Harmonic Structure Transform for Speaker Recognition

Kornel Laskowski & Qin Jin

Carnegie Mellon University, Pittsburgh PA, USA
KTH Speech Music & Hearing, Stockholm, Sweden

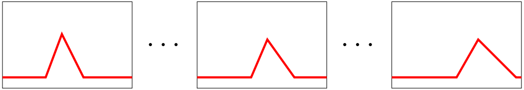
29 August, 2011

Spectral Transforms in General

Given $\mathbf{x} \equiv$ the energy spectrum of a speech frame,

$$\mathbf{y} \;=\; \mathcal{F}^{-1}\left(\log\left(\mathbf{M}^T\mathbf{x}\right)\right) - \langle\text{normalization term}\rangle$$

The matrix $\mathbf{M}$ is a filterbank, whose columns look like:



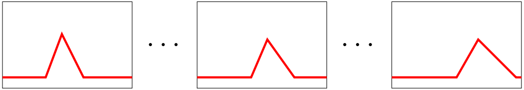
M defines the **number** of filters, and their **central frequencies**, **widths**, and **general shapes**.

Spectral Transforms in General

Given $\mathbf{x} \equiv$ the energy spectrum of a speech frame,

$$\mathbf{y} = \mathcal{F}^{-1}\left(\log\left(\mathbf{M}^T\mathbf{x}\right)\right) - \langle\text{normalization term}\rangle$$

The matrix $\mathbf{M}$ is a filterbank, whose columns look like:

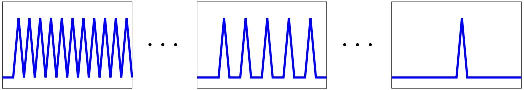


M defines the **number** of filters, and their **central frequencies**, **widths**, and **general shapes**.

Importantly here, the filters of all such filterbanks integrate energy across frequencies **related by adjacency**.

The Harmonic Structure Transform (HST)

In contrast, the HST is implemented by a matrix **H** whose columns look like:



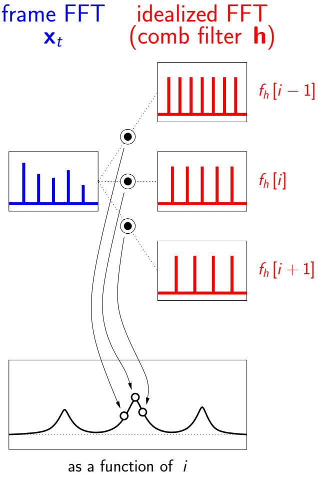
Each filter integrates energy across frequencies **related by harmonicity** (not adjacency).

- this is **novel** (Laskowski & Jin, 2010) for speaker recognition
- related to (Liénard, Barras & Signol, 2008) for pitch detection
- unknown: **number** of filters, and their **fundamental frequencies**, **tooth widths**, and **individual tooth shapes**

Outline of this Talk

1. Baseline Performance
   - What **is** known?
2. Experiments in HSCC Filterbank Design
   - linear spacing in fundamental frequency
   - piecewise linear spacing in fundamental frequency
   - logarithmic spacing in fundamental frequency
   - fundamental frequency range and density
3. Score-level Fusion with Standard MFCCs
4. Generalization
5. Conclusions

HST Processing



- analysis every 8 ms
- frames 32 ms wide
- comb filter teeth triangular (global width parameter)
- 400 filters, linearly spanning from 50 Hz to 450 Hz
- logarithm at each filter output, then normalization
- decorrelation using LDA
- yields *harmonic structure cepstral coefficients (HSCCs)*

HSCC Modeling for Classification

As simple as possible.

- one GMM per speaker
    1. assume one Gaussian element
    2. determine optimal number N_D of LDA dimensions
    3. hold N_D fixed
    4. determine optimal number of N_G Gaussians
- maximum likelihood closed-set classification (MAP under uniform prior)

Available Results (Laskowski & Jin, ODYSSEY 2010)

- Wall Street Journal data, mostly read speech
- 100-way closed-set classification, per gender
- ≈1500 10-second trials, per gender and dataset
- matched channel and matched multi-session conditions

