MITSUBISHI ELECTRIC RESEARCH LABORATORIES
Cambridge, Massachusetts

MICbots: collecting large realistic datasets for speech and audio research using mobile robots

Jonathan Le Roux (MERL)
Emmanuel Vincent (INRIA)
John R. Hershey (MERL)
Daniel P.W. Ellis (Columbia U.)

Motivation

- NEMISIG 2013, and the quest for the ideal cocktail party data
- Many datasets for robust speech processing research, but typically focus on either ASR, localization, or separation
- Hard to obtain realistic data while still having rich ground truth
  - ASR: hard/costly to record/annotate a lot of data in many environments
  - Speaker localization: requires calibrated tracking apparatus
  - Source separation: ground truth signals to measure performance and for discriminative training methods; close-talking mics not clean enough in “cocktail party” scenarios
- How could we get a large dataset of real recordings with
  - Ground truth speech signals, speaker location, and uttered words
  - Overlapping speech and/or strongly non-stationary interference
  - Source motion
  - Multi-microphone data

Data collection & acoustic emulation

Two approaches to ensure ground truth speech signals:

- Numerical simulation: from instantaneous mixing of separate mic array recordings, to application of room impulse responses (RIRs) to single-channel recordings
- Re-recording of existing data: playing single-channel recordings through loudspeakers

Different degrees of realism:

- "acoustically realistic": reflect realistic acoustic effects, e.g., mixing in mic, reverberant propagation, source motion, changes in environment, etc.
- "ecologically realistic": replicate non-acoustic properties, e.g., natural head motion, speech activity patterns, varied pronunciation & acoustic environment

Methods developed on acoustically realistic data are expected to work well on real data with similar acoustic properties

Where do existing datasets stand?

- Analysis of their characteristics regarding size, realism, availability of ground truth

		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]	¢	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

Where do existing datasets stand?

- Analysis of their characteristics regarding size, realism, availability of ground truth
- Often little/no interference

	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]	¢	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

Where do existing datasets stand?

- Analysis of their characteristics regarding size, realism, availability of ground truth
- Often little/no interference
  - Single-speaker reverberated speech datasets: lack signal ground truth or channel realism (TED, REVERB)

		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]	¢	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

Where do existing datasets stand?

- Analysis of their characteristics regarding size, realism, availability of ground truth
- Often little/no interference
  - Single-speaker reverberated speech datasets: lack signal ground truth or channel realism (TED, REVERB)
  - Overlapping speech datasets: unrealistic and small (CUAVE, Pascal SSC, SiSEC)

	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]	¢	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

