Channel Compensation for Speaker Recognition Using MAP Adapted PLDA and Denoising DNNs

Frederick Richardson, Brian Nemsick and Douglas Reynolds

Odyssey 2016

June 22, 2016

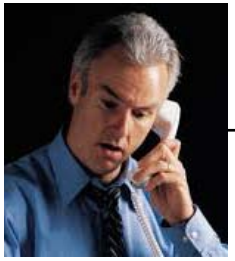
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Outline

- **Multi-channel speaker recognition using Mixer data**
- **Baseline i-vector system**
- **MAP adapted PLDA**
- **DNN channel compensation**
- **Hybrid i-vector system**
- **Results**
- **Conclusions**

Microphone Speaker Recognition



Telephone Speech

Speaker Recognition

Good Performance

Telephone Hyper Parameters

Microphone Speech

Speaker Recognition

Poor Performance

- Telephone systems generally perform poorly on mic data
- We look at two approaches to address this problem:
  - Adapting telephone hyper parameters to microphone data
  - Transforming microphone data to look like telephone data

Channel Compensation Approaches

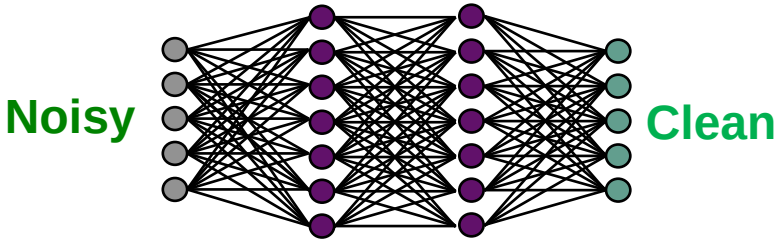
MAP adapted PLDA

- I-vectors are adapted
- Small performance gain
- Calibration issues
- Can use standard i-vectors

$$\Sigma_{adapt} = \lambda\Sigma_{tel} + (1-\lambda)\Sigma_{mic}$$

DNN enhancement

- Features are transformed
- Substantial performance gain
- Robust / better calibrated



- DNN enhancement performs better than MAP adaptation

Microphone Speaker Recognition

- Telephone data used to train speaker recognition system
  - Switchboard 1 and 2
  - 3100 speakers, 10 sessions
- Two corpora used in this work:

Mixer 2

- 2004 LDC collection
- 8 microphones + telephone
- Conversational speech
- 240 speakers, 4 sessions
- Used for development

Mixer 6

- 2008 LDC collection
- 14 microphones + telephone
- Conversations and interviews
- 540 speakers, 2 sessions
- Used for evaluation

- Both are parallel microphone corpora
- Rooms and speakers are different in each collection
  - Evaluating on unseen Mixer 6 channel conditions

Mixer Microphones

Mixer 1 and 2 (train)

Chan	Microphone
01	AT3035 (Audio Technica Studio Mic)
02	MX418S (Shure Gooseneck Mic)
03	Crown PZM Soundgrabber II
04	AT Pro45 (Audio Technica Hanging Mic)
05	Jabra Cellphone Earwrap Mic
06	Motorola Cellphone Earbud
07	Olympus Pearlcorder
08	Radio Shack Computer Desktop Mic

Mixer 6 (eval)

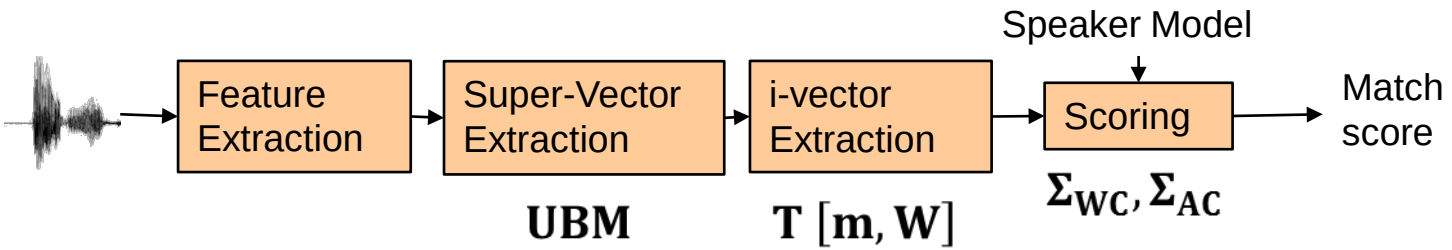
Chan	Microphone	Distance
02	**Subject Lavalier**	**8**
04	**Podium Mic**	**17**
10	R0DE NT6	21
05	**PZM Mic**	**22**
06	AT3035 Studio Mic	22
08	**Panasonic Camcorder**	**28**
11	Samson C01U	28
14	Lightspeed Headset On	34
07	**AT Pro45 Hanging Mic**	**62**
01	Interviewer Lavalier	77
03	Interviewer Headmic	77
12	AT815b Shotgun Mic	84
13	**Acoust Array Imagic**	**110**
09	R0DE NT6	124