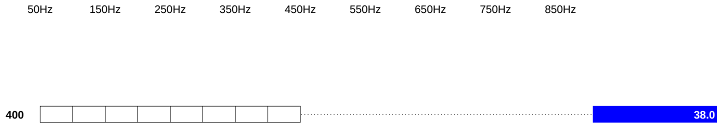
System	Female, ♀		Male, ♂	
	DEV	TEST	DEV	TEST
$F0$	17.6	18.4	26.2	27.4
HST/LDA	99.7	99.9	99.7	99.7
MEL/DCT	98.7	99.3	99.3	98.6
MEL/LDA	98.7	99.3	99.3	98.9

Session Mismatch

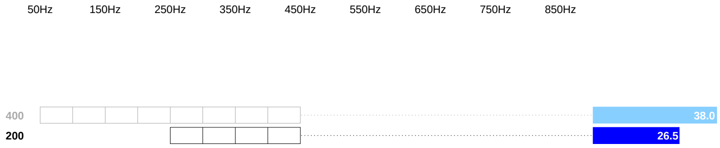
- MIXER5 data, various speaking styles
- 66-way closed-set classification
- ≈3000 10-second trials, per dataset
- matched channel and matched session: accuracies of 100%
- matched channel but **mismatched session**:

System	DEV	TEST
$F0$	14.1	16.2
HST/LDA	**59.8**	**68.1**
MEL/DCT	74.4	84.4
MEL/LDA	81.5	87.8

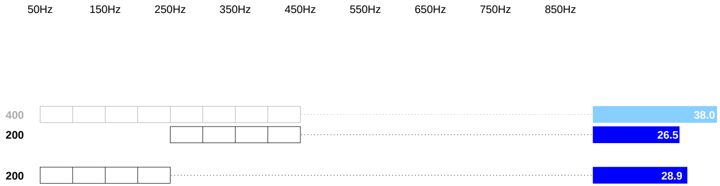
Linear Spacing of Fundamental Frequencies



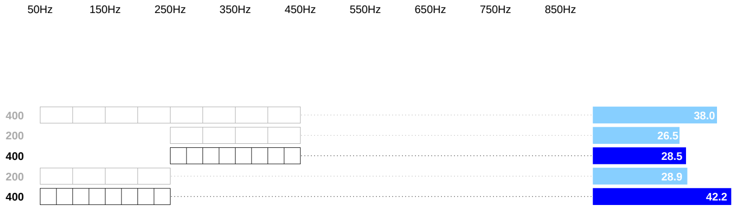
Linear Spacing of Fundamental Frequencies



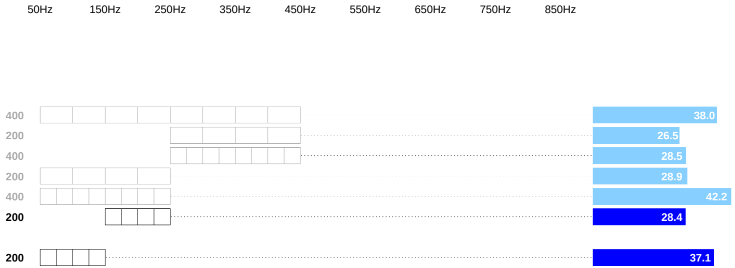
Linear Spacing of Fundamental Frequencies



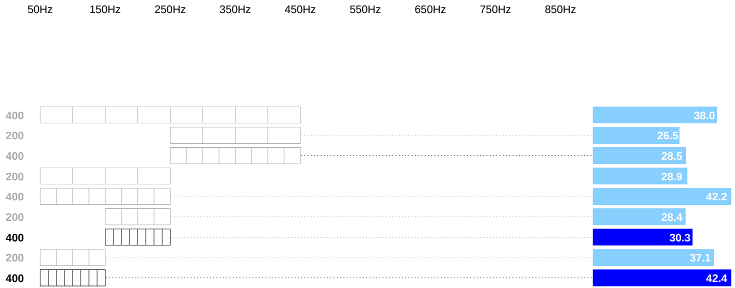
Linear Spacing of Fundamental Frequencies



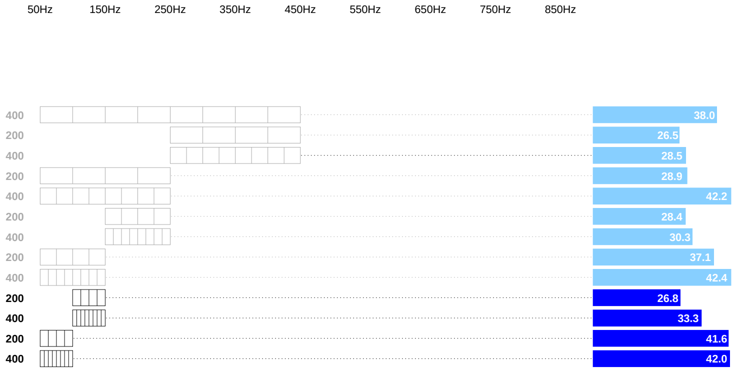
Linear Spacing of Fundamental Frequencies