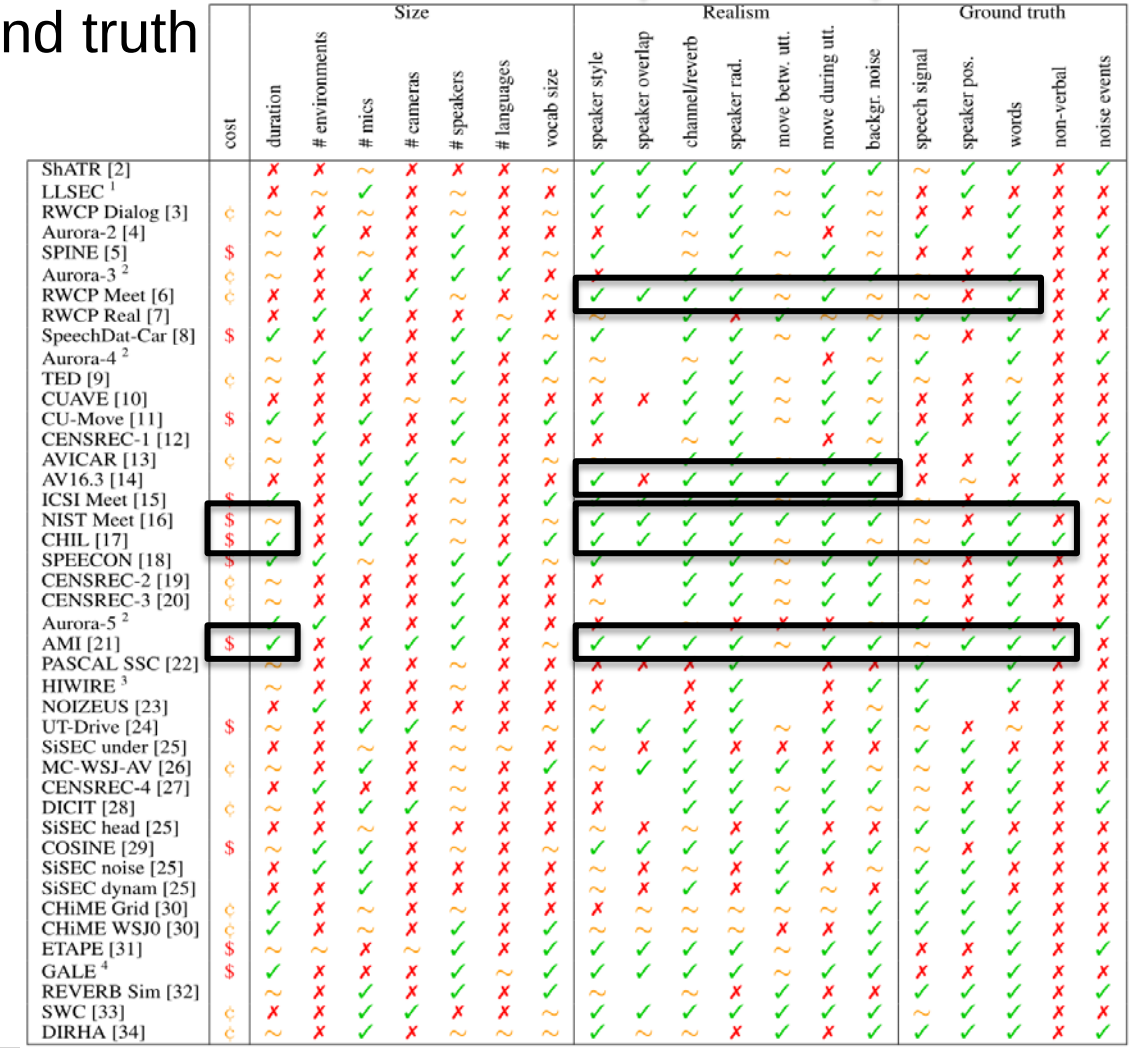
Where do existing datasets stand?

- Analysis of their characteristics regarding size, realism, availability of ground truth
- Often little/no interference
  - Single-speaker reverberated speech datasets: lack signal ground truth or channel realism (TED, REVERB)
  - Overlapping speech datasets: unrealistic and small (CUAVE, Pascal SSC, SiSEC)
  - Broadcast datasets: few mics, lack signal ground truth (GALE, ETAPE)

		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]	¢	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

Where do existing datasets stand?

- Analysis of their characteristics regarding size, realism, availability of ground truth
- Often little/no interference
  - Single-speaker reverberated speech datasets: lack signal ground truth or channel realism (TED, REVERB)
  - Overlapping speech datasets: unrealistic and small (CUAVE, Pascal SSC, SiSEC)
  - Broadcast datasets: few mics, lack signal ground truth (GALE, ETAPE)
  - Meeting/dialog datasets: rich but costly to create, hard to scale up, lack signal ground truth (ShATR, RWCP Meet, AV 16.3, NIST Meet, CHIL, AMI)



Where do existing datasets stand?

- With significant amount of noise: hard to record, limited

		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]	¢	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

Where do existing datasets stand?

- With significant amount of noise: hard to record, limited
  - Additive noise: large, but simulated mixing, scenarios limited to commands or read speech (Aurora-2, CENSREC-1)

		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]	¢	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

Where do existing datasets stand?

- With significant amount of noise: hard to record, limited
  - Additive noise: large, but simulated mixing, scenarios limited to commands or read speech (Aurora-2, CENSREC-1)
  - Real noisy recordings: only few environments (e.g., car), and a few command scenarios (Aurora-3, CU-Move, SPEECON);

		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]	€	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

Where do existing datasets stand?

