

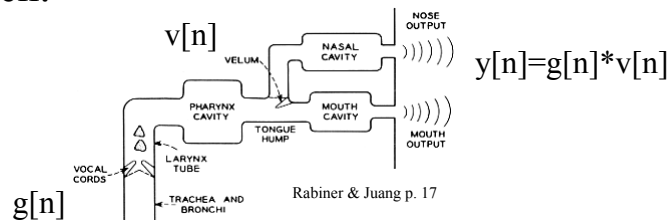
Cepstral processing

Professor Marie Roch

Huang: 6.4,6.4.1,6.4.4, 6.5.2, 9.3.3

Speech production & convolution

- Speech can be modeled as a convolution between:



- a glottal excitation source $g[n]$
- and vocal tract impulse response $v[n]$

Separating source & filter

- It is believed that the vocal tract characteristics is important for speech & speaker recognition.
- We would like to separate out this filter response.

Homomorphic transforms

- A *homomorphic transform* converts convolution:

$$y[n] = g[n] * v[n]$$

- to a sum:

$$\hat{y}[n] = \hat{g}[n] + \hat{v}[n]$$

Log frequency domain

- Recall that

$$Y(\omega) = G(\omega)V(\omega)$$

so

$$\log Y(\omega) = \log G(\omega) + \log V(\omega)$$

Cepstrum

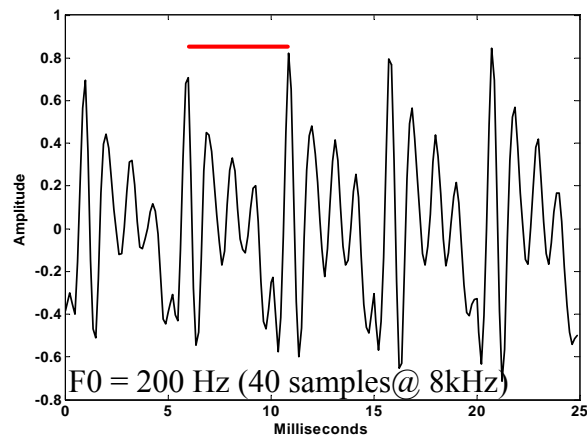
A homomorphic transform

- Compute power spectrum
 - Multiply each freq. bin by complex conjugate
- Take logarithm
 - Special steps may be needed to avoid log 0 problems.
- Apply 2nd DFT to move to cepstral domain.

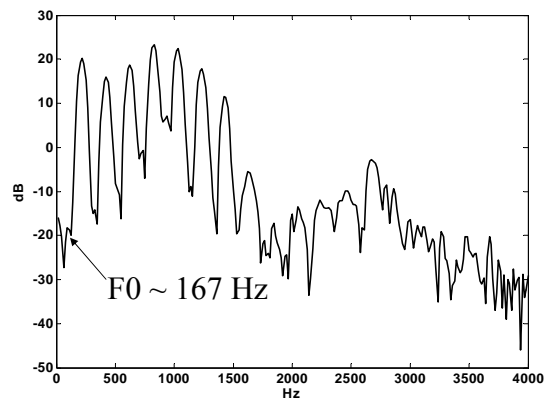
Cepstrum

- Some definitions use the inverse DFT, either one works well (duality of the Fourier transform).
- Denote the k^{th} cepstral coefficient as
 - $c[k]$ or
 - c_k

