Modeling Prosody for Speaker Recognition: Why Estimating Pitch May Be a Red Herring

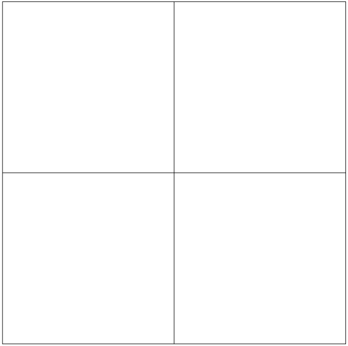
Kornel Laskowski & Qin Jin

Carnegie Mellon University
Pittsburgh PA, USA

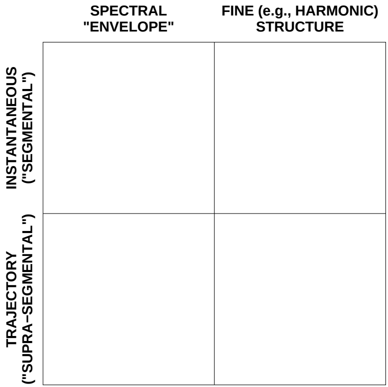
28 June, 2010

Features in Speech Processing

Features in Speech Processing



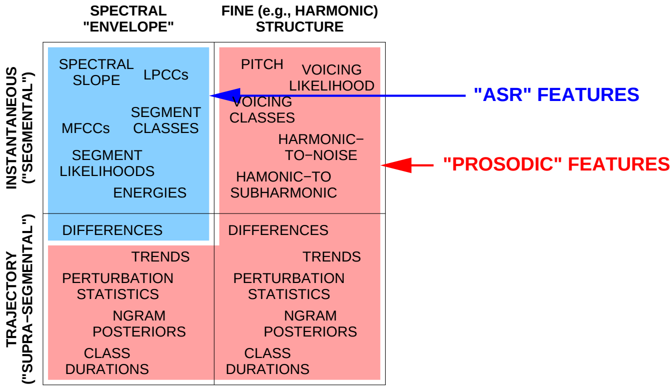
Features in Speech Processing



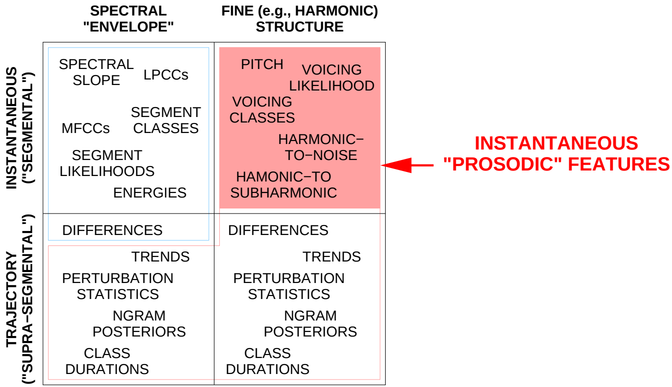
Features in Speech Processing

	SPECTRAL "ENVELOPE"	FINE (e.g., HARMONIC) STRUCTURE
INSTANTANEOUS ("SEGMENTAL")	SPECTRAL SLOPE LPCCs MFCCs SEGMENT CLASSES SEGMENT LIKELIHOODS ENERGIES	PITCH VOICING LIKELIHOOD VOICING CLASSES HARMONIC−TO−NOISE HAMONIC−TO SUBHARMONIC
TRAJECTORY ("SUPRA−SEGMENTAL")	DIFFERENCES TRENDS PERTURBATION STATISTICS NGRAM POSTERIORS CLASS DURATIONS	DIFFERENCES TRENDS PERTURBATION STATISTICS NGRAM POSTERIORS CLASS DURATIONS

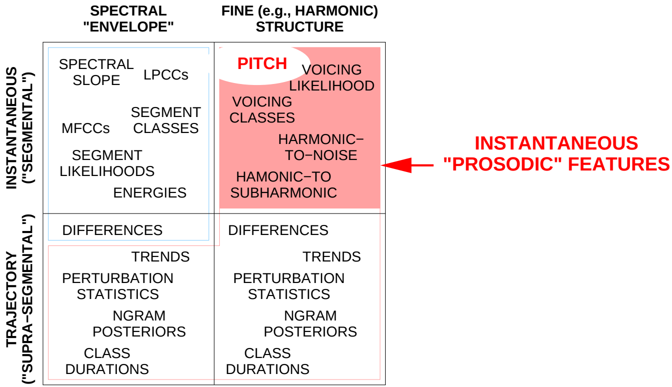
Features in Speech Processing



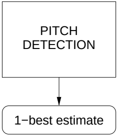
Features in Speech Processing



Features in Speech Processing

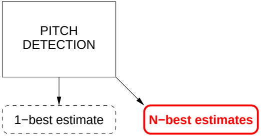


Pitch Estimation & Processing



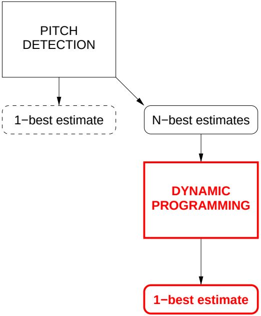
- this talk explores what happens **inside here**
- **low-level feature computation**

Pitch Estimation & Processing



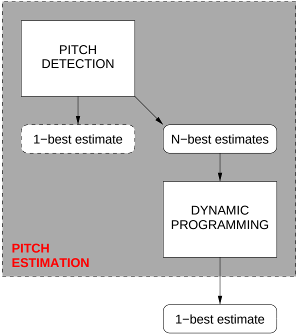
- this talk explores what happens **inside here**
- **low-level feature computation**

Pitch Estimation & Processing



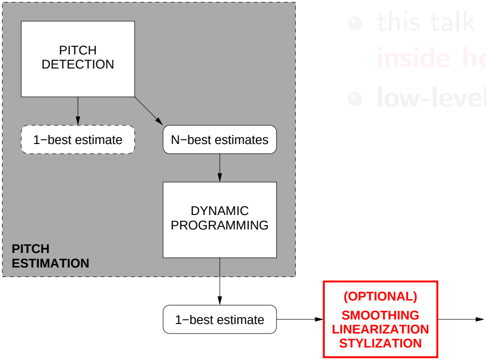
- this talk explores what happens **inside here**
- **low-level feature computation**

Pitch Estimation & Processing



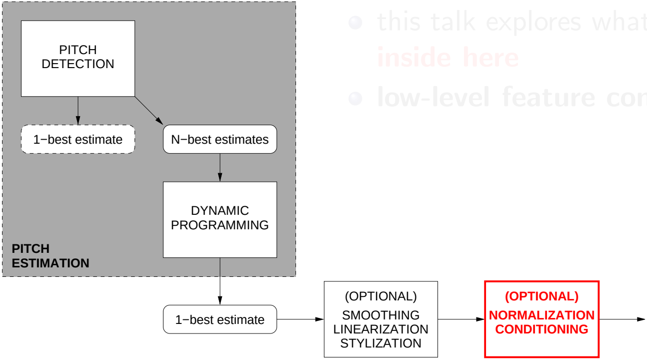
- this talk explores what happens **inside here**
- **low-level feature computation**