- **All 8 Mixer 2 mics used**
- **6 mics from Mixer 6**
  - **Selected by distance (green)**
  - **Only evaluate same mic trials**

  (same mic for enrollment and test)

Baseline System

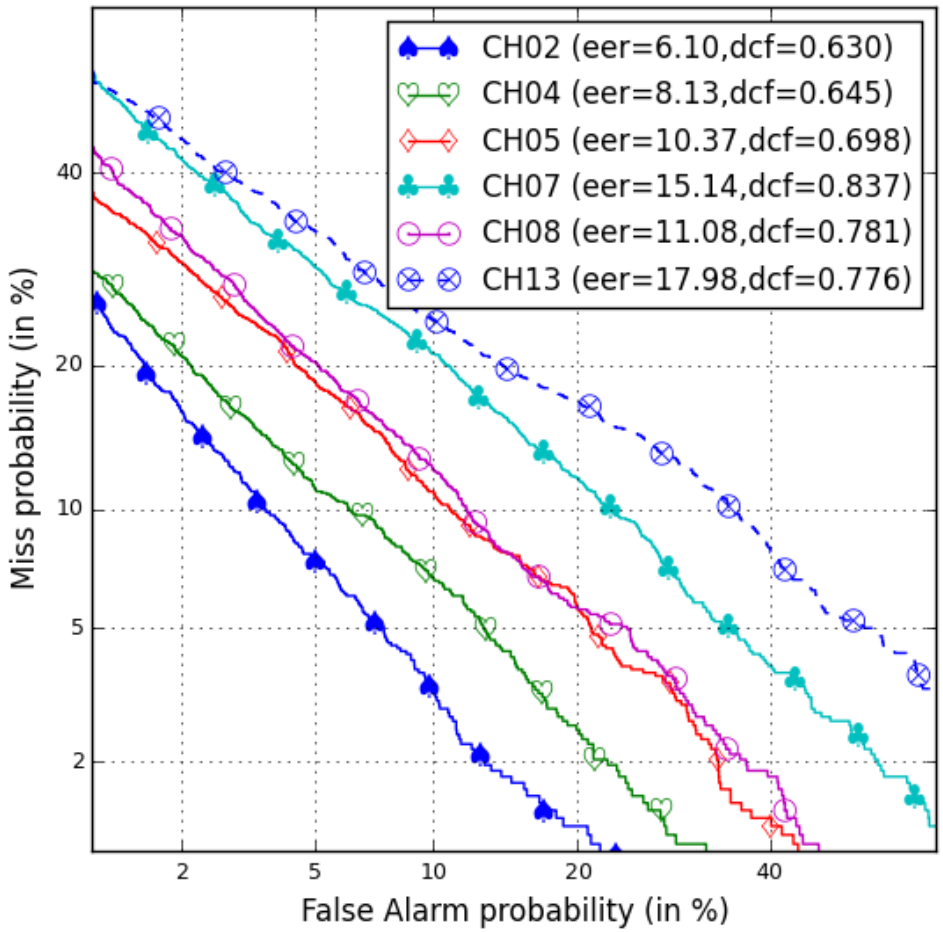
- All system trained on Switchboard 1 and 2 telephone speech
- I-vector PLDA system used for all experiments
- All systems use similar configuration:
  - 2048 Gaussian mixtures, 600 dimensional i-vectors
- Baseline system uses 40 MFCCs (including 20 deltas)



