Linear Predictive Coding

- The problem of linear prediction is the estimation of the set of coefficients a_k from the input signal x(n). The standard solution minimizes the mean square error:

$$\frac{1}{N}\sum_n e^2(n) = \frac{1}{N}\sum_n [x(n) - \sum_{k=1}^{p} a_k x(n-k)]^2$$

$$\frac{\partial E_n}{\partial a_i} = -2\sum_n [(x(n)x(n-i) - \sum_{k=1}^{p} a_k x(n-k)x(n-i)] = 0$$

$$\boxed{\sum_{k=1}^{p} a_k(n)\phi(|i-k|) = \phi(i) \quad s.t. \quad \phi(k) = \sum_n x(n)x(n+k)}$$

Linear Predictive Coding

Can be written in matrix form as:

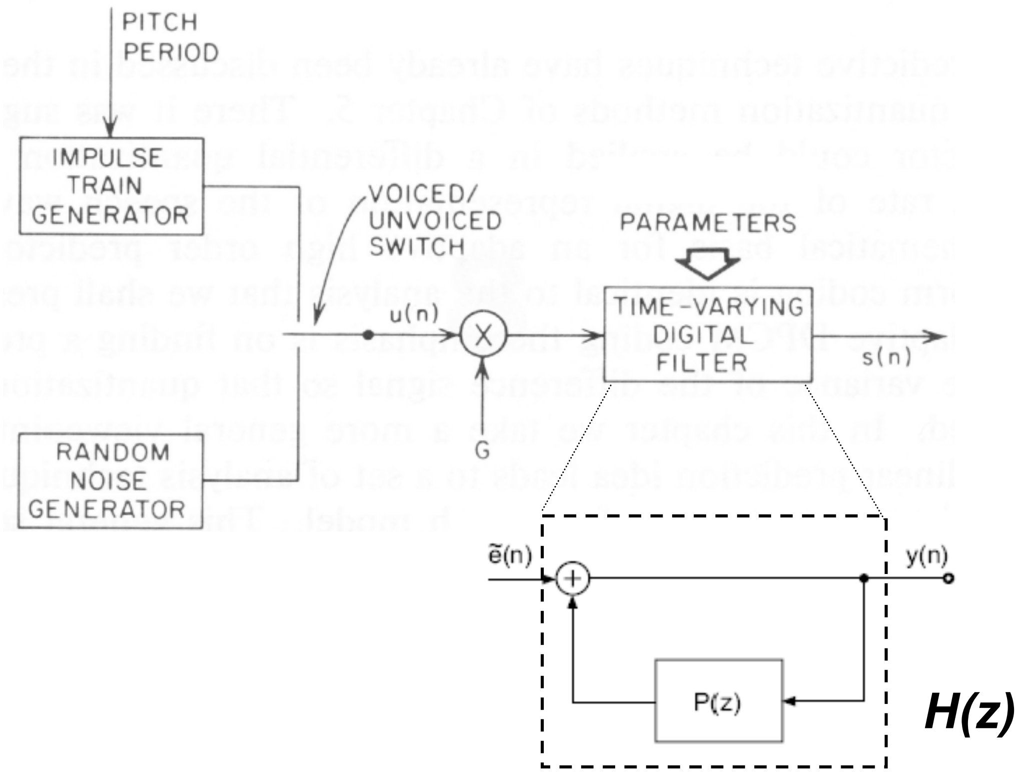
$$\mathbf{\Phi} a = \psi$$

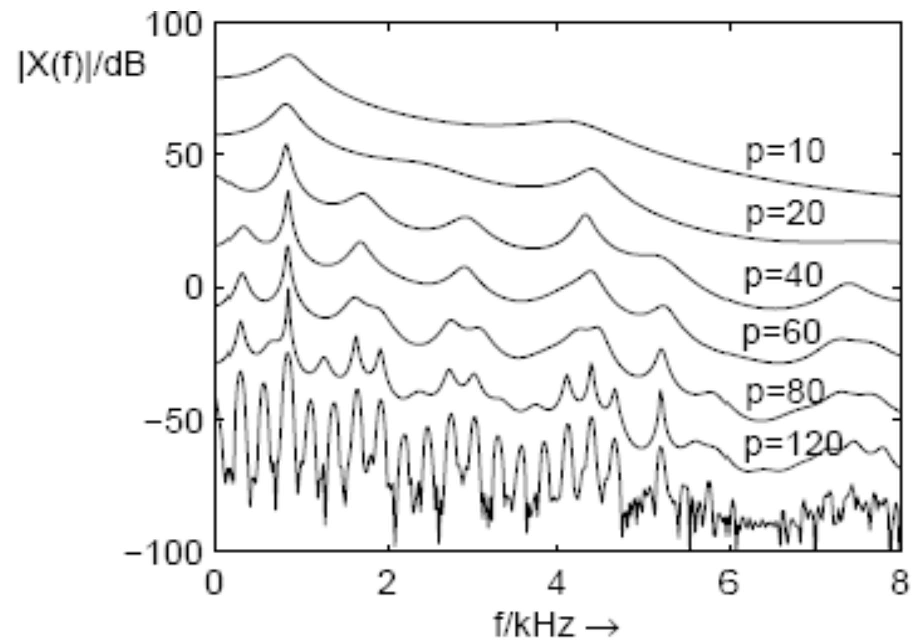
where:

$$\mathbf{\Phi_{ik}} = \phi(|i-k|)$$
$$a = a_k$$
$$\psi = \phi(k)$$

which can be solved using the Levinson-Durbin recursion

Linear Predictive Coding





Remember Cepstrum?

