



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 singlechannel 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

al face the					Size						F	Realisi	m				Gro	und t	ruth	
a trutn	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-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 noise [25] SiSEC dynam [25] CHIME Grid [30] CHIME Grid [30] CHIME WSJ0 [30] ETAPE [31] GALE ⁴ REVERB Sim [32]	0 \$ 00 \$ 0 \$ 0 \$ \$ 0 \$ \$ 0 \$ \$ 0 \$ \$ 0 \$ \$ 0 \$ \$ 0 \$ \$ 0 \$ \$ \$ 0 \$	***************************************	* ~ * ` * * * * * * * * * * * * * * * *	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	************************************	***************************************	* * * * * * * * * * * * * * * * * * * *	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	\$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$	1 1 1 1 1 1 1 1 1 1 1 1 1 1	>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	· · · · · · · · · · · · · · · · · · ·		- > > > > > > > > > > > > > > > > > > >	××××××××××××××××××××××××××××××××××××××	***************************************	· · · · · · · · · · · · · · · · · · ·	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	* * * * * * * * * * * * * * * * * * * *	××>×××××××××××××××××××××××××××××××××××
DIRHA [34]	¢	\sim	×	 Image: A set of the set of the	×	~	~	~	1	~	~	×	1	×	1	1	1	1	×	





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

					Size						F	Realis	m				Gro	und t	ruth	
d truth			12							dr	p		÷	II.						
	cost	duration	# environmen	# mics	# cameras	# speakers	# languages	vocab size	speaker style	speaker overla	channel/rever	speaker rad.	move betw. ul	move during t	backgr. noise	speech signal	speaker pos.	words	non-verbal	noise events
ShATR [2]		X	×	\sim	X	X	×	\sim	~	~	~	~	\sim	~	~	\sim	~	~	×	~
LLSEC ¹		X	\sim	1	X	\sim	×	X	1	1	1	1	\sim	1	\sim	X	1	×	×	×
RWCP Dialog [3]	¢	\sim	×	\sim	×	\sim	×	~	1	1	1	~	\sim	1	\sim	×	×	1	×	×
Aurora-2 [4]		\sim	 Image: A second s	×	×	1	×	×	×		\sim	1		×	\sim	 Image: A second s		1	×	 Image: A second s
SPINE [5]	\$	\sim	×	\sim	×	1	×	~	 Image: A set of the set of the		\sim	~	\sim	1	\sim	×	×	1	×	×
Aurora-3 ²	¢	\sim	×	×.	×	 Image: A second s	×.	×	×		1	1	\sim	1	1	~	×	1	×	×
RWCP Meet [6]	¢	X	×	×	×	\sim	×	\sim	 Image: A start of the start of	 Image: A second s	1	 	\sim	~	\sim	\sim	×	1	X	X
RWCP Real [7]	•	X	×.	1	×.	×	\sim	×	\sim		1	×	1	\sim	\sim	· ·	<i>.</i>	1	×.	<i>.</i>
SpeechDat-Car [8]	\$	 Image: A set of the set of the	×.	×	X	1	×	\sim	 Image: A second s		×	1	\sim	×	~	\sim	×	1	X	×.
Aurora-4 ²		~	×	X	X	1	×	~	~		\sim	1		×	\sim	 ✓ 		 Image: A second s	×	 Image: A second s
TED [9]	ç	\sim	X	X	×	 Image: A second s	×.	\sim	\sim		1	1	\sim	1	~	\sim	X	\sim	X	X
CUAVE [10]	dr.	X	Č.	×,	\sim	\sim	Č.	×.	X	×	1	1	\sim	1	\sim	V.	Č.	1	Č	Č.
CENEREC 1 [12]	•		· ^	.	<u>٠</u>	1	<u>٠</u>	*	*		×	×,	\sim	*			· ^	1	۰.	<u>^</u>
VICAP [12]		\sim	×.	<u></u>		×	С.	<u>^</u>	<u>^</u>		$\tilde{}$	٠,		<u></u>	\sim	1 V	~	٠,	÷.	*
AV163[14]	ç	$\widetilde{\mathbf{x}}$	Ŷ	×,	×,	\sim	Ŷ	$\tilde{\mathbf{x}}$	$\widetilde{}$	×	×,	×,	$\widetilde{\mathcal{I}}$	1	1		2	×.	Ŷ	Ŷ
CSI Meet [15]	\$	2	Ŷ	1	×.	~	Ŷ	- 2	1	2	1	1		1	1		×	- 2	2	~
NIST Meet [16]	š	~	Ŷ	1	Ŷ	~	Ŷ	~	1	1	1	1	1	1	1	~	Ŷ	1	x	×
CHIL [17]	š	1	x	1	12	\sim	x	1	1	1	1	1	~	1	~	\sim	12	1	2	x
SPEECON [18]	Š	1	1	~	x	1	1	~	1	· ·	1	1	\sim	1	1	\sim	x	1	x	X
CENSREC-2 [19]	ć	\sim	X	X	X	1	X	X	X		1	1	\sim	1	1	\sim	X	1	X	X
CENSREC-3 [20]	ė	~	X	X	X	1	X	X	~		1	1	\sim	1	1	\sim	X	1	X	X
Aurora-5 ²		1	1	X	X	1	X	X	X		\sim	X	X	X	\sim	1	X	1	X	1
AMI [21]	\$	1	X	1	1	1	x	\sim	1	1	1	1	~	1	1	\sim	1	1	1	X
PASCAL SSC [22]		~	X	×	X	\sim	X	×	X	X	X	1		×	X	 ✓ 		1	X	X
HIWIRE ³		~	X	X	X	\sim	X	X	X		X	1		X	1	1		1	X	X
NOIZEUS [23]		X	1	X	X	X	X	X	\sim		X	1		X	\sim	1		X	X	X
UT-Drive [24]	\$	\sim	×	1	1	~	×	~	1	1	1	1	\sim	1	1	\sim	×	~	X	X
SiSEC under [25]		×	×	\sim	×	\sim	\sim	X	\sim	×	1	×	×	×	×	1	1	×	×	×
MC-WSJ-AV [26]	¢	\sim	×	1	×	\sim	×	1	\sim	1	1	1	1	1	\sim	\sim	1	1	×	×
CENSREC-4 [27]		×	1	×	×	\sim	×	×	×		1	 Image: A second s	\sim	1	 Image: A second s	\sim	×	1	×	 Image: A second s
DICIT [28]	¢	\sim	×	 Image: A second s	 Image: A second s	\sim	×	×	×		 Image: A second s	 Image: A second s	 Image: A second s	1	\sim	\sim	 Image: A second s	 Image: A second s	×	 Image: A set of the set of the
SiSEC head [25]		×	×	\sim	×	×	×	×	\sim	×	\sim	×	1	×	×	 ✓ 	~	×	×	×
COSINE [29]	\$	\sim	1	1	×	\sim	×	\sim	 Image: A set of the set of the	×	 Image: A second s	×	1	×	 Image: A second s	\sim	×	×	×	×
SiSEC noise [25]		X	× .	1	X	×	×	×	\sim	×	\sim	×	1	×	\sim		1	×	×	×
SiSEC dynam [25]		×	×	 Image: A second s	X	×	×	×	\sim	×	 Image: A second s	×	 Image: A second s	\sim	×		1	×	X	X
CHIME Grid [30]	9	1	×.	\sim	×.	\sim	×.	×	×	\sim	\sim	\sim	\sim	\sim	1		1	1	×.	X
TABE [21]	ç	-	×	\sim	×	1	Č.	1	\sim	\sim	\sim	\sim	×	×.	1	1 🗸	1	1	<u> </u>	×
ETAPE [31]	\$	\sim	\sim	<u>.</u>	\sim	1	×	1	1	1	1	1	\sim	1	1	N	<u> </u>	1	<u> </u>	×
JALE	\$	 Image: A second s	Č	×,	Č	1	\sim	1	 Image: A second s	× .	1	<u> </u>	\sim	×.	×.	× 1	×.	1	×.	×
KEVERB Sim [32]		\sim	Č.	1	×	1	Č.	 Image: A set of the set of the	\sim	1	\sim	×.	1	×.	×.	 ✓ 	1	1	Č.	1
SWC [33]	ç	^	÷.	1	*	<u>^</u>	<u>^</u>	\sim	1	×	×	*	1	*	1	\sim	1	1	÷.	2
JIKHA [34]	ç	\sim	<u>^</u>	<u> </u>	<u>^</u>	~	~	~	×	~	~	<u>^</u>	<u> </u>	^	·	v	<u> </u>	<u> </u>	<u>^</u>	- ·





