Mel-frequency Cepstral coefficients

EE3662: Digital Signal Processing Lab
Lab#9 -- Nov. 14, 2016

Prof. Yi-Wen Liu
Acoustics and Hearing Group, NTHU Dept. EE

Lab purpose

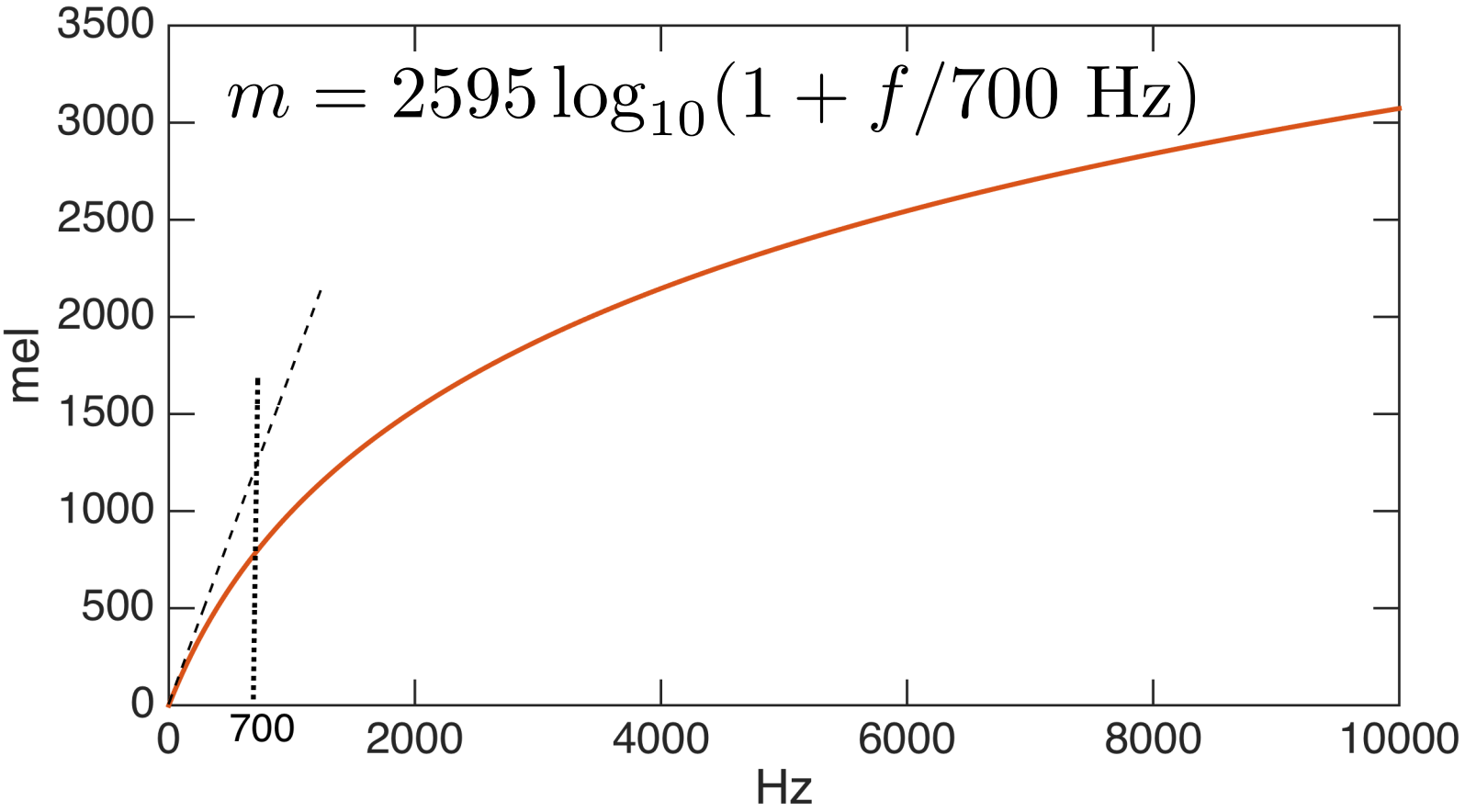
If two sounds have similar MFCCs, would they sound similar to our ears?

