ECE 417, Lecture 10: Speech Perception

Mark Hasegawa-Johnson

10/3/2017

Content

- Parseval's Theorem: Cepstral Distance = Spectral Distance
- What spectrum do people hear? The basilar membrane
- Frequency scales for hearing: mel, ERB
- Filterbank coefficients and MFCC

Parseval’s Theorem

L2 norm of a signal equals the L2 norm of its Fourier transform.

Parseval's Theorem: Examples

- Fourier Series:

$$\frac{1}{T}\int_0^T |x(t)|^2 dt = \sum_{k=-\infty}^{\infty} |X_k|^2$$

- DTFT:

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(\omega)|^2 d\omega$$

- DFT:

$$\sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |X[k]|^2$$

Parseval's Theorem: DCT

$$\frac{1}{M}\left(c[0]^2 + 2\sum_{n=1}^{M-1} c[n]^2\right) = \sum_{k=0}^{M-1} {C_k}^2$$

Where you remember that

$$C_k = \ln\left|S\left(\frac{(k+0.5)F_s}{N}\right)\right|$$

Parseval's Theorem: Vector Formulation

Suppose we define the vectors $\vec{c}$ and $\vec{C}$ as the cepstrum and the log spectrum, thus

$$\vec{c} = \begin{bmatrix} c_0 \\ \cdots \\ c_{M-1} \end{bmatrix}, \vec{C} = \begin{bmatrix} C_0 \\ \cdots \\ C_{M-1} \end{bmatrix}$$

Where for convenience we'll say

$$c_n = \begin{cases} \dfrac{c[0]}{\sqrt{M}} & n = 0 \\ \dfrac{c[n]}{\sqrt{M/2}} & 1 \le n \le M-1 \end{cases}$$

Parseval's Theorem: Vector Formulation

That way Parseval's theorem can be written very simply as

$$\sum_{n=0}^{M-1} {c_n}^2 = \sum_{k=0}^{M-1} {C_k}^2$$

…or even more simply as…

$$\|\vec{c}\|^2 = \left\|\vec{C}\right\|^2$$

i.e., the L2 norm of the cepstrum equals the L2 norm of the log spectrum.

