VoiceHome-2, an extended corpus for multichannel speech processing in real homes

Nancy Bertin, Ewen Camberlein, Romain Lebarbenchon, Emmanuel Vincent, Sunit Sivasankaran, Irina Illina, Frédéric Bimbot

VoiceHome-2, an extended corpus for multichannel speech processing in real homes[☆]

Nancy Bertin[a,*], Ewen Camberlein[a], Romain Lebarbenchon[a], Emmanuel Vincent[b], Sunit Sivasankaran[b], Irina Illina[b], Frédéric Bimbot[a]

[a] *IRISA - CNRS UMR 6074, Rennes, France*
[b] *Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

Abstract

We present a new, extended version of the voiceHome corpus for distant-microphone speech processing in domestic environments. This 5-hour corpus includes short reverberated, noisy utterances (smart home commands) spoken in French by 12 native French talkers in diverse realistic acoustic conditions and recorded by an 8-microphone device at various angles and distances and in various noise conditions. Noise-only segments before and after each utterance are included in the recordings. Clean speech and spontaneous speech recorded in 12 real rooms distributed in 4 different homes are also available. All data have been fully annotated. At last, we provide baseline software for speaker and noise localization, enhancement by source separation, and automatic speech recognition. 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, and we report baseline results.

Keywords: distant-microphone, reverberation, noise, robustness, localization, enhancement, ASR

1. Introduction

On October 18th, 2016, Microsoft announced that their automatic speech recognition (ASR) technology named Cortana had "reached human parity" in conversational speech recognition, a "historical achievement", to borrow the words of the press release[1], followed by a large number of articles in the general press and supported by later publications (Xiong et al., 2016). As system performance has increased in the last few years, notably thanks to the introduction of deep learning based technologies, commercial products have reached the market and the interest for the numerous applications of speech technology has grown: applications on mobile phones or tablets, games and toys, conference call systems, hands-free systems, hearing aids and other aids for people with disabilities, smart homes, etc.

In order to accompany the development and deployment of such end-user applications with wider and wider usage scenarios, it is now impossible to limit oneself to near-field capture of the user's speech. While this remains the most frequent use case today, speaker localization, speech enhancement, and ASR in distant-microphone scenarios remain challenging (Baker et al., 2009; Wölfel and McDonough, 2009; Cohen et al., 2010; Virtanen et al., 2012; Li et al., 2015; Vincent et al., 2018).

The development of robust techniques able to alleviate reverberation and noise, the main challenges of distant-speech processing "in the wild", 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

[☆] This work has received the support of BpiFrance (FUI voiceHome)
[*] Corresponding author
Email address: nancy.bertin@irisa.fr (Nancy Bertin)
[1] https://blogs.microsoft.com/next/2016/10/18/

for cars (Aurora-3, 2000; Hansen et al., 2001; Lee et al., 2004) and in public spaces (Barker et al., 2015), automatic transcription of lectures (Lamel et al., 1994), meetings (Janin et al., 2003; Mostefa et al., 2007; Renals et al., 2008), dialogs (LLSEC, 1996; Stupakov et al., 2011) and other public gatherings (Lincoln et al., 2005; Fox et al., 2013), and automatic transcription of noisy or overlapped speech in broadcast media (Gravier et al., 2012; Bell et al., 2015).

More recently, distant-microphone speech processing in domestic environments has drawn much interest. Emblematic devices such as Amazon Echo or Google Home embody this interest, which is explained not only by the financial stakes behind voice-controlled home automation and multimedia systems (or other applications such as human-robot communication, speech monitoring and surveillance systems, among others) 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, among others.

1.1. Comparable corpora

The CHiME series of challenges and corpora have contributed to popularizing research on robust speech processing in domestic environments. The first two corpora, CHiME-1 (Barker et al., 2013) and CHiME-2 (Vincent et al., 2013) 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 (Cristoforetti et al., 2014) 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 (Vincent et al., 2013; Brutti et al., 2014). 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 (Brutti et al., 2008) features a constrained scenario, with talkers sitting in front of a smart TV. The DIRHA-English corpus (Ravanelli et al., 2015) and the Sweet-Home corpus (Vacher et al., 2014) relax this constraint, but are not publicly available[2]. The DIRHA_AEC corpus (Zwyssig et al., 2015), specifically designed for echo cancellation, includes semi-simulated data and real data, with a larger number of scenarios of increasing complexity, but a limited number of noise conditions. The ATHENA corpus (Tsiami et al., 2014) provides real noisy speech data, together with video and Kinect recordings, with a larger diversity (20 speakers, 16 speaker positions, 20 noise conditions, heterogeneous and spatially distributed microphone network), recorded mostly in one room, plus 2 speaker positions and 2 microphones in a second. Crucially, all these 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 ASR (Weninger et al., 2015; Yu and Deng, 2014), which require data collected in distinct homes for training, validation, and testing. In addition, none of these corpora provides data in French, which is, according to the French government[3], the sixth most spoken language in the world (220 million speakers, 77 million native speakers, official language in 29 countries.)

[2]Ravanelli et al. (2015) plan to distribute DIRHA-English via the LDC for a fee. Samples of DIRHA-II, an extension of DIRHA-English to 3 more languages, can be downloaded from `https://shine.fbk.eu/resources/dirha-ii-simulated-corpus`.

[3]`https://www.diplomatie.gouv.fr/en/french-foreign-policy/francophony-and-the-french-language/the-status-of-french-in-the-world/`

1.2. The voiceHome and voiceHome-2 corpora

Motivated by these observations, we introduced in 2016 a new corpus named voiceHome (Bertin et al., 2016) for distant-microphone speech processing in domestic environments, which is publicly available at no cost[4]. Developed in the scope of the voiceHome project, whose target application is distant-microphone command of multimedia and smart home appliances via natural dialog, its first release included live speech from 3 native French talkers in reverberant and noisy conditions recorded in a smart room experimental facility furnished and equipped to mimic a real home[5], as well as room impulse responses and noise signals recorded in various homes with an 8-microphone device.

In order to enrich this corpus, as initially planned, we collected and annotated more data, including a larger number of homes, rooms and speakers. The result of this new campaign constitutes the voiceHome-2 corpus presented in this paper. For homogeneity and usability reasons, data from the first voiceHome corpus was not included in voiceHome-2. Some notable differences can be emphasized to explain this choice:

- In the voiceHome corpus, each speaker uttered the same sentence at 5 positions in each room, leading to little lexical and phonological diversity, and possible biases in experimental results if the corpus was to be separated in training and test subsets. In voiceHome-2, all uttered sentences are different.
- The duration of noise-only segments before and after each utterance was downsized (from 15 s before and after the utterance in voiceHome, to 5 s before and a short, variable duration after in voiceHome-2) to reduce recording time and storage size, while experimentally preserving performance.
- Two speakers would have appeared in two different homes if the two corpora had been merged. By keeping them separate, each speaker appears only in one home in voiceHome-2.
- In voiceHome-2, we recorded noisy speech in 3 rooms in each home, compared to only 1 in voiceHome.
- VoiceHome-2 includes 2 utterances per speaker and noise condition instead of 5 in voiceHome, which allows a larger variety of noises and rooms for the same total recording duration.
- VoiceHome-2 includes spontaneous speech. Only short, read utterances were recorded in voiceHome.

