Speaker Variability as a Source of Error in Forensic Speaker ID

Finnian Kelly & John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

http://crss.utdallas.edu (Finnian.Kelly,John.Hansen)@utdallas.edu

NIST Forensic Science Error Management

July 20-24, 2015

Outline

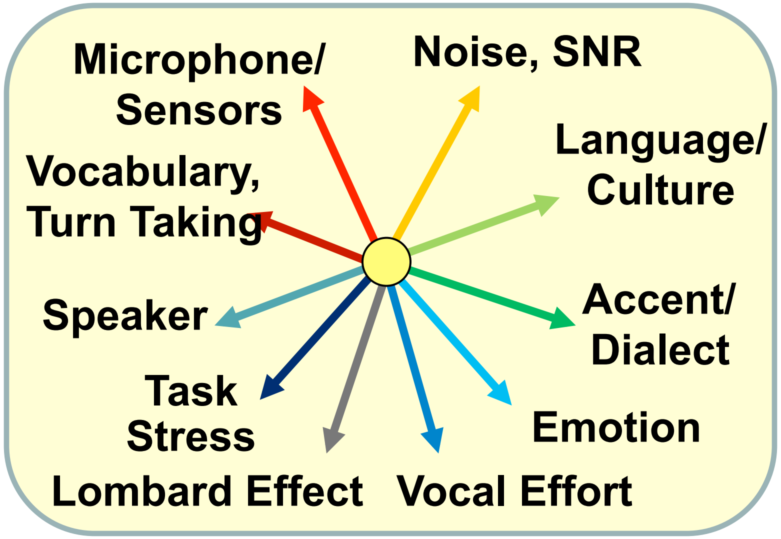
- Sources of Variability in Speech and Speaker for SID
- **Sample Research Efforts:**
  - Vocal Effort & Whisper Speech and SID
  - Lombard Effect "Flavors" and SID
  - Prof-Life-Log: naturalistic longitudinal speech variability
- **In-Depth:** Longitudinal & Aging for SID / Voice Forensics
- Summary & Conclusions

Speech Production: Variability

SPEAKER:

- Task Stress
- Emotion
- Vocal Effort
- Accent/Dialect
- Speaker

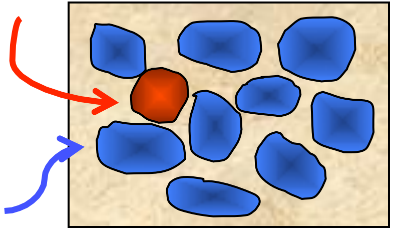
Neutral	Stress
🔊	🔊
🔊	🔊
🔊	🔊



ENVIRONMENT:

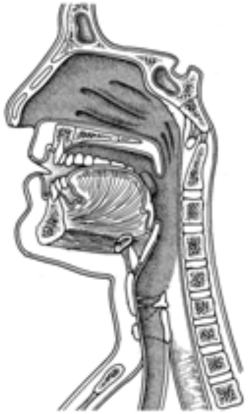
- Noise
- Mic/Envir ext
- Style: clo
  rea us
- “Ground-
- Simulate

Variability – within speaker



Variability – across speakers

Speaker Based



Technology Based

Conversation Based

Human-to-Human
Human-to-Machine

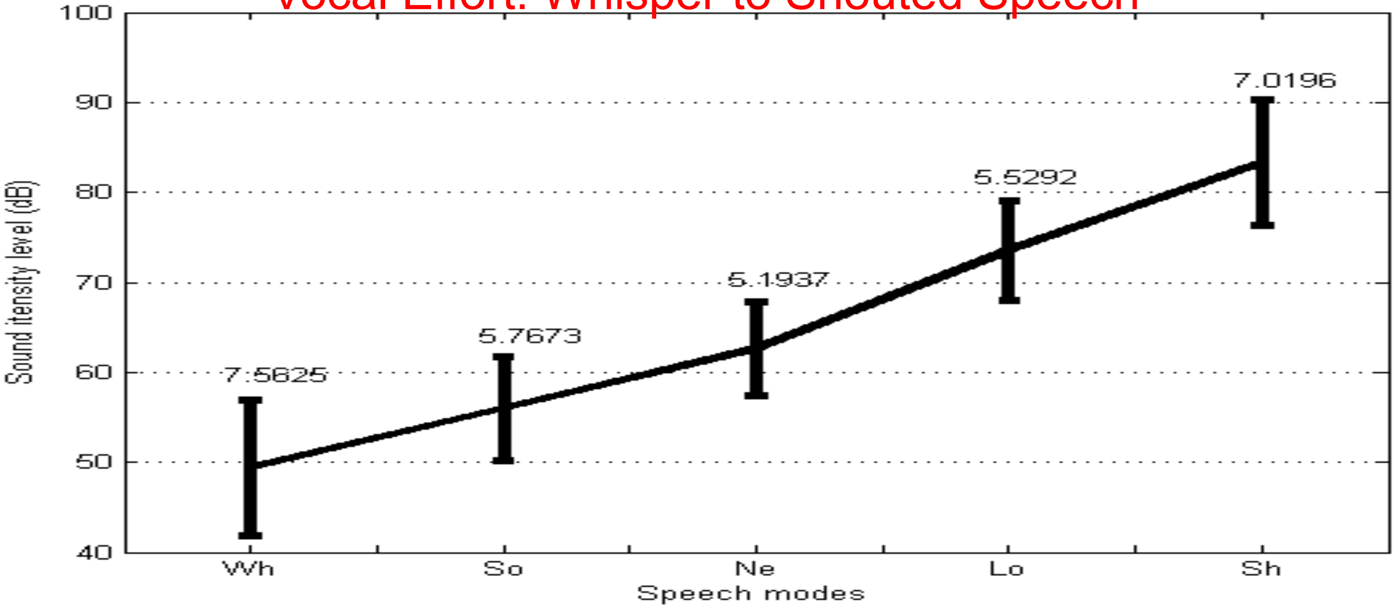
Prompted/Read Speech
Spontaneous Speech

Monologue
2-way conversation
Group Discussion

Vocal Effort Speech Analysis:

Sound Intensity Level (SIL)

Vocal Effort: Whisper to Shouted Speech



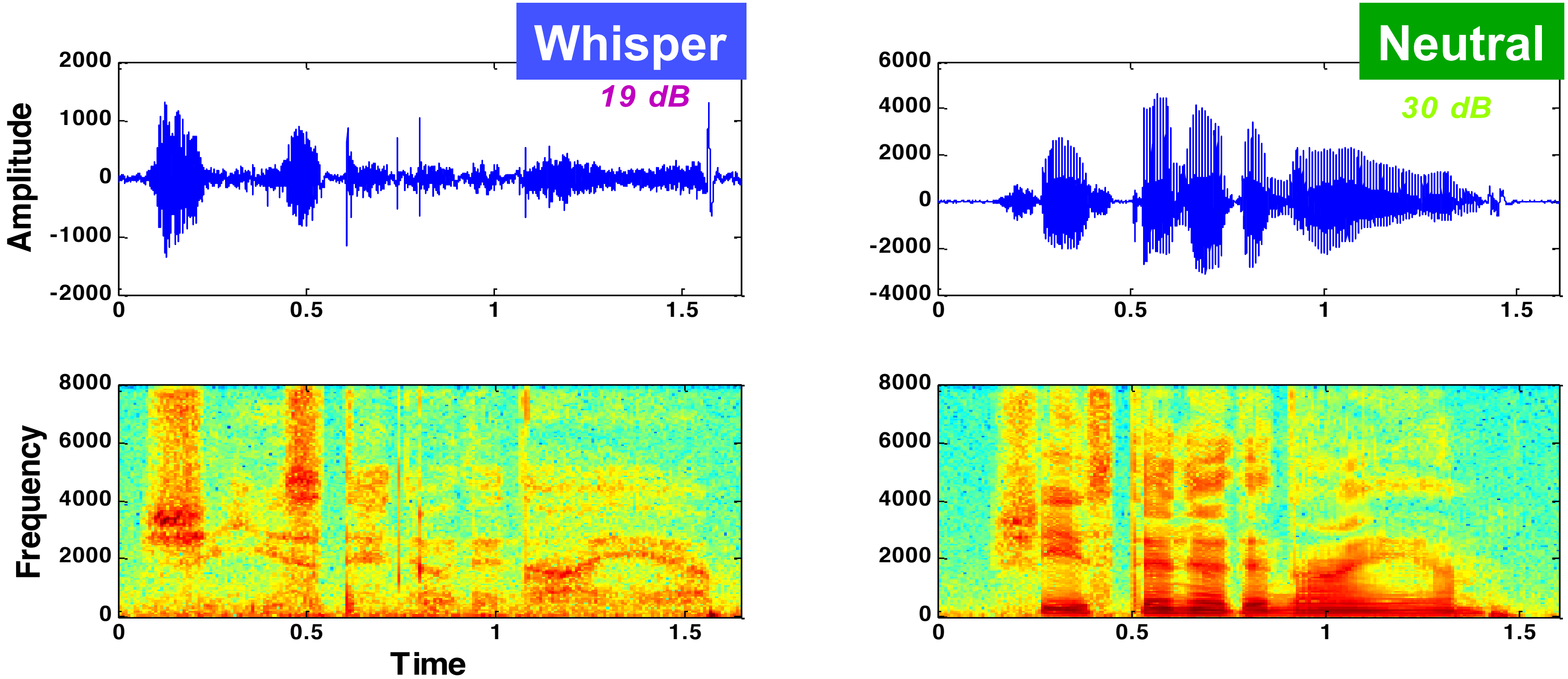
Whisper **Soft** **Neutral** **Loud** **Shout**

Mean and standard deviation of sound intensity level of sentences under five speech modes.

- Increasing SIL: speech mode changes from whispered to shouted.
- Standard deviation of SIL in five speech modes indicate that variation of SIL in neutral mode is lower than that in the other four speech modes

Whisper Speech

- Problem: Whisper - alternative speech production presents unique challenges to speaker ID systems
  - *Absence of periodic excitation (F0) and existence of formant shifts*
  - *Reduced signal energy*



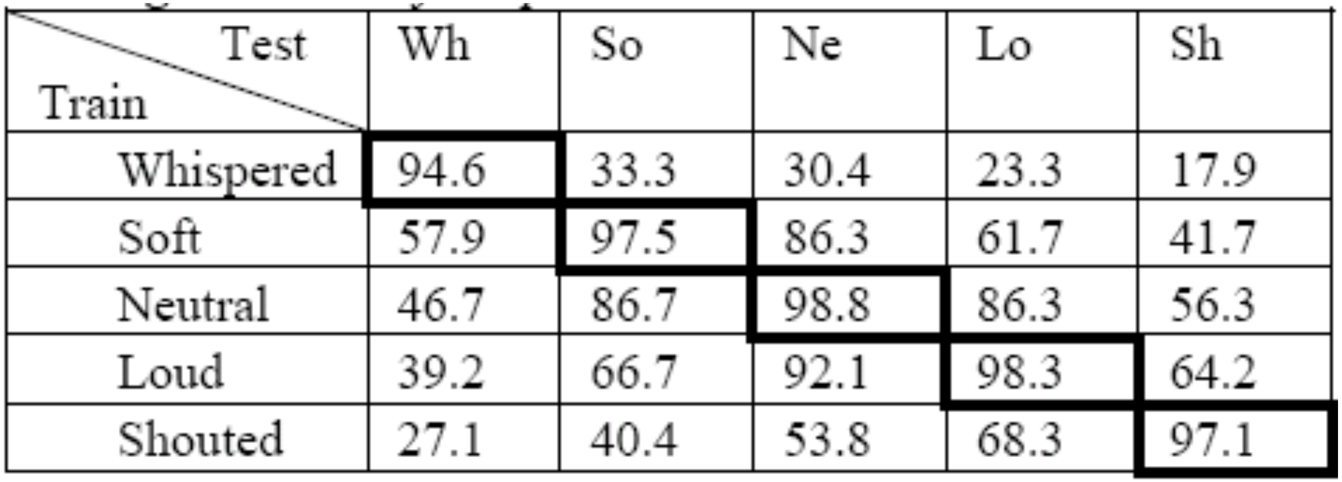
[1] X. Fan, J.H.L. Hansen, "Speaker Identification within Whispered Speech Audio Streams," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19(5), pp. 1408-1421, July 2011.

[2] C. Zhang, J.H.L. Hansen, "Whisper-Island Detection Based on Unsupervised Segmentation with Entropy-Based Speech Feature Processing," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19(4), pp. 883-894, May 2011.

Vocal Effort: Impact on In-Set Speaker ID

(Accuracy: In-Set/Out-of-Set SID)

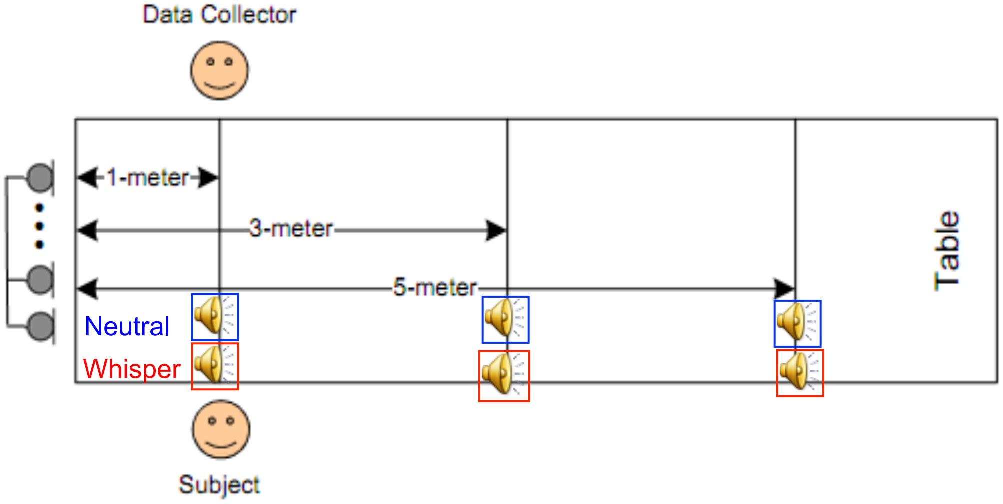
- Speaker ID Systems & Vocal Effort Impact:
  - In-Set Speaker ID System; GMM based with UBM; MAP adaptation using In-Set speaker data (110 sentences); MFCC; 10-12 sec train per spkr, ~8 sec test per speaker



