

# **Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition**

**Xinhui Zhou, Daniel Garcia-Romero  
Ramani Duraiswami, Carol Espy-Wilson  
Shihab Shamma**

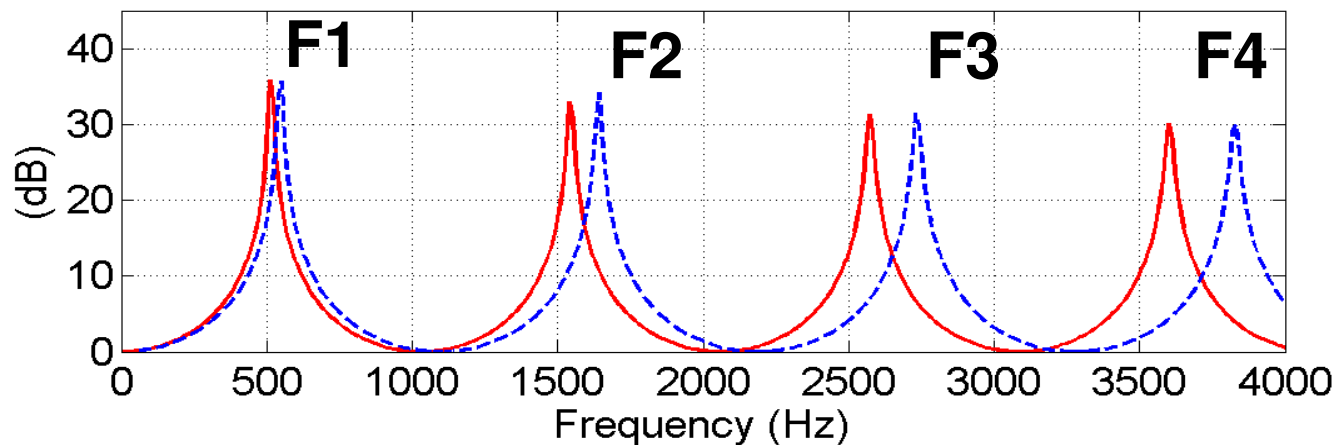
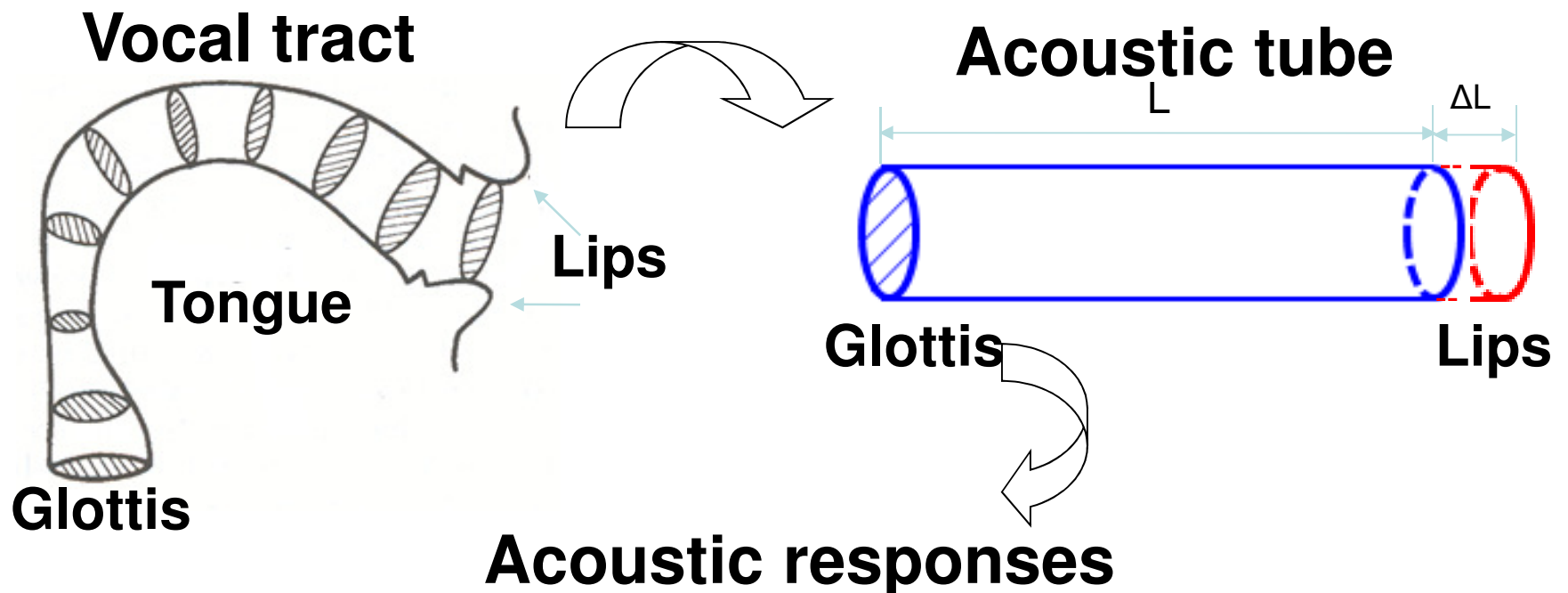
University of Maryland, College Park

**ASRU 2011**

# Introduction

- Mel-frequency cepstral coefficients (MFCC) have been dominantly used in both speaker recognition and speech recognition.
- This is counterintuitive since speech recognition and speaker recognition seek different types of information from speech.
- In speech production theory, speaker characteristics associated with structures of the vocal tract are reflected more in high frequency region of speech [1,2].

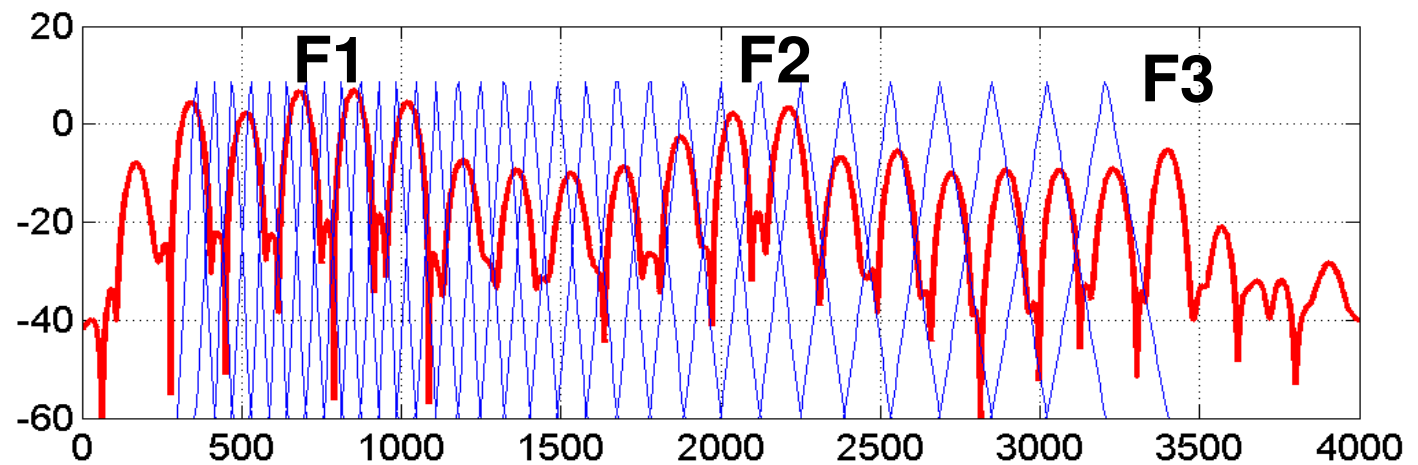
- **Example:** a change in vocal tract length results in more of a shift in higher formants.