It must be noted that all voiceHome-2 data have been recorded in different homes than those used for voiceHome. The newly recorded data do not include additional room impulse responses and noise-only data. Indeed, the room impulse responses and the noise-only data in voiceHome can be used to generate simulated data for training (Ravanelli et al., 2016) while the new, real voiceHome-2 data can be used for testing. Table 1 comparatively summarizes the main features of the two corpora.

Corpus features	**voiceHome**	**voiceHome-2**
Homes (noisy speech)	1	4
Rooms (noisy speech)	1	12
Speakers (noisy speech)	3	12
Noise conditions (noisy speech)	3	36
Noise context duration (noisy speech)	15 seconds	5 seconds
Total duration* (noisy speech)	2.5 hours	5 hours
Number of utterances	360 (60 different)	1560 (all different)
Spontaneous speech	**no**	72 min
Room impulse responses	188 (12 rooms)	**no**
Noise-only signals	120 min (12 rooms)	**no**

Table 1: Comparative summary of voiceHome and voiceHome-2 corpora contents. * *These durations include the noise context before and after each utterance.*

[4] `https://zenodo.org/record/1252143`

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

All the above mentioned differences make the voiceHome-2 corpus a new corpus, which is described in detail in the rest of this paper. Used together as complementary, companion corpora, voiceHome and voiceHome-2 constitute a complete, self-contained dataset allowing for training, development and test of robust speech processing techniques. In this paper, particular emphasis is put on the new data from the voiceHome-2 corpus, while only necessary reminders about voiceHome are included. In Section 2, we present the corpus specifications and annotations, the recording protocol, and an exhaustive description of the newly recorded data. In Sections 3 and 4, we describe the baseline software tools for source localization and for speech enhancement and ASR, respectively, and the resulting performance. We conclude in Section 5 by outlining possible future uses and applications of this data.

2. Specifications, recording and annotations

The voiceHome-2 corpus contains audio recordings, annotations and transcriptions of speech utterances from several speakers in various realistic, reverberant and noisy domestic environments.

Put together, the two corpora gather various types of audio signals:

- room impulse responses (voiceHome),
- noise-only signals (voiceHome),
- short reverberated, noisy utterances following a scenario of home automation or multimedia appliance control (voiceHome, voiceHome-2),
- clean utterances of the same nature (voiceHome, voiceHome-2),
- spontaneous speech (voiceHome-2).

2.1. Specifications

The short utterances and the spontaneous speech in voiceHome-2 were recorded in 4 different homes. Each home was assigned to a group of 3 speakers (2 males, 1 female). In each home, recordings were performed in 3 different rooms, for a total of 12 different rooms. Each speaker assigned to a given home was recorded in each of the 3 rooms of the home.

2.1.1. Short utterances

The short utterances were generated from two distinct grammars, one for home automation applications and one for multimedia applications, designed from industry specifications and user studies and spanning the basic functionalities expected in a smart home. Both grammars were written in the Augmented Backus-Naur Form (ABNF) format, respecting the W3C Speech Recognition Grammar Specification V1.0 standard[6]. All utterances start with the keyword *« OK Vesta »* to allow for future wake-up-word technology deployment. This keyword is followed either by:

- a question: *« Qu'est-ce qu'il y a ce matin à la télé ? »* ("What's on TV this morning?"),
- a wish: *« Je veux éteindre tous les luminaires. »* ("I want to switch all the lights off."),
- a command: *« Redémarre le programme ! »* ("Restart the program!"),

[6] JSpeech Grammar Format files containing the full grammars are enclosed in the baseline recognition scripts, downloadable from the corpus webpage.

with possible adjuncts of time, space, or other adjuncts specifying the query. A small number of utterances are short commands (*« Pause ! »* — "Pause!") or possible followups to a first round of dialog (*« Et sur France 2 ? »* — "And what about Channel 2?").

The vocabulary includes 345 words, including a few named entities (mostly names of French TV channels). The dataset only contains utterances generated by the grammar (no true negative), as spontaneous speech excerpts can be used for this purpose. All sentences in the corpus differ from each other, resulting in a total of 1560 different utterances. Thus, each sentence can be straightforwardly associated to a quintuplet: home, room, speaker, speaker position, and noise condition.

Each recording chronologically contains 5 s of the predetermined noise condition, then the *« OK Vesta »* keyword immediately followed by the specific sentence uttered, and finally a short noise of variable duration depending on when the operator turned the recording off.

2.1.2. Spontaneous speech

In addition, spontaneous speech was also collected in each room. The speakers were asked to pick one topic among a list of suggested topics (your best vacation, a cooking recipe, etc.) or to choose one from their own imagination. They were given free time to prepare themselves and, once ready, spoke about the selected topic during 2 min.

2.1.3. Noises and room impulse responses

We remind here some information about the recordings from voiceHome, that may act as a companion training dataset for the new data.

Impulse responses were obtained by processing recordings of a 6 s chirp from 0 to 8 kHz, played by a loudspeaker[7], in 12 different rooms of 3 real homes (4 rooms per home: living room, kitchen, bedroom, bathroom). In each room, recordings were performed for 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.

In addition, in each of the 12 rooms, 5 complex, everyday noise scenes relevant to the function of the room (background speech, television, footsteps, meal preparation, shutters opening or closing, water flowing and so on) were recorded, at the same 2 array positions as above. The noise sources are different in each home.

We recall that no home nor room is shared between the voiceHome and voiceHome-2 corpora.

2.2. Speech recording protocol

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

Speech data were recorded at a pace of one daily session per home. The relative position of the microphones on the cube was fixed during the whole session. In order to ensure realism, the noise conditions were decided at the beginning of each recording session. They consisted of 1 "quiet" and 3 "noisy" conditions per room, determined by selecting objects present in the room and respecting the room's function and natural usage (hairdryer in the bathroom, cooking appliances in the kitchen, and so on). 5 speaker positions per room, including standing and sitting postures, were chosen to cover a large range of angles and distances with respect to the microphone array, which was placed at a single, fixed position for all the recordings in the room. The speaker positions were marked with duct tape and their coordinates were measured with a laser telemeter and written down, as well as the noise source locations, before proceeding with the recordings.

In each home, each of the 3 speakers (2 males, 1 female) was assigned a list of 120 sentences. He/she was equipped with a tie-clip microphone to ease the transcription process[9] and the approximate height of

[7] KEF IQ3 120W 8Ω loudspeaker.
[8] MP34DT01 Digital MEMS by ST Microelectronics.
[9] These data are not included in the corpus, since they are not synchronized with the microphone array recordings.