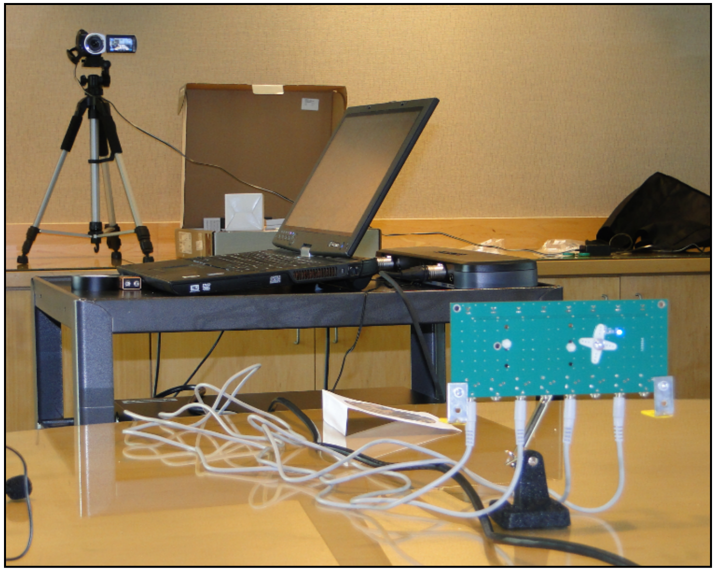
Test Train	Wh	So	Ne	Lo	Sh
Whispered	94.6	33.3	30.4	23.3	17.9
Soft	57.9	97.5	86.3	61.7	41.7
Neutral	46.7	86.7	98.8	86.3	56.3
Loud	39.2	66.7	92.1	98.3	64.2
Shouted	27.1	40.4	53.8	68.3	97.1

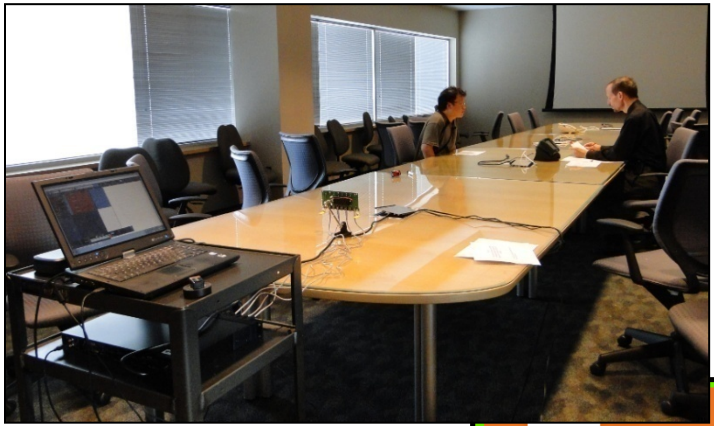
- Matched Vocal Effort conditions: In-Set Spkr ID performance is good
- Significant Reduction for In-Set Spkr ID for mismatched conditions

Distance Speech & Whisper



Room Setup for Data Collection

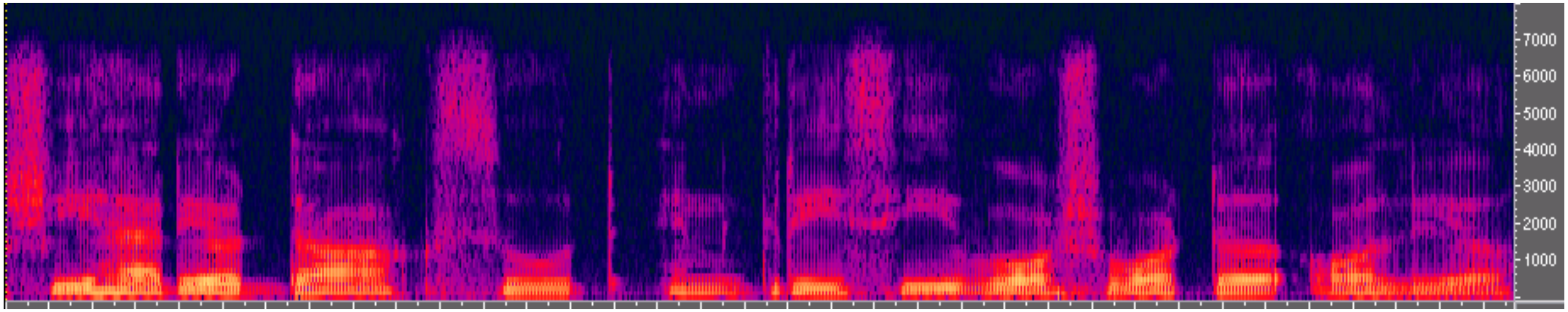




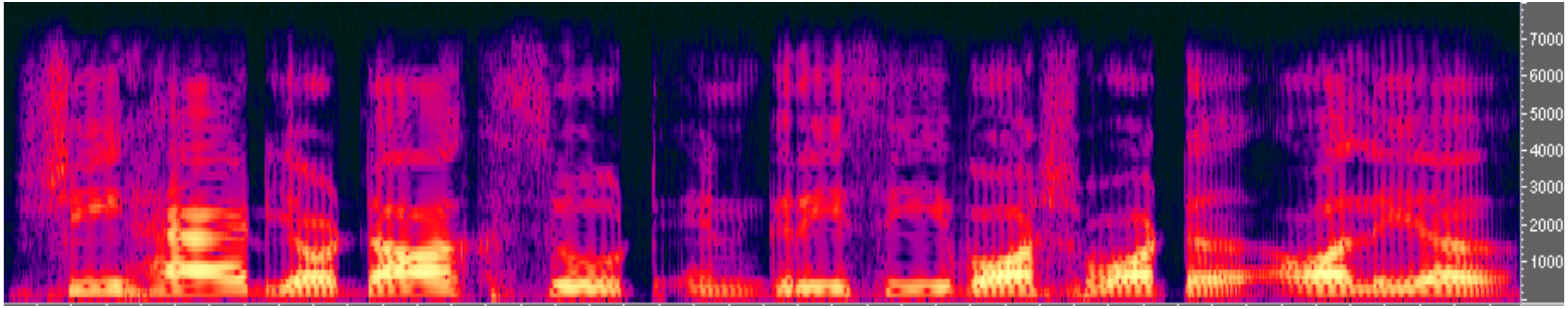
Speech Under Lombard Effect

- Lombard Effect: speech produced in noise
- Automatic & Perceptual experiments show "flavors" of LE based on Noise type and level

Male – Neutral



Male – Lombard



J.H.L. Hansen, V.S.Varadarajan, "Analysis and Normalization of Lombard Speech under different types and levels of noise with application to In-Set/Out-of-Set Speaker Recognition, *IEEE Trans. Audio, Speech & Language Processing,* vol. 17, no. 2, pp. 366-378, Feb. 2009

In-Set Speaker ID: Lombard Effect

- Use Clean 3 & 12 sec Test Tokens (9 Lombard conditions)
- EER improved for neutral but *NOT* Lombard speech

Equal error rate (%) for 3 sec clean test tokens. Neutral EER: 14.67%

Noise Type	Noise Level 1	Noise Level 2	Noise Level 3
HWY	23.16	32.67	34.83
LCR	25.83	29.5	30.33
PNK	22.17	25	31.5

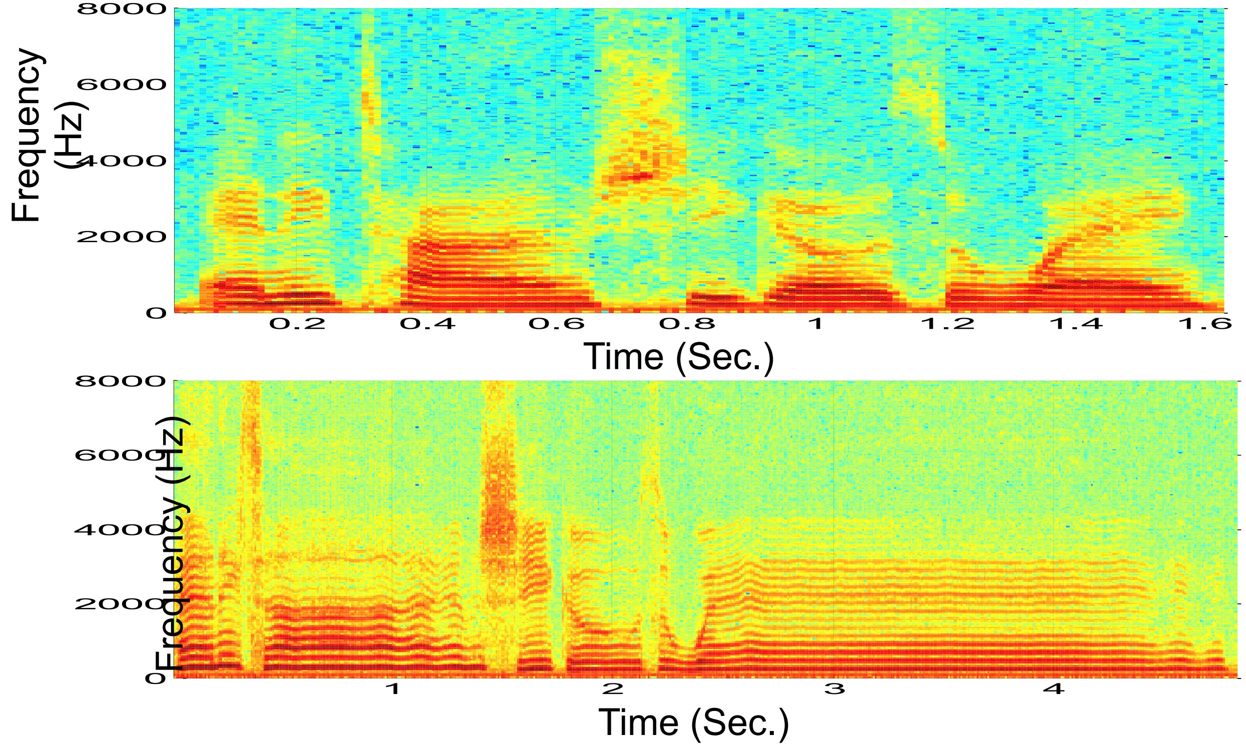
Equal error rate (%) for 12 sec clean test tokens. Neutral EER: 7.2%

Noise Type	Noise Level 1	Noise Level 2	Noise Level 3
HWY	20.0	29.5	34.0
LCR	24.5	30.17	28.83
PNK	16.8	22.16	31.5

Singing vs. Speaking

Speaking

Singing



- Reading and singing spectrograms of "Any time she goes away"
- Corpus: UT-Sing Corpus – 81 subjects (4 languages).

➡ *Significant harmonic structure in singing vs. reading.*

GMM Based Closed-Set Speaker ID Results

Singing & Speaking: Hindi & Mandarin Speakers

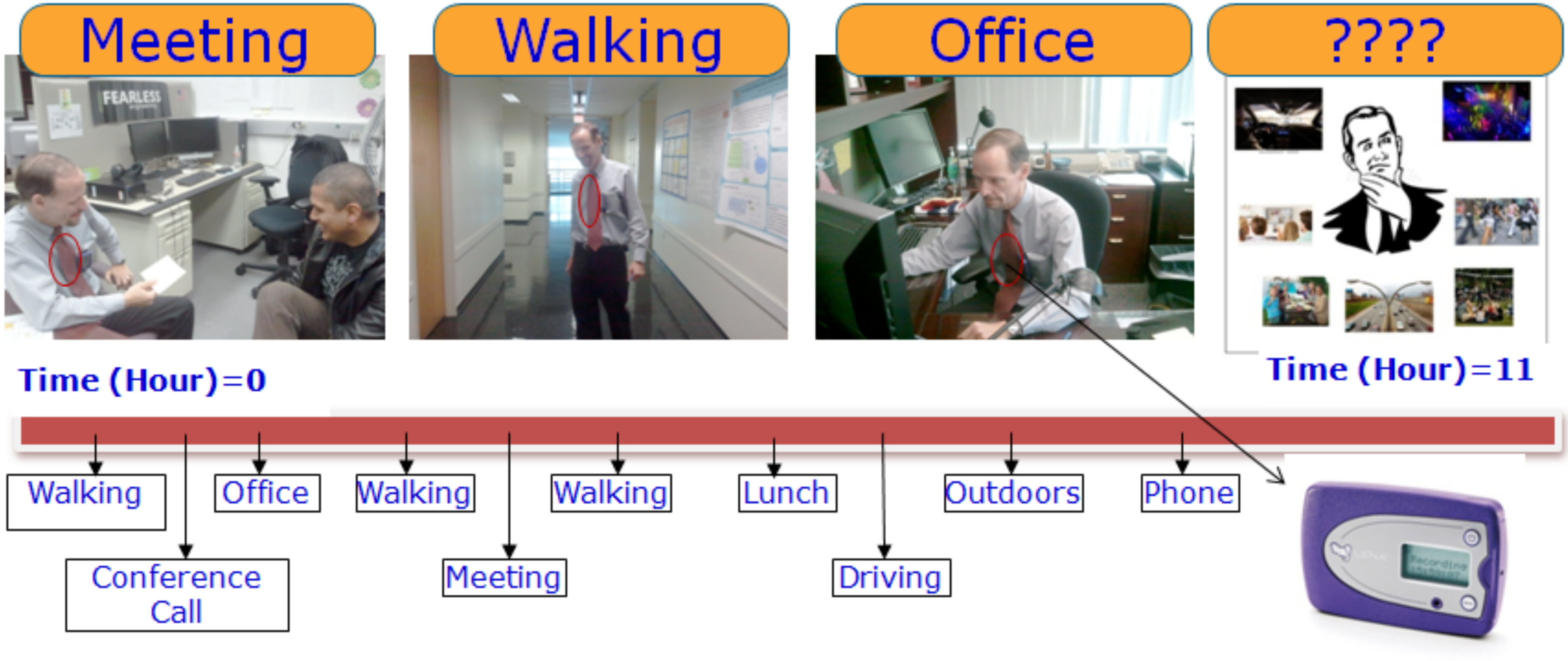
Train	Test	Hindi Accuracy	Mandarin Accuracy
Spoken	Spoken	100%	100%
Singing	Singing	96.3%	95.7%
Spoken	Singing	32.6%	38.5%
Singing	Spoken	63.7%	69.6%

⇨ *Profound difference in Closed-Set Speaker ID with Train/Test mismatch*

(Note: Singing only contains speech (i.e., no music))

Prof-Life-Log: monitoring human interactions

Prof-Life-Log: Corpus Development



Meeting

Walking

Office

????