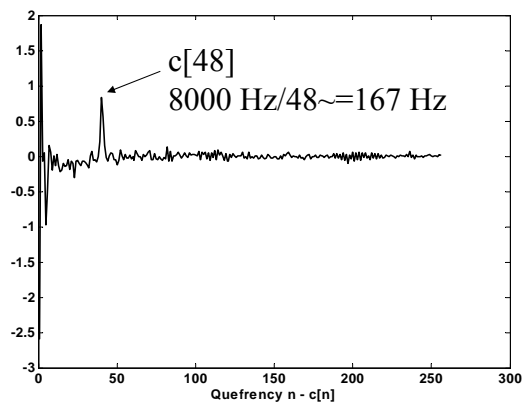
/ao, ɔ : / as in *Octopus*



Spectrum of /ao, ɔː/

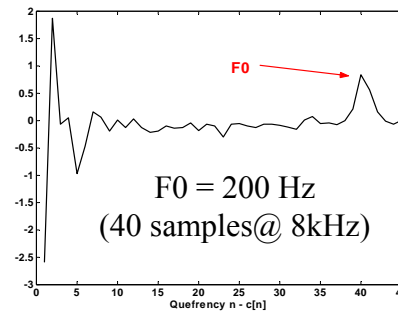


Cepstrum of /ao, ɔː/



Linear separation of source & filter

- When the signal is periodic, a pulse will occur every F_s/F_0 frames.
- $c[1]$ up to the pulse can be thought of as the response of the vocal tract.
- $c[40]$ contains the pulse and vocal tract information as we add the two.



Cepstrum

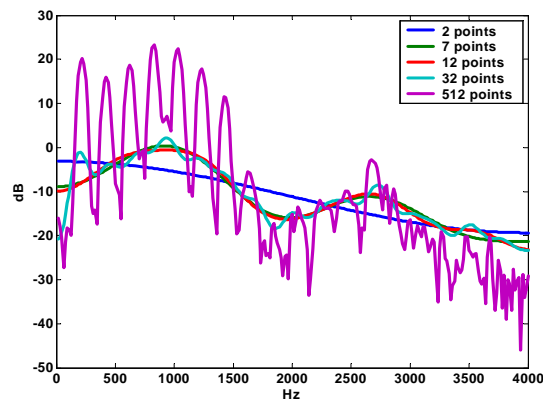
- $c[0]$ is similar to log energy (or DC portion) of the signal and is **not** referred to as a cepstral coefficient.
- Spectrum can be reconstructed by taking the inverse DFT of the cepstrum.
- Infinite series, but...
 - a DFT will only produce the first N points where N is the transform size.

Significance of quefrequency

- Lower quefrequencies are responsible for overall slope.
- Higher quefrequencies introduce higher frequency components.
- Retaining the first N quefrequencies, gives a smoothed approximation of the spectrum.

Reconstruction of spectrum

original spectrum 512 frequency bins



Practical issues

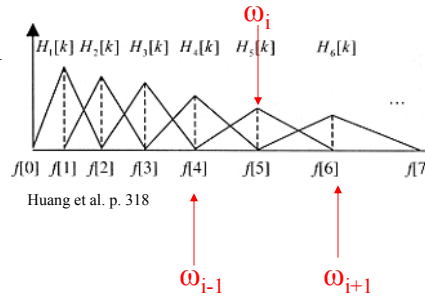
- The discrete cosine transform (DCT) is usually used instead of the inverse DFT
- There are several strategies for avoiding log 0 problems, i.e.:
 - Flooring
 - Set all log spectral values $< -k$ to $-k$ (i.e. $k=50$)
 - Adding a bias
 - Add k to all bins prior to logarithm (i.e. $k=1$).

Mel cepstrum

- Motivated by human perception
- Uses a filterbank to separate the spectrum into channels:
 - Lower frequency channels linearly spaced.
 - Higher frequency channels logarithmically spaced.

Mel-filter bank

- Filters spaced uniformly on Mel scale \rightarrow logarithmic on Hz scale
- Triangular-shaped filters emphasize center frequency ω_i and span to the next center frequency.



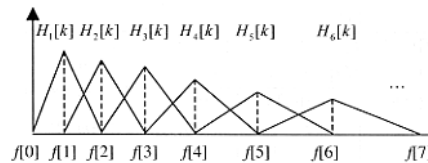
Mel-filter bank

- The area under each filter is constant and is sometimes scaled to sum to 1.
- Let M = desired number of filter banks.
- Distribute these uniformly across the Mel frequency space.
- Convert to Hz to get ω_i 's on linear scale:

$$Mel2Hz(mel) = 700 \left(e^{\frac{mel}{1125}} - 1 \right)$$

Constructing Mel filters

- Consider $H_4[k]$
- Asymmetric, spans
 - $f[3]$ to $f[4]$
 - $f[4]$ to $f[5]$
- Want normalized coefficients (e.g. all bin coefficients should add to 1).