Activities

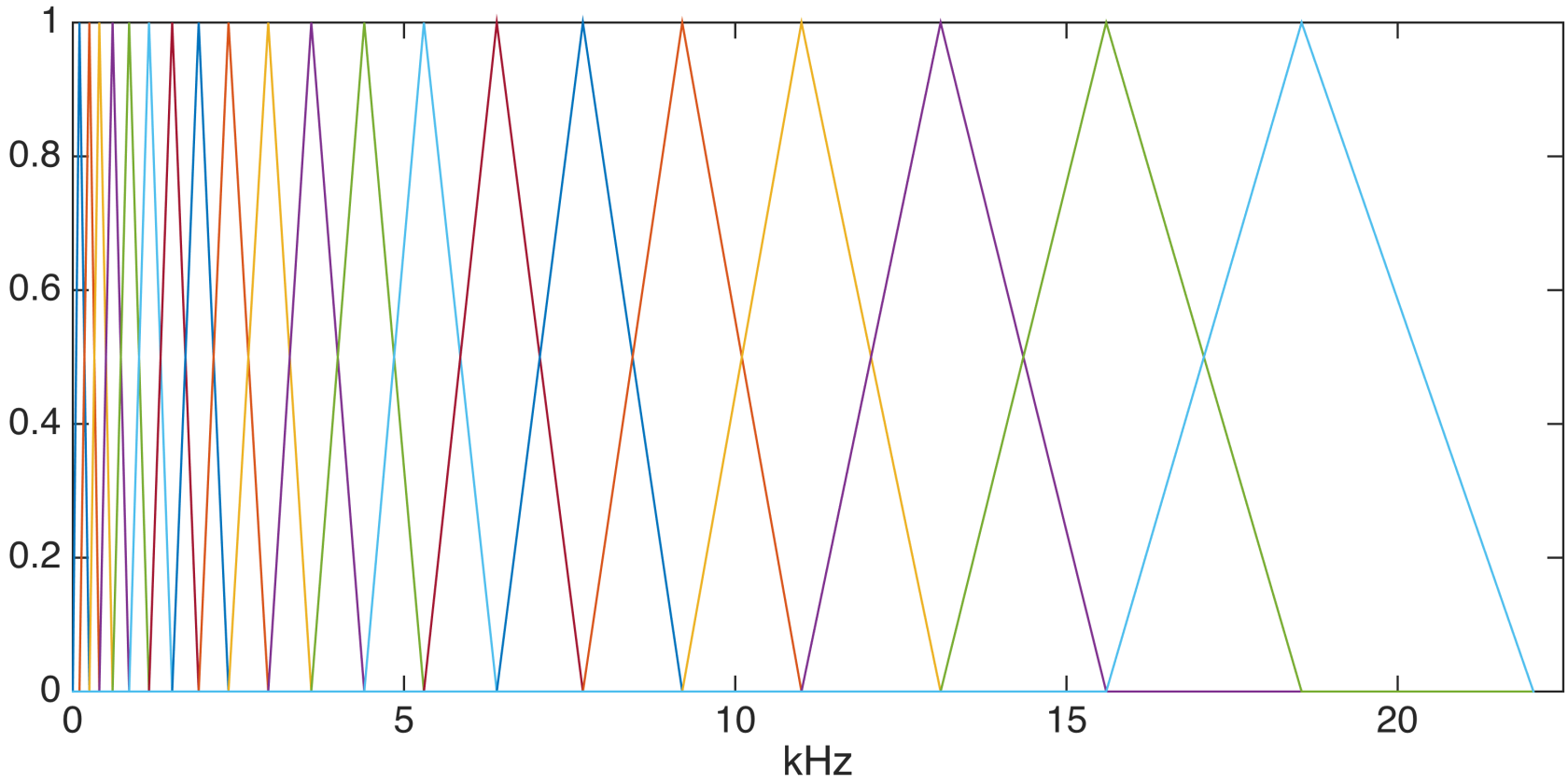
- MFCC
- Inverse MFCC?

