L9: Cepstral analysis

The cepstrum

Homomorphic filtering

The cepstrum and voicing/pitch detection

Linear prediction cepstral coefficients

Mel frequency cepstral coefficients

This lecture is based on [Taylor, 2009, ch. 12; Rabiner and Schafer, 2007, ch. 5; Rabiner and Schafer, 1978, ch. 7]

The cepstrum

Definition

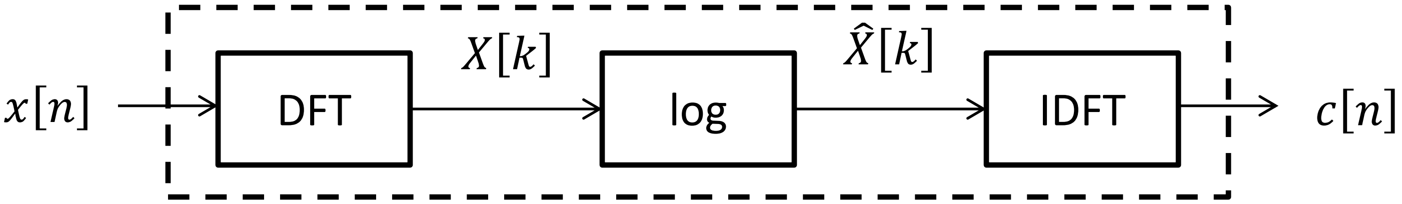
- The cepstrum is defined as the inverse DFT of the log magnitude of the DFT of a signal

$$c[n] = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]\}|\}$$

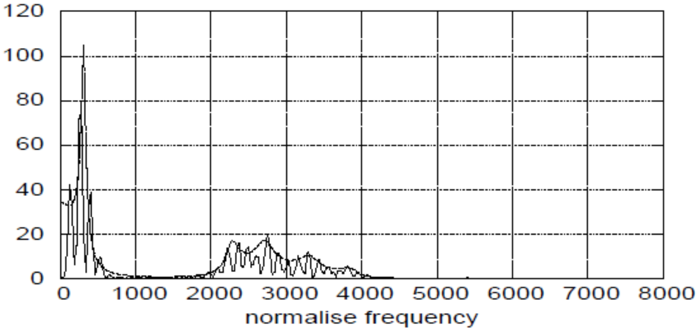
  - where $\mathcal{F}$ is the DFT and $\mathcal{F}^{-1}$ is the IDFT

- For a windowed frame of speech $y[n]$, the cepstrum is

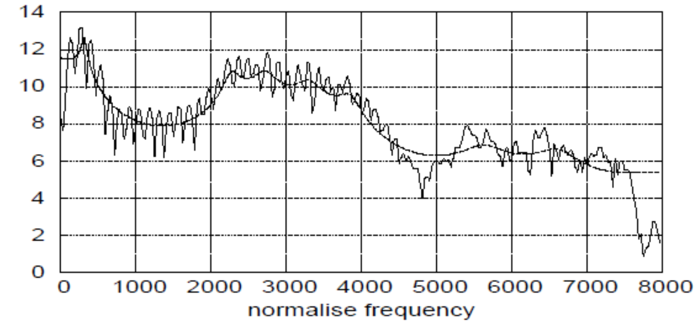
$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}\right|\right) e^{j\frac{2\pi}{N}kn}$$



