A categorization of robust speech processing datasets

Jonathan Le Roux, Emmanuel Vincent

MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

A categorization of robust speech processing datasets

Le Roux, J.; Vincent, E.

TR2014-116 August 2014

Abstract

Speech and audio signal processing research is a tale of data collection efforts and evaluation campaigns. While large datasets for automatic speech recognition (ASR) in clean environments with various speaking styles are available, the landscape is not as picture- perfect when it comes to robust ASR in realistic environments, much less so for evaluation of source separation and speech enhancement methods. Many data collection efforts have been conducted, moving along towards more and more realistic conditions, each mak- ing different compromises between mostly antagonistic factors: financial and human cost; amount of collected data; availability and quality of annotations and ground truth; natural- ness of mixing conditions; naturalness of speech content and speaking style; naturalness of the background noise; etc. In order to better understand what directions need to be explored to build datasets that best support the development and evaluation of algorithms for recognition, separation or localization that can be used in real-world applications, we present here a study of existing datasets in terms of their key attributes.

Mitsubishi Electric Research Labs

A categorization of robust speech processing datasets

Jonathan Le Roux
Mitsubishi Electric Research Labs (MERL)
Cambridge, MA, USA
leroux@merl.com

Emmanuel Vincent
INRIA
Nancy, France
emmanuel.vincent@inria.fr

Abstract

Speech and audio signal processing research is a tale of data collection efforts and evaluation campaigns. While large datasets for automatic speech recognition (ASR) in clean environments with various speaking styles are available, the landscape is not as picture-perfect when it comes to robust ASR in realistic environments, much less so for evaluation of source separation and speech enhancement methods. Many data collection efforts have been conducted, moving along towards more and more realistic conditions, each making different compromises between mostly antagonistic factors: financial and human cost; amount of collected data; availability and quality of annotations and ground truth; naturalness of mixing conditions; naturalness of speech content and speaking style; naturalness of the background noise; etc. In order to better understand what directions need to be explored to build datasets that best support the development and evaluation of algorithms for recognition, separation or localization that can be used in real-world applications, we present here a study of existing datasets in terms of their key attributes.

Version: This is version **v2014-09** of the report, last updated on September 05, 2014. The original version of the report was published in August 2014. Please cite it as:

```
@TechReport{LeRouxVincent2014TRdatasets,
author =       {Le Roux, Jonathan and Vincent, Emmanuel},
title =        {A categorization of robust speech processing datasets},
institution =  {Mitsubishi Electric Research Labs},
year =         {2014},
number =    {TR2014-116},
address =   {Cambridge, MA, USA},
month =     aug,
note =      {v2014-09}
}
```

1 List of robust speech processing datasets

This technical report aims to provide a list of speech datasets with detailed attributes and links to software baselines and evaluation results. Each dataset may be used for one or more applications: automatic speech

recognition, speaker identification and verification, source localization, speech enhancement and separation...

The core of the report is Table 1, which compares a wide range of datasets for robust speech processing based on their key attributes[1]. A list of links relevant to each dataset (to download/purchase, obtain baselines and results, etc.) is provided separately in Table 2.

Only datasets that are publicly available, (at least partially) annotated, suitable for research on robustness, and longer than 5 min are listed. Other relevant datasets are listed in Section 2.

The list of considered attributes and their meaning is detailed below.

General attributes

- year of release
- scenario: car, cocktail party, domestic, lecture, meeting, office, public space, TV...
- total duration (h) (multiple channels counted only once)
- sampling rate (kHz)
- number of distant or noisy microphones
- number of video cameras
- cost for non-members of ELRA and LDC (cost for members is lower or free)
- links: download data, reference papers, software baselines, evaluation results...

Speech attributes

- duration of speech (h) (overlapping speech counted only once)
- number of unique speakers
- language
- number of unique words (differs from assumed vocabulary size, which is somewhat arbitrary)
- speaking style: digits, command, read, spontaneous...
- number of speakers present in the room
- type of speaker overlap: no overlap, simulated overlap, dialogue, meeting, full overlap...

Channel attributes

- channel type: none, simulated room impulse response, convolution by a recorded room impulse response, reverberant recording...
- speaker radiation: loudspeaker, dummy head with mouth simulator, human...
- speaker location: at a fixed position in the room, at a quasi-fixed position (e.g., seated), at different positions...
- speaker movements: no movement, head movements, walking...