- 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)

100 109	u			9		~	,			ŧ					+					
1 4 . 41					Size						R	ealisr	n		-		Gro	und ti	uth	
d truth		ion	ironments	S	neras	akers	guages	b size	ter style	ker overlap	nel/reverb	ker rad.	e betw. utt.	during utt.	gr. noise	ch signal	cer pos.	s	verbal	events
	post	durat	‡ env	# mi	‡ car	f spe	f lan	voca	peak	peak	chan	peak	nove	nove	ack	beec	peak	vord	-uot	loise
01 4770 (0)	3		-++-		-++-	-#=	-#F	-	×	×	~	×	-			×	×		-	_
ShATR [2]		X	×	\sim	X	×	X	\sim	×.	×.	×.	×.	\sim	1	 Image: A second s	\sim	×.	×	X	~
LLSEC '		×	\sim	× .	×	\sim	X	×	×.	×.	×.	×.	\sim	1	\sim	X	×.	×	X	X
RWCP Dialog [3]	ç	\sim	×.	\sim	×.	\sim	×.	\sim	×	~	~	1	\sim	×.	\sim	× .	×	×.	×.	×
Aurora-2 [4]		\sim	×.	×	×.	×,	×.	×	×		\sim	×,		×	\sim	1 *		×,	×.	×
SPINE [5]	\$	\sim	×	\sim	×	×.	×	\sim	× .		\sim	×.	\sim	×.	\sim	×	×	×.	×	×
Aurora-3 ²	¢	\sim	×	×	×	~	 Image: A second s	×	×		1	1	\sim	1	~	~	×	1	×	×
RWCP Meet [6]	¢	×	×	×	 Image: A second s	\sim	×	\sim	 Image: A second s	 Image: A second s	1	 Image: A second s	\sim	 Image: A second s	\sim	\sim	×	1	×	×
RWCP Real [7]		×	×.	1	×	×	\sim	×	\sim		1	×	 Image: A second s	\sim	\sim	 ✓ 	×.	1	×	~
SpeechDat-Car [8]	\$	× .	×	~	×	~	~	\sim	~		~	~	\sim	 Image: A second s	~	\sim	×	~	×	×
Aurora-4 ²		\sim	 Image: A set of the set of the	×	×	1	×	I	\sim		\sim	1		×	\sim	<u> </u>		 Image: A second s	×	~
TED [9]	¢	\sim	×	×	×	~	×	\sim	\sim		1	1	\sim	1	~	\sim	×	\sim	×	×
CUAVE [10]		×	×	×	\sim	\sim	×	×	×	×	1	×.	\sim	1	\sim	~	×	1	×	×
CU-Move [11]	\$	1	×	~	×	1	×	~	~		~	1	\sim	 Image: A second s	~	X	×	1	×	×
CENSREC-1 [12]		\sim	×.	×	×	× .	X	×	×		\sim	1		×	\sim			1	×	×
AVICAR [13]	ç	\sim	×	×.	1	\sim	×	\sim	\sim		×.	1	\sim	×.	1	X	×	×	×	X
AV16.3 [14]		X	X	1	×.	\sim	X	X	1	X	1	×.	 Image: A set of the set of the	1	1	×	\sim	×	×	×
ICSI Meet [15]	\$	 Image: A second s	Č	×,	<u>.</u>	\sim	<u>.</u>	 Image: A second s	×,	×,	1	1	\sim	×,	1	\sim	<u>.</u>	×,	×.	\sim
NIST Meet [16]	\$	\sim	Č	×,	×.	\sim	×.	\sim	×,	×,	×.	1	 Image: A second s	×,	~	\sim	×.	×,	×.	×.
CHIL [1/]		1	<u></u>	× .	<u>ک</u>	\sim	<u></u>	· ·	×,	× .	1	1	\sim	1	\sim	\sim	<u>ک</u>	1	<u>ک</u>	- Č
SPEECON [18]	3	•	×.	\sim	÷.	×,	*	\sim	×.		×,	×,	\sim	×,	1	\sim	<u>۰</u>	×,	÷.	- C
CENSREC-2 [19]	ç	\sim	- C	- C	С.	×,	С.	- C	<u> </u>		×,	×,	\sim	×,	×,	\sim	<u>۰</u>	×,	С.	0
$\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{i$	ç	\sim	<u></u>	С.	С.	٠,	0	0	\sim		· ·	٠	\sim	٠.	•	\sim	0	٠,	0	2
Aurora-5 -		1	×.	<u></u>	×.	1	<u> </u>	×	<u></u>		\sim	<u></u>	×	×.	\sim	· ·	<u></u>	1	×.	<u> </u>
AMI [21]	Э	× .	- Č	×.	×.	× .	<u> </u>	\sim	×.	×.	×.	1	\sim	×.	×.	\sim	× .	1	×.	- Č
PASCAL SSC [22]		\sim	- <u>C</u>	- C	<u></u>	\sim	<u></u>	- <u>0</u>	- <u>C</u>	<u>^</u>	<u></u>	×.		<u></u>	<u></u>	×.		×.	<u></u>	
HIWIRE "		\sim	×	X	X	\sim	×.	X	×		×.	1		×.	~			×.	X	X
NOIZEUS [23]		~	×.	×.	×.	×	Č.	×	\sim		×.	1		×.	\sim	· ·		×	Č.	<u> </u>
UI-Drive [24]	•	\sim	÷.	× .	×.	\sim	<u>^</u>	\sim	× .	*	1	*	\sim	×.	×.	\sim	<u></u>	\sim	÷.	- Č
SISEC under [25]		<u>^</u>	÷.	\sim	÷.	\sim	\sim	· ^	\sim	1	×,	<u></u>	<u></u>	<u></u>	<u> </u>	· ·	×,	<u></u>	÷.	- C
MC-WSJ-AV [20]	ç	\sim	<u></u>	×.	<u>٠</u>	\sim	<u>٠</u>	*	\sim	 Image: A second s	1	1	× .	1	\sim	\sim	<u>ئ</u>	1	۰.	2
CENSREC-4 [27]		<u></u>	×.	· 2	<u></u>	\sim	÷.	- Ç	÷.		×,	×,	\sim	×,	×	\sim	<u></u>	×,	÷.	× 1
SISEC hand [25]	ç	\sim	С.		×.	\sim	С.	С I	<u></u>	~		*	×,	*	\sim	\sim	٠,	×.	<u>ې</u>	v
COSINE [20]	¢	2	2	$\tilde{}$	С,	<u></u>	С,	<u></u>	$\tilde{}$	2	\sim	2	×,	2	2		¥.	2	С,	- Ç
SiSEC noise [25]	Φ	\sim	×,	×,	С,	\sim	÷.	\sim		÷.		÷.	×,	¥.		\sim	2	×.	С,	- Ç
SiSEC hoise [25]		÷.	Š.	1	÷.	÷.	Ŷ	Ŷ		÷.	$\widetilde{\mathcal{L}}$	÷.	×,	<u></u>	$\widetilde{\mathbf{x}}$	1	×.	- Ç	÷.	- Ç
CHIME Grid [20]		2	С,		÷.	<u></u>	÷.	Q	$\widetilde{\mathbf{v}}$	2	×	<u></u>		\sim	2	1	×,	2	÷.	- Ç
CHIME UNU [50]	X	1	Ŷ	~	Ŷ	2	Ŷ	2	2	~	\sim	~	Ŷ	Ĩ	1	1	1	1	Ŷ	<u></u>
ETAPE [31]	š	~	2	×	2	2	Ŷ	1	1	1	1	1	2	2	1	×	x	1	Ŷ	2
GALE ⁴	¢	1	×	\$	×	1	1	1	1	1				1	1	y .	ý.	1	S.	· ·
DALE REVERB Sim [22]	Ф	×	Ŷ	2	Ŷ	1	$\widetilde{\mathbf{x}}$	1	× .	× 1		ý	$\widetilde{}$	×.	×.	2	2	1	Ŷ	2
SWC [33]		x	Ŷ	×.	2	x	Ŷ	~	2	1	Ľ,	2	×.	2	2	~	1	1	Ŷ	X
DIRHA [34]	ž	\sim	x	1	x	\sim	\sim	\sim	1	~	~	x	1	x	1	1	1	1	Ŷ	2
	T											-		-	-		-			-