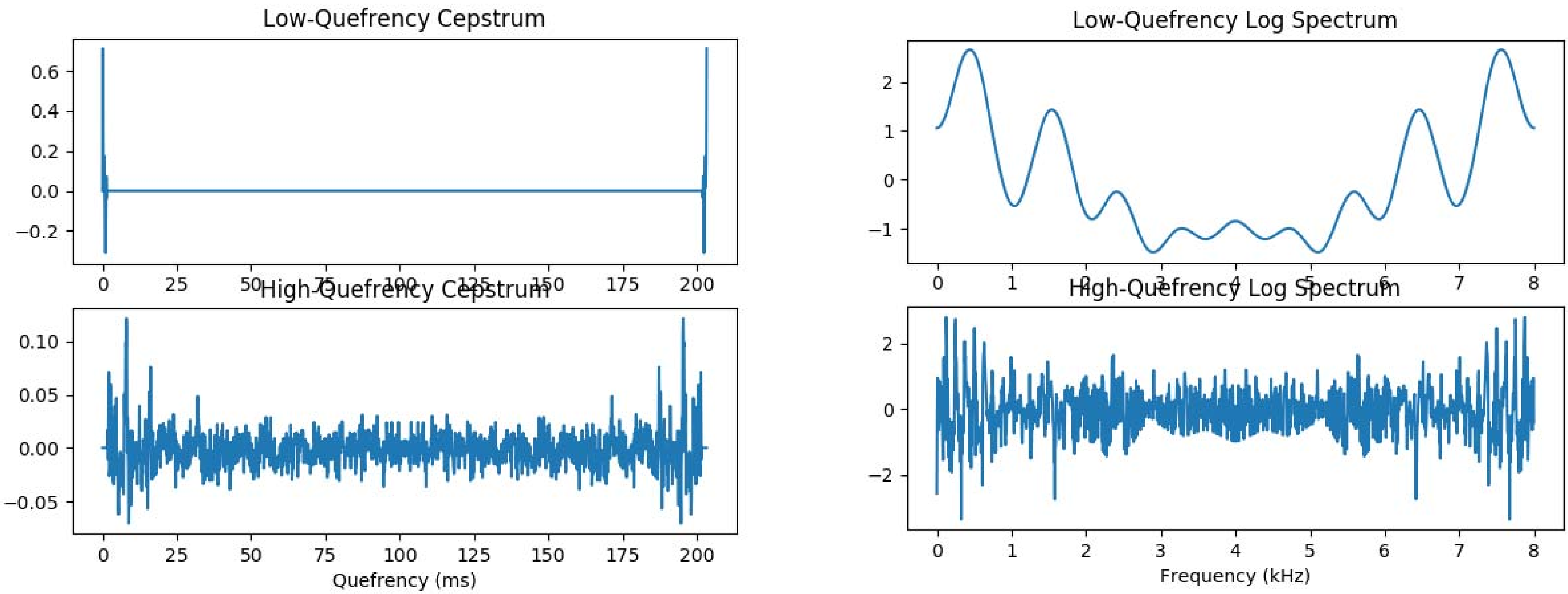
What it means for KNN

Suppose we have two acoustic signals $x(t)$ and $y(t)$, and we want to find out how different they sound. If they have static spectra, then a good measure of their difference is the L2 difference between their log spectra:

$$D = \sum_{k=0}^{M} \left(\ln \left| X\left(\frac{(k+0.5)F_s}{N}\right)\right| - \ln \left| Y\left(\frac{(k+0.5)F_s}{N}\right)\right| \right)^2$$

$$= \sum_{k=0}^{M-1} (X_k - Y_k)^2 = \sum_{n=0}^{M-1} (x_n - y_n)^2 = \|\vec{x} - \vec{y}\|^2 = \left\|\vec{X} - \vec{Y}\right\|^2$$

Low-pass liftering smooths the spectrum



Low-pass liftered L2 norm

If you want to know whether two signals are the same vowel, then you want to know how different their smoothed spectra are. Let H(k) be your smoothing function. You smooth the log spectrum, then find the L2 distance:

$$\sum_{k=0}^{M}\left(H(k)*\ln\left|X\left(\frac{(k+0.5)F_s}{N}\right)\right|-H(k)*\ln\left|Y\left(\frac{(k+0.5)F_s}{N}\right)\right|\right)^2$$
$$=\sum_{k=0}^{M-1}(H(k)*X_k-H(k)*Y_k)^2=\sum_{n=0}^{M-1}h^2[n](x_n-y_n)^2$$