Pitch Estimation & Processing



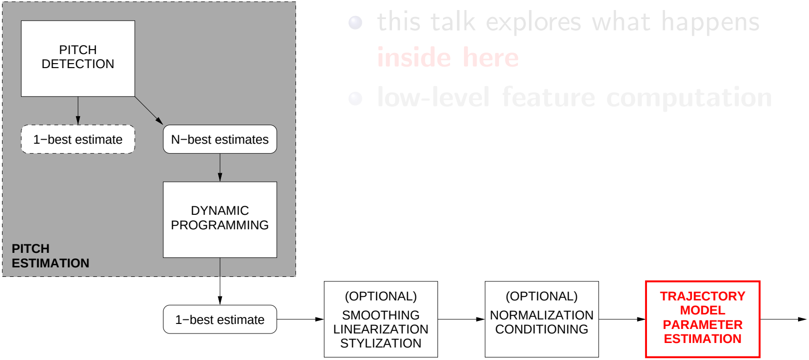
- this talk explores what happens **inside here**
- **low-level feature computation**

Pitch Estimation & Processing



- this talk explores what happens **inside here**
- **low-level feature computation**

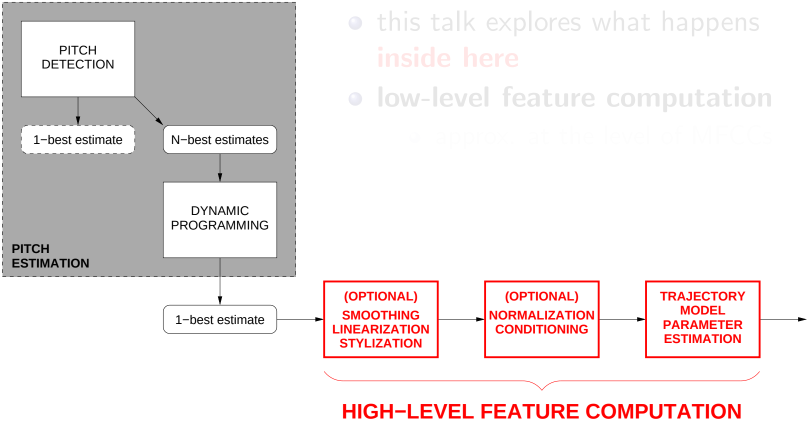
Pitch Estimation & Processing



- this talk explores what happens **inside here**
- low-level feature computation

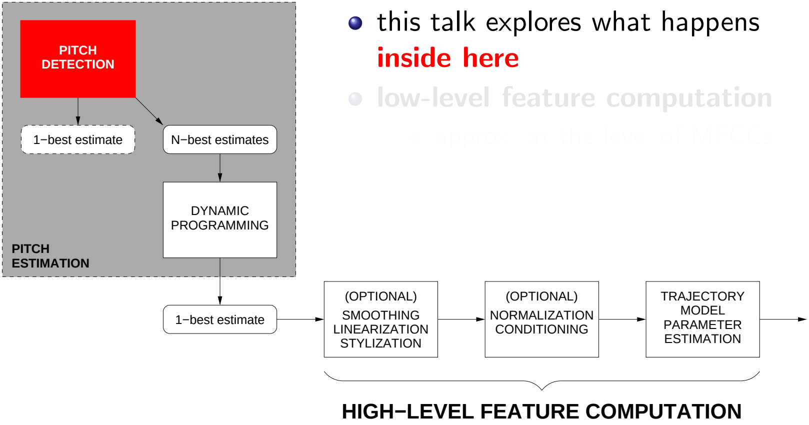
Pitch Estimation & Processing

- this talk explores what happens **inside here**
- **low-level feature computation**
  - approx. at the level of MFCCs



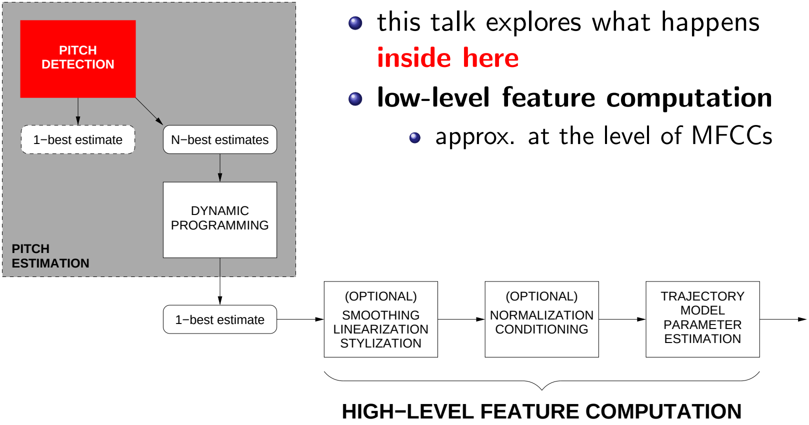
Pitch Estimation & Processing

- this talk explores what happens **inside here**
- low-level feature computation
  - approx. at the level of MFCCs



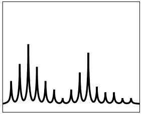
Pitch Estimation & Processing

- this talk explores what happens **inside here**
- **low-level feature computation**
  - approx. at the level of MFCCs



Inside (Per-Frame) Pitch Detection

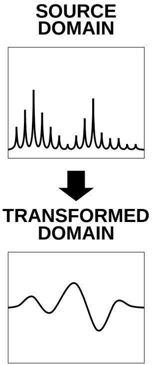
SOURCE DOMAIN



Essentially a 2-step process:

0. begin with a **source-domain x**
   - typically, the short-time FFT
1. compute the **transformed-domain** $\mathbf{y} = f(\mathbf{x})$
   - autocorrelation spectrum
   - real cepstrum
   - comb filterbank energies
   - and many others
2. find the supremum of $\mathbf{y}$, $F_0 = \arg\max \mathbf{y}$

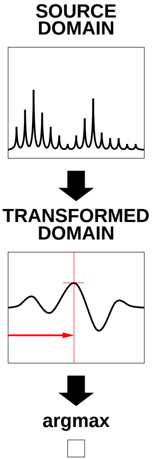
Inside (Per-Frame) Pitch Detection



Essentially a 2-step process:

0. begin with a **source-domain x**
   - typically, the short-time FFT
1. compute the **transformed-domain** $\mathbf{y} = f(\mathbf{x})$
   - autocorrelation spectrum
   - real cepstrum
   - comb filterbank energies
   - and many others
2. find the supremum of $\mathbf{y}$, $F_0 = \arg\max \mathbf{y}$

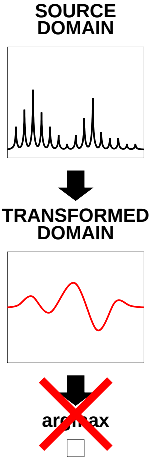
Inside (Per-Frame) Pitch Detection



Essentially a 2-step process:

0. begin with a **source-domain x**
   - typically, the short-time FFT
1. compute the **transformed-domain** $\mathbf{y} = f(\mathbf{x})$
   - autocorrelation spectrum
   - real cepstrum
   - comb filterbank energies
   - and many others
2. find the supremum of $\mathbf{y}$, $F_0 = \arg\max \mathbf{y}$

Inside (Per-Frame) Pitch Detection



Essentially a 2-step process:

0. begin with a **source-domain x**
   - typically, the short-time FFT
1. compute the **transformed-domain** $\mathbf{y} = f(\mathbf{x})$
   - autocorrelation spectrum
   - real cepstrum
   - comb filterbank energies
   - and many others
2. find the supremum of $\mathbf{y}$, $F_0 = \arg\max \mathbf{y}$

Outline of this Talk

1. Harmonic Structure Transform
2. Experiment: closed-set classification, 10-second trials
   - matched-multisession, matched-channel conditions
   - contrast with `get_f0`-estimated pitch
   - contrast with MFCCs
3. Analysis
   - simulated perturbations
   - spectral envelope ablation
4. Conclusions

Schroeder's "Harmonic Product Spectrum"

Given a continuous short-time spectrum $S(f)$, Schroeder proposed

$$\Sigma(f) = 20 \log_{10} \sum_{n=1}^{N} |S(n\, f)|$$

A M. R. Schroeder, 1968. "Period histogram and product spectrum: New methods for fundamental-frequency measurement", *J. Acoust. Soc. Am.* **43**(4):829–834.

