MICBOTS: COLLECTING LARGE REALISTIC DATASETS FOR SPEECH AND AUDIO RESEARCH USING MOBILE ROBOTS

Jonathan Le Roux
MERL
Cambridge, MA, USA

Emmanuel Vincent
Inria
Nancy, France

John R. Hershey
MERL
Cambridge, MA, USA

Daniel P.W. Ellis
Columbia University
New York, NY, USA

ABSTRACT

Speech and audio signal processing research is a tale of data collection efforts and evaluation campaigns. Large benchmark datasets for automatic speech recognition (ASR) have been instrumental in the advancement of speech recognition technologies. However, when it comes to robust ASR, source separation, and localization, especially using microphone arrays, the perfect dataset is out of reach, and many different data collection efforts have each made different compromises between the conflicting factors in terms of realism, ground truth, and costs. Our goal here is to escape some of the most difficult trade-offs by proposing MICbots, a low-cost method of collecting large amounts of realistic data where annotations and ground truth are readily available. Our key idea is to use freely moving robots equiped with microphones and loudspeakers, playing recorded utterances from existing (already annotated) speech datasets. We give an overview of previous data collection efforts and the trade-offs they make, and describe the benefits of using our robot-based approach. We finally explain the use of this method to collect room impulse response measurement.

Index Terms— Mobile robots, resources, robust ASR, source separation, room acoustics

1. INTRODUCTION

Over the years, many datasets have been created for robust speech processing research [1]. However, they are typically designed with a focus on either automatic speech recognition (ASR) or source separation/localization, but not all three tasks at the same time. In particular, there exists no dataset of real recordings, that simultaneously provides the ground truths for the speech signals, the speaker location, and the uttered words, in scenarios with overlapping speech and/or strongly non-stationary interference, and recorded with multiple microphones. There is a strong need for such real datasets, in order to integrate ASR, separation, and localization, and to separately measure performance on the three tasks.

ASR, source localization, and speech separation generally present different problems for data collection. For ASR, it is typically difficult and costly to record and later annotate large amounts of data in a wide range of environments. For speech localization, ground truth location can be obtained, but it requires a calibrated tracking apparatus for natural head movements. For speech separation, ground truth speech signals are needed to measure performance, as well as to use discriminative training methods. The presence of interference makes these ground truth signals impossible to obtain in normal recordings: for example, in a “cocktail party” scenario, the signal-to-noise ratio (SNR) of a close-talking microphone might be under 10 dB, whereas 30 dB is desirable.

To provide ground truth for the speech signal in a multi-microphone corpus, one can resort to acoustic emulation techniques including both numerical simulation and re-recording of existing data. Simulation ranges from instantaneous mixing of different microphone array recordings, to simulation of the whole reverberation effect by applying estimated room impulse responses (RIRs) to single-channel recordings. Re-recording, by playing single-channel recordings through loudspeakers, allows the mixing and reverberation to be real, but makes realistic motion and speech radiation patterns difficult to achieve.

Each of these methods may be realistic to different degrees. Here we distinguish between two different notions of realism. The data can be “acoustically realistic”, to a certain extent so that the signals reflect realistic acoustic effects such as mixing in the microphone, reverberant signal propagation, and potentially source motion and time-varying source radiation patterns, changes in the environment due to moving bodies, avoidance of artifacts and so on. Methods that work well when trained and tested on acoustically realistic data can be reasonably expected to work well when re-trained and tested on real recordings with the same acoustic properties. Data that is “ecologically realistic” would additionally replicate the non-acoustic properties of the target application. For human speakers, this would include natural head motion and speech activity patterns, variety of pronunciation and environmental conditions, Lombard effect, and so on. With such levels of realism, one could reasonably expect that a model trained on the approximate data would generalize well to real test data.

