Delta-Spectral Cepstral Coefficients for Robust Speaker Recognition

MERIT FAIR
BIEN 2011

Jonathan Deutsche, Xinhui Zhou, Carol Espy-Wilson

The Institute for Systems Research

Introduction

•**Speaker Recognition** - recognizing a person from his/her unique voice characteristics (physical differences and manner of speaking)

- •Physical differences:
  - •Vocal tract shape and size
  - •Larynx shape and size
- •Manners of speaking:
  - •Accent and Dialect
  - •Rhythm
  - •Word Usage

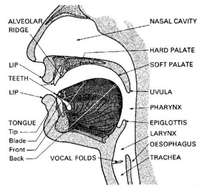

Figure 1: Mid-Sagittal View of Human Vocal Tract [1]

•Uses: forensics, security, telephone services, information searching

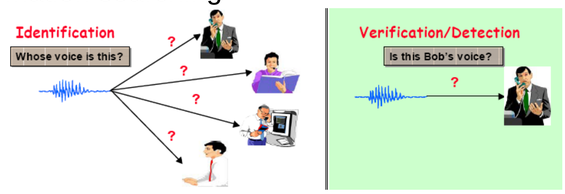

Figure 2: Speaker Identification vs. Speaker Verification [1]

Motivation

•Most current speaker recognition systems use mel-frequency cepstral coefficients in conjunction with delta/double-delta cepstral coefficients (**MFCC + DCC's**) as front-ends – these features are not robust to noise, reverberation, and channel effects [2]

•Recently, it has been shown that delta-spectral cepstral coefficients (**DSCC's**) are more robust than DCC's for speech recognition [3]

Objective

*Figure 4 illustrates process behind 1-4

•Test whether **MFCC + DSCC's** are more robust than **MFCC + DCC's** for speaker recognition

Clean Speech **Speech in 0 dB White Noise**

1. **Spectrograms of Speech Signal**
2. **Short-Time Power Plot of Mel-Filtered Speech Signal**
3. * w/ Delta Operation **Temporal Difference Operation over the Logarithmic Power of Speech Signal**
4. **Gaussianization Operation over the Temporal Difference of Speech Signal**

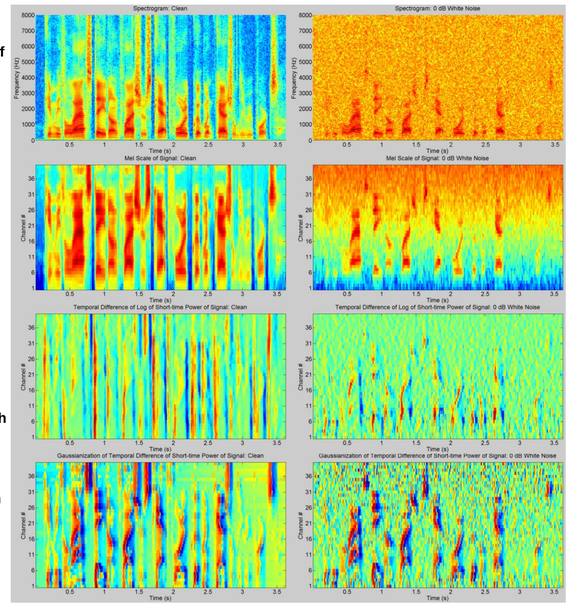

Figure 3: Spectrograms and Short-Time Power Plots of Speech Signal

Methods

*Figure 3 illustrates features from 1-4

•**Front-End System**

- •**Feature extraction** – speech signal is transformed into feature vectors in which speaker-specific properties are emphasized

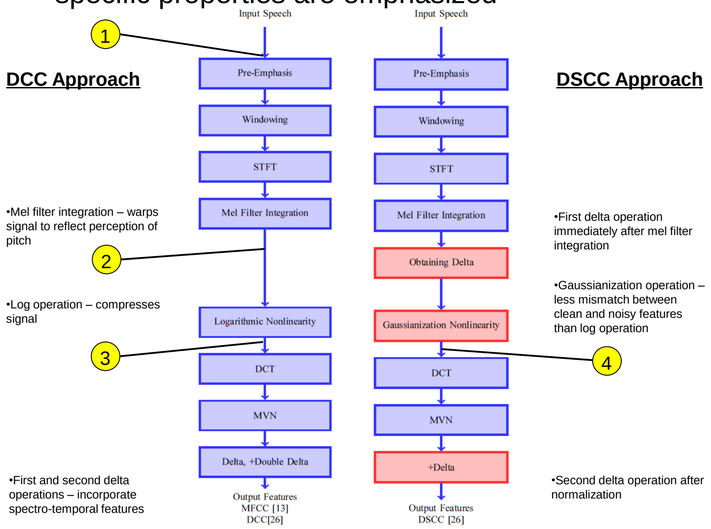

Figure 4: Feature extraction process: DCC vs. DSCC (with Gaussianization) [3]