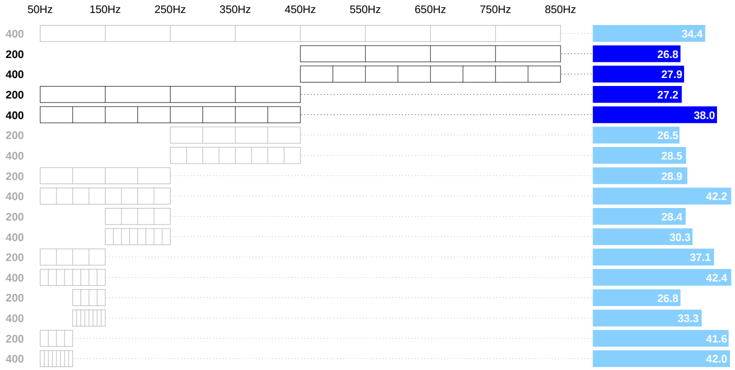
Linear Spacing of Fundamental Frequencies



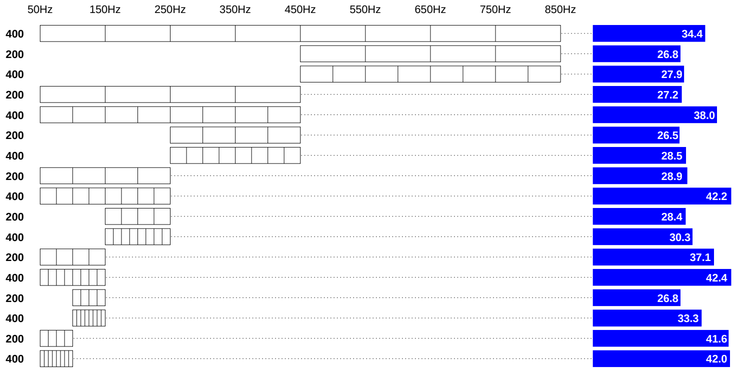
Linear Spacing of Fundamental Frequencies



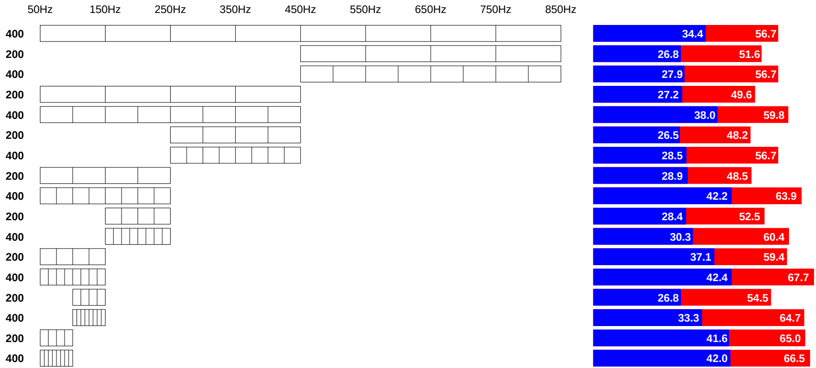
Linear Spacing of Fundamental Frequencies



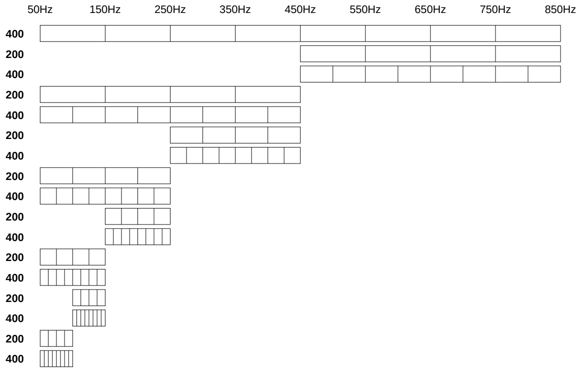
Linear Spacing of Fundamental Frequencies