- 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)

ics reg	Ja	IU		y	21	Zt	ラ,			ł					ł					
ما ۲. س. ۲. ام					Size						R	ealisr	n				Gro	und ti	uth	
a 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] AVI6.3 [14] ICSI Meet [15] NIST Meet [15] CHIL [17] SPEECON [18] CENSREC-2 [19] 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]	\$00	* 2 * 2 * 2 * 2 * 2 * 2 * 2 * 2 * * 2 * * 2 * * 4 m	13# × ~× 、 × × × × × × × × × × × × × × × ×	m# ~ 、 ~ × ~ 、 × 、 × × × × × × × 、 × × × ×	5# × × × × × × × × × × × × × × × × × × ×	は# × ~~~~×~~~~~~~~~~~~~~~~×~~~×~~~×~	#	<pre>/ ************************************</pre>	۲۲ ××۲ ۲ × ۲ × ۲ × ۲ × ۲ × ۲ × ۲ × ۲ ×	× ××< × × ×× ×× ×× ×× ××	A < < < < < < < < < < < < < < < < <	ads >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	۲. ۲. ۲. ۲. ۲. ۲. ۲. ۲. ۲. ۲. ۲. ۲. ۲. ۲	■ 「		۲ × ۲ × ۲ × ۲ × ۲ × ۲ × × × × × × × × ×	ads >>> × × × × × × × × × × × × × × × × ×	> × < > < > < < < < < < < < < < < < < < <	uuu × × × × × × × × × × × × × × × × × ×	××<<××××××××××××××××××××××××××××××××××
SiSEC noise [25] SiSEC dynam [25] CHiME Grid [30] CHIME WSJ0 [30] ETAPE [31] GALE ⁴ REVERB Sim [32] SWC [33] DIRHA [34]	00 \$ \$ 00	****	`	>> ? ? × × > > >	**** ~ ***	< < < < < < × <	****	< < < < × × × × × × × < < < < < < × × × × × × × < < < < < < < × × × × × × × × < < < < < < < × × × × × × × × × < < < < < < < < < < < < < × × × × × × × × × × × × < < < < < < < < < < < < < < < < < < ×	~~~~~~	×× 2 2 × × 2	1~1~~1~~1~	** ? ? ` ` * *	>>>>>>>>>>	* ~ ~ * * * * * *	~~~~~~	>>>> × ×>>>	>>>> × >>>	******	****	****





Where do existing datasets stand?

Analysis of their characteristics regarding size realism, availability of ground

- 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)

ics reg	a	IU		y	31	Zt	,			ŧ					ł					
					Size						R	tealisr	n				Gro	und ti	uth	
a 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] AVI6.3 [14] ICSI Meet [15] NIST Meet [15] 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]	\$00	× × × × × × × × × × × × × × × × × × ×	## × ~ × × × × × × × × × × × × × × × × ×	田井 ~ 、 ~ × ~ 、 × 、 × × × × × × × × · × × × × × × × ×	5# × × × × × × × × × × × × × × × × × × ×	\$\$# × 2 2 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	* * * * ^ * * * * * * * * * * * * * * *	××××××××××××××××××××××××××××××××××××××	Specific and the second se	x x x x x x x x x x x x x x x x x x x	3 くくくく××××く? くくくくくくくくくくくくく くくく? くくく? 4 and	ads <-> <> <> <> <> <> <> <> <> <> <> <> <> <>	×<××× ×× ×× ×× ×× ×× ×× ×× ×× ×× ×× ×× ×	MOL	× 2 × 2 × 4 2 × 4 2 × 4 2 × 4 4 4 4 4 5 5 5 5 7 5 7 5 8 9 ad	Solution	ads >>> × × × × × × × × × × × × × × × × ×	×<>×>	non × × × × × × × × × × × × × × × × × ×	<pre>x < x x x x x x x x x x x x x < x x x x</pre>
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]	\$ ¢¢\$ \$ \$ ¢¢	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	·>>×××~××××	>>> ~ ~ ~ × >>>	*****	· · · · · · · · · · · · · · · · · · ·	******	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	\$??*?\$\$?\$\$	5××77×× >7	~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~	****	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	·>×~~×>>×>×	~~~~~~~~~~~~	~~~~~~	*****	~~~~~~~~~~~	****	*****