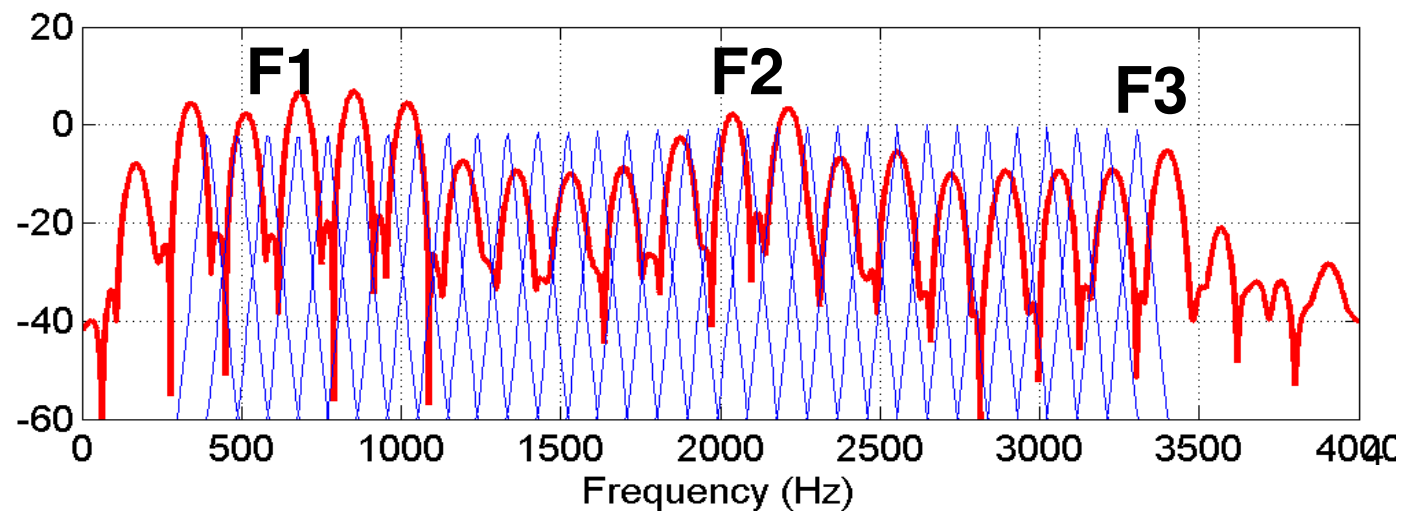
# Mel versus Linear Filterbank

- Linear filterbank has better resolution in higher frequency region.

Mel  
filterbank



Linear  
filterbank



# Objectives

- To compare the performances between MFCC and LFCC (linear frequency cepstral coefficients) on state-of-the-art back-end systems using the NIST 2010 speaker recognition evaluation (SRE) [3].
- To evaluate the noise (additive and convolutive) robustness of both features.
- Our long-term goal: to find an optimal frequency-warping function for speaker recognition.

# Feature Extraction

- MFCC and LFCC features are based on the revised functions in the RASTAMAT toolbox [4]. Both have the same parameters except for the frequency scale.
- Speech signal is band-limited to 300-3400 Hz. 32 filterbanks are used. The 19 cepstral coefficients plus its delta makes the 38 dimension feature vector.
- The MFCC/LFCC code is available online at [http://www.glue.umd.edu/~zxinhui/LFCC\\_ASRU2011](http://www.glue.umd.edu/~zxinhui/LFCC_ASRU2011)

# Two State-of-the-art Back-end Systems

- The Joint Factor Analysis (JFA) system [5]
  - Two separate gender-dependent universal background models (UBM) with 2048 mixtures and hyper-parameter sets gender-dependent
  - The eigenvoice and eigenchannel matrices were trained independently
- The I-vector Probabilistic Linear Discriminant Analysis (PLDA) system [6]:
  - Both the i-vector extractor and the PLDA systems were gender-dependent. Baum-Welch sufficient statistics were collected using the same 2048 mixture UBMs as in JFA.
  - The subspace matrix  $T$  with 400 columns

# NIST SRE10

- About 6.5 million trials were tested, each belonging to one of the nine conditions :

**C1:** Interview-Interview same mic, **C2:** Interview-Interview diff mic

**C3:** Interview -Phonecall

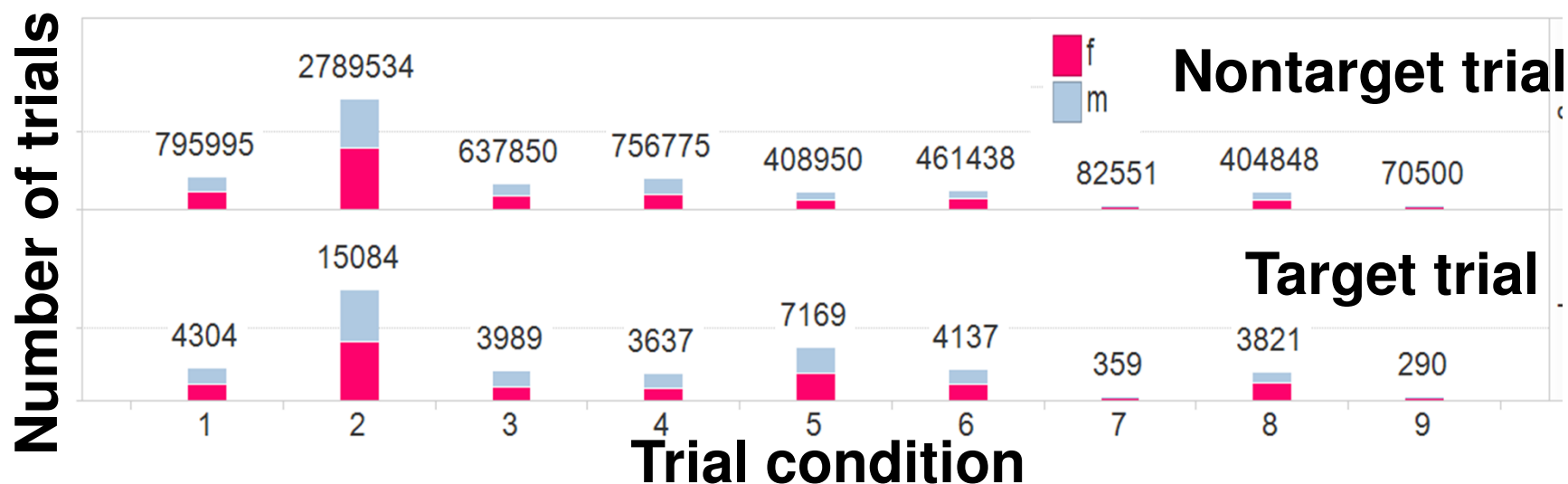
**C4:** Interview -Phonecall recorded by mic, **C5:** Phonecall-Phonecall