Low-pass liftered L2 norm

In particular, suppose

$$h[n] = \begin{cases} 1 & 0 < n \le 15 \\ 0 & n > 15 \end{cases}$$

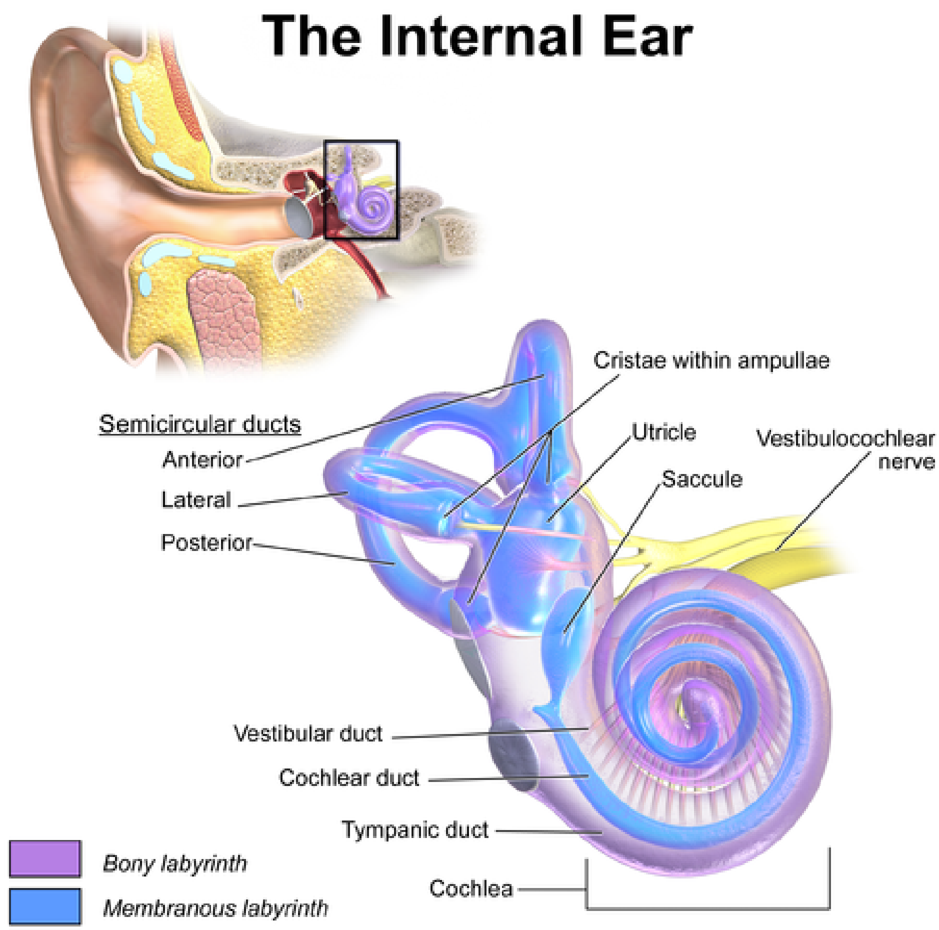
Then

$$\sum_{k=0}^{M} \left(H(k) * \ln \left| X\left(\frac{(k+0.5)F_s}{N}\right) \right| - H(k) * \ln \left| Y\left(\frac{(k+0.5)F_s}{N}\right) \right| \right)^2$$

$$= \sum_{n=1}^{15} (x_n - y_n)^2$$

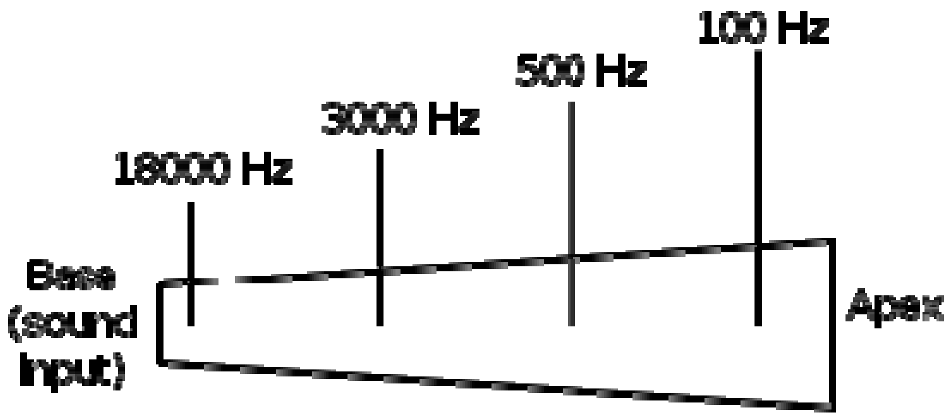
What spectrum do people hear? Basilar membrane

