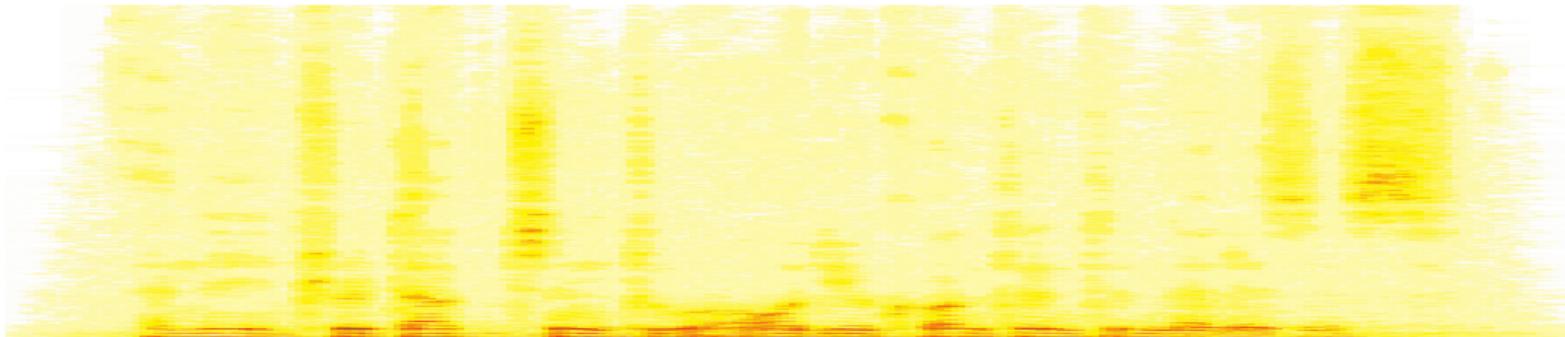


# Introduction to Audio Content Analysis

## Module 3.2: Feature Extraction — Spectral Shape Features

alexander lerch



# introduction

## overview

corresponding textbook section

**Chapter 3 — Instantaneous Features:** pp. 41–53

### ● lecture content

- timbre
- spectral shape instantaneous features

### ● learning objectives

- describe the general impact of spectral shape on timbre perception
- summarize features, describe their computation, and discuss their meaning



# introduction

overview

corresponding textbook section

[Chapter 3 — Instantaneous Features: pp. 41–53](#)

## ● lecture content

- timbre
- spectral shape instantaneous features

## ● learning objectives

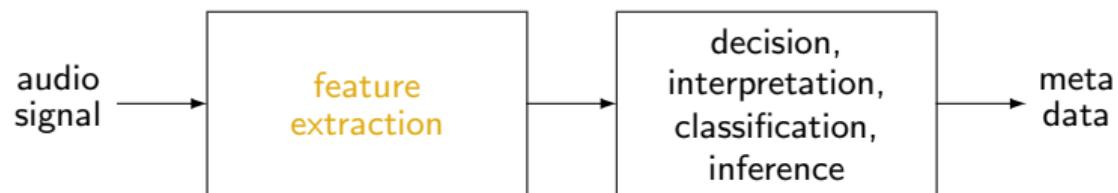
- describe the general impact of spectral shape on timbre perception
- summarize features, describe their computation, and discuss their meaning



# introduction

## Audio Content Analysis flowchart

remember the flow chart of a general ACA system:



# introduction

## timbre 1/2

### definition (American Standards Association)

...that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar

What is the problem with this definition?



---

<sup>1</sup>A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1994.

<sup>2</sup>S. McAdams and A. Bregman, "Hearing Musical Streams," *Computer Music Journal*, vol. 3, no. 4, pp. 26–60, Dec. 1979, ISSN: 0148-9267.

# introduction

## timbre 1/2

### definition (American Standards Association)

...that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar

### What is the problem with this definition?

Bregman:<sup>1</sup>

- ① implies that timbre *only* exists for sounds with pitch!
- ② only says that timbre *is not* loudness and pitch



---

<sup>1</sup>A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1994.

<sup>2</sup>S. McAdams and A. Bregman, "Hearing Musical Streams," *Computer Music Journal*, vol. 3, no. 4, pp. 26–60, Dec. 1979, ISSN: 0148-9267.

# introduction

## timbre 1/2

### definition (American Standards Association)

...that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar

### What is the problem with this definition?

Bregman:<sup>1</sup>

- ① implies that timbre *only* exists for sounds with pitch!
  - ② only says that timbre *is not* loudness and pitch
- [timbre is] "...the psychoacoustician's multidimensional waste-basket category for everything that cannot be labeled pitch or loudness."<sup>2</sup>

<sup>1</sup>A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1994.

