Dimension Reduction of the Modulation Spectrogram for Speaker Verification

Tomi Kinnunen
Speech and Image Processing Unit
Department of Computer Science
University of Joensuu, Finland

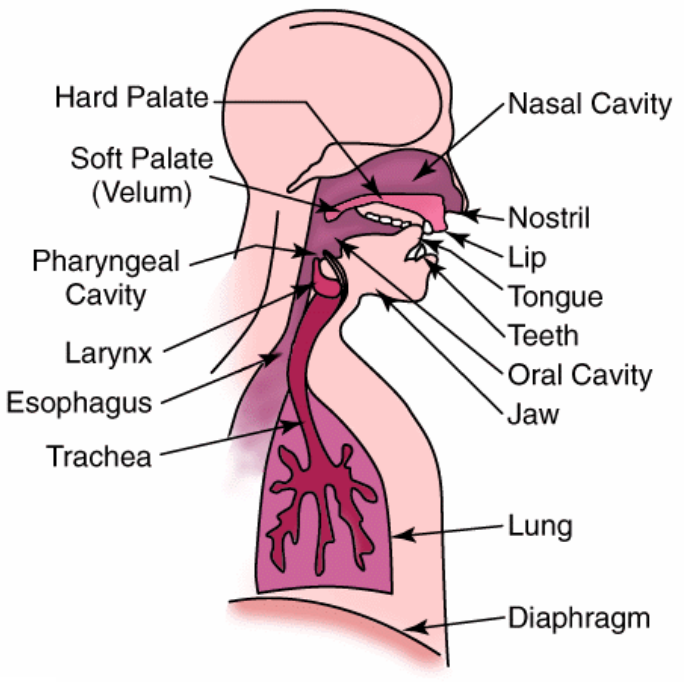
Kong Aik Lee and **Haizhou Li**
Human Language Technology Department
Speech and Dialogue Processing Lab
Institute for Infocomm Research (I^{2}R), Singapore

Odyssey 2008 - The Speaker and Language Recognition Workshop, Stellenbosch, South Africa, 21.-24.1.2008)

Speaker Recognition

Recognizing persons from their voices

Physiology and anatomy
("Speech hardware")



Manner of speaking
("Speech software")

Speaker recognition systems

	Physical features (physiology)	Stylistic features (manner of speaking)
Front-end (Feature extractor)	Short-term spectrum (MFCC, LPCC)	Tokenizer (HMM …nizer, prosodic …tor)
Back-end (Classifier)	Gaussia… model (G… vector machi… neural nets	

Is it possible to extract stylistic features 'directly' from the signal, without a complex tokenizer ?

+ Computationally efficient

+ Simple implementation

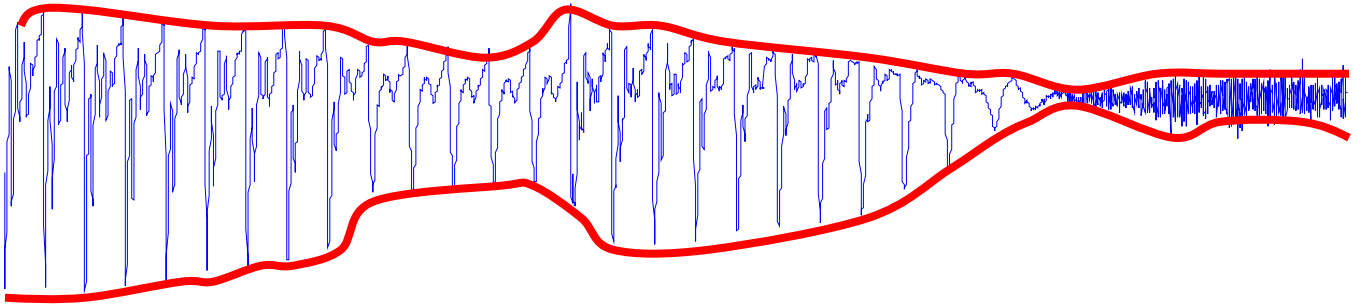
- Computationally expensive

- Complex front-end

- Speaking style assumed to be discrete and categorical

Speech: a low bandwidth process which modulates higher bandwidth carriers

- Lips, jaw and tongue movements are low-frequency processes that modulate the glottal airflow
  - Energy oscillations at syllabic rates
  - Formant transitions
- Syllable rate of continuous speech ~4 Hz



H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, no. 1-3, pp. 3–27, Aug. 1998.