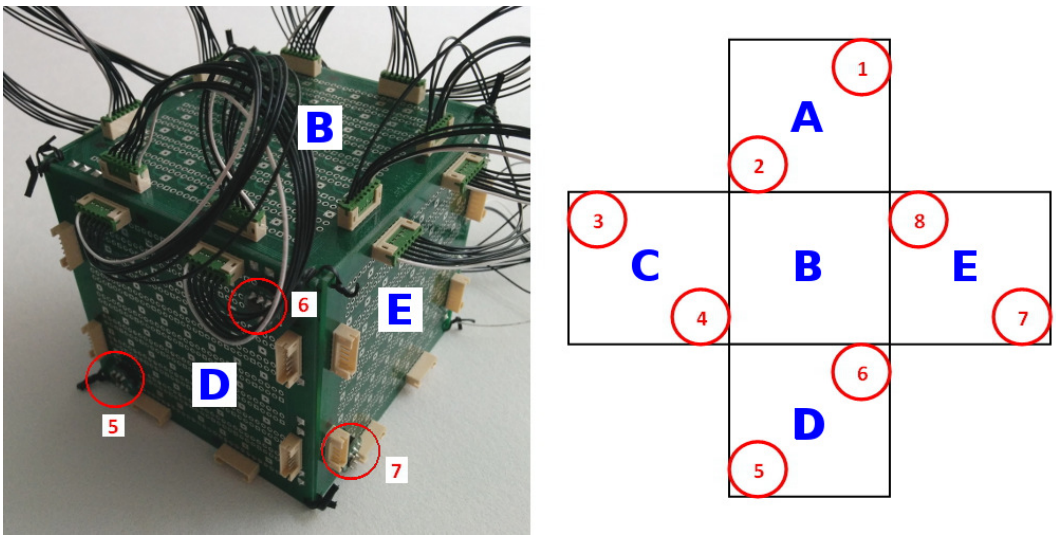

Figure 1: Microphone array (left) and schematic placement of microphones on the cube when unfolded (right). Microphones are identified by a number and faces of the cube by a letter. The precise positions of the microphones used for the recordings are available in the corpus annotations.

his/her mouth (sitting and standing) was measured. In order to avoid additional uncontrolled noise, only 2 people were present in the room besides the speaker: one operator who turned recording on and off from the control laptop, and one operator who produced the desired noise condition. The 2 other speakers for the session stayed in another room and remained as quiet as possible to avoid additional background noise.

The recordings were collected in the following order:

1. room by room;
2. in a given room, speaker by speaker;
3. in a given room and for a given speaker, position by position;
4. for a given speaker at a given position in the room, noise condition by noise condition.

For each position and noise condition, each speaker was asked to utter 2 sentences, one from the home automation grammar and one from the multimedia appliance control grammar, starting at the recording operator's hand sign (after recording 5 s of noise). The speakers were asked to face the microphone array and to avoid moving away from the target position, which is realistic given the short duration of the utterances, so that their actual position matches the annotation. They were however given no instruction regarding the way to speak. As a result, they adapted their voice in different, natural ways against reverberation, noise, and distance.

Once all 40 sentences had been recorded for a given speaker in a room, the speaker was assigned one position and one noise condition randomly drawn from the 5 positions and the 4 conditions associated with the room, excluding the "quiet" condition, and spontaneous speech was recorded for that speaker for approximately 2 min. These random assignments were constrained so that, in a given room, the 3 speakers were recorded at different positions and under different noise conditions.

The operation was repeated speaker by speaker and eventually room by room. In addition, all speakers were asked to utter 10 different sentences from the same grammars in clean conditions. These clean data were recorded in noiseless, low reverberation, close-microphone conditions with a single-channel AKG CK91 microphone with a pop filter and an AKG SE300B pre-amplifier.

All speakers signed a consent form for recording, dissemination and use of the corpus for research purposes.

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 of the mouth in the room coordinate system, orientation of the mouth in azimuth and elevation);
- The type of the room (free text description, such as "kitchen");
- The noise condition (noise type, approximate noise position when fixed and known).

For speech data, the transcriptions include the start and end time of each utterance (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. The transcriptions were manually reviewed by listening to the clip-tie microphone recording. All text files are encoded in UTF-8.

2.4. File naming conventions, documentation, and download

Directory structure and file naming conventions of the dataset have been made explicit to allow for easy use and automated parsing. The structure is identical to the first voiceHome corpus, except for the removal of the subdirectories `audio/rirs/` (room impulse responses) and `audio/noises/` (noise-only signals) and the addition of `audio/spontaneous/` and `transcriptions/spontaneous/` (spontaneous speech). Each file of the dataset, irrespective of its nature, follows a naming convention describing its contents, built from a general pattern from which irrelevant fields are discarded. Each audio file can be automatically matched with the corresponding annotations and transcriptions. Filetypes depend on the contents (.wav for audio files, .txt for annotations, .pdf for documentation).

An exhaustive list of filenames and information about each field can be found in the documentation, which can be downloaded separately or together with the data. The complete corpus and the documentation can be freely downloaded at no cost at `https://zenodo.org/record/1252143`.

2.5. Summary contents

Overall, the voiceHome-2 corpus contains:

- 120 clean utterances, 360 reverberated utterances in quiet, and 1080 reverberated utterances in noise from 12 different speakers recorded in 12 rooms distributed in 4 real homes, for a total duration of about 4 h (including background noise before and after the actual utterance);
- 36 chunks of spontaneous speech of 2 min each, for the same speakers and rooms;
- full annotations for all data.

Together with the voiceHome corpus, the two corpora form a complete set allowing for training, development and testing of speech processing applications. To highlight the complementarity of the two corpora, let us recall the contents of the voiceHome corpus:

- 8-channel impulse responses from 12 different rooms of 3 real homes,
- 120 min of various noises recorded in the same rooms,
- 60 clean utterances, 75 reverberated utterances in quiet and 225 reverberated utterances in noise from 3 different speakers,

for a total duration of about 2.5 h.

In particular, the corpora include challenging situations (high reverberation, low signal-to-noise ratio, nonstationary noises with variable position or diffuse spatial distribution, obstacles between the speaker and the microphone array), which makes them particularly suited for the development and testing of next-generation speech processing techniques. The diversity and realism of room types and noise conditions can be assessed from the list in Table 2. Reverberated noisy speech is mainly intended for the testing of source localization, speech enhancement and ASR, while clean speech, impulse responses and noise-only signals are intended for generating simulated data for training.