[1] A maintained version of this table is available on the wiki of ISCA's Special Interest Group (SIG) on Robust Speech Processing: `https://wiki.inria.fr/rosp/`. This report is frequently updated to reflect changes on the wiki.

Noise attributes

- noise type: stationary background noise (e.g., air-conditioning), car noise, meeting noises, domestic noises, outdoor noises...

Available ground truth

- reference speech signal: original (at the mouth), headset or lapel (slightly differs from the signal at the mouth), spatial image (at the microphones)...
- speaker location and orientation
- words uttered
- paralinguistic attributes: nodding, gaze, communication intent, emotion...
- noise events: type and time of individual noise events

Table 1: Comparison of robust speech processing datasets

	General attributes						Speech							Channel				Noise		Ground truth				
Datasets	rel. year	use case	total time (h)	sam. rate (kHz)	dist. or noisy mics	video cams	cost (non- memb)	speak. time (h)	uniq. speak.	lang.	uniq. words (k)	speak. style	speak. / rec.	overl. type	chan. type	speak. radiat.	speak. loc.	speak. moves	noise type	avg. SNR	ref. signal	speak. loc., orient.	words	non- verb. traits	noise events
ShATR [1]	1994	meeting	0.6	48	3	no	free	0.6	5	UK English	1	spontaneous	5	multiple dialogs	reverb	human	quasi-fixed	head	meeting	high	headset	yes	yes	no	yes
LLSEC	1996	dialog	1.4	16	4	no	free	?	12	N/S	N/S	read, spontaneous	2	dialog	reverb	human	quasi-fixed	head	hallway, restaurant (scenarized)	medium	no	yes	no	no	no
RWCP Spoken Dialog Corpus [2]	1996 - 1997	dialog	10	16	2	no	free	10	39	Japanese	?	spontaneous	1 - 2	dialog	reverb (low)	human	quasi-fixed	head	stationary background	high	no	no	yes	no	no
Aurora-2 [3]	2000	public spaces	33	8 - 16	1	no	free given TIDigits (0.5 k$)	33	214	US English	0.01	digits	1	no	simulated phone	human	N/S	no	various real environments (added)	low	original	N/S	yes	no	yes
SPINE1, SPINE2 [4]	2000 - 2001	military	38	16	2	no	7.4 k$	?	100	US English	1	command, spontaneous	1 - 2	no	simulated radio	human	quasi-fixed	head	military (added)	low	no	no	yes	no	no
Aurora-3 (subset of SpeechDat- Car) [5]	2000 - 2003	car	?	16	4	no	1 k	?	730	various	0.01	digits	1	no	reverb	human	quasi-fixed	head	car	low	headset	no	yes	no	no
RWCP Meeting Speech Corpus [6]	2001	meeting	3.5	16 - 48	1	3	free	3.5	?	Japanese	?	spontaneous	1 - 5	meeting	reverb (low)	human	quasi-fixed	head	stationary background	high	headset	no	yes	no	no
RWCP Real Environment Speech and Acoustic Database [7]	2001	domestic, office	?	16 - 48	84	no	free	?	5	US English, Japanese	?	read	1	no	real rir, reverb	loudspeaker	various	no, pivoting arm	various (sum of events)	medium	original	yes	yes	no	yes
SpeechDat- Car [8]	2001 - 2011	car	?	16	4	no	39 - 182 k per lang	?	300 per lang	various	?	digits, command, read, spontaneous	1	no	reverb	human	quasi-fixed	head	car	low	headset	no	yes	no	no
Aurora-4 [9]	2002	public spaces	?	8 - 16	1	no	free given WSJ0 (1.5 k$)	?	101	US English	10	read	1	no	simulated phone	human	N/S	no	various real environments (added)	low	original	N/S	yes	no	yes