UBM	Universal background model
T	Total variability matrix
m, W	Whitening parameters
Σ_{WC}, Σ_{AC}	Within-class and across-class covariance

Baseline Results on Mixer 6

Mixer 6 Microphone Results



SRE10 (CTS) vs Mixer 6 (MIC)

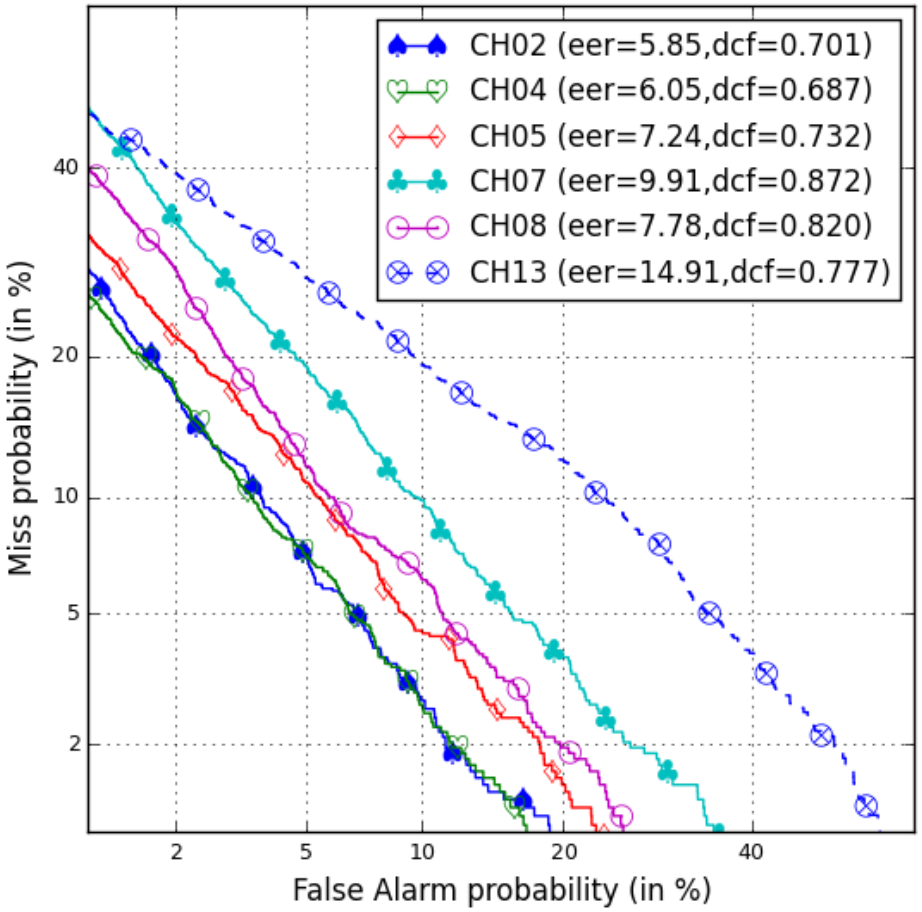
Test	EER	Min DCF
SRE10	5.77	0.662
Mixer 6 AVG	11.5	0.728
Mixer 6 POOL	18.8	0.875

- Switchboard trained system
- AVG uses threshold per channel
- POOL uses only one threshold
  - Reflects channel calibration
  - More practical
- Remaining results will use POOL

Baseline performs poorly (AVG) and is poorly calibrated (POOL)