Time (Hour)=0

Time (Hour)=11

Walking

Conference Call

Office

Walking

Meeting

Walking

Lunch

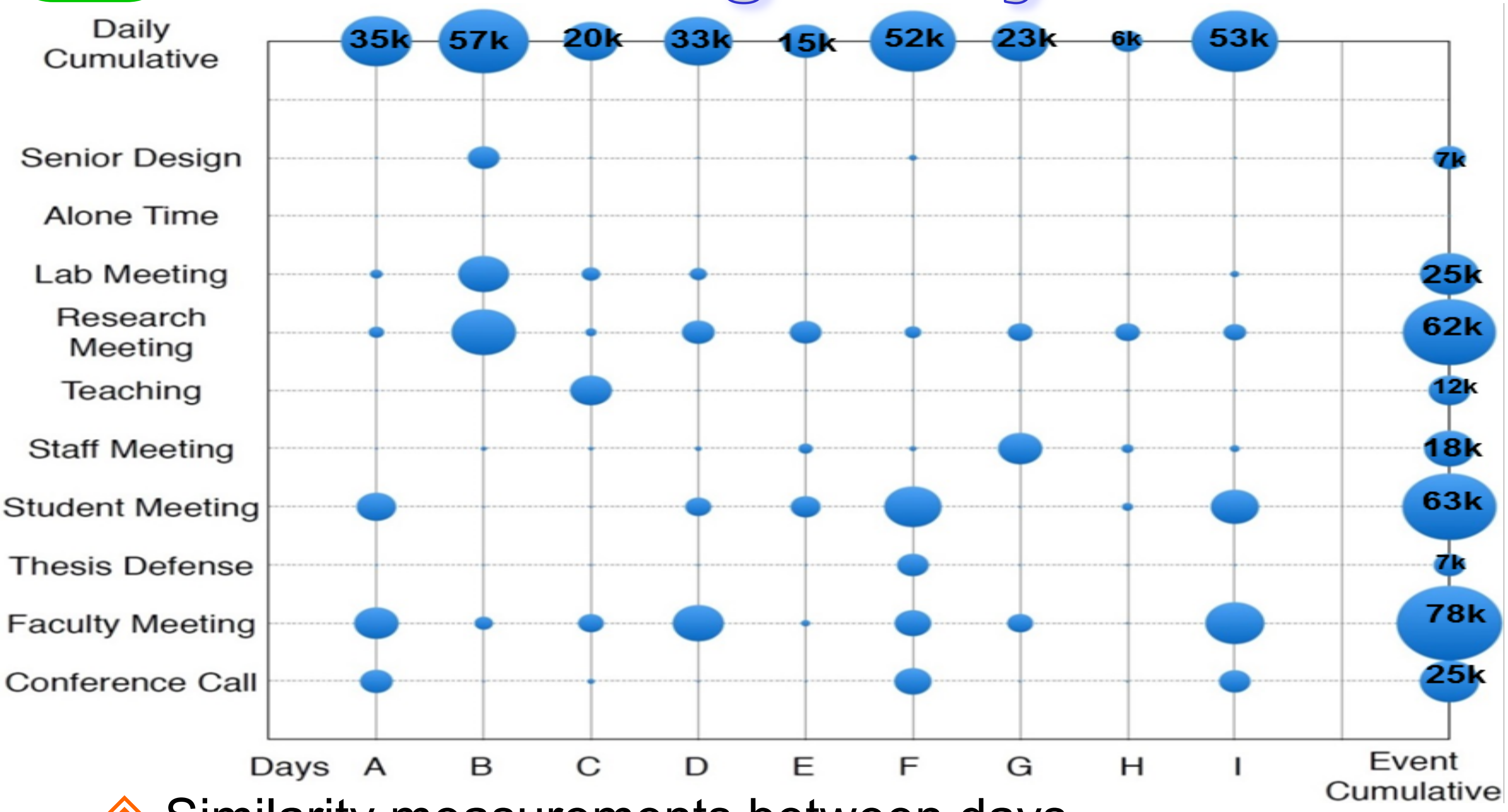
Driving

Outdoors

Phone

- ◈Unscripted speech collection in natural environments
- ◈Unrestricted topics, vocabulary and language use
- ◈Good for: Co-Speaker research; Diarization; SID; KWS

Prof-Life-Log Daily Word Count

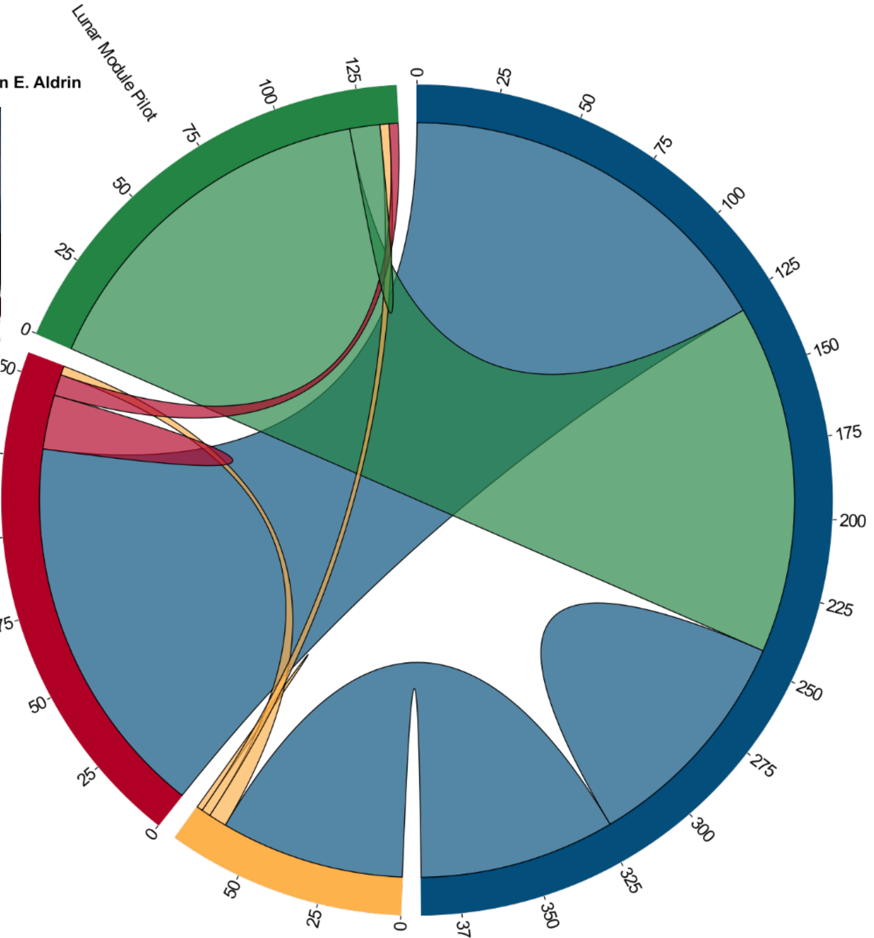


- Similarity measurements between days
- Most similar days: A and I → Correlation =0.87
- Most diverse days: B and F → Correlation =0.27

Apollo-11: Who's Talking to Whom?

Lunar Module Pilot

Edwin E. Aldrin



Command Module Pilot

Michael Collins

Commander

Neil A. Armstrong

Capsule Communicator

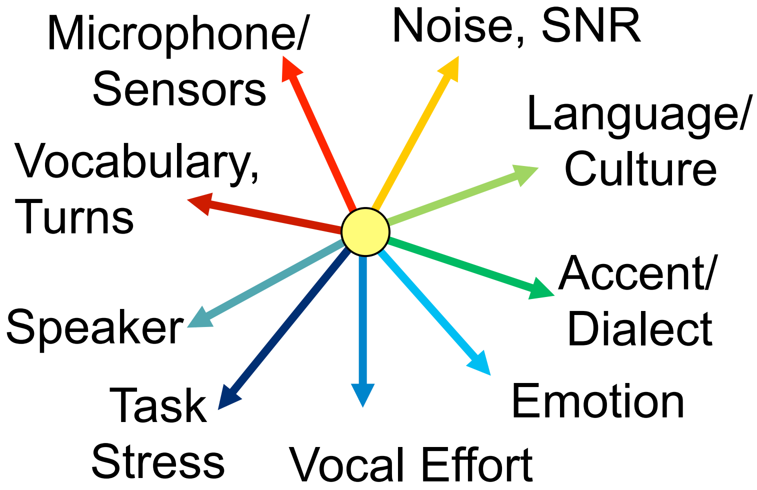
Bruce McCandless

Parameter	Value
Conversation Count	10
Word Count	1050
# Turns Taken	60
Topic of discussion	Lift-off

Parameter	Value
Sentiment	Positive
Emotion	Positive
Stress Levels	High

Speaker Traits & Characteristics

- Speaker Modeling over the Mission
- Aging process of the speakers



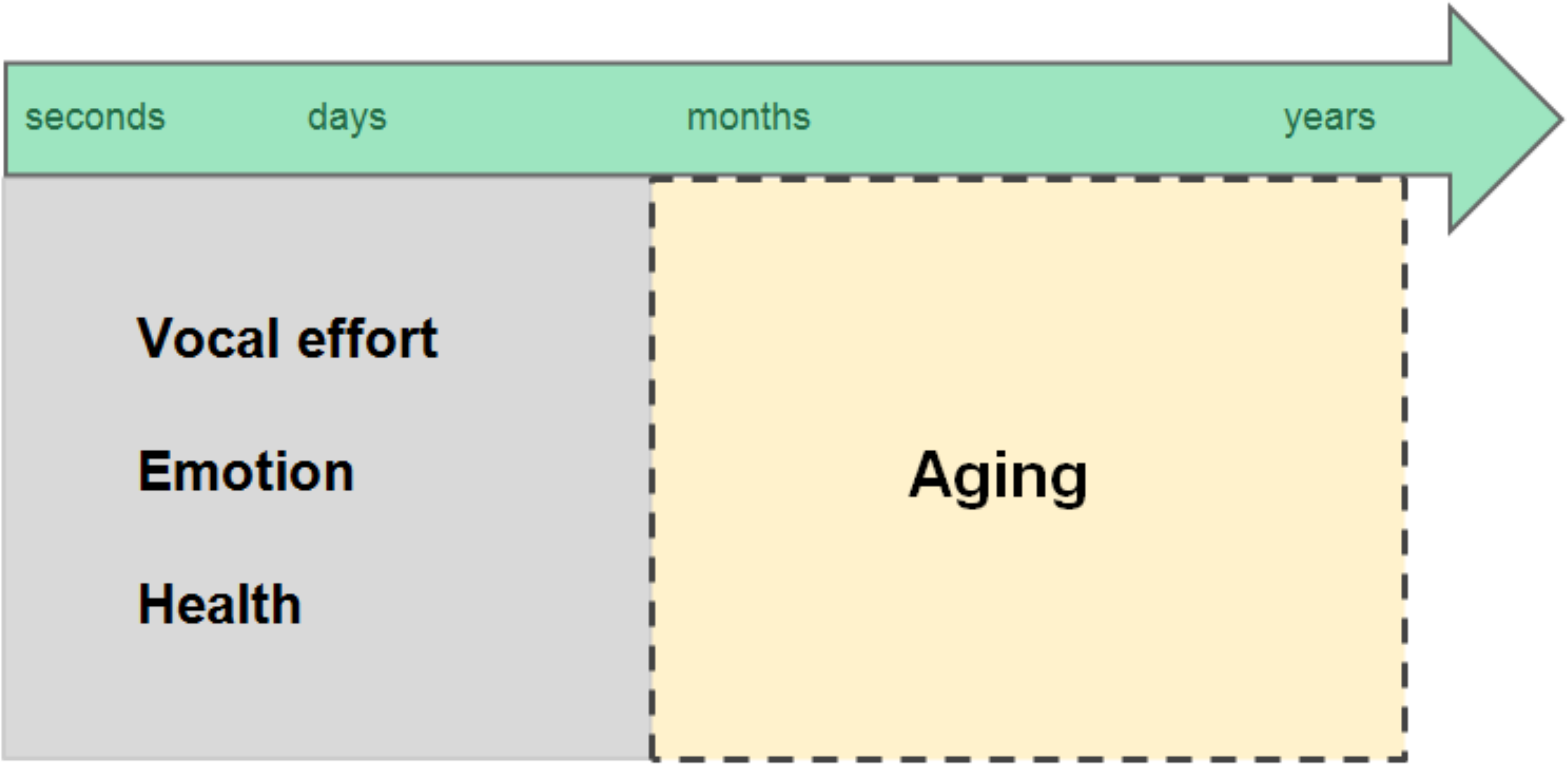
Allen Bean (Apollo 12)

Harrison Schmitt (Apollo 17)

Eugene Cernan (Apollo 10)

Variability in Speaker Recognition

Finnian Kelly



seconds days months years

Vocal effort

Emotion

Health

Aging

Longitudinal Speech Corpora

- RedDots [Lee15]: 1 year range (weekly), 45+ speakers
- MARP [Lawson09] : 3 year range (2 month interval), 73 speakers
- Greybeard [Brandschain10] : 2-14 year range, 172 speakers
- Up Series [Rhodes13]: 7-28 year range, 8 speakers
- TCDSA [Kelly13] : 1-58 year range, 26 speakers

[Lee15] K.A. Lee et al., "The RedDots Data Collection for Speaker Recognition," to appear at InterSpeech 2015, Dresden, Germany, September
[Lawson09] A. D. Lawson, A. R. Stauffer, E. J. Cupples, W. S.J., W. P. Bray, J. Grieco, "The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial Findings," ISCA InterSpeech-09, Brighton, 2009.
[Brandschain10] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Greybeard – Voice and Aging," *7th Conf. on Inter. Language Resources and Evaluation (LREC '10)*, Valletta, Malta, 2010.
[Rhodes13], R. Rhodes, "Assessing non-contemporaneous forensic speech evidence: acoustic features, formant frequency-based likelihood ratios and ASR performance," *The International Journal of Speech, Language and the Law*, 20, 147-150, 2013.
[Kelly13] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language,* vol. 27, pp. 1068-1084, 2013.