- Treat the log magnitude spectrum as if it were a signal -> take its (I)DFT
- Measures rate of change across frequency bands (Bogert et al., 1963)
- For a real-valued signal it's defined as:

$$\boxed{c_x(l) = real(IFFT(log(|FFT(x)|)))}$$

- Followed by low-pass "liftering" in the cepstral domain

Cepstrum

- The real cepstrum can be weighted using a low-pass window of the form:

$$w_{LP}(l) = \begin{cases} 1 & \text{if } l = 0, L_1 \\ 2 & \text{if } 1 \leq l < L_1 \\ 0 & \text{if } L_1 < l \leq L - 1 \end{cases}$$

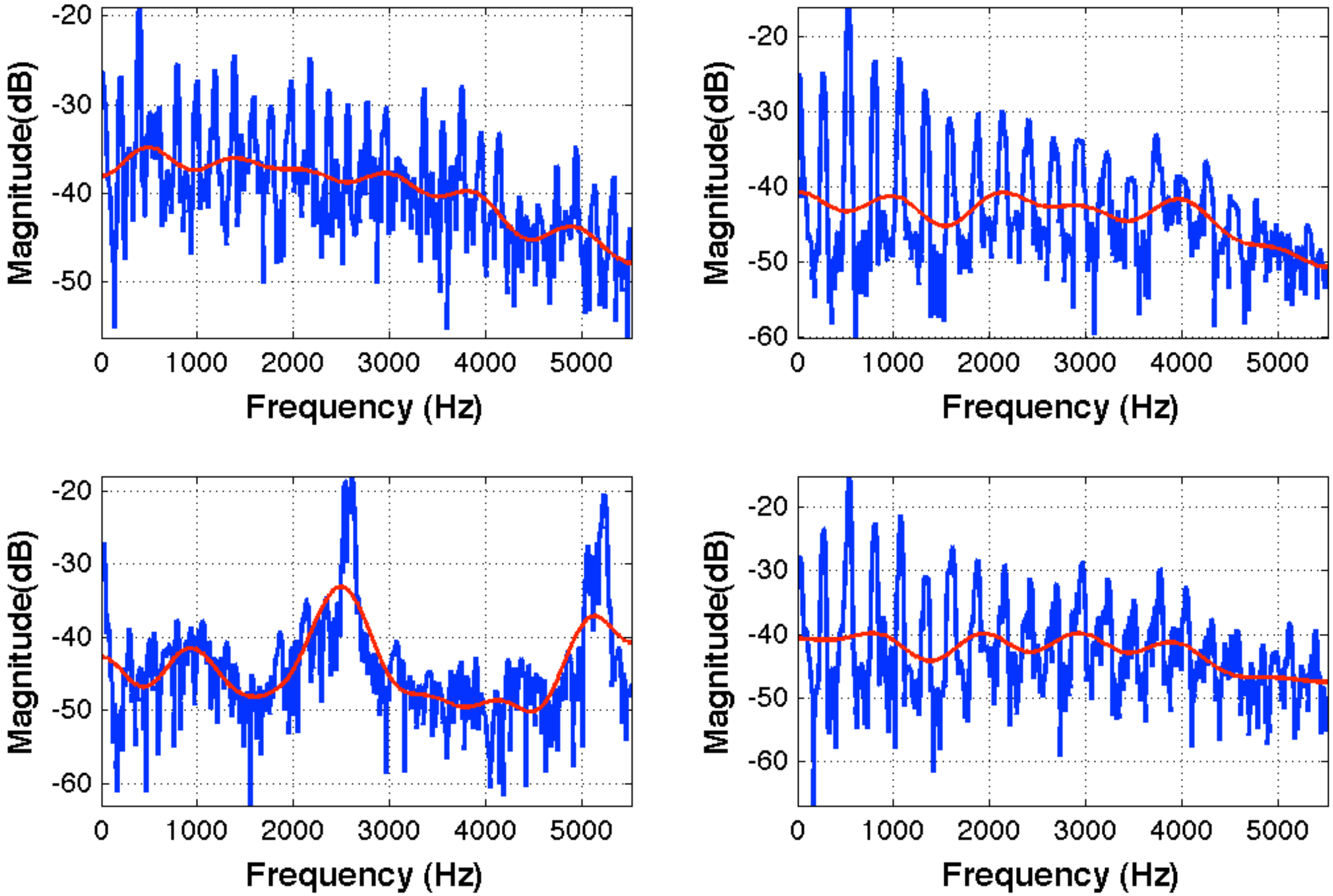
$$c_{LP}(l) = c_x(l) \times w_{LP}(l)$$

$$C_{LP}(k) = e^{\mathbb{R}[\mathrm{FFT}(c_{LP}(l))]}$$

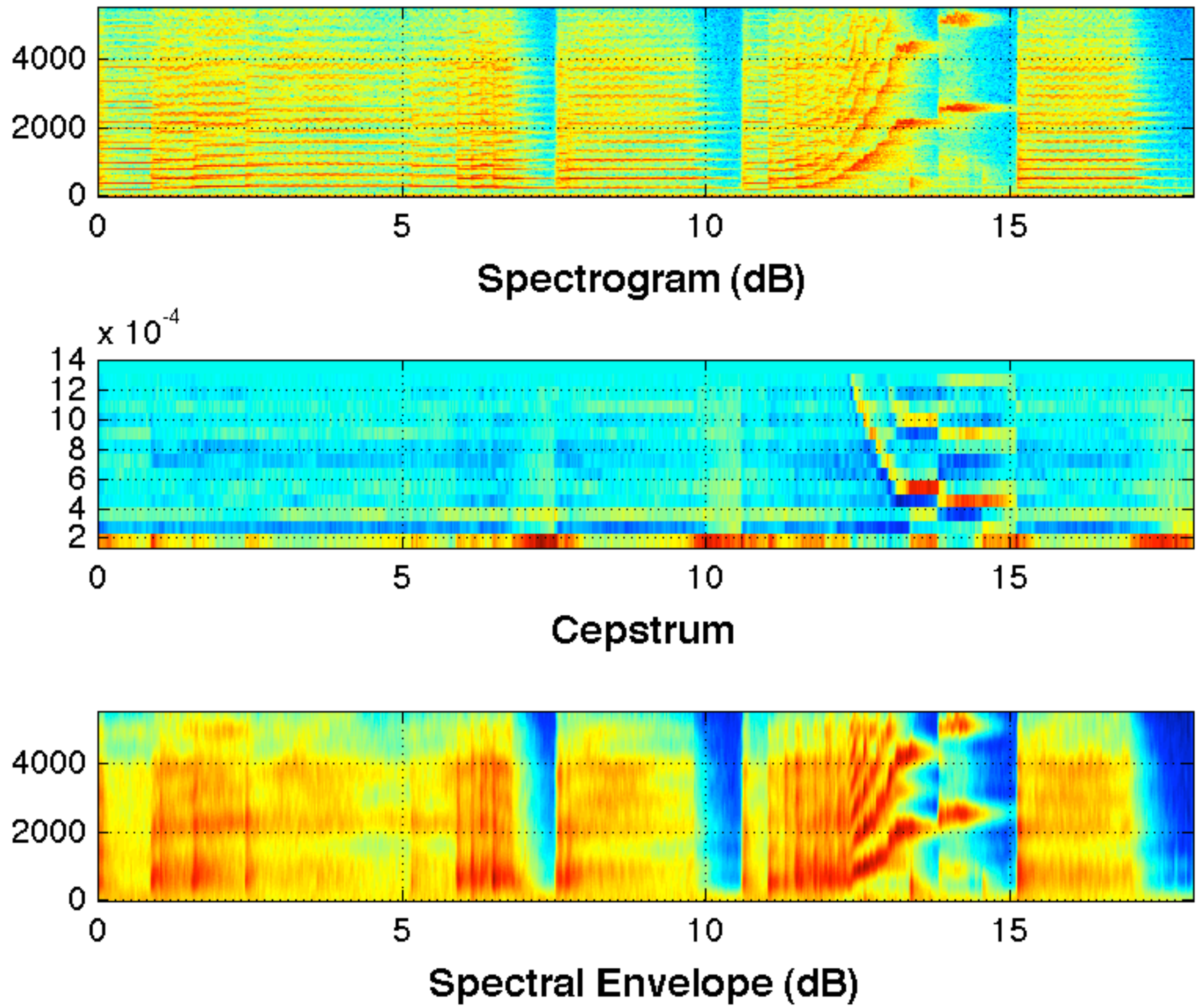
- Such that $L_1 \leq L/2$, and C_{LP} is the spectral envelope.

Cepstrum

- The spectral envelope approximation is coarser/finer depending on L_1