<sup>2</sup>S. McAdams and A. Bregman, "Hearing Musical Streams," *Computer Music Journal*, vol. 3, no. 4, pp. 26–60, Dec. 1979, ISSN: 0148-9267.

# introduction

## timbre 2/2

timbre is

- a function of **temporal envelope**
  - attack time characteristics
  - amplitude modulations
  - ...
- a function of **spectral distribution**
  - spectral envelope
  - number of partials
  - energy distribution of partials
  - ...

when dealing with complex mixtures of sound, it is very difficult (maybe impossible?) to extract detailed temporal information for individual tones

⇒ timbre features typically focus on the **spectral shape**

# introduction

## timbre 2/2

timbre is

- a function of **temporal envelope**
  - attack time characteristics
  - amplitude modulations
  - ...
- a function of **spectral distribution**
  - spectral envelope
  - number of partials
  - energy distribution of partials
  - ...

when dealing with complex mixtures of sound, it is very difficult (maybe impossible?) to extract detailed temporal information for individual tones

⇒ timbre features typically focus on the **spectral shape**

# introduction

## timbre 2/2

timbre is

- a function of **temporal envelope**
  - attack time characteristics
  - amplitude modulations
  - ...
- a function of **spectral distribution**
  - spectral envelope
  - number of partials
  - energy distribution of partials
  - ...

when dealing with complex mixtures of sound, it is very difficult (maybe impossible?) to extract detailed temporal information for individual tones

⇒ timbre features typically focus on the **spectral shape**

# introduction

## timbre 2/2

timbre is

- a function of **temporal envelope**
  - attack time characteristics
  - amplitude modulations
  - ...
- a function of **spectral distribution**
  - spectral envelope
  - number of partials
  - energy distribution of partials
  - ...

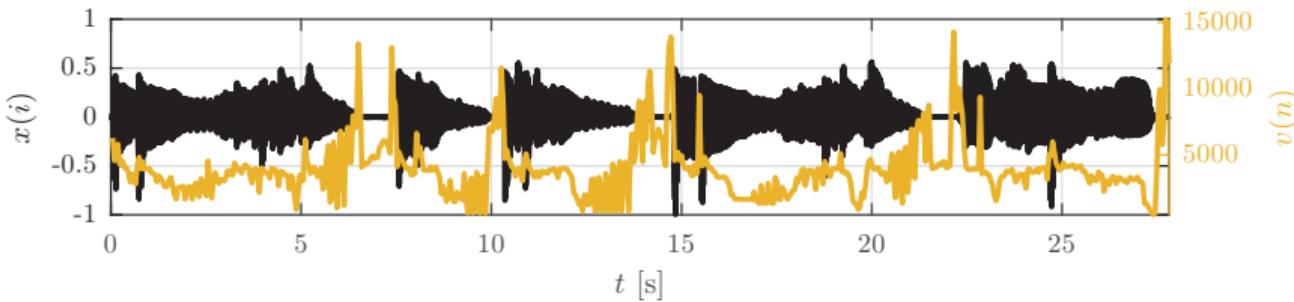
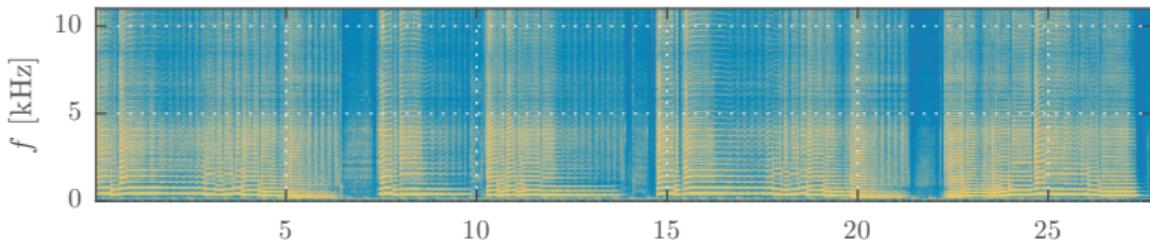
when dealing with complex mixtures of sound, it is very difficult (maybe impossible?) to extract detailed temporal information for individual tones

⇒ timbre features typically focus on the **spectral shape**

# spectral shape features

## spectral rolloff

$$v_{SR}(n) = i \quad \text{at} \sum_{k=0}^i |X(k, n)| = \kappa \cdot \sum_{k=0}^{\kappa/2-1} |X(k, n)|$$



# spectral shape features

## spectral rolloff

$$v_{SR}(n) = i \quad \text{at} \sum_{k=0}^i |X(k, n)| = \kappa \cdot \sum_{k=0}^{\kappa/2-1} |X(k, n)|$$