We propose to use mobile robots equipped with microphones and loudspeakers, which we call MICbots, to produce large re-recorded datasets that can help answer many of the ques-

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

tions raised above at low cost. We believe MICbots can be used to create datasets of real speech recordings with the three types of ground truths, with greater acoustic realism than ever achieved before. They can also be used to collect RIRs in many locations and conditions, in particular allowing one to test the validity of acoustic simulations, and investigate the influence of various acoustic factors such as radiation patterns and the movement of bodies.

In the rest of the paper, we overview the trade-offs in existing speech processing and RIR datasets and discuss our proposed methodology and its advantages and disadvantages.

2. DATA COLLECTION & ACOUSTIC EMULATION

Speech data collection is done with a variety of scenarios in mind. Those that are collected primarily for recognition can use real acoustic recordings and cope with a close-talking microphone as the ground-truth. In contrast, for speech enhancement/separation, a close-talking microphone is inadequate, and priority is placed on obtaining clean speech as the ground truth, using acoustic emulation at the expense of absolute acoustic realism. There is considerable debate about how realistic various aspects of a recording need to be in order for research on a dataset to be applicable to real recordings. Here we discuss different aspects of a recording that affect its acoustic realism, especially in the context of acoustic emulation. 1) The motion of the sources or sensors may cause difficulties for both methods. In simulation, source movements are approximated by interpolating estimated RIRs, which may introduce some small errors (e.g., approximately -19dB for interpolating between 2-cm laterally displaced RIRs [30]). 2) Loudspeaker distortion, which occurs in both re-recording and when estimating RIRs for simulation. In typical use, loudspeakers and microphones can add a small distortion and noise (1% total harmonic distortion plus noise (THD+N) [35] is equivalent to -40dB). 3) The radiation pattern of human speech depends both on head orientation, and upon the distribution of energy between nose and mouth as a function of the phonemes during speech. The importance of this effect is unknown. 4) Other speakers or objects moving in the environment can also have an impact on re-recording and upon the estimation of RIRs. This effect may be significant if they cross the line between the target speaker and the microphones. 5) Thermal fluctuation, air movement, room vibration, and other physical effects may have various impacts, some of which are analyzed in [36]. However these are smaller than those of head movements [30].

In terms of ecological realism, factors which may play an important role in acoustic emulation include realism of the head movements and the Lombard effect, in which speakers compensate for noise by modulating their voices. Both realistic head movements and Lombard effect may be difficult to produce using acoustic emulation. Although Lombard speech simulators have been proposed, e.g., in [37], it is not clear how realistic these are. Recording with natural head movement and Lombard effect are both relatively easy to do (using headphones in the case of Lombard to induce the effect without corrupting the recordings). So Lombard speech can be used with both methods, at the expense of performing recordings with human subjects and transcribing the data. Other factors that pertain to the ecological realism regardless of the acoustics, are the variety of reverberant and noisy environments, voice characteristics, pronunciations, vocabulary, spontaneity of speech, speech activity patterns, and so on.

2.1. Robust ASR and audio source separation

There are many existing datasets intended for robust ASR, and audio source separation. We highlight here some of their relevant characteristics. We focus on three main directions: size, realism, and availability of ground truth. Table 1 shows a qualitative assessment of the appropriateness of each dataset for a thorough evaluation of robust speech processing algorithms in realistic environments. This assessment is given for a set of key attributes in each direction, and boiled down to three levels: good (✓), intermediate (∼), bad (✗). Their meaning is the obvious one: for size attributes, the more the better; for realism attributes, the more realistic the better; for ground truth attributes, the more complete and the closer to the original the better. Thresholds are set to distribute the markers roughly uniformly whenever possible. We also give a crude estimate of the production cost, including data collection and annotation, mainly based on the size of the dataset and the amount and type of annotations. See [1] for more complete information and the exact value of each attribute.