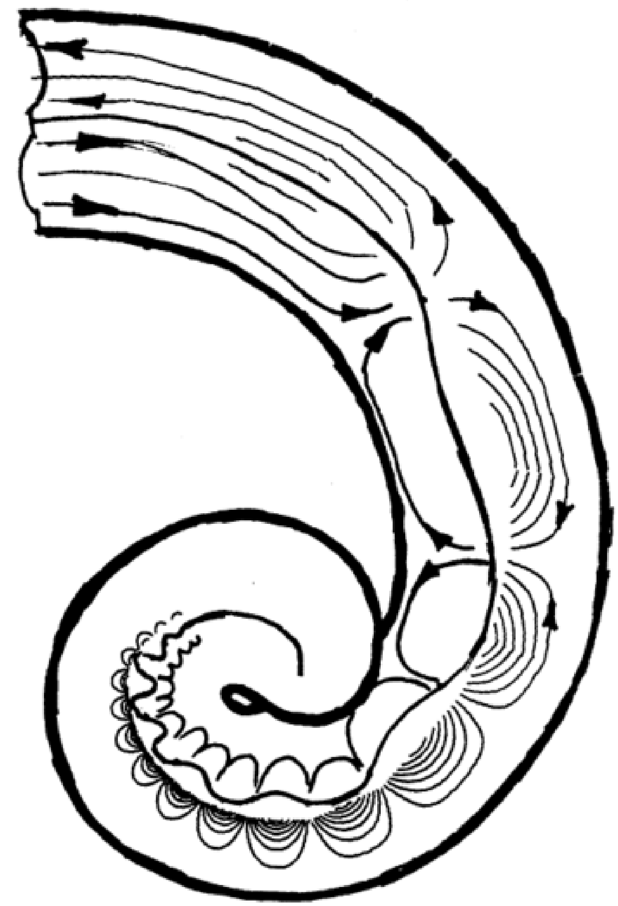
Inner ear



The Internal Ear
Cristae within ampullae
Semicircular ducts
Anterior
Lateral
Posterior
Utricle
Vestibulocochlear nerve
Saccule
Vestibular duct
Cochlear duct
Tympanic duct
Cochlea
Bony labyrinth
Membranous labyrinth

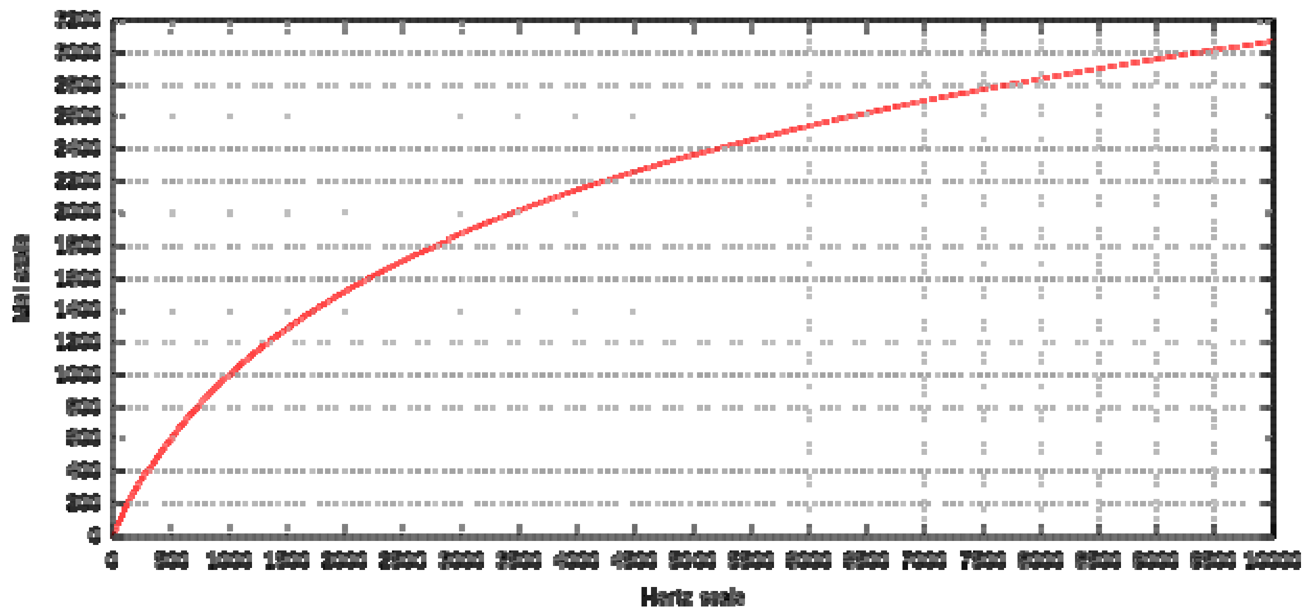
Basilar membrane of the cochlea = a bank of mechanical bandpass filters