- 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)

ics reg	Ja	ra	In	g	SI	Ze	Э,			ţ					ł					
ما ۲۰۰۰ ما					Size						R	lealisi	n		-		Gro	und ti	uth	
a 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]	-	X	X	~	X	X	X	\sim	~	~	~	~	\sim	~	~	\sim	~	~	X	
LLSEC ¹ RWCP Dialog [3] Aurora-2 [4] SPINE [5] Aurora-3 ²	¢ \$	× ~ ~ ~ ~	~ ×	< < < <	* * * *	~~~~	****	x { x { x	> > × > >	1	~ ~ ~ ~ ~ ~ ~	1111	~~~~	× × × ×	1111	× × √ ×	× × ×	* > > > >	× × × ×	× × × × × ×
RWCP Meet [6]	ž	X	x	x	2	~	x	\sim	1	~	~	~	2	~	2	2	X	~	x	x
RWCP Real [7]		×	1	1	×	×	~	×	~		v	~	v	\sim	\sim	v	-	~	×	1
SpeechDat-Car [8]	\$	1	×	1	×	1	1	\sim	1		~	1	\sim	1	1	~	×	1	×	×
Aurora-4 ²		~	<u> </u>	×	×	1	×	1	~		\sim	1		×	\sim	 Image: A start of the start of		~	×	
TED [9]	ç	\sim	Č.	÷.	×	×	Č.	\sim	\sim	~	1	1	\sim	1	×	\sim	Č.	\sim	Č.	- Ç
CU-Move [11]	\$	2	Ŷ	2	x	2	Ŷ	2	2	^	×	×,	\sim	×,	$\widetilde{\mathcal{I}}$	Ŷ	Ŷ	×,	Ŷ	Ŷ
CENSREC-1 [12]	₩	\sim	2	x	x	1	x	x	x		~	1		x	~	2	· ·	1	x	2
AVICAR [13]	ċ	\sim	×	1	1	~	×	~			/	1		1	-	X	×	1	×	X
AV16.3 [14]		×	×	1	1	\sim	×	×	× -	×	1	1	1	1	×	×	\sim	×	×	X
ICSI Meet [15]	\$	1	×	1	×	\sim	×	1			-		-	-	- 1		×	1	-	~
NIST Meet [16]	\$	\sim	X	1	×	\sim	×	\sim	1	×.	×.	×.	 Image: A second s	1	× .	\sim	×	×.	×	X
CHIL [17]	- \$	 Image: A start of the start of	X	~	<i>.</i>	\sim	×	1	 Image: A second s	 Image: A start of the start of	~	 Image: A start of the start of	~	~	~	~	 Image: A start of the start of	 Image: A start of the start of	 Image: A start of the start of	X
SPEECON [18]	3	~	×.	\sim	÷.	1	<i>.</i>	\sim	<i>.</i>		1	1	\sim	1	1	\sim	÷.	1	0	<u> </u>
CENSREC-2 [19]	ΙX.	\sim	Ŷ	Ŷ	Ŷ	×,	Ŷ	Ŷ	2		×,	1	\sim	1	1	\sim	Ŷ	1	Ŷ	Ŷ
Aurora-5 ²	Y	1	. 2	Ŷ.	Ŷ.	1	Ŷ	Ŷ	^v			<u>`</u>	~	<u>`</u>			<u> </u>	· /	<u> </u>	2
AMI [21]	\$	1	x	2	2	1	Ŷ	2	1	1	1	1	~	1	1	~	1	1	1	x
PASCAL SSC [22]		\sim	X	x	x	~	X	X				÷.				~	·	-	×	X
HIWIRE ³		~	×	X	X	\sim	X	X	X		×	1		×	1	1		1	×	X
NOIZEUS [23]		X	1	×	×	×	×	×	\sim		×	1		×	~	1		×	×	X
UT-Drive [24]	\$	\sim	×	1	1	\sim	×	\sim	1	1	1	1	\sim	1	1	\sim	×	\sim	×	X
SiSEC under [25]		×	×	\sim	×	\sim	\sim	×	\sim	×	1	×	×	×	×	 Image: A start of the start of	1	×	×	×
MC-WSJ-AV [26]	¢	\sim	×	 Image: A second s	×	\sim	×	 Image: A second s	\sim	✓	1	1	 Image: A second s	1	\sim	\sim	 Image: A second s	1	×	X
CENSREC-4 [27]	Ι.	×	×.	×	×	\sim	X	X	X		1	1	\sim	1	 Image: A second s	\sim	×	1	X	
DICIT [28]	ç	\sim	Č.	× .	~	\sim	Č.	- Č	×	~	×	1	1	<i>.</i>	\sim	\sim	1	1	Č.	~
COSINE [20]	¢	1	2	$\tilde{}$	<u>ې</u>	<u>^</u>	<u>ې</u>	<u>^</u>	\sim	1	\sim	1	1	2	1	× .	×.	^	<u>ې</u>	- 🗘
SiSEC noise [25]	•	x .	×.	×,	Ŷ	Ŷ	Ŷ	ĩ	~	×.	~	×.	×.	×.	~	$\tilde{\mathbf{z}}$	2	×.	Ŷ.,	<u></u>
SiSEC dynam [25]		x	x	1	Ŷ	Ŷ	Ŷ	Ŷ	\sim	Ŷ	1	x	1	2	x		1	Ŷ	x	Ŷ
CHiME Grid [30]	ć	1	X	~	X	\sim	X	X	X	\sim	~	\sim	~	\sim	1		1	1	X	X
CHiME WSJ0 [30]	l č	1	×	\sim	X	1	X	1	~	\sim	\sim	\sim	×	×	1	1	1	1	X	X
ETAPE [31]	\$	~	~	×	~	1	X	1	1	1	1	1	~	1	1	×	×	1	×	1
GALE ⁴	\$	1	×	×	×	1	\sim	1	1	1	1	1	\sim	1	1	×	×	1	×	×
REVERB Sim [32]		\sim	×	1	×	1	×	1	\sim		\sim	×	1	×	×	 Image: A second s	1	1	×	1
SWC [33]	¢	×	×	1	✓	×	×	\sim	1	1	1	1	1	✓	1	\sim	1	1	×	×
DIRHA [34]	¢	\sim	×	 Image: A set of the set of the	×	\sim	\sim	\sim	 Image: A set of the set of the	\sim	\sim	×	<u> </u>	×	 Image: A second s	 Image: A second s	<u> </u>	<u> </u>	×	 Image: A set of the set of the



s

Where do existing datasets stand?

