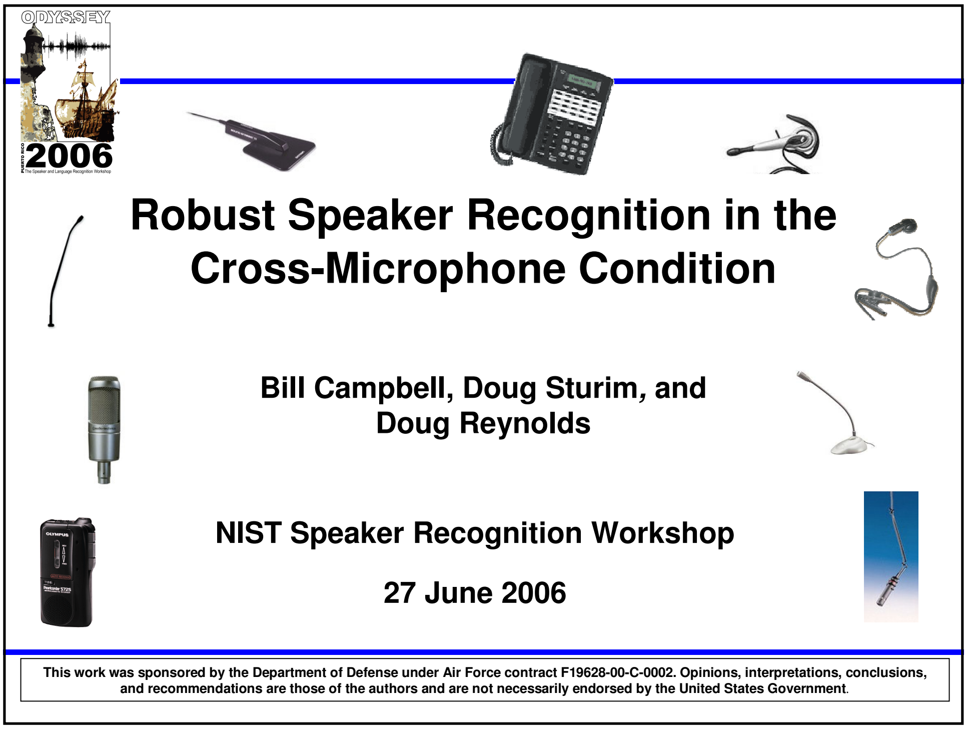


ODYSSEY
2006
The Speaker and Language Recognition Workshop

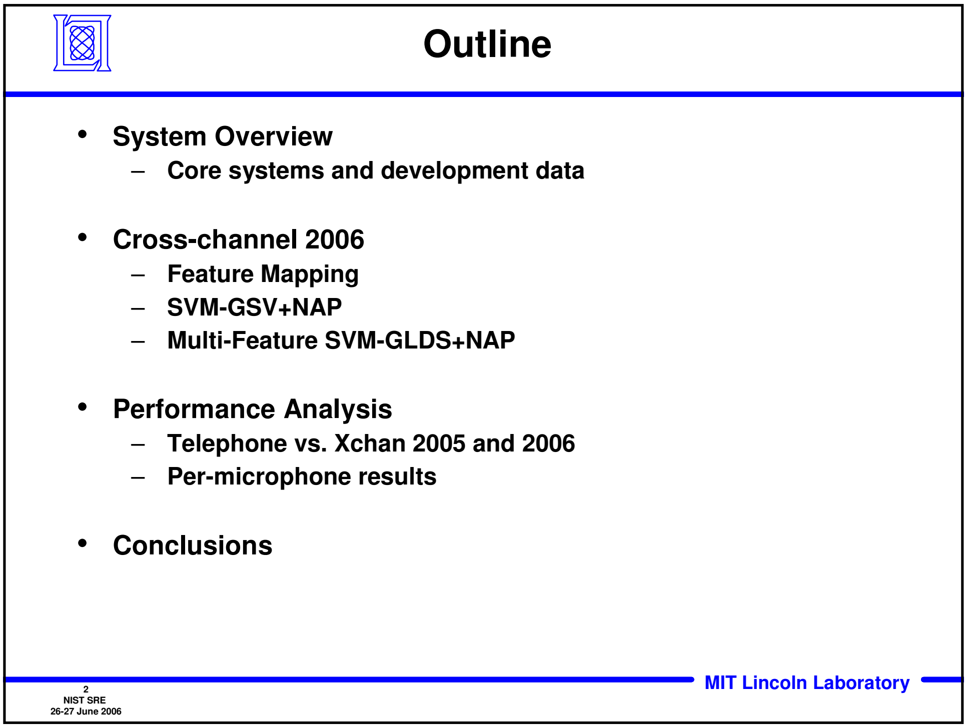
Robust Speaker Recognition in the Cross-Microphone Condition

Bill Campbell, Doug Sturim, and
Doug Reynolds

NIST Speaker Recognition Workshop

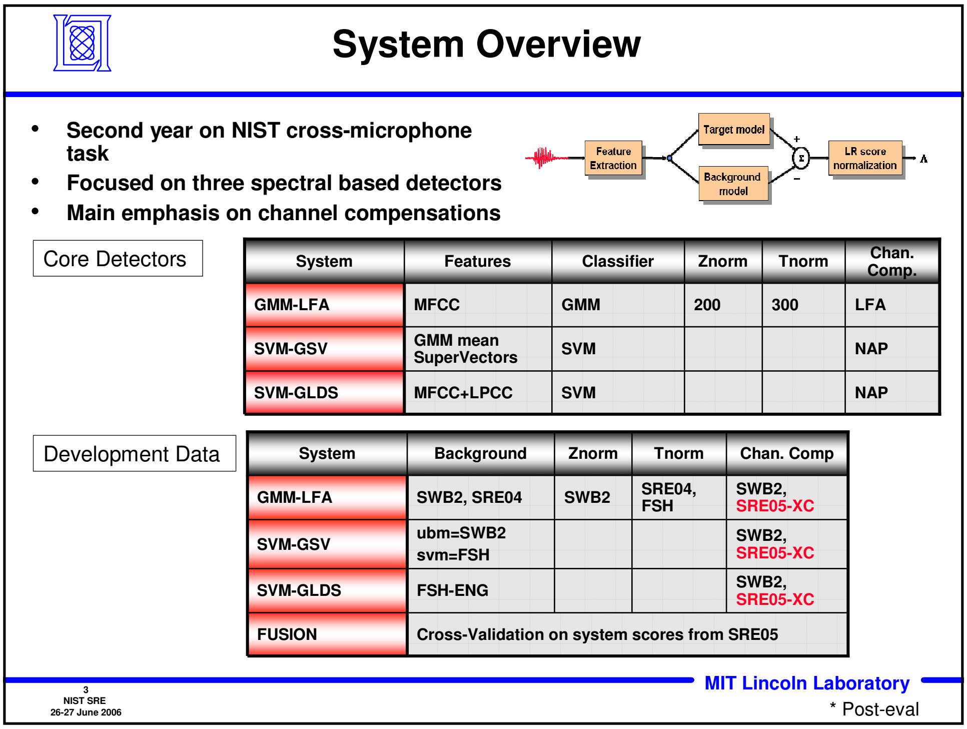
27 June 2006

This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.



