ICSI’s Efforts on the Altmic Condition

Nikki Mirghafori, Lara Stoll, Andy Hatch, and Howard Lei

With special thanks to:

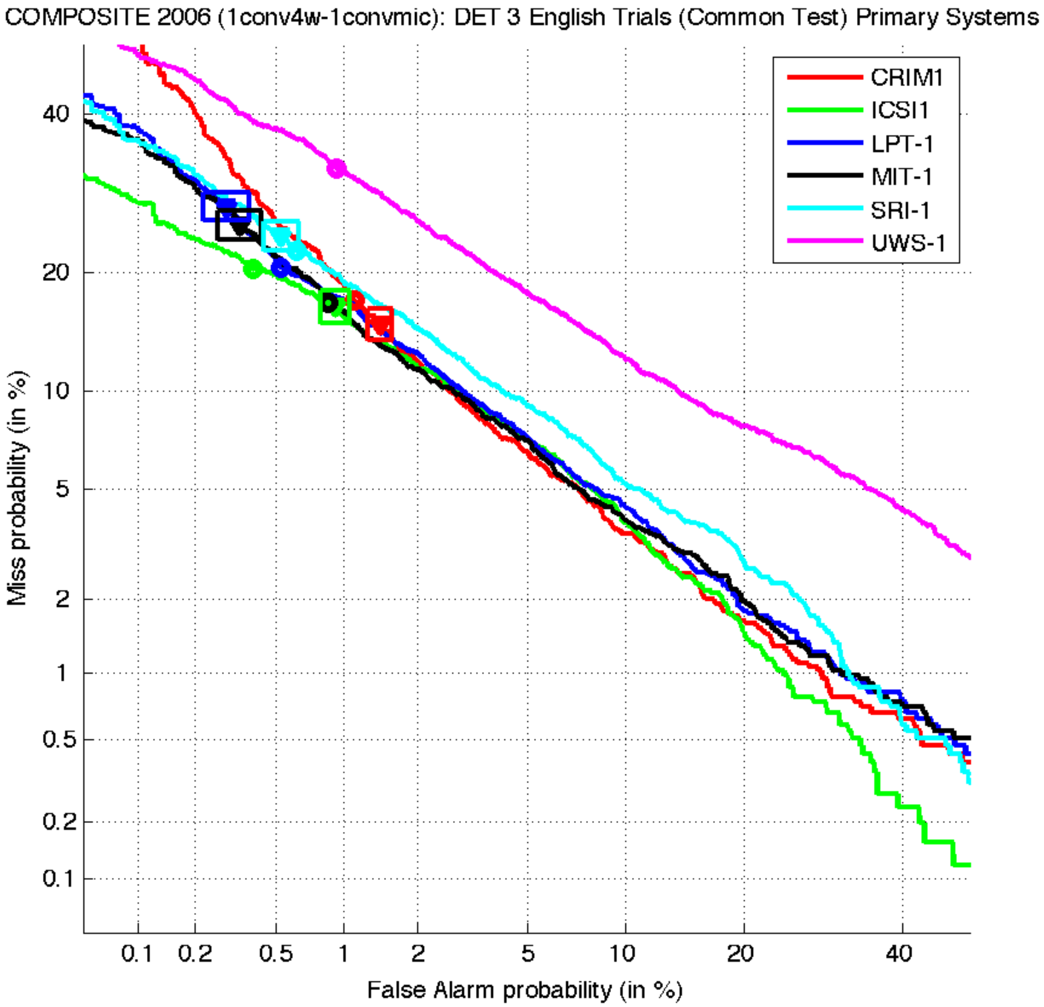
our collaborators at SRI

&

our advisor George Doddington

Presentation can be downloaded from
http://www.icsi.berkeley.edu/~nikki/ICSI2.pdf.gz

We Have a Color!



COMPOSITE 2006 (1conv4w-1convmic): DET 3 English Trials (Common Test) Primary Systems