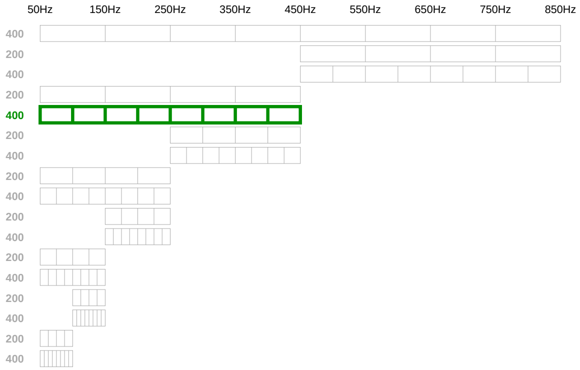
Linear Spacing of Fundamental Frequencies



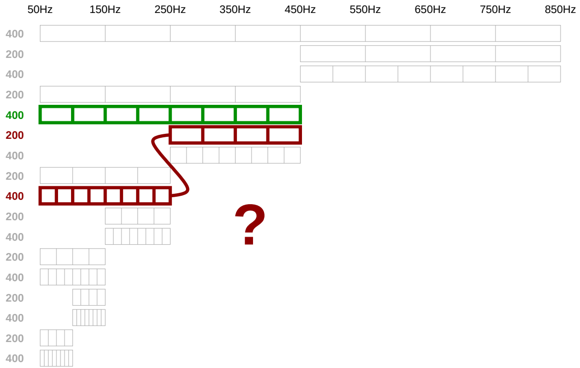
Linear Spacing of Fundamental Frequencies



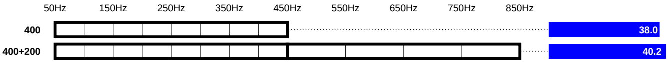
Linear Spacing of Fundamental Frequencies



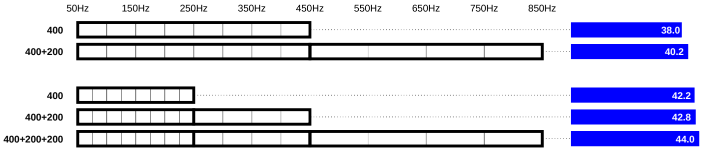
Linear Spacing of Fundamental Frequencies



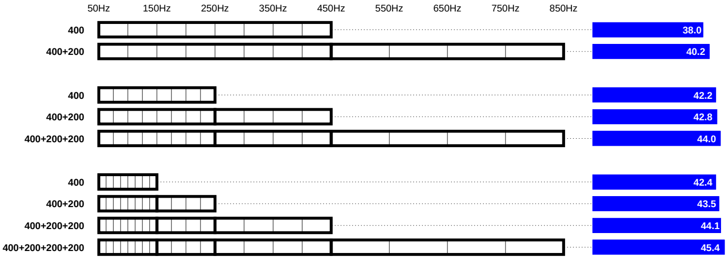
Piecewise Linear Spacing of Fundamental Frequencies



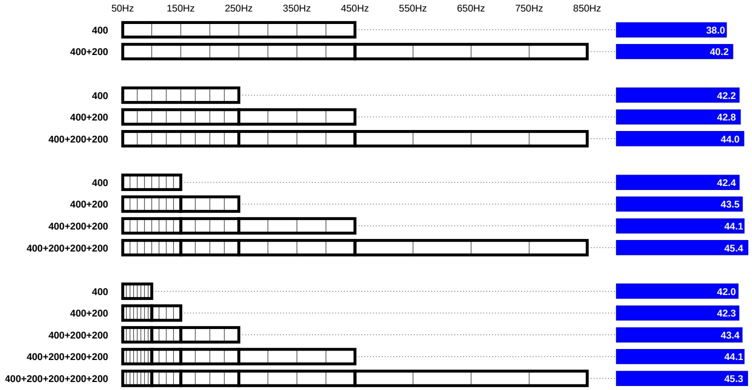
Piecewise Linear Spacing of Fundamental Frequencies