B A. M. Noll, 1970. "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate", *Symposium on ComputerProcessing in Communication*, Microwave Institute (University of Brooklyn, New York), **19**:779–797.

Schroeder's "Harmonic Product Spectrum"

Given a continuous short-time spectrum $S(f)$, Schroeder proposed

$$\Sigma(f) = 20 \log_{10} \sum_{n=1}^{N} |S(n\,f)|$$

Noll dubbed this "harmonic compression".
(Distinctly **non-linear**.)

A M. R. Schroeder, 1968. "Period histogram and product spectrum: New methods for fundamental-frequency measurement", *J. Acoust. Soc. Am.* **43**(4):829–834.

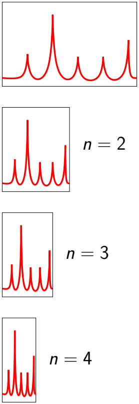
B A. M. Noll, 1970. "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate", *Symposium on ComputerProcessing in Communication*, Microwave Institute (University of Brooklyn, New York), **19**:779–797.

Schroeder's "Harmonic Product Spectrum"

Given a continuous short-time spectrum $S(f)$, Schroeder proposed

$$\Sigma(f) \quad = \quad 20 \log_{10} \sum_{n=1}^{N} |S(n\,f)|$$

Noll dubbed this "harmonic compression".
(Distinctly **non-linear**.)

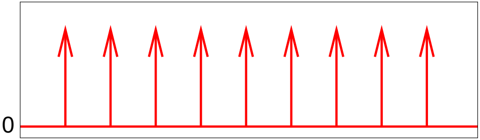


A M. R. Schroeder, 1968. "Period histogram and product spectrum: New methods for fundamental-frequency measurement", *J. Acoust. Soc. Am.* **43**(4):829–834.

B A. M. Noll, 1970. "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate", *Symposium on ComputerProcessing in Communication*, Microwave Institute (University of Brooklyn, New York), **19**:779–797.

Dirac Comb Filterbank

- the alternative: design a continuous-frequency **comb filter**
  - for each candidate fundamental frequency of interest

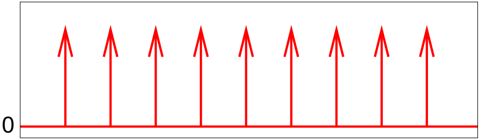


- no "compression difficulties" during discretization
  - filtering is a linear operation
- here: each filter is defined over 300–8000 Hz
- a set of such comb filters (here: 400) yields a **filterbank**
  - from 50 Hz to 450 Hz, spaced 1 Hz apart

A J. A. Moorer, 1974. "The optimum comb method for pitch period analysis of continuous digitized speech", *IEEE Trans. Acoustics, Speech, and Signal Proc.* **22**(5):330–338.

Dirac Comb Filterbank

- the alternative: design a continuous-frequency **comb filter**
  - for each candidate fundamental frequency of interest

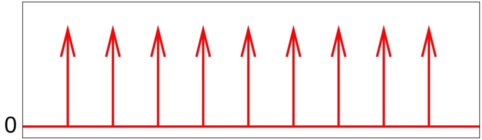


- no "compression difficulties" during discretization
  - filtering is a **linear** operation
- here: each filter is defined over 300–8000 Hz
- a set of such comb filters (here: 400) yields a **filterbank**
  - from 50 Hz to 450 Hz, spaced 1 Hz apart

A J. A. Moorer, 1974. "The optimum comb method for pitch period analysis of continuous digitized speech", *IEEE Trans. Acoustics, Speech, and Signal Proc.* **22**(5):330–338.

Dirac Comb Filterbank

- the alternative: design a continuous-frequency **comb filter**
  - for each candidate fundamental frequency of interest

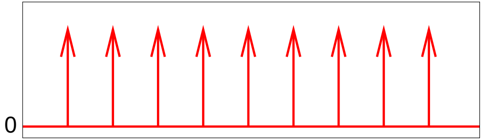


- no "compression difficulties" during discretization
  - filtering is a **linear** operation
- here: each filter is defined over 300–8000 Hz
- a set of such comb filters (here: 400) yields a **filterbank**
  - from 50 Hz to 450 Hz, spaced 1 Hz apart

A J. A. Moorer, 1974. "The optimum comb method for pitch period analysis of continuous digitized speech", *IEEE Trans. Acoustics, Speech, and Signal Proc.* **22**(5):330–338.

Dirac Comb Filterbank

- the alternative: design a continuous-frequency **comb filter**
  - for each candidate fundamental frequency of interest

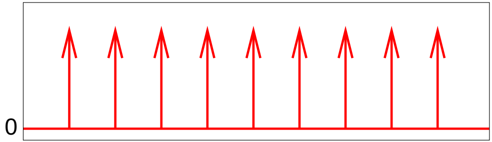


- no "compression difficulties" during discretization
  - filtering is a **linear** operation
- here: each filter is defined over 300–8000 Hz
- a set of such comb filters (here: 400) yields a **filterbank**
  - from 50 Hz to 450 Hz, spaced 1 Hz apart

A J. A. Moorer, 1974. "The optimum comb method for pitch period analysis of continuous digitized speech", *IEEE Trans. Acoustics, Speech, and Signal Proc.* **22**(5):330–338.

Discrete Comb Filterbank

- in software, have a **discrete** FFT **x**
  - sampling frequency: 16 kHz
  - frame size: 32 ms
  - 257 discrete real, non-negative frequencies (bins)



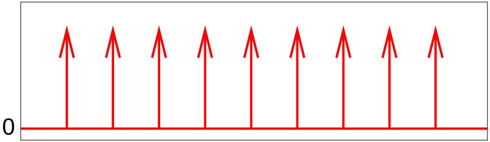
1. here: assume each comb tooth is triangular
2. Riemmann sample the triangular comb filter

- note: the resulting discrete comb filters are **not** harmonic

A J.-S. Liénard, C. Barras & F. Signol, 2008. "Using sets of combs to control pitch estimation errors", *Proc. 155th Meeting ASA*, Paris, France.

Discrete Comb Filterbank

- in software, have a **discrete** FFT **x**
  - sampling frequency: 16 kHz
  - frame size: 32 ms
  - 257 discrete real, non-negative frequencies (bins)



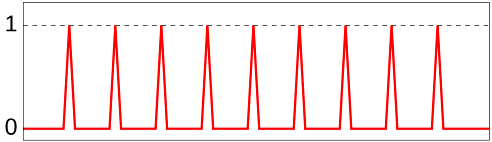
1. here: assume each comb tooth is triangular
2. Riemmann sample the triangular comb filter

- note: the resulting discrete comb filters are **not** harmonic

A J.-S. Liénard, C. Barras & F. Signol, 2008. "Using sets of combs to control pitch estimation errors", *Proc. 155th Meeting ASA*, Paris, France.