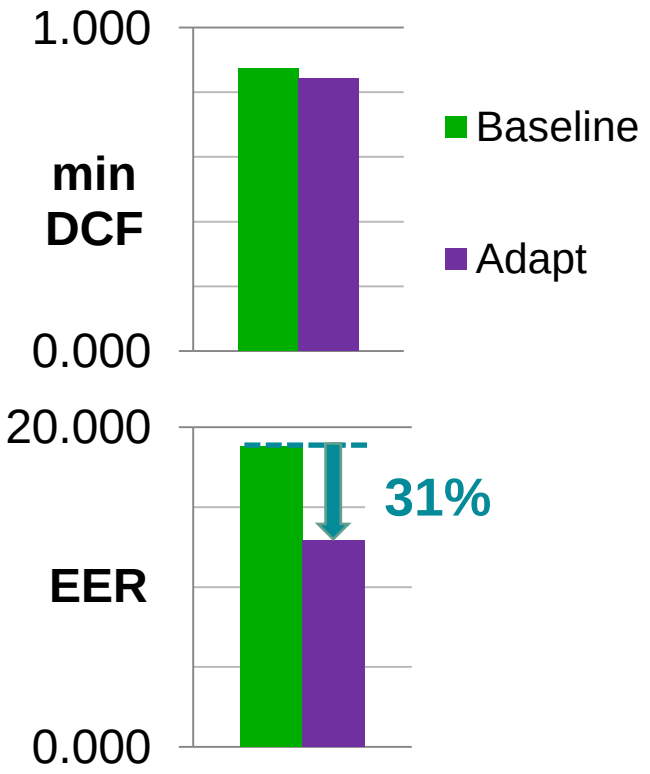
MAP Adapted PLDA Performance

Mixer 6 Results



Lambda = 0.5

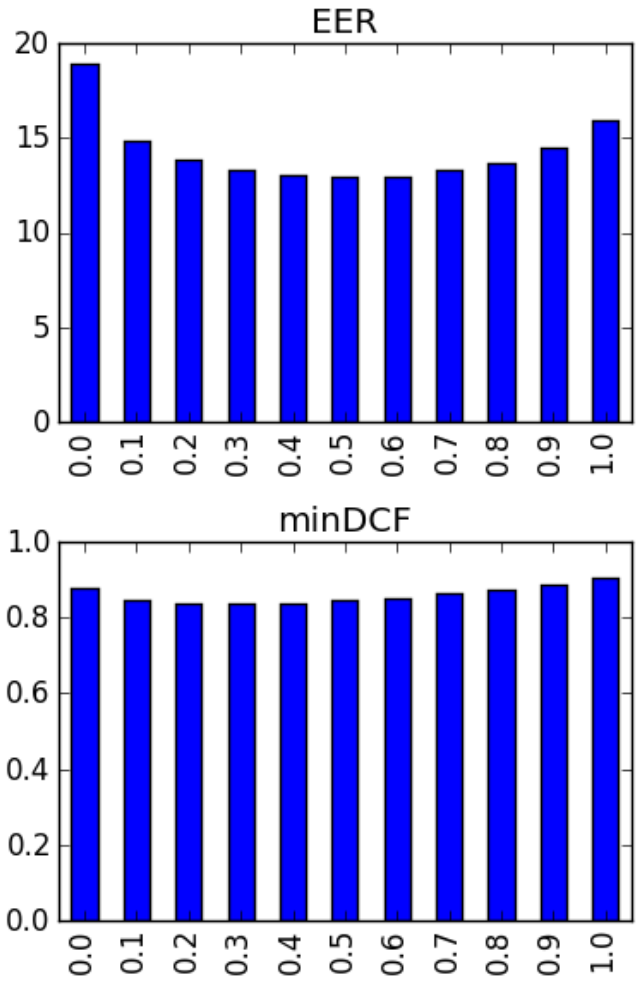
POOL Results

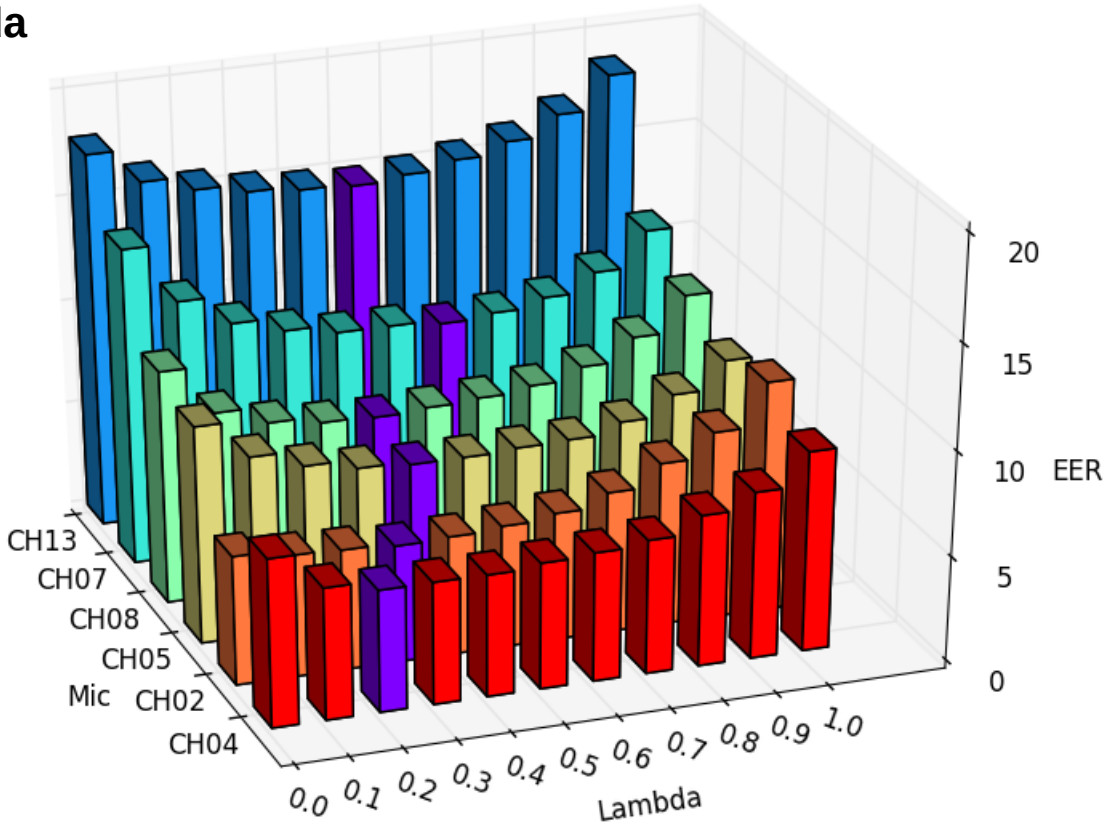


Big reduction in EER
But not Min DCF!