Mel-to-Hz: nonlinear mapping

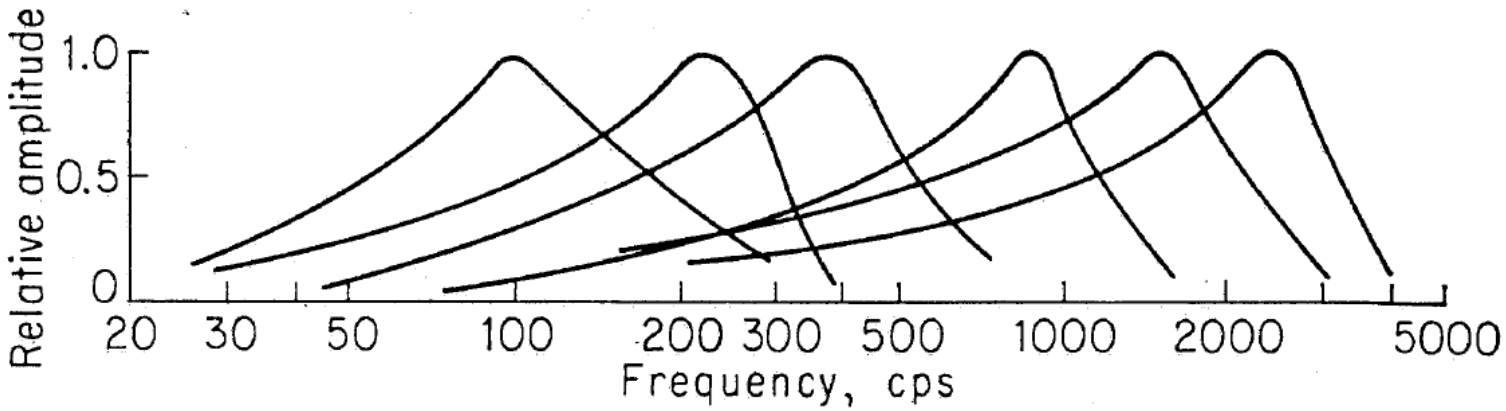
Hz	40	161	200	404	693	867	1000	2022	3000	3393	4109	5526	6500	7743	12000
mel	43	257	300	514	771	928	1000	1542	2000	2142	2314	2600	2771	2914	3228



Mel-scale triangular filters (n=20)

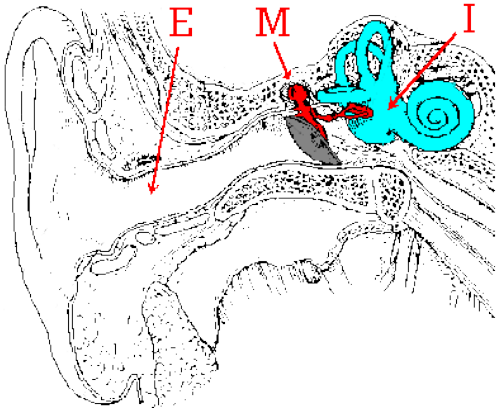


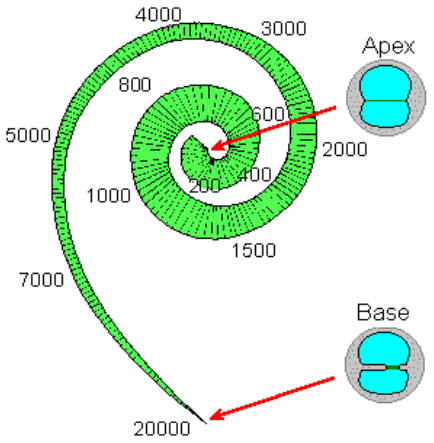
Response to pure tones: approximately constant-Q