S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, 2001, pp. 473–476.

L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, pp. 117–132, 1998.

Modulation spectrum in speech technology

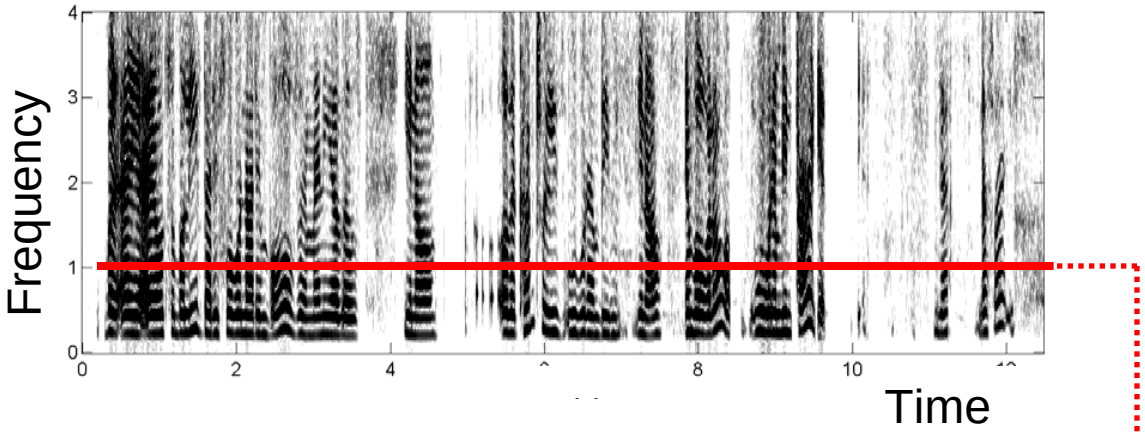
- RASTA filtering
  [Hermansky, *IEEE T Speech & Audio Proc*, 1994]
- Improving speech recognition by modulation filtering
  [Kingsbury, Morgan & Greenberg, *Speech Communication*, 1998]
- Speaker separation from a single-channel audio
  [Schimmel, Atlas & Nie, *ICASSP* 2007]
- Age and gender classification [Ajmera & Burkhardt, *Odyssey 2008*]
- Many others: speech enhancement, voice activity detection, audio compression

In **speaker recognition** :

• Filtering in the modulation domain to improve conventional cepstral systems [v. Vuuren & Hermansky, *ICSLP* 1998], RASTA filtering

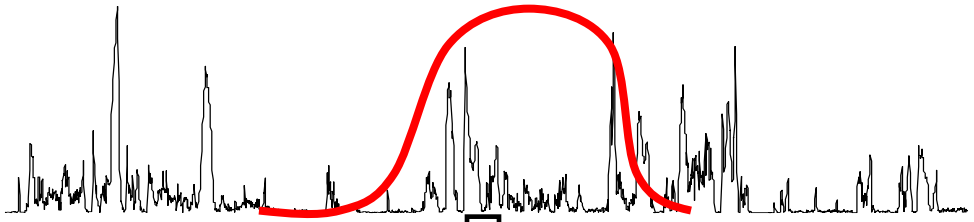
Our proposal: using joint acoustic and modulation spectrum, or *modulation spectrogram*, as a **feature** [*ICASSP* 2006]