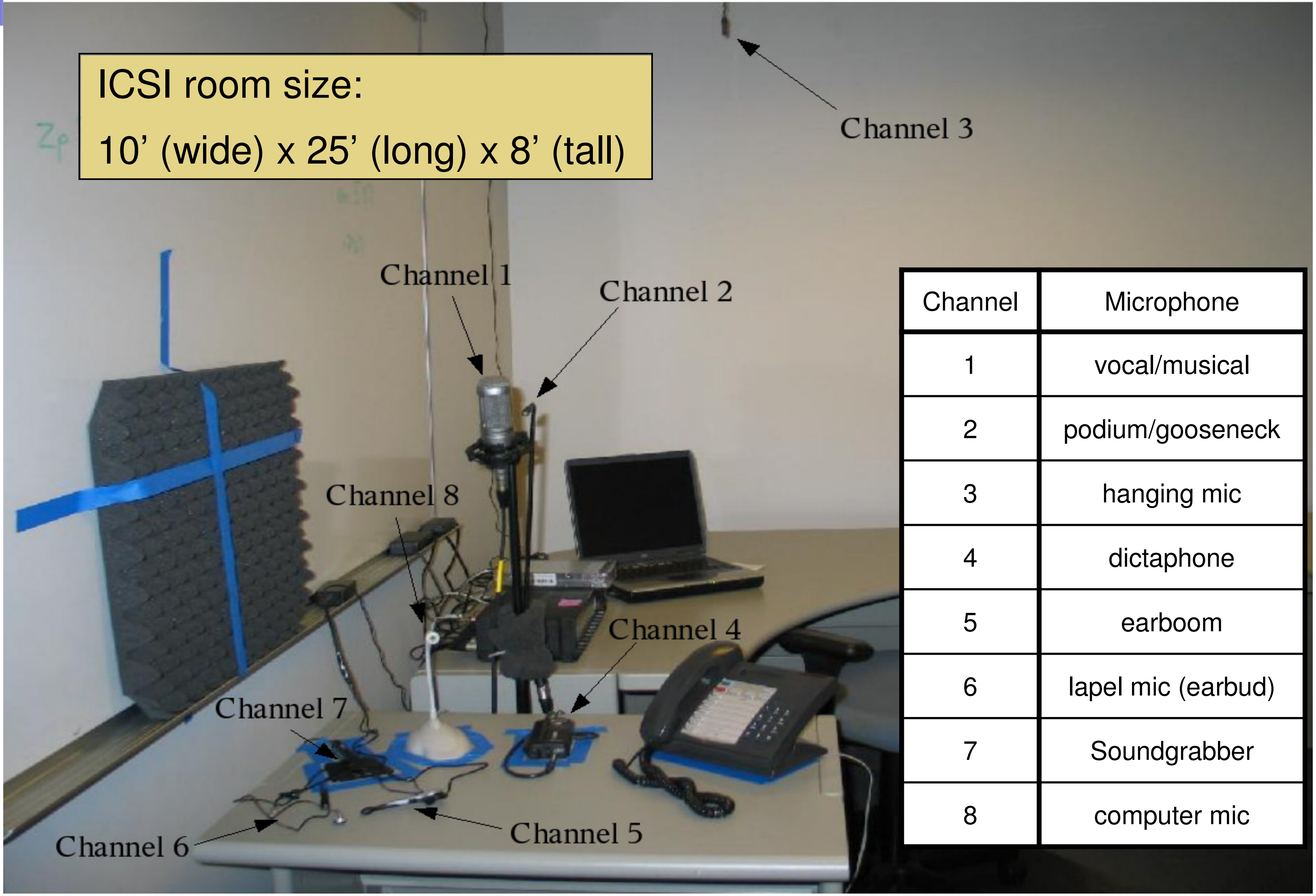
Overview

- Multi-microphone data
  - Microphone types
  - Example setup (ICSI)
- ICSI’s altmic submission
  - Description of individual sub-systems
  - Combination strategy
  - System results and breakdown of individual contributions
- 2005 vs. 2006 – channel and site differences
  - Comparison of system performance
  - Audio samples
- Conclusions

Multi-microphone data

- 8 types of microphones:
  1. Audio Technica AT 3035 – vocal/musical recording
  2. Shure MX418s – podium recording (gooseneck)
  3. Audio Technica AT Pro45 – hanging microphone
  4. Olympus Pearlcorder 725S – dictaphone (microcassette)
  5. Jabra Earboom – cell phone accessory
  6. Motorola Earbud – cell phone accessory (lapel mic)
  7. Crown Soundgrabber II – impromptu recordings
  8. Radio Shack Computer Mic – computer accessory
- Data collected from 3 different sites: LDC (released in SRE05 and SRE06), ISIP (released in SRE05), and ICSI (released in SRE06)
  - Setup instructions were only guidelines – led to variation among recordings from different sites

Example Microphone Setup (ICSI) [1/4]