Outline

- System Overview
  - Core systems and development data
- Cross-channel 2006
  - Feature Mapping
  - SVM-GSV+NAP
  - Multi-Feature SVM-GLDS+NAP
- Performance Analysis
  - Telephone vs. Xchan 2005 and 2006
  - Per-microphone results
- Conclusions



System Overview

- Second year on NIST cross-microphone task
- Focused on three spectral based detectors
- Main emphasis on channel compensations

Feature Extraction
Target model
Background model
+
–
Σ
LR score normalization
Λ

Core Detectors

System	Features	Classifier	Znorm	Tnorm	Chan. Comp.
GMM-LFA	MFCC	GMM	200	300	LFA
SVM-GSV	GMM mean SuperVectors	SVM			NAP
SVM-GLDS	MFCC+LPCC	SVM			NAP

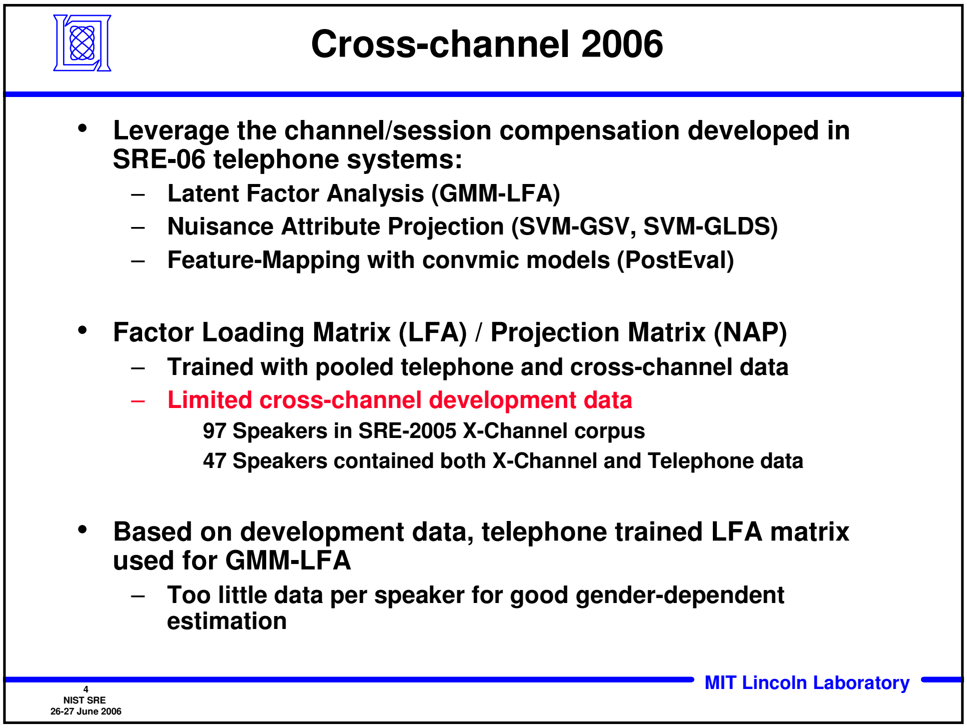
Development Data

System	Background	Znorm	Tnorm	Chan. Comp
GMM-LFA	SWB2, SRE04	SWB2	SRE04, FSH	SWB2, SRE05-XC

SVM-GSV	ubm=SWB2
	svm=FSH
				SWB2, SRE05-XC

SVM-GLDS	FSH-ENG			SWB2, SRE05-XC
FUSION	Cross-Validation on system scores from SRE05			

* Post-eval



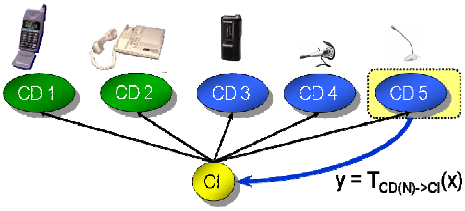
Cross-channel 2006

- Leverage the channel/session compensation developed in SRE-06 telephone systems:
  - Latent Factor Analysis (GMM-LFA)
  - Nuisance Attribute Projection (SVM-GSV, SVM-GLDS)
  - Feature-Mapping with convmic models (PostEval)
- Factor Loading Matrix (LFA) / Projection Matrix (NAP)
  - Trained with pooled telephone and cross-channel data
  - Limited cross-channel development data
    - 97 Speakers in SRE-2005 X-Channel corpus
    - 47 Speakers contained both X-Channel and Telephone data
- Based on development data, telephone trained LFA matrix used for GMM-LFA
  - Too little data per speaker for good gender-dependent estimation

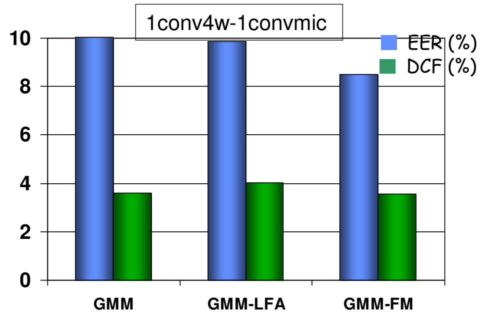
Feature Mapping

Post Evaluation System

- Added microphone dependent models to feature mapper
  - Trained with SRE05 xchan data
  - Channels c1-c8
  - Gender dependent
- Total of 22 models
  - 6 telephone models (cell, cordless, regular)
- Appears to be better than LFA for 1c/1c
  - Perhaps using limited xchan data more effectively
- Currently coupling with other systems

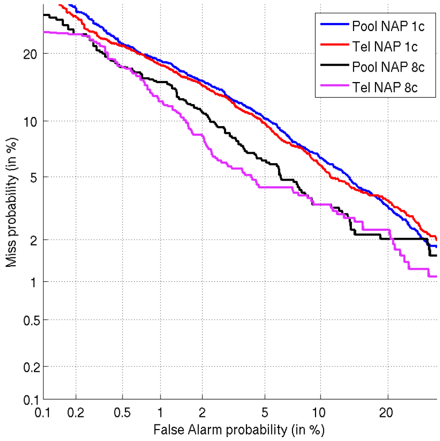


$y = T_{CD(N)\text{->}CI}(x)$



SVM-GSV+NAP

- Eval strategy:
  - Pool NAP: Pool telephone data with xchan microphone data
  - Design NAP projection to eliminate all variation
  - Cons: Development data reused for cross-validation (fusion, thresholds)
- Alternate strategy:
  - Tel NAP: Use models with default telephone session NAP projection
  - Cons: No modeling of xchan microphones