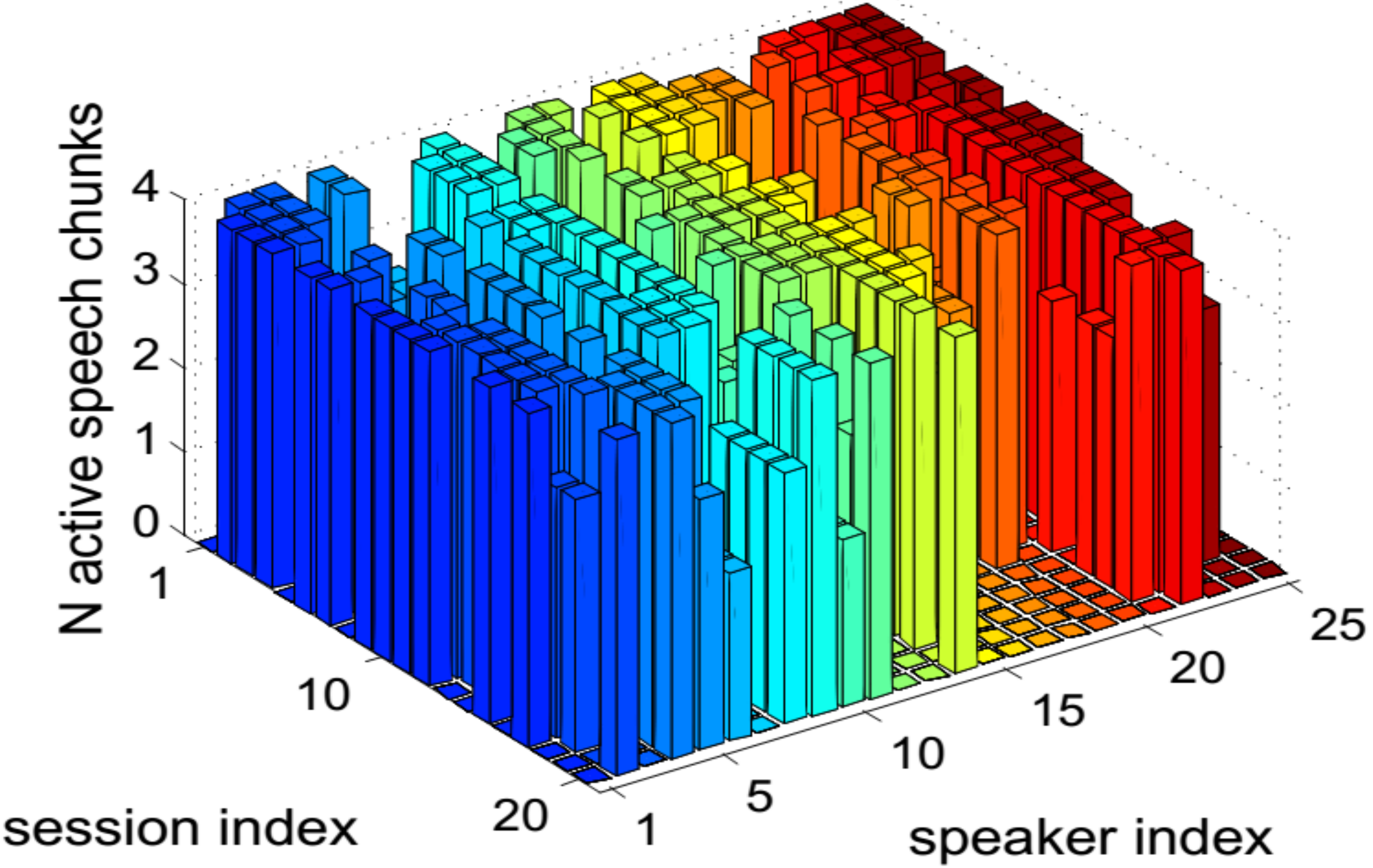
MARP Corpus

- 73 speakers (46 male, 27 female) *
- 21 sessions recorded over 3 years, at intervals of approximately 2 months *
- Each session included a pair of speakers conversing freely for 10 minutes
- Recording environment and equipment remained consistent throughout: soundproof booth + headset microphones
- Data released as 8 kHz, 16 bit, raw mono audio

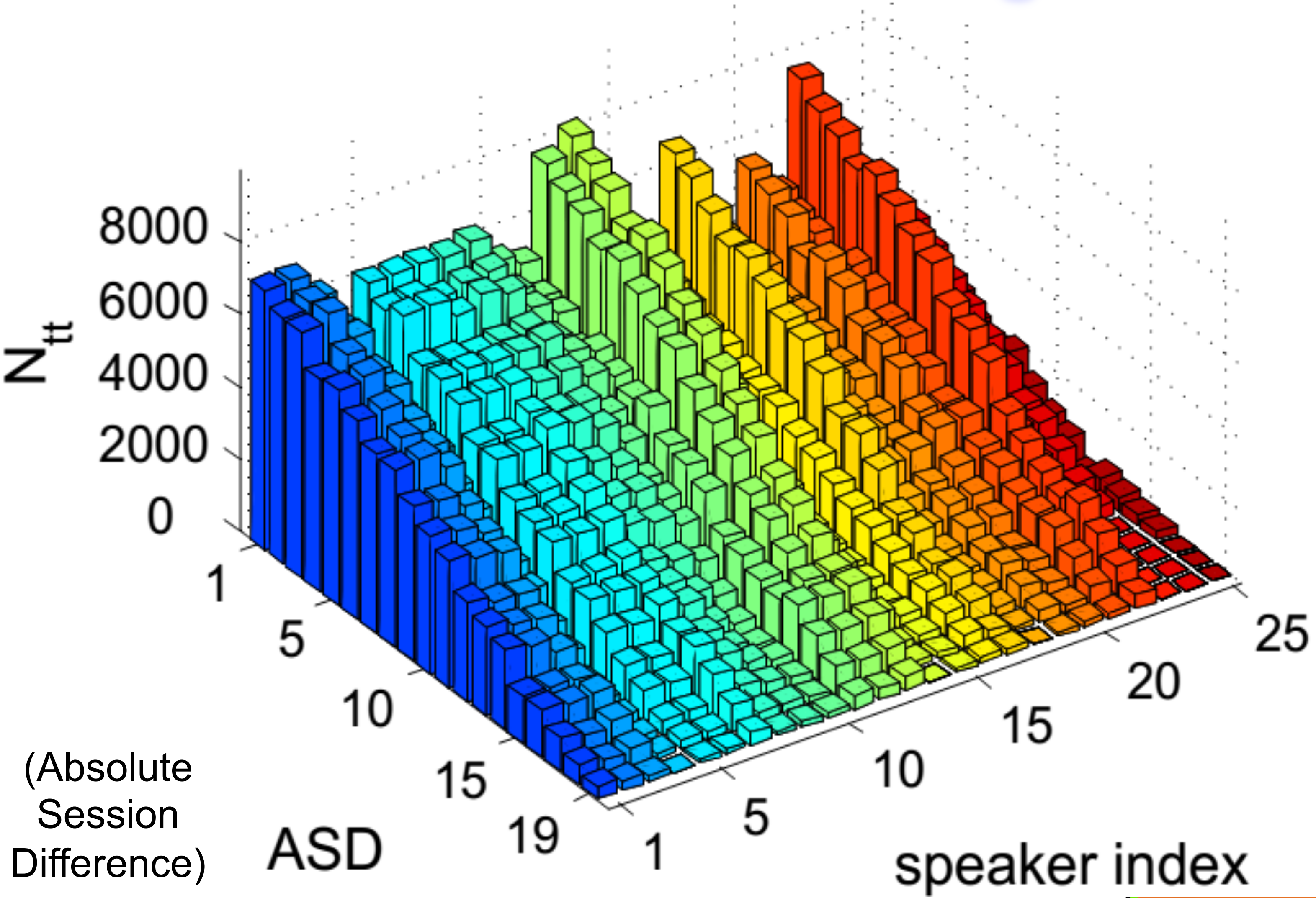
* not all speakers participated in all sessions
1st and 6th session data not released
exact recording dates unavailable

MARP Corpus

25 selected female speakers

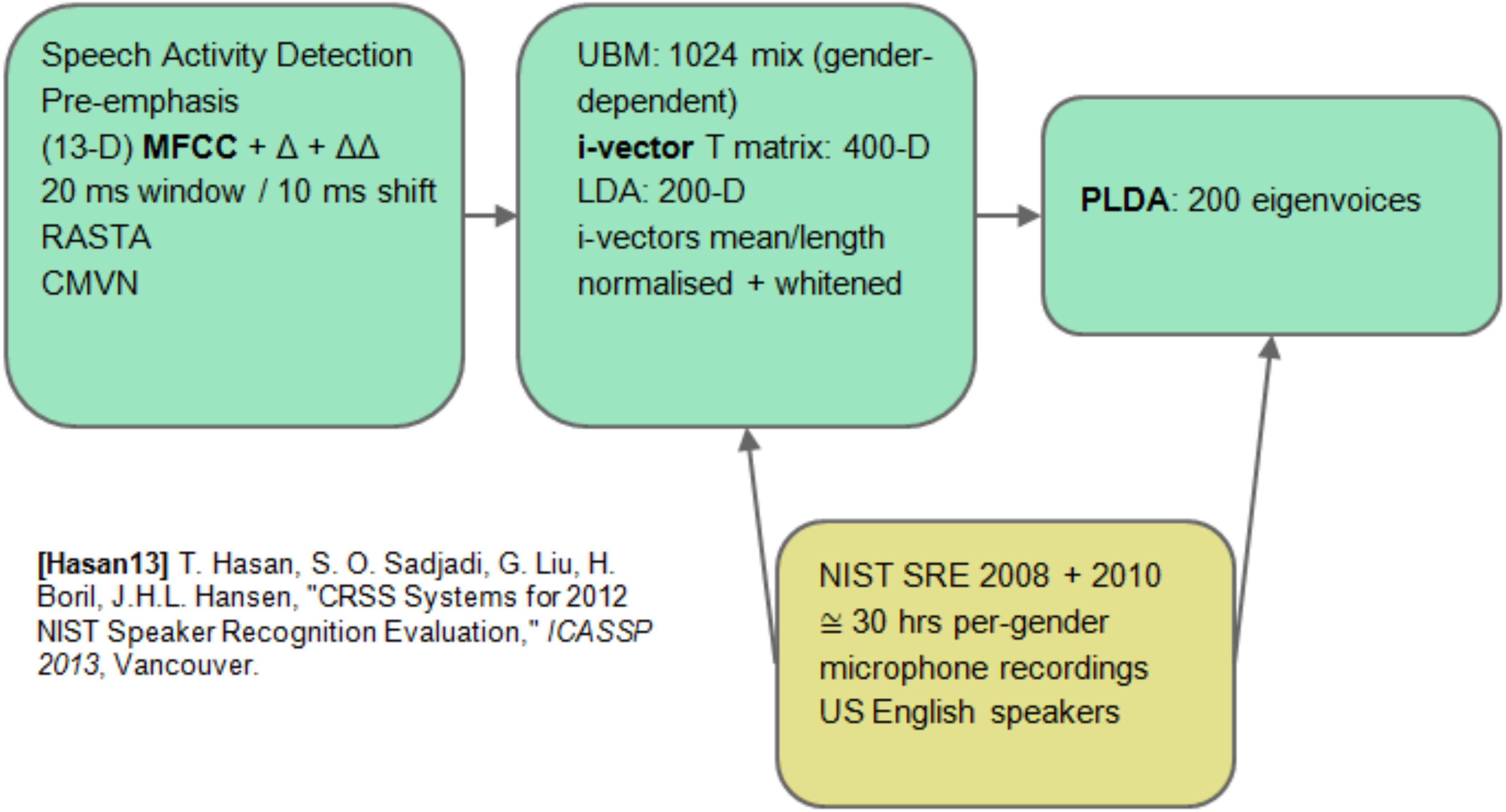


All-vs-All Protocol: Target Trials



(Absolute Session Difference)

Speaker Recognition System



Speech Activity Detection
Pre-emphasis
(13-D) **MFCC** + Δ + ΔΔ
20 ms window / 10 ms shift
RASTA
CMVN

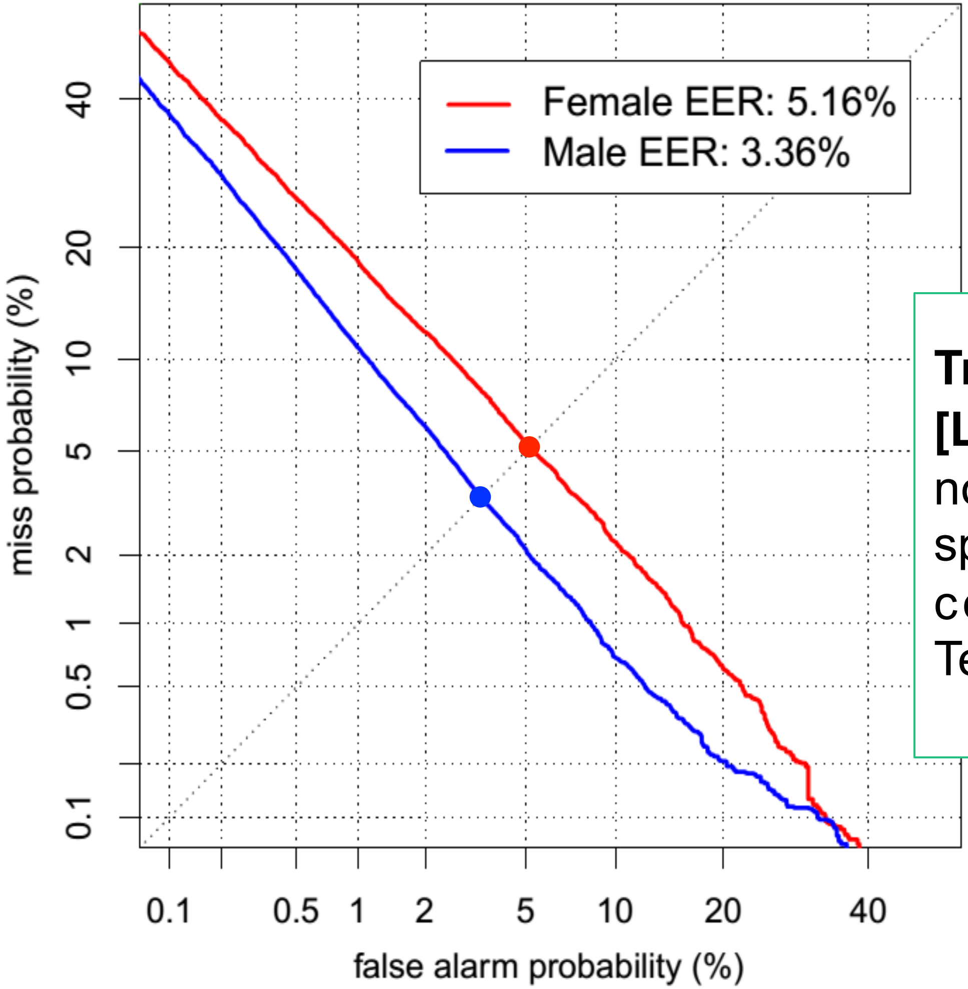
UBM: 1024 mix (gender-dependent)
i-vector T matrix: 400-D
LDA: 200-D
i-vectors mean/length normalised + whitened

PLDA: 200 eigenvoices

NIST SRE 2008 + 2010
≅ 30 hrs per-gender
microphone recordings
US English speakers

[Hasan13] T. Hasan, S. O. Sadjadi, G. Liu, H. Boril, J.H.L. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," *ICASSP 2013*, Vancouver.