Discrete Comb Filterbank

- in software, have a **discrete** FFT **x**
  - sampling frequency: 16 kHz
  - frame size: 32 ms
  - 257 discrete real, non-negative frequencies (bins)



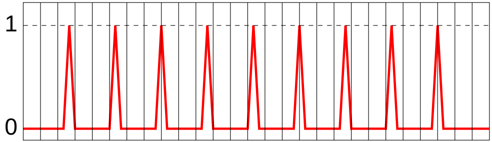
1. here: assume each comb tooth is triangular
2. Riemmann sample the triangular comb filter

- note: the resulting discrete comb filters are **not** harmonic

A J.-S. Liénard, C. Barras & F. Signol, 2008. "Using sets of combs to control pitch estimation errors", *Proc. 155th Meeting ASA*, Paris, France.

Discrete Comb Filterbank

- in software, have a **discrete** FFT **x**
  - sampling frequency: 16 kHz
  - frame size: 32 ms
  - 257 discrete real, non-negative frequencies (bins)



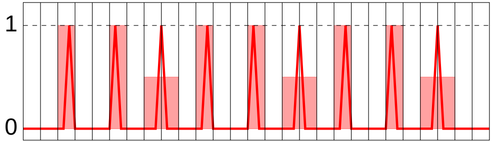
1. here: assume each comb tooth is triangular
2. Riemmann sample the triangular comb filter

- note: the resulting discrete comb filters are **not** harmonic

A J.-S. Liénard, C. Barras & F. Signol, 2008. "Using sets of combs to control pitch estimation errors", *Proc. 155th Meeting ASA*, Paris, France.

Discrete Comb Filterbank

- in software, have a **discrete** FFT **x**
  - sampling frequency: 16 kHz
  - frame size: 32 ms
  - 257 discrete real, non-negative frequencies (bins)



1. here: assume each comb tooth is triangular
2. Riemmann sample the triangular comb filter

- note: the resulting discrete comb filters are **not** harmonic

A J.-S. Liénard, C. Barras & F. Signol, 2008. "Using sets of combs to control pitch estimation errors", *Proc. 155th Meeting ASA*, Paris, France.

Normalizing Harmonic Energy by Non-Harmonic Energy

- the discrete comb filterbank forms a matrix $\mathbf{H}$
  1. its application to FFT $\mathbf{x}$ is a matrix multiplication $\left(\mathbf{H}^T\mathbf{x}\right)$
  2. we take the logarithm at the output (as for Mel energies)
  3. and subtract the log-energy found everywhere else in $\mathbf{x}$

$$
\begin{aligned}
\tilde{\mathbf{H}} &\equiv 1 - \mathbf{H} \\
\mathbf{y} &= \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)
\end{aligned}
$$

- $\mathbf{y}$ is effectively a vector of **harmonic-to-noise ratios** (HNRs)

A E. Yumoto & W. Gould, 1982. "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* **71**(6):1544–1550.

Normalizing Harmonic Energy by Non-Harmonic Energy

- the discrete comb filterbank forms a matrix $\mathbf{H}$
  1. its application to FFT $\mathbf{x}$ is a matrix multiplication $\left(\mathbf{H}^T\mathbf{x}\right)$
  2. we take the logarithm at the output (as for Mel energies)
  3. and subtract the log-energy found everywhere else in $\mathbf{x}$

$$\begin{aligned} \tilde{\mathbf{H}} &= 1 - \mathbf{H} \\ \mathbf{y} &= \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right) \end{aligned}$$

- $\mathbf{y}$ is effectively a vector of **harmonic-to-noise ratios** (HNRs)

A E. Yumoto & W. Gould, 1982. "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* **71**(6):1544–1550.

Normalizing Harmonic Energy by Non-Harmonic Energy

- the discrete comb filterbank forms a matrix $\mathbf{H}$
  1. its application to FFT $\mathbf{x}$ is a matrix multiplication $\left(\mathbf{H}^T\mathbf{x}\right)$
  2. we take the logarithm at the output (as for Mel energies)
  3. and subtract the log-energy found everywhere else in $\mathbf{x}$

$$\begin{aligned}\tilde{\mathbf{H}} &\equiv 1 - \mathbf{H} \\ \mathbf{y} &= \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)\end{aligned}$$

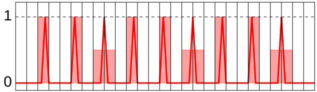
- $\mathbf{y}$ is effectively a vector of **harmonic-to-noise ratios** (HNRs)

A E. Yumoto & W. Gould, 1982. "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* **71**(6):1544–1550.

Normalizing Harmonic Energy by Non-Harmonic Energy

- the discrete comb filterbank forms a matrix $\mathbf{H}$
  1. its application to FFT $\mathbf{x}$ is a matrix multiplication $\left(\mathbf{H}^T\mathbf{x}\right)$
  2. we take the logarithm at the output (as for Mel energies)
  3. and subtract the log-energy found everywhere else in $\mathbf{x}$

$$
\begin{aligned}
\tilde{\mathbf{H}} &\equiv 1 - \mathbf{H} \\
\mathbf{y} &= \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)
\end{aligned}
$$



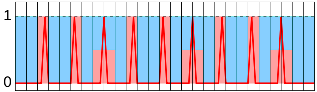
- $\mathbf{y}$ is effectively a vector of **harmonic-to-noise ratios** (HNRs)

A E. Yumoto & W. Gould, 1982. "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* **71**(6):1544–1550.

Normalizing Harmonic Energy by Non-Harmonic Energy

- the discrete comb filterbank forms a matrix $\mathbf{H}$
  1. its application to FFT $\mathbf{x}$ is a matrix multiplication $\left(\mathbf{H}^T\mathbf{x}\right)$
  2. we take the logarithm at the output (as for Mel energies)
  3. and subtract the log-energy found everywhere else in $\mathbf{x}$

$$\begin{aligned} \tilde{\mathbf{H}} &\equiv 1 - \mathbf{H} \\ \mathbf{y} &= \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right) \end{aligned}$$



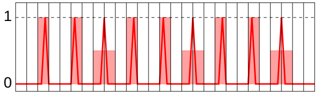
- $\mathbf{y}$ is effectively a vector of **harmonic-to-noise ratios** (HNRs)

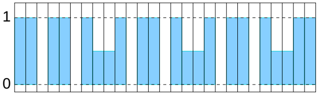
A E. Yumoto & W. Gould, 1982. "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* **71**(6):1544–1550.

Normalizing Harmonic Energy by Non-Harmonic Energy

- the discrete comb filterbank forms a matrix $\mathbf{H}$
  1. its application to FFT $\mathbf{x}$ is a matrix multiplication $\left(\mathbf{H}^T\mathbf{x}\right)$
  2. we take the logarithm at the output (as for Mel energies)
  3. and subtract the log-energy found everywhere else in $\mathbf{x}$

$$\tilde{\mathbf{H}} \equiv 1 - \mathbf{H}$$
$$\mathbf{y} = \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)$$





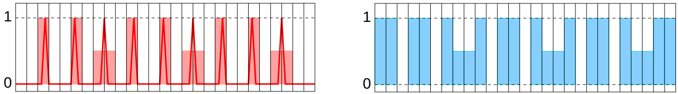
- $\mathbf{y}$ is effectively a vector of **harmonic-to-noise ratios** (HNRs)

A E. Yumoto & W. Gould, 1982. "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* **71**(6):1544–1550.