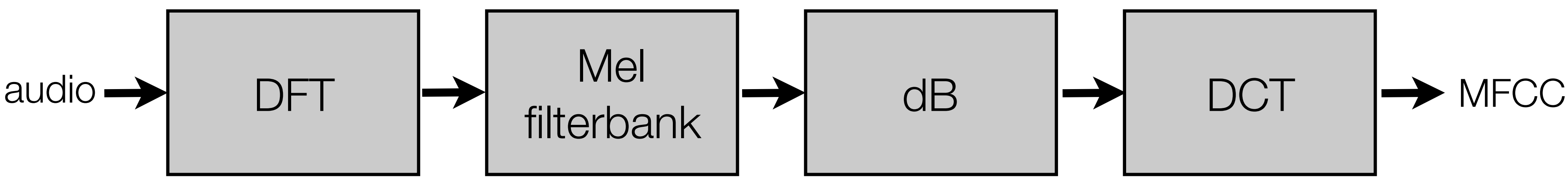
Cepstrum


4000
2000
0
0
5
10
15
Spectrogram (dB)
x 10^-4
14
12
10
8
6
4
2
0
5
10
15
Cepstrum
4000
2000
0
0
5
10
15
Spectral Envelope (dB)

MFCC

- Mel-frequency Cepstral Coefficients (MFCC): variation of the linear cepstrum, widely used in audio analysis.

- Most popular features in speech (Gold et al, 2011): due to their ability to compactly represent the audio spectrum. Introduced to music DSP by Logan (ISMIR, 2000). Ubiquitous in environmental sound analysis.



MFCC

- The Mel scale is a non-linear perceptual scale of pitches judged to be equidistant:

$$\mathrm{mel} = 1127.01028 \times \log\left(1 + \frac{f}{700}\right)$$

$$f = 700 \times (e^{\frac{\mathrm{mel}}{1127.01028}} - 1)$$

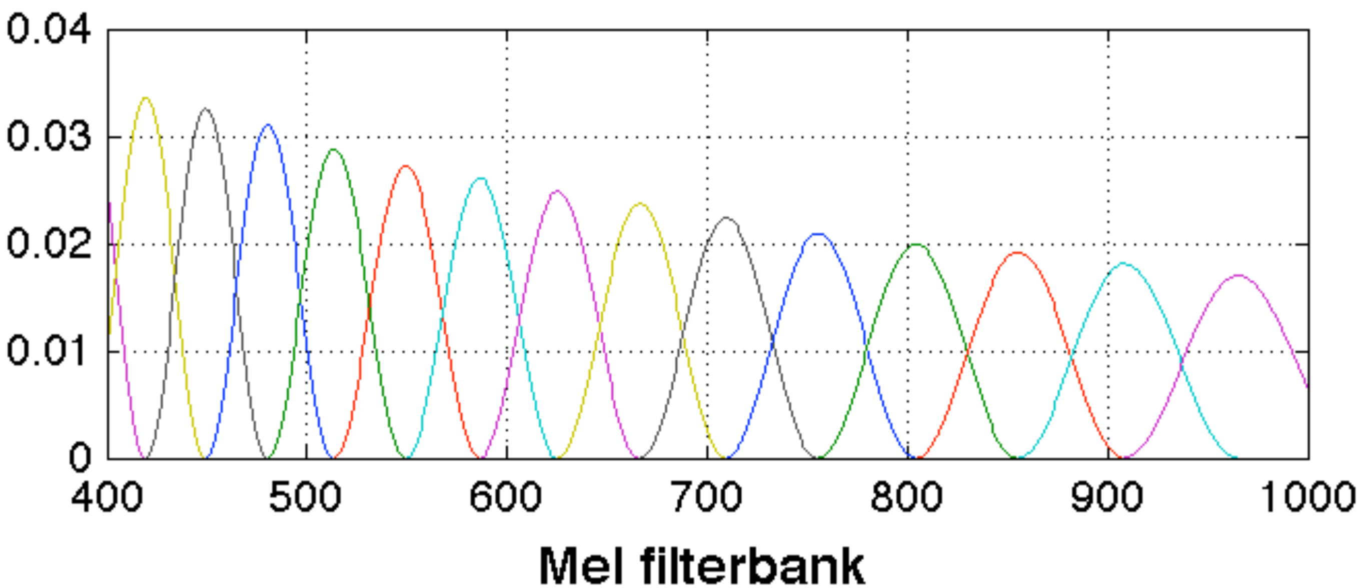
- Approx. linear $f < 1$kHz; logarithmic above that.