TED [10]	2002	seminar	47	16	1	no	0.5 k$	47	188	non-native English	?	lecture	1 or more	seminar	reverb	human	quasi-fixed	head	stationary background	high	lapel	no	partial	no	no
CUAVE [11]	2002	speech overlap	3	44	1	1	free	3	36	US English	0.01	digits	1 - 2	full	reverb	human	quasi-fixed	head	stationary background	high	no	no	yes	no	no
CU-Move Microphone Array Data [12]	2002 - 2011	car	286	44	6 - 8	no	25 k$	286	172	US English	12	digits, command, read, dialog	1	no	reverb	human	quasi-fixed	head	car	low	no	no	yes	no	no
CENSREC-1 (Aurora-2J) [13]	2003	public spaces	?	8	1	no	free	?	214	Japanese	0.01	digits	1	no	simulated phone	human	N/S	no	various real environments (added)	low	original	N/S	yes	no	yes
AVICAR [14]	2004	car	29	16	7	4	free	29	86	US English, non-native English	1	read	1	no	reverb	human	quasi-fixed	head	car	low	no	no	yes	no	no
AV16.3 [15]	2004	meeting	1.5	16	16	3	free	1.5	12	N/S	N/S	spontaneous	1 - 3	full	reverb	human	various	head, walk	stationary background	high	no	partial	no	no	no
ICSI Meeting Corpus [16]	2004	meeting	72	16	6	no	2.8 k$	72	53	US English, other English	13	meeting	3 - 10	meeting	reverb	human	quasi-fixed	head	meeting	high	headset, lapel	no	yes	yes	ad-hoc
NIST Meeting Pilot Corpus Speech [17]	2004	meeting	15	16	7	no	5.5 k$	15	61	US English	6	meeting	3 - 9	meeting	reverb	human	various	head, walk	stationary background	high	headset, lapel	no	yes	no	no
CHIL Meetings [18]	2004 - 2007	seminar, meeting	60	44	79 - 147	6 - 9	3.5 k	?	?	non-native English	?	seminar, meeting	3 - 20	seminar, meeting	reverb	human	quasi-fixed	head	meeting (scenarized)	high	headset	yes	yes	yes	no
SPEECON [19]	2004 - 2011	public space, domestic, office, car	?	16	3	no	75 k per lang	?	600 per lang	various	?	command, read, spontaneous	1	no	reverb	human	quasi-fixed	head	various real environments	medium	headset	no	yes	no	no
CENSREC-2 [20]	2005	car	?	16	1	no	free	?	214	Japanese	0.01	digits	1	no	reverb	human	quasi-fixed	head	car	low	headset	no	yes	no	no
CENSREC-3 [21]	2005	car	?	16	1	no	21 k	?	311	Japanese	0.05	read	1	no	reverb	human	quasi-fixed	head	car	low	headset	no	yes	no	no
Aurora-5 [22]	2006	public spaces, domestic, office, car	?	8	1	no	free given TIDigits (0.5 k$)	?	225	US English	0.01	digits	1	no	no, simulated rir, real rir	loudspeaker	fixed	no	various real environments (added)	low	original	no	yes	no	yes
AMI [23]	2006	meeting	100	16	16	6	free	?	189	UK English, other English	8	meeting	most often 4	meeting (18% overlap)	reverb	human	quasi-fixed	head	stationary background	high	headset, lapel	yes	yes	yes	no
PASCAL SSC [24]	2006	speech overlap	8.8	25	1	no	free	8.8	34	UK English	0.05	command	2	full	no	human	N/S	no	no	N/S	original	N/S	yes	no	no
HIWIRE [25]	2007	airplane	21	16	1	no	0.05 k	21	81	non-native English	0.1	command	1	no	no	human	N/S	no	airplane (added)	low	original	N/S	yes	no	no
NOIZEUS [26]	2007	public spaces	0.6	8	1	no	free	0.6	6	US English	0.1	read	1	no	simulated phone	human	N/S	no	various real environments (added)	low	original	N/S	no	no	no
UT-Drive [27]	2007	car	40	25	5	2	25 k$	40	25	US English	2.4	command, dialog	1 - 2	dialog	reverb	human	quasi-fixed	head	car	low	headset (low quality)	no	partial	no	no