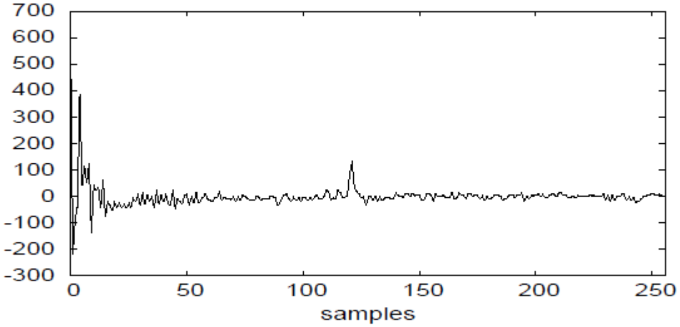
$\mathcal{F}\{x[n]\}$



$\log|\mathcal{F}\{x[n]\}|$



$\mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]\}|\}$

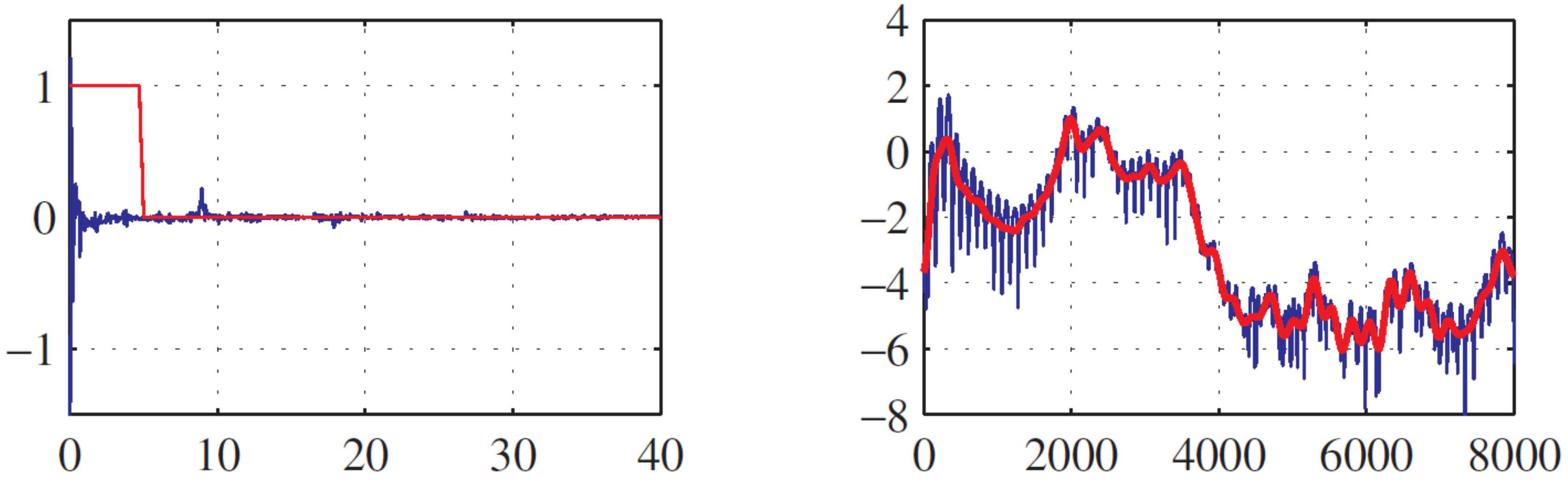


[Taylor, 2009]

Motivation

- Consider the magnitude spectrum $|\mathcal{F}\{x[n]\}|$ of the periodic signal in the previous slide
  - This spectra contains harmonics at evenly spaced intervals, whose magnitude decreases quite quickly as frequency increases
  - By calculating the log spectrum, however, we can compress the dynamic range and reduce amplitude differences in the harmonics
- Now imagine we were told that the log spectra was a waveform
  - In this case, we would we would describe it as quasi-periodic with some form of amplitude modulation
  - To separate both components, we could then employ the DFT
  - We would then expect the DFT to show
    - A large spike around the "period" of the signal, and
    - Some "low-frequency" components due to the amplitude modulation
  - A simple filter would then allow us to separate both components (see next slide)

Liftering in the cepstral domain



[Rabiner & Schafer, 2007]

Cepstral analysis as deconvolution

- As you recall, the speech signal can be modeled as the convolution of glottal source $u[n]$, vocal tract $v[n]$, and radiation $r[n]$

$$y[n] = u[n] \otimes v[n] \otimes r[n]$$

  - Because these signals are convolved, they cannot be easily separated in the time domain