Modulation spectrum



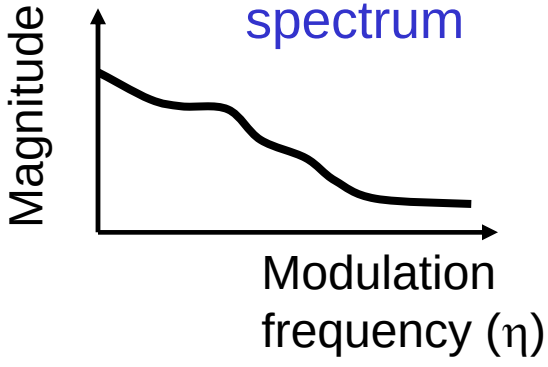
FFT spectrogram

Temporal trajectory of one subband

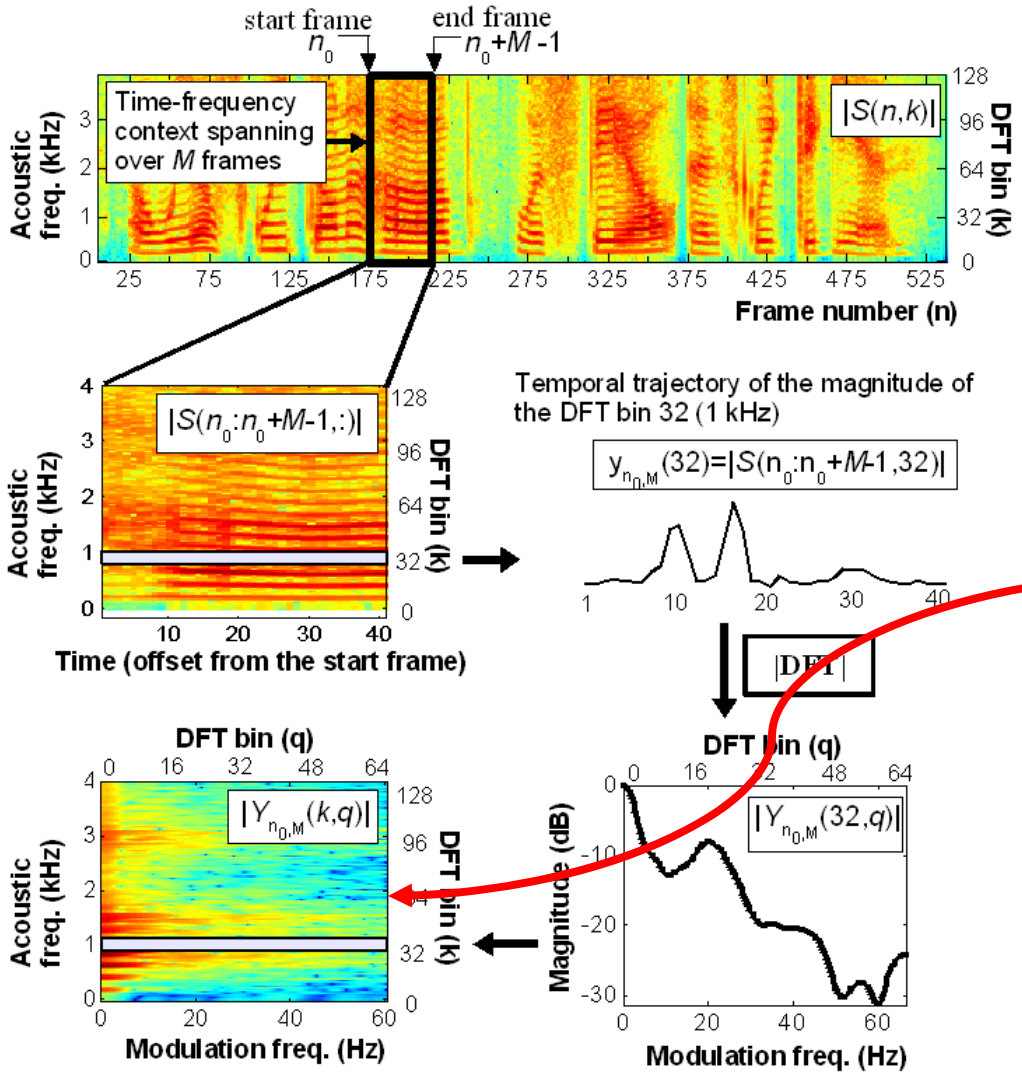


Another short-term FFT

Modulation spectrum



What is modulation spectrogram ?



Spectrogram:

short-term (~30ms) distribution of the energy across different "acoustic" frequencies

A practical problem: high dimensionality! (10^3~10^4)

Modulation spectrogram:

Longer-term (200~300 ms) joint distribution of the energy across different "acoustic" and "modulation" frequencies

Dimensionality reduction

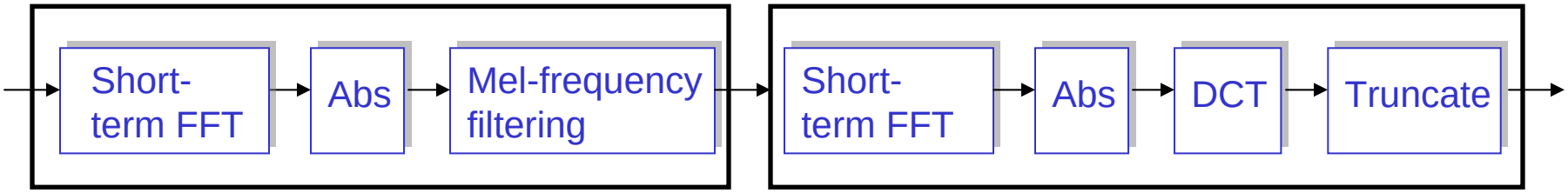
1. “Acoustic” frequency dimension:
   - A bank of triangular shaped mel-frequency filters as usual