SASSEC, SiSEC under- determined [28]	2007 - 2011	cocktail party	0.3	16	2	no	free	0.3	16	N/S	N/S	read	3 - 4	full	simulated rir, real rir, reverb	no, loudspeaker	fixed	no	no	N/S	original, spatial image	yes	no	no	no
MC-WSJ-AV, PASCAL SSC2, 2012..MMA, REVERB RealData [29] [30]	2007 - 2014	speech overlap	10	16	8 - 40	partial	1.5 k$	?	45	UK English	10	read	1 - 2	full	reverb	human	various	head, walk	stationary background	high	headset, lapel	yes	yes	no	no
CENSREC-4 (Simulated) [31]	2008	public spaces, domestic, office, car	?	16	1	no	free	?	214	Japanese	0.01	digits	1	no	real rir	dummy	fixed	no	various real environments (added)	low	original	no	yes	no	yes
CENSREC-4 (Real) [31]	2008	public spaces, domestic, office, car	?	16	1	no	free	?	10	Japanese	0.01	digits	1	no	reverb	human	quasi-fixed	head	various real environments	low	headset	no	yes	no	yes
DICIT [32]	2008	domestic	6	48	16	2	free	1	?	Italian	?	command	4	no	reverb	human	various	head, walk	domestic (scenarized)	medium	headset, tv	yes	yes	no	yes
SiSEC head-geometry [28]	2008	speech overlap	1.9	16	2	no	free	1.9	?	N/S	N/S	read	2	full	real rir	loudspeaker	various	no	no	N/S	original, spatial image	yes	no	no	no
COSINE [33]	2009	dialog	38	48	20	no	free	11	91	US English, non-native English	5	spontaneous	2 - 7	dialog	reverb	human	various	head, walk	various real environments	low	headset, throat mic	no	yes	no	no
SiSEC real-world noise [28]	2010	public spaces	0.3	16	2 - 4	no	free	0.3	6	N/S	N/S	read	1 - 3	full	no, reverb (other room)	loudspeaker	various	no	various real environments (added)	low	original, spatial image	yes	no	no	no
SiSEC dynamic [28]	2010 - 2011	cocktail party	0.2	16	2 - 4	no	free	0.2	?	N/S	N/S	read	?	full (2 at a time)	reverb	loudspeaker	various	simulated	no	N/S	original, spatial image	yes	no	no	no
CHiME 1, CHiME 2 Grid [34]	2011 - 2012	domestic	70	16 - 48	2	no	free	12	34	UK English	0.05	command	1	no	real rir	dummy	quasi-fixed	simulated head	domestic	low	yes	yes	yes	no	no
CHiME 2 WSJ0 [34]	2012	domestic	78	16	2	no	free given WSJ0 (1.5 k$)	33	101	US English	11	read	1	no	real rir	dummy	fixed	no	domestic	low	yes	yes	yes	no	no
ETAPE [35]	2012	TV/radio debates, outdoor interviews	42	16	1	1	?	32	347	French	16	spontaneous	1 or more	dialog (up to 10% overlap)	reverb (some)	human	quasi-fixed	head	various real environments	high	no	N/S	yes	no	yes
GALE	2013	TV dialog	120 - 251 per lang	16	1	no	3.5 - 7 k$ per lang	108 - 234 per lang	?	Mandarin, Arabic	?	spontaneous	1 or more	dialog	no	human	quasi-fixed	head	no	N/S	no	N/S	yes	no	no
REVERB SimData [36]	2013	domestic, office	25	16	8	no	free given WSJCAM0 (1.75 k$)	25	130	UK English	10	read	1	no	real rir	loudspeaker	various	no	random noise (added)	high	original, spatial image	yes	yes	no	yes
Sheffield Wargames Corpus [37]	2013	cocktail party	7	48	92	3	free	?	9	UK English	?	spontaneous	4	multiple dialogs	reverb	human	various	head, walk	background music	medium	headset	yes	yes	no	no
DIRHA [38]	2014	domestic	11	48	40	no	free (partial avail.)	4	90	various	3.8	command, read, spontaneous	1 or more	simulated	real rir	loudspeaker	various	no	domestic	low	yes	yes	yes	no	yes