**C6:** Phonecall-Phonecall in high vocal effort

**C7:** Phonecall-Phonecall in high vocal effort (both recorded by Mic)

**C8:** Phonecall-Phonecall in low vocal effort

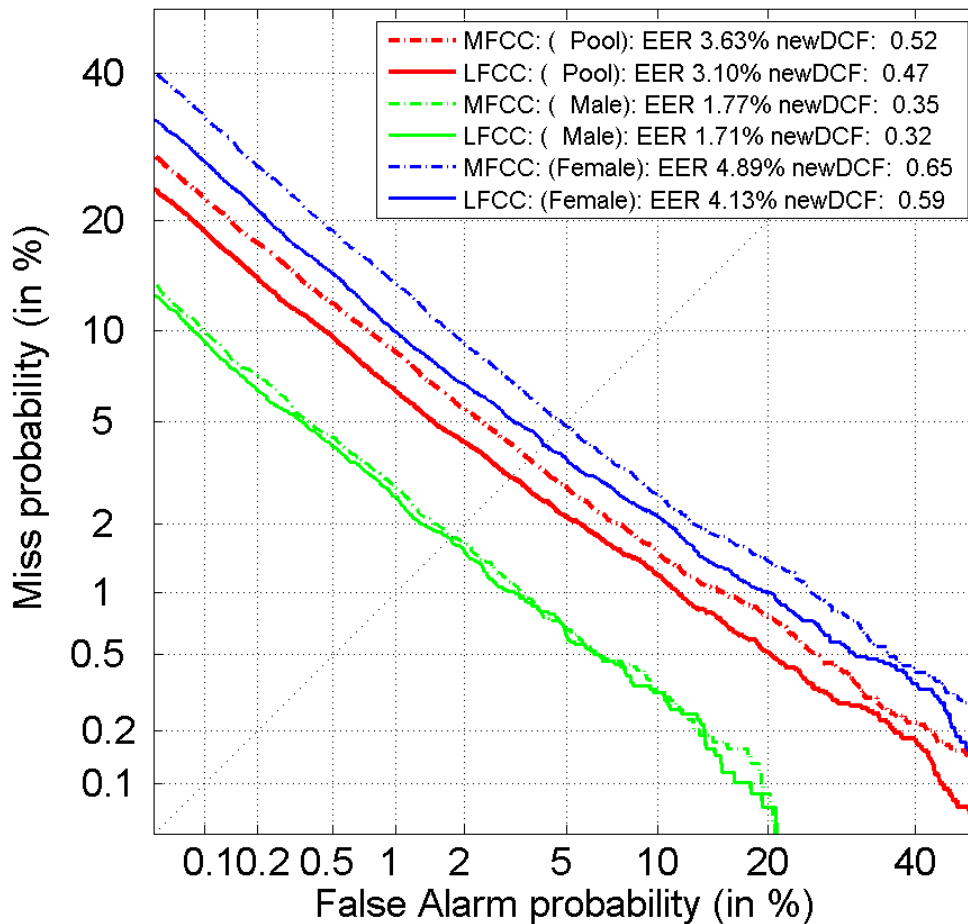
**C9:** Phonecall-Phonecall in low vocal effort (both recorded by Mic)



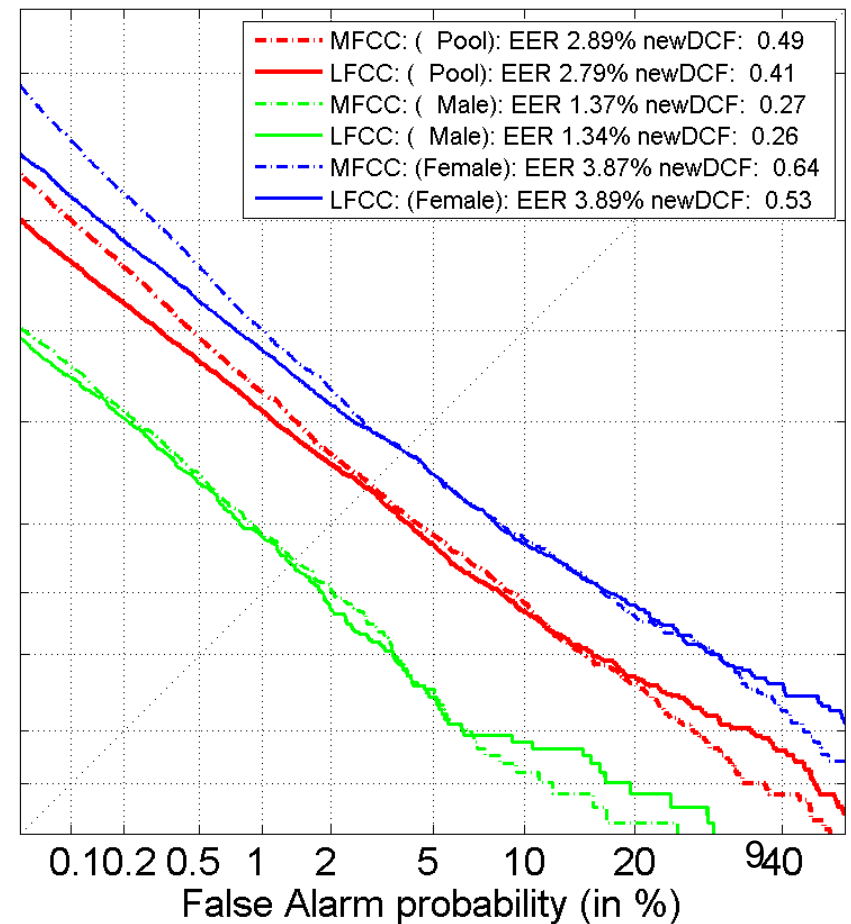
# DET Curves in NIST SRE10

(C2: Interview-Interview diff mic)