DET Curves: all trials

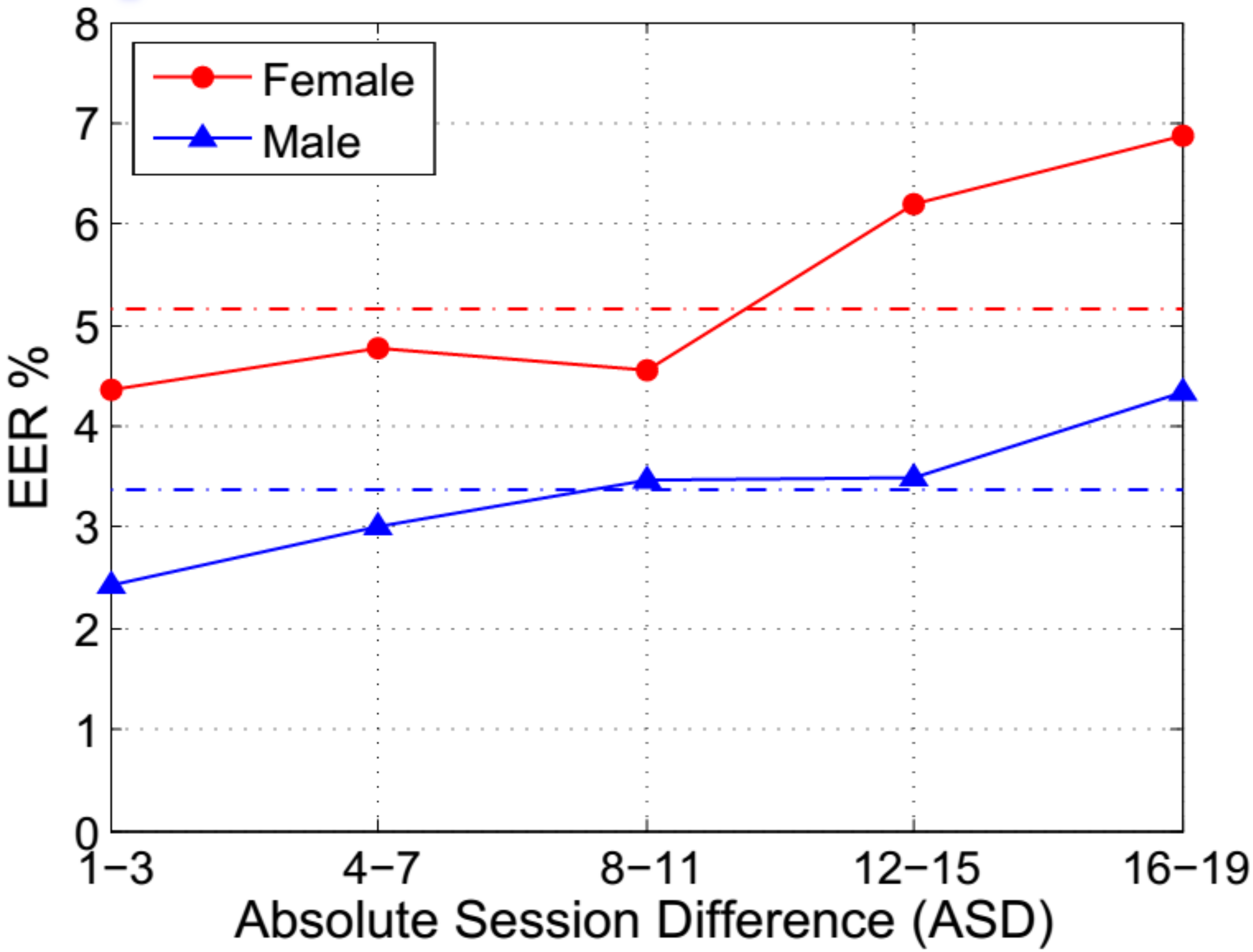


DET = Detection Error Tradeoff
EER = Equal Error Rate

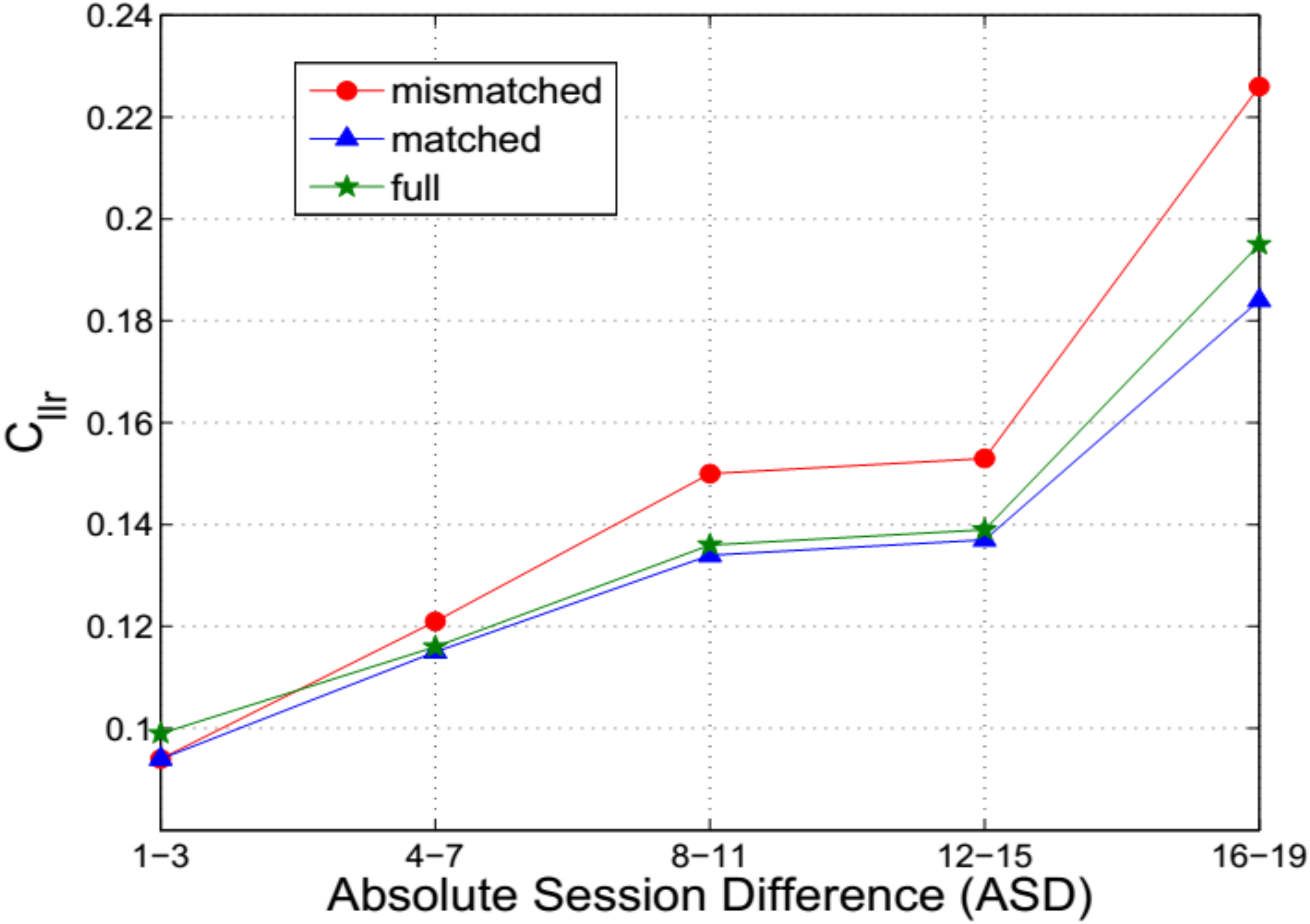
Trials weighted by speaker + ASD

[Leeuwen07] D. van Leeuwen, “A note on performance metrics for speaker recognition using multiple conditions in an evaluation”, Technical Report, 2007