2. “Modulation” frequency dimension:
   - Heavy damping of frequencies above 20 Hz
   - Smooth shape, no harmonic structure

   ==> Apply discrete cosine transform (DCT)

   to approximate the envelope

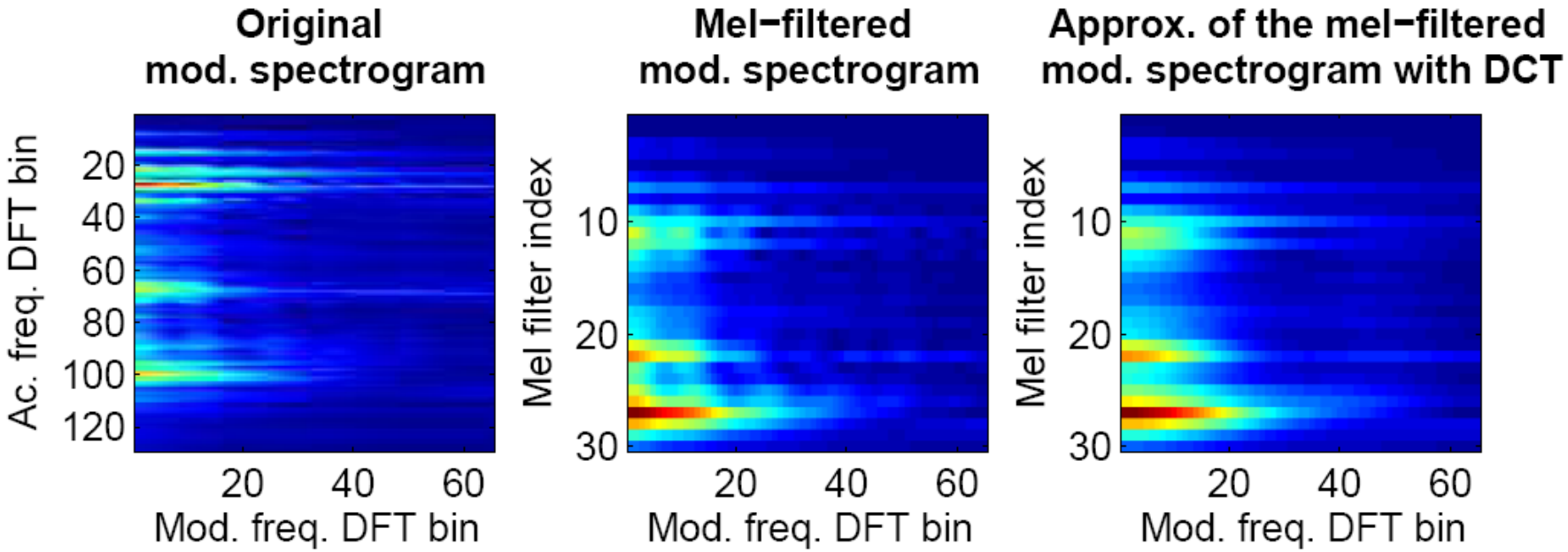
Summary of the steps



Mel spectrogram computation

Modulation spectral analysis

Dimensionality reduction



Dimensionality
129 x 65 = 8385

Dimensionality
30 x 65 = 1950

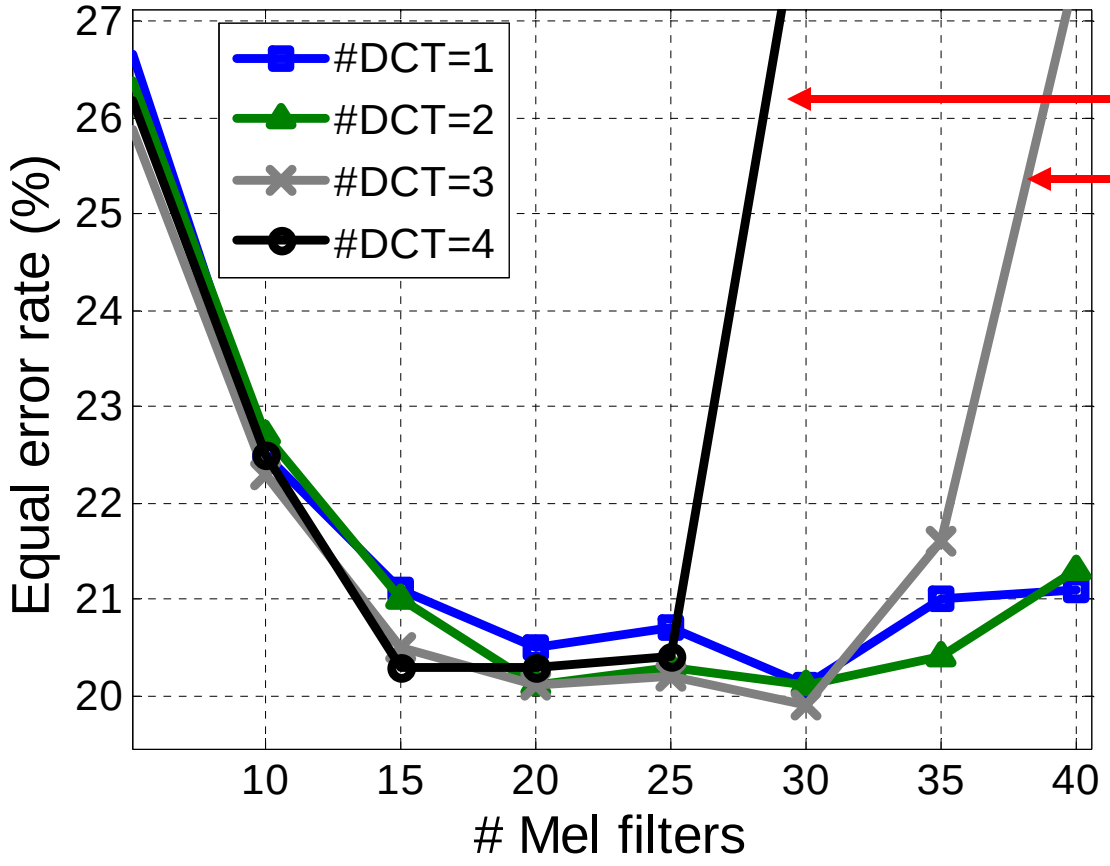
Dimensionality
30 x 4 = 120

Experiments