Normalizing Harmonic Energy by Non-Harmonic Energy

- the discrete comb filterbank forms a matrix $\mathbf{H}$
  1. its application to FFT $\mathbf{x}$ is a matrix multiplication $\left(\mathbf{H}^T\mathbf{x}\right)$
  2. we take the logarithm at the output (as for Mel energies)
  3. and subtract the log-energy found everywhere else in $\mathbf{x}$

$$
\begin{aligned}
\tilde{\mathbf{H}} &\equiv 1 - \mathbf{H} \\
\mathbf{y} &= \log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)
\end{aligned}
$$



- $\mathbf{y}$ is effectively a vector of **harmonic-to-noise ratios** (HNRs)

A E. Yumoto & W. Gould, 1982. "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* **71**(6):1544–1550.

Feature Vector Decorrelation

- the elements of $\mathbf{y}$ are correlated
- transform $\mathbf{y}$ by
  1. subtracting global mean
  2. orthogonalizing (rotating) via data-dependent $\mathcal{F}_{CORR}^{-1}$
  3. truncating non-positive eigenvalue dimensions
- yields the *harmonic structure cepstral coefficients*

$$\begin{aligned}
\text{HSCC} &= \mathcal{F}_{CORR}^{-1}\left(\log\left(\mathbf{H}^T\mathbf{x}\right) - \log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)\right) \\
&= \mathcal{F}_{CORR}^{-1}\left(\log\left(\mathbf{H}^T\mathbf{x}\right)\right) - \underbrace{\mathcal{F}_{CORR}^{-1}\left(\log\left(\tilde{\mathbf{H}}^T\mathbf{x}\right)\right)}_{\text{normalization term}}
\end{aligned}$$

- two options for $\mathcal{F}_{CORR}^{-1}$:
  1. PCA: conditionally independent of labels
  2. LDA: conditioned on labels

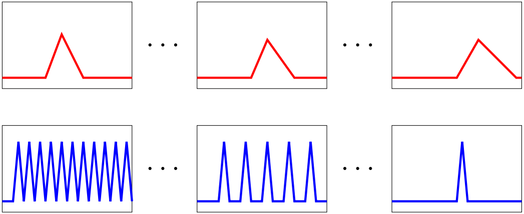
Similarities with the Mel Filterbank, M

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

$$\text{HSCC} = \mathcal{F}^{-1}_{CORR}\left(\log\left(\mathbf{H}^T\mathbf{x}\right)\right) - \langle\text{normalization term}\rangle$$

columns of **M**

columns of **H**



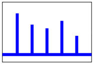
Similarities with the FFV Spectrum

HST (here) **FFV (previous work)**

Similarities with the FFV Spectrum

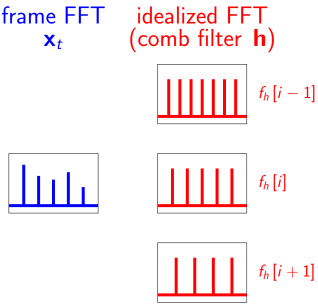
HST (here) **FFV (previous work)**

frame FFT
$\mathbf{x}_t$



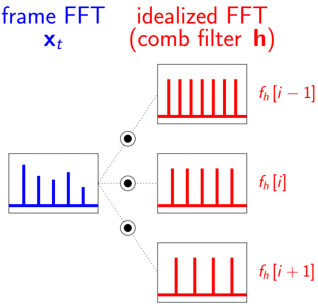
Similarities with the FFV Spectrum

HST (here) **FFV (previous work)**



Similarities with the FFV Spectrum

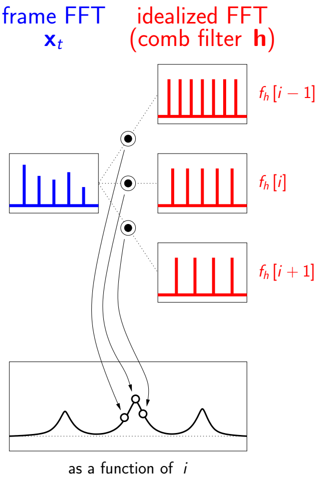
HST (here)



FFV (previous work)

Similarities with the FFV Spectrum

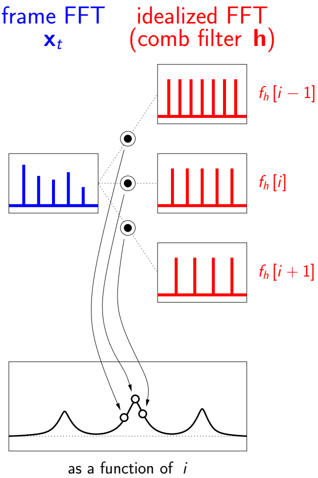
HST (here)



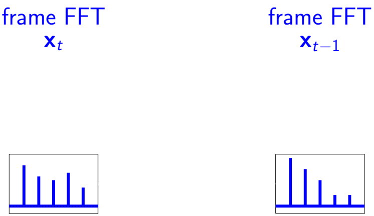
FFV (previous work)

Similarities with the FFV Spectrum

HST (here)

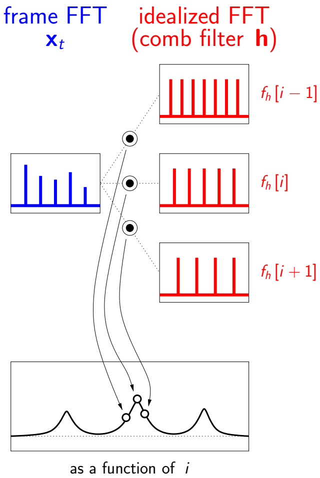


FFV (previous work)

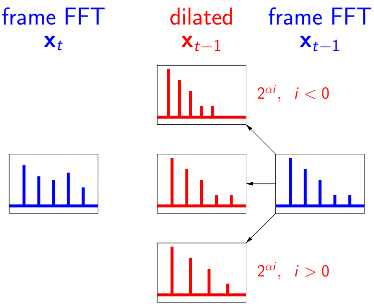


Similarities with the FFV Spectrum

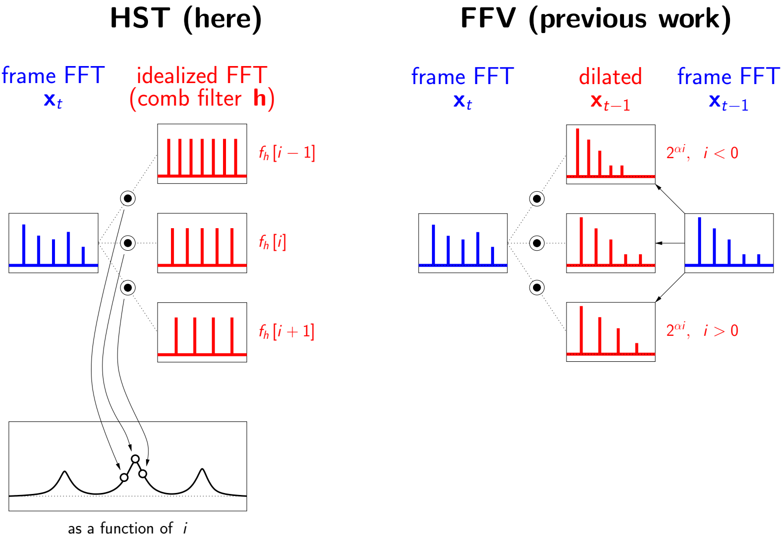
HST (here)