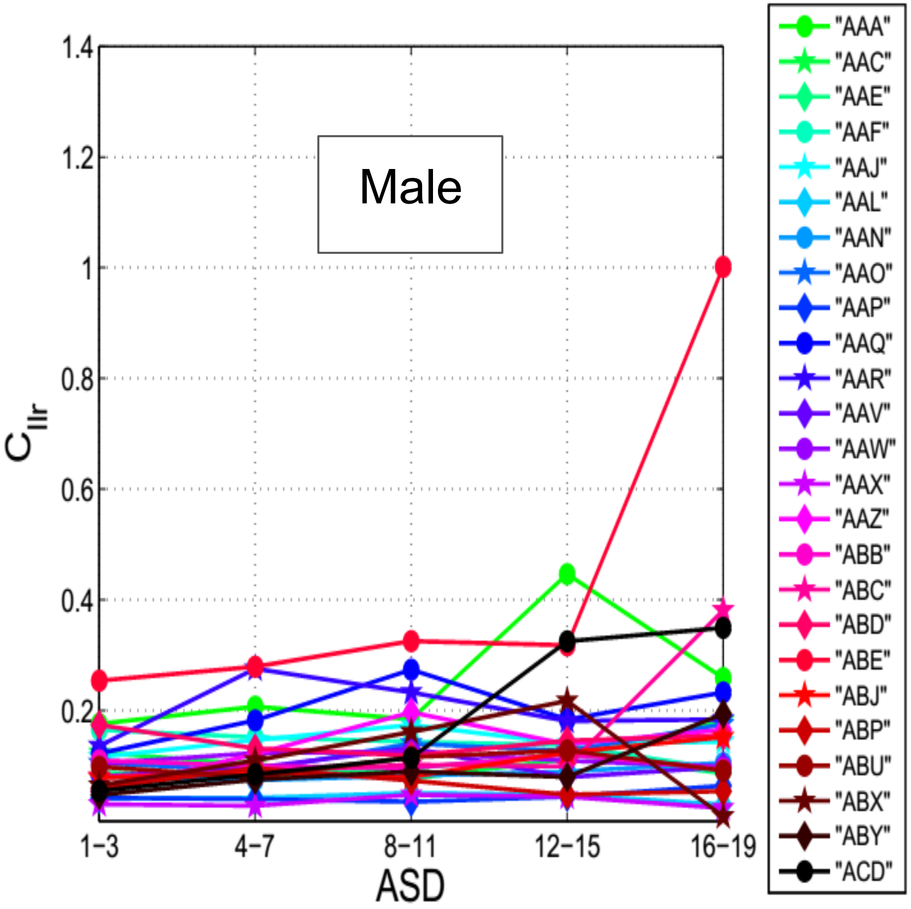
Equal Error Rate vs. Time Lapse

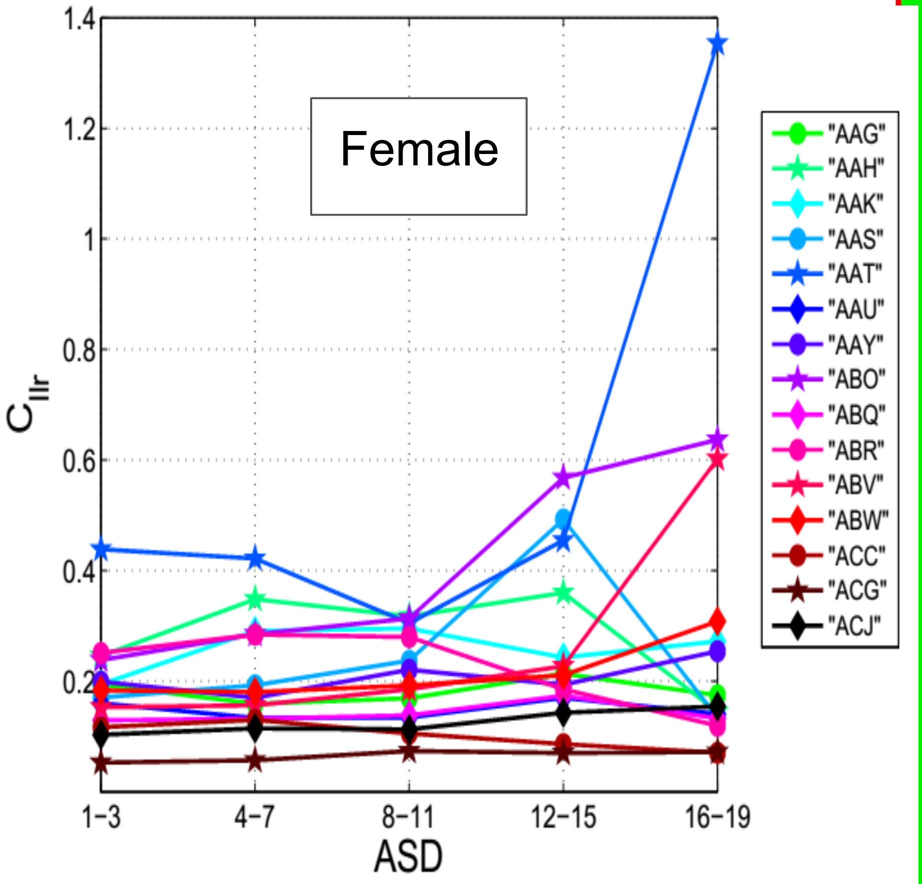


$C_{\ell\ell r}$ vs. Time Lapse

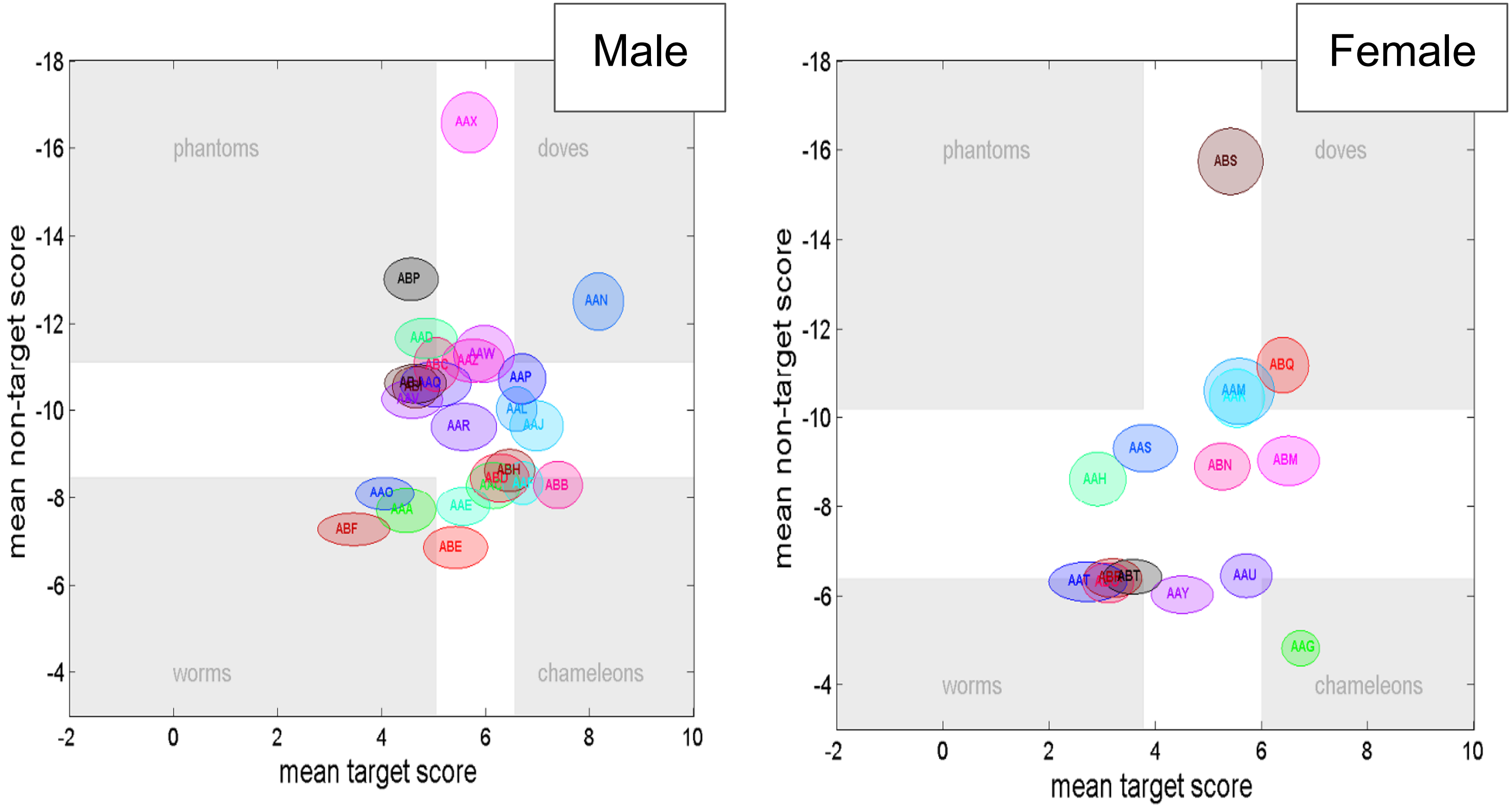


$C_{\ell\ell r}$ per Speaker



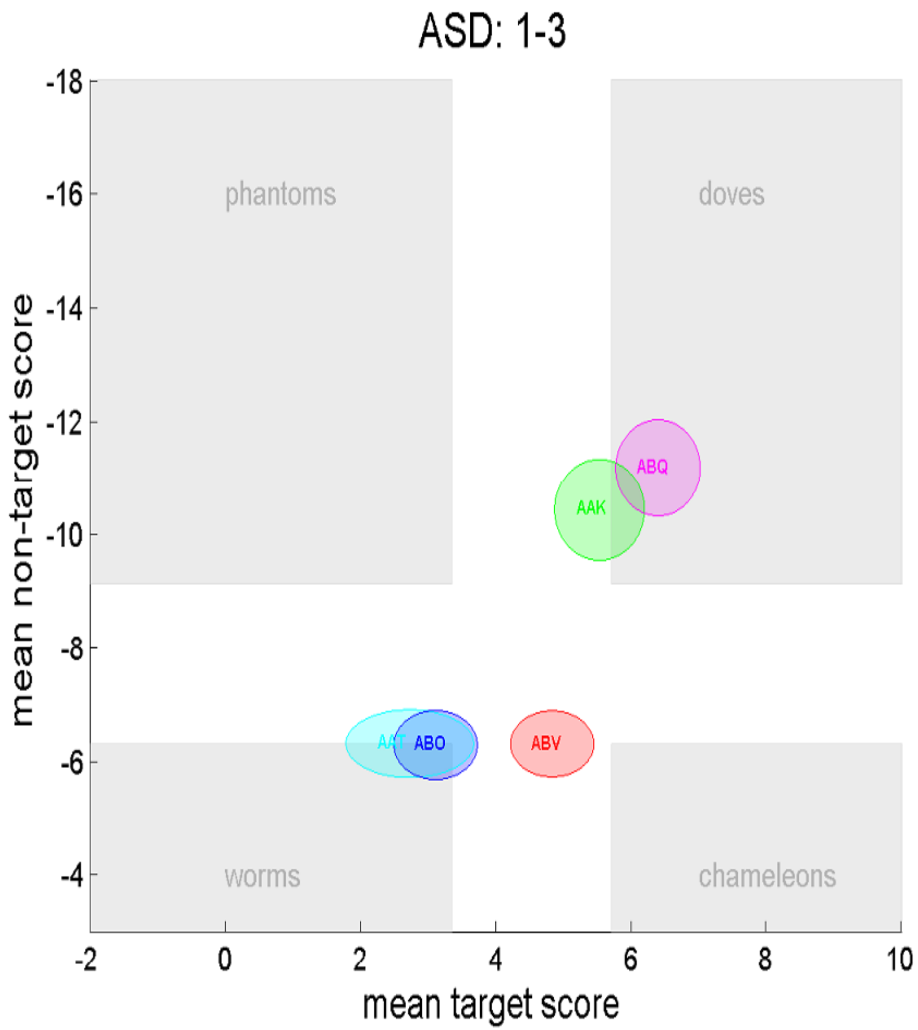


Speaker "Zoo" Classification

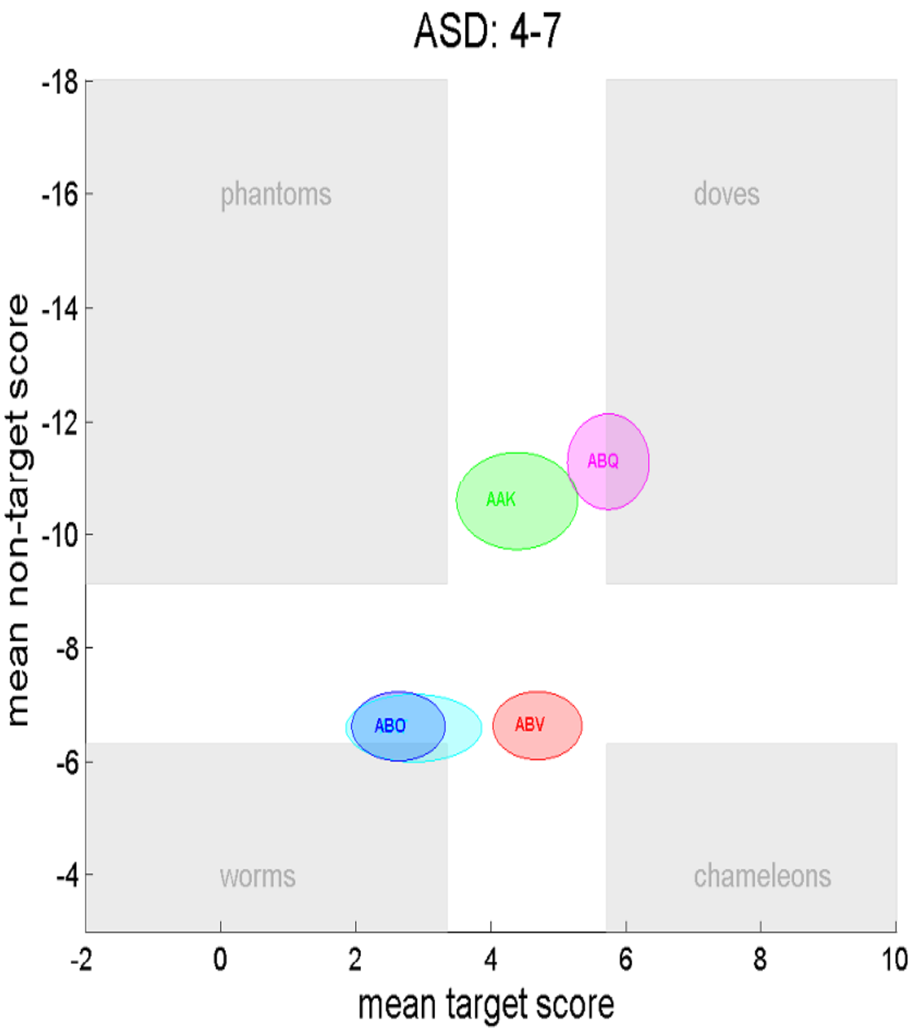


[Alexander14] A. Alexander, O. Forth, J. Nash, N, Yager, "Zooplots for Speaker Recognition with Tall and Fat Animals," *International Association of Forensic Phonetics and Acoustics (IAFPA)* 2014 Zurich, Switzerland.