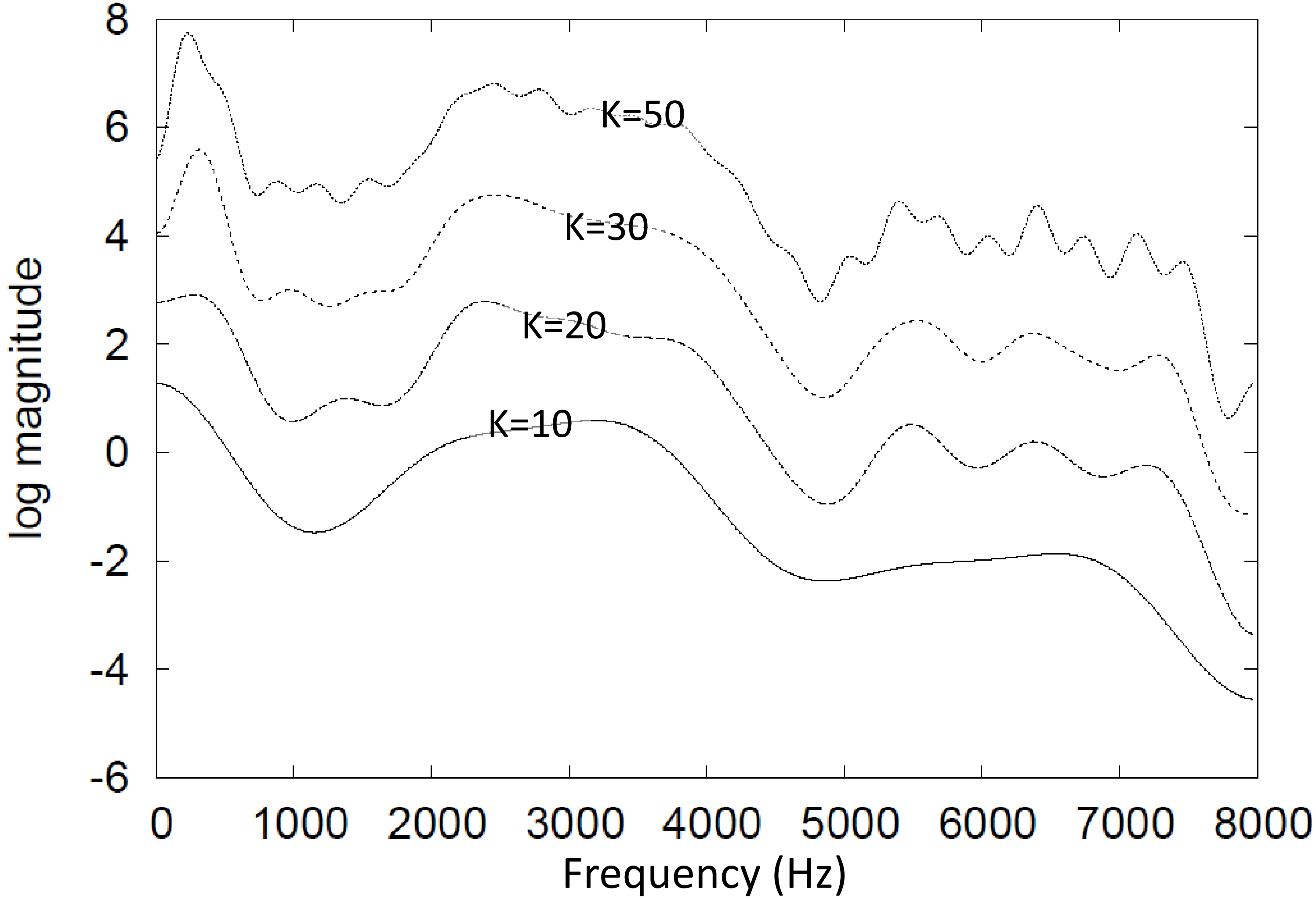
- We can however perform the separation as follows
  - For convenience, we combine $v'[n] = v[n] \otimes r[n]$, which leads to
$$y[n] = u[n] \otimes v'[n]$$
  - Taking the Fourier transform
$$Y(e^{j\omega}) = U(e^{j\omega})V'(e^{j\omega})$$
  - If we now take the log of the magnitude, we obtain
$$\log(|Y(e^{j\omega})|) = \log(|U(e^{j\omega})|) + \log(|V'(e^{j\omega})|)$$
    - which shows that source and filter are now just added together
  - We can now return to the time domain through the inverse FT
$$c[n] = c_u[n] + c_v[n]$$

Where does the term "cepstrum" come from?

- The crucial observation leading to the cepstrum terminology is that the log spectrum can be treated as a waveform and subjected to further Fourier analysis
- The independent variable of the cepstrum is nominally time since it is the IDFT of a log-spectrum, but is interpreted as a frequency since we are treating the log spectrum as a waveform
- To emphasize this interchanging of domains, Bogert, Healy and Tukey (1960) coined the term cepstrum by swapping the order of the letters in the word spectrum
- Likewise, the name of the independent variable of the cepstrum is known as a quefrency, and the linear filtering operation in the previous slide is known as liftering