## JFA



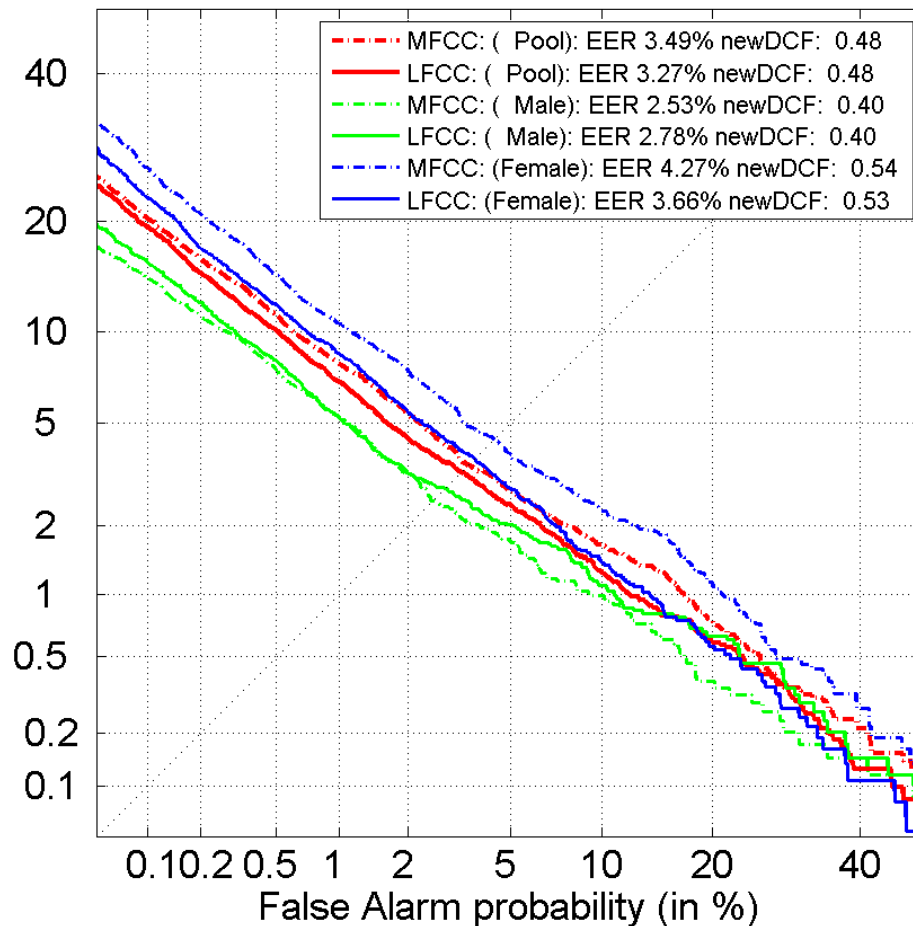
## PLDA



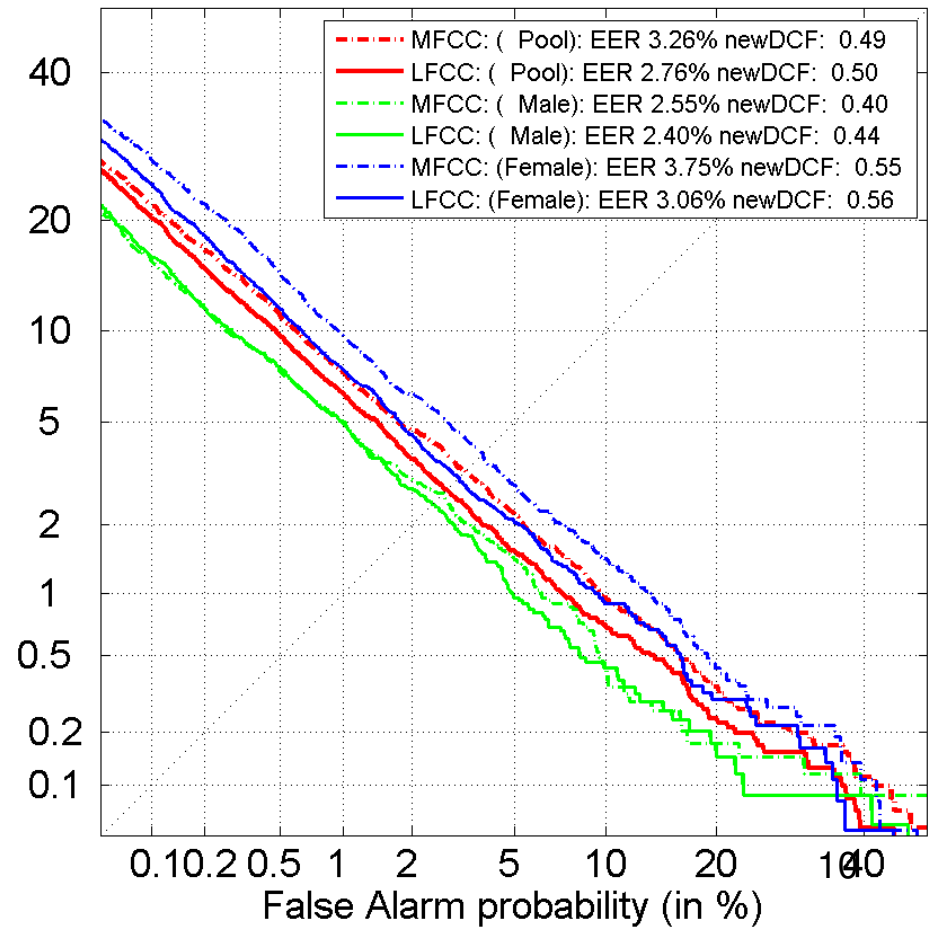
# DET Curves in NIST SRE10

(C5: Phonecall-Phonecall)