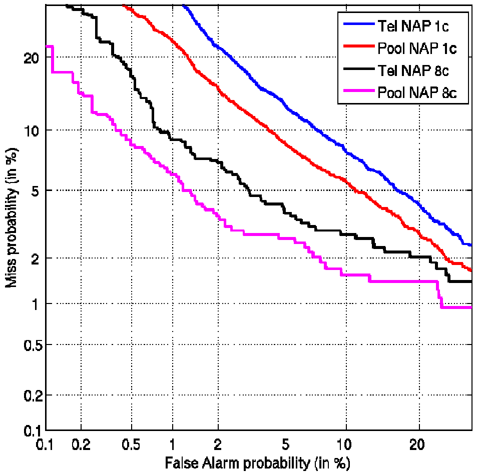


Conclusion: Not much difference for SVM-GSV between pooled and telephone NAP projection at minDCF.

Multi-Feature SVM-GLDS+NAP

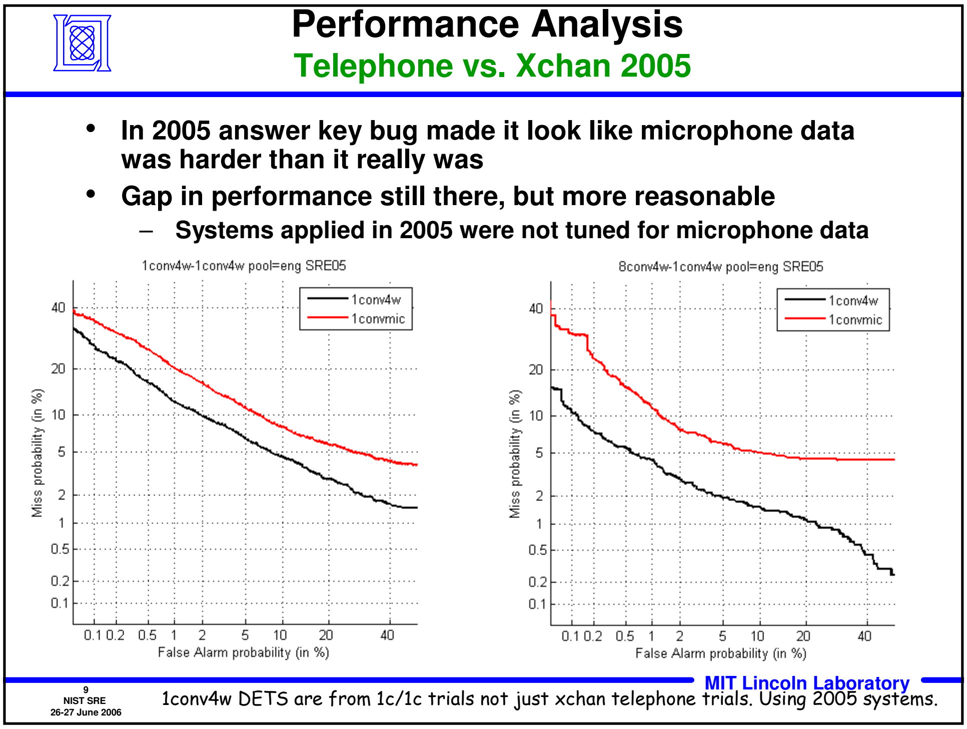
- Strategies for NAP:
  - Pool telephone data with xchan microphone data
  - Telephone only
- Analysis:
  - Pooled NAP works significantly better
  - NAP interacts differently with different feature sets
  - LPCCs are not as good as MFCCs under mismatch; need NAP to make them fuse well

	MFCC EER (%)	LPCC EER (%)	Fuse EER (%)
1c, Tel NAP	10.04	14.07	8.84
1c, Pool NAP	9.22	10.34	6.88
8c, Tel NAP	4.04	7.91	4.19
8c, Pool NAP	3.72	5.27	2.90



Outline

- System Overview
  - Core systems and development data
- Cross-channel 2006
  - Feature Mapping
  - SVM-GSV+NAP
  - Multi-Feature SVM-GLDS+NAP
- Performance Analysis
  - Telephone vs. Xchan 2005 and 2006
  - Per-microphone results
- Conclusions

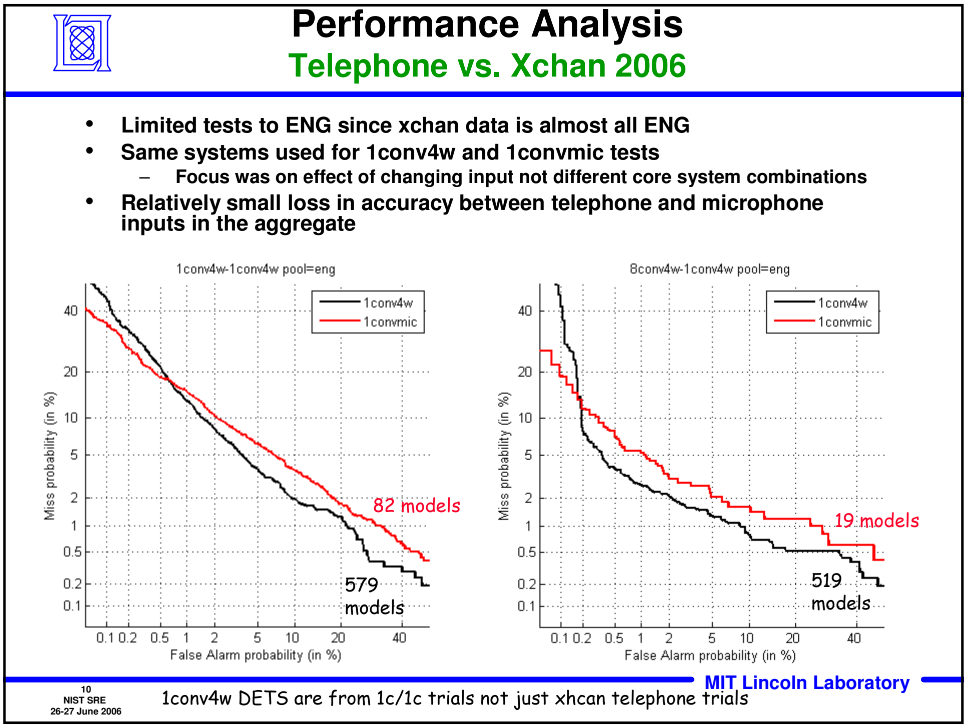


Performance Analysis

Telephone vs. Xchan 2005

- In 2005 answer key bug made it look like microphone data was harder than it really was
- Gap in performance still there, but more reasonable
  - Systems applied in 2005 were not tuned for microphone data

1conv4w DETS are from 1c/1c trials not just xchan telephone trials. Using 2005 systems.