- With significant amount of noise: hard to record, limited
  - Additive noise: large, but simulated mixing, scenarios limited to commands or read speech (Aurora-2, CENSREC-1)
  - Real noisy recordings: only 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)

		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]	€	~	✗	✓	✗	~	~	~	✓	~	~	✗	✓	✗	✓	✓	✓	✓	✗	✓

For more complete information

- **RoSP wiki**: https://wiki.inria.fr/rosp/Datasets
- Jonathan Le Roux, Emmanuel Vincent, "**A categorization of robust speech processing datasets**", Mitsubishi Electric Research Laboratories Technical Report, TR2014-116, Aug. 2014

	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	no	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

Our proposal: MICbots

- Use freely-moving robots to re-record existing human speech datasets

Our proposal: MICbots

- Use freely-moving robots to re-record existing human speech datasets
- Pros/Cons:
  - Low-cost: can re-use annotations
  - Scalable: just let the robots run
  - Ground truth for signals and location available
  - Transport setup to new environments and reproduce similar collection
  - Acoustic realism: real room acoustics, moving sources/mics/bodies
  - Acoustic realism: no time-varying radiation pattern;
    robot noise (but that can be good!)

Our proposal: MICbots

- Use freely-moving robots to re-record existing human speech datasets
- Pros/Cons:
  - Low-cost: can re-use annotations
  - Scalable: just let the robots run
  - Ground truth for signals and location available
  - Transport setup to new environments and reproduce similar collection
  - Acoustic realism: real room acoustics, moving sources/mics/bodies
  - Acoustic realism: no time-varying radiation pattern; robot noise (but that can be good!)
- Example scenario: cocktail-party
  - Re-record clean speech dataset such as WSJ0, TIMIT
  - Use multiple MICbots, each playing utterances from a separate subset of speakers

Constructing the first MICbots

Many thanks to John Barnwell (MERL)
for his help with the design and construction

Keywords: simple, low-cost

- Mostly off-the-shelf components

iRobot Create 2
$200

Raspberry Pi
$35

PlayStation Eye
(4ch mic array)
$8 (!!)