There are many scenarios including little to no interference. Single-speaker reverberated speech datasets (TED, REVERB) tend to lack either signal ground truth or channel realism. Overlapping speech datasets (CUAVE, PASCAL SSC, SiSEC) tend to be unrealistic, and small. Some lack word ground truth. Broadcast datasets (GALE, ETAPE) tend to have few microphones and lack signal ground truth. Meeting/dialog datasets (ShATR, RWCP Meet, AV 16.3, NIST Meet, ICSI Meet, CHIL, AMI) tend to be rich but also costly to create; they are typically recorded in specially equipped rooms, making it difficult to record in different environments and virtually impossible to record large amounts of data.

Datasets with significant amounts of noise are hard to record properly, and tend to be quite limited. Those that use additive noise can be large, but the mixing of speech and noise is simulated, and scenarios are limited to commands or read speech (Aurora-2, CENSREC-1). Those that involve real noisy recordings have only a few environments (e.g., car) and a few command scenarios (Aurora-3, CU-Move, SPEECON); some use spontaneous speech, but tend to be small (CENSREC-4 Real, COSINE).

2.2. Room impulse response data

We similarly analyze existing RIR datasets, as summarized in Table 2. Most datasets suffer from a limited number of rooms

Table 1. *Comparison of various robust speech processing datasets*

		Size							Realism							Ground truth				
	cost	duration	# environments	# mics	# cameras	# speakers	# languages	vocab size	speaker style	speaker overlap	channel/reverb	speaker rad.	move betw. utt.	move during utt.	backgr. noise	speech signal	speaker pos.	words	non-verbal	noise events
ShATR [2]		✗	✗	∼	✗	✗	✗	∼	✓	✓	✓	✓	∼	✓	✓	∼	✓	✓	✗	✓
LLSEC [1]		✗	∼	✓	✗	∼	✗	✗	✓	✓	✓	✓	∼	✓	∼	✗	✓	✗	✗	✗
RWCP Dialog [3]	¢	∼	✗	∼	✗	∼	✗	∼	✓	✓	✓	✓	∼	✓	∼	✗	✗	✓	✗	✗
Aurora-2 [4]		∼	✓	✗	✗	✓	✗	✗	✗		∼	✓		✗	∼	✓		✓	✗	✓
SPINE [5]	$	∼	✗	∼	✗	✓	✗	∼	✓		∼	✓	∼	✓	∼	✗	✗	✓	✗	✗
Aurora-3 [2]	¢	∼	✗	✓	✗	✓	✓	✗	✗		✓	✓	∼	✓	✓	∼	✗	✓	✗	✗