Discussion

- If we take the DFT of a signal and then take the inverse DFT of that, we of course get back the original signal (assuming it is stationary)
- The cepstrum calculation is different in two ways
  - First, we only use magnitude information, and throw away the phase
  - Second, we take the IDFT of the log-magnitude which is already very different since the log operation emphasizes the “periodicity” of the harmonics
- The cepstrum is useful because it separates source and filter
  - If we are interested in the glottal excitation, we keep the high coefficients
  - If we are interested in the vocal tract, we keep the low coefficients
  - Truncating the cepstrum at different quefrency values allows us to preserve different amounts of spectral detail (see next slide)


8
6
4
2
0
-2
-4
-6
log magnitude
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequency (Hz)
K=50
K=30
K=20
K=10

[Taylor, 2009]

- Note from the previous slide that cepstral coefficients are a very compact representation of the spectral envelope
- It also turns out that cepstral coefficients are (to a large extent) uncorrelated
  - This is very useful for statistical modeling because we only need to store their mean and variances but not their covariances
- For these reasons, cepstral coefficients are widely used in speech recognition, generally combined with a perceptual auditory scale, as we see next

ex9p1.m

Computing the cepstrum

Liftering vs. linear prediction

Show uncorrelatedness of cepstrum

Homomorphic filtering

Cepstral analysis is a special case of homomorphic filtering

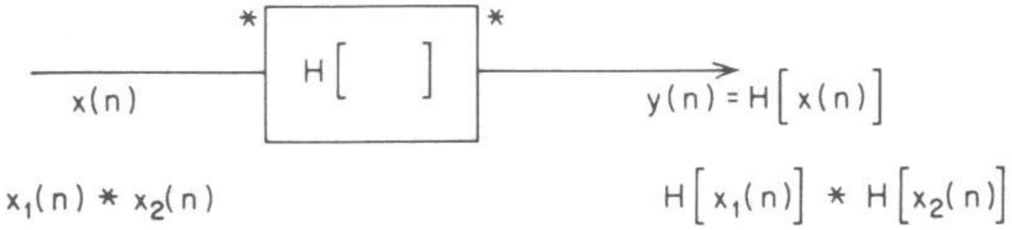
- Homomorphic filtering is a generalized technique involving (1) a nonlinear mapping to a different domain where (2) linear filters are applied, followed by (3) mapping back to the original domain
- Consider the transformation defined by $y(n) = L[x(n)]$
  - If L is a linear system, it will satisfy the principle of superposition

$$L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)]$$

- By analogy, we define a class of systems that obey a generalized principle of superposition where addition is replaced by convolution

$$H[x(n)] = H[x_1(n) * x_2(n)] = H[x_1(n)] * H[x_2(n)]$$

  - Systems having this property are known as homomorphic systems for convolution, and can be depicted as shown below



[Rabiner & Schafer, 1978]

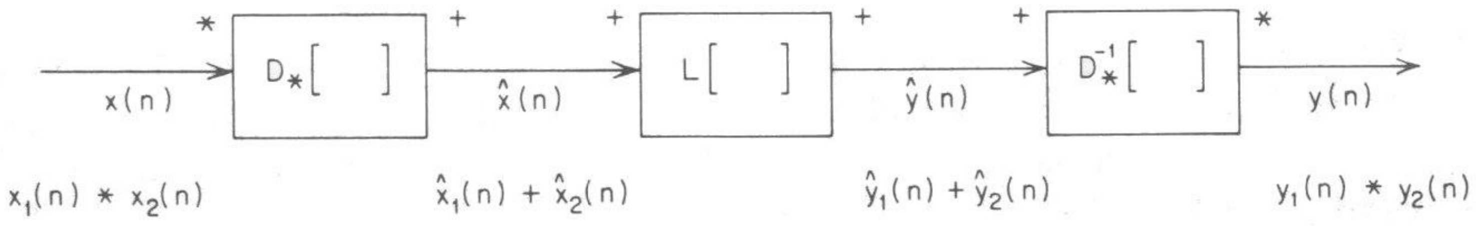
- An important property of homomorphic systems is that they can be represented as a cascade of three homomorphic systems
  - The first system takes inputs combined by convolution and transforms them into an additive combination of the corresponding outputs

$$D_*[x(n)] = D_*[x_1(n) * x_2(n)] = D_*[x_1(n)] + D_*[x_2(n)] = \hat{x}_1(n) + \hat{x}_2\ (n) = \hat{x}(n)$$

  - The second system is a conventional linear system that obeys the principle of superposition
  - The third system is the inverse of the first system: it transforms signals combined by addition into signals combined by convolution

$$D_*^{-1}[\hat{y}(n)] = D_*^{-1}[\hat{y}_1(n) + \hat{y}_2(n)] = D_*^{-1}[y_1(n)] * D_*^{-1}[y_2(n)] = y_1(n) * y_2\ (n) = y(n)$$