- NIST 2001 speaker recognition evaluation (SRE) corpus
  - 174 target speakers
  - 22,418 verification trials (90% impostors, 10% genuine)
  - Training data: 2 minutes / speaker
  - Test data: 0~60 sec
- Gaussian mixture model - universal background model (GMM-UBM) recognizer
- Background model trained from the development set of the NIST 2001 corpus

How many mel filters and DCT coefficients?

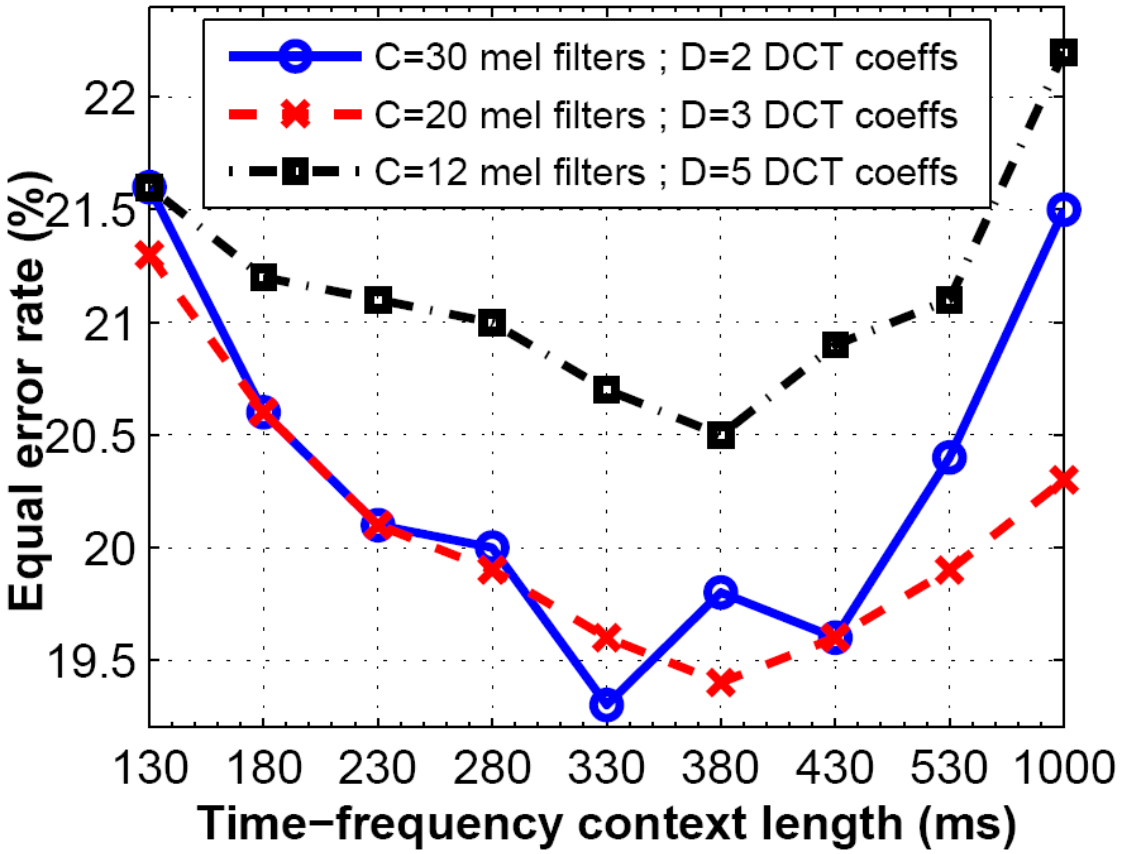
NIST 2001 corpus, GMM-UBM recognizer
Context length = 27 frames = 225 milliseconds