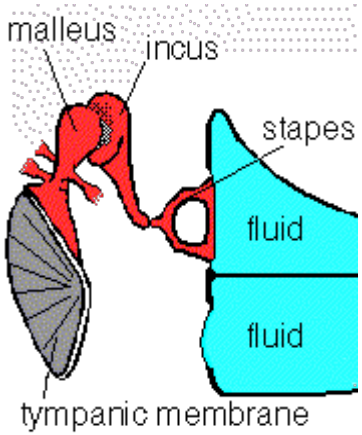


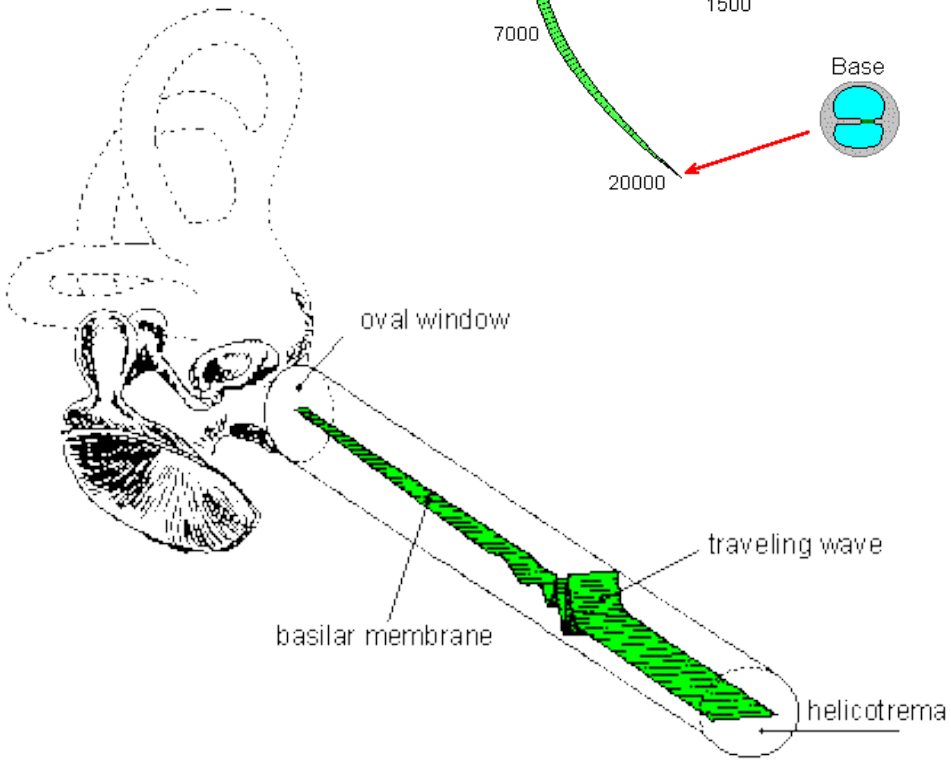
von Békésy (1943)