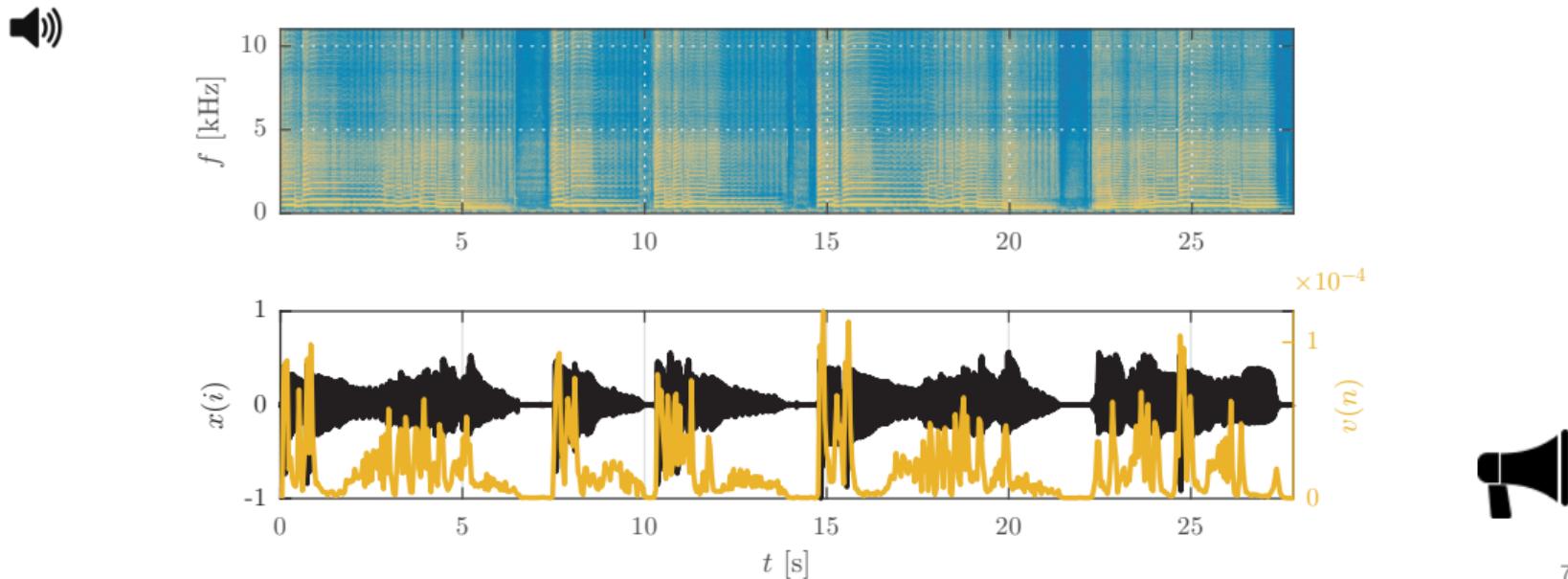
### common variants:

- scaled to frequency
- power spectrum

## spectral shape features

**Georgia Tech**  **Center for Music Technology**  
College of Design

$$v_{\text{SF}}(n) = \frac{\sqrt{\sum_{k=0}^{\kappa/2-1} (|X(k, n)| - |X(k, n-1)|)^2}}{\kappa/2}$$



# spectral shape features

## spectral flux

$$v_{SF}(n) = \frac{\sqrt{\sum_{k=0}^{\mathcal{K}/2-1} (|X(k, n)| - |X(k, n-1)|)^2}}{\mathcal{K}/2}$$

common variants:

$$v_{SF}(n, \beta) = \frac{\sqrt[\beta]{\sum_{k=0}^{\mathcal{K}/2-1} (|X(k, n)| - |X(k, n-1)|)^\beta}}{\mathcal{K}/2}$$