- Reference point is at f = 1kHz, which corresponds to 1000 Mel: a tone perceived to be half as high is 500 Mel, twice as high is 2000 Mel, etc.

MFCC

- Filterbank of overlapping windows

- Center frequencies uniformly distributed in mel scale, s.t. the center frequency of one window: starting point of next window and end point of previous window.



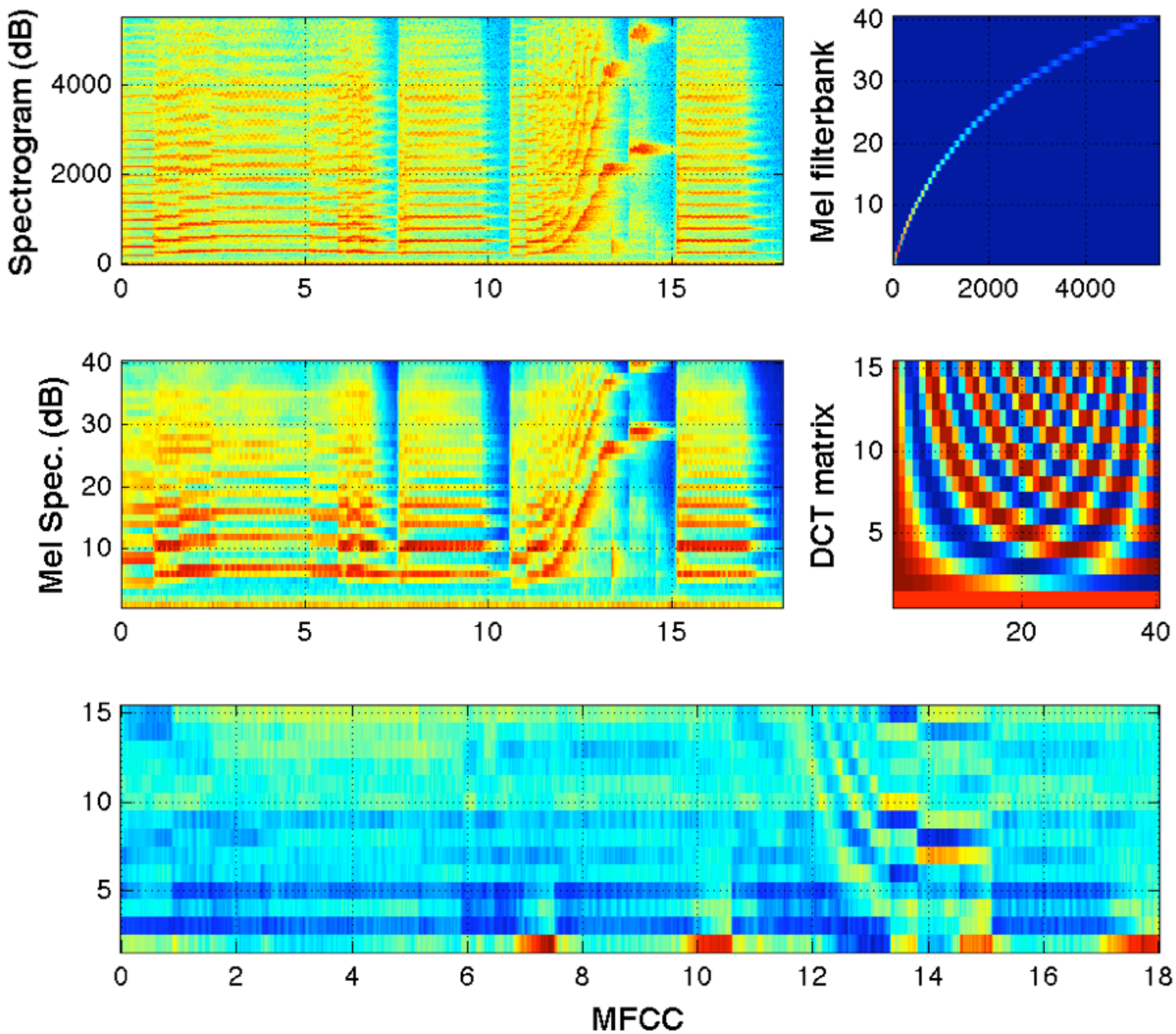
- All windows are normalized to unity sum.