RWCP Meet [6]	¢	✗	✗	✗	✓	∼	✗	∼	✓	✓	✓	✓	∼	✓	∼	∼	✗	✓	✗	✗
RWCP Real [7]		✗	✓	✓	✗	✗	∼	✗	∼		✓	✗	✓	∼	∼	✓	✓	✓	✗	✓
SpeechDat-Car [8]	$	✓	✗	✓	✗	✓	✓	∼	✓		✓	✓	∼	✓	✓	∼	✗	✓	✗	✗
Aurora-4 [2]		∼	✓	✗	✗	✓	✗	✓	∼		∼	✓		✗	∼	✓		✓	✗	✓
TED [9]	¢	∼	✗	✗	✗	✓	✗	∼	∼		✓	✓	∼	✓	✓	∼	✗	∼	✗	✗
CUAVE [10]		✗	✗	✗	∼	∼	✗	✗	✗	✗	✓	✓	∼	✓	∼	✗	✗	✓	✗	✗
CU-Move [11]	$	✓	✗	✓	✗	✓	✗	✓	✓		✓	✓	∼	✓	✓	✗	✗	✓	✗	✗
CENSREC-1 [12]		∼	✓	✗	✗	✓	✗	✗	✗		∼	✓		✗	∼	✓		✓	✗	✓
AVICAR [13]	¢	∼	✗	✓	✓	∼	✗	∼	∼		✓	✓	∼	✓	✓	✗	✗	✓	✗	✗
AV16.3 [14]		✗	✗	✓	✓	∼	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗	∼	✗	✗	✗
ICSI Meet [15]	$	✓	✗	✓	✗	∼	✗	✓	✓	✓	✓	✓	∼	✓	✓	∼	✗	✓	✓	∼
NIST Meet [16]	$	∼	✗	✓	✗	∼	✗	∼	✓	✓	✓	✓	✓	✓	✓	∼	✗	✓	✗	✗
CHIL [17]	$	✓	✗	✓	✓	∼	✗	✓	✓	✓	✓	✓	∼	✓	∼	∼	✓	✓	✓	✗
SPEECON [18]	$	✓	✓	∼	✗	✓	✓	∼	✓		✓	✓	∼	✓	✓	∼	✗	✓	✗	✗
CENSREC-2 [19]	¢	∼	✗	✗	✗	✓	✗	✗	✗		✓	✓	∼	✓	✓	∼	✗	✓	✗	✗
CENSREC-3 [20]	¢	∼	✗	✗	✗	✓	✗	✗	∼		✓	✓	∼	✓	✓	∼	✗	✓	✗	✗
Aurora-5 [2]		✓	✓	✗	✗	✓	✗	✗	✗		∼	✗	✗	✗	∼	✓	✗	✓	✗	✓
AMI [21]	$	✓	✗	✓	✓	✓	✗	∼	✓	✓	✓	✓	∼	✓	✓	∼	✓	✓	✓	✗
PASCAL SSC [22]		∼	✗	✗	✗	∼	✗	✗	✗	✗	✗	✓		✗	✗	✓		✓	✗	✗
HIWIRE [3]		∼	✗	✗	✗	∼	✗	✗	✗		✗	✓		✗	✓	✓		✓	✗	✗
NOIZEUS [23]		✗	✓	✗	✗	✗	✗	✗	∼		✗	✓		✗	∼	✓		✗	✗	✗
UT-Drive [24]	$	∼	✗	✓	✓	∼	✗	∼	✓	✓	✓	✓	∼	✓	✓	∼	✗	∼	✗	✗
SiSEC under [25]		✗	✗	∼	✗	∼	∼	✗	∼	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗
MC-WSJ-AV [26]	¢	∼	✗	✓	✗	∼	✗	✓	∼	✗	✓	✓	✓	✓	∼	∼	✓	✓	✗	✗
CENSREC-4 [27]		✗	✓	✗	✗	∼	✗	✗	✗		✓	✓	∼	✓	✓	∼	✗	✓	✗	✓
DICIT [28]	¢	∼	✗	✓	✓	∼	✗	✗	✗		✓	✓	✓	✓	∼	∼	✓	✓	✗	✓
SiSEC head [25]		✗	✗	∼	✗	✗	✗	✗	∼	✗	∼	✗	✓	✗	✗	✓	✓	✗	✗	✗
COSINE [29]	$	∼	✓	✓	✗	∼	✗	∼	✓	✓	✓	✓	✓	✓	✓	∼	✗	✓	✗	✗
SiSEC noise [25]		✗	✓	✓	✗	✗	✗	✗	∼	✗	∼	✗	✓	✗	∼	✓	✓	✗	✗	✗
SiSEC dynam [25]		✗	✗	✓	✗	✗	✗	✗	∼	✗	✓	✗	✓	∼	✗	✓	✓	✗	✗	✗
CHiME Grid [30]	¢	✓	✗	∼	✗	∼	✗	✗	✗	∼	∼	∼	∼	∼	✓	✓	✓	✓	✗	✗
CHiME WSJ0 [30]	¢	✓	✗	∼	✗	✓	✗	✓	∼	∼	∼	∼	✗	✗	✓	✓	✓	✓	✗	✗
ETAPE [31]	$	∼	∼	✗	∼	✓	✗	✓	✓	✓	✓	✓	∼	✓	✓	✗	✗	✓	✗	✓
GALE [4]	$	✓	✗	✗	✗	✓	∼	✓	✓	✓	✓	✓	∼	✓	✓	✗	✗	✓	✗	✗
REVERB Sim [32]		∼	✗	✓	✗	✓	✗	✓	∼		∼	✗	✓	✗	✗	✓	✓	✓	✗	✓
SWC [33]	¢	✗	✗	✓	✓	✗	✗	∼	✓	✓	✓	✓	✓	✓	✓	∼	✓	✓	✗	✗
DIRHA [34]	¢	∼	✗	✓	✗	∼	∼	∼	✓	∼	∼	✗	✓	✗	✓	✓	✓	✓	✗	✓