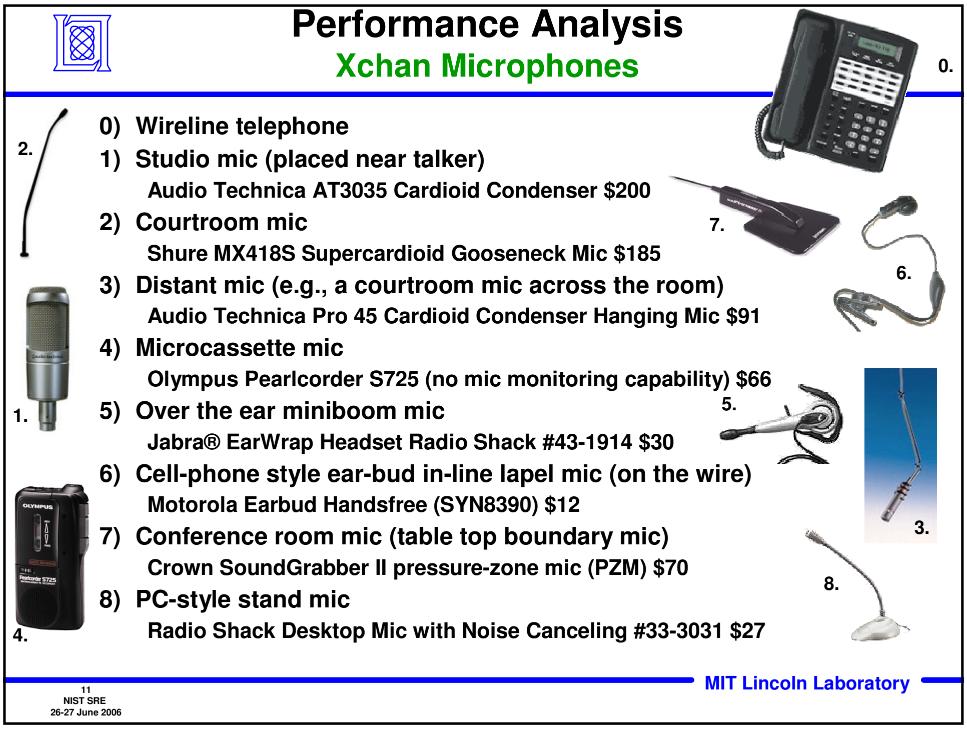


Performance Analysis

Telephone vs. Xchan 2006

- Limited tests to ENG since xchan data is almost all ENG
- Same systems used for 1conv4w and 1convmic tests
  - Focus was on effect of changing input not different core system combinations
- Relatively small loss in accuracy between telephone and microphone inputs in the aggregate

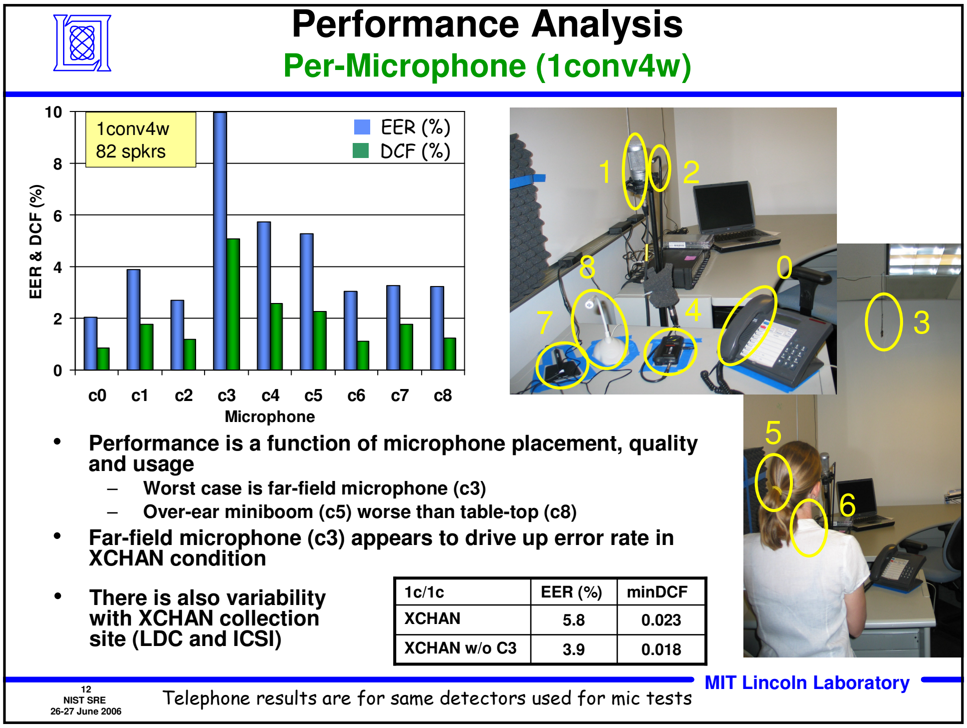
1conv4w DETS are from 1c/1c trials not just xhcan telephone trials



Performance Analysis
Xchan Microphones

0) Wireline telephone
1) Studio mic (placed near talker)
   Audio Technica AT3035 Cardioid Condenser $200
2) Courtroom mic
   Shure MX418S Supercardioid Gooseneck Mic $185
3) Distant mic (e.g., a courtroom mic across the room)
   Audio Technica Pro 45 Cardioid Condenser Hanging Mic $91
4) Microcassette mic
   Olympus Pearlcorder S725 (no mic monitoring capability) $66
5) Over the ear miniboom mic
   Jabra® EarWrap Headset Radio Shack #43-1914 $30
6) Cell-phone style ear-bud in-line lapel mic (on the wire)
   Motorola Earbud Handsfree (SYN8390) $12
7) Conference room mic (table top boundary mic)
   Crown SoundGrabber II pressure-zone mic (PZM) $70
8) PC-style stand mic
   Radio Shack Desktop Mic with Noise Canceling #33-3031 $27