## JFA



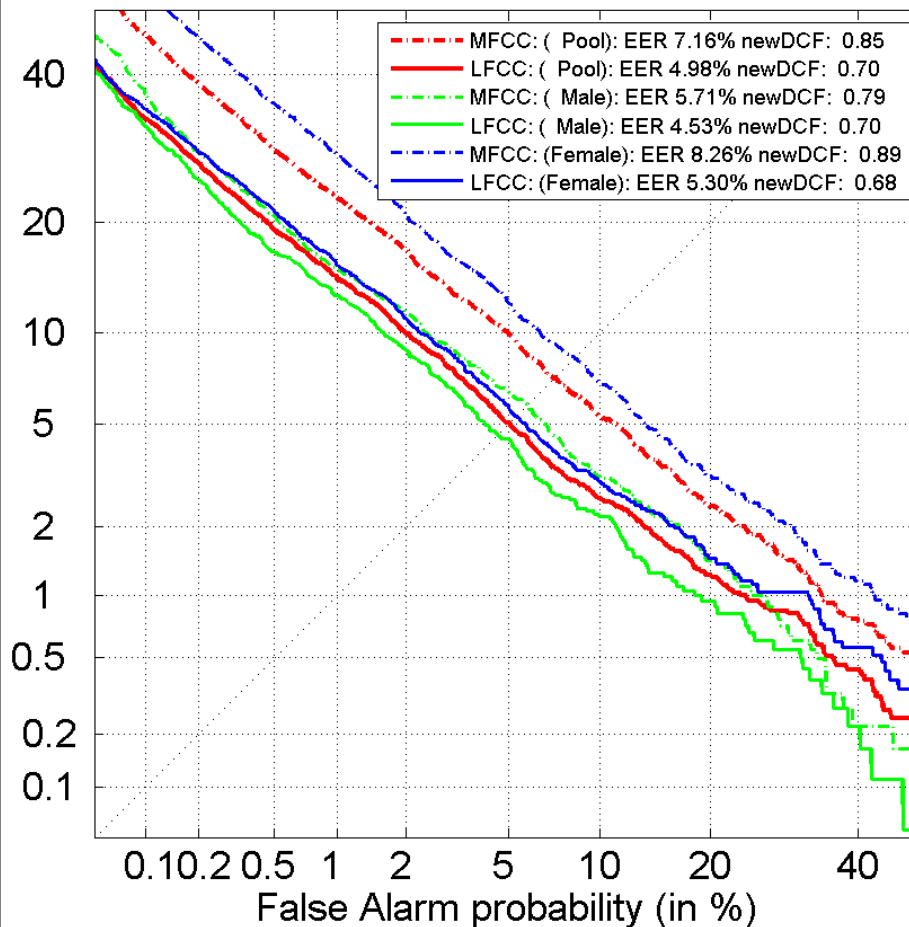
## PLDA



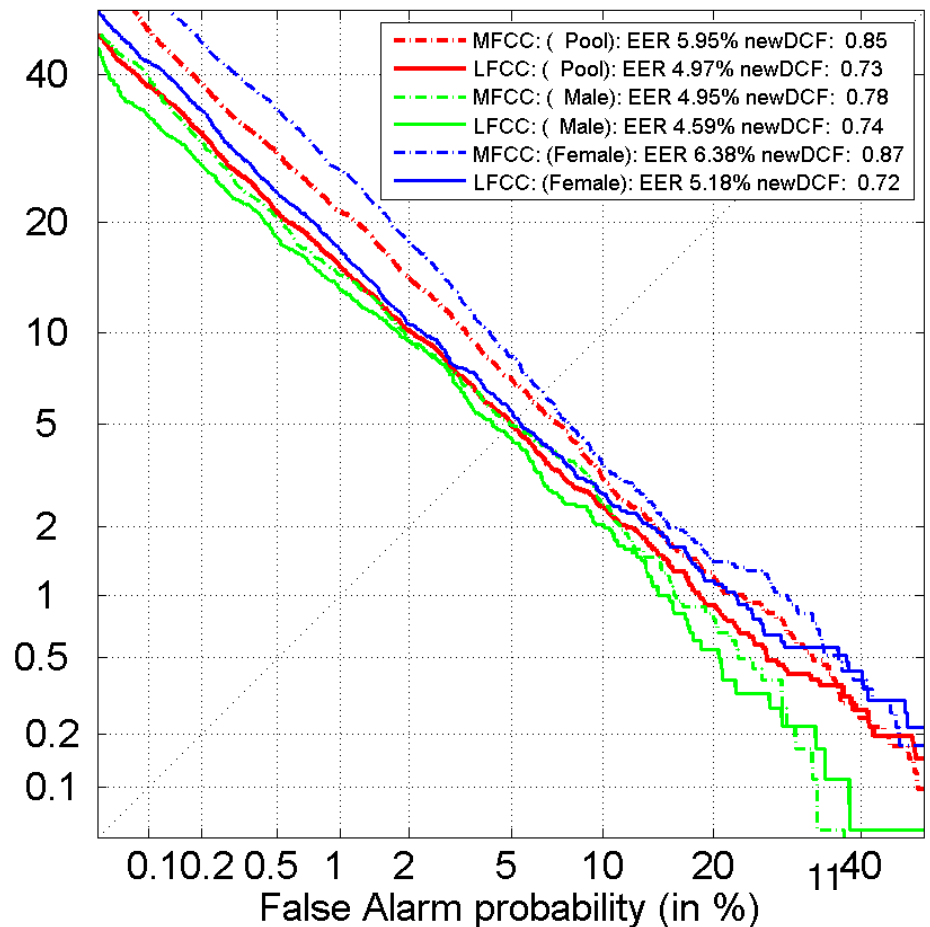
# DET Curves in NIST SRE10

(C6: Phonecall-Phonecall in high vocal effort)

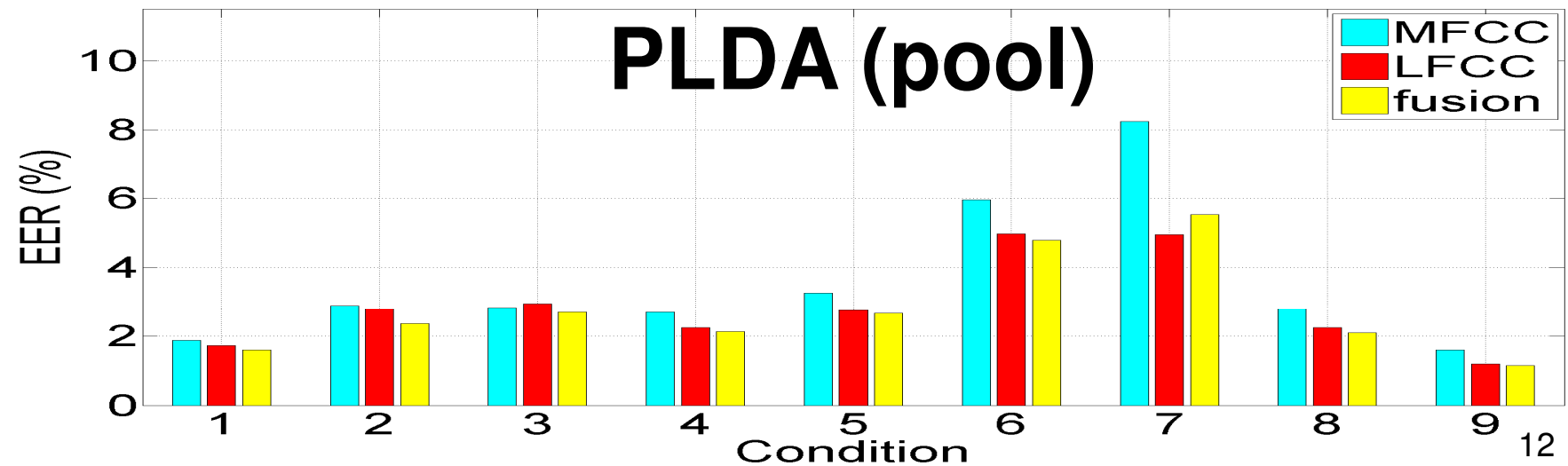
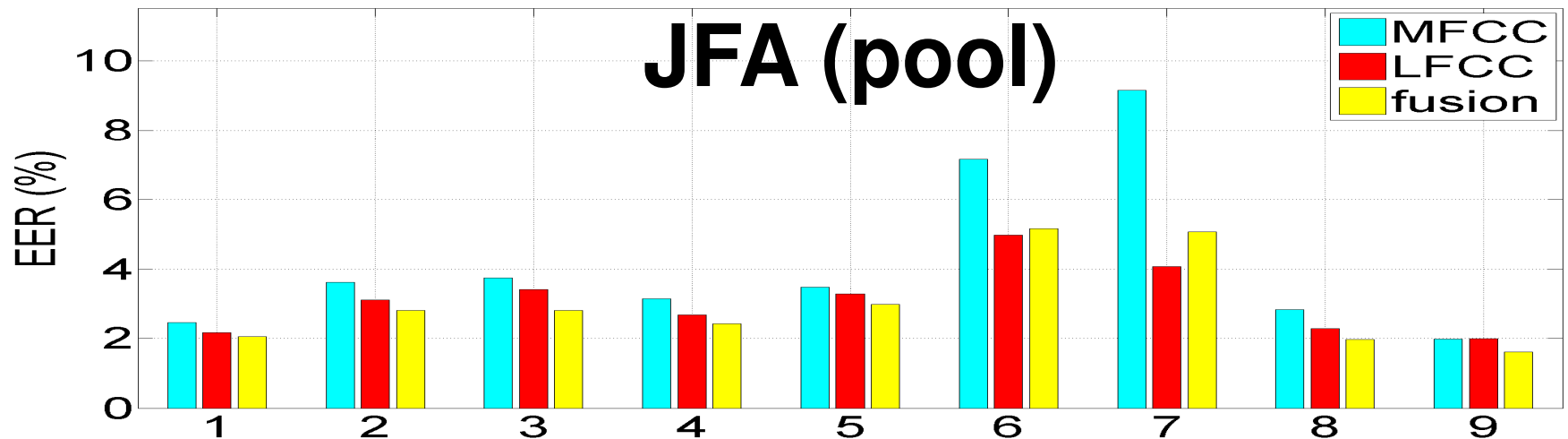
## JFA