and, except for CAMIL, microphone locations. No dataset considers many locations for both microphones and speakers. Speaker radiation emulation using a stationary mannequin head is done in CHiME 2 Grid. Speaker movements are only considered in RWCP RE, and microphone movements in CAMIL. There is clearly a gap to be filled in terms of a large dataset involving many environments and rooms, with a large number of microphone and speaker locations. The only dataset that comes close in terms of size in CAMIL, with slightly more than 30k RIRs. It was made use of a robotic head, which greatly accelerated data collection time.

3. MICBOTS

3.1. Concept and advantages

We have seen that there are few large realistic datasets for robust speech processing, given the cost of recording and an-

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

[2] http://aurora.hsnr.de/

[3] http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/

[4] https://catalog.ldc.upenn.edu/search?q[name_cont]=GALE

Table 2. *Comparison of various RIR datasets*

		Size					Realism			
	cost	# rooms	# IRs	# mics	# speaker locs	# mic locs	channel type	speaker rad.	speaker move	mic moves
RWCP Real Env. [7]	¢	∼	∼	✓	✗	✗	✓	∼	✓	✗
SASSEC, SiSEC und. [25]		∼	✗	∼	∼	✗	✓	✗	✗	✗
SiSEC head [25]		✗	✗	∼	∼	✗	✓	✗	✗	✗
Aachen Imp. Resp. [38]	¢	∼	∼	∼	∼	✗	✓	✗	✗	✗
CAMIL [39]		✗	✓	∼	✗	✓	✓	✗	✗	✓
CHiME 2 Grid [30]		✗	∼	∼	✓	✗	✓	∼	✗	✗
AVASM [40]		✗	∼	∼	✓	✗	✓	✗	✗	✗
DIRHA [34]	¢	∼	✓	✓	∼	✗	✓	✗	✗	✗

notation, and the difficulties with ground truth. To integrate ASR, separation and localization, large acoustically realistic datasets must be recorded in various environments and featuring word, signal, and location ground truth.

Our proposal is to use freely-moving robots equipped with microphones and loudspeakers to re-record human data. This has several key advantages. It is a very low cost solution: existing annotated clean speech or conversational speech datasets can be used as speech signals, bypassing the need for primary collection and transcription. The main cost is to buy and equip the robots, but off-the-shelf equipment can be used. The re-recording of the data is fully automatic. The ground truth recordings of clean signals can be easily provided. The acoustic realism will include real room acoustics, complete with moving sources, microphones, and bodies.

The disadvantages in terms of acoustic realism would include generating radiation patterns of a similar complexity to real speech. The robots themselves will also produce noise during movements, which can be seen as either a curse or a blessing in terms of noise robustness. Ecological realism in terms of human speech will also be difficult to produce without significant effort. However, the acoustic realism will already be far better than any corpus of overlapping speech recorded to date with ground truth speech signals, and this may be enough to inspire useful research. In addition if the target application is robot speech, then the ecological realism is very good.

The robot platform we plan to use is Kobuki[5], which is similar to the popular Roomba vacuum cleaners. The robot will be fitted with a platform to hold a laptop, a Kinect-like camera, a microphone array, and a loudspeaker. We are also envisioning introducing head-like movements by mounting the robot's loudspeaker on a turntable.

3.2. Challenges

Localization also poses some technological challenges. "Simultaneous localization and mapping" (SLAM) [41] using laser sensors may be a reasonable solution. Its relative lack of precision in unknown environments should not be an issue in our scenarios, where we can build a map prior to the experiment. In such a case, one case expect to estimate position with an accuracy of the order of 3 to 10 cm [42].