MIT Lincoln Laboratory



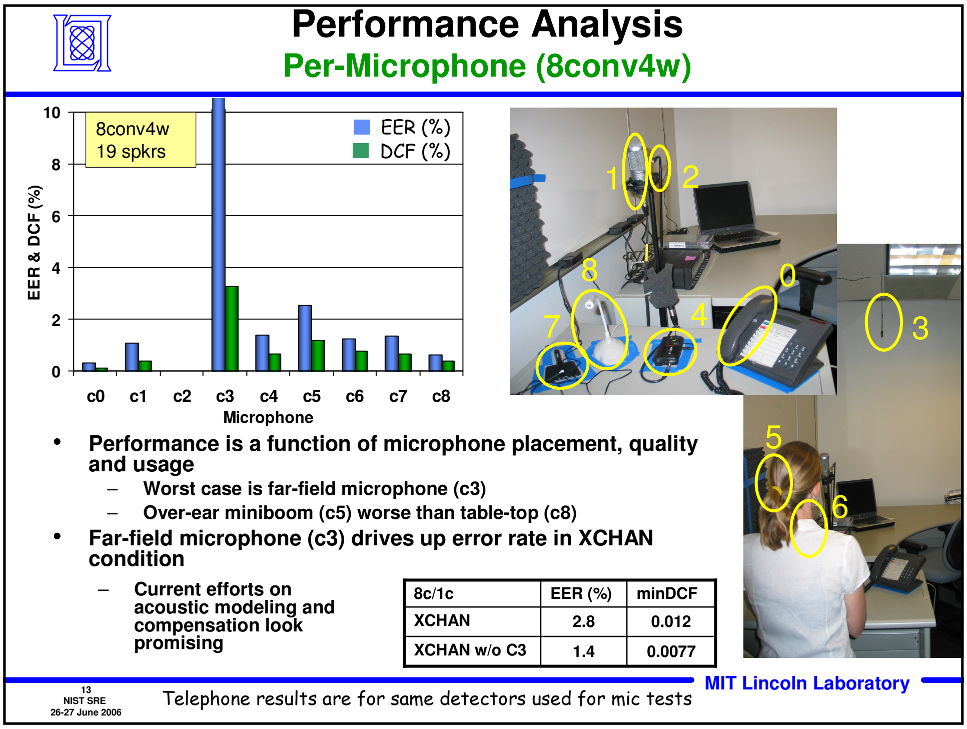
Performance Analysis
Per-Microphone (1conv4w)

- Performance is a function of microphone placement, quality and usage
  - Worst case is far-field microphone (c3)
  - Over-ear miniboom (c5) worse than table-top (c8)
- Far-field microphone (c3) appears to drive up error rate in XCHAN condition
- There is also variability with XCHAN collection site (LDC and ICSI)

1c/1c	EER (%)	minDCF
XCHAN	5.8	0.023
XCHAN w/o C3	3.9	0.018

Telephone results are for same detectors used for mic tests

MIT Lincoln Laboratory



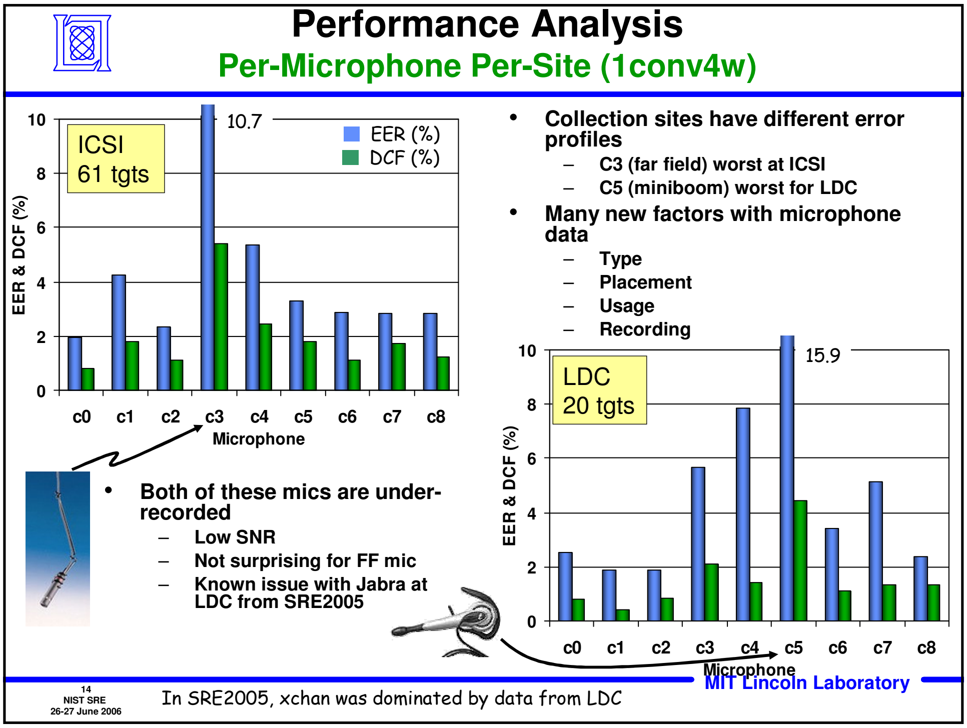
Performance Analysis

Per-Microphone (8conv4w)

- Performance is a function of microphone placement, quality and usage
  - Worst case is far-field microphone (c3)
  - Over-ear miniboom (c5) worse than table-top (c8)
- Far-field microphone (c3) drives up error rate in XCHAN condition
  - Current efforts on acoustic modeling and compensation look promising

8c/1c	EER (%)	minDCF
XCHAN	2.8	0.012
XCHAN w/o C3	1.4	0.0077

Telephone results are for same detectors used for mic tests



Performance Analysis

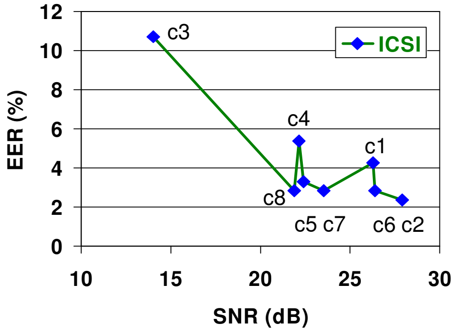
Per-Microphone Per-Site (1conv4w)

- Collection sites have different error profiles
  - C3 (far field) worst at ICSI
  - C5 (miniboom) worst for LDC
- Many new factors with microphone data
  - Type
  - Placement
  - Usage
  - Recording
- Both of these mics are under-recorded
  - Low SNR
  - Not surprising for FF mic
  - Known issue with Jabra at LDC from SRE2005

In SRE2005, xchan was dominated by data from LDC

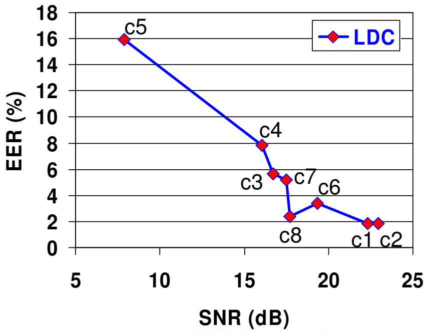
Performance Analysis

EER versus SNR



- EER is from 1conv4w train, 1convmic test
- Simple SNR calculation: SAD marks, SNR=(total speech energy)/(total non-speech energy)
- Correlation between EER and SNR seen again