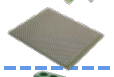
USB cable,
batteries…

Wifi
dongle
$18

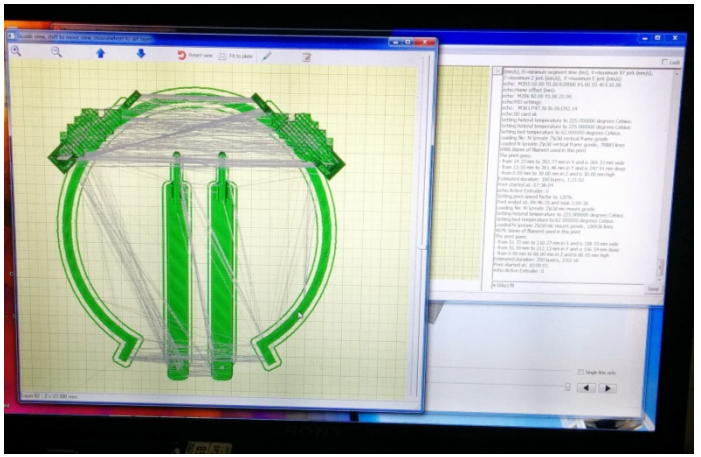
Speaker
$88

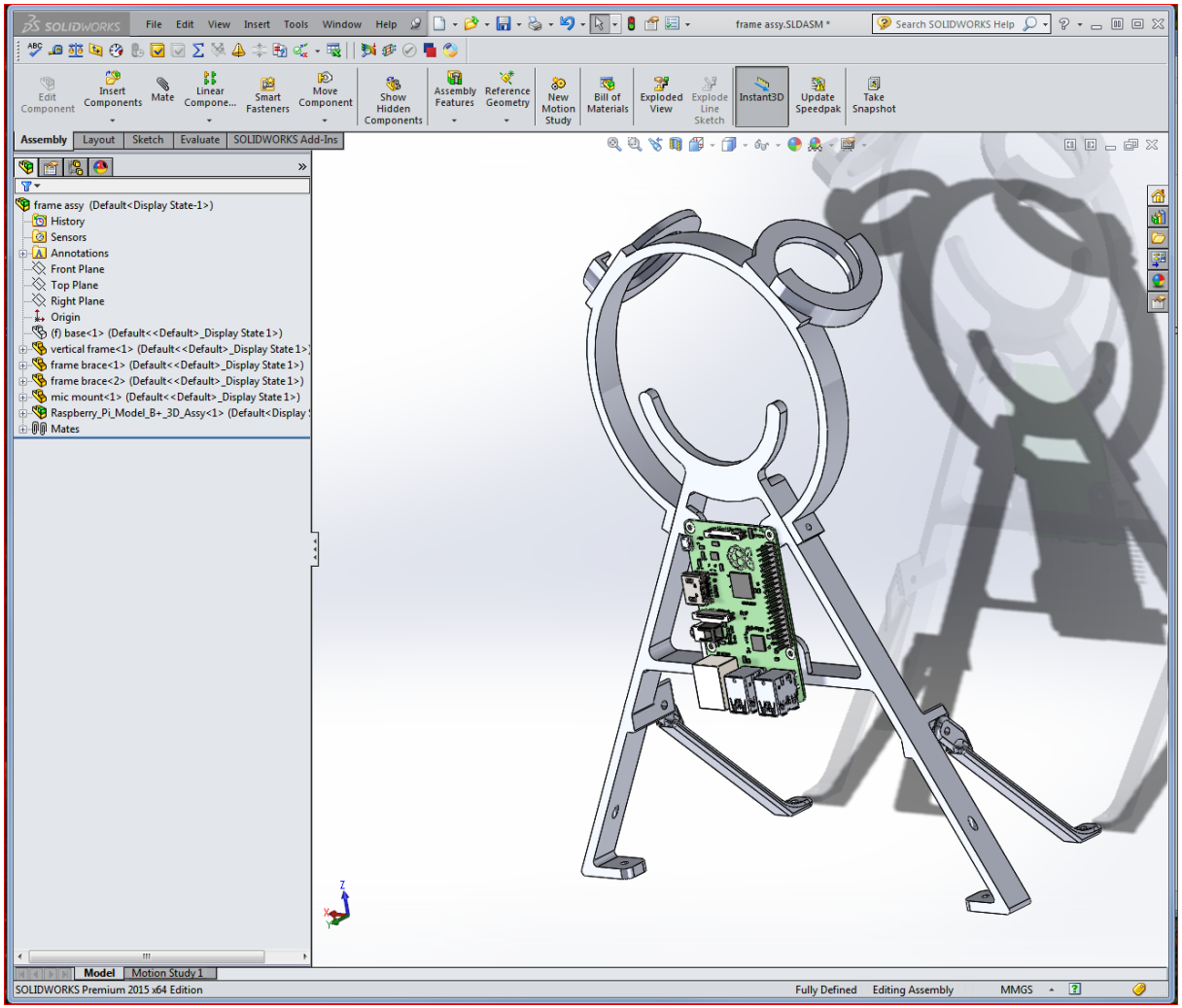
- Mounted on 3D-printed structure

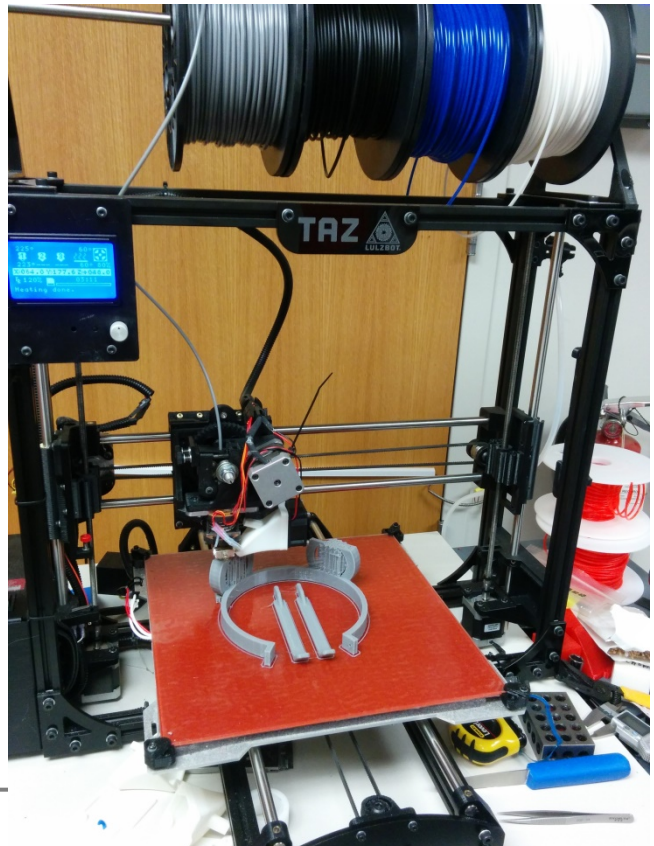
Shopping list for one robot

1 iRobot Create 2		$200
2 Playstation Eye 4ch Mic Array+Camera		$16
1 Jabra 410 Speaker		$88
1 Raspberry Pi	$35	
1 Raspberry Pi Case	$8	
1 Micro SD Card	$16	
1 Wifi dongle	$18	
1 USB-Micro USB Cable	$3	
Raspberry Pi+Accessories Total		$80
 1 Switching buck converter (OKI-78sr-5/1.5-W36-C)	$6	
 1 Vector PC Board (V2018-ND)	$6	
 2 Terminal blocks (A98334-ND, A98335-ND)	$4	
Buck converter for 5V power Total		$12
6 D Cells + 2 battery cases		$12
1 3D printing order (~$5 if self-printing)		$30
		$438