FFV (previous work)

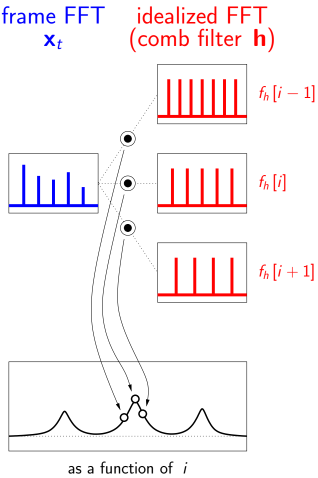


Similarities with the FFV Spectrum

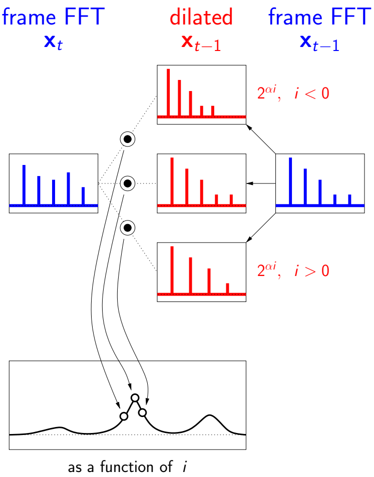


Similarities with the FFV Spectrum

HST (here)



FFV (previous work)



Experiments: Data

- WSJ: LDC CSR-I (WSJ0) & LDC CSR-II (WSJ1)
- 102 female (♀) speakers, 95 male (♂) speakers
- closed-set classification, 10-second trials
  - TRAINSET: 5 minutes
  - DEVSET: 3 minutes, # trials: 1775 (♀) and 1660 (♂)
  - TESTSET: 3 minutes, # trials: 1510 (♀) and 1412 (♂)
- matched channel, Sennheiser HMD414 (`.wv1`)
- matched multi-session:
  - 4–20 sessions per speaker
  - TRAIN-/DEV-/TEST- SETs drawn from most sessions

F_0/GMM Baseline System (not in paper)

1. extract F_0 using `get_f0`
   - Snack Sound Toolkit: ESPS, default settings
   - note: relies on dynamic programming
2. transform voiced frames to $\log_2$ domain
   - ignore unvoiced frames

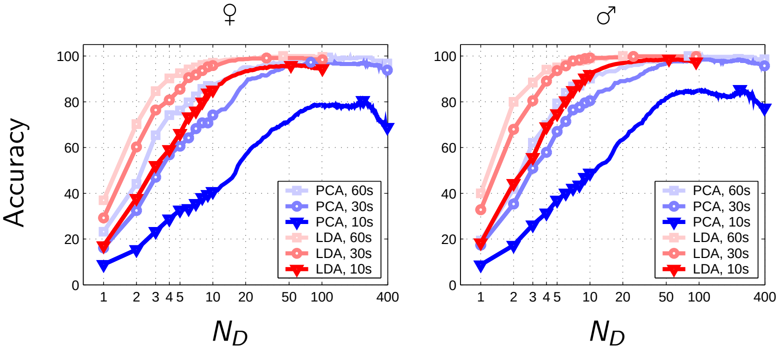
N_G	Female		Male	
	DevSet	EvalSet	DevSet	EvalSet
1	12.31	12.71	17.15	17.41
8	17.48	17.94	25.91	27.62
16	16.70	17.44	**26.21**	**27.44**
256	**17.62**	**18.36**	25.91	26.02

HSCC System Configuration

Parameter/Aspect	HSCC System
pre-emphasis	no
framing	8ms/32ms
window	Hann
N_D	*to optimize*
N_G	*to optimize*
UBM	no
SAD	no

HSCC Vector Rotation and Truncation

- pick number of dimensions N_D
  - set number of (diagonal-covariance) Gaussians $N_G = 1$
  - train PCA, LDA on TRAINSET
  - choose N_D to maximize accuracy on DEVSET



Results I

- with N_D fixed, find N_G to maximize DevSet accuracy $\rightarrow$ 256

System	Female, ♀		Male, ♂	
	Dev	Test	Dev	Test
`get_f0`	17.62	18.36	26.21	27.44
HSCC/LDA	99.72	99.87	99.70	99.65

1. there **is** speaker-discriminative information in the transformed-domain, **beyond the** $\arg\max$
   - discarding it leads to much worse performance
2. improving $\arg\max$ estimation appears unnecessary
   - $\arg\max$ estimation = pitch estimation

Contrastive MFCC/GMM System

Parameter/Aspect	HSCC System	MFCC System
pre-emphasis	no	yes
framing	8ms/32ms	8ms/32ms
window	Hann	Hamming
N_D	52-53 (opt)	20
N_G	256 (opt)	256 (opt)
UBM	no	no
SAD	no	no

Results II

System	Female, ♀		Male, ♂	
	Dev	Test	Dev	Test
HSCC/LDA	99.72	99.87	99.70	99.65
MFCC	98.66	99.27	99.34	98.58
MFCC/LDA	98.71	99.27	99.34	98.87
HSCC/LDA $\oplus$ MFCC	100.00	100.00	99.70	99.87

1. HSCC performance comparable to MFCC performance
   - in these experiments, always better
2. equal-weight score-level fusion can yield improvement
   - HSCC and MFCC appear complementary

Some Perturbations

Evaluate several types of perturbation:

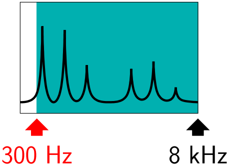
1. source-domain frequency range ablation
   - low frequency (LF) cutoff
   - high frequency (HF) cutoff
2. transformed-domain frequency resolution
3. source-domain spectral envelope ablation

Simplify analysis suite by:

- using $N_G = 1$ diagonal-covariance Gaussian per speaker
- computing accuracy $\textsc{DevSet}$ only
- plotting accuracy as a function of N_D