Numerical problems due to high dimensionality

Context length

Dimensionality fixed to
30x2 = 20x3 = 12x5 = 60



Better time resolution, stationarity

Better mod. spectrum resolution

Comparison with our previous result

[ICASSP 2006]: EER = 25.1 %

Classifier: Long-term averaging classifier with Kullback-Leibler distance + T-norm score normalization

Dimensionality = 3200

[This study] EER = 17.4 %

Classifier: GMM-UBM (256 Gaussians), no score normalization

Dimensionality = 60

Comparison with MFCCs

Test duration (s)	MFCC	Mod.spec.	Fusion
0–20	10.5	18.6	10.5
20–30	8.5	17.6	8.4
30–40	7.6	16.6	7.3
40–60	7.7	15.8	7.3

Expected result: better fusion for longer samples

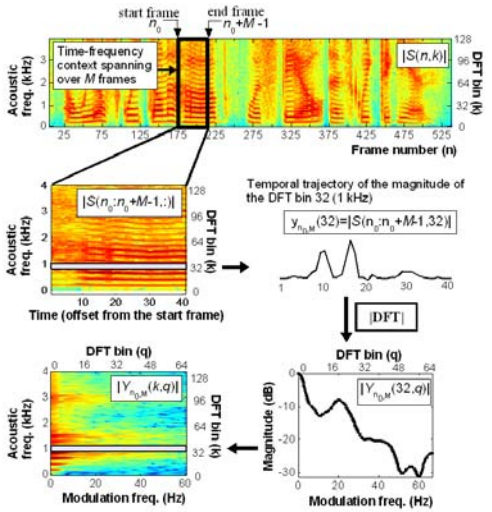
Fusion: linear score fusion with the weights optimized using logistic regression (FoCal toolkit)

… but the improvement is relatively modest

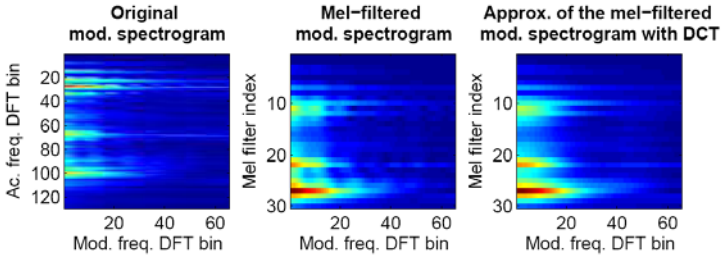
- Would the benefit be better for significantly longer training and test data ?
- Fusion too simplistic ?
- Phase differences of the subbands should be retained as well ?

Summary

Modulation spectrogram as a feature for speaker recognition



Added mel filtering and DCT to reduce dimensionality



Demonstrated accuracy improvement on NIST 2001 compared to our previous result

EER = 25.1 % ==> EER = 17.4 %

Fusion gain with MFCCs was minor, cannot be recommended for applications yet

… but we will not give up yet :-)