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

					Size						R	Realisi	n				Gro	und ti	ruth	
oise:	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
hATR [2]		×	X	\sim	X	×	X	~	~	~	~	~	\sim	~	~	\sim	~	~	×	~
LSEC ¹		X	\sim	1	X	\sim	×	X	1	1	1	1	\sim	1	\sim	X	1	×	×	×
RWCP Dialog [3]	¢	\sim	×	\sim	×	\sim	×	~	1	1	1	1	\sim	1	\sim	×	×	1	×	×
Aurora-2 [4]		\sim	 Image: A second s	×	×	~	×	×	×		\sim	 Image: A second s		×	\sim	 ✓ 		~	×	 Image: A second s
PINE [5]	\$	\sim	×	\sim	×	1	×	~	1		\sim	1	\sim	1	\sim	×	×	1	×	×
Aurora-3 ²	¢	\sim	×	1	×	1	1	×	×		1	1	\sim	1	1	~	×	1	×	×
RWCP Meet [6]	¢	×	×	×	 Image: A second s	\sim	×	\sim	 Image: A second s	 Image: A second s	×.	 Image: A second s	\sim	 Image: A second s	\sim	\sim	×	1	×	×
WCP Real [7]		×	×	1	×	×	\sim	×	\sim		1	×	~	\sim	\sim	 Image: A set of the set of the	×	1	×	×.
peechDat-Car [8]	\$	~	×	~	×	×.	~	\sim	1		~	×	\sim	~	~	~	×	 Image: A second s	×	×
Aurora-4 ²		\sim	 Image: A second s	×	×	1	×	~	\sim		~	1		×	\sim	 ✓ 		~	×	 Image: A second s
ED [9]	¢	\sim	×	×	×	~	×	\sim	\sim		1	1	\sim	1	~	\sim	X	\sim	X	X
UAVE [10]	đ	×.	Č.	×.	\sim	\sim	Č.	×.	×.	×	1	1	\sim	1	\sim	X.	Č.	1	Č.	Č.
U-Move [11]	э	×	<u></u>	<i>.</i>	÷.	1	٠	*	Ý.,		×	1	\sim	*	×		^	1	÷.	2
VICAP [13]	~	\sim	¥.	2	2	×	÷.	2	1		$\widetilde{}$	1		2	$\tilde{}$	× 1	×	1	÷.	Š.
V163[14]	Ŷ	x	Ŷ	×,	×,	$\tilde{\sim}$	Ŷ	x	2	x	×.	×,	2	×,	×.	Ŷ	2	x	Ŷ	Ŷ
CSI Meet [15]	\$	2	Ŷ	1	x	~	Ŷ	- 2	1	2	1	1	~	1	1	\sim	x	2	2	\sim
JIST Meet [16]	š	÷	X	1	X	~	X	~	1	1	1	1	1	1	1	\sim	X	1	x	X
HIL [17]	š	1	X	1	1	\sim	X	1	1	1	1	1	~	1	~	~	1	1	1	X
PEECON [18]	Š	1	1	~	X	1	1	~	1		1	1	\sim	1	1	~	X	1	X	X
ENSREC-2 [19]	ċ	\sim	×	×	X	1	×	×	×		1	1	\sim	1	1	\sim	X	1	X	X
ENSREC-3 [20]	¢.	\sim	×	×	X	1	×	X	\sim		1	1	\sim	1	1	~	×	1	×	×
urora-5 ²		1	1	×	X	1	x	X	X		\sim	X	×	×	\sim	1	X	1	×	1
MI [21]	\$	1	X	1	1	1	X	\sim	1	1	1	1	\sim	1	1	\sim	1	1	1	×
ASCAL SSC [22]		\sim	×	×	×	\sim	×	×	×	×	×	1		×	×	1		1	×	×
IIWIRE ³		\sim	×	×	×	\sim	×	×	×		×	1		×	1	1		1	×	×
NOIZEUS [23]		×	1	×	×	×	×	X	\sim		×	1		×	\sim	1		×	×	×
JT-Drive [24]	\$	\sim	×	1	1	\sim	×	~	1	1	1	1	\sim	1	1	~	×	\sim	×	×
iSEC under [25]		×	×	\sim	×	\sim	\sim	×	\sim	×	1	×	×	×	×	1	1	×	×	×
AC-WSJ-AV [26]	¢	\sim	×	 Image: A second s	×	\sim	×		\sim	✓	1	1	 Image: A set of the set of the	1	\sim	~	 Image: A second s	1	×	×
CENSREC-4 [27]		×	×	×	×	\sim	×	×	X		×.	×.	\sim	1	× .	\sim	×	×.	×	×.
DICIT [28]	ç	\sim	×	 Image: A second s	×.	\sim	×	X	X		 Image: A second s	×.	1	×.	\sim	\sim	1	×.	X	×.
SISEC head [25]	æ	×	×.	\sim	Č.,	×	Č.,	×	\sim	×.	\sim	×.	1	×.	×.	 Image: A start of the start of	<u>ک</u>	×.	Č.,	Č.,
USINE [29]	Э	\sim	×,	×,	٠.	\sim	÷.	\sim	× .	<u>ن</u>	×	*	×,	×.	×	\sim	<u>^</u>	*	۰.	÷.
SEC dynam [25]		÷.	*	×,	<u>ې</u>	С,	<u>ې</u>	С (\sim	÷.	$\widetilde{}$	<u>ې</u>	×,	<u></u>	\sim	1	×,	<u>ې</u>	÷.	÷.
THIME Grid [20]	~	2	Ŷ		Ŷ	2	Ŷ	<u></u>	Ŷ	2		2		2	2	1	1	2	Ŷ	Ŷ
HIME WSI0 [30]	X	1	Ŷ	\sim	Ŷ	1	Ŷ	- 2	2	\sim	\sim	~	×	×	1	2	1	1	Ŷ	Ŷ
TAPE [31]	š	~	\sim	x	\sim	1	x	1	1	1	1	1	\sim	2	1	x	x	1	x	2
ALE ⁴	\$	1	x	x	x	1	~	1	1	1	1	1	~	1	1	x	x	1	x	x
EVERB Sim [32]	Ψ	~	x	2	x	2	x	1	~	•	~	x	1	x	x	2	2	2	x	2
WC [33]	ć	X	X	1	1	x	X	\sim	1	1	1	1	1	1	1	~	1	1	X	X
DIRHA [34]	č	\sim	×	1	×	\sim	\sim	~	1	~	~	×	1	×	1	1	1	1	×	1



- With significant amount of no hard to record, limited
 - Additive noise: large, but simulated mixing, scenarios limited to commands or read speech (Aurora-2, CENSREC-1)

					Size						R	ealis	m				Gro	und t	ruth	
oise:	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 dynam [25] CHIME Grid [30] CHIME WI30 [30] ETAPE [31]	0 \$ 0 0 \$ 0 \$ \$ \$ \$ \$ \$ 0 0 \$ \$ \$ \$ \$ \$	**?????**\$??*\$??*\$??*\$??*\$??*??*?*?*?*?	* ~ * * * * * * * * * * * * * * * * * *	<pre>2 \$ 2 \$ 2 \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$</pre>	* * * * * * * * * * * * * * * * * * * *	* ? ? > > > > > > > > > > ? ? ? > > ? ? ? ? > > > > > > ? ? ? * ? *	* * * * * * ~ * * * * * * * * * * * * *	{<************************************	\$	555 5 X X555 5X 5X5 X5XX275	Solution (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	2×2×××××××××××××××××××××××××××××××××××	>>>×>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	> ????>?>?>?>?>?>?>?>?>?>?>?>?>?>?>?>?>	2×××× 22×2×2×××××22222×2×2×2×2×2×2×2×2×	<pre>>> × ××> × ×× × ~×>>××> ×>>×>>>>>>>>>></pre>	> × > > > > > > > > > > > > > > > > > >	* * * * * * * * * * * * * * * * * * * *	××××××××××××××××××××××××××××××××××××××
GALE ⁴ REVERB Sim [32] SWC [33] DIRHA [34]	9 \$ \$ \$	✓ ✓ ✓	× × × ×	****	××××	· · · · × ~	; ~ × × ~		~ ~ ~ ~ ~ ~ ~	• • • • •	• • • • •	> × > ×	~~~~	× × ×	× × × ×	; ×	(× / / / /	1111	* * * *	× / × /



- With significant amount of no 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 s (Aurora-3, CU-Move, SPEECON);