3D printing the parts







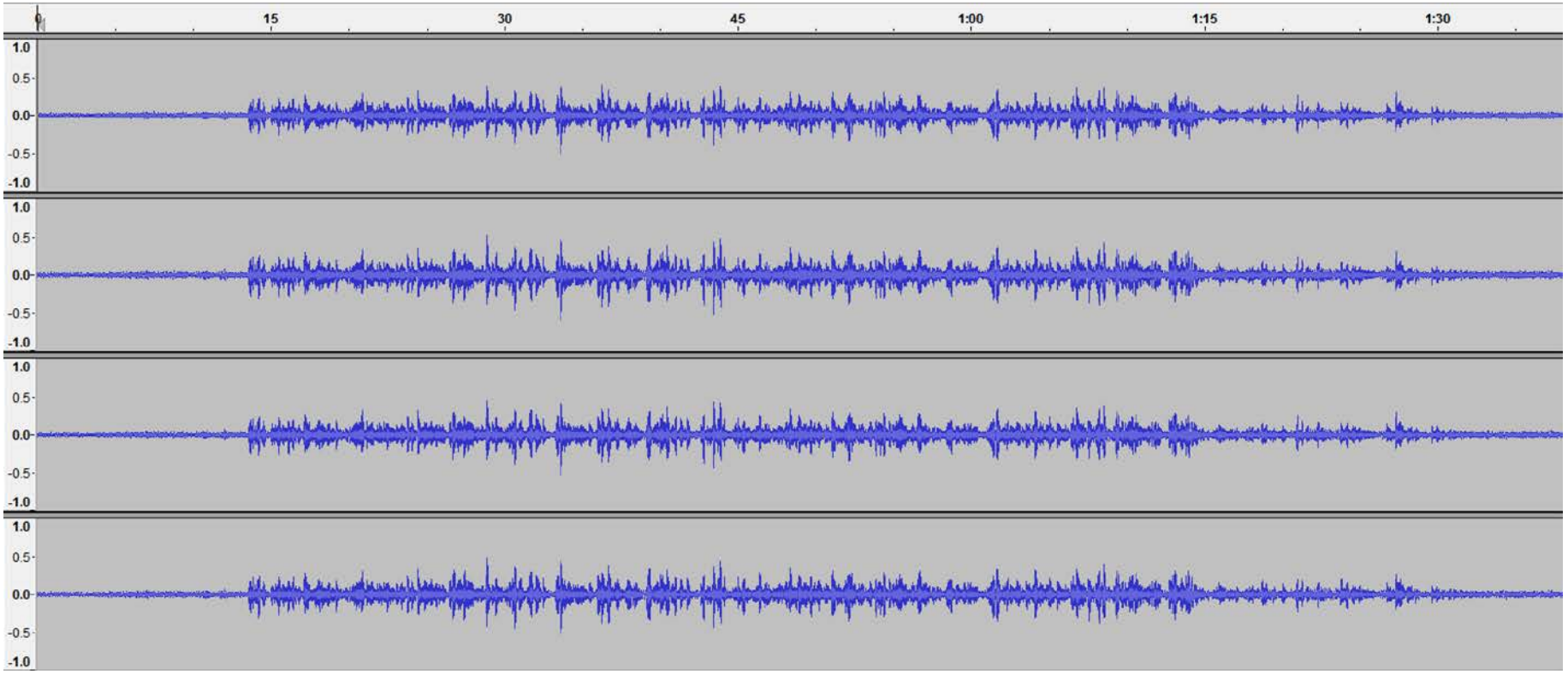
Assembling

Et voila!

Meet Dot, Hot and Lot

- https://youtu.be/cyTaDiVsPPM

Lot's recording



Challenges

- Alignment of all audio streams

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision
- Robot noise: could investigate reduction, but already limited and similar to HVAC noise at low speeds

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision
- Robot noise: could investigate reduction, but already limited and similar to HVAC noise at low speeds
- Determine collection protocol:

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision
- Robot noise: could investigate reduction, but already limited and similar to HVAC noise at low speeds
- Determine collection protocol:
  - Positions of robots in the room and with respect to each other

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision
- Robot noise: could investigate reduction, but already limited and similar to HVAC noise at low speeds
- Determine collection protocol:
  - Positions of robots in the room and with respect to each other
  - Allow movements during utterances or only between

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision
- Robot noise: could investigate reduction, but already limited and similar to HVAC noise at low speeds
- Determine collection protocol:
  - Positions of robots in the room and with respect to each other
  - Allow movements during utterances or only between
  - Reproduce “head” movements in addition to “body” movements

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision
- Robot noise: could investigate reduction, but already limited and similar to HVAC noise at low speeds
- Determine collection protocol:
  - Positions of robots in the room and with respect to each other
  - Allow movements during utterances or only between
  - Reproduce “head” movements in addition to “body” movements
  - Schedule timing of utterances to reproduce realistic speech overlap patterns

Challenges

- Alignment of all audio streams
- Accounting for the loudspeaker channel
- Ground truth location: SLAM w/ depth sensor → mm precision
- Robot noise: could investigate reduction, but already limited and similar to HVAC noise at low speeds
- Determine collection protocol:
  - Positions of robots in the room and with respect to each other
  - Allow movements during utterances or only between
  - Reproduce “head” movements in addition to “body” movements
  - Schedule timing of utterances to reproduce realistic speech overlap patterns
  - Suggestions?

Future plans

Stay tuned: www.jonathanleroux.org

- Finalize design and recording protocol
- Record and release large cocktail party dataset
- Release setup tutorial with CAD designs
- Investigate use of MICbots for RIR recordings
  - Detailed sampling of source/mic locations within a room
  - see our paper for more details
- Dataset of speech plus random moving sounds

Future plans

Stay tuned: www.jonathanleroux.org

- Finalize design and recording protocol
- Record and release large cocktail party dataset
- Release setup tutorial with CAD designs
- Investigate use of MICbots for RIR recordings
  - Detailed sampling of source/mic locations within a room
  - see our paper for more details
- Dataset of speech plus random moving sounds

- Try our methods on the data! <shameless plug>
  - Le Roux et al., "**Deep NMF for speech separation**" this afternoon at 15:50, AASP-L3.2, Mezzanine M2
  - Erdogan et al., "**Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks**", Friday at 10:50, AASP-P10, Poster Area G