- This is important because the design of such system reduces to the design of a linear system



[Rabiner & Schafer, 1978]

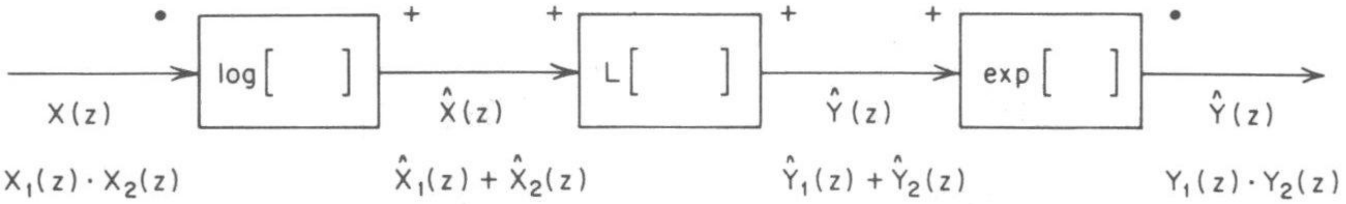
- Noting that the z-transform of two convolved signals is the product of their z-transforms

$$x(n) = x_1(n) * x_2(n) \Leftrightarrow X(z) = X_1(z)X_2(z)$$

- We can then take logs to obtain

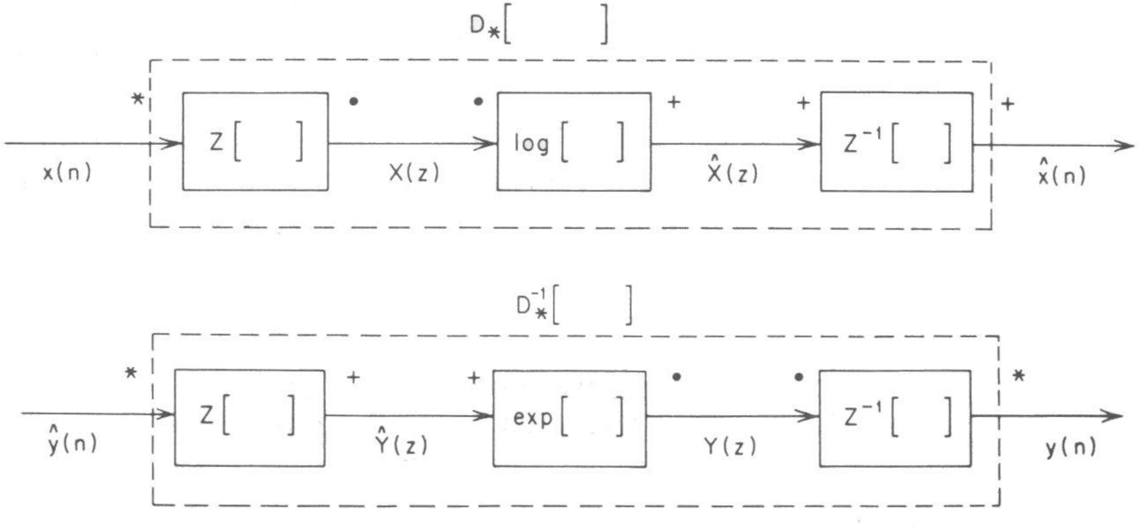
$$\hat{X}(z) = \log[X(z)] = \log[X_1(z)X_2(z)] = \log[X_1(z)] + \log[X_2(z)]$$

- Thus, the frequency domain representation of a homomorphic system for deconvolution can be represented as



[Rabiner & Schafer, 1978]

- If we wish to represent signals as sequence rather than in the frequency domain, then the systems $D_*[\quad]$ and $D_*^{-1}[\quad]$ can be represented as shown below
  - Where you can recognize the similarity between the cepstrum with the system $D_*[\quad]$
  - Strictly speaking, $D_*[\quad]$ defines a complex cepstrum, whereas in speech processing we generally use the real cepstrum
- Can you find the equivalent system for the liftering stage?