Frequency scales for hearing: mel scale, ERB scale

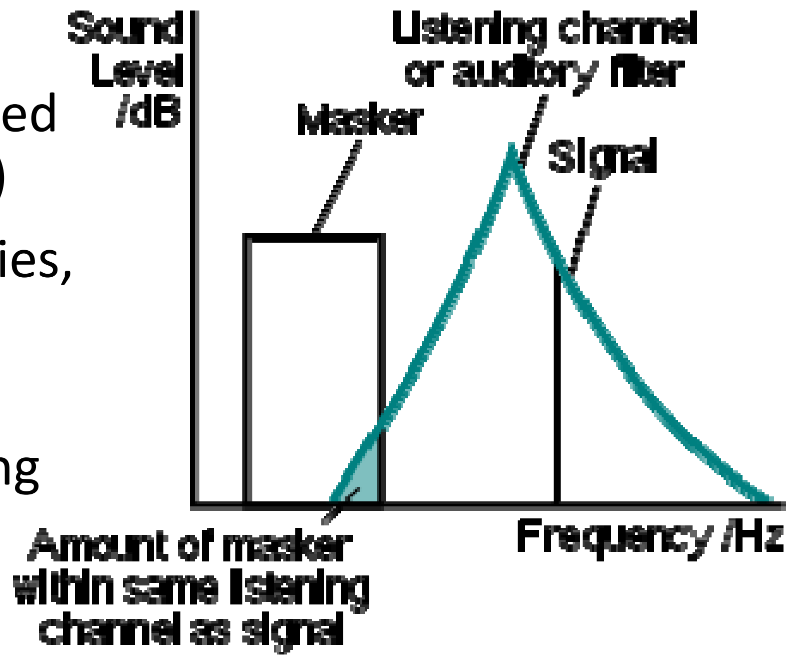
Mel-scale



- The experiment:
  - Play tones A, B, C
  - Let the user adjust tone D until pitch(D)-pitch(C) sounds the same as pitch(B)-pitch(A)
- Analysis: create a frequency scale m(f) such that m(D)-m(C) = m(B)-m(A)
- Result: $m(f) = \frac{1}{2595}\log_{10}\left(1 + \frac{f}{700}\right)$

Critical bands

- When two tones play at exactly the same frequency, users can't tell the difference between x(t) versus x(t)+y(t) if y(t) is about 14dB below x(t) (in other words, the summed power is 1.03 times the power of x(t) alone)
- When x(t) and y(t) are at different frequencies, the masking power of x(t) is reduced
- Model: assume that the reduced masking power of x(t) is caused because x(t) is coming in through the tails of the bandpass filter centered at y(t).



ERB scale

- The experiment: find out the widths, B(f), of the critical-band filters centered at every frequency f.
- Analysis: create a scale e(f) such that e(f+0.5B(f)) – e(f-0.5B(f)) = 1, for all frequencies
- Result: $\mathrm{e}(f) = 21.4\log_{10}(1 + 0.00437f)$

MFCC

Mel filterbank coefficients: convert the spectrum from Hertz-frequency to mel-frequency

- Goal: instead of computing

$$C_k = \ln\left|S\left(\frac{(k+0.5)F_s}{N}\right)\right|$$

We want

$$C_k = \ln|S(f_k)|$$

Where the frequencies f_k are uniformly spaced on a mel-scale, i.e., $\mathrm{m}(f_{k+1}) - \mathrm{m}(f_k)$ is a constant across all k.

The problem with that idea: we don’t want to just sample the spectrum. We want to summarize everything that’s happening within a frequency band.