MAP Adapted PLDA – Tuning Lambda

Mixer 6 EER / min DCF vs. Lambda

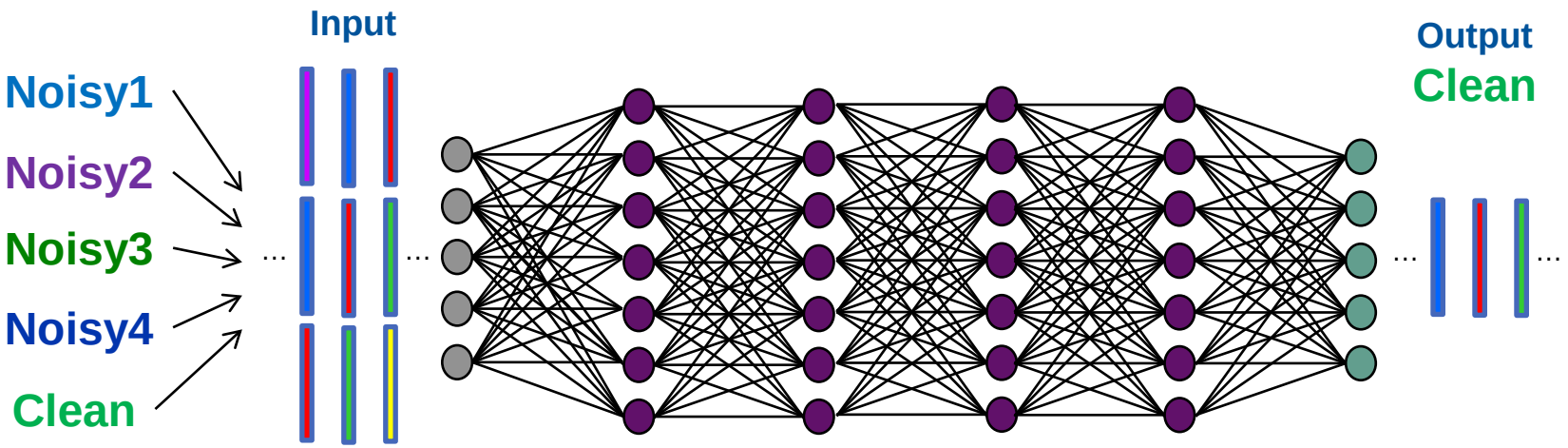




Lambda has a big impact on EER
But not on min DCF

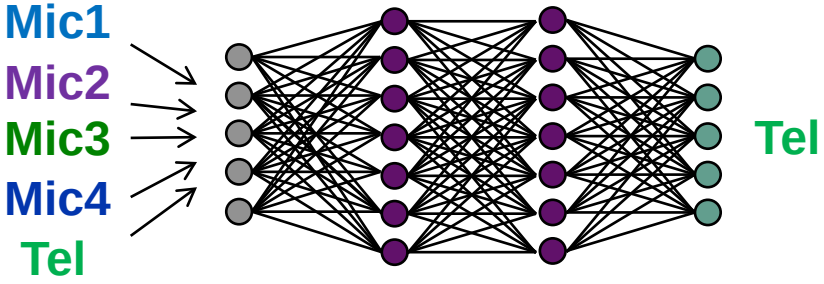
DNN Speech Enhancement

- Another approach to enhancement is to use a DNN
- The DNN is trained as a regression
- Parallel clean and noisy data is needed for this
- Objective is to reconstructed clean data from a noisy version

