 \times 11 = 440$. No cepstral mean and variance normalization (CMVN) was performed. The DNN output represents the senone states. The number of senones used was 4113. All the hidden layers were pretrained using restricted Boltzmann machines (RBMs). The RBMs are stacked together to form a deep belief network (DBN). The weights of the DBN are updated using the backpropagation algorithm with cross-entropy as the loss function. The two deterministic grammars (*cf.* Sec. 2.1) were used as language models for test. Table 2 sums up the first results obtained with this approach.

Test scenario	Home automation	Multimedia
WER (%)	67.93	71.30

Table 2: Baseline speech recognition performance.

Improved speech recognition baselines involving signal enhancement, model adaptation, and multi-condition training will be made available in the near future.

4. Conclusion

The voiceHome corpus, which will be fully completed in the near future, already provides a variety of multichannel distant-microphone speech, noise and room data, in realistic domestic environments. This data can serve for the development and testing of robust speech processing technology, including source localization, speech enhancement, and speech recognition. As real speech data will always be limited in quantity, this data is fundamentally meant to serve as a test set. By contrast, the possibility to mix collected room impulse responses with clean utterances and recorded noises seems to be the key solution to obtain enough data for efficient multi-condition training. In the coming months, we plan to enrich this corpus by extending the (currently insufficient) number of speakers, as well as collecting speech uttered in natural dialog scenarios and more ecological situations, for instance through a Wizard-of-Oz scheme. New data will be included in successive releases of the voiceHome corpus and made available at the same URL.

5. Acknowledgments

We acknowledge the support of Bpifrance (FUI voiceHome). We also wish to thank eSoftThings and Deltadore for implementing the microphone array prototype.

6. References

[1] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, May 2009.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[3] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech processing in modern communication: Challenges and perspectives*. Springer, 2010.

[4] E. Vincent and Y. Deville, "Audio applications," in *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010, pp. 779–819.

[5] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[6] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition — A Bridge to Practical Applications*. Elsevier, 2015.

[7] http://aurora.hsnr.de/aurora-3/reports.html.

[8] J. H. L. Hansen, P. Angkititrakul, J. Plucienkowski, S. Gallant, U. Yapanel *et al.*, ""CU-Move": Analysis & corpus development for interactive in-vehicle speech systems," in *Proc. Eurospeech*, 2001, pp. 2023–2026.

[9] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: audio-visual speech corpus in a car environment," in *Proc. Interspeech*, 2004, pp. 2489–2492.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 504–511.

[11] L. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillman, "The translingual English database (TED)," in *Proc. 3rd Int. Conf. on Spoken Language Processing (ICSLP)*, 1994.

[12] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. 2003 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, pp. 364–367.

[13] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language Resources and Evaluation*, vol. 41, no. 3–4, pp. 389–407, 2007.

[14] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The AMI and AMIDA projects," in *Proc. 2nd Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008, pp. 115–118.

[15] https://www.ll.mit.edu/mission/cybersec/HLT/corpora/SpeechCorpora.html.

[16] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments," *Computer Speech and Language*, vol. 26, no. 1, pp. 52–66, 2011.

[17] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. 2005 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2005, pp. 357–362.

[18] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The Sheffield wargames corpus," in *Proc. Interspeech*, 2013, pp. 1116–1120.

[19] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *Proc. 8th Int. Conf. on Language Resources and Evaluation (LREC)*, 2012, pp. 114–118.

[20] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *Proc. 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 687–693.

[21] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, May 2013.

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

[23] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC)*, 2014, pp. 2629–2634.

[24] A. Brutti, M. Ravanelli, P. Svaizer, and M. Omologo, "A speech event detection and localization task for multiroom environments," in *Proc. 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 157–161.

[25] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo, "WOZ acoustic data collection for interactive TV," in *Proc. 6th Int. Conf. on Language Resources and Evaluation (LREC)*, 2008, pp. 2330–2334.

[26] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *Proc. 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 275–282.

[27] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The sweet-home speech and multimodal corpus for home automation interaction," in *Proc. of the 9th edition of the Language Resources and Evaluation Conference*, 2014.

[28] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. 12th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2015, pp. 91–99.

[29] D. Yu and L. Deng, *Automatic Speech Recognition - A Deep Learning Approach*. Springer, 2014.

[30] J. Dibiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 8.

[31] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[32] A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, Florence, Italy, Sep. 2011. [Online]. Available: https://hal.inria.fr/inria-00598734

[33] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The Flexible Audio Source Separation Toolbox Version 2.0," ICASSP, May 2014, poster. [Online]. Available: https://hal.inria.fr/hal-00957412

[34] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news," in *Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC)*, 2006.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. 2011 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.