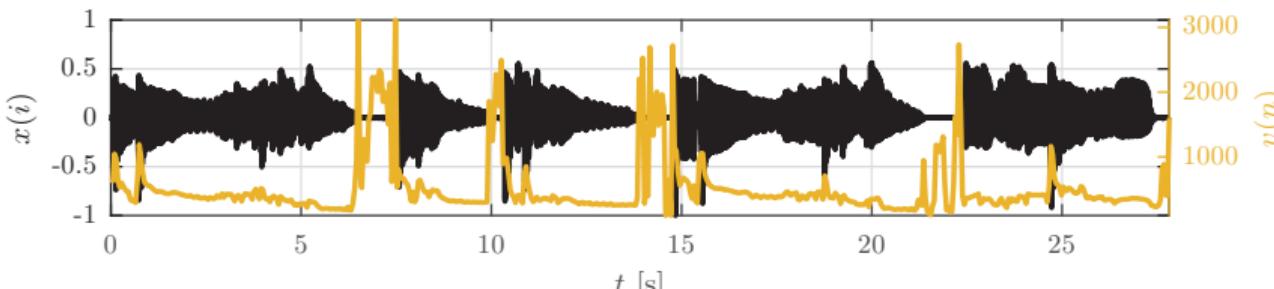
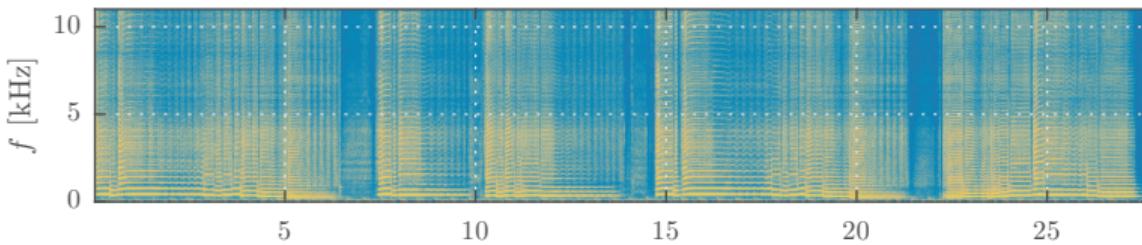
$$v_{SF,\sigma}(n) = \sqrt{\frac{2}{\mathcal{K}} \sum_{k=0}^{\mathcal{K}/2-1} (\Delta X(k, n) - \mu_{\Delta X})^2}$$

$$v_{SF,\log}(n) = \frac{2}{\mathcal{K}} \sum_{k=0}^{\mathcal{K}/2-1} \log_2 \left( \frac{|X(k, n)|}{|X(k, n-1)|} \right)$$

# spectral shape features

## spectral centroid

$$v_{SC}(n) = \frac{\sum_{k=0}^{\mathcal{K}/2-1} k \cdot |X(k, n)|}{\sum_{k=0}^{\mathcal{K}/2-1} |X(k, n)|}$$



# spectral shape features

## spectral centroid

$$v_{SC}(n) = \frac{\sum_{k=0}^{\mathcal{K}/2-1} k \cdot |X(k, n)|}{\sum_{k=0}^{\mathcal{K}/2-1} |X(k, n)|}$$

### common variants:

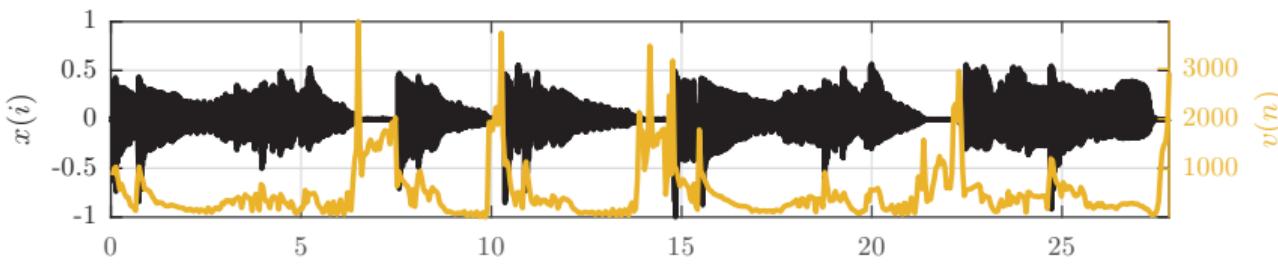
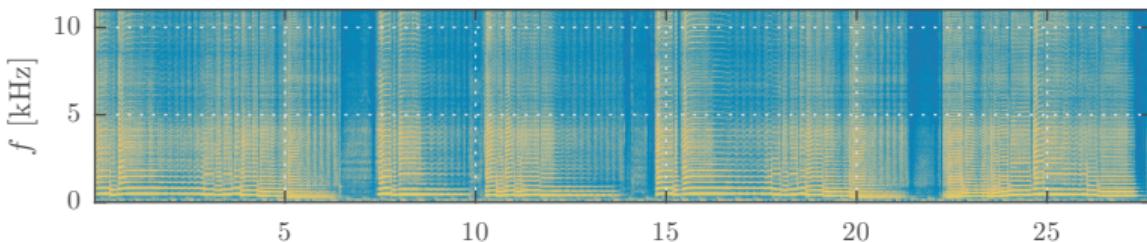
- power spectrum
- logarithmic frequency scale

$$v_{SC,\log}(n) = \frac{\sum_{k=k(f_{\min})}^{\mathcal{K}/2-1} \log_2 \left( \frac{f(k)}{f_{\text{ref}}} \right) \cdot |X(k, n)|^2}{\sum_{k=k(f_{\min})}^{N/2-1} |X(k, n)|^2}$$