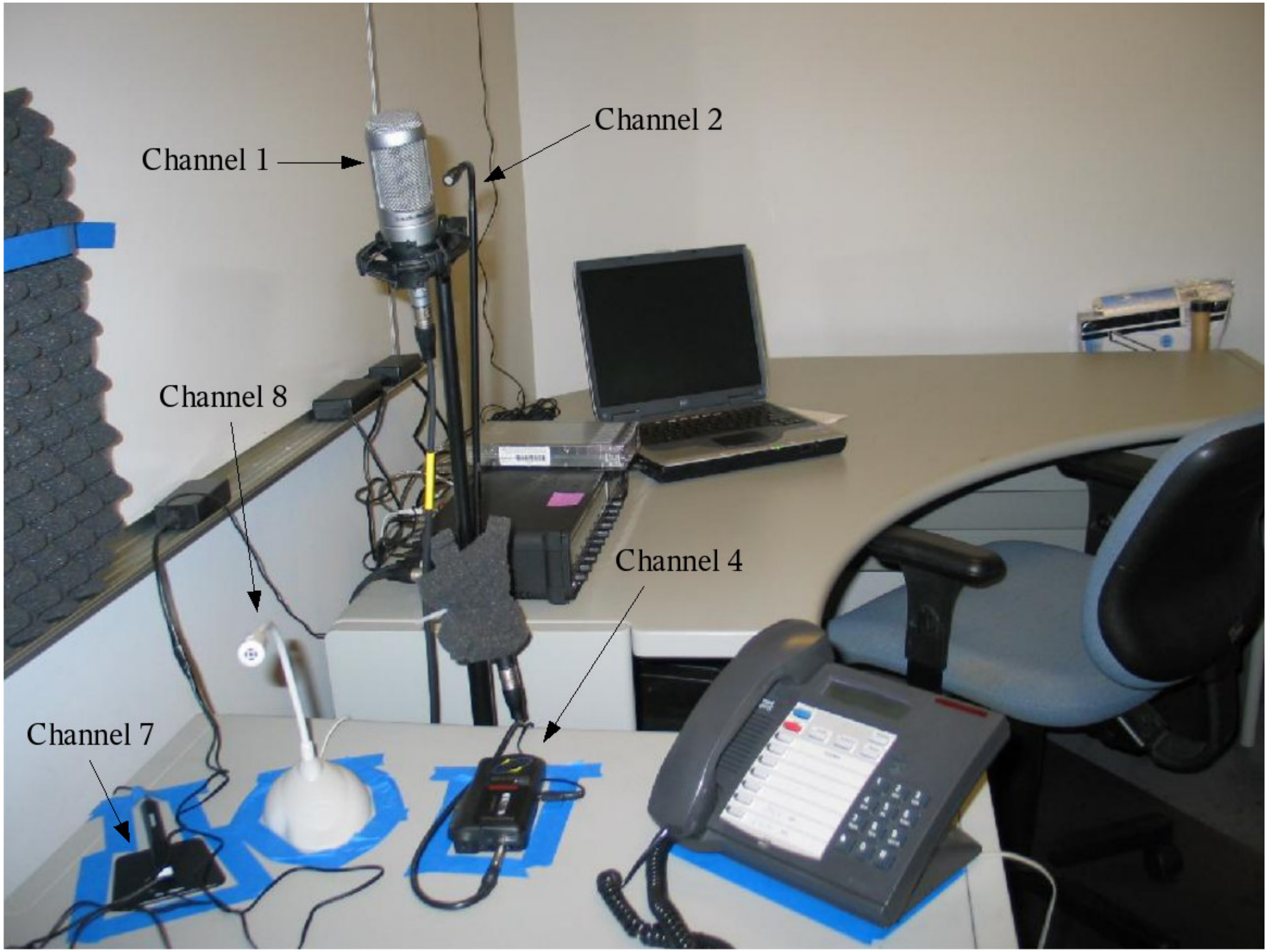


ICSI room size:

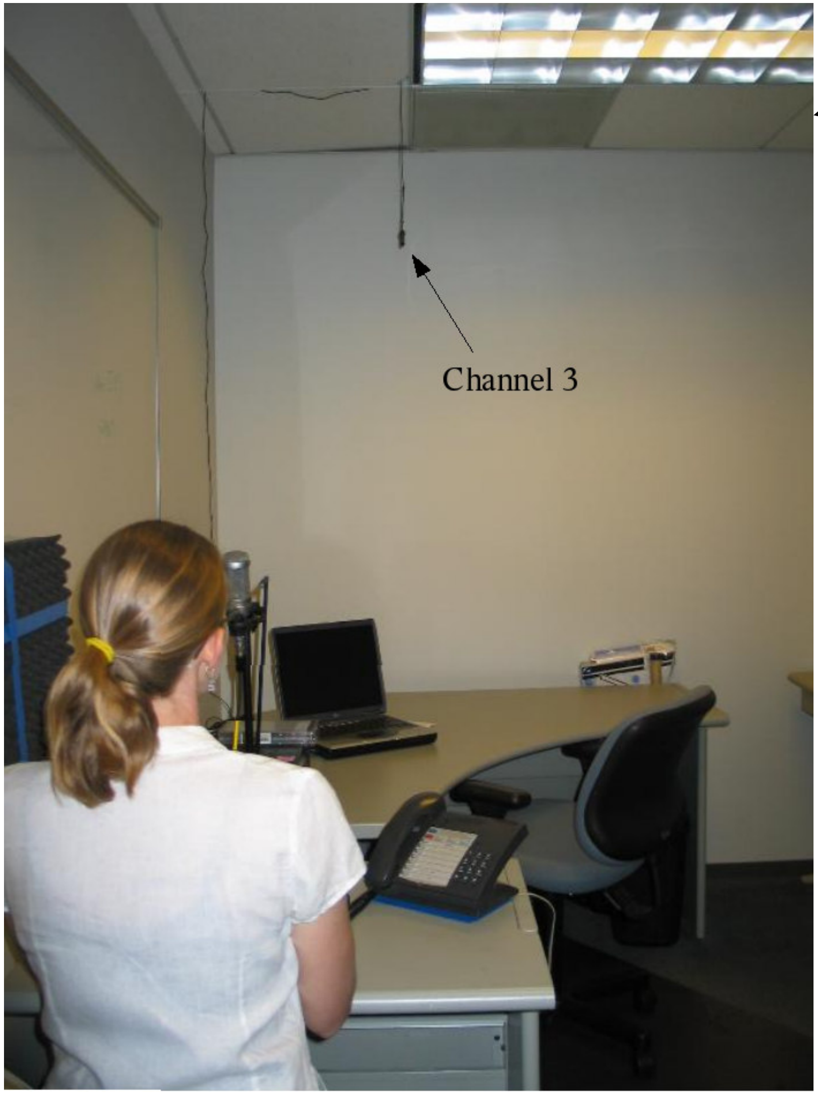
10’ (wide) x 25’ (long) x 8’ (tall)