Reduction of mechanical noise is potentially a goal, with solutions ranging from insulation to improved actuators. Preliminary experiments with a Kobuki robot showed however that the noise from the robot's wheels during slow movements was limited, and qualitatively comparable to noise from the air conditioning.

Many potential variations can be envisioned in the design of the setup and are left to be investigated: what positions to use for the robots with respect to each other; what positions in the room; allowing movements during the utterances or only between; allowing "head" motion in addition to "body" motion. recording multiple takes in the same setting to investigate the influence of noise or temperature variations; attempting to use Lombard speech as a function of the noise level; scheduling the timing of utterances by each robot to reproduce realistic speech overlap patterns.

3.3. Example recording protocols

We first plan to consider a cocktail party scenario, in which multiple robots are used to play (and record) multiple concurrent speech tracks from the WSJ0 clean speech corpus. We plan to let four to five robots move freely, each within its own section of a room separated from the others by a physical boundary on the floor, and play random utterances, each from a separate subset of speakers in the the WSJ0 corpus, with pauses of about the same length as the utterance. Microphone arrays will be mounted on each robot as well as at multiple locations in the room. The location of the robots will be inferred from fixed cameras in the room, as well as by SLAM from the laser sensors. Although WSJ0 is not spontaneous speech, the ecological validity can be addressed at a later stage.

Regarding RIRs, we plan to record an order of magnitude more than in the largest existing dataset, CAMIL. By discretizing a room of size 3 m by 4 m in 10 cm cells, we obtain 1131 cells in the horizontal plane, which would already amount to $1.3 \cdot 10^6$ RIRs. For this task a different kind of track-based robot must be used for precise positioning. One goal will be to test the acoustic realism of interpolation used in simulation of source movements using RIRs.

4. CONCLUSION

We presented MICbots, a method of collecting large amounts of realistic noisy speech recordings with rich ground truth at low cost. The method uses mobile robots to re-record existing clean speech datasets in noisy or multi-speaker environments. It can also be used to record RIRs, in order to investigate the validity of simulation-based datasets.

[5] http://kobuki.yujinrobot.com/

5. REFERENCES

[1] J. Le Roux and E. Vincent, “A categorization of robust speech processing datasets,” Mitsubishi Electric Research Labs, Tech. Rep. TR2014-116, Aug. 2014.

[2] M. D. Crawford, G. J. Brown, M. P. Cooke, and P. D. Green, “Design, collection and analysis of a multi-simultaneous-speaker corpus,” *Proceedings of the IOA*, vol. 16, no. 5, 1994.

[3] K. Tanaka, S. Hayamizu, Y. Yamashita, K. Shikano, S. Itahashi, and R. Oka, “Design and data collection for a spoken dialog database in the Real World Computing (RWC) program,” *J. Acoust. Soc. Am.*, vol. 100, 1996.

[4] H.-G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. ASR2000*, 2000.

[5] T. H. Crystal, A. Schmidt-Nielsen, and E. Marsh, “Speech in noisy environments (SPINE) adds new dimension to speech recognition R&D,” in *Proc. HLT*, 2002.

[6] K. Tanaka, K. Itou, M. Ihara, and R. Oka, “Constructing a meeting speech corpus,” *IPSJ Tech. Rep.*, 2001.

[7] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” in *Proc. LREC*, 2000.

[8] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri *et al.*, “SPEECHDAT-CAR. a large speech database for automotive environments,” in *Proc. LREC*, 2000.

[9] L. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillman, “The translingual English database (TED),” in *Proc. ICSLP*, 1994.

[10] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, “Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus,” *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 208541, 2002.

[11] 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.

[12] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa *et al.*, “Aurora-2J, an evaluation framework for Japanese noisy speech recognition,” *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 3, 2005.