Cochlea: frequency to place mapping is roughly log-linear

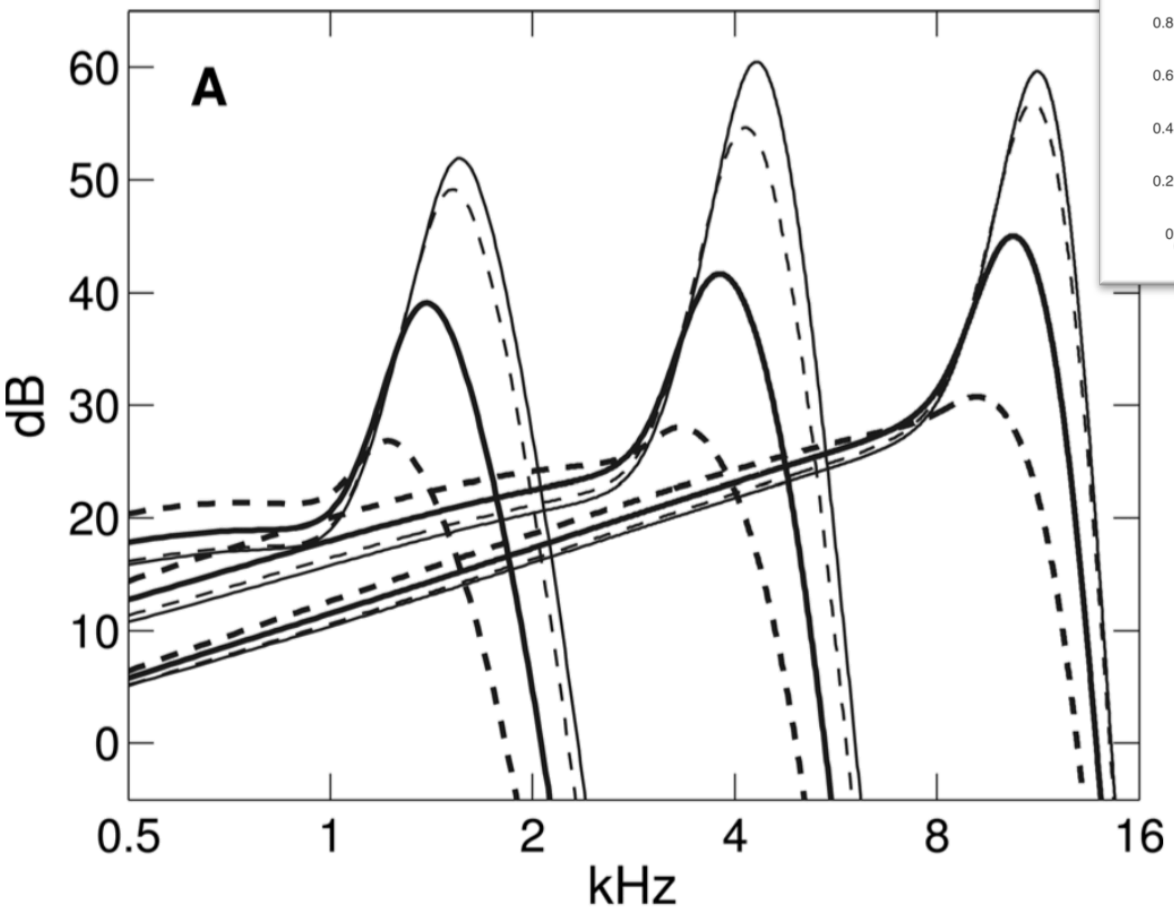


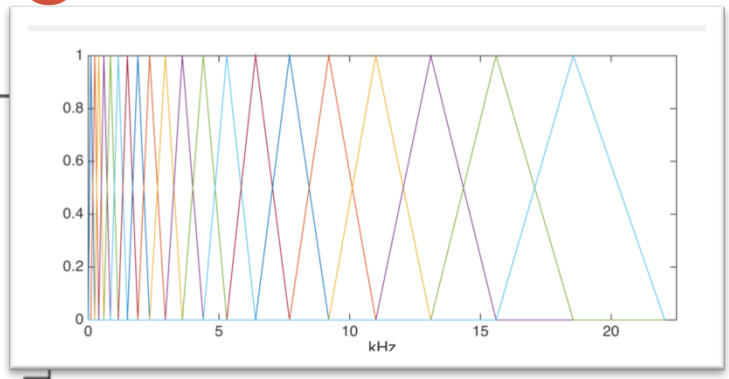






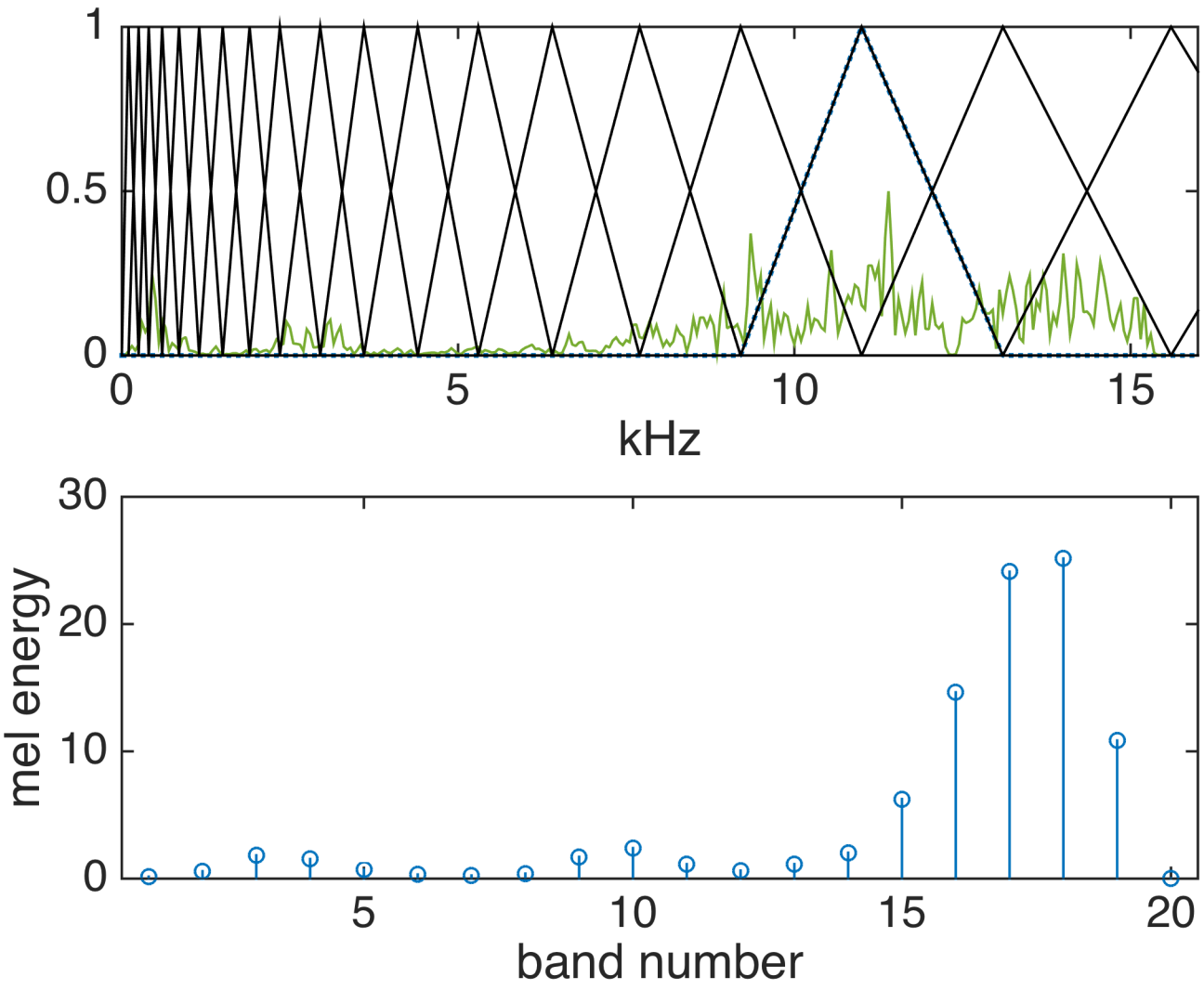