Channel	Microphone
1	vocal/musical
2	podium/gooseneck
3	hanging mic
4	dictaphone
5	earboom
6	lapel mic (earbud)
7	Soundgrabber
8	computer mic

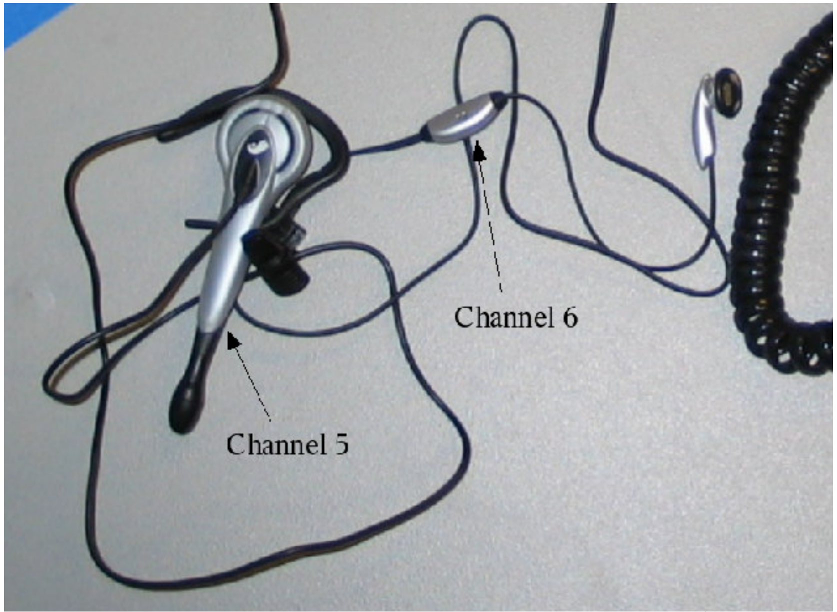
ICSI Microphone Setup [2/4]



ICSI Microphone Setup [3/4]