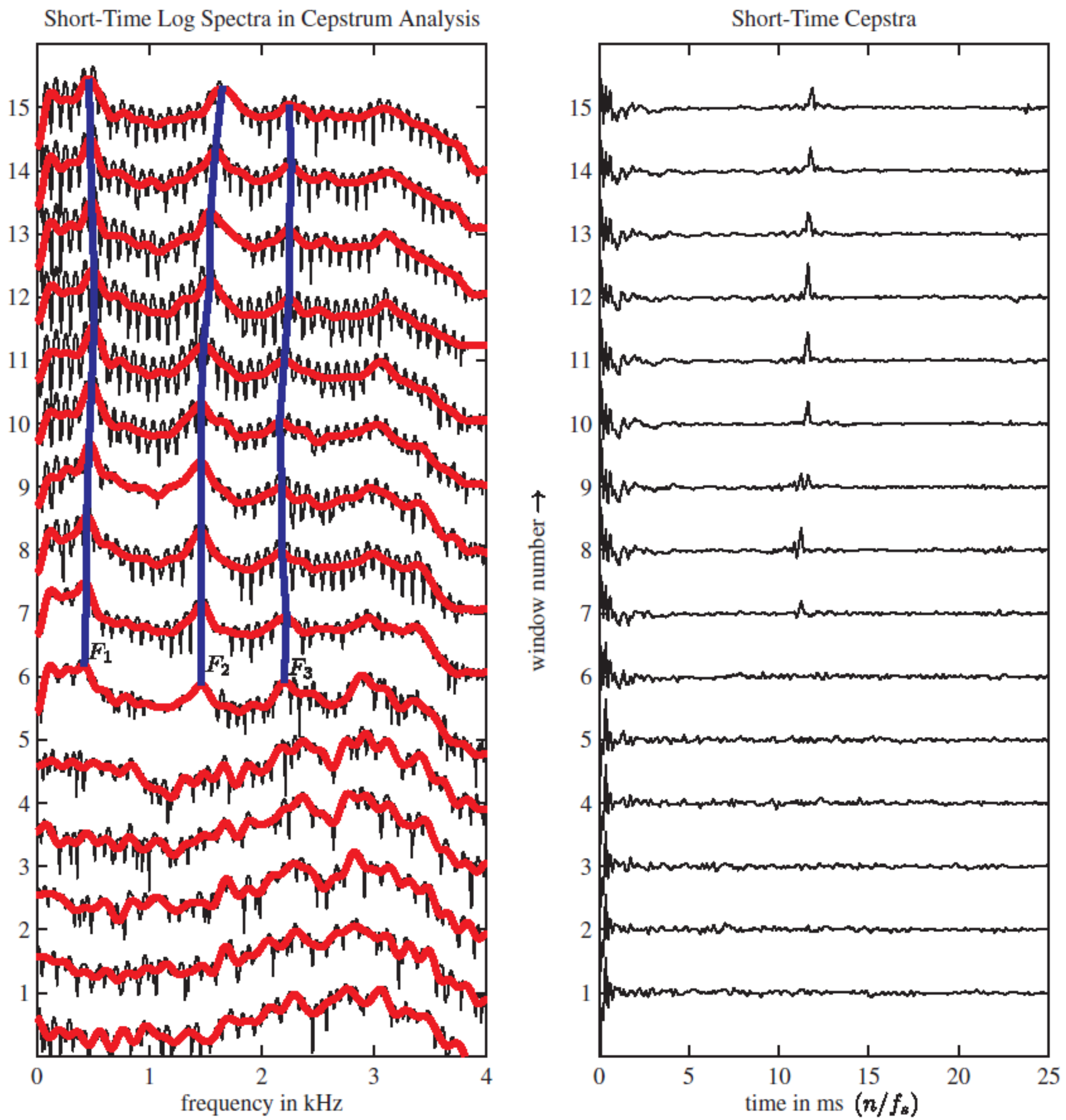


[Rabiner & Schafer, 1978]

Voicing/pitch detection

Cepstral analysis can be applied to detect local periodicity

- The figure in the next slide shows the STFT and corresponding spectra for a sequence of analysis windows in a speech signal (50-ms window, 12.5-ms shift)
- The STFT shows a clear difference in harmonic structure
  - Frames 1-5 correspond to unvoiced speech
  - Frames 8-15 correspond to voiced speech
  - Frames 6-7 contain a mixture of voiced and unvoiced excitation
- These differences are perhaps more apparent in the cepstrum, which shows a strong peak at a quefrency of about 11-12 ms for frames 8-15
- Therefore, presence of a strong peak (in the 3-20 ms range) is a very strong indication that the speech segment is voiced
  - Lack of a peak, however, is not an indication of unvoiced speech since the strength or existence of a peak depends on various factors, i.e., length of the analysis window