-					Size						R	lealisi	n				Gro	und t	ruth	
oise:	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]		×	×	\sim	×	×	×	~	~	~	~	~	\sim	~	~	~	~	~	×	
LLSEC ¹ RWCP Dialog [3]		×	$\tilde{\mathbf{x}}$	~	X	\sim	X	×	1	1	1	1	\sim	1	\sim	X	4	×	X	X
Aurora-2 [4]	Ý	\sim	2	X	x	1	Ŷ	X	x		~	1		x	\sim	2	· ·	1	Ŷ	21
SPINE [5]	\$	\sim	×	~	×	1	×	~	<u>`</u>		\sim	1	\sim	1	\sim	×	×	1	×	×
Aurora-3 ²	¢	~	×	1	×	1	1	×	×		1	1	\sim	1	1	~	×	1	×	X
RWCP Meet [6]	¢.	X	~	×	1	\sim	×	~		1	1	1	\sim	1	\sim	\sim	×	1	×	×
RWCP Real [7]		×	1	1	×	×	\sim	×	\sim		1	×	1	\sim	\sim	1	1	1	×	1
SpeechDat-Car [8]	\$	1	×	1	×	1	1	~	1		~	~	\sim	1	1	~	×	1	×	×
Aurora-4 ²		\sim	1	×	×	1	×	1	\sim		\sim	1		×	\sim	 Image: A second s		1	×	 Image: A second s
TED [9]	¢	\sim	×	×	×	~	×	\sim	\sim		1	1	\sim	1	1	\sim	×	\sim	×	X
CUAVE [10]		<u> </u>	Č	X	\sim	\sim	×	X	×	×	1	1	\sim	1	\sim	X	×.	1	×	× I
CENSPEC 1 [12]	•	*	<u>^</u>	1 *	<u>ې</u>	1	٠.	*	<i>.</i>		×	1	\sim	*	×		^	1	÷.	2
AVICAR [13]	~	\sim	×	2	2	×	Ŷ	2	2		2	1	~	2	$\widetilde{\mathcal{I}}$	×	×	1	Ŷ	¥ I
AV163[14]	Y	x	Ŷ	÷.	÷.	\sim	Ŷ	x	2	x	÷.	÷.	1	×.	×.	x	2	x	Ŷ	x
ICSI Meet [15]	\$	1	X	1	x	\sim	X	1	1	1	1	1	~	1	1	\sim	X	1	1	\sim
NIST Meet [16]	\$	\sim	X	1	X	\sim	X	\sim	1	1	1	1	1	1	1	\sim	X	1	X	×
CHIL [17]	\$	1	×	1	1	\sim	×	1	1	1	1	1	\sim	1	\sim	~	1	1	1	X
SPEECON [18]	\$	1	1	\sim	×	1	1	\sim	1		1	1	\sim	1	1	\sim	×	1	×	×
CENSREC-2 [19]	¢	\sim	×	×	×	1	×	×	×		1	1	\sim	1	1	\sim	×	1	×	×
CENSREC-3 [20]	¢	\sim	×	×	×	~	×	×	\sim		~	~	\sim	~	 Image: A second s	~	×	1	×	×
Aurora-5 ²		1	1	×	×	1	×	×	×		\sim	×	×	×	\sim	 Image: A start of the start of	×	1	×	
AMI [21]	\$	 Image: A set of the set of the	×	1	1	1	×	\sim	×.	×.	1	1	\sim	×.	1	\sim	1	1	×.	X
PASCAL SSC [22]		\sim	×	×	×	\sim	×	X	×	×	×	1		×	×	1		1	×	X
HIWIRE '		\sim	×	×	×	\sim	×	X	×		×.	1		×	~			 	×	X
NOIZEUS [23]	¢	×	<i>.</i>	×.	×.	×	Č.	×	\sim	1	×.	1		×.	\sim	· ·	~	~	Č.	<u> </u>
SiSEC under [25]	•	\sim	÷.	×	*	\sim	<u>^</u>	\sim	¥.,	Ý.,	1	*	\sim	*	*	\sim	2	\sim	÷.	_ ≎
MC-WSI-AV [26]	~	2	Ŷ	2	Ŷ	$\tilde{\sim}$	ĩ	2	$\widetilde{\sim}$	2	×,	2	2	2	2	×	×,	2	Ŷ	<u></u>
CENSREC-4 [27]	Y	x	2	x	x	\sim	x	x	X	•	1	2	~	1	1	~	x	1	x	21
DICIT [28]	ć	\sim	X	1	1	\sim	×	X	x		1	1	1	1	~	~	1	1	×	1
SiSEC head [25]		×	X	\sim	×	×	X	X	\sim	×	\sim	×	1	×	×	1	1	×	X	×
COSINE [29]	\$	\sim	1	1	×	\sim	×	\sim	1	1	1	1	1	1	1	\sim	×	1	×	×
SiSEC noise [25]		×	1	1	×	×	×	×	\sim	×	\sim	×	1	×	\sim	 Image: A second s	1	×	×	×
SiSEC dynam [25]		×	×	1	×	×	×	×	\sim	×	~	×	1	\sim	×	1	1	×	×	×
CHiME Grid [30]	¢	1	X	\sim	X	\sim	×	X	×	\sim	\sim	\sim	\sim	\sim	1	1	1	1	×	X
CHIME WSJ0 [30]	ç	 Image: A second s	×	\sim	×	1	×.	1	\sim	\sim	\sim	\sim	×	×	1	1	×.	1	×.	×
EIAPE[31]	3	\sim	\sim	<u></u>	\sim	1	^	1	1	1	1	1	\sim	1	1	^	<u></u>	1	<u></u>	*
GALE	\$	× .	<i>.</i>	^	<i>\</i>	1	\sim	1	× .	× .	× .	1	\sim	<u>ک</u>	<u>ک</u>		^	1	Č.	×
SWC [33]		$\widetilde{\mathbf{x}}$	Ŷ	1	2	×	Ŷ	×.	$\widetilde{}$	1	$\widetilde{}$	2	1	2	2		1	1	Ŷ	Y I
DIRHA [34]	×.	2	Ŷ	1	x	2	2	$\widetilde{\sim}$	1	~	~	x	1	x	1	2	1	1	Ŷ	2
Dirata [54]	Y		r	•	r				•			r	•	<u> </u>	•	•	•	•	<u> </u>	•



Where do existing datasets stand?

- With significant amount of no 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						R	lealisi	m				Gro	und t	ruth	
oise:	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] CU-Move [11] CU-Move [12] Aurora-5 ² AMI [21] PASCAL SSC [22] HWIRE ³ NOIZEUS [23] UT-Drive [24] SiSEC under [25] COSINE [29] SiSEC coise [25] SiSEC dynam [25] CHIME Grid [30]		***************************************	+ × ~×> × ××> × ××××××××××××××××××××××××	+ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~	+ × × × × × × × × × × × × × × × × × × ×	# × ? ? > > ? ? > > > ? ? ? ? ? ? > > > ? ? ? ? * ? *	# * * * * * * * * * * * * * * * * * * *	****	× \ \ × × \ × \ × \ × \ × \ × \ × \ × \	s >>> > > > > > > >>> >>> >>>>>>>>>>>>	\$ \$ \$ \$ 2 2 5 5 5 5 2 5 5 5 2 5 5 5 5 5	s > > > > > > > > > > > > > > > > > > >	1 222 22252 223 2552 223 25252 2×2 2×25555555	I >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	× 1 × 1 × 1 × 1 × 1 × 1 × 1 × 1 × 1 × 1	s 2 ×××× 2 2 5 2 × ×××× 2 2 2 2 2 × ××× 2 × 2	s >>> × × ××>× ××× × ~××××××× >> ×>>×>>×>>	· × > > > > > > > > > > > > > > > > > >	- * * * * * * * * * * * * * * * * * * *	-
CHIME WSJ0 [30] ETAPE [31] GALE ⁴ REVERB Sim [32] SWC [33]	¢ \$ \$	√ < √ < ×	* ~ * * *	~ * * * *	* ~ * * *	\$ \$ \$ \$ \$ \$ \$ \$ \$ \$	× × ~ × ×	2222	~~~~	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	~ ~ ~ ~	~ > > > >	*~~~	× × × × × ×	5 5 X X 5	× × > ~	× × × ×	1111	* * * * *	× / × / × /
DIRHA [34]	Ç	\sim	×	 Image: A second s	~	\sim	\sim	\sim	 Image: A second s	\sim	\sim	×	 Image: A second s	×	 Image: A second s	 Image: A second s	 Image: A second s	 Image: A second s	×	 Image: A second s