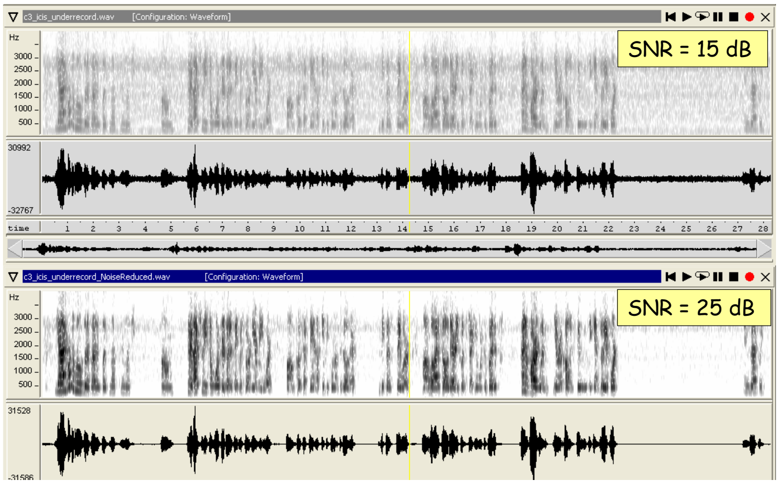
- SNR profile different @ different sites
- LDC order of channels the same as last year
- EER variation shows up most below about 17 dB SNR

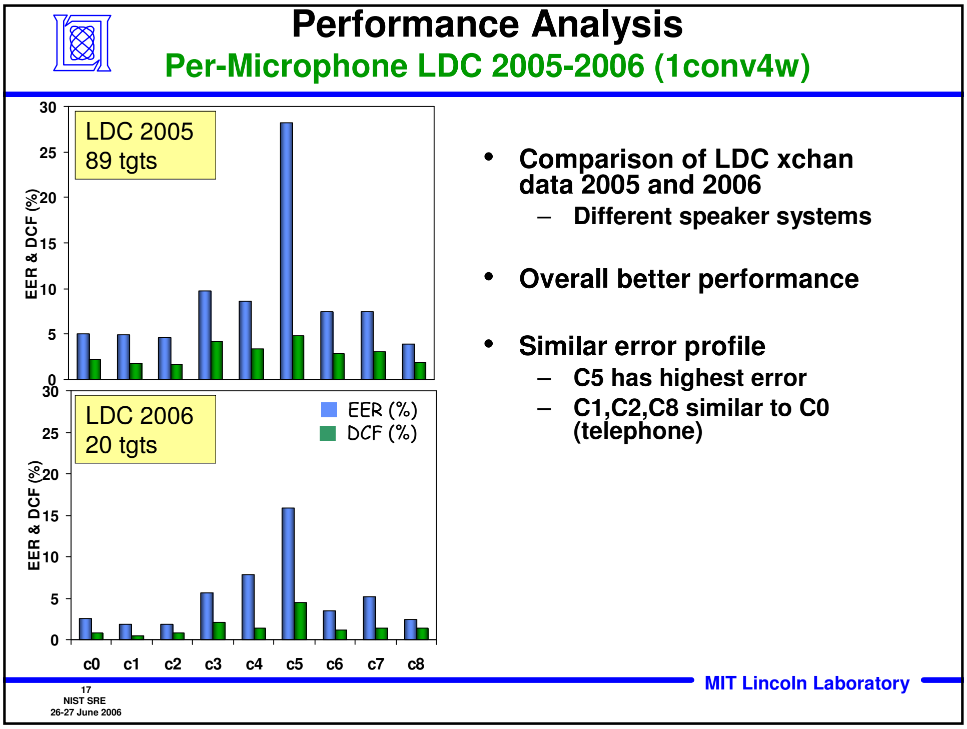


Performance Analysis

Enhancement with LLEnhance

- C3 example from ICSI
- Processed with LLEnhance toolkit for wideband noise reduction
- Will better SNR lead to reduced error on this microphone?

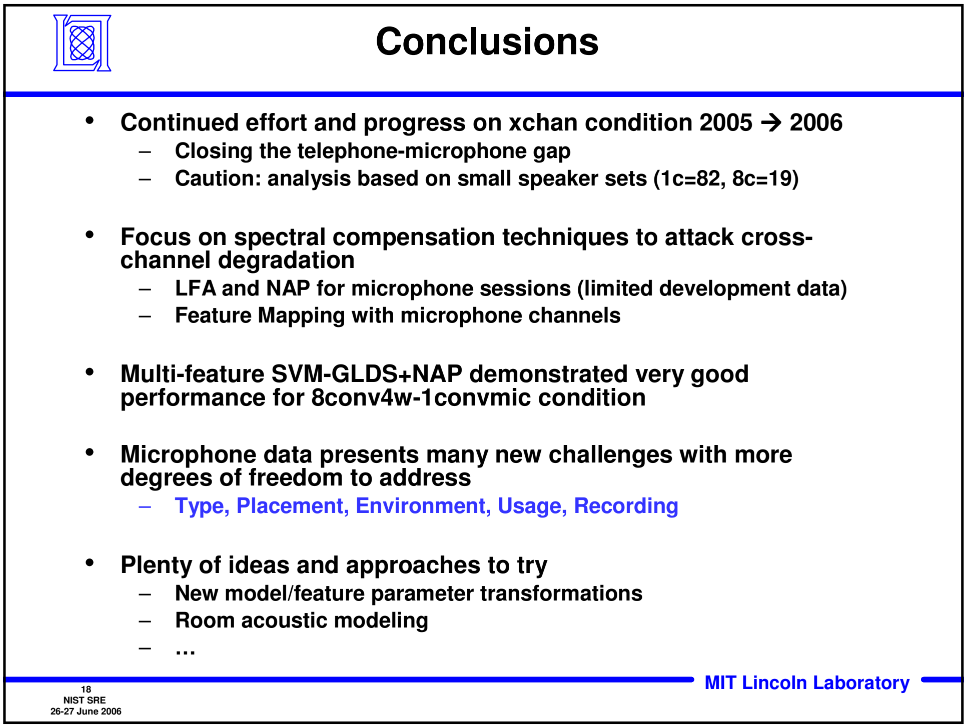

c3_icis_underrecord.wav [Configuration: Waveform]
SNR = 15 dB
Hz
3000
2500
2000
1500
1000
500
30992
-32767
time
c3_icis_underrecord_NoiseReduced.wav [Configuration: Waveform]
SNR = 25 dB
31528
-31586