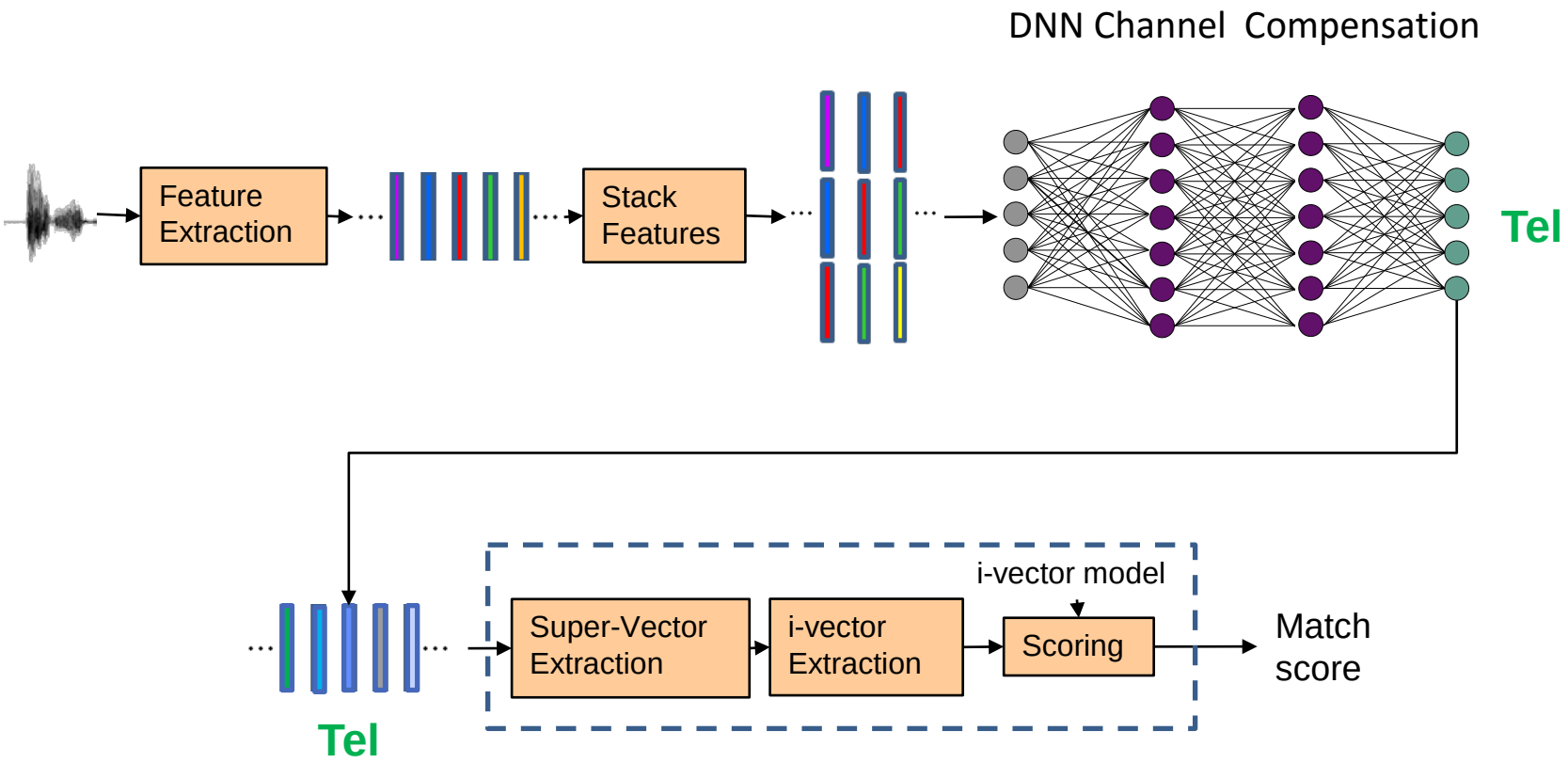


Speech Enhancement

- LDC Mixer data was collected over microphones in a room
- Different mics placed in different locations
- Clean data comes from telephone handset
- Expensive approach – limited to specific rooms and mics