Home	Room	Noise conditions
Home1: house	Room1: kitchen	Cooker hood
		Water flowing in a sink filled with dishes
		Someone crumpling up a sheet of paper
	Room2: living room	Music played on a loudspeaker
		Vacuum cleaner (fixed position)
		Electric roller shutter going up and down
	Room3: library	Someone playing with the strings of a ukulele
		Someone rubbing polystyrene pieces
		Someone shaking a bag full of gaming plastic tiles
Home2: flat	Room1: living room	Someone flipping through a book
		Music played on a loudspeaker
		Someone using an indoor bike
	Room2: kitchen	Someone using a manual roller shutter
		Someone washing the dishes
		Someone opening and closing the oven door
	Room3: bedroom	Someone opening and closing the storage cupboard door
		Someone pressing the keys of a laptop
		Someone playing with the window handle
Home3: house	Room1: living room	Someone reading the newspaper
		Music played on a loudspeaker
		Someone shaking a box of pencils
	Room2: dining room	Someone playing with a toy
		Someone using a bell
		Metronome
	Room3: kitchen	Someone playing with cutlery
		Someone shaking a jar of seeds
		Beeps played by a microwave
Home4: house	Room1: living room	Music played on a loudspeaker
		Someone shaking keys attached to a keyring
		Vacuum cleaner (fixed position)
	Room2: kitchen	Blender
		Cooker hood
		Someone shaking gently a crate of empty glass bottles
	Room3: bathroom	Electric razor
		Water flowing in the sink
		Hairdryer (fixed position)

Table 2: Types of homes, rooms and noises in the voiceHome-2 corpus. Rooms span floor areas from roughly 8 to 30 m². The "quiet" noise condition in each room is omitted.

3. Multichannel source localization

In addition to the above data, we distribute baseline software tools for source localization and for speech enhancement and ASR. Indeed, the direction-of-arrival (DOA) of the speaker with respect to the microphone array is a valuable piece of information for subsequent signal processing, in particular for certain speech enhancement and ASR methods.

3.1. Baseline Localization System

As a baseline for speaker localization, we use our own implementation of the state-of-the-art steered response power with phase transform (SRP-PHAT) algorithm (Dibiase et al., 2001). We made this implementation freely available, together with 7 other angular spectrum-based localization techniques (Blandin et al., 2012), in a Matlab toolbox named Multichannel BSS Locate[10].

The general principle of SRP-PHAT is to compute a function $\Phi(\theta, \varphi)$ termed "angular spectrum", where θ and ϕ are azimuth and elevation variables, which is expected to exhibit local maxima in the directions of active sources. The angular spectrum is first computed for each microphone pair and then aggregated across pairs. The computation consists of the following steps:

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

1. Define the search space, *i.e.* a grid of possible DOAs (θ_j, φ_k) for which we want to evaluate Φ in the global coordinate system;
2. For each microphone pair n:
   (a) Compute the corresponding angles of arrival (AOAs) $\{\alpha_{jk}^{(n)}\}_{jk}$ with respect to the microphone pair;
   (b) Resample $\{\alpha_{jk}^{(n)}\}_{jk}$ into a smaller set $\{\alpha_i^{(n)}\}_i$ in order to reduce computation time;
   (c) Compute the time differences of arrival $\{\tau_i^{(n)}\}_i$ between the two microphones corresponding to the AOAs $\{\alpha_i^{(n)}\}_i$;
   (d) Compute the generalized cross-correlation with phase transform (GCC-PHAT) $\Phi_n(t, \tau_i^{(n)})$ (Knapp and Carter, 1976) between the signals recorded at the two microphones in each time frame t and for each time difference of arrival $\tau_i^{(n)}$;
   (e) Interpolate it back to the original angle resolution to obtain the local angular spectrum $\Phi_n(t, \alpha_{jk}^{(n)})$.
3. Compute the global spectrum $\Phi(\theta_j, \varphi_k)$ by pooling the local angular spectra $\Phi_n(t, \alpha_{jk}^{(n)})$ over all time frames t and across all microphone pairs n. A pooling method (such as maximum or sum) must be chosen for this purpose.
4. Find the indexes j and k of the largest peak (single source case) or peaks (multiple source case) of $\Phi(\theta_j, \varphi_k)$, yielding the estimated source azimuth(s) θ_j and elevation(s) φ_k.

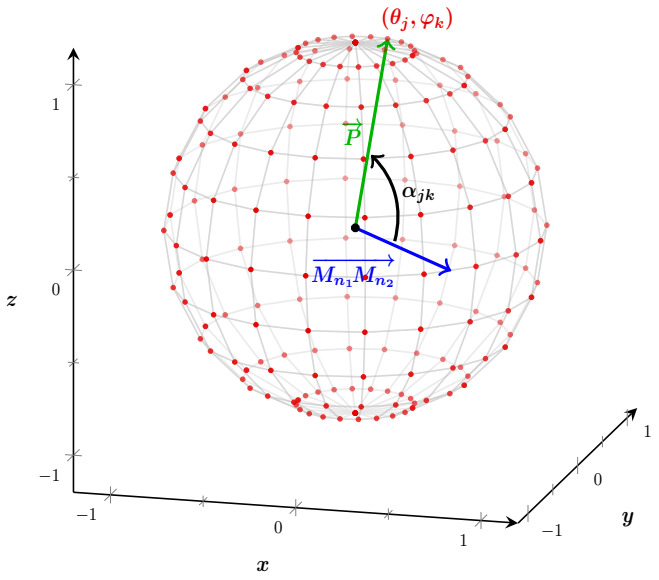

Figure 2: Sampling of the DOA sphere and AOA computation for SRP-PHAT. $\overrightarrow{M_{n_1}M_{n_2}}$ is the vector defined by the considered microphone pair. The DOA (θ_j, φ_k) and the corresponding AOA α_{jk} for this microphone pair are illustrated for one point on the sphere. The number of sampled azimuths is constant per elevation.

The sampling of the DOA sphere and the computation of the AOA are illustrated in Fig. 2. More details about interpolation and pooling can be found in the toolbox documentation. These operations, as well as the definition of the search space, can be tuned by the user and depends on the desired resolution. Experimental results presented in Sec. 3.3 were obtained by limiting the search space to the upper half sphere, sampling it at a resolution of 1° in azimuth and elevation in the global frame, then resampling them at 5° in the microphone pair local frame. This results in a subsampling factor of approximately 32.