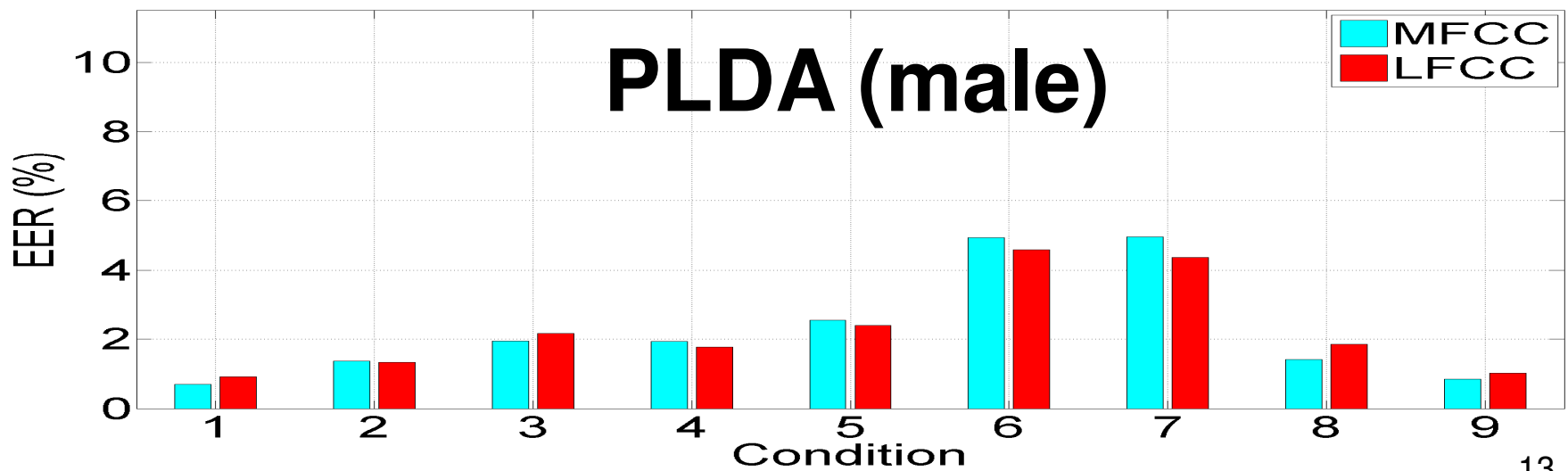
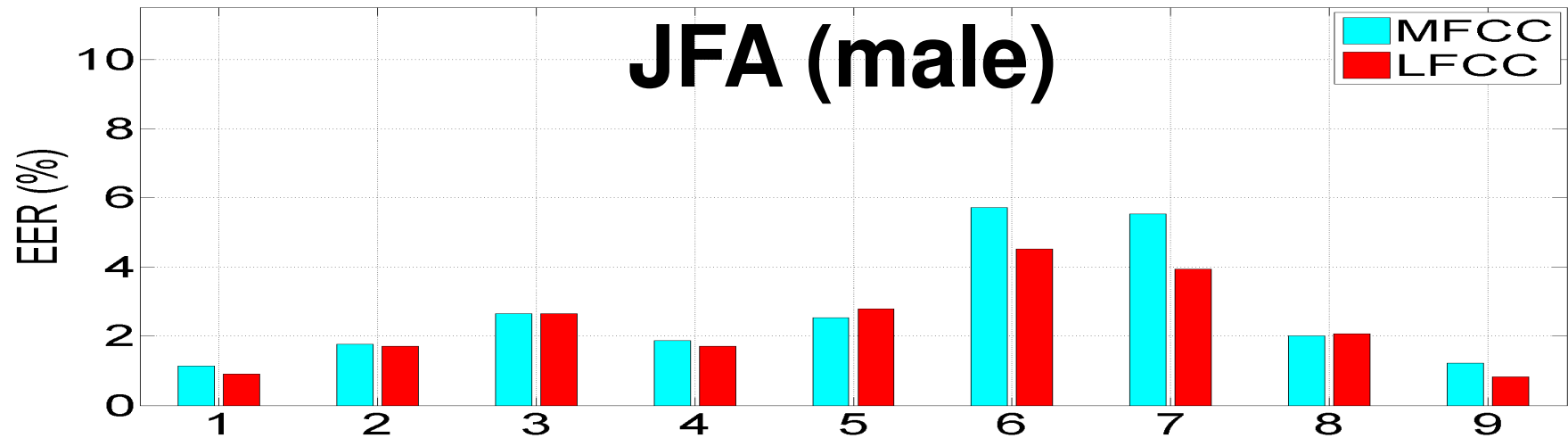
## PLDA