•**Back-End System**

- •Training and testing data from **NIST 2008 SRE Plan**: 8 different conditions [4]
- •Large set of background speakers is used to train **universal background model** (UBM) [2]
- •Data from specific speaker and UBM is used to train **target/speaker model** [2]
- •Models are trained using **Gaussian mixture models** (GMM's) [2]
- •Test data is compared to UBM and target model, categorized as "speaker" or "not the speaker" [2]

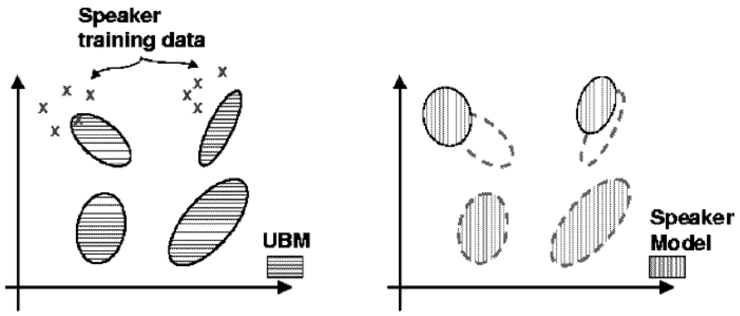

Figure 5: GMM Target Model Training Using MAP Adaptation [2]

•**Performance Evaluation**

- •**Detection error tradeoff (DET) curve** – the probability of false acceptance vs. the probability of false alarm [2]
- •**Equal error rate (EER)** – accuracy at decision threshold for which probability of false acceptance and false alarm are equal [2]

Results

Condition 2: Interview Speech – Same Mic **Condition 3: Interview Speech – Different Mic**

With White Noise **With Babble Noise** **With Reverberation**

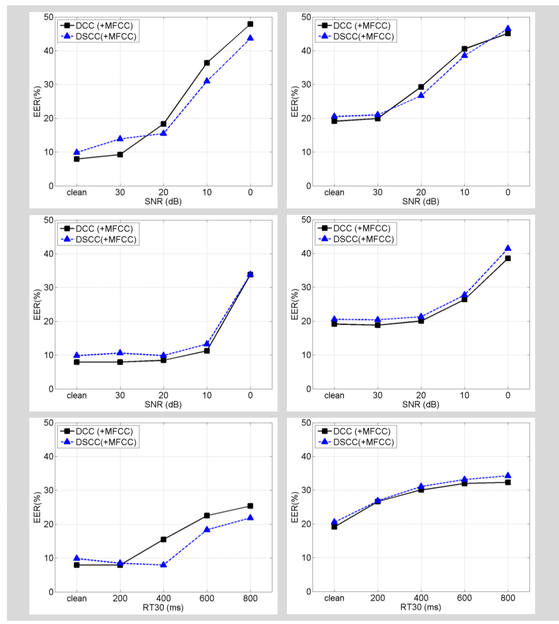

Figure 6: Equal Error Rates for Different Evaluation Conditions and Test Noise Types

Conclusions

•In general, MFCC+DSCC's are more robust to white noise and reverberation than MFCC+DCC's when training and test data are recorded on same channel type

•When there is a channel mismatch between training and test data, MFCC+DSCC's show no improvement over MFCC+DCC's in all conditions

•MFCC+DSCC's show no improvement over MFCC+DCC's in babble noise

•DCC's (with logarithmic nonlinearity) may be more robust to channel mismatch than DSCC's (with Gaussianization nonlinearity)

Acknowledgments

•National Science Foundation OCI award #1063035
•Daniel Garcia-Romero
•Tarun Pruthi

References

[1] Garcia-Romero, Daniel. "Speaker Recognition Using Gaussian Mixture Models (GMM's)." Powerpoint Presentation. 2006.
[2] Kinnunen, Tomi, and Haizhou Li. "An Overview of Text-independent Speaker Recognition: From Features to Supervectors." *Speech Communication* 52.1 (2010): 12-40.
[3] Kumar, Kshitiz, Chanwoo Kim, and Richard M. Stern. "Delta-Spectral Cepstral Coefficients for Robust Speech Recognition." *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2011): 4784-787.
[4] "The NIST Year 2008 Speaker Recognition Evaluation Plan." http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf. (2008).