MFCC

- An efficient representation of the log-spectrum can be obtained by applying a transform that decorrelates the Mel dB spectrum (see Rabiner and Juang, 93).
- This decorrelation is commonly approximated by means of the Discrete Cosine Transform (DCT)
- DCT: real-valued transform, similar to the DFT. Most of its energy is concentrated on a few low coefficients (effectively compressing the spectrum)

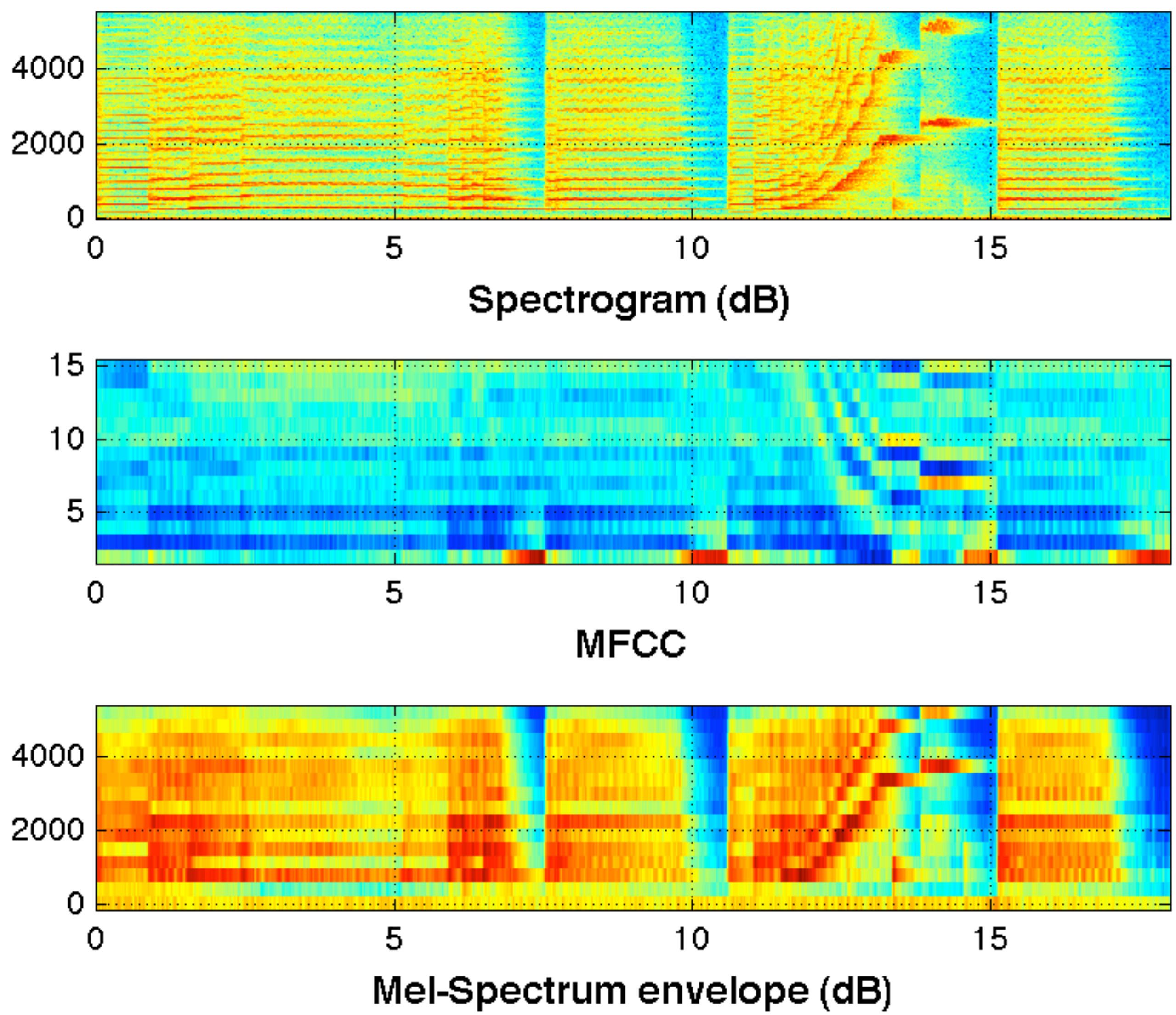
$$X_{DCT}(k) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi k}{N}\left(n - \frac{1}{2}\right)\right]$$

MFCC



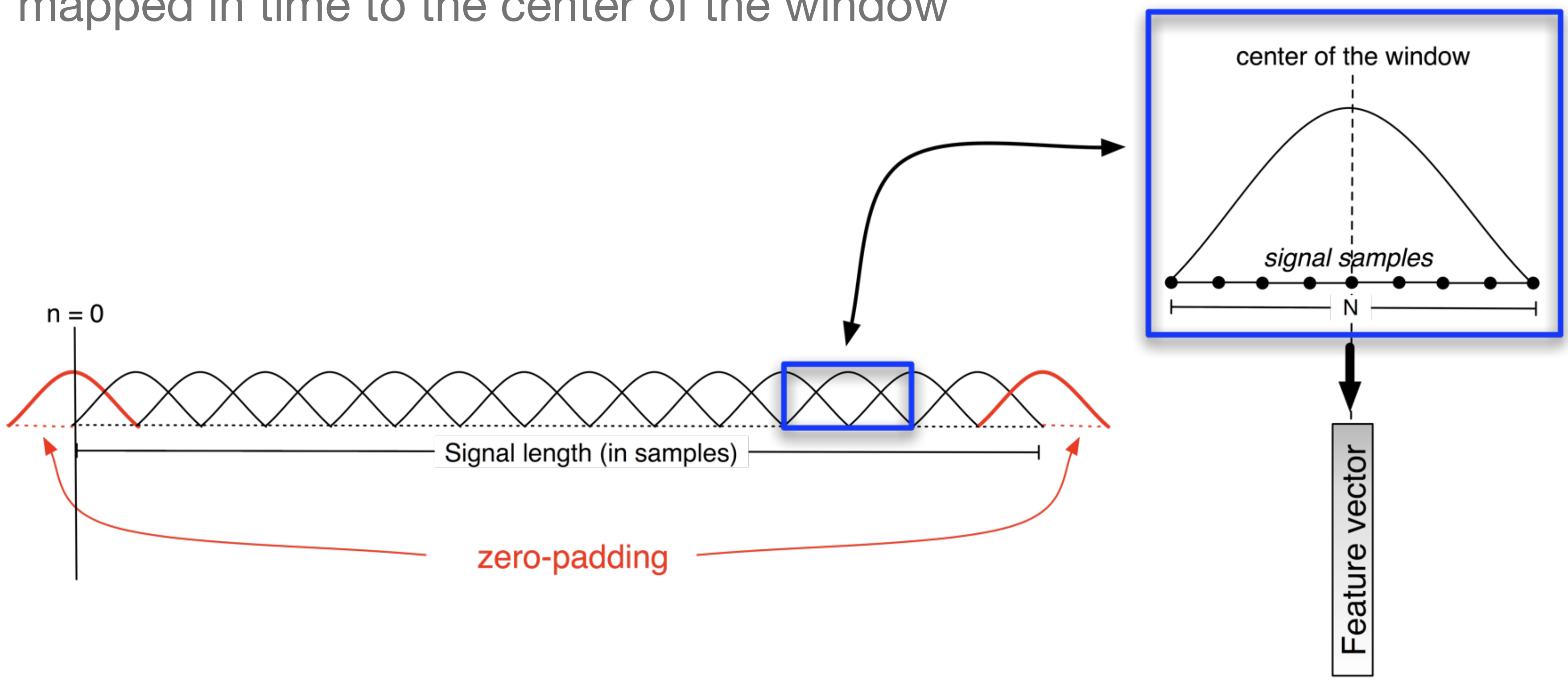
Spectrogram (dB)
4000
2000
0
0
5
10
15
Mel filterbank
40
30
20
10
0
2000
4000
Mel Spec. (dB)
40
30
20
10
0
5
10
15
DCT matrix
15
10
5
20
40
15
10
5
0
2
4
6
8
10
12
14
16
18
MFCC