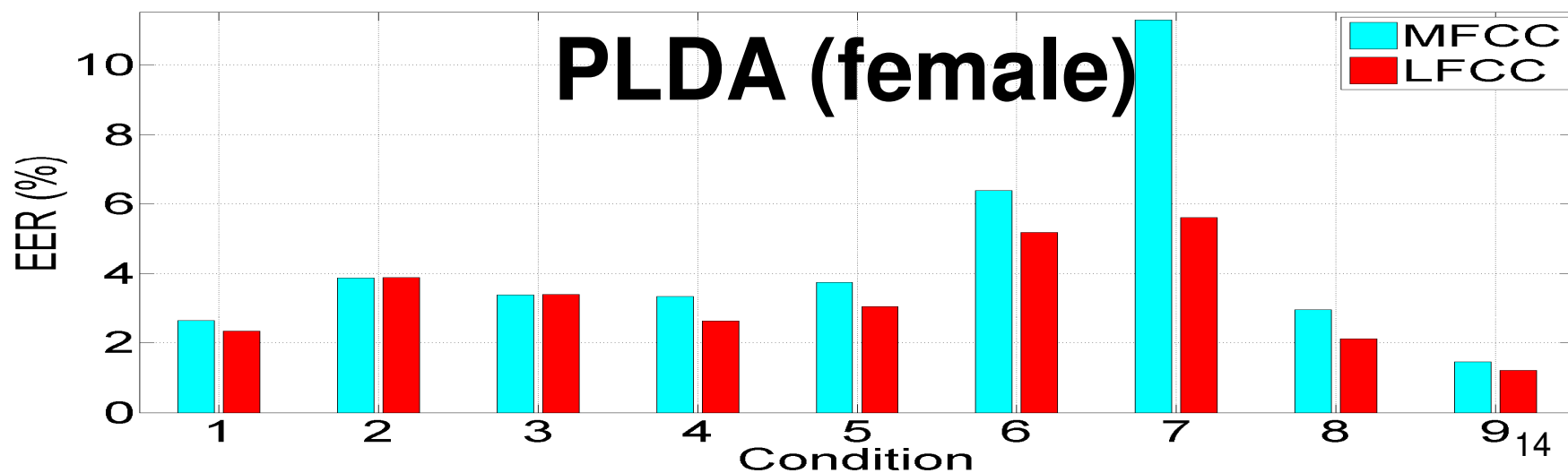
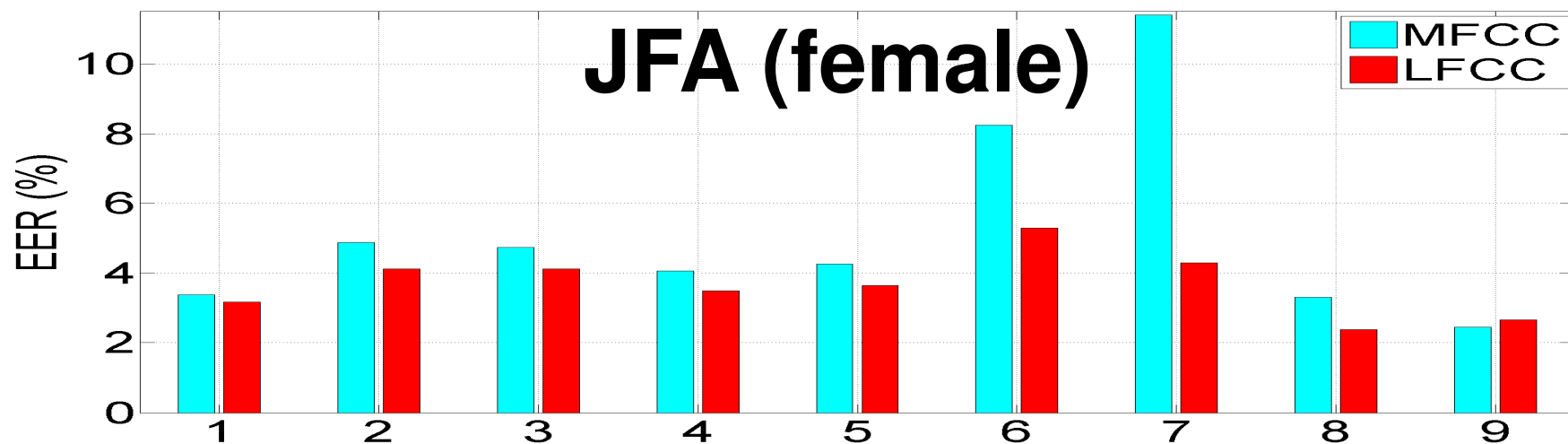
# Equal Error Rates (all trials)



# Equal Error Rates (male trials)

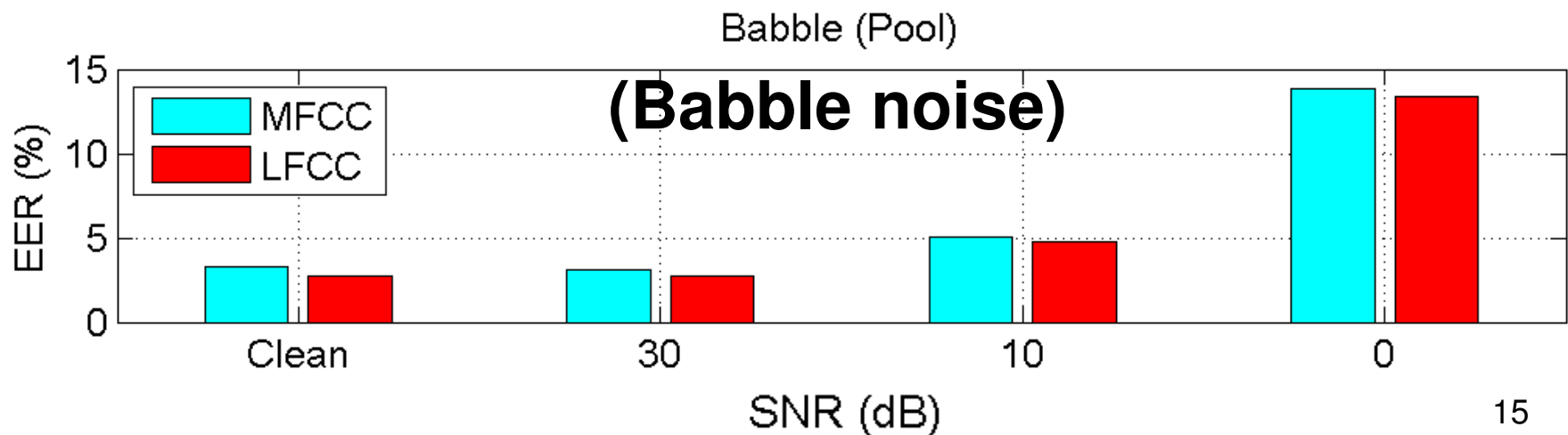
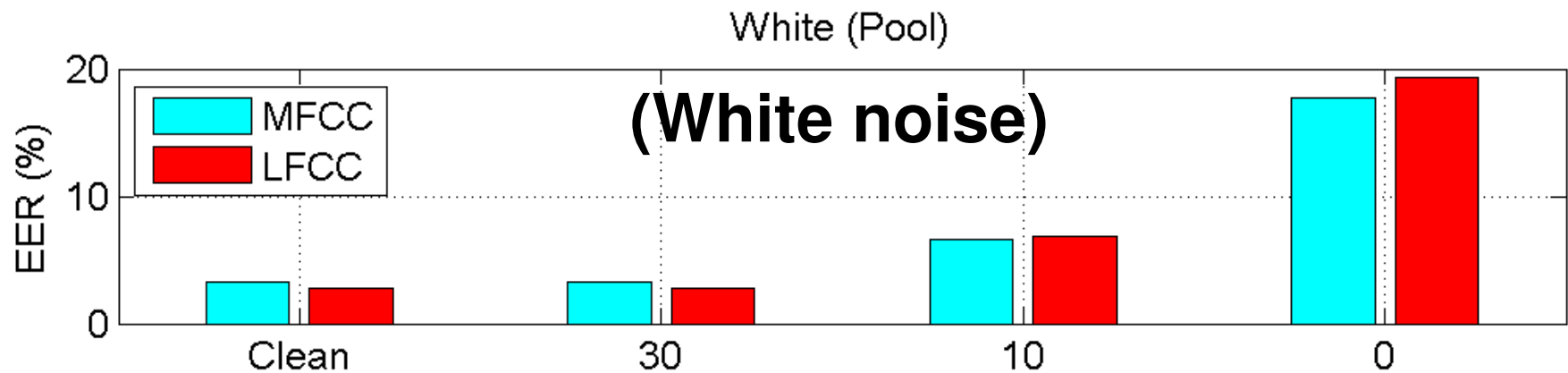