Performance Analysis

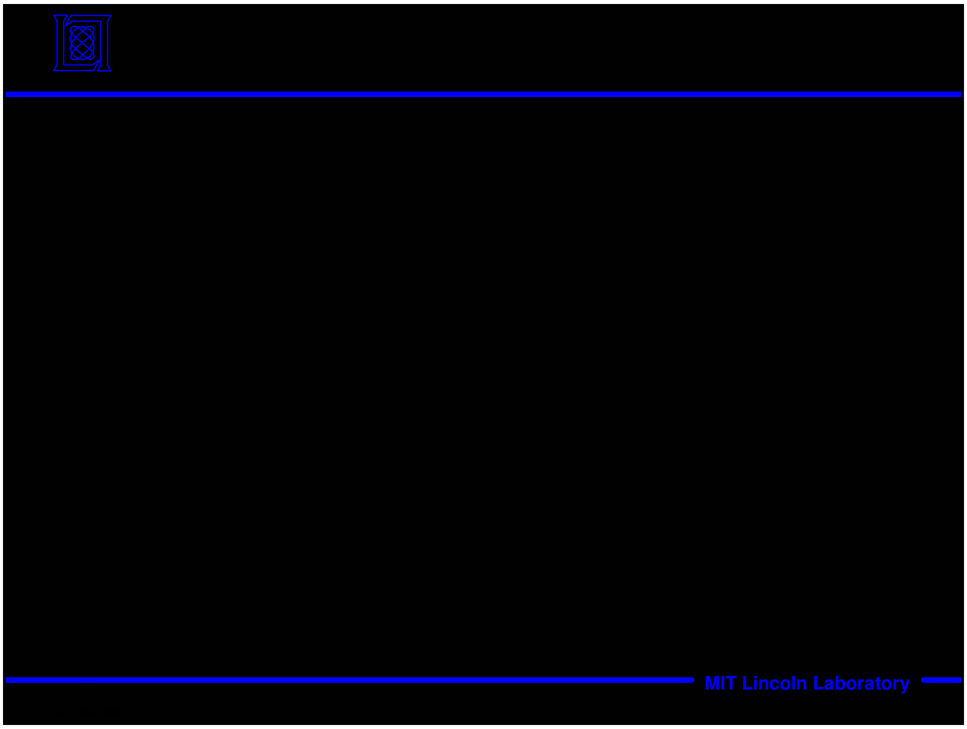
Per-Microphone LDC 2005-2006 (1conv4w)

- Comparison of LDC xchan data 2005 and 2006
  - Different speaker systems
- Overall better performance
- Similar error profile
  - C5 has highest error
  - C1,C2,C8 similar to C0 (telephone)

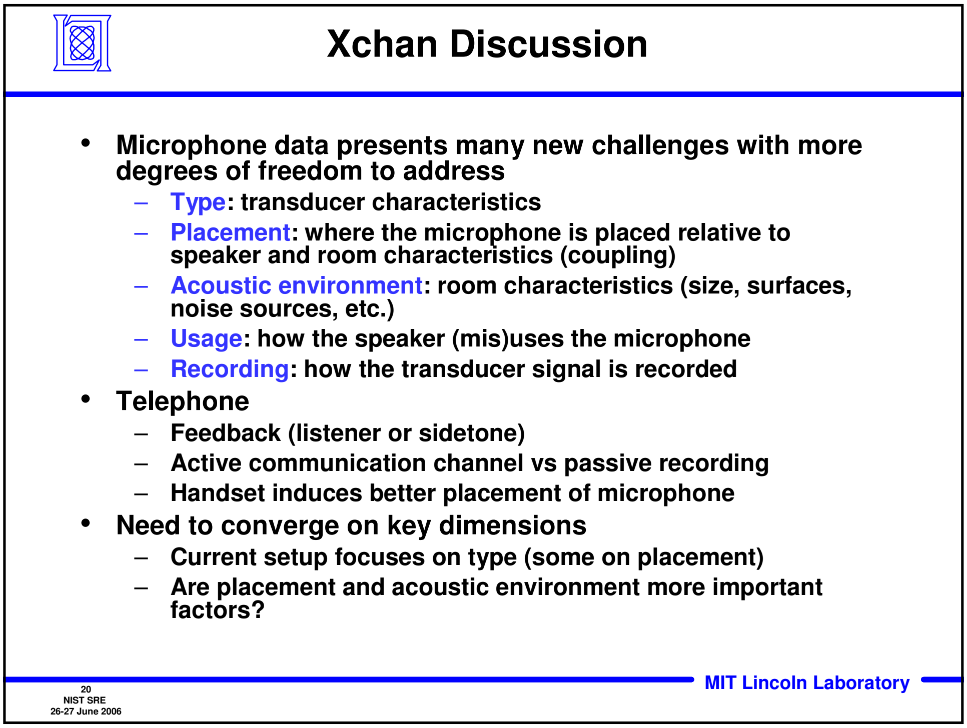


Conclusions

- Continued effort and progress on xchan condition 2005 → 2006
  - Closing the telephone-microphone gap
  - Caution: analysis based on small speaker sets (1c=82, 8c=19)
- Focus on spectral compensation techniques to attack cross-channel degradation
  - LFA and NAP for microphone sessions (limited development data)
  - Feature Mapping with microphone channels
- Multi-feature SVM-GLDS+NAP demonstrated very good performance for 8conv4w-1convmic condition
- Microphone data presents many new challenges with more degrees of freedom to address
  - Type, Placement, Environment, Usage, Recording
- Plenty of ideas and approaches to try
  - New model/feature parameter transformations
  - Room acoustic modeling
  - …