MFCC


4000
2000
0
0
5
10
15
Spectrogram (dB)
15
10
5
0
5
10
15
MFCC
4000
2000
0
0
5
10
15
Mel-Spectrum envelope (dB)

A reminder

- The feature vector is representing an N-long time segment, and is best mapped in time to the center of the window



- Zero-padding can be used to map the first vector to n = 0, and ensure all the signal is analyzed

Post-processing

- We can characterize the short-term temporal dynamics of feature coefficients by using delta and acceleration coefficients:

$$\Delta y = \frac{y(n) - y(n-h)}{h}$$

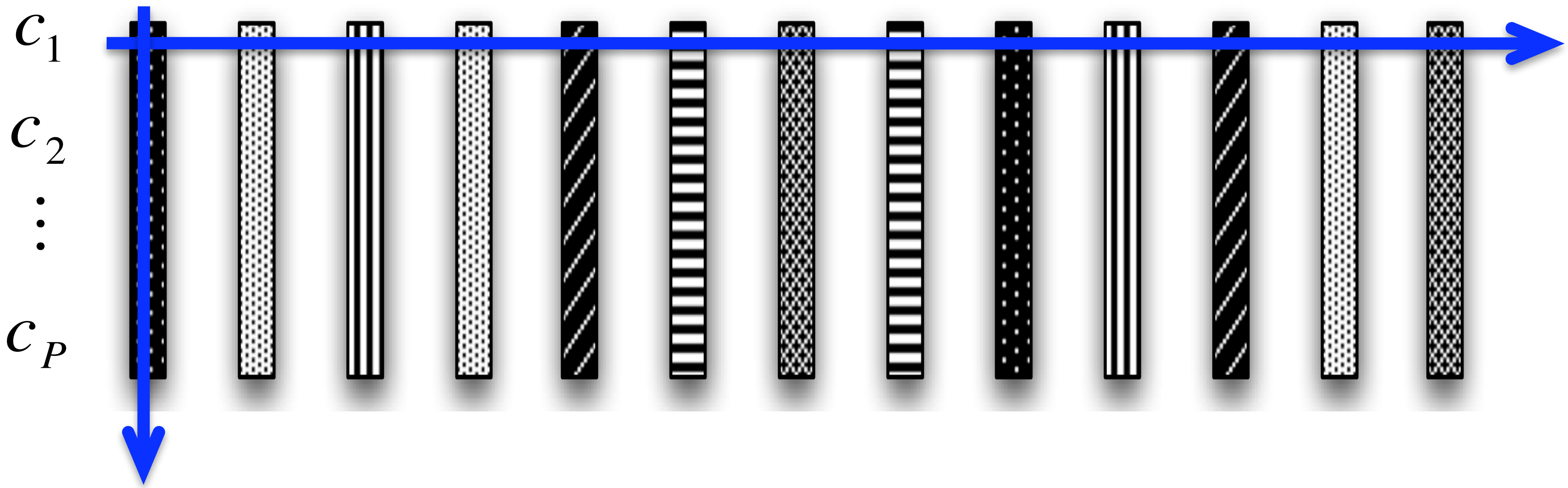
$$\Delta\Delta y = \frac{y(n) - 2y(n-h) + y(n-2h)}{h^2}$$

- Normalization is often necessary/beneficial:

$$\hat{y} = \frac{y - \min(y)}{\max(y - \min(y))} \qquad \hat{y} = \frac{y - \mu_y}{\sigma_y}$$