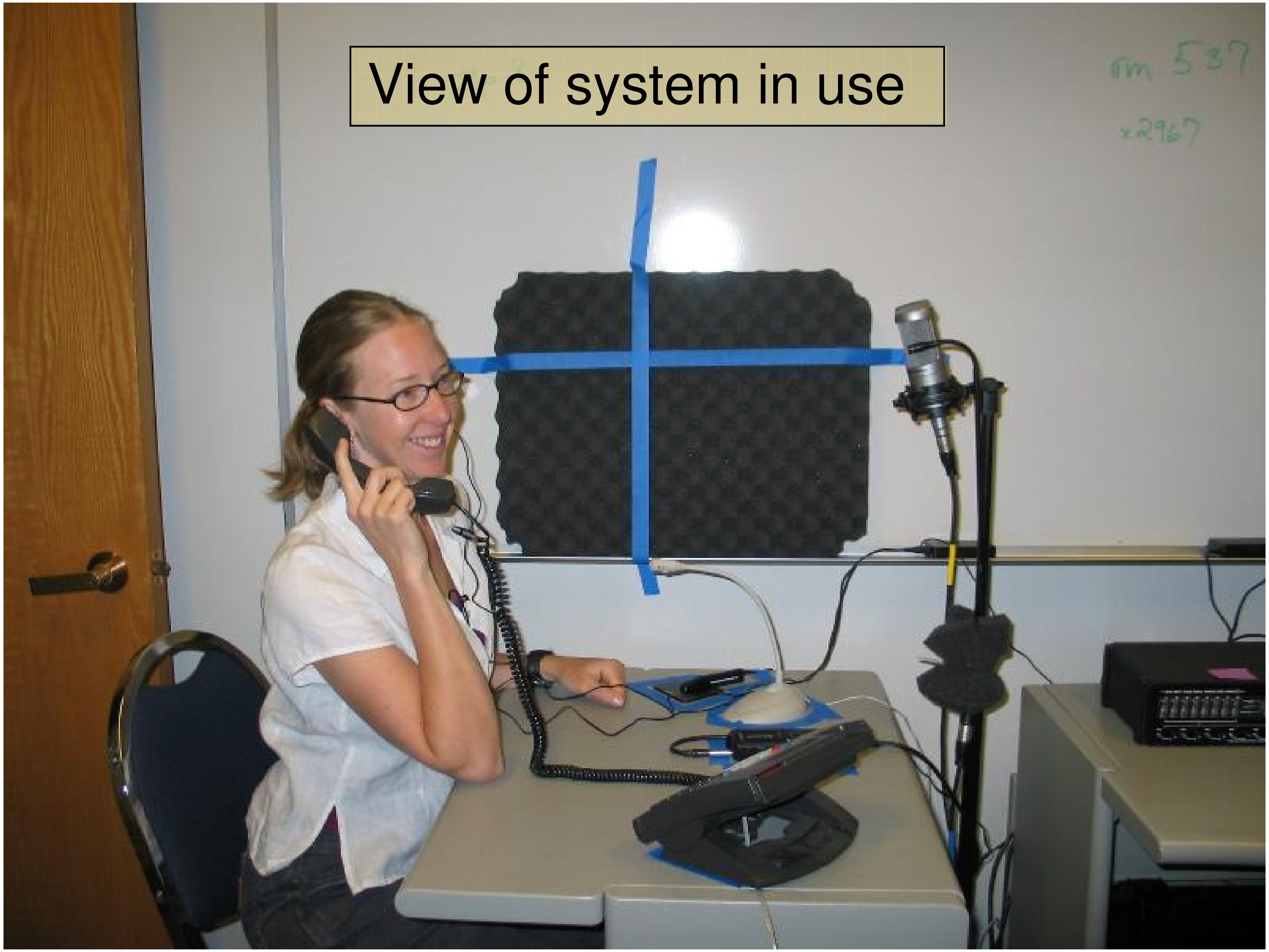


View of hanging mic from speaker location



Close-up view of earboom and earbud/lapel mics

ICSI Microphone Setup [4/4]


View of system in use

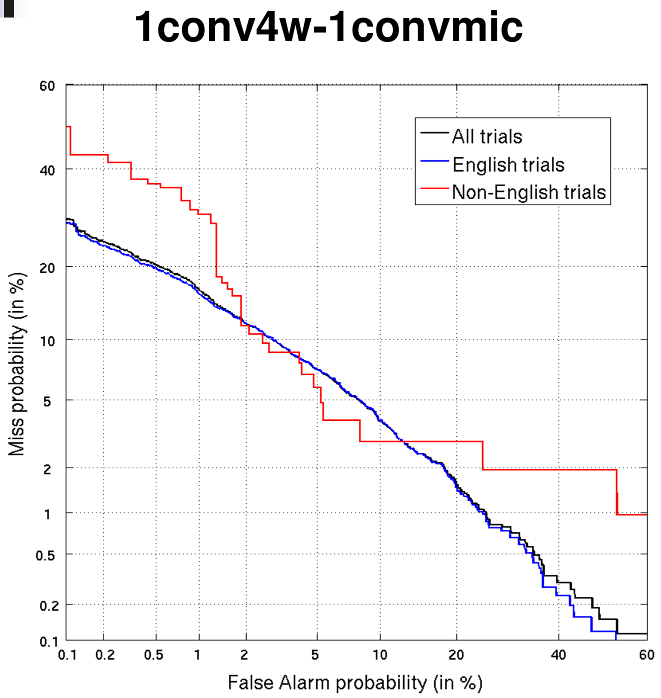
ICSI Altmic System [1/2]

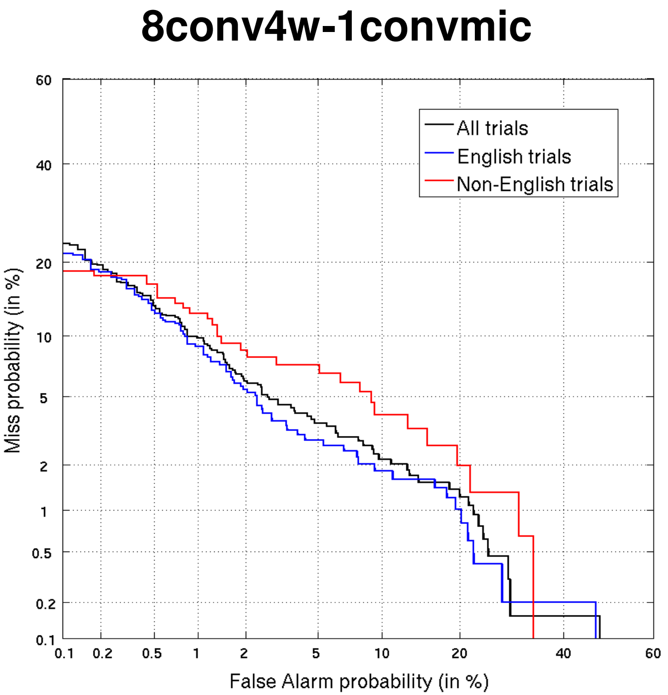
- Same system as for telephone
  - Only difference: applied Wiener filtering before ASR
- Combination of 6 sub-systems for English trials:
  - SRI’s cepstral GMM
  - Keyword conditional HMM (WordHMM)
    - Uses whole-word HMMs for frequent keywords that are rich with speaker characteristic cues (19 total – discourse markers, filled pauses, backchannels)
  - SVM-based Lattice Phone n-grams (PhoneNgram) with WCCN
    - Uses relative frequencies of phone n-grams as features in SVM
  - Word-conditioned phone n-grams with WCCN (WCPhoneNgram)
    - Only considers phone n-grams as conditioned on particular (52 frequently occurring) word unigrams
  - Word-conditioned part-of-speech n-grams (WCPOSNgram)
    - Combination of word n-grams and part-of-speech n-grams
  - Lexical statistics (LexStats)
    - 8 features, measuring speaking rate, number of words, number of conversation turns, number of characters