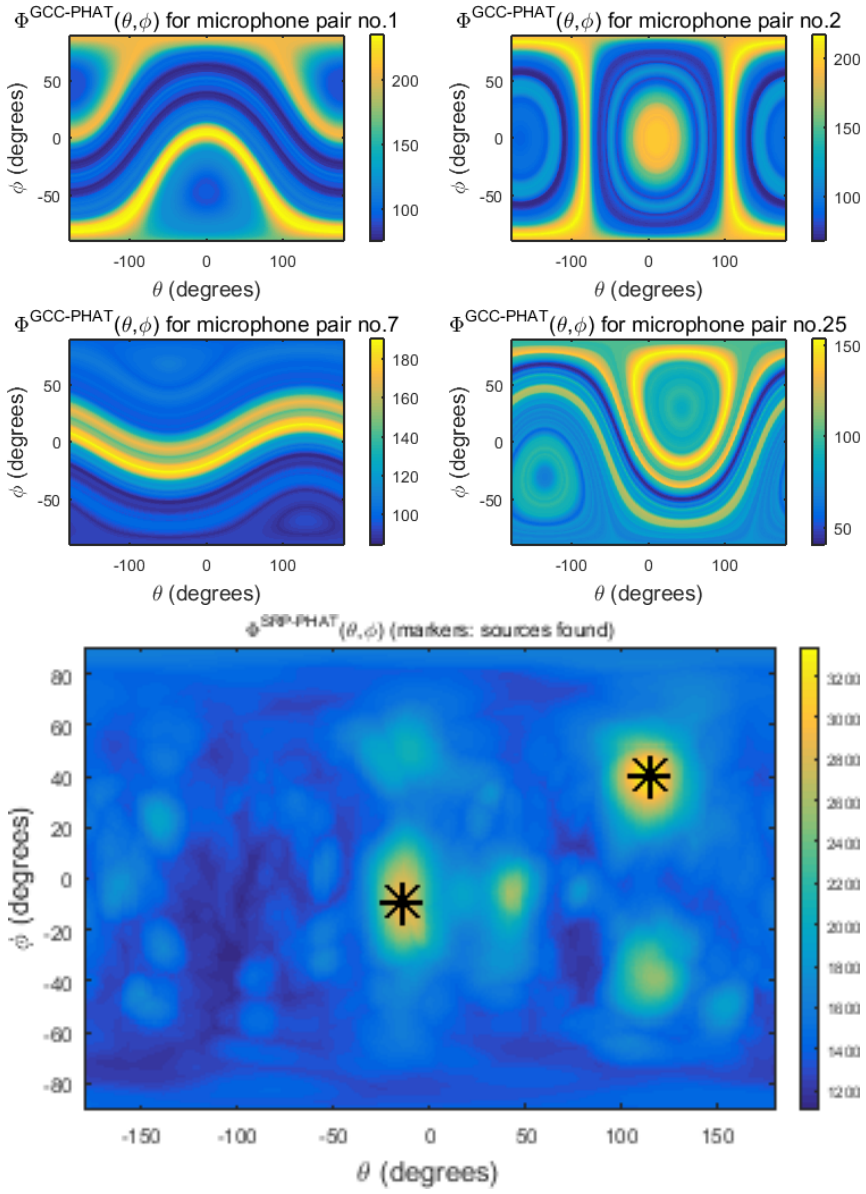

Figure 3: Top: 4 examples of local GCC-PHAT angular spectra (2-channel, *i.e.* 1 microphone pair, each). Bottom: global SRP-PHAT angular spectrum (8-channel, *i.e.* 28 microphone pairs). All angular spectra are computed on utterance 1295 and pooled over time. All angles (θ, φ) are expressed in the same global coordinate system.

The application of SRP-PHAT to one noisy speech file of the voiceHome-2 corpus is illustrated in Fig. 3. The top four plots represent the local angular spectra for four different pairs of microphones. The "projection" in the global coordinate system of the cone of confusion that defines the intrinsic localization indeterminacy with 2 microphones is clearly visible. The bottom plot represents the global angular spectrum after aggregation of all 28 microphone pairs and the locations of the two detected sources. Thanks to the multichannel setting, the indeterminacies have been resolved and the contrast has improved.

3.2. Experimental setting and performance measures

We investigated the capability of SRP-PHAT to return the correct DOA of the speaker and the noise source. Based on the ground truth timestamps, *joint speaker and noise localization* is performed on the time

interval containing the keyword and the sentence. The results are reported separately on average over all "quiet" conditions and all noisy conditions. The speaker DOA and the noise DOA (in noisy conditions) are assumed to be the largest and the second largest peak of the angular spectrum, respectively. In addition, *noise-only localization* is performed on the first 4 s of the recording which contain noise only. The noise DOA is then assumed to be the largest peak.

The number of detected DOAs is varied from 1 to 8. The results are evaluated in terms of recall, that is the proportion of correct DOAs returned, and localization error in degrees, computed only for those DOAs which are considered as correct. These metrics are computed in three different ways:

- in azimuth only — an estimated DOA is considered as correct if the absolute difference with the ground truth azimuth is less than 10° and the error is the average absolute difference for correct DOAs;
- in elevation only — an estimated DOA is considered as correct if the absolute difference with the ground truth elevation is less than 10° and the error is the average absolute difference for correct DOAs;
- in both azimuth and elevation — an estimated DOA is considered as correct if the angle it forms with the correct DOA is less than 10°, and the error is the average angle between the unit norm vectors pointing in the true and estimated directions[11].

3.3. Results

3.3.1. Speaker localization

The performance of speaker localization in quiet and in noise is displayed in Fig. 4. Due to the 3D array geometry, the recall and the error are similar in azimuth and in elevation. In quiet, correct speaker localization is systematically achieved, with 99% recall for one returned DOA and 100% for 2 or more returned DOAs, and the correct DOAs are within 2° of the ground truth. This is enough for subsequent enhancement, *e.g.*, by beamforming. In noise, the recall severely degrades when compared to the performance in quiet but the loss of accuracy (in the order of 0.2°) is insignificant.

Due to thresholding effects (we recall that average errors are computed only on estimated DOAs considered as correct with respect to the 10° tolerance on the considered criterion), the average localization error in total angle apparently doesn't benefit from increasing the number of estimated DOAs, even while azimuth and elevation errors are improved separately. This is of little significance when compared to the precision of ground truth measurements (5°), the resolution of the sphere sampling (1°) and the variance of the errors (about 1.5 to 2°, not shown here.)

The diversity and size of voiceHome-2 allow deeper investigation. As an example, we divided the corpus into two subsets of 720 utterances, depending on whether the speaker distance is below or above 2 m. The differentiated results highlight the impact of the speaker distance. Obviously, the recall in quiet remains very close or equal to 100% irrespective of the distance. However, the recall in noise (see Fig. 5) decreases for larger distances. The average loss of accuracy (not shown here) from small to large distances is insignificant.

Overall, this indicates that the target DOA is generally among the estimated DOAs but it is not always the first one in noisy conditions, especially when the speaker is far from the microphone array and the noise source may be closer or stronger. Possible strategies to retrieve the correct DOA among the returned DOAs could be: i) to perform localization on noise-only intervals (see below) and exclude the resulting noise DOAs, or ii) to classify the corresponding signals to determine which one is the targeted speech.

[11] The angle is defined as the acos of the dot product between two unit vectors pointing in the estimated and true directions. Note that this criterion is more restrictive than would be a logical "AND" condition from the two previous. Indeed, when the errors in azimuth only and elevation only are close enough to the 10° threshold, azimuth is considered as correct, elevation as correct, but the DOA "in both" is not.