Bio-mimetic filter spacing



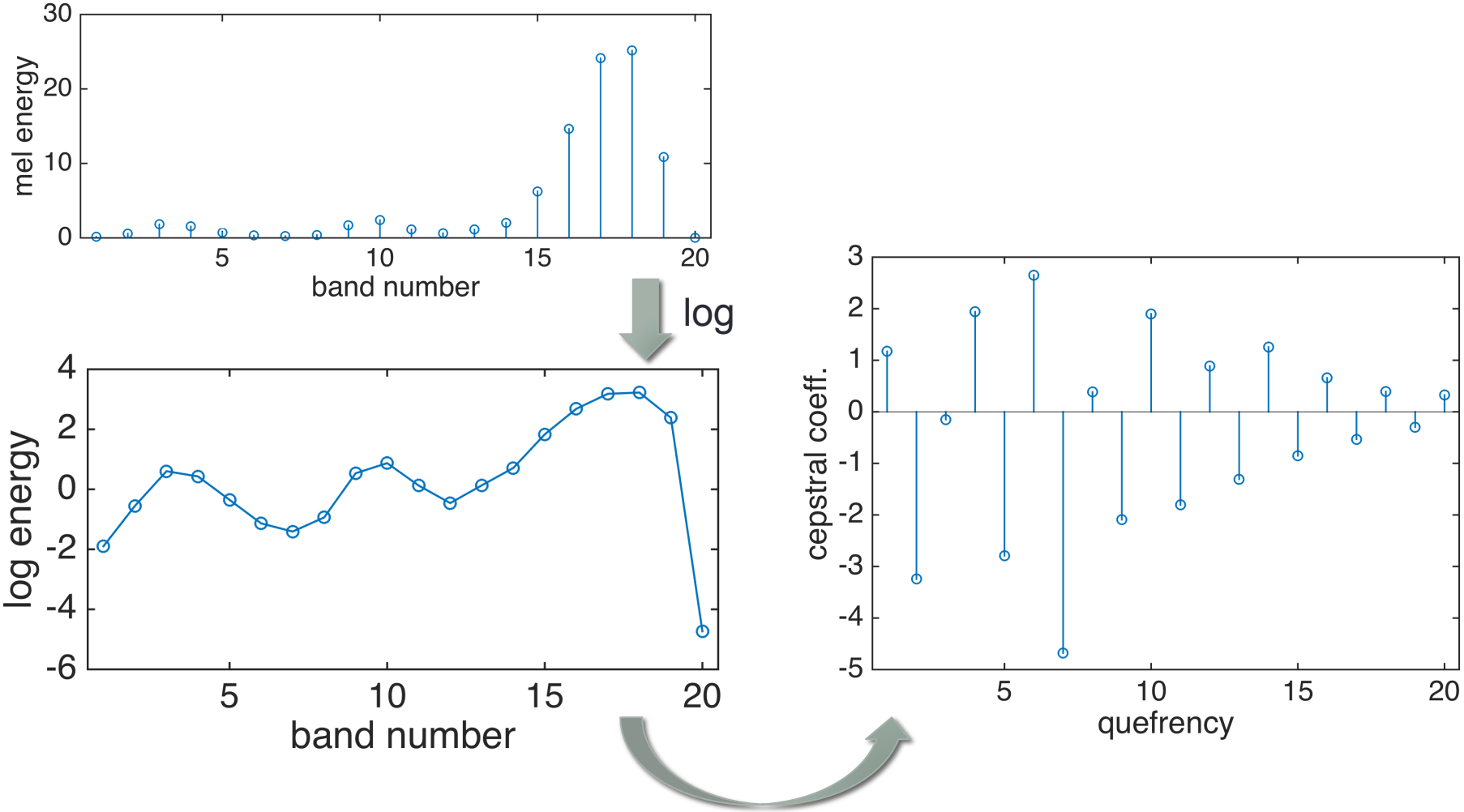


Nonlinear cochlear frequency responses at three different places: a computer modeling study (Liu, 2014).