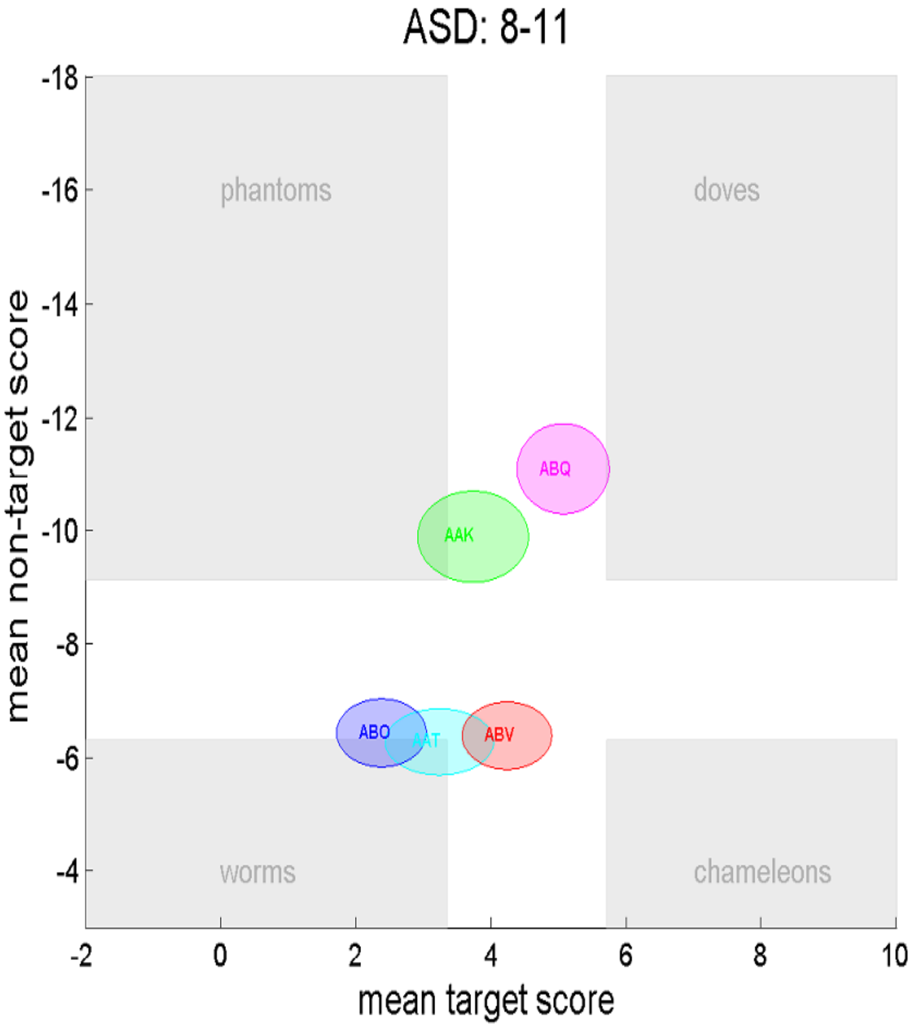
Visualizing Aging Score Trajectories



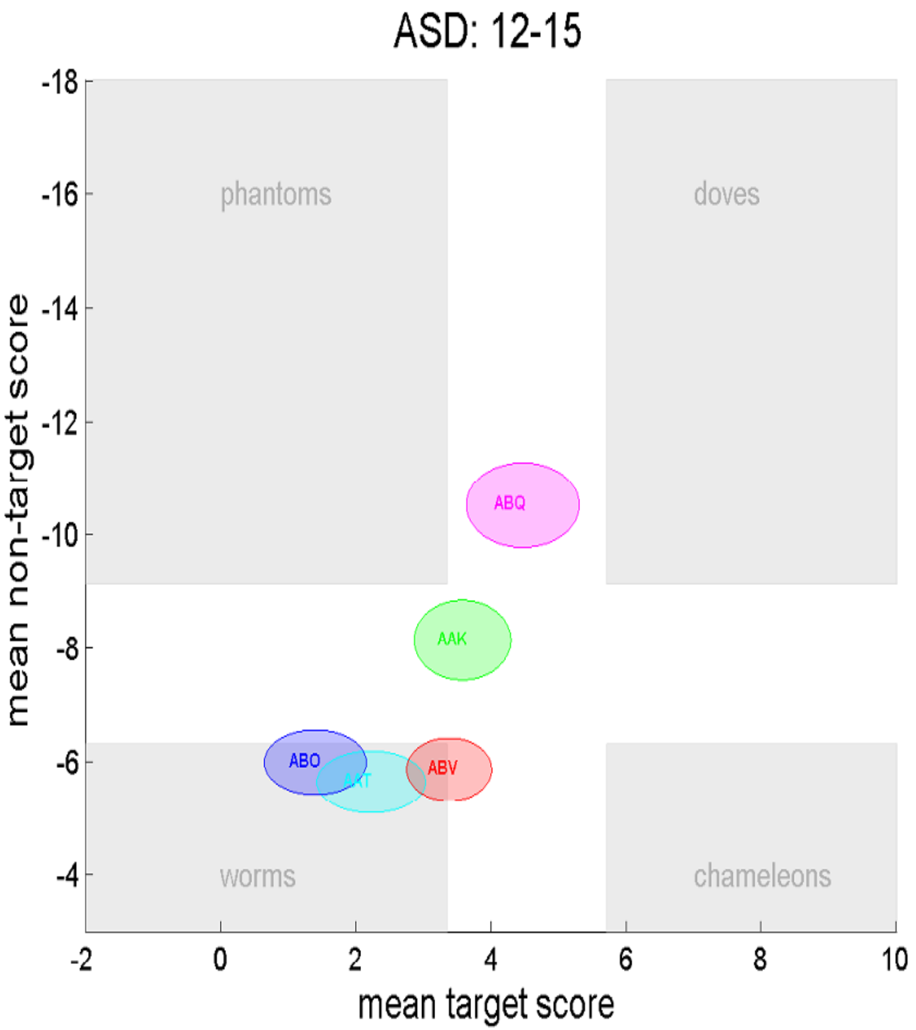
Visualizing Aging Score Trajectories



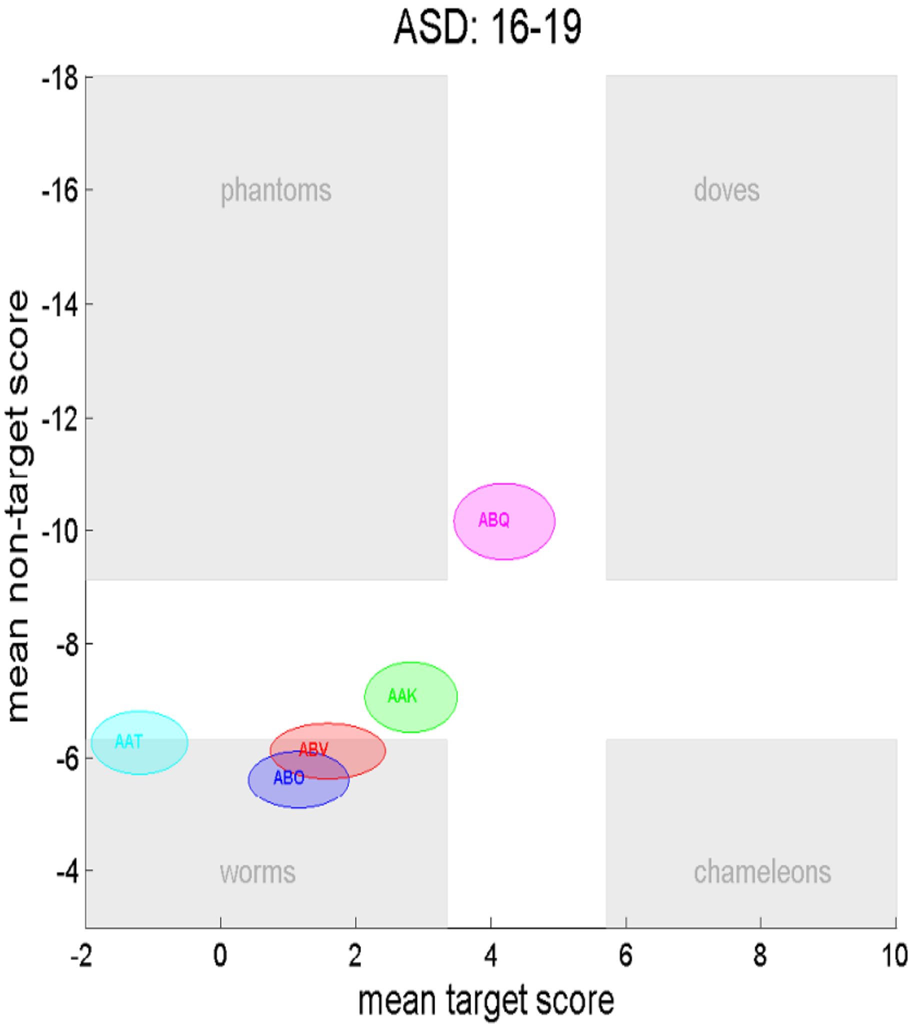
Visualizing Aging Score Trajectories



Visualizing Aging Score Trajectories

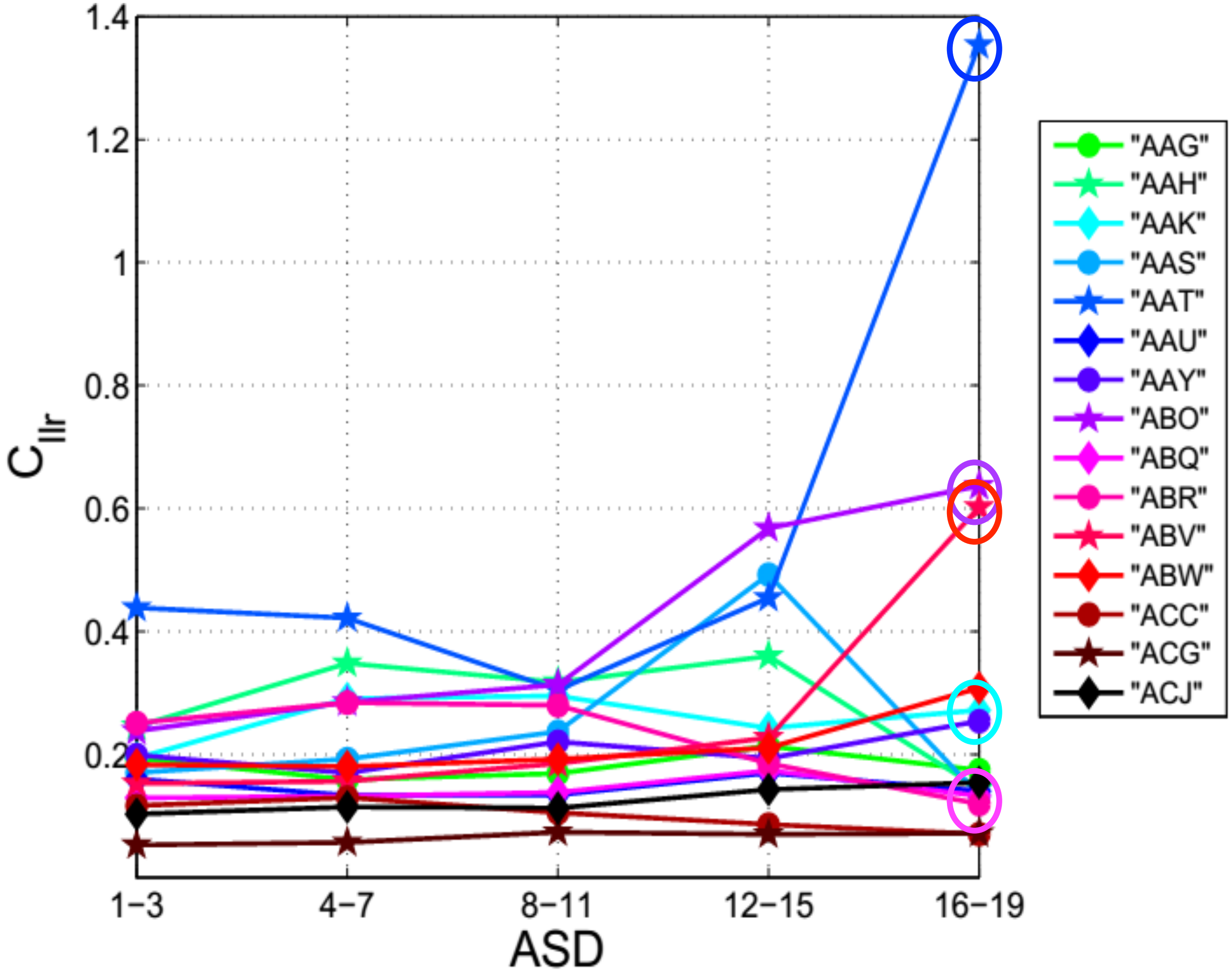


Visualizing Aging Score Trajectories



$C_{\ell\ell r}$ per Speaker

(highlighting the most mobile female speakers)

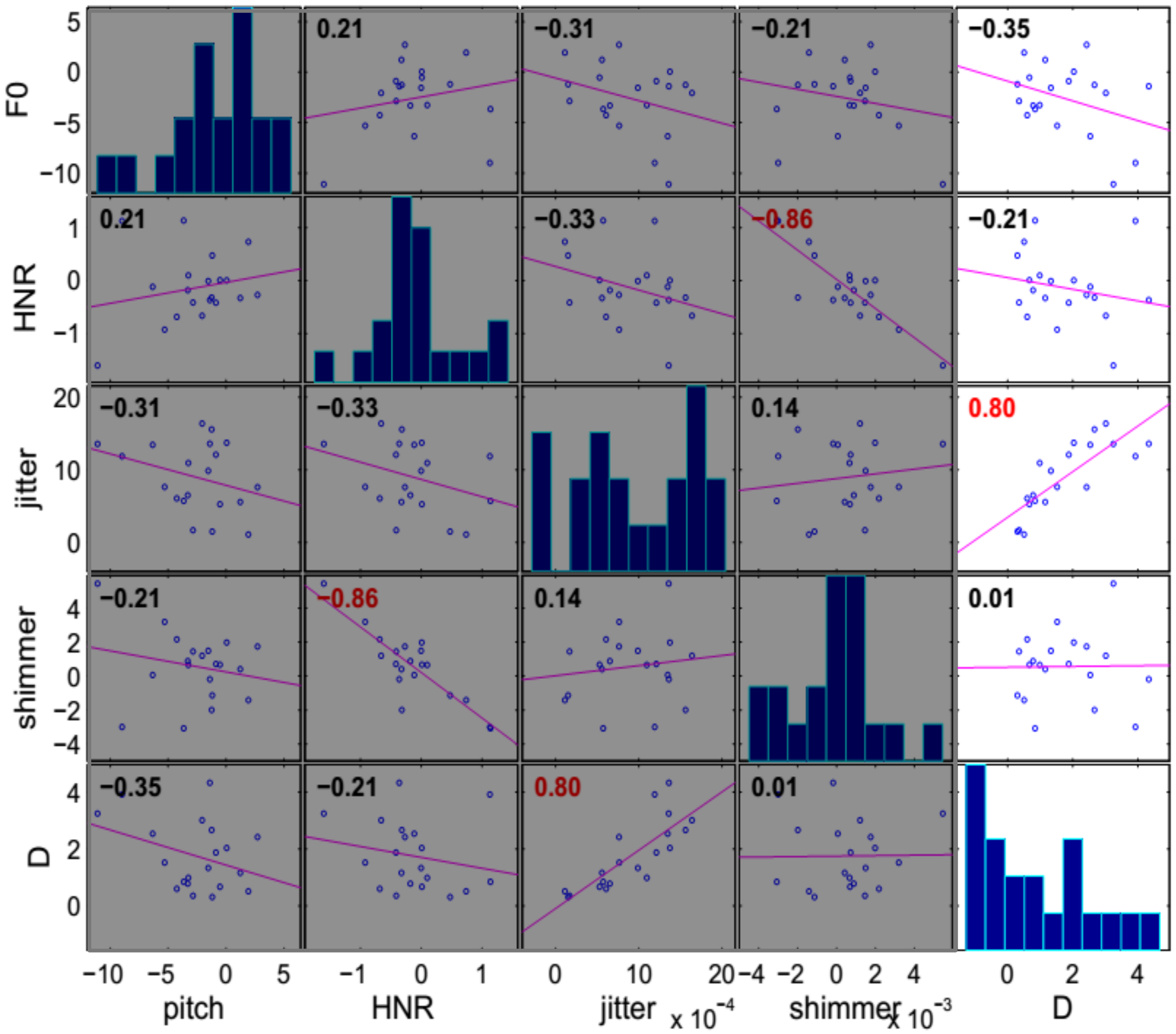


Vocal Feature Analysis: Long-term Averages

1. F0
2. HNR (harmonic-to-noise ratio)
3. Local Shimmer
4. Local Jitter

Vocal Feature Analysis

Female speakers with most mobile score distributions



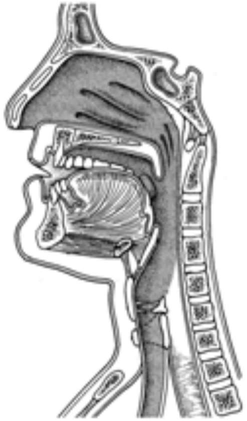
Conclusions

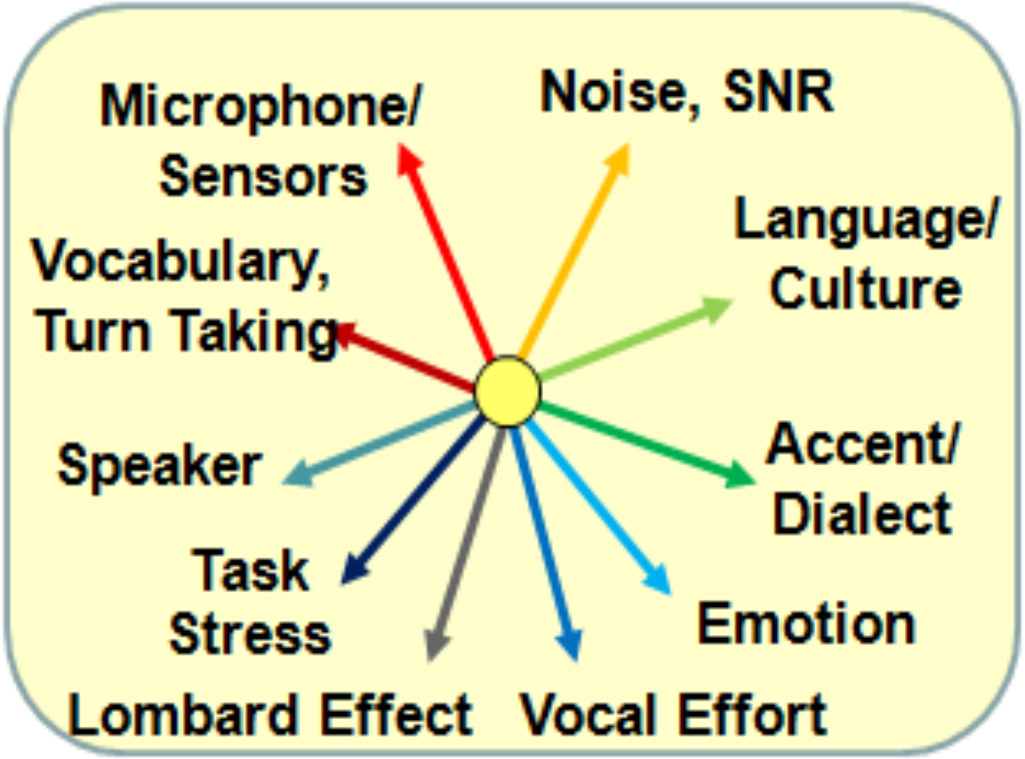
- Speaker Variability is EVERYWHERE!
- Aging process affects automatic speaker recognition in a speaker-dependent way
- Score-aging calibration can improve discrimination and calibration performance **[kelly15]**
- Analysis of score trajectories can flag speakers with most rapidly changing voices
- Feature development may be informed by characteristics of these speakers' voices

[kelly15] F. Kelly and J. H. L. Hansen, "Evaluation and calibration of short-term aging effects in speaker verification", to appear in *InterSpeech 2015*, Dresden, Germany, September

Questions?

Speaker Based

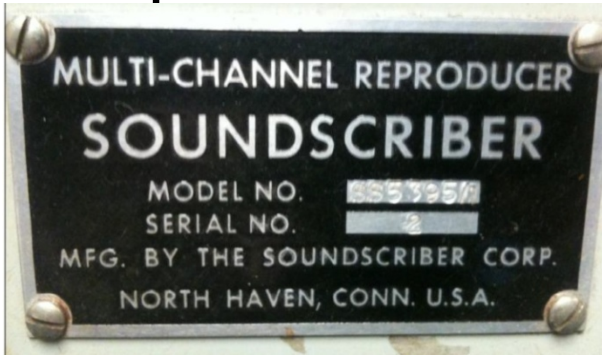




(Finnian.Kelly,John.Hansen)@utdallas.edu

Corpus Development: SoundScriber

- All loops exist on 30-track tapes
- Air to Ground & Flight Director Loops have been digitized
- Lunar & Command Module: more digitizing needed
- Backroom Loops: little exists
- Current effort: digitizing original 30-track tapes from archives



MULTI-CHANNEL REPRODUCER

SOUNDSCRIBER

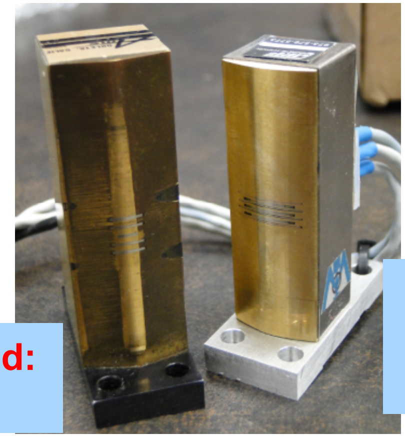
MODEL NO. SS5395/1

SERIAL NO. 2

MFG. BY THE SOUNDSCRIBER CORP.

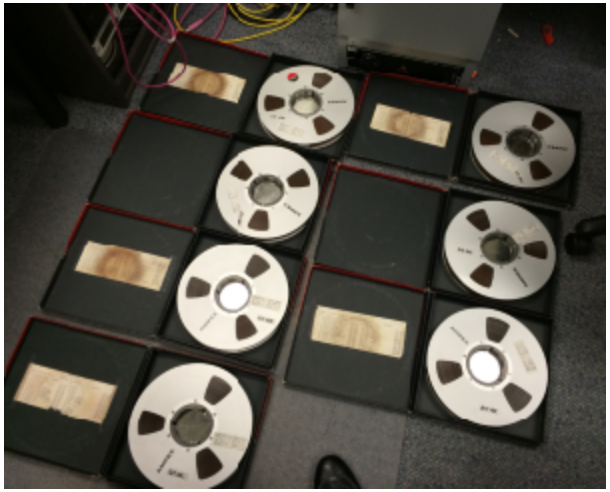
NORTH HAVEN, CONN. U.S.A.

