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.
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
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**

- MFCC: EER 3.63% newDCF: 0.52
- LFCC: EER 3.10% newDCF: 0.47
- MFCC: EER 1.77% newDCF: 0.35
- LFCC: EER 1.71% newDCF: 0.32
- MFCC: (Female) EER 4.89% newDCF: 0.65
- LFCC: (Female) EER 4.13% newDCF: 0.59

**PLDA**

- MFCC: (Pool) EER 2.89% newDCF: 0.49
- LFCC: (Pool) EER 2.79% newDCF: 0.41
- MFCC: (Male) EER 1.37% newDCF: 0.27
- LFCC: (Male) EER 1.34% newDCF: 0.26
- MFCC: (Female) EER 3.87% newDCF: 0.64
- LFCC: (Female) EER 3.89% newDCF: 0.53
DET Curves in NIST SRE10

(C5: Phonecall-Phonecall)

JFA

PLDA
DET Curves in NIST SRE10

(C6: Phonecall-Phonecall in high vocal effort)

JFA

PLDA

False Alarm probability (in %)

0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00

0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

MFCC: (Pool): EER 7.16% newDCF: 0.85
LFCC: (Pool): EER 4.98% newDCF: 0.70
MFCC: (Male): EER 5.71% newDCF: 0.79
LFCC: (Male): EER 4.53% newDCF: 0.70
MFCC: (Female): EER 8.26% newDCF: 0.89
LFCC: (Female): EER 5.30% newDCF: 0.68

MFCC: (Pool): EER 5.95% newDCF: 0.85
LFCC: (Pool): EER 4.97% newDCF: 0.73
MFCC: (Male): EER 4.95% newDCF: 0.78
LFCC: (Male): EER 4.59% newDCF: 0.74
MFCC: (Female): EER 6.38% newDCF: 0.87
LFCC: (Female): EER 6.18% newDCF: 0.72
Equal Error Rates (all trials)

**JFA (pool)**

**PLDA (pool)**
Equal Error Rates (male trials)

JFA (male)

PLDA (male)
Equal Error Rates (female trials)

JFA (female)

PLDA (female)
MFCC vs. LFCC in additive noise
(C5 in PLDA system)

White (Pool)

(White noise)

Babble (Pool)

(Babble noise)
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


Acknowledgement

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