Mel filterbank coefficients: convert the spectrum from Hertz-frequency to mel-frequency

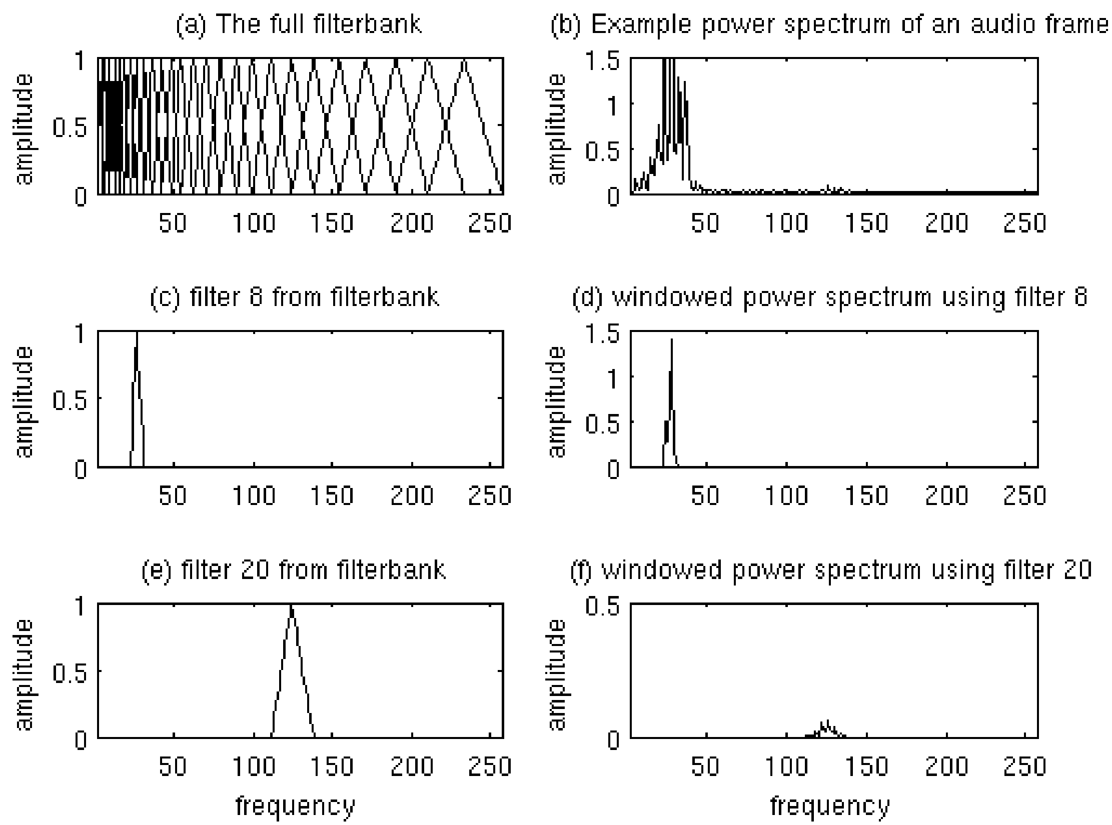
The solution:

$$C_m = \ln \sum_{k=0}^{\frac{N}{2}-1} W_m(k) \left| S\left(\frac{kF_s}{N}\right) \right|$$

Where

$$W_m(k) = \begin{cases} \dfrac{\frac{kF_s}{N} - f_{m-1}}{f_m - f_{m-1}} & f_m \geq \dfrac{kF_s}{N} \geq f_{m-1} \\ \dfrac{f_{m+1} - \frac{kF_s}{N}}{f_{m+1} - f_m} & f_{m+1} \geq \dfrac{kF_s}{N} \geq f_m \\ 0 & otherwise \end{cases}$$

Mel filterbank coefficients: convert the spectrum from Hertz-frequency to mel-frequency



MFCC: the full process

- Divide the acoustic signal into frames
- Compute the magnitude FFT of each frame
- Filterbank coefficients: $C_m = \ln \sum_{k=0}^{\frac{N}{2}-1} W_m(k) \left| S\left(\frac{kF_s}{N}\right) \right|$
- MFCC: $c[n] = \sum_{m=0}^{M-1} C_m \cos\left(\frac{\pi(m+0.5)n}{M}\right)$
- Liftering: keep only the first 12-15 MFCC coefficients, set the rest to zero.

Summary

- L2 distance(cepstra) = L2 distance(log magnitude spectra)
- L2 distance(windowed cepstrum) = L2 distance(smoothed log magnitude spectrum)