# spectral shape features

## spectral spread

$$v_{SS}(n) = \sqrt{\frac{\sum_{k=0}^{\kappa/2-1} (k - v_{SC}(n))^2 \cdot |X(k, n)|^2}{\sum_{k=0}^{\kappa/2-1} |X(k, n)|^2}}$$



# spectral shape features

## spectral spread

$$v_{SS}(n) = \sqrt{\frac{\sum_{k=0}^{\mathcal{K}/2-1} (k - v_{SC}(n))^2 \cdot |X(k, n)|^2}{\sum_{k=0}^{\mathcal{K}/2-1} |X(k, n)|^2}}$$

### common variants:

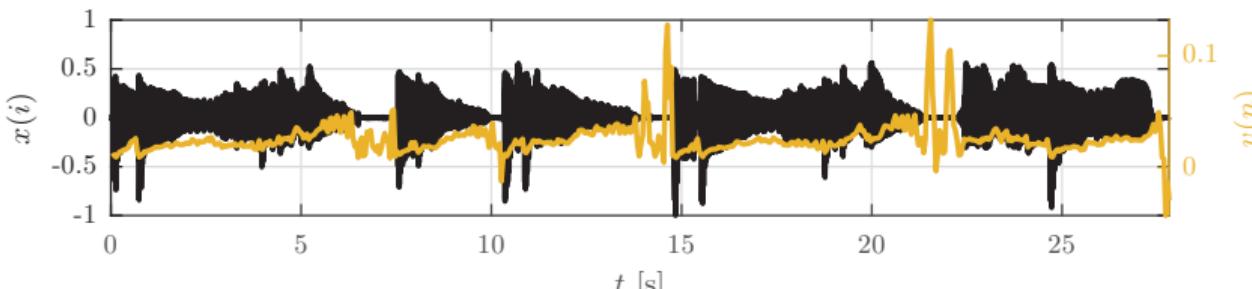
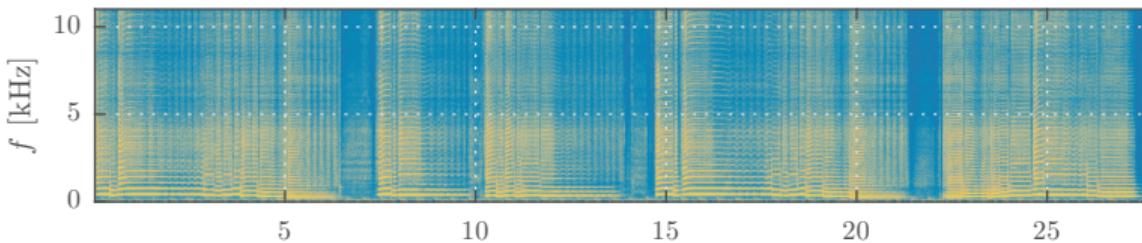
- same variants as with *Spectral Centroid*, e.g. logarithmic:

$$v_{SS,\log}(n) = \sqrt{\frac{\sum_{k=k(f_{min})}^{\mathcal{K}/2-1} \left( \log_2 \left( \frac{f(k)}{1000 \text{ Hz}} \right) - v_{SC}(n) \right)^2 \cdot |X(k, n)|^2}{\sum_{k=k(f_{min})}^{\mathcal{K}/2-1} |X(k, n)|^2}}$$

# spectral shape features

## spectral decrease

$$v_{SD}(n) = \frac{\sum_{k=1}^{\kappa/2-1} \frac{1}{k} \cdot (|X(k, n)| - |X(0, n)|)}{\sum_{k=1}^{\kappa/2-1} |X(k, n)|}$$



# spectral shape features

## spectral decrease

$$v_{SD}(n) = \frac{\sum_{k=1}^{\kappa/2-1} \frac{1}{k} \cdot (|X(k, n)| - |X(0, n)|)}{\sum_{k=1}^{\kappa/2-1} |X(k, n)|}$$

### common variants:

- restricted frequency range:

$$v_{SD}(n) = \frac{\sum_{k=k_l}^{k_u} \frac{1}{k} \cdot (|X(k, n)| - |X(k_l - 1, n)|)}{\sum_{k=k_l}^{k_u} |X(k, n)|}$$