Table 2: Miscellaneous links for each dataset

Datasets	Links
ShATR	download: http://spandh.dcs.shef.ac.uk/projects/shatrweb/
LLSEC	download: https://www.ll.mit.edu/mission/cybersec/HLT/corpora/SpeechCorpora.html
RWCP Spoken Dialog Corpus	download: http://research.nii.ac.jp/src/en/RWCP-SP96.html
Aurora-2	purchase (incl. HTK): http://catalog.elra.info/product_info.php?cPath=37_40&products_id=693 features: http://aurora.hsnr.de/download.html
SPINE1, SPINE2	purchase: https://catalog.ldc.upenn.edu/search?q%5Bname_cont%5D=SPINE
Aurora-3 (subset of SpeechDat- Car)	purchase (incl. HTK): http://catalog.elra.info/index.php?cPath=37_40
RWCP Meeting Speech Corpus	download: http://research.nii.ac.jp/src/en/RWCP-SP01.html
RWCP Real Environment Speech and Acoustic Database	download: http://research.nii.ac.jp/src/en/RWCP-SSD.html
SpeechDat- Car	purchase: http://catalog.elra.info/search.php
Aurora-4	purchase: http://catalog.elra.info/index.php?cPath=37_40 HTK: http://www.keithv.com/software/htk/
TED	purchase: https://catalog.ldc.upenn.edu/LDC2002S04
CUAVE	download: http://www.clemson.edu/ces/speech/cuave.htm
CU-Move Microphone Array Data	purchase: http://crss.utdallas.edu/
CENSREC-1 (Aurora-2J)	download: http://research.nii.ac.jp/src/en/CENSREC-1.html
AVICAR	download: http://www.isle.illinois.edu/sst/AVICAR/
AV16.3	download: http://www.idiap.ch/dataset/av16-3/
ICSI Meeting Corpus	purchase: https://catalog.ldc.upenn.edu/search?q%5Bname_cont%5D=ICSI info: http://www1.icsi.berkeley.edu/Speech/mr/
NIST Meeting Pilot Corpus Speech	purchase: https://catalog.ldc.upenn.edu/search?q%5Bname_cont%5D=NIST%20Meeting
CHIL Meetings	purchase: http://catalog.elra.info/search.php
SPEECON	purchase: http://catalog.elra.info/search.php
CENSREC-2	download: http://research.nii.ac.jp/src/en/CENSREC-2.html
CENSREC-3	purchase: http://research.nii.ac.jp/src/en/CENSREC-3.html

Aurora-5	purchase (incl. HTK): http://catalog.elra.info/product_info.php?cPath=37_40&products_id=1015
AMI	download: http://groups.inf.ed.ac.uk/ami/
PASCAL SSC	download: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm
HIWIRE	purchase: http://catalog.elra.info/product_info.php?products_id=1088&language=en
NOIZEUS	download: http://ecs.utdallas.edu/loizou/speech/noizeus/
UT-Drive	download: http://crss.utdallas.edu/
SASSEC, SiSEC under-determined	download: http://sisec2011.wiki.irisa.fr/tiki-index.php?page=Underdetermined+speech+and+music+mixtures
MC-WSJ-AV, PASCAL SSC2, 2012_MMA, REVERB RealData	purchase: https://catalog.ldc.upenn.edu/LDC2014S03 info: http://www.cstr.ed.ac.uk/corpora/2012_MMA/ video: http://scholar.google.co.uk/citations?view_op=view_citation&hl=en&user=8J_nG0wAAAAJ&citation_for_view=8J_nG0wAAAAJ:08ZZubdj9fEC HTK: http://reverb2014.dereverberation.com/tools/REVERB_TOOLS_FOR_ASR_ver2.0.tgz Kaldi: http://www.mmk.ei.tum.de/~wen/REVERB_2014/kaldi_baseline.tar.gz results: http://reverb2014.dereverberation.com/result_se.html results: http://reverb2014.dereverberation.com/result_asr.html
CENSREC-4 (Simulated)	download: http://research.nii.ac.jp/src/en/CENSREC-4.html
CENSREC-4 (Real)	download: http://research.nii.ac.jp/src/en/CENSREC-4.html
DICIT	download: http://shine.fbk.eu/resources/dicit-acoustic-woz-data
SiSEC head-geometry	download: http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Head-geometry%20mixtures%20of%20two%20speech%20sources%20in%20real%20environments,%20impinging%20from%20many%20directions
COSINE	download: http://melodi.ee.washington.edu/cosine/
SiSEC real-world noise	download: http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Source+separation+in+the+presence+of+real-world+background+noise
SiSEC dynamic	download: http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Determined+convolutive+mixtures+under+dynamic+conditions