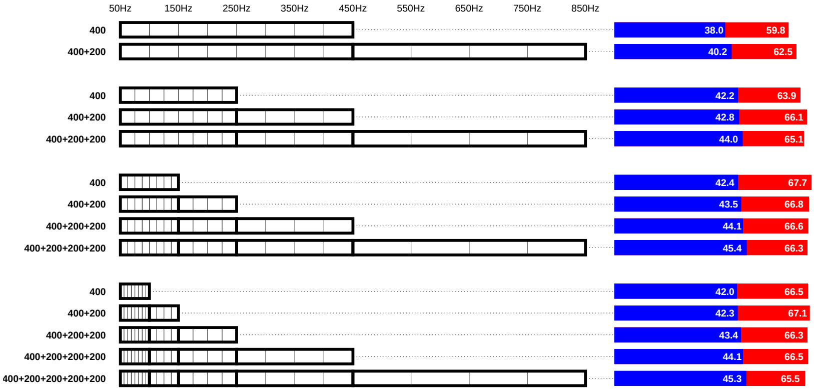
Piecewise Linear Spacing of Fundamental Frequencies



Piecewise Linear Spacing of Fundamental Frequencies



Piecewise Linear Spacing of Fundamental Frequencies

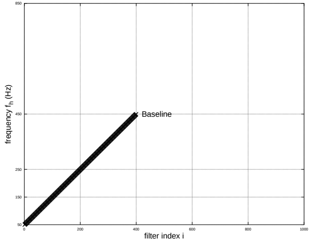


Logarithmic Spacing of Fundamental Frequencies

In the linear case:

- N_h fundamental frequencies f_h between f_{min} and f_{max}:

$$f_h[i] \quad = \quad f_h^{min} + \frac{i-1}{N_h-1}\left(f_h^{max} - f_h^{min}\right) \qquad 1 \leq i \leq N_h$$

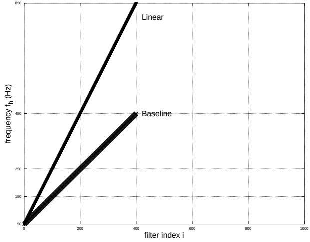


Logarithmic Spacing of Fundamental Frequencies

In the linear case:

- N_h fundamental frequencies f_h between f_{min} and f_{max}:

$$f_h[i] \;=\; f_h^{min} + \frac{i-1}{N_h - 1}\left(f_h^{max} - f_h^{min}\right) \qquad 1 \leq i \leq N_h$$

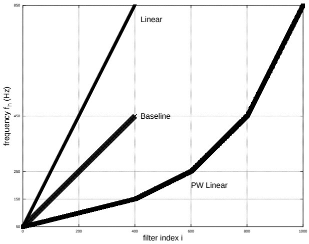


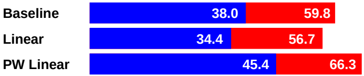
Logarithmic Spacing of Fundamental Frequencies

In the linear case:

- N_h fundamental frequencies f_h between f_{min} and f_{max}:

$$f_h\left[i\right] \quad = \quad f_h^{min} + \frac{i-1}{N_h - 1}\left(f_h^{max} - f_h^{min}\right) \qquad 1 \leq i \leq N_h$$



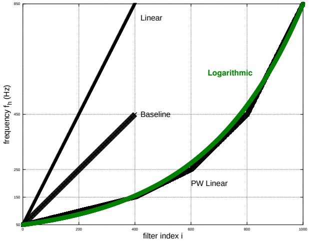


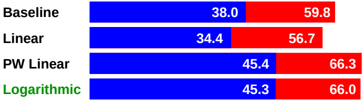
Logarithmic Spacing of Fundamental Frequencies

In the linear case:

- N_h fundamental frequencies f_h between f_{min} and f_{max}:

$$f_h[i] = f_h^{min} + \frac{i-1}{N_h-1}\left(f_h^{max} - f_h^{min}\right) \quad 1 \le i \le N_h$$



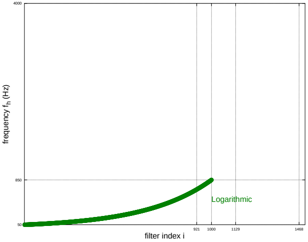


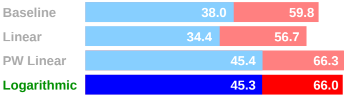
$$f_h[i] = f_h^{min}\left(\frac{f_h^{max}}{f_h^{min}}\right)^{\frac{i-1}{N_h-1}}$$

Range and Density of Fundamental Frequencies

Three manipulations:

- f_{min}: raise to 62.5 Hz
- f_{max}: raise to maximize accuracy (to 4000 Hz)
- N_h: lower to maximize accuracy (to 1129)