ICSI Altmic System [2/2]

- Combination of 2 sub-systems for non-English trials:
    - SRI's cepstral GMM
    - Phone n-gram with WCCN
- WCCN: within class covariance normalization
- Used SVM combiner (SVMLite)
    - Classification mode
    - Linear kernel
    - Trained on SRE05 multi-microphone data
        - 8conv4w-1convmic non-English combiner had to be trained on SRE05 8conv4w-1convmic English trials (due to lack of data)
    - Estimated decision threshold using SRE05 multi-microphone data as well

SRE06 ICSI System Results





1conv4w-1convmic trials

23952 English

1018 Non-English

24970 Total

	1-side training			8-side training		
	EER	aDCF	mDCF	EER	aDCF	mDCF
English trials	6.25%	0.257	0.243	3.43%	0.186	0.174
Non-English trials	4.81%	0.414	0.300	6.58%	0.215	0.193
All trials	6.16%	0.263	0.249	4.02%	0.193	0.180

8conv4w-1convmic trials

4620 English

1278 Non-English

5898 Total

Individual ICSI System Contributions

1conv4w-1convmic (English)

Best	GMM	Phone n-gram	WordHMM	WC phone n-gram	Lexical stats	WC POS n-gram	Min DCF
1	X						0.31619
2	X	X					0.27143
3	X	X	X				0.24535
4	X	X	X	X			0.23635
5	X	X	X	X	X		**0.23531**
6	X	X	X	X	X	X	0.24279

Relative improvements

- **+14.2%**
- **+9.6%**
- **+3.7%**
- **+0.4%**
- **-3.2%**

Switched places

8conv4w-1convmic (English)

Best	GMM	WordHMM	WC POS n-gram	WC phone n-gram	Lexical stats	Phone n-gram	Min DCF
1	X						0.25028
2	X	X					0.19851
3	X	X	X				0.18109
4	X	X	X	X			**0.16771**
5	X	X	X	X	X		0.17049
6	X	X	X	X	X	X	0.17491

Relative improvements

- **+20.7%**
- **+8.8%**
- **+7.4%**
- **-3.8%**
- **-2.6%**

Word conditioning helps more on 8 side training

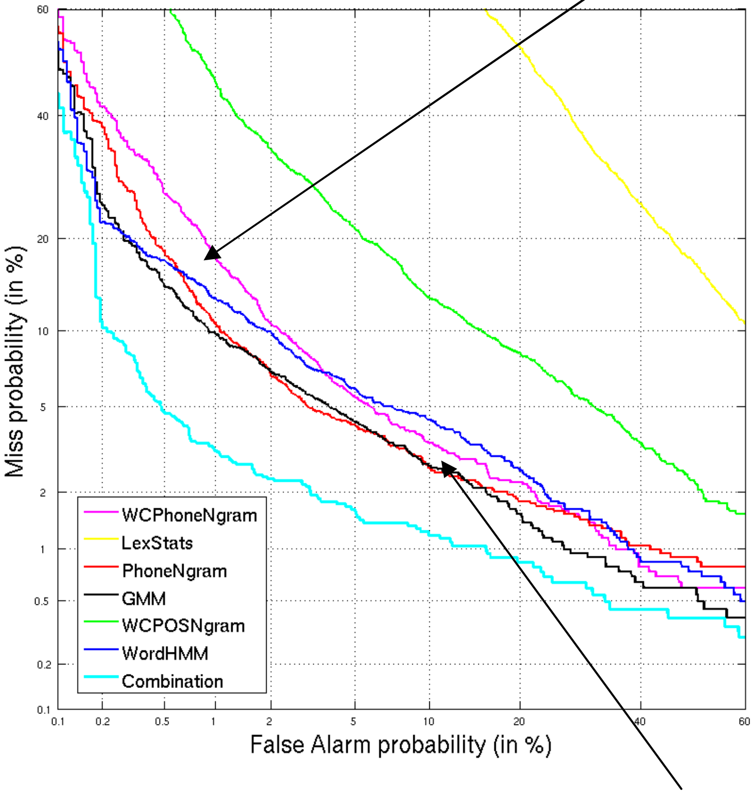
ICSI Systems – Altmic vs. Telephone

- For 1 side training, individual systems perform similarly for telephone and multi-mic
- Not so for 8 sides
  - WordHMM is more channel robust than both phone n-gram systems
    - Open-loop phone recognition more sensitive to channel variation
    - Word conditioning appears to provide a constraint that improves channel robustness