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 at	ttributes							S	Speech				Channel			Noise		G	ound truth		
				-	-9		9	-00			-											1	24	
				Hz	УII		10	(E)			×.					-		25				100	7	é .
	100		e (j.	6	si o	E .	-	8	숺		eds	2	LCC.	2	×	- P	2	NO8		<i>64</i>	7	2		£ 18
	La la	201	.5	12	5	8	D0	-	de		ũ.	8	1	AL.	161	2	1	-	141	N	6	28	4 7	0 0
	2	10 e	3			deo	at (cok	1	2 ¹	3	eak	cak	Ti a	1	cak	cak	eat	ai ai	80	3	eak	pu /	
Datasets	12	sn	2	a	-6	1	8	8	5	. д	5	8	₽.	8	-8	8	8-	8	8	ř.	8	₽.	8 8	á ă
ShATR [1]	1994	meeting	0.6	-48	3	10	free	0.6	5	UK English	1	spontaneous	5	multiple dialogs	reverb	human	quasi-fixed	head	meeting	high	headset	yes	yes n	to yes
LLSEC	1996	dialog	1.4	16	4	DO	free	. *	12	N/S	N/S	read, spontaneous	2	dialog	reverb	human	quasi-fixed	head	hallway, restaurant (scenarized)	medium	00	yes	10 J	08 00
RWCP Spoken Dialog Corpus [2]	1996 - 1997	dialog	10	16	- 2	no	free	10	39	Japanese	19	spontaneous	1 - 2	dialog	reverb (low)	human	quasi-fixed	head	stationary background	high	no	no	yes p	10 110
Aurora-2 [3]	2000	public spaces	33	8 - 16	1.	no	free given TIDigits (0.5 kS)	33	214	US English	0.01	digits	1	no	simulated phone	human	N/S	no	various real environments (added)	low	original	N/S	yes #	to yes
SPINE1, SPINE2 [4]	2000 - 2001	military	38	16	2	DO	7.4 kS	2	100	US English	1	command, spontaneous	1 - 2	no	simulated radio	human	quasi-fixed	head	military (added)	low	во	BO	yes u	10 100
Aurora-3 (subset of SpeechDut- Car) [5]	2000 - 2003	car	2	16	4	DO	1 k	2	730	various	0.01	digits	1	BO	reverb	buman	quasi-fixed	head	car	low	headset	no	yes #	10 10
RWCP Meeting Speech Corpus [6]	2001	meeting	3.5	16 - 48	1.	3	lice	3.5	. 7	Japanese	-2.1	spontaneous	1-5	meeting	reverb (low)	human	quasi-fixed	head	stationary background	high	hradset	80	yes p	0 no
RWCP Real Environment Speech and Acoustic Database [7]	2001	domestic, office	3	16 - 48	84	no	free		5	US English, Japanese	12	read	1	no	real rir, reverb	loudspeaker	various	no, pivoting arm	various (sum of events)	medium	original	yes	yes P	to yes
SpeechDut- Car [8]	2001 - 2011	Car	2	16	4	00	39 - 182 k per lang	2	300 per lang	various	394	digits, command, read, spontaneous	1	DO.	reverb	buman	quasi-fixed	heal	car	low	headset	no.	VEN P	00 00
Autora-4 [9]	2002	nublic spaces	7	8 - 16	1	DO	free given WSJ0 (1.5 kS)	2	101	US English	10	read	i i	BO	simulated phone	human	N/S	BO	various real environments (added)	low	original	N/S	YES P	to yes
TED (10)	2002	seminar	47	16	î.	no	0.5 kS	47	188	non-native English	293	lecture	1 or more	seminar	reserb	human	uuasi-fixed	head	stationary background	hirb	lanel	00	partial P	10 00
CUAVE III	2002	speech overtan	3	44	i.	1	firm	3	36	US English	0.01	distre	1.2	full	reverb	human	oursi-fixed	bend.	stationary background	high		80	100 F	00 00
CU-Move Microphone Array Data [12]	2002 - 2011	car.	286	44	6-8	DO	2515	286	172	US English	12	digits command read dialog	1	Dia.	reset	burnan	anai-fixed	brad	car	how	80	100	100 1	Ars 1945
CENSREC-1 (Autora 21) [13]	2003	public spaces		8	1	70	Inc	-	214	Japanese	0.01	diaits		00	simulated phone	buman	N/S	80	various real environments (added)	Inn	original	N/S	100 F	-
AVICAR (14)	2004	present operates	20	16			feest	20	86	US English non-native English	1	rend		100	month	harmon	onnei-fixed	hand		how		- BO		
47163115	2004	manting	1.5	16	16		free	15	12	N/S	NUS	constantion .	1.1	full	reverb	horman	- quasi-nice o	hand walk	stationary background	hinh	100	matial	100 1	
ICSI Manting Corners [16]	2004	mating	73	16	6		2816	72	51	US English other English	11	mating	7.10	manting	reverb	human	auxi-fixed	hand.	mating	high	handrat land	- mer		ad hos
NICT Meeting Dilat Comme Frank (17)	2004	meeting	12	16		100	2.5.15	12	11	LIP English	-	incentry .	1.0	arccung	and the second s	and the second sec	dime.ured	hand mult	and a second	high	headers ined		100 11	
CON Maning (10)	2004 2007	incering	60	10	70	100 E	242		DI.	Co Caginal	-	Income	3 . 20	necong	in the second se	harriste	THE REAL PROPERTY.	head a	Automaty out aground	high	bander, sayes	100	365 10	10 INT
Critic Meetings (18)	2004 - 2007	senurar, meeting			19-141	0-9	TE h mar have		and some first	non-manye rangnan	14	seminar, meening	3-20	seminar, insecting	reverb	human	quest-fixed	Incast	meening (scenarized)	nign	headset	yes	yes ye	CA IIO
CENERG 2 CON	2004-2011	public space, domestic, onice, car	4	10	- 1	00	/3 k per tang		oto per tang	various	0.01	contriand, read, spontaneous	÷ .	BO	revero	human	quasi-fixed	incast.	various real environments	incontain	headset	10	yes m	0 10
CENSIGE-2 20	2003	can	1	10		00	nee	2	214	Jupanese	0.01	augus	- C	100	reserv	ISTUISTU	quasi-inxed	nead	car	ROW	neauset	10	yes in	10 110
CENSREC-3 [21]	2005	car	1	10	1	no.	21 K		311	Japanese	0.05	read	÷ .	no	reverb	human	quasi-fixed	ficad	car	low	headset	no	yes n	a 110
Autora-5 [22]	2006	public spaces, domestic, otnice, car		- 8	1.	no	tree given Tithighs (0.5 kS)	1 2	445	US English	0.01	aigns	1.1	во	no, simulated nr, reat nr	loudspeaker	nxed	0.0	various real environments (adoed)	BOW	onginai	no	yes n	io yes
AMI [23]	2006	meeting	100	10	10	0	lice		189	UK English, other English	8	meeting	most often 4	meeting (18% overlap)	reverb	buman	quasi-fixed	head	stationary background	nigh	headset, tapet	yes.	yes ye	cs no
PASCAL SSC [24]	2006	speech overlap	8.8	42	1	no	tree	8.8		UK English	0.05	command		run	80	numan	20/5	no	RO	N/S	original	N/S	yes n	0 10
HIWIKE [25]	2007	airptune	21	10	1	00	0.05 K	- 21	81	non-native English	0.1	command		80	20	numon	34/5	no	urplane (added)	tow	original	N/S	yes in	10 100
NOLZEUS [26]	2007	public spaces	0.6		1	DO	Tree	0.6	0	US English	0.1	read		Bo	simulated phone.	Buman	31/3	80	various real environments (added)	10W	original	N/S	no na	40 100
UT-Drive [27]	2007	car	40	25	- 2	- 2	25 kS	40	-25	US English	2.4	command, dialog	1 - 2	dialog	reverb	human	quasi-fixed	head	Car	low	fieadset (low quality)	DO F	parnal ne	10 NO
SASSEC, SiSEC under- determined [28]	2007 - 2011	cocktail purty	0.3	16	2	- 00	free	0.3	16	N/S	N/S	read	3-4	full	simulated nr, real nr, reverb	no, loudspeaker	fixed	80	no	N/S	original, spatial image	yes	no n	10 10
MC-WSJ-AV, PASCAL SSC2, 2012.MMA, REVERB RealData [29] [3]	2007 - 2014	speech overlap	10	10	8 - 40	partial	1.5 kS	3	45	UK English	10	read	1 - 2	full	reverb	human	various	head, walk	stationary background	high	beadset, lapel	yes	yes n	10 80
CENSREC-4 (Simulated) [31]	2008	public spaces, domestic, office, car	2	16	1	no	free	2	214	Japanese	0.01	digits	1	100	real rir	dummy	fixed	no	various real environments (added)	low	original	00	yes n	to yes
CENSREC-4 (Real) [31]	2008	public spaces, domestic, office, car	2	16	1	no	free	2	10	Japanese	0.01	digits	1	no	reverb	human	quasi-fixed	head	various real environments	low	headset	no	yes n	ao yes
DICIT [32]	2008	domestic	6	48	16	- 2	free		2	Italian	1.7.3	command	4	50	reverb	human	various	head, walk	domestic (scenarized)	medium	headset, tv	yes	yes n	30 yes
SiSEC head-geometry [28]	2008	speech overlap	1.9	16	2	DO.	free	1.9	- 2	N/S	N/S	read	2	full	real rir	loudspeaker	various	80	BO	N/S	original, spatial image	yes	DO B	40 BO
COSINE [33]	2009	dialog	38	48	20	no	free	u	91	US English, non-native English	5	spontaneous	2-7	dialog	reverb	human	various	head, walk	various real environments	low	headset, throat mic	BO	yes p	10 no.
SiSEC real-world noise [28]	2010	public spaces	0.3	16	2-4	no	free	0.3	Ď.	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 n	10 BO
SiSEC dynamic [28]	2010 - 2011	cocktail purty	0.2	16	2-4	no	free	0.2	3.	N/S	N/S	read	2	full (2 at a time)	reverb	loudspeaker	various	simulated	во	N/S	original, spatial image	yes	10 17	10 80
CHEME 1, CHEME 2 Grid [34]	2011 - 2012	domestic	70	16 - 43	2	no	free	12	- 34	UK English	0.05	command	1	no	real rir	dummy	quasi-fixed	simulated head	domestic	low	yes	yes	yes p	00 100
CHEME 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 #	10 10
ETAPE (35)	2012	TV/radio debates, outdoor interviews	42	16	1	1	1	- 32	347	French	16	spontaneous	1 or more	dialog (up to 10% overlap)	reverb (some)	humon	quasi-fixed	head	various real environments	high	110	N/S	yes II	to yes
GALE	2013	TV dialog	120 - 251 per la	ng 16	1.	DO.	3.5 - 7 k\$ per lang	108 - 234 per lang	2 7	Mandarin, Arabic	120	spontaneous	1 or more	dialog	80	buman	quasi-fixed	head	bo	N/S	no	N/S	yes #	00 80
REVERB SimData [36]	2013	domestic, office	25	16	8	no	free given WSJCAM0 (1.75 kS)	25	130	UK English	10	read	1	no	real rir	loudspeaker	various	no	random noise (added)	high	original, spatial image	yes	yes II	so yes
Sheffield Wargames Corpus [37]	2013	cocktail party	7	-48	92	3	free	2	9	UK English	12	spontaneous	4	multiple dialogs	reverb	human	various	head, walk	background music	medium	headset	yes	yes II	10 BO
DIRHA (38)	2014	domestic	Ú.	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 p	to 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:
 - Contemporal Contension Content Content
 - Scalable: just let the robots run
 - Cround truth for signals and location available
 - Transport setup to new environments and reproduce similar collection
 - Coustic 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:
 - Cow-cost: can re-use annotations
 - Scalable: just let the robots run
 - Cround truth for signals and location available
 - Transport setup to new environments and reproduce similar collection
 - Coustic 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