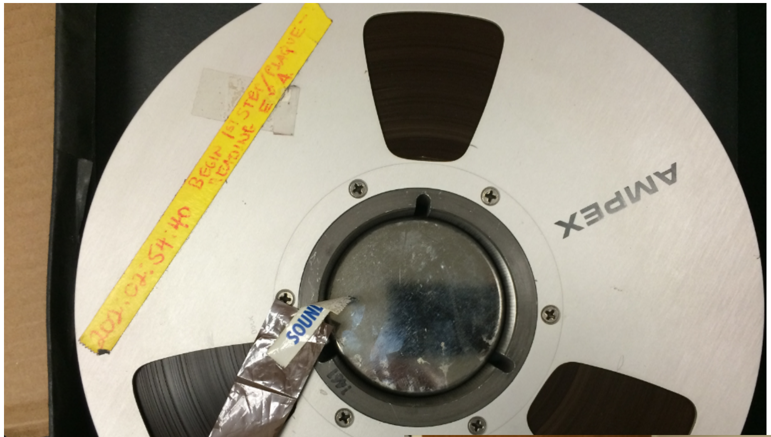
Original 1-track Read Head: from SoundScriber #2

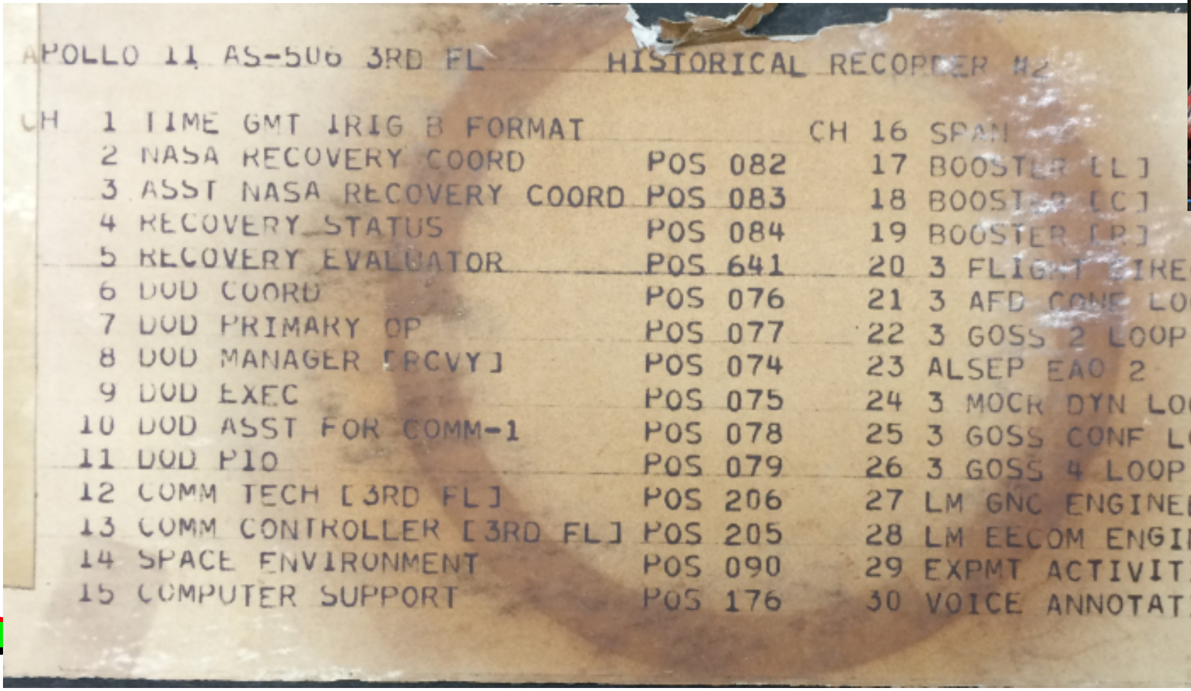


New Designed 3-track Read Head; proto-type for 30/60 Track Head

Corpus:





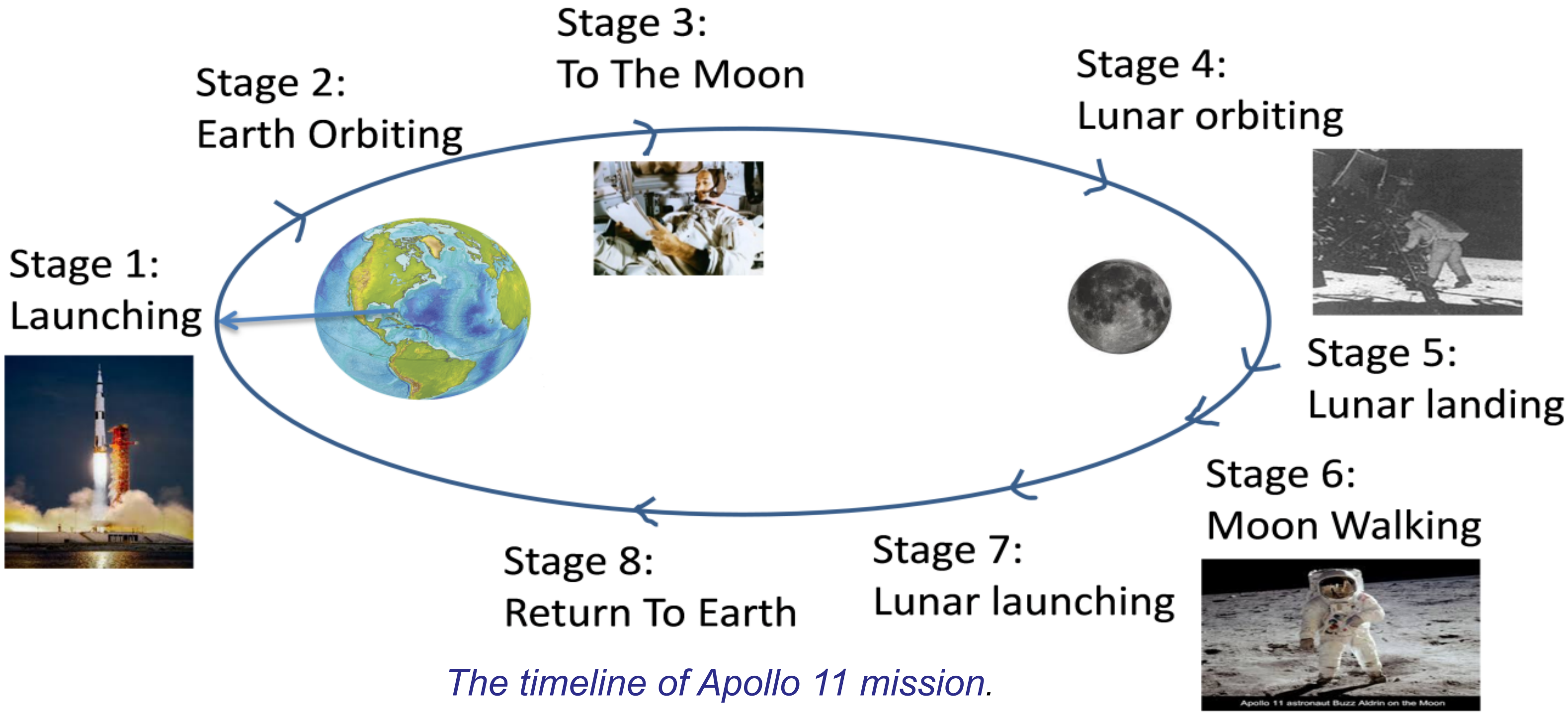


APOLLO 11 AS-506 3RD FL HISTORICAL RECORDER #2

CH	1 TIME GMT IRIG B FORMAT		CH	16 SPAN	
	2 NASA RECOVERY COORD	POS 082		17 BOOSTER [L]	
	3 ASST NASA RECOVERY COORD	POS 083		18 BOOSTER [C]	
	4 RECOVERY STATUS	POS 084		19 BOOSTER [R]	POS 012
	5 RECOVERY EVALUATOR	POS 641		20 3 FLIGHT DIRECTOR	
	6 DOD COORD	POS 076		21 3 AFD CONF LOOP	
	7 DOD PRIMARY OP	POS 077		22 3 GOSS 2 LOOP	
	8 DOD MANAGER [RCVY]	POS 074		23 ALSEP EAO 2	
	9 DOD EXEC	POS 075		24 3 MOCR DYN LOOP	
	10 DOD ASST FOR COMM-1	POS 078		25 3 GOSS CONF LOOP	
	11 DOD PIO	POS 079		26 3 GOSS 4 LOOP	
	12 COMM TECH [3RD FL]	POS 206		27 LM GNC ENGINEER	
	13 COMM CONTROLLER [3RD FL]	POS 205		28 LM EECOM ENGINEER	
	14 SPACE ENVIRONMENT	POS 090		29 EXPMT ACTIVITIES OFFICER	POS 603
	15 COMPUTER SUPPORT	POS 176		30 VOICE ANNOTATION	

1.0 Apollo-11: Audio Analysis

- Corpus: air-to-ground from Apollo-11 mission
- Mission Duration: **8 days, 3 hours 18 minutes, 35 seconds.**
- Voice of 3 astronauts: Neil Armstrong, Buzz Aldrin, Michael Collins.
- Apollo-11: separated into 8 stages



Stage 1: Launching

Stage 2: Earth Orbiting

Stage 3: To The Moon

Stage 4: Lunar orbiting

Stage 5: Lunar landing

Stage 6: Moon Walking

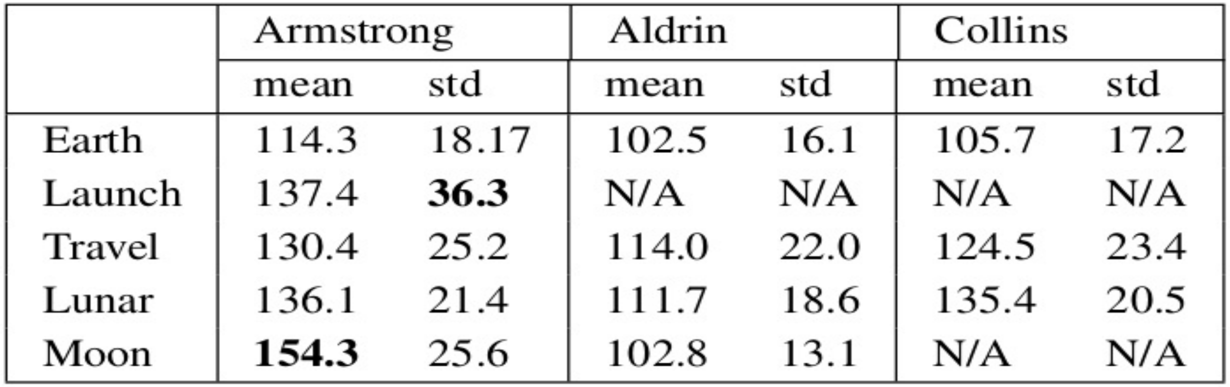
Stage 7: Lunar launching

Stage 8: Return To Earth

The timeline of Apollo 11 mission.

1.2 Fundamental Frequency

Mean & Standard deviation of f0 over mission stages



	Armstrong		Aldrin		Collins	
	mean	std	mean	std	mean	std
Earth	114.3	18.17	102.5	16.1	105.7	17.2
Launch	137.4	**36.3**	N/A	N/A	N/A	N/A
Travel	130.4	25.2	114.0	22.0	124.5	23.4
Lunar	136.1	21.4	111.7	18.6	135.4	20.5
Moon	**154.3**	25.6	102.8	13.1	N/A	N/A

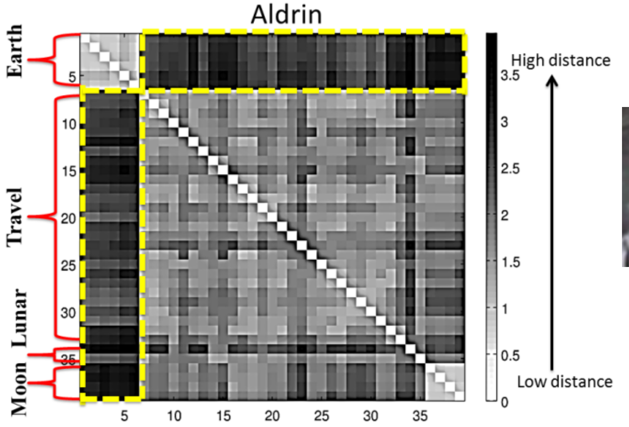
- Mean & Standard deviation of f0 consistently higher in space.
- Armstrong's f0 significantly higher on the moon compared to other conditions; same effect not observed for Aldrin's f0.
- Armstrong's f0 reached 160Hz when he uttered the famous quote:

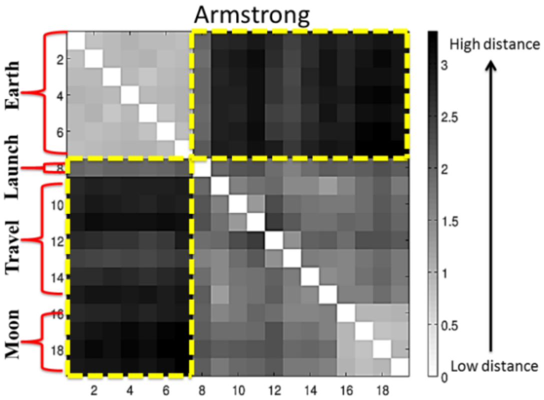
"That's one small step for man, one giant leap for mankind"

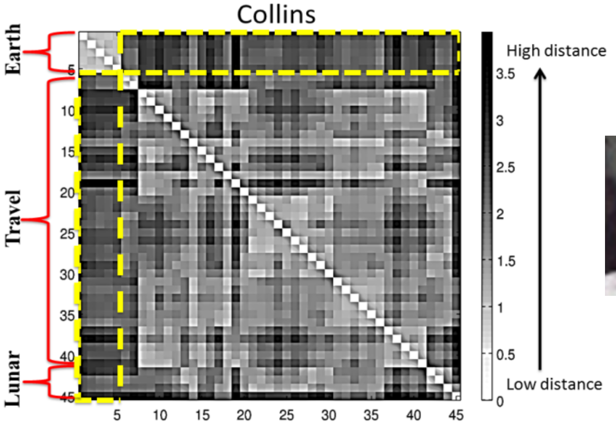
1.3 Acoustic Model Analysis

Analysis of Speaker Acoustic Models over Mission:

Acoustic Model comparison using models trained from 60-sec audio blocks with different conditions using GMM and KL divergence.







Speaker Models in Space Varies Significantly compared to the ones in Earth.