# fundamentals

## cepstrum 1/3

### signal model:

convolution of *excitation signal* and *transfer function*

$$x(i) = e(i) * h(i)$$

$$X(j\omega) = E(j\omega) \cdot H(j\omega)$$

$$\begin{aligned}\log(X(j\omega)) &= \log(E(j\omega) \cdot H(j\omega)) \\ &= \log(E(j\omega)) + \log(H(j\omega))\end{aligned}$$

# fundamentals

## cepstrum 1/3

### signal model:

convolution of *excitation signal* and *transfer function*

$$x(i) = e(i) * h(i)$$

$$X(j\omega) = E(j\omega) \cdot H(j\omega)$$

$$\begin{aligned}\log(X(j\omega)) &= \log(E(j\omega) \cdot H(j\omega)) \\ &= \log(E(j\omega)) + \log(H(j\omega))\end{aligned}$$

# fundamentals

## cepstrum 1/3

### signal model:

convolution of *excitation signal* and *transfer function*

$$x(i) = e(i) * h(i)$$

$$X(j\omega) = E(j\omega) \cdot H(j\omega)$$

$$\begin{aligned}\log(X(j\omega)) &= \log(E(j\omega) \cdot H(j\omega)) \\ &= \log(E(j\omega)) + \log(H(j\omega))\end{aligned}$$

# fundamentals

## cepstrum 2/3

$$\begin{aligned}c_x(i) &= \mathfrak{F}^{-1}\{\log(X(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega)) + \log(H(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega))\} + \mathfrak{F}^{-1}\{\log(H(j\omega))\} \\ \hat{c}_x(i_s(n) \dots i_e(n)) &= \sum_{k=0}^{K/2-1} \log(|X(k, n)|) e^{\jmath k i \Delta \Omega}\end{aligned}$$

# fundamentals

## cepstrum 2/3

$$\begin{aligned}c_x(i) &= \mathfrak{F}^{-1}\{\log(X(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega)) + \log(H(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega))\} + \mathfrak{F}^{-1}\{\log(H(j\omega))\} \\ \hat{c}_x(i_s(n) \dots i_e(n)) &= \sum_{k=0}^{K/2-1} \log(|X(k, n)|) e^{\jmath k i \Delta \Omega}\end{aligned}$$

# fundamentals

## cepstrum 2/3

$$\begin{aligned}c_x(i) &= \mathfrak{F}^{-1}\{\log(X(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega)) + \log(H(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega))\} + \mathfrak{F}^{-1}\{\log(H(j\omega))\}\end{aligned}$$

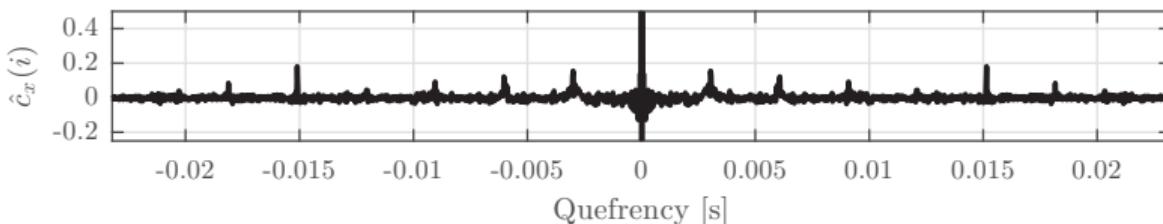
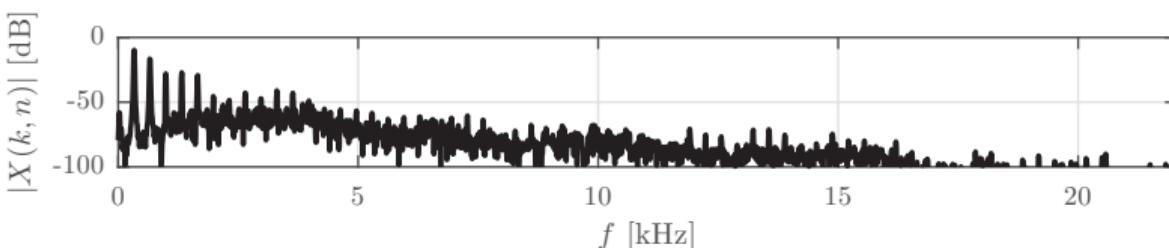
$$\hat{c}_x(i_s(n) \dots i_e(n)) = \sum_{k=0}^{\mathcal{K}/2-1} \log(|X(k, n)|) e^{jki\Delta\Omega}$$