Calculating the mel-“energy”



MFCC is DCT(log(mel_energies))



Definition of DCT

$$y(m) = C \sum_{l=1}^{L} x[l] \cos\left(\frac{m\pi}{L}\left(l - \frac{1}{2}\right)\right), m = 0, 1, 2, \ldots L - 1$$

C: normalization factor such that DCT is *unitary*.
m: quefrency
x(*l*): log of mel energies
y(*m*): cepstral coefficients

MFCC is a dimension reduction procedure

Signal =>

Windowing

FFT

Take absolute value

Inner product with triangular filters

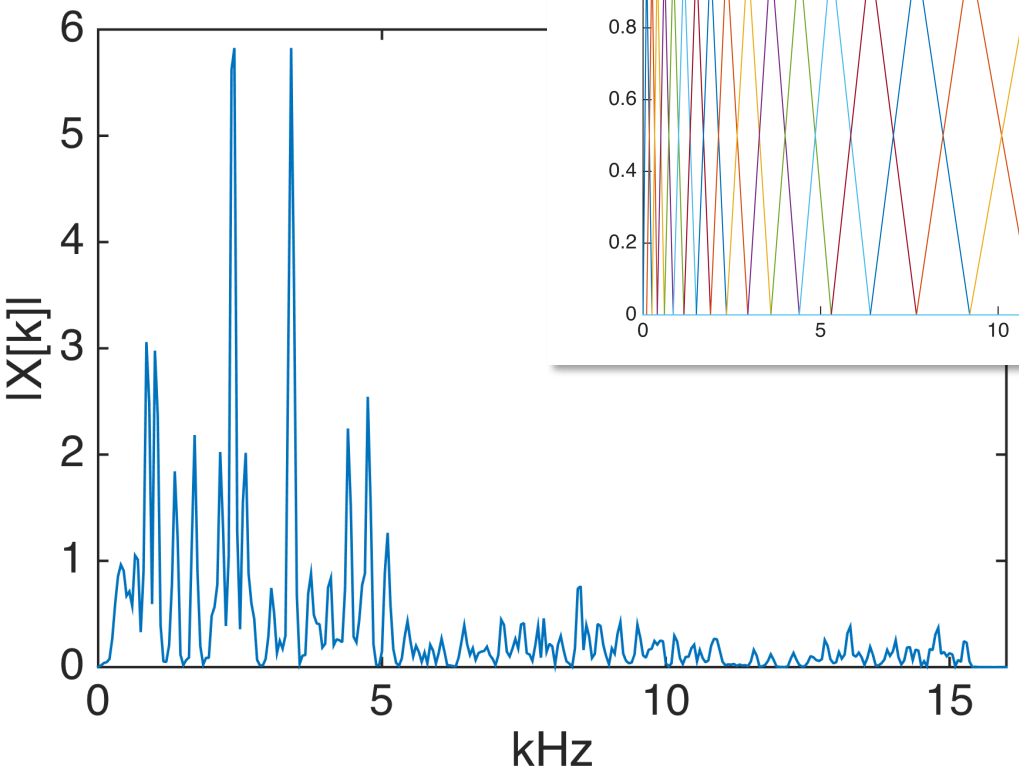
Take log

Apply DCT

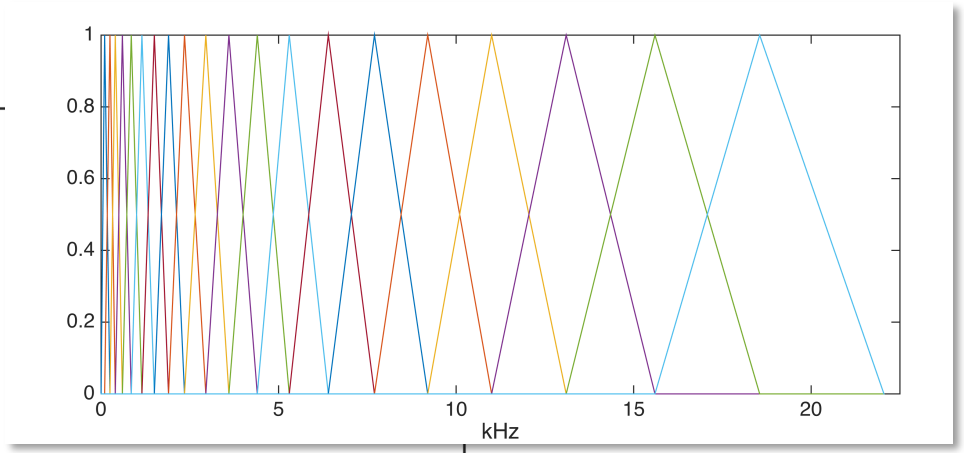
=> “feature vector”

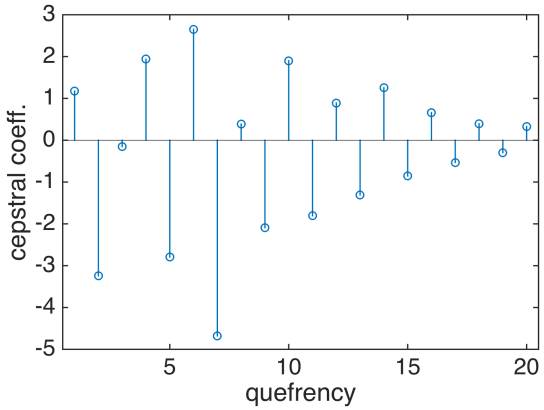
Can MFCC be inverted?

$$\mathbb{R}^{N_{\mathrm{FFT}}} \xrightarrow{\mathrm{MFCC}} \mathbb{R}^{n} \xrightarrow{?} \mathbb{R}^{N_{\mathrm{FFT}}}$$



a typical short-time spectrum of speech





This lab: Pseudo inverse of MFCC

Signal =>
Windowing
FFT
Take absolute value
triangular filters
Take log
Apply DCT

=> MFCC

MFCC =>
Inverse DCT
Exponential
Least square synthesis
random phase assignment
IFFT
Overlap and add
=> Signal

Spectral synthesis from mel-energies

$$W \in \mathsf{M}_{L\times M}(\mathbb{R}), y \in \mathbb{R}^L$$