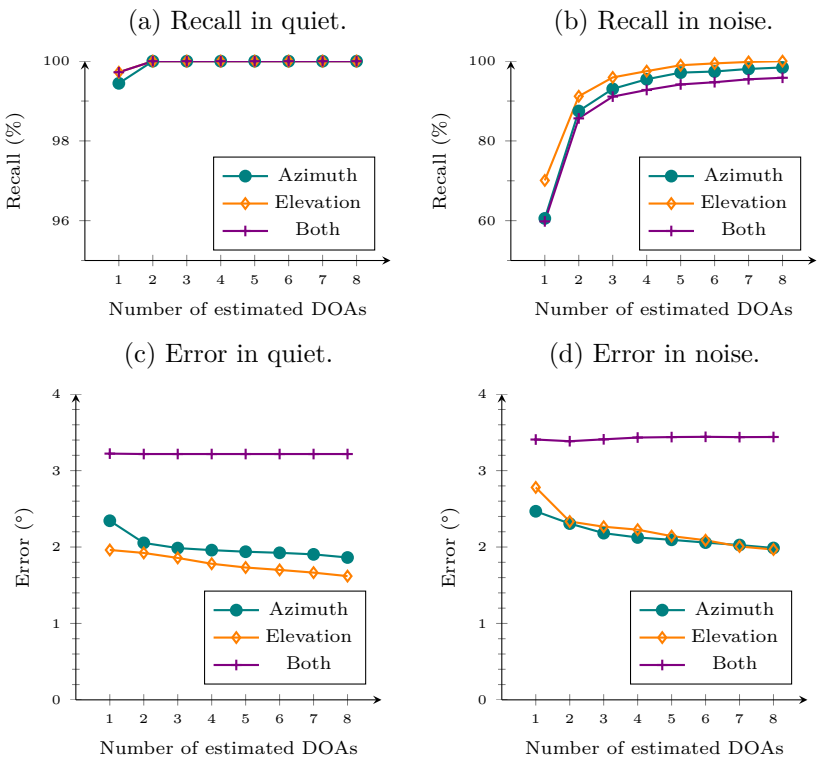

Figure 4: Speaker localization performance in quiet and in noise.

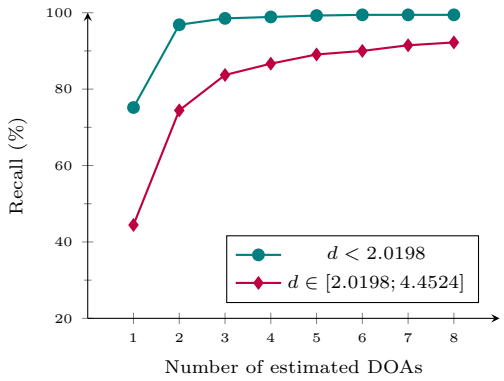

Figure 5: Speaker localization recall in noise as a function of the speaker distance d.

3.3.2. Noise localization

Besides speaker localization, noise localization is also useful for source separation and speech enhancement. As seen in Fig. 6, this task is harder and requires returning a larger number of DOAs. Joint speaker and noise localization on the same time interval (keyword and sentence) is difficult: the recall in both azimuth and elevation reaches 75% only when 5 or more DOAs are returned. Much better results can be obtained if localization is performed on the first 4 s of each recording, with 75% recall in both azimuth and elevation for 1 DOA returned and 95% recall in either azimuth or elevation with 4 DOAs returned. Thus, in realistic applications, estimating the speaker and noise DOAs on different time intervals would be preferable. In practice, such a strategy could be deployed if: i) the last few seconds of noise before the keyword can be

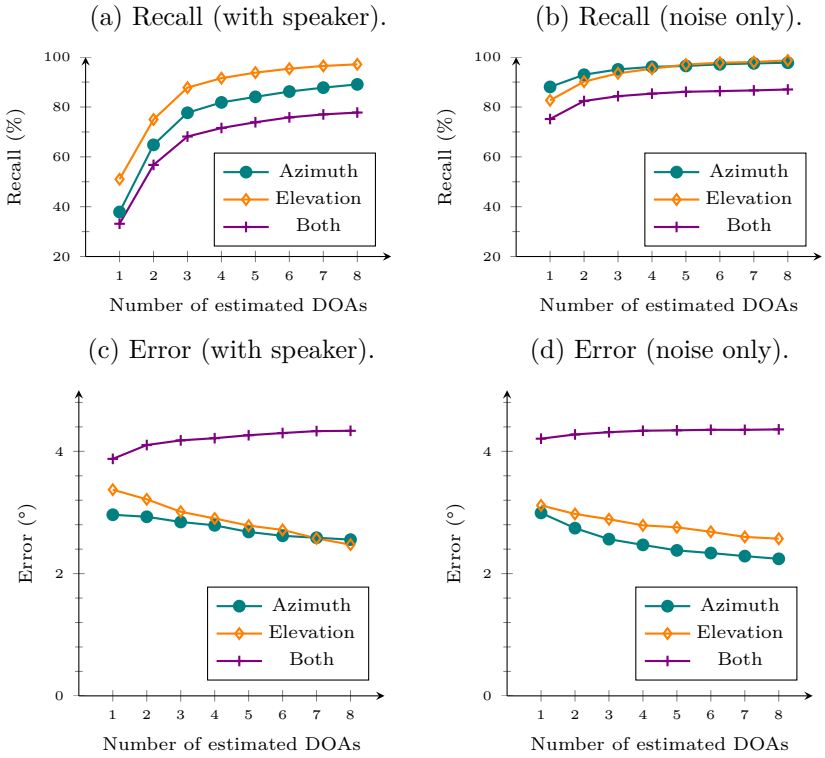

Figure 6: Noise localization performance on either noisy speech or noise-only intervals.

accessed via a buffer, or ii) a voice activity detection module or an event classification module is employed to detect noise-only intervals during the utterance.

We stress here the diversity of the noise conditions encountered in the corpus, which span different levels of difficulty or complexity of the scene. In particular, the recorded sounds include diffuse noise sources, complex noises with several DOAs, and cases where no direct path was observed between the noise source and the microphones.

4. Speech enhancement and ASR

4.1. Baseline Speech Enhancement System

Source separation has proved to be a valuable speech enhancement strategy for subsequent ASR, as evaluated for instance in the CHiME challenges (Barker et al., 2013; Vincent et al., 2013). We replicated the strategy deployed by Ozerov and Vincent (2011) and adapted it to perform source separation on the voiceHome-2 corpus using the FASST toolbox (Salaün et al., 2014)[12]. 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 clean data (`audio/clean`) are used to train speech 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

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

magnitude spectrogram. Second, for each utterance to be enhanced, a spatial and spectral model of the noise is trained from the first 4 s of the recording which contain noise only. This is achieved by a second call to FASST, where the spatial covariance matrices are initialized via a rank-1 model computed from the DOA of the noise source. Noise template spectra are learned by 16-component NMF. Finally, actual source separation is performed by a last call to FASST, using the previously trained models and the speaker DOA 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. Given the template spectra and the estimated time activation coefficients and spatial covariance matrices, separation is achieved by multichannel Wiener filtering.