Mounted on 3D-printed structure

© MERL 2015



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	\$3 <i>5</i>	
1 Raspberry Pi Case	\$8	
1 Micro SD Card	\$16	
1 Wifi dongle	\$18	
1 USB-Micro USB Cable	\$3	
Raspberry Pi+Accessories Total		\$8 <i>0</i>
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	Ģ	\$12
6 D Cells + 2 battery cases		\$12
1 3D printing order (~\$5 if self-printing)		\$30
		\$438



MITSUBISHI ELECTRIC RESEARCH LABORATORIES

for a greener tomorrow



3D printing the parts





© MERL 2015



co

Assembling







Et voila!





Meet Dot, Hot and Lot

https://youtu.be/cyTaDiVsPPM



Lot's recording

91	15 30 45 1:00 1:15 1:30
1.0	
0.5-	
0.0-	de viel have deren deren deren deren bei der ihren bleren belle in der and deren der der der der der der einen deren deren deren deren an der einen deren d
0.0	and the second
-0.5-	
-1.0	
1.0	
0.5-	a start to be the total solution of the second start and the second start and the second start second start starts and the second starts an
0.0-	
.0.5.	
10	
1.0	
0.5-	
	all sheep and all and be de les les de se tables s'he se to describ the seles described bais a state of the second se
0.0	
-0.5	
-1.0	
1.0	
0.5-	en an an an an Arana an
0.0-	the state to be and the second state of the se
	the state of the
-0.5	
-1.0	



• Alignment of all audio streams

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



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





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



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



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



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





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





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





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