[Rabiner & Schafer, 2007]

Cepstral-based parameterizations

Linear prediction cepstral coefficients

- As we saw, the cepstrum has a number of advantages (source-filter separation, compactness, orthogonality), whereas the LP coefficients are too sensitive to numerical precision
- Thus, it is often desirable to transform LP coefficients $\{a_n\}$ into cepstral coefficients $\{c_n\}$
- This can be achieved as [Huang, Acero & Hon, 2001]

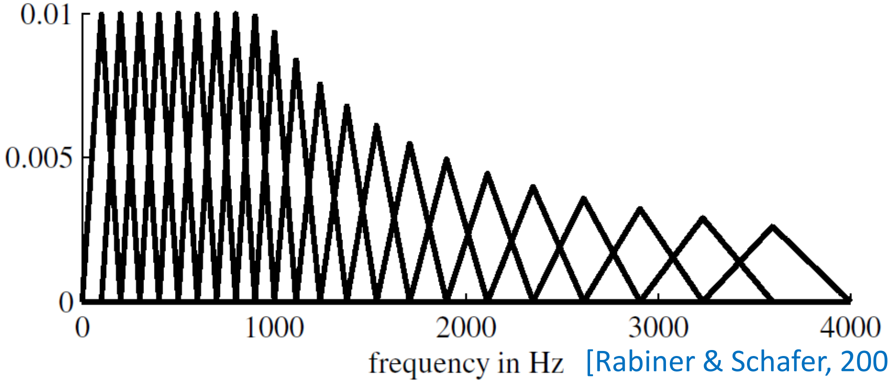
$$c_n = \begin{cases} \ln(G) & n = 0 \\ a_n + \frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} & 1 < n \leq p \end{cases}$$

Mel Frequency Cepstral Coefficients (MFCC)

- Probably the most common parameterization in speech recognition
- Combines the advantages of the cepstrum with a frequency scale based on critical bands

Computing MFCCs

- First, the speech signal is analyzed with the STFT
- Then, DFT values are grouped together in critical bands and weighted according to the triangular weighting function shown below
  - These bandwidths are constant for center frequencies below 1KHz and increase exponentially up to half the sampling rate



[Rabiner & Schafer, 2007]

- The mel-frequency spectrum at analysis time n is defined as

$$MF[r] = \frac{1}{A_r}\sum\nolimits_{k=L_r}^{U_r} |V_r[k]X(n,k)|$$

- where $V_r[k]$ is the triangular weighting function for the r-th filter, ranging from DFT index L_r to U_r and