MIT Lincoln Laboratory

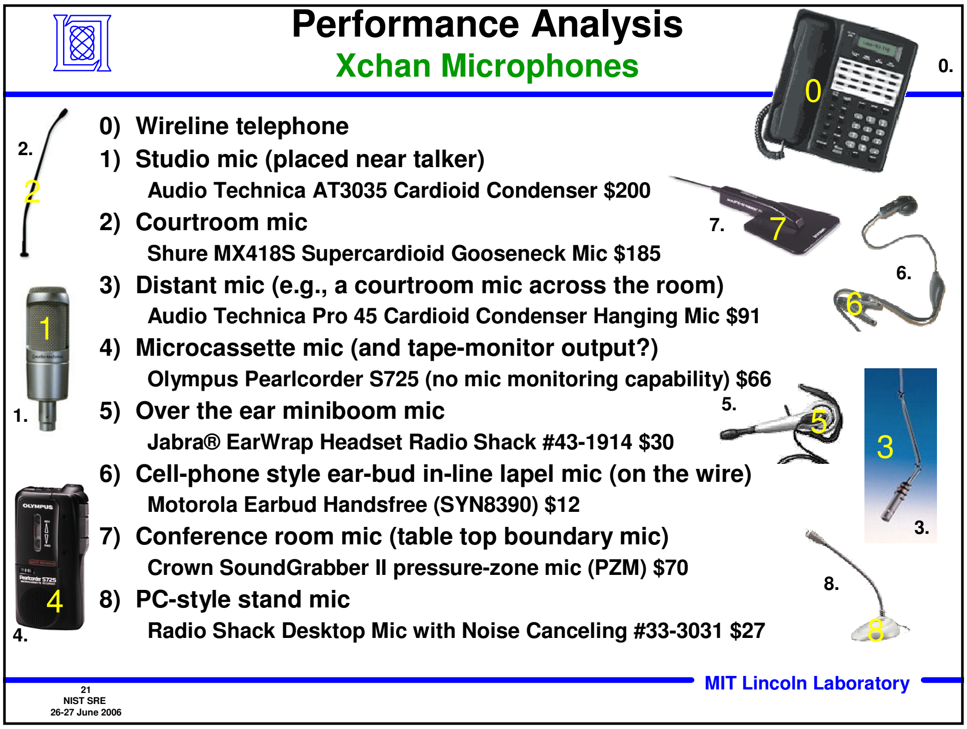


Xchan Discussion

- Microphone data presents many new challenges with more degrees of freedom to address
  - Type: transducer characteristics
  - Placement: where the microphone is placed relative to speaker and room characteristics (coupling)
  - Acoustic environment: room characteristics (size, surfaces, noise sources, etc.)
  - Usage: how the speaker (mis)uses the microphone
  - Recording: how the transducer signal is recorded
- Telephone
  - Feedback (listener or sidetone)
  - Active communication channel vs passive recording
  - Handset induces better placement of microphone
- Need to converge on key dimensions
  - Current setup focuses on type (some on placement)
  - Are placement and acoustic environment more important factors?

20
NIST SRE
26-27 June 2006

MIT Lincoln Laboratory



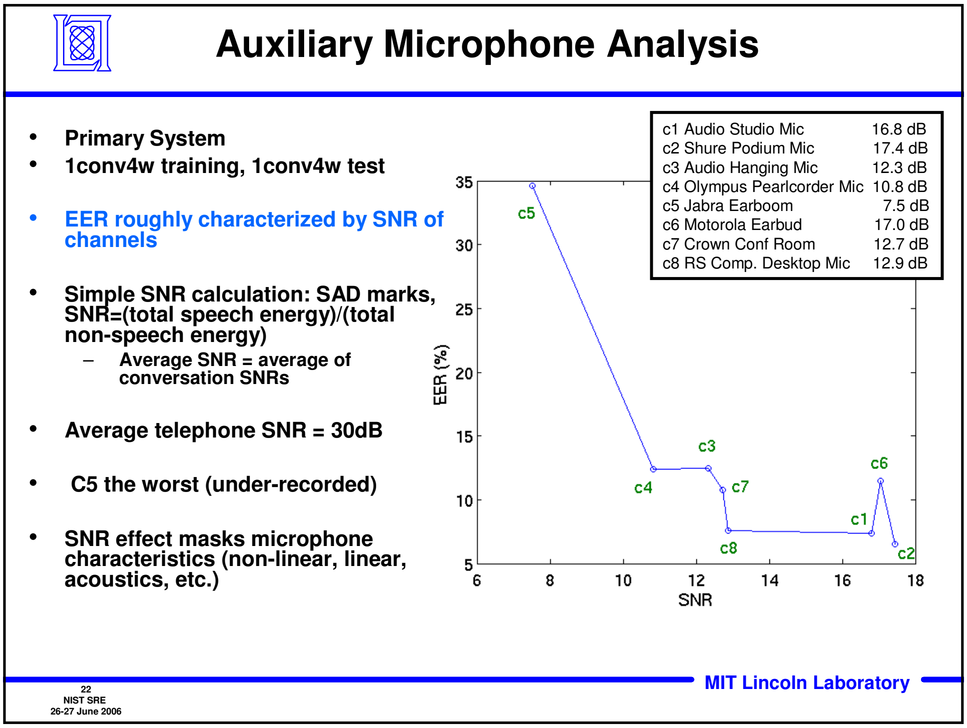
Performance Analysis

Xchan Microphones

0) Wireline telephone
1) Studio mic (placed near talker)
   Audio Technica AT3035 Cardioid Condenser $200
2) Courtroom mic
   Shure MX418S Supercardioid Gooseneck Mic $185
3) Distant mic (e.g., a courtroom mic across the room)
   Audio Technica Pro 45 Cardioid Condenser Hanging Mic $91
4) Microcassette mic (and tape-monitor output?)
   Olympus Pearlcorder S725 (no mic monitoring capability) $66
5) Over the ear miniboom mic
   Jabra® EarWrap Headset Radio Shack #43-1914 $30
6) Cell-phone style ear-bud in-line lapel mic (on the wire)
   Motorola Earbud Handsfree (SYN8390) $12
7) Conference room mic (table top boundary mic)
   Crown SoundGrabber II pressure-zone mic (PZM) $70
8) PC-style stand mic
   Radio Shack Desktop Mic with Noise Canceling #33-3031 $27

MIT Lincoln Laboratory

21
NIST SRE
26-27 June 2006



Auxiliary Microphone Analysis

- Primary System
- 1conv4w training, 1conv4w test
- EER roughly characterized by SNR of channels
- Simple SNR calculation: SAD marks, SNR=(total speech energy)/(total non-speech energy)
  - Average SNR = average of conversation SNRs
- Average telephone SNR = 30dB
- C5 the worst (under-recorded)
- SNR effect masks microphone characteristics (non-linear, linear, acoustics, etc.)

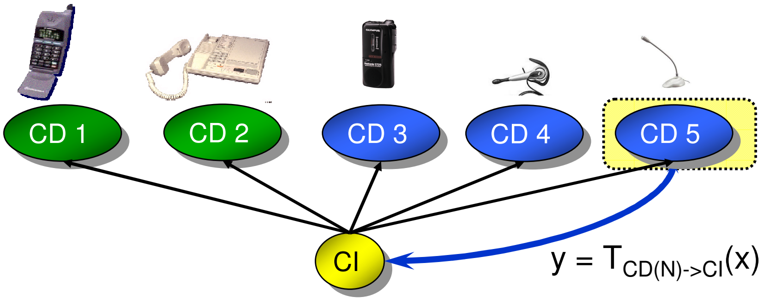
c1 Audio Studio Mic	16.8 dB
c2 Shure Podium Mic	17.4 dB
c3 Audio Hanging Mic	12.3 dB
c4 Olympus Pearlcorder Mic	10.8 dB
c5 Jabra Earboom	7.5 dB
c6 Motorola Earbud	17.0 dB
c7 Crown Conf Room	12.7 dB
c8 RS Comp. Desktop Mic	12.9 dB

MIT Lincoln Laboratory

22
NIST SRE
26-27 June 2006

Feature Domain Compensation

Feature Mapping



$y = T_{CD(N)\text{->}CI}(x)$