Post-processing

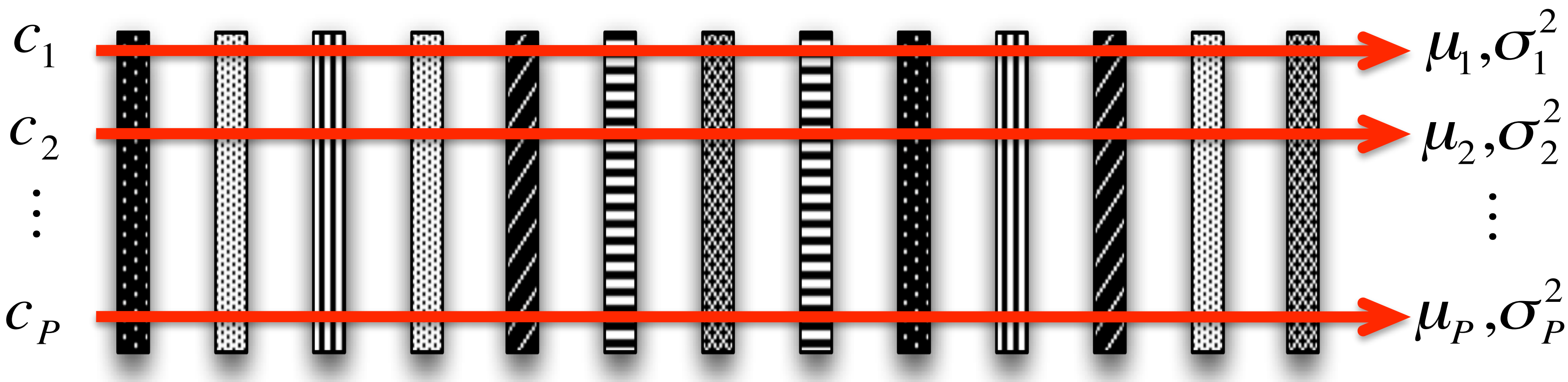
- Normalizing features across time avoids bias towards high-range features



- Normalizing feature vectors make them more comparable to each other
- Looses dynamic change information

Summarization

- Global (song/sound) features can be obtained by summarizing frame-level features:

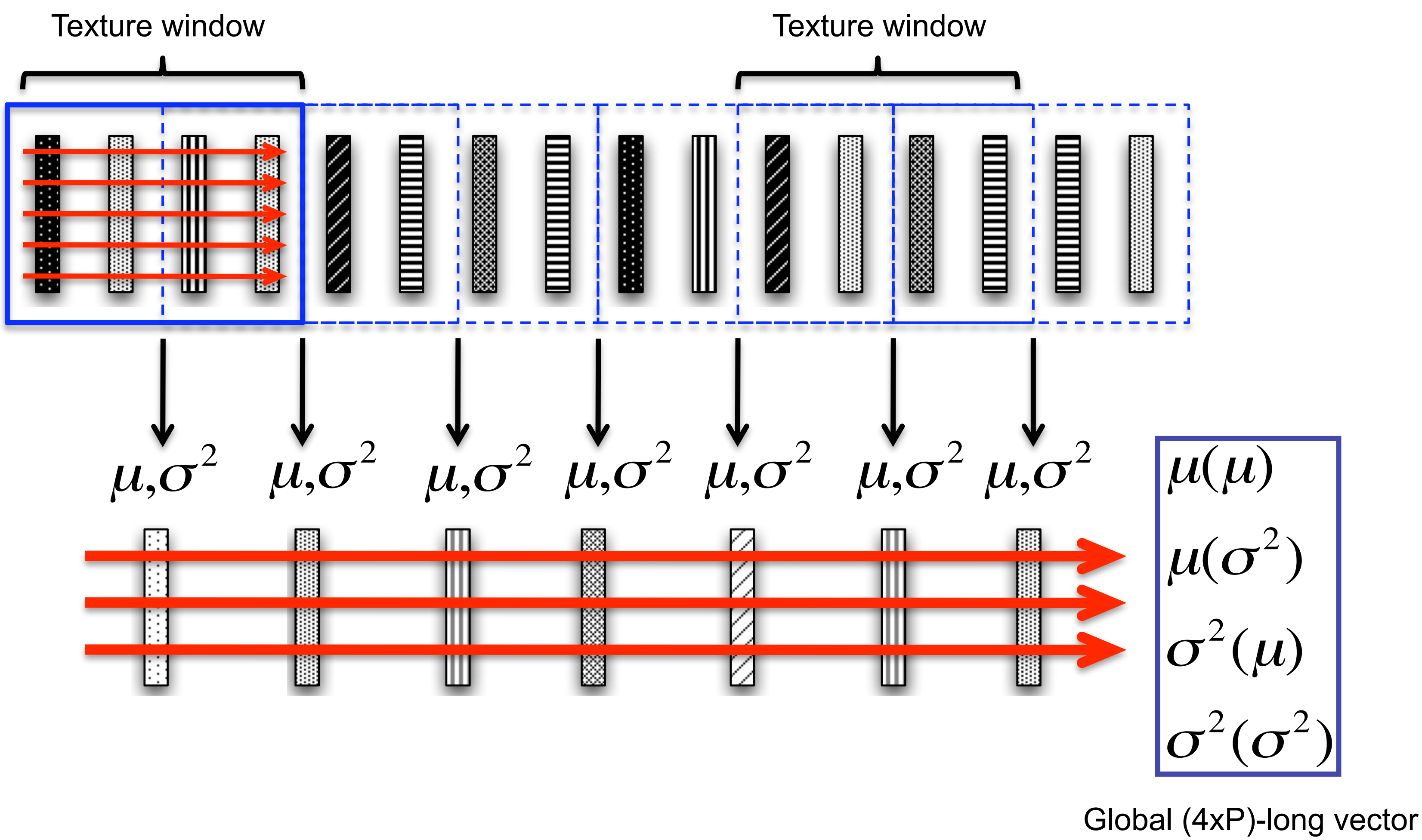


- Resulting on a single 2xP-long feature vector of means and variances.
- If not independent we measure the covariance:

$$\text{cov} = \sum_m (y - \mu_y)(y - \mu_y)^T / M$$

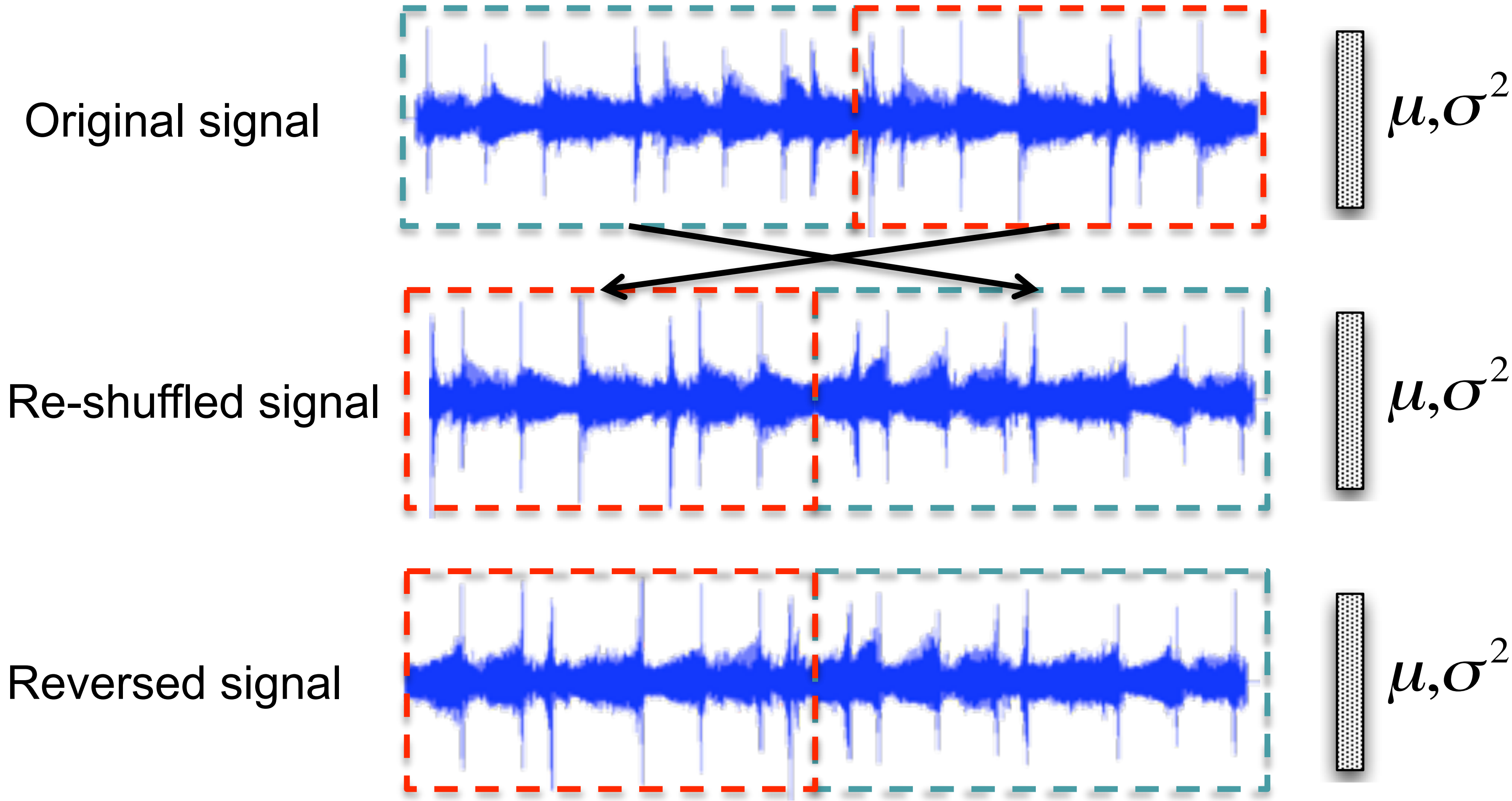
Summarization

- Texture windows can be used to capture local behavior:



Summarization

- Computing simple statistics across time ignores temporal ordering. Same global features for:



References

- Gold, B., Morgan, N. and Ellis, D. (Eds) “Speech and Audio Signal Processing”. Wiley (2011): chapters 19-22.
- Grey, J. “An Exploration of Musical Timbre”. CCRMA, Stanford University, Report # STAN-M-2.
- Klapuri, A. and Davy, M. (Eds) “Signal Processing Methods for Music Transcription”. Springer (2006): chapter 6, Herrera, P., Klapuri, A. and Davy, M. “Automatic Classification of Pitched Instrument Sounds”.
- Zölzer, U. (Ed). “DAFX: Digital Audio Effects”. John Wiley and Sons (2002): chapter 8, Arfib, D., Keiler, F. and Zölzer, U., “Source-filter Processing”.
- Pampalk, E. “Computational Models of Music Similarity and their Application in Music Information Retrieval”. PhD Thesis, Vienna University of Technology, Vienna, Austria (2006). PDF available at: http://staff.aist.go.jp/elias.pampalk/mir-phds/
- Logan, B. “Mel Frequency Cepstral Coefficients for Music Modeling”, Proceedings of the ISMIR International Symposium on Music Information Retrieval, Plymouth, MA (2000).

References

- Peeters, G. “A large set of audio features for sound description (similarity and classification) in the CUIDADO project”. CUIDADO I.S.T. Project Report (2004)
- Smith, J.O. “Mathematics of the Discrete Fourier Transform (DFT)”. 2nd Edition, W3K Publishing (2007): Appendix A.6.1, “The Discrete Cosine Transform”.
- McKinney, M. and Breebaart, J. “Features for audio and music classification”. Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 03), Baltimore, Maryland, USA, October 27-30, 2003.
- Vincent, E. “Instrument Recognition”. Lecture notes ELEM-035: Music Analysis and Synthesis, Queen Mary University of London (2006)
- Kim, H-G., Moreau, N. and Sikora, T. “MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval”. John Wiley & Sons (2005): chapter 2: “Low-Level Descriptors”
- Cook, P. (Ed) “Music, Cognition and Computerized Sound”, The MIT Press (2001): chapter 7, Mathews, M. “Introduction to Timbre”.