[13] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys *et al.*, “Avicar: audio-visual speech corpus in a car environment.” in *Proc. Interspeech*, 2004.

[14] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, “AV16.3: an audio-visual corpus for speaker localization and tracking,” in *Proc. MLMI*, 2004.

[15] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart *et al.*, “The ICSI meeting corpus,” in *Proc. ICASSP*, 2003.

[16] J. S. Garofolo, C. D. Laprun, M. Michel, V. M. Stanford, and E. Tabassi, “The NIST meeting room pilot corpus,” in *Proc. LREC*, 2004.

[17] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu *et al.*, “The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms,” *Lang. Resour. Eval.*, vol. 41, no. 3–4, 2007.

[18] D. J. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, “SPEECON - speech databases for consumer devices: Database specification and validation.” in *Proc. LREC*, 2002.

[19] S. Nakamura, M. Fujimoto, and K. Takeda, “CENSREC2: Corpus and evaluation environments for in car continuous digit speech recognition,” in *Proc. Interspeech*, 2006.

[20] M. Fujimoto, K. Takeda, and S. Nakamura, “CENSREC-3: An evaluation framework for Japanese speech recognition in real driving-car environments,” *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 11, 2006.

[21] S. Renals, T. Hain, and H. Bourlard, “Interpretation of multiparty meetings: The AMI and AMIDA projects,” in *Proc. HSCMA*, 2008.

[22] M. P. Cooke, J. R. Hershey, and S. J. Rennie, “Monaural speech separation and recognition challenge,” *Comput. Speech Lang.*, vol. 24, no. 1, 2010.

[23] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech Commun.*, vol. 49, no. 7–8, 2007.

[24] P. Angkititrakul, J. H. L. Hansen, S. Choi, T. Creek, J. Hayes *et al.*, “UTDrive: The smart vehicle project,” in *In-vehicle corpus and signal processing for driver behavior*. Springer, 2009.

[25] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill *et al.*, “The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges,” *Signal Process.*, vol. 92, 2012.

[26] 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. ASRU*, 2005.

[27] T. Nishiura, M. Nakayama, Y. Denda, N. Kitaoka, K. Yamamoto *et al.*, “Evaluation framework for distant-talking speech recognition under reverberant environments — newest part of the CENSREC series —,” in *Proc. LREC*, 2008.

[28] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo, “WOZ acoustic data collection for interactive TV,” in *Proc. LREC*, 2008.

[29] 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,” *Comput. Speech Lang.*, vol. 26, no. 1, 2011.

[30] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, “The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines,” in *Proc. ICASSP*, 2013.

[31] 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. LREC*, 2012.

[32] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets *et al.*, “The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *Proc. WASPAA*, 2013.

[33] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, “The Sheffield wargames corpus,” in *Proc. Interspeech*, 2013.

[34] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad *et al.*, “The DIRHA simulated corpus,” in *Proc. LREC*, 2014.

[35] W. M. Hartmann, *Signals, sound, and sensation*. Springer, 1997.

[36] G. W. Elko, E. Diethorn, and T. Gänsler, “Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation,” in *Proc. IWAENC*, 2003.

[37] D.-Y. Huang, S. Rahardja, and E. P. Ong, “Lombard effect mimicking,” in *Proc. SSW*, 2010.

[38] M. Jeub, M. Schafer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proc. DSP*, 2009.

[39] A. Deleforge, F. Forbes, and R. Horaud, “Variational EM for binaural sound-source separation and localization,” in *Proc. ICASSP*, 2013.

[40] A. Deleforge, V. Drouard, L. Girin, R. Horaud *et al.*, “Mapping sounds on images using binaural spectrograms,” in *Proc. EUSIPCO*, 2014.

[41] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*. MIT press, 2005.

[42] Hokuyo, “Scanning range finder (SOKUIKI sensor),” https://www.hokuyo-aut.jp/02sensor/07scanner/urg_04lx_ug01.html, [Online; accessed 19-July-2008].