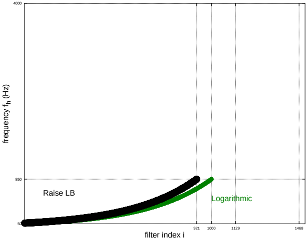


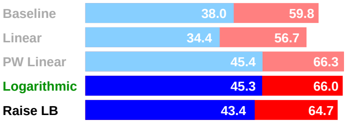


Range and Density of Fundamental Frequencies

Three manipulations:

- f_{min}: raise to 62.5 Hz
- f_{max}: raise to maximize accuracy (to 4000 Hz)
- N_h: lower to maximize accuracy (to 1129)

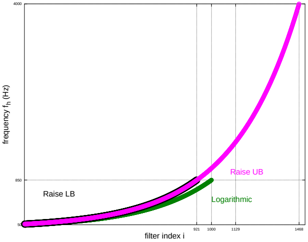


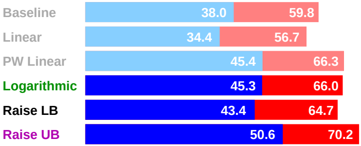


Range and Density of Fundamental Frequencies

Three manipulations:

- f_{min}: raise to 62.5 Hz
- f_{max}: raise to maximize accuracy (to 4000 Hz)
- N_h: lower to maximize accuracy (to 1129)

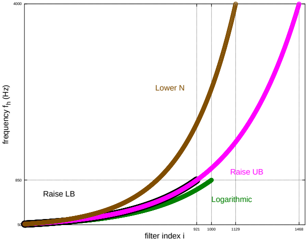


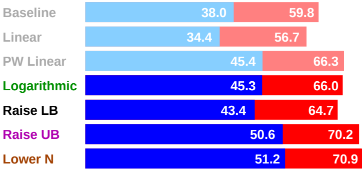


Range and Density of Fundamental Frequencies

Three manipulations:

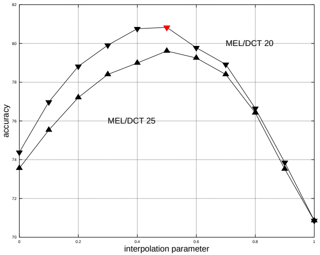
- f_{min}: raise to 62.5 Hz
- f_{max}: raise to maximize accuracy (to 4000 Hz)
- N_h: lower to maximize accuracy (to 1129)





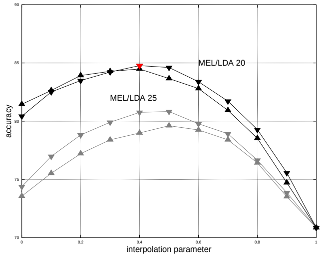
Interpolation with Standard MFCCs

- linear interpolation with DCT-decorrelated log-Mel energies
  - 20 coefficients
  - 25 coefficients (always worse)

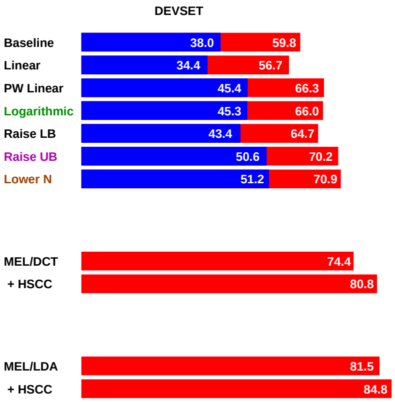


Interpolation with LDA-Rotated MFCCs

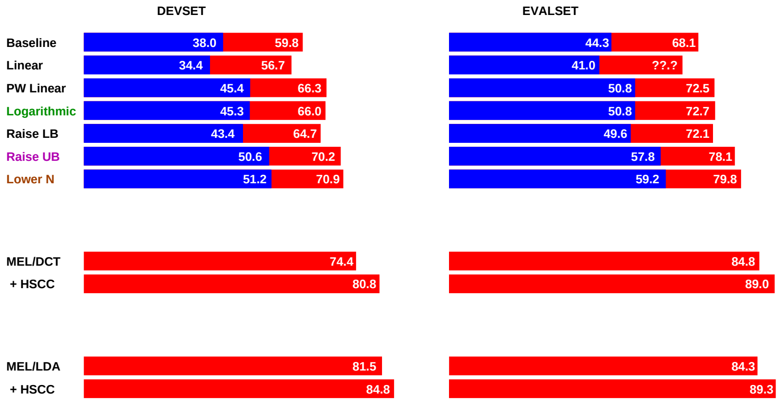
- linear interpolation with LDA-decorrelated log-Mel energies
  - 20 coefficients (better in combination)
  - 25 coefficients (better alone)