$$A_r = \sum\nolimits_{k=L_r}^{U_r} |V_r[k]|^2$$

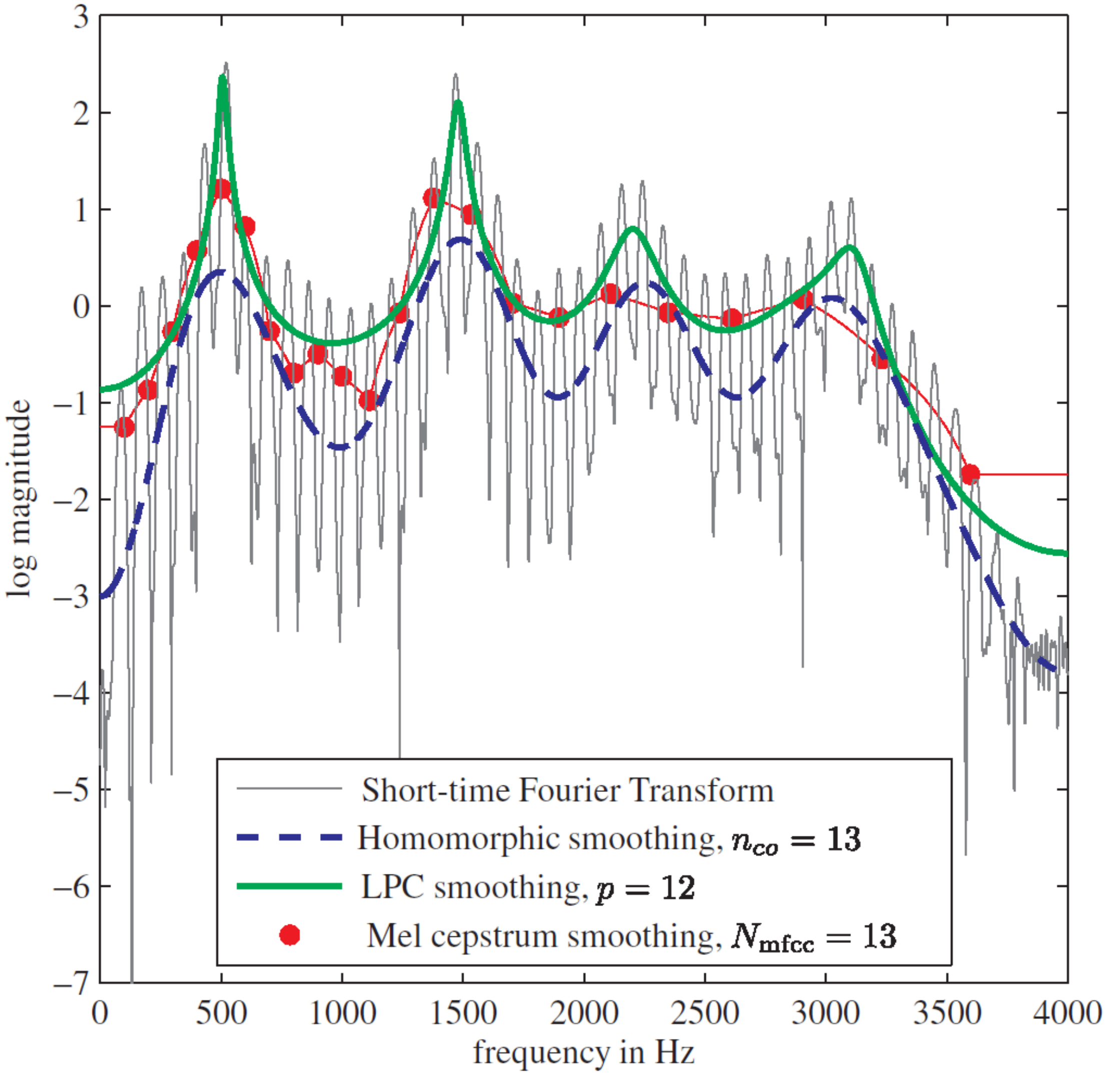
  - which serves as a normalization factor for the r-th filter, so that a perfectly flat Fourier spectrum will also produce a flat Mel-spectrum

- For each frame, a discrete cosine transform (DCT) of the log-magnitude of the filter outputs is then computed to obtain the MFCCs

$$MFCC[m] = \frac{1}{R}\sum\nolimits_{r=1}^{R} \log(MF[r]) \cos\left[\frac{2\pi}{R}\left(r+\frac{1}{2}\right)m\right]$$

- where typically $MFCC[m]$ is evaluated for a number of coefficients N_{MFCC} that is less than the number of mel-filters R
  - For $F_s = 8KHz$, typical values are $N_{MFCC} = 13$ and $R = 22$

Comparison of smoothing techniques: LPC, cepstral and mel-cepstral



[Rabiner & Schafer, 2007]

Notes

- The MFCC is no longer a homomorphic transformation
  - It would be if the order of summation and logarithms were reversed, in other words if we computed

$$\frac{1}{A_r}\sum_{k=L_r}^{U_r} \log|V_r[k]X(n,k)|$$

  - Instead of

$$\log\left(\frac{1}{A_r}\sum_{k=L_r}^{U_r} |V_r[k]X(n,k)|\right)$$

  - In practice, however, the MFCC representation is approximately homomorphic for filters that have a smooth transfer function
  - The advantages of the second summation above is that the filter energies are more robust to noise and spectral properties

- MFCCs employ the DCT instead of the IDFT
  - The DCT turns out to be closely related to the Karhunen-Loeve transform
    - The KL transform is the basis for PCA, a technique that can be used to find orthogonal (uncorrelated) projections of high dimensional data
    - As a result, the DCT tends to decorrelate the mel-scale frequency log-energies
  - Relationship with the DFT
    - The DCT transform (the DCT-II, to be exact) is defined as

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

    - whereas the DFT is defined as

$$X_{DFT}(k) = \sum_{n=0}^{N-1} x[n] e^{j\frac{2\pi}{N}kn}$$

    - Under some conditions, the DFT and the DCT-II are exactly equivalent
    - For our purposes, you can think of the DCT as a “real” version of the DFT (i.e., no imaginary part) that also has a better energy compaction properties than the DFT

ex9p2.m

Computing MFCCs