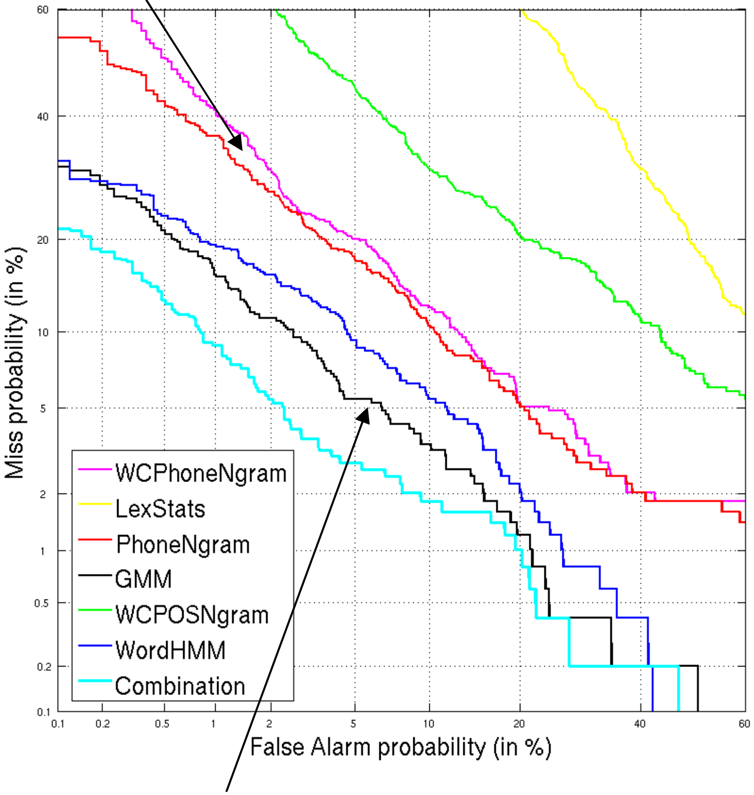
8conv4w – Telephone vs. Altmic

Gap between phone n-gram (red) and WC phone n-gram (magenta) narrows for altmic

8 sides - Telephone



8 sides - Altmic

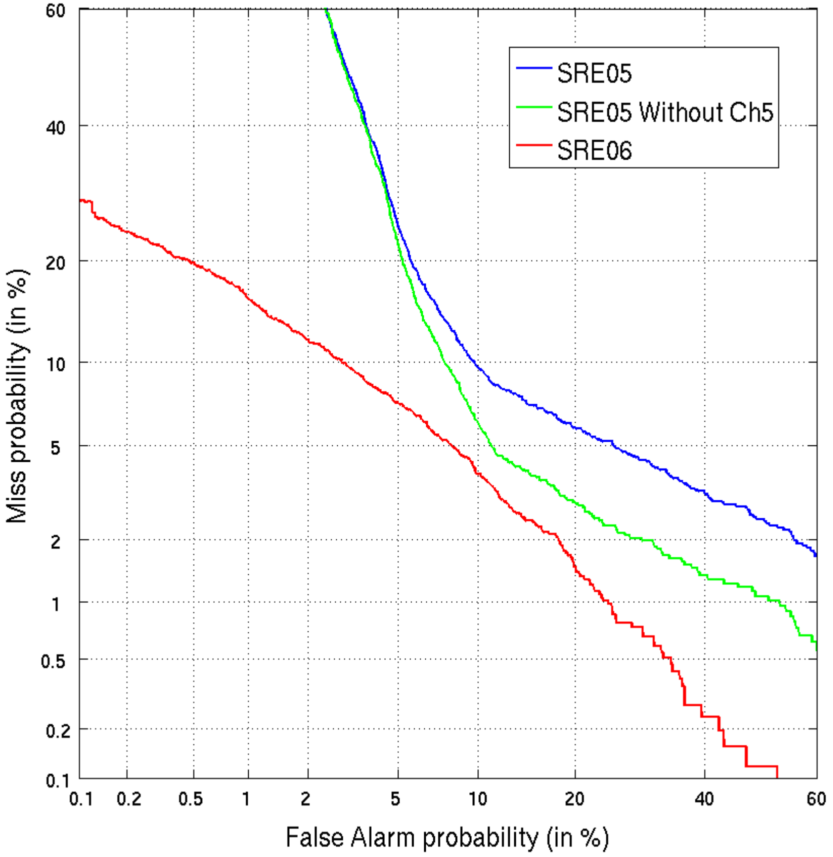


WordHMM (blue) and GMM (black) better than phone n-gram systems (red and magenta) for altmic