We implemented several configurations based on this general approach:

- Two time-frequency representations can be used: either a 1024-bin short time Fourier transform (STFT) or an 8-band equivalent rectangular bandwidth (ERB) transform (Vincent et al., 2010). These choices respectively correspond to a high-quality configuration with high frequency resolution and a real-time configuration with lower frequency resolution but real-time operating capability.

- The speech template spectra can be either speaker-dependent (only clean data from the test speaker are used for training) or speaker-independent (clean data from the other 11 speakers excluding the test speaker are used for training).

- Either the ground truth speaker and noise DOAs or the estimated DOAs are used for initialization of the spatial covariance matrices of speech and noise. The estimated DOAs are the first DOA returned by SRP-PHAT on the time interval containing the keyword and the sentence (speaker localization) and the first DOA returned on the first 4 s of the recording (noise localization).

- The estimated short-term power spectra of speech and noise, that are the product of the estimated template spectra and time activation coefficients, are smoothed over time using a sliding rectangular window as described by Vincent (2010, eq. (8)) before computing the multichannel Wiener filter. The length of the window is set to 1 (no smoothing), 5, 9, or 17 frames. Smoothing helps tuning the tradeoff between noise reduction and speech distortion (Vincent, 2010). By reducing speech distortion (at the cost of increased residual noise), better ASR performance can often be obtained (Virtanen et al., 2012).

This resulted in 32 different configurations in total.

4.2. Baseline ASR System

We conducted ASR experiments on noisy and enhanced data to serve as a baseline for further use of the corpus. The ASR baseline was implemented using Kaldi (Povey et al., 2011).

The acoustic features are 13 mel frequency cepstral coefficients (MFCCs), which are concatenated with 3 frames of left and right context, reduced to 40 dimensions by linear discriminant analysis (LDA), transformed by feature-space maximum likelihood linear regression (fMLLR), and eventually concatenated with 5 frames of left and right context. The effectiveness of this feature pipeline for distant-microphone ASR was shown by Tachioka et al. (2013). The acoustic model is a deep neural network (DNN) with 7 hidden layers and 2048 sigmoid units per layer. The DNN outputs represent 4113 hidden Markov model (HMM) states. The parameters are pretrained using restricted Boltzmann machines and updated by backpropagation using cross-entropy as the loss function. Early stopping is conducted based on a validation set. The two deterministic grammars (cf. Section 2.1.1) are used as language models.

Training is performed on a subset of the ESTER corpus (Galliano et al., 2006), which contains 43 h of "clean" broadcast speech randomly split into 90% for training and 10% for validation. A Gaussian mixture based acoustic model (GMM-HMM) is first trained on these clean data to obtain alignments. A simulated multi-condition corpus is then generated by convolving each utterance of ESTER with a room impulse response and adding a noise signal. The room impulse response and the noise signal are randomly chosen from the voiceHome corpus and they are distinct for every utterance, in the limit of the total number

Smoothing	Speaker independent		Speaker dependent	
	Est. DOA	True DOA	Est. DOA	True DOA
None	6.03	4.90	5.30	4.86
5 frames	5.89	5.04	**5.26**	**4.61**
9 frames	**5.83**	4.88	5.30	4.73
17 frames	6.03	**4.85**	5.36	4.70

Table 3: Baseline WER (%) after speech enhancement using a 1024-bin STFT representation.

of impulse responses and noise signals. The signal-to-noise ratio is set such that the intensity of speech matches that actually recorded in these homes. We recall again that these homes are different from those used in voiceHome-2 and that the speakers are totally disjoint. The DNN acoustic model is trained on this multi-condition corpus using the clean alignments obtained via the GMM-HMM as targets. Note that the model is not retrained on enhanced training data (as suggested by (Yoshioka et al., 2015), this should not be detrimental to ASR performance.)

4.3. Performance measures

In contrast with popular semi-simulated corpora cited in the introduction, voiceHome-2 consists of real data only. This favors realism over the availability of a ground truth speech signal. As such, the computation of classical source separation metrics such as SDR, SAR, SIR (Vincent et al., 2006) is excluded here. In line with the target application, we retain the word error rate (WER) as the main performance metric for ASR but also for speech enhancement. The quality of separation is indirectly assessed through the WER improvement it brings when used as a preprocessing step for ASR. The 95% confidence interval on the WERs reported below is in the order of ±0.3% (for the smallest WER on enhanced data) to ±0.4% (for the largest WER).

4.4. Results

Without enhancement, our baseline ASR system yields a WER of 2.15% on clean data and 8.15% on noisy, unprocessed data. The results obtained after speech enhancement are presented in Table 3 for STFT based separation and in Table 4 for ERB based separation.

As now well established, despite the considerable progress lately made in ASR thanks to the introduction of DNN-based technologies and multi-condition training, STFT-based source separation still improves the WER compared to processing the original noisy data. The improvement is as large as 43% relative using speaker-dependent models initialized with the true DOAs and 28% relative using speaker-independent models initialized with the estimated DOAs.

In line with the above localization results, the use of the ground truth speech and noise DOAs provides notably better performance than the estimated DOAs. Indeed, the recall in the order of 60% for the first returned speaker DOA in noisy conditions (see Fig. 4) translates into the fact that the initial spatial covariance matrices sometimes point to a noise source instead of the speaker. FASST can partially recover from such bad initialization thanks to the information provided by the pretrained template spectra, but not always. The strategies listed at the end of Section 3 to improve speaker localization are expected to improve the WER too.

Disappointingly, and in contrast with previous preliminary experiments (unpublished), the real-time ERB-based separation configuration does not preserve the WER improvement to some extent compared to the STFT-based configuration. Although the WER decreases compared to noisy speech processing when the true DOAs are used, the accumulation of localization errors and separation results in a WER increase otherwise. This suggests that further research is needed to identify the best tradeoff between quality and computation time, by exploring more finely the range of parameters controlling this tradeoff.

Also, despite previous evidence of its benefit for improving the separation quality (Vincent, 2010), temporal smoothing does not improve the WER significantly. One possible explanation, which remains to be

Smoothing	Speaker independent		Speaker dependent	
	Est. DOA	True DOA	Est. DOA	True DOA
None	8.65	**6.68**	8.91	6.83
5 frames	8.58	6.72	**8.61**	**6.66**
9 frames	8.68	6.96	8.66	6.96
17 frames	**8.32**	6.81	8.40	6.84

Table 4: Baseline WER (%) after speech enhancement using an 8-band ERB representation.

investigated, is that the benefit of smoothing was greater for GMM-HMM based ASR than it is for the latest DNN-based technologies.