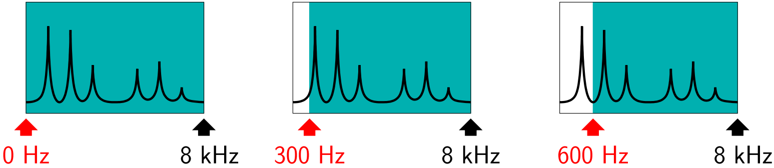
Source-Domain Low Frequency (LF) Range

- modify the low-frequency cutoff for source-domain (FFT) **x**



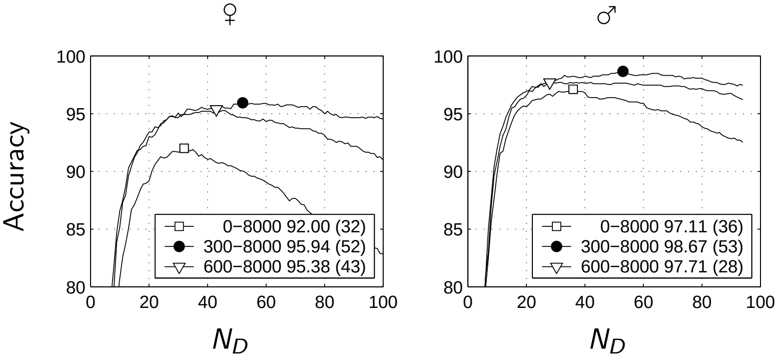
Source-Domain Low Frequency (LF) Range

- modify the low-frequency cutoff for source-domain (FFT) **x**



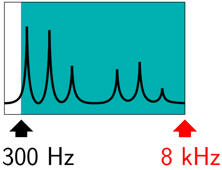
Source-Domain Low Frequency (LF) Range

- modify the low-frequency cutoff for source-domain (FFT) **x**



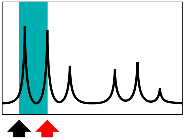
Source-Domain High Frequency (HF) Range

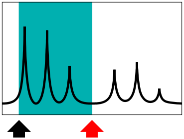
- modify the high-frequency cutoff for source-domain (FFT) **x**

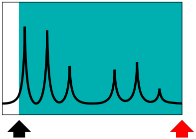


Source-Domain High Frequency (HF) Range

- modify the high-frequency cutoff for source-domain (FFT) **x**

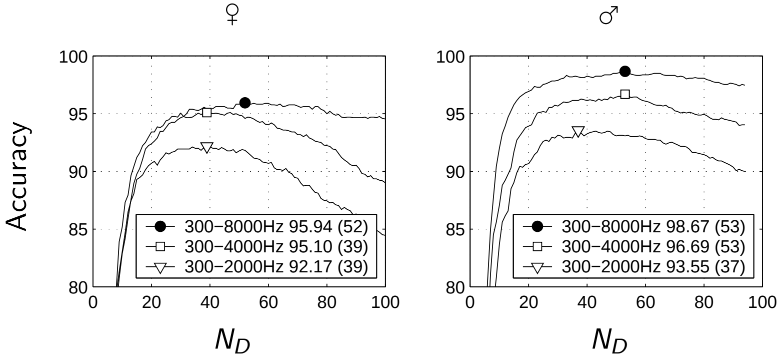






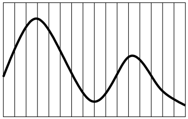
Source-Domain High Frequency (HF) Range

- modify the high-frequency cutoff for source-domain (FFT) **x**



Transformed-Domain Frequency Resolution

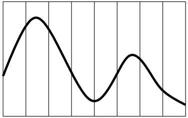
- modify the resolution of the transformed-domain **y**



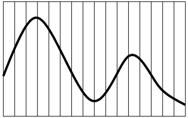
400 filters
1.0 Hz apart

Transformed-Domain Frequency Resolution

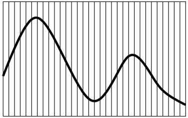
- modify the resolution of the transformed-domain **y**



200 filters
2.0 Hz apart



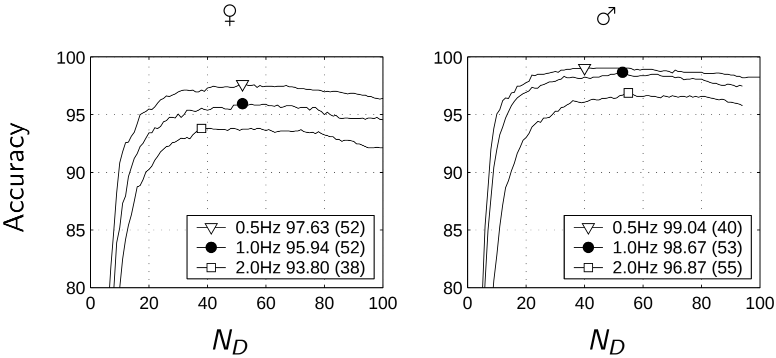
400 filters
1.0 Hz apart



800 filters
0.5 Hz apart

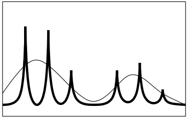
Transformed-Domain Frequency Resolution

- modify the resolution of the transformed-domain **y**



Source-Domain Spectral Envelope Ablation

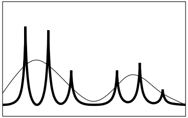
- lifter the low-quefrency components of source-domain (FFT) $\mathbf{x}$
- low-order CCs approximate low-order MFCCs



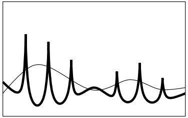
lifter 0 CCs

Source-Domain Spectral Envelope Ablation

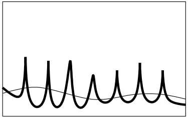
- lifter the low-quefrency components of source-domain (FFT) **x**
- low-order CCs approximate low-order MFCCs



lifter 0 CCs



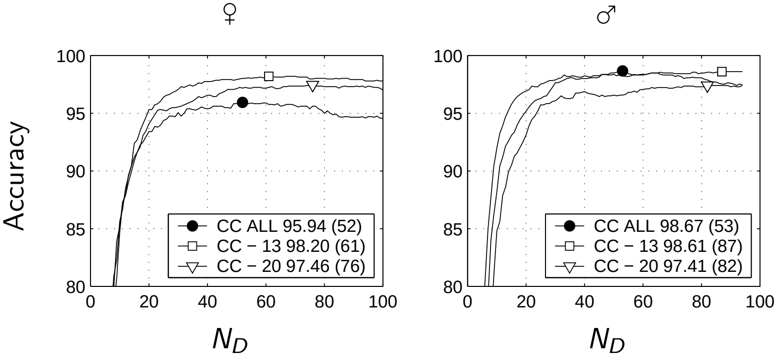
lifter 13 CCs



lifter 20 CCs

Source-Domain Spectral Envelope Ablation

- lifter the low-quefrency components of source-domain (FFT) **x**
- low-order CCs approximate low-order MFCCs



Analysis Findings

- HSCC representation appears to be robust to perturbation
  - low-frequency source-domain range (♀: 4%, ♂: 1.5%)
  - high-frequency source-domain range (♀: 4%, ♂: 5%)
  - transformed domain resolution (♀: 4%, ♂: 2%)
  - source-domain envelope ablation (♀: 2.5%, ♂: 1.5%)
- generally, performance for ♀ speakers more sensitive
- even under perturbed conditions, vastly outperform the system based on pitch alone
- not known how a pitch tracker would perform

Summary of Findings

1. **Information available to (but discarded by) (some) pitch trackers is valuable.**
2. HSCC performance is comparable to MFCC performance.
3. HSCC information is complimentary to MFCC information.
4. HSCC modeling is as easy as MFCC modeling.

Recommendations/Impact

The presented evidence suggests:

1. should not invest time in improving estimation of the transformed-domain arg max (i.e., pitch)
   - **simply model the entire transformed-domain**
2. if require pitch for other ("high-level") features
   - **should not discard transformed-domain** following arg max estimation
3. using the entire transformed-domain may lead to a **paradigmatic shift in the modeling of prosody**

Of Immediate Interest ...