$$\text{Find } x \in \mathbb{R}^M, \text{ s.t. } ||x||^2 \text{ is minimized}$$

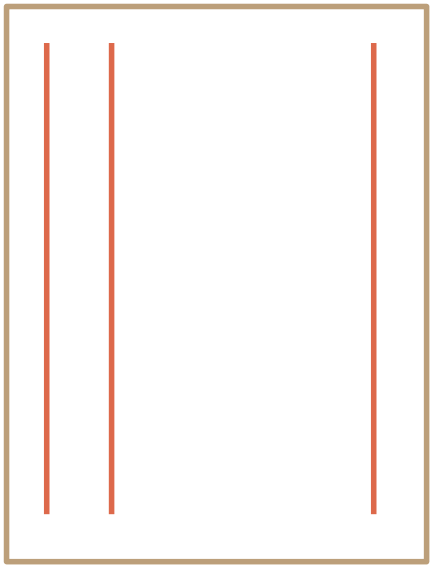
$$\text{subject to } Wx = y.$$

L: # of energy bands
M: length of spectrum from DC to Nyquist rate
W: representation of triangular filtering
y: mel energy vector

Solution: $$\hat{x} = W^t \left(WW^t\right)^{-1} y$$

Pseudo-inverse: the interpolation interpretation

$$\hat{x} = W^t \left(WW^t\right)^{-1} y$$



W^t

$\left(WW^t\right)^{-1} y$

= combination coeff.

Demo

Signal =>
Windowing
FFT
Take absolute value
triangular filters
Take log
Apply DCT

=> MFCC

MFCC =>
Inverse DCT
Exponential
Least square synthesis
random phase assignment
IFFT
Overlap and add
=> Signal

Questions for you to ponder

- Why not using rectangular filters for “energy” calculation?
- Why should mel-filters be overlapping?
- Why are high-*quefrency* MFCCs usually abandoned when doing speech recognition?
- Do you think MFCC is good for speaker identification purposes?
- If two sounds have similar MFCCs, does that imply they sound similar to our ears?
- Is $(WW^t)^{-1}y$ guaranteed to be non-negative given that all entries in y are non-negative?