Channel Compensation I-vector System

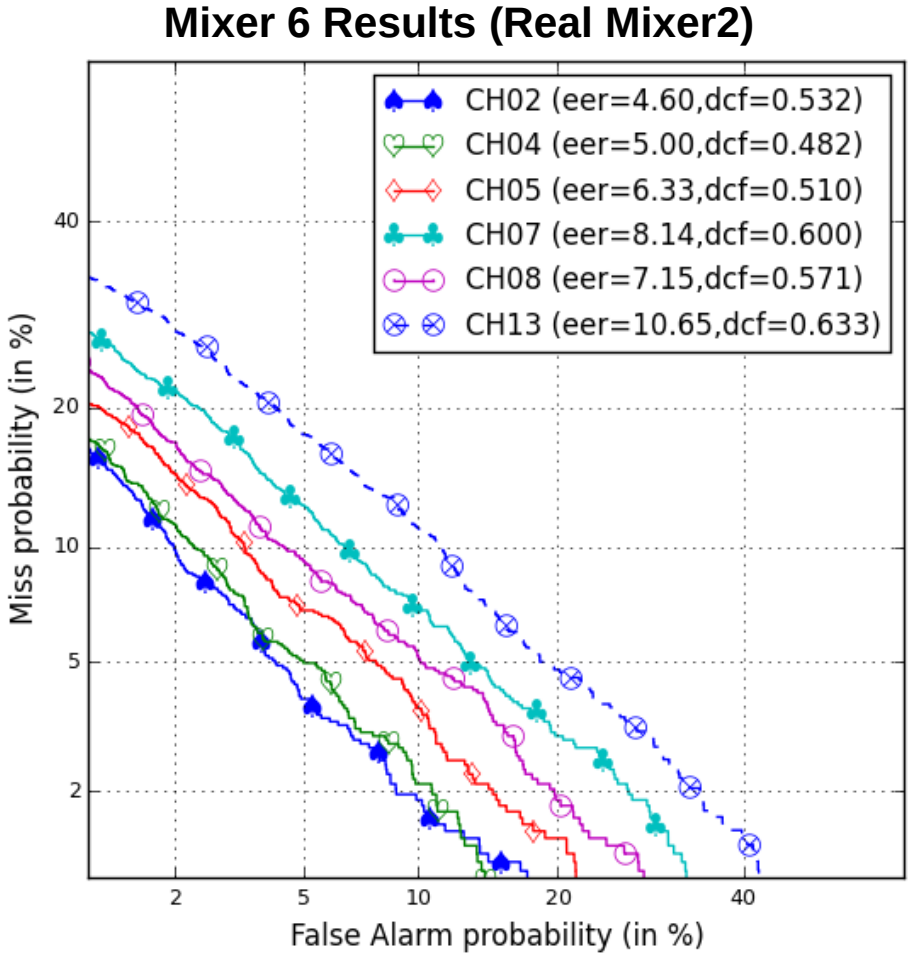


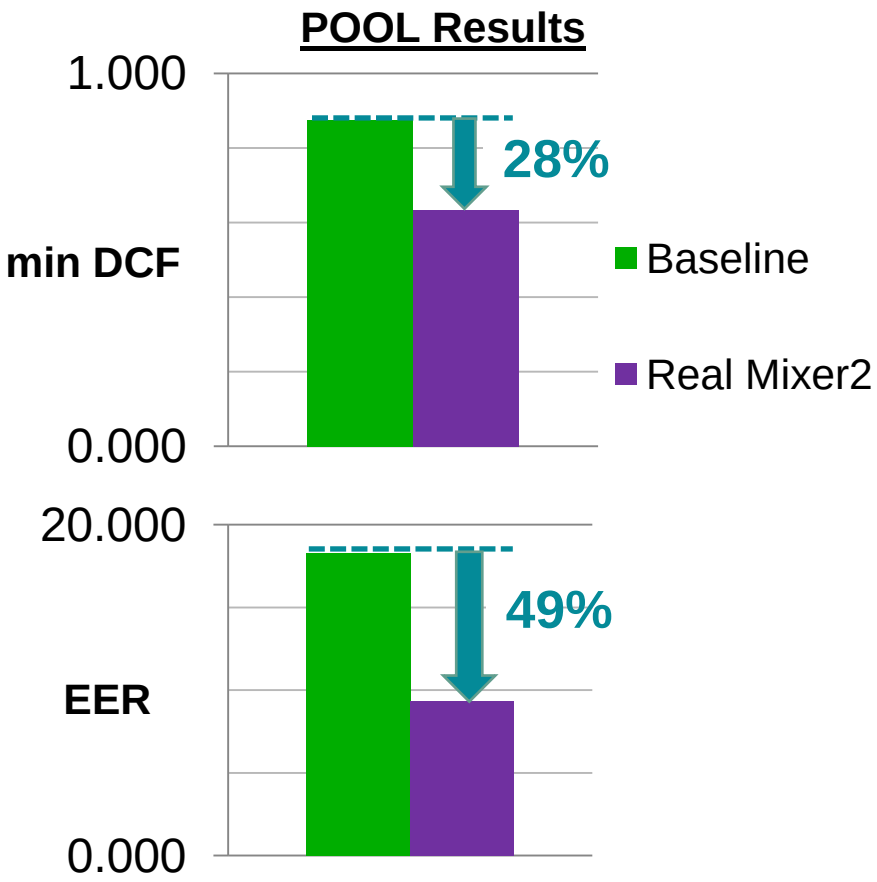
DNN Feature Enhancement

- DNN trained using Mixer 2 parallel data

- DNN has the following architecture
  - 40 MFCCs (which includes 20 delta MFCCs)
  - 5 layers and 2048 nodes / layer (5 x 2048)
  - 21 frame input (+/- 10 frames around center frame)
  - 1 frame output (center frame of clean channel)
  - Input is either clean or one of 8 noisy parallel versions

DNN Feature Enhancement Performance





Big reduction in EER and in Min DCF!

DNN Performance Tuning

- We found several things that impact performance:
- Log Mel frequency banks – these did not work as well as MFCCs
- Mean and variance normalization of input and output is critical
- DNN architecture has a big impact
- 2048 x 5 (nodes x layers) is best performing
  - But is much more expensive to train than 1024 x 5

POOL Mixer 6 Performance

DNN Arch	EER	Min DCF
512 x 5	11.4	0.711
1024 x 5	10.3	0.667
2048 x 5	8.16	0.633

Telephone Performance

- Map adapted PLDA does not perform well on telephone data
- DNN compensation gives a gain on telephone data
  - Almost 10% relative gain
- DNN compensation can be used without channel detection

SRE10 Telephone Performance

Task	EER	DCF
Baseline	5.77	0.662
MAP adapt PLDA	11.9	0.824
2048x5 DNN	5.20	0.615

Conclusions

- DNN channel compensation works very well
  - 28% reduction in Min DCF, 49% reduction in EER
- No loss on telephone data
  - Actually a small gain (~10%)
  - No need to detect channel or switch front-ends
- MAP adapted PLDA does not work as well
  - Gains at EER but does not improve min DCF
  - Performance issues on telephone data
  - But… easy to implement – only uses i-vectors
- Note that real parallel data is expensive to collect
  - Synthetic parallel data would be much more practical