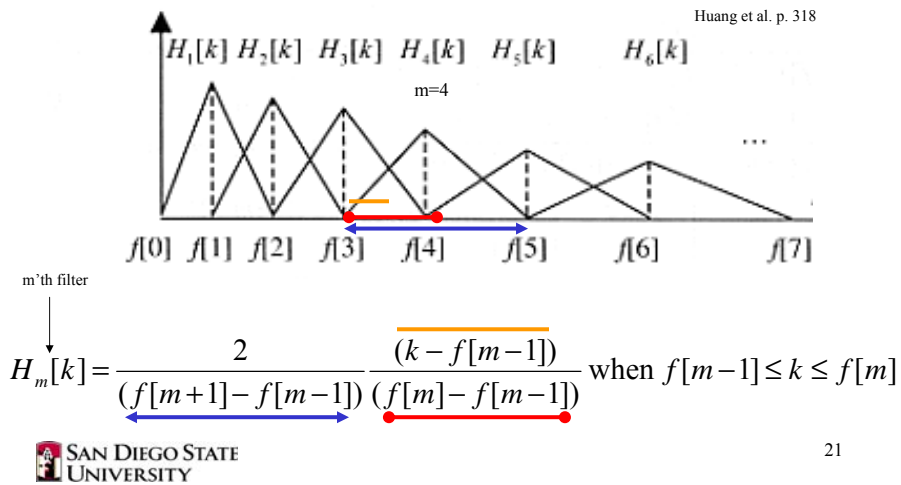
Huang et al. p. 318

Mel-filter bank

- Construct filters
 - $f[m]$ is the frequency bin associated with the ω_m 'th center frequency:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m+1] - f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases}$$

Example



Mel cepstrum

- Apply Mel filterbank

$$S[m] = \log \left[\sum_{k=0}^{N-1} |X[k]|^2 H_m[k] \right] \quad 0 < m \leq M$$

- Obtain Mel cepstrum by applying DCT

$$\begin{aligned}
 c[n] &= \text{dct}(S[m]) \\
 &= \sum_{m=0}^{M-1} S[m] \cos \left(\frac{\pi n (m - \frac{1}{2})}{M} \right) \quad 0 \leq n < M
 \end{aligned}$$

Mel cepstrum

- As we take the log of the summed filters, no longer homomorphic, but approximately so.
- Mel cepstrum is the most popular feature set today for speech/speaker recognition.

Mel cepstrum

- Common choices
 - At least 24 Mel filters for 16 kHz speech
 - Speech recognition, ~ 12 MFCC
 - Speaker recognition, >12, e.g. 18
- Frequently abbreviated MFCC for “Mel-filtered cepstral coefficients.”

Cepstral derivatives

- All of the modeling techniques that we have discussed do not capture how a signal changes.
- Cepstral derivatives, also known as *dynamic features*, are an attempt to capture information related to the evolution of the signal.

Cepstral derivatives

- The first derivative is computed by fitting a curve to the feature vectors that occur between N ms interval around the current feature vector.
- Typically, N is about 40-50 ms.

$$\underbrace{\begin{bmatrix} c_1 & c_1 & c_1 & c_1 & c_1 \\ c_1 & c_1 & c_1 & c_1 & c_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_d & c_d & c_d & c_d & c_d \end{bmatrix}}_{N \text{ms}}$$

Delta cepstrum

- This derivative is known as the delta cepstrum and usually denoted by the Greek letter Δ .
- While the fit can be complicated, the easiest fit is simply to subtract the earliest feature from the latest.

Delta cepstrum

- With a 10 ms advance, we could define a 40 ms first derivative as follows (more complex approximations are possible):

$$\Delta c_k = c_{k+2} - c_{k-2}$$

- The deltas are appended to the feature vector:

$$x = (c_1 \ c_2 \ \dots \ c_N \ \Delta c_1 \ \Delta c_2 \ \dots \ \Delta c_N)'$$

Delta-delta cepstrum

- The delta-delta cepstrum, or acceleration coefficients, are approximations of the second derivatives:

$$\Delta\Delta c_k = \Delta c_{k+1} - \Delta c_{k-1}$$

and can be similarly appended to the feature vector.

Feature extraction summary

Feature Set	Relative Error Reduction
13th-order LPC cepstrum coefficients	Baseline
13th-order MFCC	+10%
16th-order MFCC	+0%
+1st- and 2nd-order dynamic features	+20%
+3rd-order dynamic features	+0%

Huang et al. p. 426

- Note: LPC Cepstrum (not covered) less effective than cepstrum.