# fundamentals

## cepstrum 2/3

$$\begin{aligned}c_x(i) &= \mathfrak{F}^{-1}\{\log(X(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega)) + \log(H(j\omega))\} \\&= \mathfrak{F}^{-1}\{\log(E(j\omega))\} + \mathfrak{F}^{-1}\{\log(H(j\omega))\}\end{aligned}$$

$$\hat{c}_x(i_s(n) \dots i_e(n)) = \sum_{k=0}^{\kappa/2-1} \log(|X(k, n)|) e^{jki\Delta\Omega}$$



# fundamentals

## cepstrum 3/3

- **summary:**

- cepstrum 'replaces' time domain convolution operation with addition
- result is the *unfiltered* excitation signal *plus* the filter IR (both logarithmic)
- can be used for, e.g., *spectral envelope extraction* or *pitch detection*
- more naming silliness:  
cepstrum, quefrency, liftering, ...

# fundamentals

## cepstrum 3/3

- **summary:**

- cepstrum 'replaces' time domain convolution operation with addition
- result is the *unfiltered* excitation signal *plus* the filter IR (both logarithmic)
- can be used for, e.g., *spectral envelope extraction* or *pitch detection*
- more naming silliness:  
cepstrum, quefrency, liftering, ...

# spectral shape features

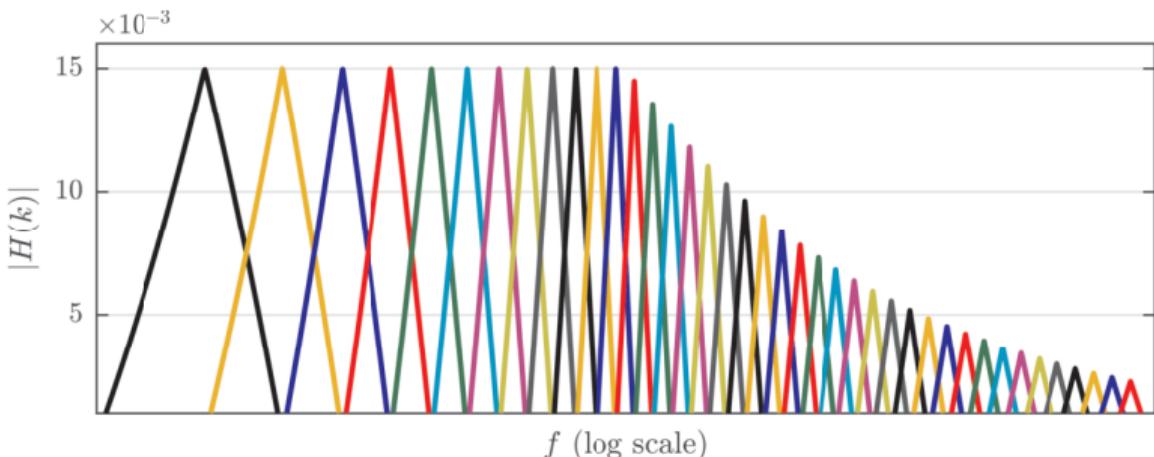
## mel frequency cepstral coefficients 1/4

- typical processing steps for the mel frequency cepstral coefficients (MFCCs):
  - compute magnitude spectrum
  - convert linear frequency scale to logarithmic
  - group bins into bands
  - apply logarithm to all bands
  - compute (inverse) cosine transform (DCT)

$$v_{\text{MFCC}}^j(n) = \sum_{k'=1}^{K'} \log(|X'(k', n)|) \cdot \cos\left(j \cdot \left(k' - \frac{1}{2}\right) \frac{\pi}{K'}\right)$$