1. don't know how the HSCC vector compares to other "instantaneous" prosody vectors
2. don't know how the HSCC vector performs under session, channel, distance, or vocal effort mismatch conditions
3. other classifiers might be better-suited to the size of the transformed-domain (SVMs, etc.)
4. existing prosody systems employ high-level features
    - first-, second-, Nth-order differences
    - modulation spectrum
5. would prefer data-independent feature rotation/compression
    - would significantly improve understanding
    - would permit UBMing
    - would allow use in large-dataset tasks (e.g., NIST SRE)

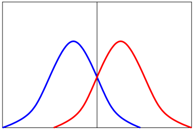
Thank You!

This work was particularly inspired by:

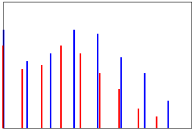
1. J.-S. Liénard, C. Barras & F. Signol, 2008. "Using sets of combs to control pitch estimation errors", *Proc. 155th Meeting ASA*, Paris, France.
2. M. R. Schroeder, 1968. "Period histogram and product spectrum: New methods for fundamental-frequency measurement", *JASA* **43**(4):829–834.
3. A. F. Huxley, 1969. "Is resonance possible in the cochlea after all?", *Nature* **221**:935-940.

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}\left[\rho\right]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



time domain

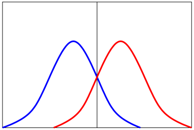


freq domain

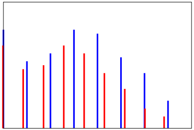
- pass $\mathbf{g}\left(\rho\right)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



time domain

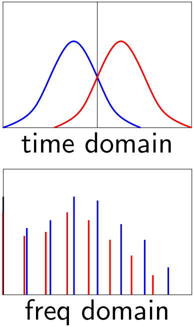


freq domain

- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

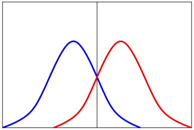
- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



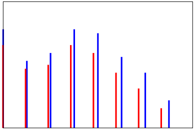
- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



time domain

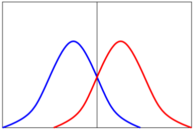


freq domain

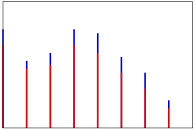
- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



time domain

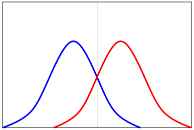


freq domain

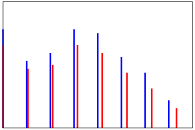
- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



time domain

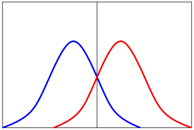


freq domain

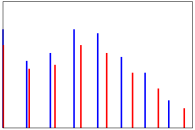
- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



time domain

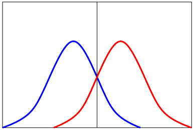


freq domain

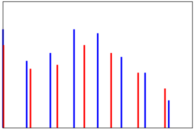
- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values



time domain

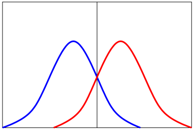


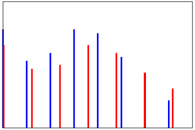
freq domain

- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values

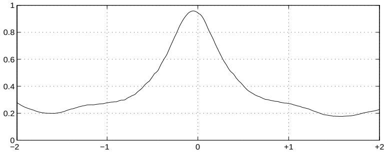

time domain

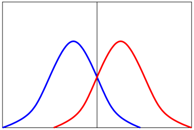

freq domain

- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

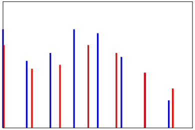
Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}\left[\rho\right]$
  - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
  - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
  - dot product with undilated $\mathbf{F}_L$
  - repeat for a continuum of ρ values





time domain

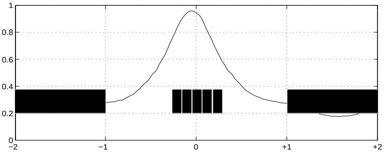


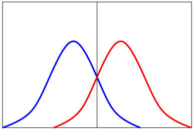
freq domain

- pass $\mathbf{g}\left(\rho\right)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

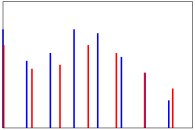
Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}\,[\rho]$
    - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
    - dilate $\mathbf{F}_R$ by a factor 2^ρ, $\rho > 0$
    - dot product with undilated $\mathbf{F}_L$
    - repeat for a continuum of ρ values





time domain

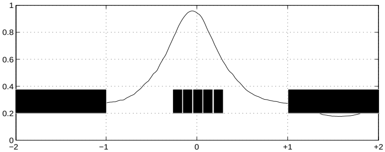


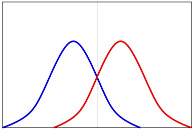
freq domain

- pass $\mathbf{g}\,(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

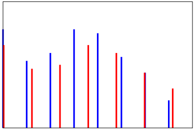
Fundamental Frequency Variation

- estimate the FFV spectrum $\mathbf{g}[\rho]$
    - estimate the power spectra $\mathbf{F}_L$ and $\mathbf{F}_R$
    - dilate $\mathbf{F}_R$ by a factor 2^{ρ}, $\rho > 0$
    - dot product with undilated $\mathbf{F}_L$
    - repeat for a continuum of ρ values





time domain



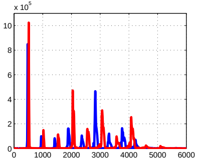
freq domain

- pass $\mathbf{g}(\rho)$ through a filterbank to yield $\mathbf{G} \in \mathbb{R}^7$
- decorrelate $\mathbf{G}$

Fundamental Frequency Variation (3)

$\underline{\rho = 2^{-0.0342} = 0.9766}$

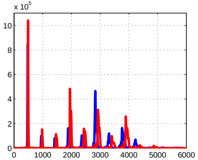
leave left FFT as is
dilate right FFT by ρ



$g(\rho) = 0.0261$

$\underline{\rho = 2^{0} = 1}$

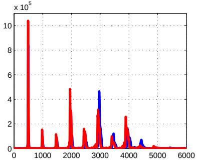
leave left FFT as is
leave right FFT as is



$g(\rho) = 0.2299$

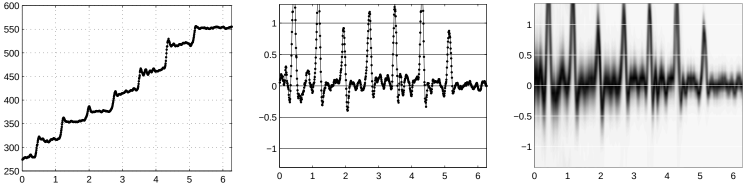
$\underline{\rho = 2^{+0.0342} = 1.0240}$

dilate left FFT by ρ
leave right FFT as is



$g(\rho) = 0.6877$

Fundamental Frequency Variation (3)



Some Distant Numbers ?

	EvalSet1 (Sess Mat)		EvalSet2 (Sess Mis)	
	Chan Mat	Chan Mis	Chan Mat	Chan Mis
MFCC	100.0	95.2	77.3	66.2
HSCC_{old}	100.0	67.0	52.5	31.9
HSCC_{new}	100.0	78.3	67.5	48.1
err (%rel)	0	34.2	31.6	24.8

Table: Classification accuracy (in %) using several different feature types, including the improved harmonic structure cepstral coefficients HSCC_{new}, in matched ("Mat") and mismatched ("Mis") session ("Sess") and channel ("Chan") conditions. "err (%rel)" indicates the relative reduction of error, in percent, from HSCC_{old} to HSCC_{new}.

What Do HSCCs Represent?