# Equal Error Rates (female trials)



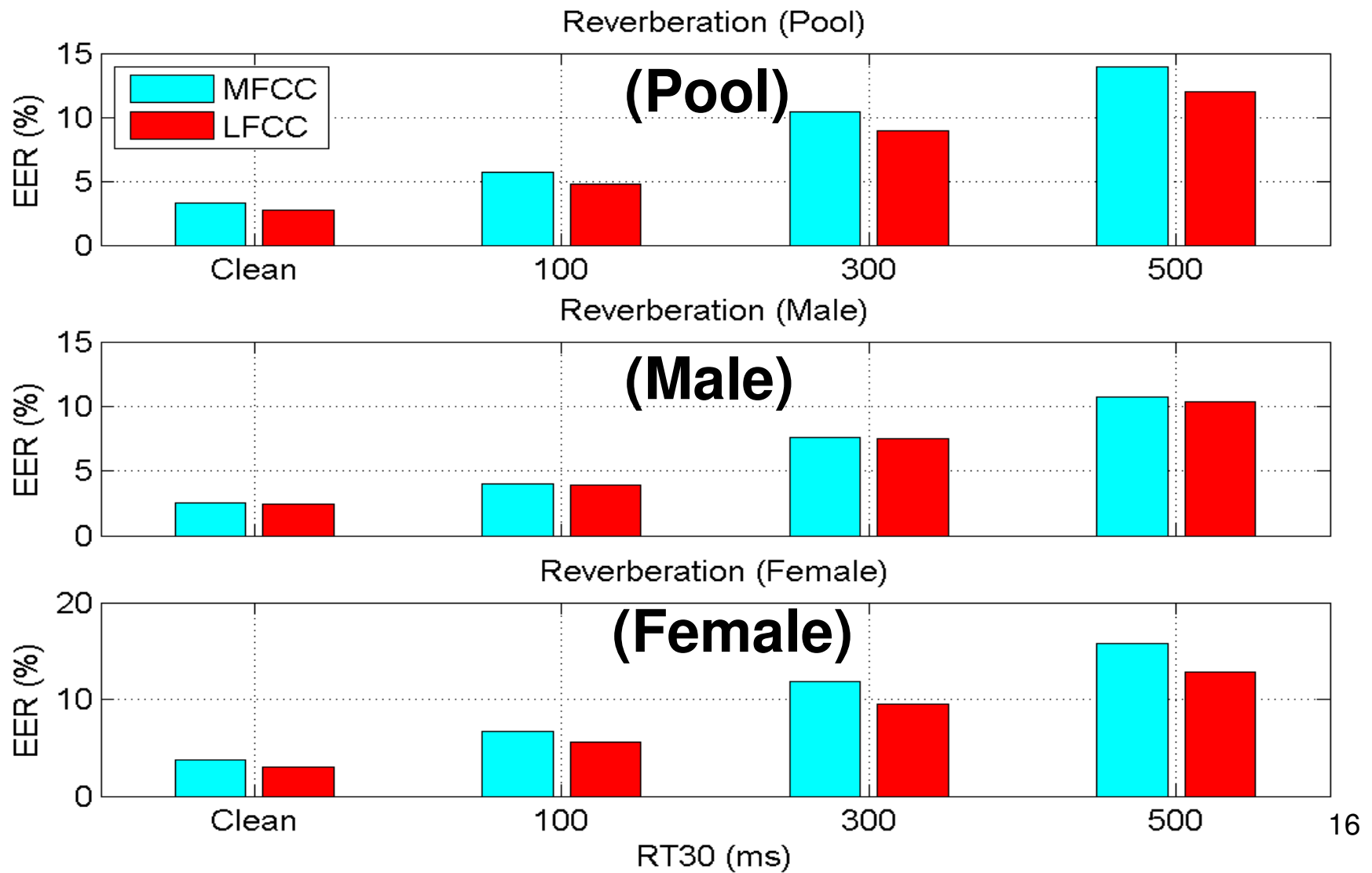
# MFCC vs. LFCC in additive noise

(C5 in PLDA system)



# MFCC vs. LFCC in reverberation

(C5 in PLDA system)



# Summary

- LFCC consistently outperforms MFCC in the female trials.
- There is some advantage of LFCC over MFCC in reverberant speech. LFCC is as robust as MFCC in the babble noise, but not in the white noise.
- Our results suggest that LFCC should be more often used, at least for the female trials, by the mainstream of the speaker-recognition community.

# References

1. K. N. Stevens, Acoustic phonetics. Cambridge, Mass.: MIT Press, 1998.
2. Story, B.H., “Using imaging and modeling techniques to understand the relation between vocal tract shape and acoustic characteristics”, Proceedings of the Stockholm Music Acoustics Conference, 6-9 August, 2003.
3. 2010 NIST Speaker Recognition Evaluation, [http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST\\_SRE10\\_evalplan.r6.pdf](http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf)
4. D. Ellis (2005), PLP and RASTA (and MFCC, and inversion) in Matlab, available online: <http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/>.
5. D. Garcia-Romero and C. Espy-Wilson, “Joint factor analysis for speaker recognition reinterpreted as signal coding using overcomplete dictionaries,” Proceedings of Odyssey Speaker and Language Recognition Workshop, June 2010.
6. D. Garcia-Romero and C. Espy-Wilson, “Analysis of I-vector Length Normalization in Speaker Recognition Systems”, INTERSPEECH 2011, Florence, Italy, pp.249-252.

# Acknowledgement

This work was supported by an IARPA grant and a NSF grant