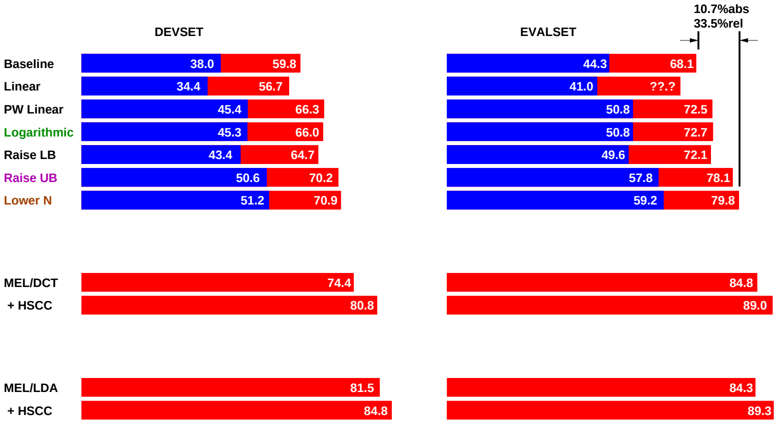
Summary of Accuracy on DevSet



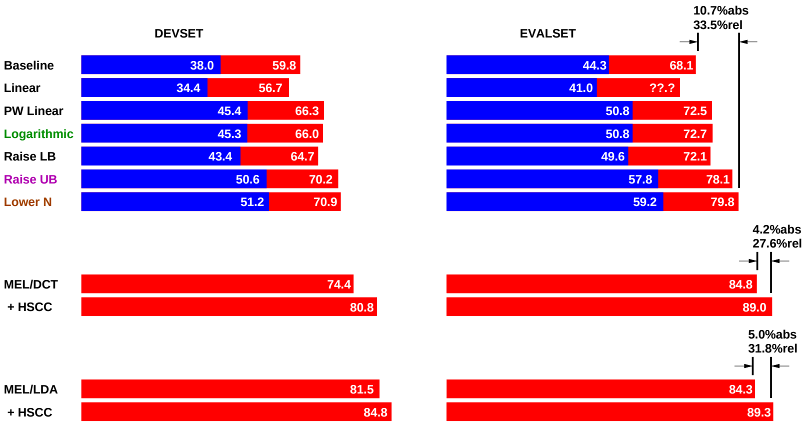
Accuracy on EvalSet



Accuracy on EVALSET



Accuracy on EVALSET



Conclusions

1. evaluated the baseline transform in **session mismatch**
   - viable: twice as many errors an equivalent MFCC system
2. errors can be reduced **by a third (33.5%rel)** by optimizing:
   - the number of filters in the filterbank
   - the fundamental frequency corresponding to each filter
3. **logarithmic spacing** of fundamental frequencies is better than linear spacing
   - more filters for low fundamental frequencies
   - fewer filters for high fundamental frequencies
4. in an equivalent MFCC system, errors can be reduced **by almost a third (27.6-31.8%rel)** via score-level fusion with the improved HST system

Future Directions

1. change **framing policy** from 8 ms/32 ms to something longer
   - intonation and voice quality are **supra**-segmental
   - larger temporal support $\longrightarrow$ greater spectral resolution
2. optimize **the tooth shape** of comb filters
3. find a **data-independent** decorrelation transform
   - leading to a compact ($<$ 25 coefficients) representation
4. explore adaptation from a **universal background model**
5. generalize to **binary speaker verification** (and NIST SREs)

Potential Impact

1. a new **general** representation of the spectrum
2. deliberately **orthogonal** to spectral envelope features (MFCCs, LPCCs, etc.)
   - but computed in an identical manner
3. likely beneficial not only for speaker recognition, but also:
   - online **speaker diarization**
   - classification of “emotional speech”
   - clinical voice quality assessment

Some Interesting Insights ...

1. currently in HST, spectral energy < 300 Hz is zeroed out
    - this improves closed-set speaker classification
    - but **the fundamental (zeroth harmonic) is ignored**
    - the fundamental is thought to play a role in **emotional expression**
2. that optimal filter spacing is logarithmic is curious
    - independent of the logarithmic tonotopicity of the basilar membrane
    - greater acuity in discriminating among harmonic sounds (not just pure tones)

THANK YOU