5. Conclusion

The voiceHome and voiceHome-2 corpora provide a variety of multichannel noise and room data and distant-microphone speech, in an unprecedented number of realistic domestic environments. These datasets can serve for the development and testing of robust speech processing technology, including speech and noise localization, speech enhancement, and ASR. As real speech data will always be limited in quantity, the real speech data in voiceHome-2 are fundamentally meant to serve as a test set. By contrast, the possibility to mix clean utterances with the room impulse responses and the recorded noises in voiceHome is the key to obtain enough data for efficient multi-condition training. For these reasons, the effort in enriching the first voiceHome corpus was primarily put on the diversity of recorded noisy, reverberant speech, resulting in the new voiceHome-2 corpus presented here, which includes 12 speakers, 12 rooms inside 4 different real homes, and 48 realistic noise conditions.

The acoustic diversity in noises and environments provided in voiceHome-2 represent an additional resource in the current common effort of the community towards more realism in the development and evaluation of robust speech processing techniques. The baseline localization, enhancement, and ASR software provided together with the data can be used as the basis for the development of future technology and the recall, localization error, and WER results presented in this article as a basis for evaluating progress. The experiments presented in this article were conducted on short utterances only, since spontaneous speech data were manually transcribed in parallel to writing the article and the transcriptions were finalized too late to run the experiments. We will however be running localization, enhancement, and ASR experiments on spontaneous speech data in the near future and release the corresponding baselines.

In the future, the evaluation of the latest DNN-based enhancement algorithms on voiceHome-2, and the collection of speech uttered in natural dialog scenarios and in more ecological situations (for instance through a Wizard-of-Oz scheme) should allow pursuing this effort towards realistic speech processing techniques.

6. Acknowledgements

We acknowledge the support of Bpifrance (FUI voiceHome). We also wish to thank eSoftThings and Deltadore for implementing the microphone array prototype, D. Fohr and O. Mella for their help with setting up the clean ESTER baseline, and all the voluntary speakers and those who shared their homes for the recording sessions.

7. References

Aurora-3, 2000. `http://aurora.hsnr.de/aurora-3/reports.html`.

Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., O'Shaughnessy, D., 2009. Research developments and directions in speech recognition and understanding, part 1. IEEE Signal Processing Magazine 26 (3), 75–80.

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

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The PASCAL CHiME speech separation and recognition challenge. Computer Speech and Language 27 (3), 621–633.

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

Bertin, N., Camberlein, E., Vincent, E., Lebarbenchon, R., Peillon, S., Lamandé, É., Sivasankaran, S., Bimbot, F., Illina, I., Tom, A., Fleury, S., Jamet, E., 2016. A French corpus for distant-microphone speech processing in real homes. In: Proc. Interspeech. pp. 2781–2785.

Blandin, C., Ozerov, A., Vincent, E., 2012. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Processing 92 (8), 1950–1960.

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

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

Cohen, I., Benesty, J., Gannot, S. (Eds.), 2010. Speech Processing in Modern Communication: Challenges and Perspectives. Springer.

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

Dibiase, J., Silverman, H., Brandstein, M., 2001. Robust localization in reverberant rooms. In: Microphone Arrays: Signal Processing Techniques and Applications. Springer, pp. 157–180.

Fox, C., Liu, Y., Zwyssig, E., Hain, T., 2013. The Sheffield wargames corpus. In: Proc. Interspeech. pp. 1116–1120.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., Choukri, K., 2006. 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). pp. 139–142.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., Galibert, O., 2012. 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). pp. 114–118.

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

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

Knapp, C., Carter, G., 1976. The generalized cross-correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing 24 (4), 320–327.

Lamel, L., Schiel, F., Fourcin, A., Mariani, J., Tillman, H., 1994. The translingual English database (TED). In: Proc. 3rd Int. Conf. on Spoken Language Processing (ICSLP). pp. 1795–1798.

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

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

Lincoln, M., McCowan, I., Vepa, J., Maganti, H. K., 2005. 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). pp. 357–362.

LLSEC, 1996. Lincoln laboratory speech enhancement corpus. `https://www.ll.mit.edu/mission/cybersec/HLT/corpora/SpeechCorpora.html`.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S., Tyagi, A., Casas, J., Turmo, J., Cristoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., Rochet, C., 2007. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. Language Resources and Evaluation 41 (3–4), 389–407.

Ozerov, A., Vincent, E., 2011. Using the FASST source separation toolbox for noise robust speech recognition. In: Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 86–87.

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

Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A., Omologo, M., 2015. 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). pp. 275–282.

Ravanelli, M., Svaizer, P., Omologo, M., 2016. Realistic multi-microphone data simulation for distant speech recognition. In: Proc. Interspeech. pp. 2786–2790.

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

Salaün, Y., Vincent, E., Bertin, N., Souviraà-Labastie, N., Jaureguiberry, X., Tran, D. T., Bimbot, F., 2014. The Flexible Audio Source Separation Toolbox Version 2.0. ICASSP Show & Tell.

Stupakov, A., Hanusa, E., Vijaywargi, D., Fox, D., Bilmes, J., 2011. The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments. Computer Speech and Language 26 (1), 52–66.

Tachioka, Y., Watanabe, S., Le Roux, J., Hershey, J. R., 2013. Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark. In: Proc. 2nd International Workshop on Machine Listening in Multisource Environments (CHiME). pp. 19–24.

Tsiami, A., Rodomagoulakis, I., Giannoulis, P., Katsamanis, A., Potamianos, G., Maragos, P., 2014. ATHENA: a Greek multi-sensory database for home automation control. In: INTERSPEECH. ISCA, pp. 1608–1612.

Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., Bonnefond, N., 2014. The Sweet-Home speech and multimodal corpus for home automation interaction. In: Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC). pp. 4499–4509.

Vincent, E., 2010. An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation. In: Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA). pp. 157–164.

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

Vincent, E., Bertin, N., Badeau, R., 2010. Adaptive harmonic spectral decomposition for multiple pitch estimation. IEEE Transactions on Audio, Speech, and Language Processing 18 (3), 528–537.

Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14 (4), 1462–1469.

Vincent, E., Virtanen, T., Gannot, S. (Eds.), 2018. Audio Source Separation and Speech Enhancement. Wiley.

Virtanen, T., Singh, R., Raj, B. (Eds.), 2012. Techniques for Noise Robustness in Automatic Speech Recognition. Wiley.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., Schuller, B., 2015. 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). pp. 91–99.

Wölfel, M., McDonough, J., 2009. Distant Speech Recognition. Wiley.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G., 2016. Achieving human parity in conversational speech recognition. CoRR abs/1610.05256.

Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S., Nakatani, T., 2015. The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In: ASRU. IEEE, pp. 436–443.

Yu, D., Deng, L., 2014. Automatic Speech Recognition - A Deep Learning Approach. Springer.

Zwyssig, E., Ravanelli, M., Svaizer, P., Omologo, M., 2015. A multi-channel corpus for distant-speech interaction in presence of known interferences. In: Proc. 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 4480–4484.