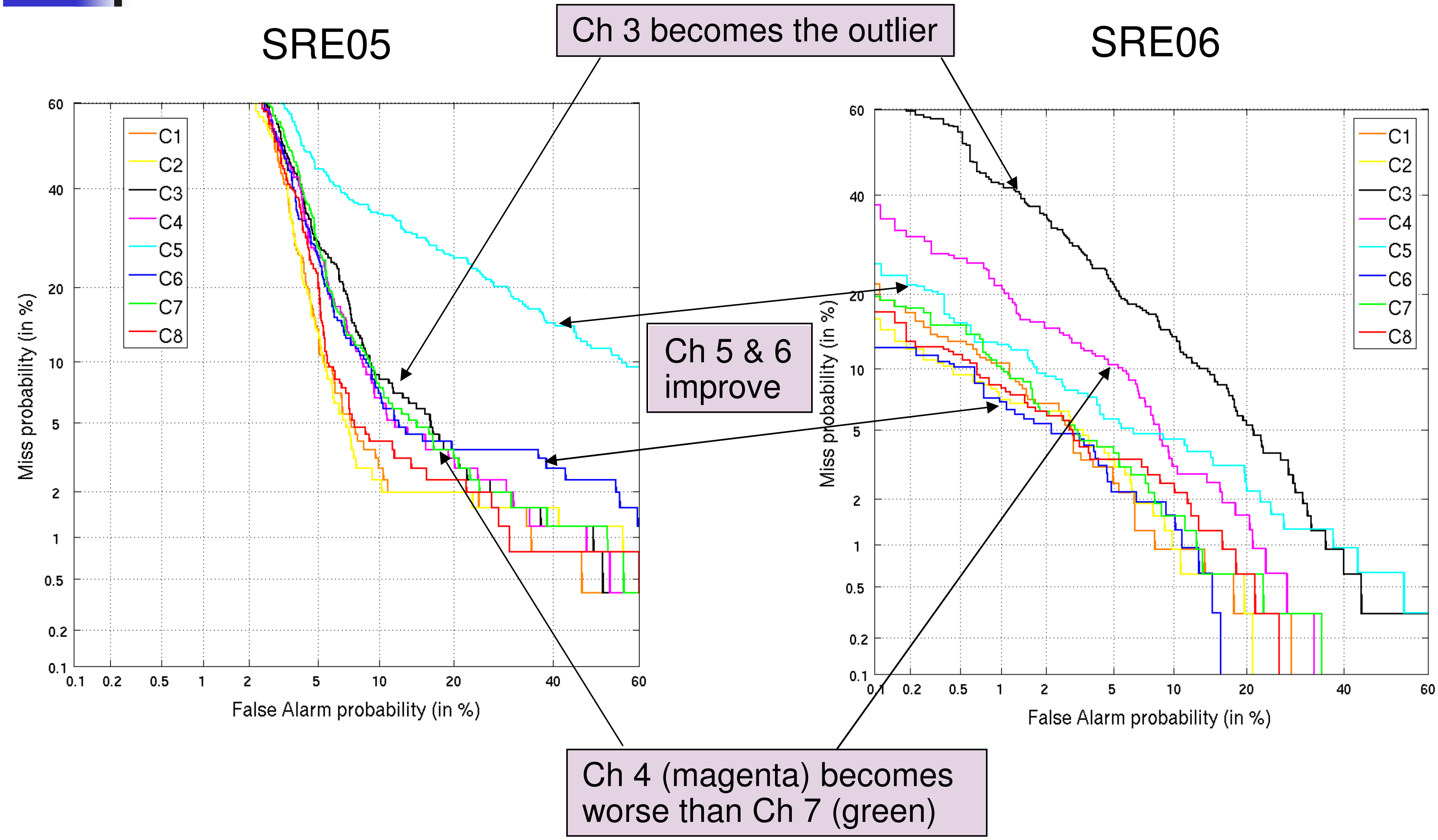
SRE05 vs. SRE06 Multi-mic

- SRE06 performance is much better than SRE05
- Still true even without the corrupted SRE05 channel 5
- Difference could be due to recording site variations
    - Same type of microphone can sound different: e.g., distance of lapel mic from speaker's mouth

1 side training



Performance by Channel



SRE05

SRE06

Ch 3 becomes the outlier

Ch 5 & 6 improve

Ch 4 (magenta) becomes worse than Ch 7 (green)

Audio samples

- Multi-mic website

Channel 1:	LDC	🔈	ISIP	🔈
Channel 5:	ICSI	🔈	LDC	🔈
Channel 6:	ISIP	🔈	LDC	🔈

Conclusions [1/2]

- Multi-microphone data can vary greatly
  - Same channel from different recording sites has different acoustic characteristics
  - Due to, e.g., room reverberation, changes in microphone placement (relative to speaker), room noise
- Looking to the future –
  - LDC should give better specifications for multi-microphone data collections – make rules instead of guidelines
  - Recording sites should better document their setups
    - Might want to include impulse response or reverberation response measurements

Conclusions [2/2]

- Data collection differences make progress tracking difficult
  - Better controlled data collection would limit variations in system performance
- Implicit expectation is that channels would behave similarly across sites
- The desirable variability is the difference between channels, not recording site variability for the same channel

- ICSI goal: to develop features or systems that are robust to channel differences, due to both microphone types and room acoustics
  - Continue to utilize word conditioning as a means of improving multi-mic performance, especially for 8 side training

Thank You!

- Any questions or comments?