CHiME 1, CHiME 2 Grid	download: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2_task1.html HTK: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2_task1.html#tools results: http://spandh.dcs.shef.ac.uk/projects/chime/PCC/results.html results: http://spandh.dcs.shef.ac.uk/chime_challenge/track1_results.html
CHiME 2 WSJ0	download: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2_task2.html HTK: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2_task2.html#tools Kaldi: http://spandh.dcs.shef.ac.uk/chime_challenge/WSJ0public/CHiME2012-WSJ0-Kaldi_0.03.tar.gz results: http://spandh.dcs.shef.ac.uk/chime_challenge/track2_results.html
ETAPE	download: http://www.afcp-parole.org/etape.html
GALE	purchase: https://catalog.ldc.upenn.edu/search?q%5Bname_cont%5D=GALE
REVERB SimData	purchase: http://reverb2014.dereverberation.com/ HTK: http://reverb2014.dereverberation.com/tools/REVERB_TOOLS_FOR_ASR_ver2.0.tgz Kaldi: http://www.mmk.ei.tum.de/~wen/REVERB_2014/kaldi_baseline.tar.gz results: http://reverb2014.dereverberation.com/result_se.html results: http://reverb2014.dereverberation.com/result_asr.html
Sheffield Wargames Corpus	download: http://mini.dcs.shef.ac.uk/data-2/
DIRHA	download: http://shine.fbk.eu/resources/dirha-ii-simulated-corpus

2 Other datasets

The following datasets were considered but not included in the table for the reasons described below:

- BABEL[2] (not yet available)
- Broadcast news, HUB4 [3] (no noise and 4.5 % speaker overlap, less than ETAPE)
- CIAIR In-Car Speech Database [39] (availability unknown)
- Dyrholm/Sawada/Parra[4] (about 1 min long)

[2] http://www.iarpa.gov/index.php/research-programs/babel
[3] https://catalog.ldc.upenn.edu/byproject
[4] http://bme.ccny.cuny.edu/faculty/parra/bss/

- NEMISIG [40] (unavailable)
- RATS[5] (not yet available)
- Rich Transcription (RT) (dataset gathered from other sets, e.g. CHIL, ICSI, ISL, AMI...)
- Settlers of Catan[6] [41] (unannotated)
- Flying MEMS microphone array[7] [41] (unannotated)

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Various, http://aurora.hsnr.de/aurora-3/reports.html, 1999–2001.

[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] N. Parihar and J. Picone, "DSR front-end large vocabulary continuous speech recognition evaluation," Mississippi State University, Tech. Rep., 2002.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[5] http://www.darpa.mil/Our_Work/I2O/Programs/Robust_Automatic_Transcription_of_Speech_%28RATS%29.aspx

[6] http://meetingdiarisation.wordpress.com/2013/05/09/ready-for-recording-settlers-of-cattan-with-the-dmma-2-and-dmma-3/

[7] http://meetingdiarisation.wordpress.com/2014/08/11/flying-digital-mems-microphone-array-dmma-3/

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

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

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

[18] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu *et al.*, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language Resources and Evaluation*, vol. 41, no. 3–4, 2007.

[19] 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.

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

[21] 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.

[22] H.G.Hirsch, "Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments," Niederrhein University of Applied Sciences, Tech. Rep., 2007, version 2.1.

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

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

[25] J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos, "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication," Tech. Rep., 2007, http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/WebHome/HIWIRE_db_description_paper.pdf.

[26] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, 2007.

[27] 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.

[28] 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 Processing*, vol. 92, 2012.

[29] 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.

[30] E. Zwyssig, F. Faubel, S. Renals, and M. Lincoln, "Recognition of overlapping speech using digital MEMS microphone arrays," in *Proc. ICASSP*, 2013.

[31] 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.

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

[33] 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.

[34] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2013.

[35] 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.

[36] 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.

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

[38] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. LREC*, 2014.

[39] N. Kawaguchi, S. Matsubara, Y. Yamaguchi, K. Takeda, and F. Itakura, "Ciair in-car speech database," in *Proc. Interspeech*, 2004.

[40] D. Ellis, H. Satoh, and Z. Chen, "Detecting proximity from personal audio recordings," in *Proc. Interspeech*, Sep. 2014.

[41] E. Zwyssig, "Speech processing using digital mems microphones," Ph.D. dissertation, University of Edinburgh, Nov. 2013.