# spectral shape features

## mel frequency cepstral coefficients 2/4

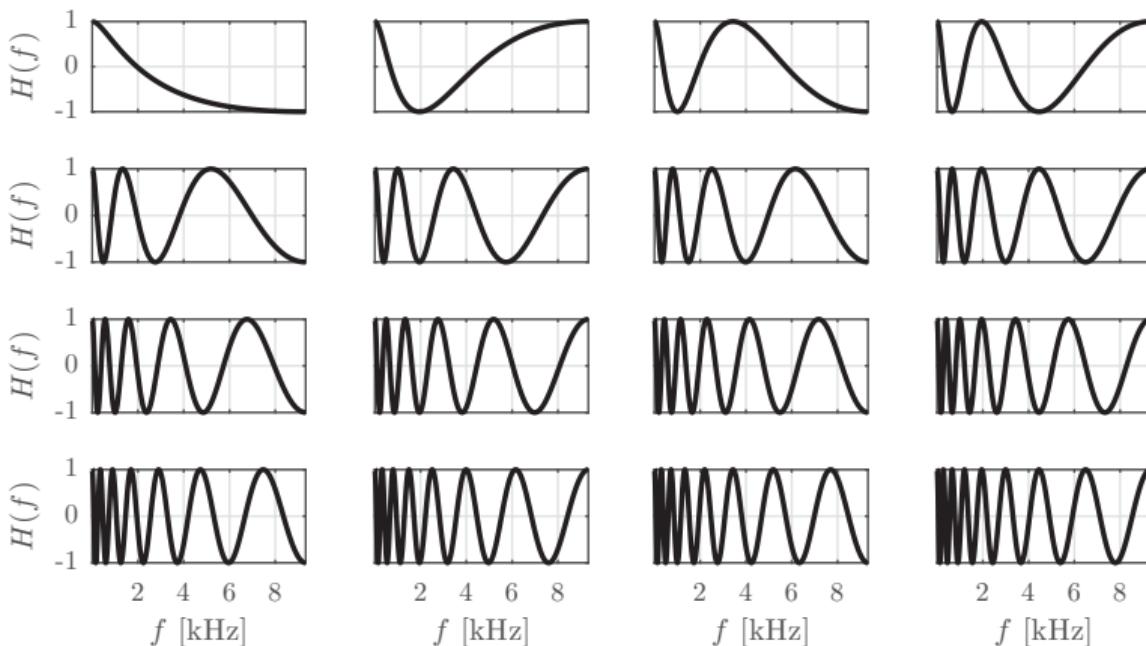


- constant Q filter spacing for higher frequencies (mel scale)
- FFT values are weighted and summed over bins for each band

# spectral shape features

## mel frequency cepstral coefficients 3/4

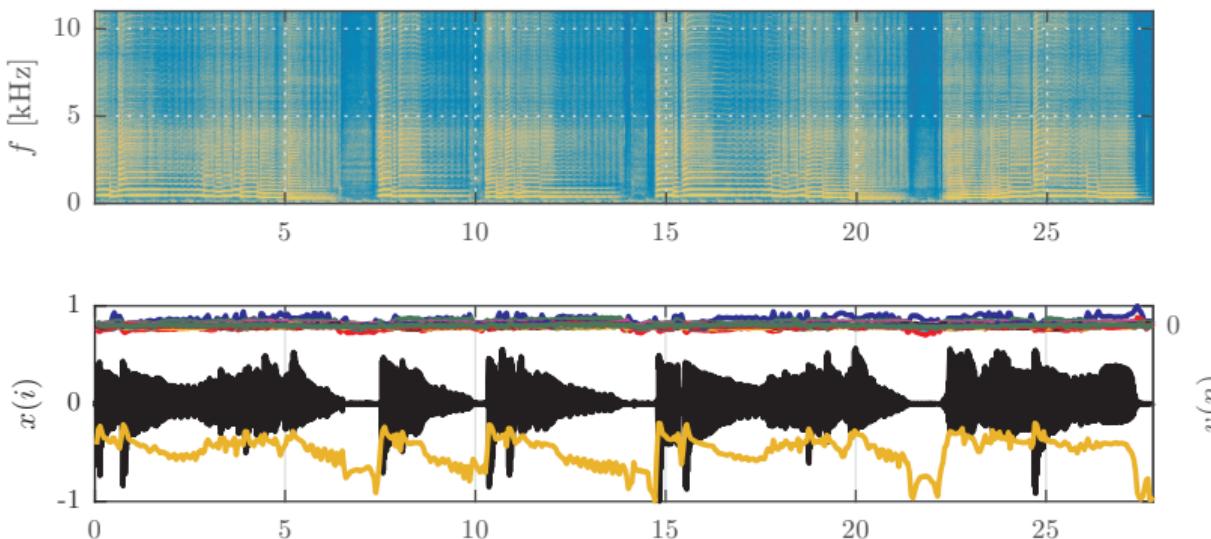
mel-warped cosine bases for DCT



# spectral shape features

## mel frequency cepstral coefficients 4/4

Property	DM	HTK	SAT
Num. filters	20	24	40
Mel scale	lin/log	log	lin/log
Freq. range	[100; 4000]	[100; 4000]	[200; 6400]
Normalization	Equal height	Equal height	Equal area



- **timbre**

- mostly dependent on both spectral shape and time domain envelope characteristics
- multi-dimensional perceptual property not as clearly defined as pitch or loudness

- **instantaneous spectral shape features**

- established set of baseline features
- usually extracted from the magnitude spectrum
- condensing various properties of the spectral shape into single values
- there exist multiple variants of “the same” feature

- **cepstrum**

- approach to “separate” excitation from filter signal
- makes use of the logarithm turning spectral multiplication into addition

