Speech Communication, Spring 2006

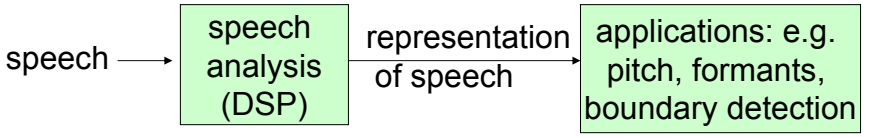
Lecture 2: Speech Analysis

Zheng-Hua Tan

Department of Communication Technology
Aalborg University, Denmark
zt@kom.aau.dk

Speech analysis

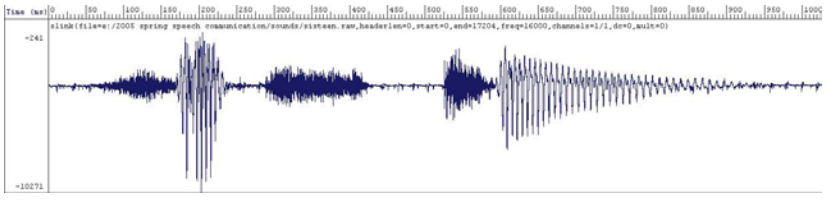
- Previous study:
  - Production speech
  - Properties of speech signals
- Most applications of speech processing must exploit the properties of speech signals → Speech Analysis: the process of extracting such properties from a speech signal.



Part I: Short-time speech analysis

- Short-time speech analysis
- Time-domain processing
- Frequency-domain (spectral) processing
- Linear predictive coding (LPC) analysis
- Cepstral analysis

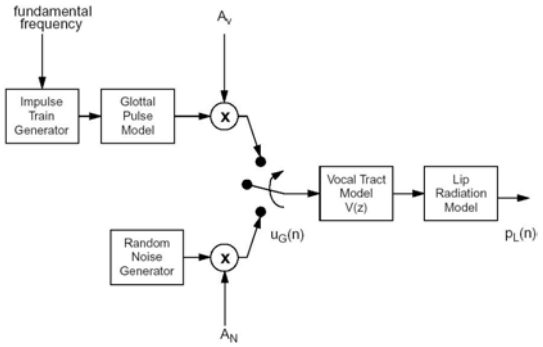
Properties of speech signals



Speech is a time-varying signal:

- excitation
- pitch
- amplitude

Discrete-time filter model for speech



Time-varying parameters: fundamental freqency (pitch), voiced/unvoiced/silence, gain, formants, vocal tract area functions, etc

Short-time processing solution

Assuming that speech has non-time-varying properties (fixed excitation and vocal tract) within short intervals →

Processing short segments (frames) of the speech signal each time

$$f_x(n,m) = x(m)w(n-m)$$

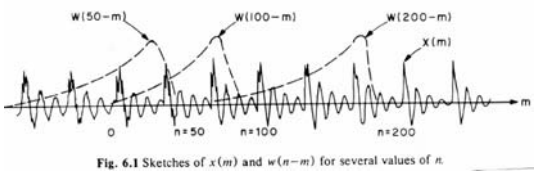
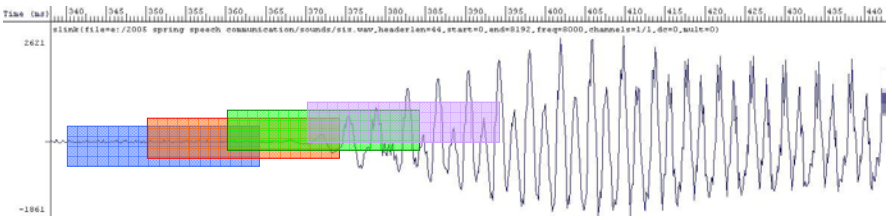


Fig. 6.1 Sketches of $x(m)$ and $w(n-m)$ for several values of n.

Frame-by-frame processing

- frames (segments) often overlap one another

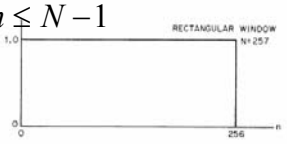


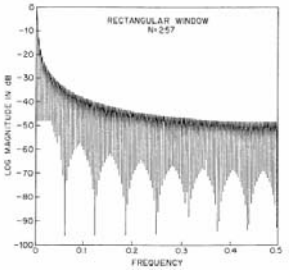
- The frame-based analysis yields a time-varying sequence as a new representation of the speech signal
  - samples at 8000/sec → vectors at 100/sec

Windows

- Rectangular window

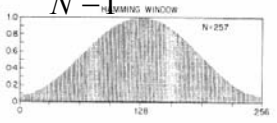
$$w[n] = 1, \qquad 0 \le n \le N-1$$

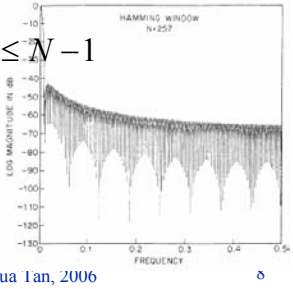




- Hamming window

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1$$





Choice of window

- Window type
  - Bandwidth of Hamming window is about twice the bandwidth of Rectangular
  - Attenuation of more than 40dB for Hamming as compared with 14 dB for Rectangular, outside passband
- Window duration - N
  - Increase N = decrease window bandwidth
  - N should be larger than a pitch period, but smaller than a sound duration

Part II: Time-domain processing

- Short-time speech analysis
- Time-domain speech processing
- Frequency-domain (spectral) processing
- Linear predictive coding (LPC) analysis
- Cepstral analysis

Time-domain parameters

- Short-time energy
- Short-time average magnitude
- Short-time zero crossing rate
- Short-time autocorrelation
- Short-time average magnitude difference

Short-time energy

- The long term energy definition is not useful for time-varying signals

$$E = \sum_{m=-\infty}^{\infty} x^2(m)$$

- Short-time energy of weighted signal around n is defined as

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$

Examples of short-time energy

- It can be used to detection voiced/unvoiced/silence
  - Effects of window type, duration N (bandwidth), why?

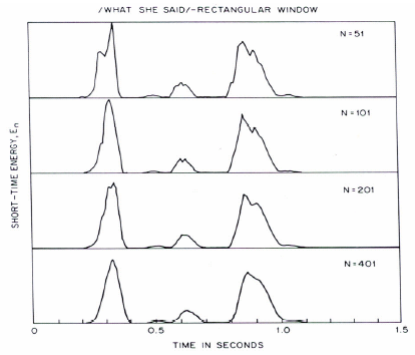


Fig. 4.6 Short-time energy functions for rectangular windows of various lengths.

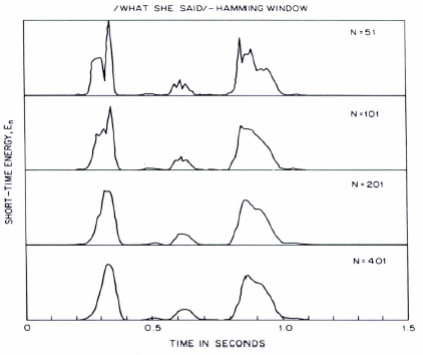


Fig. 4.7 Short-time energy functions for Hamming windows of various lengths.

Short-time magnitude

- Less sensitive to large signal levels as compared to energy where $x^2(n)$ terms is used.

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|\, w(n-m)$$

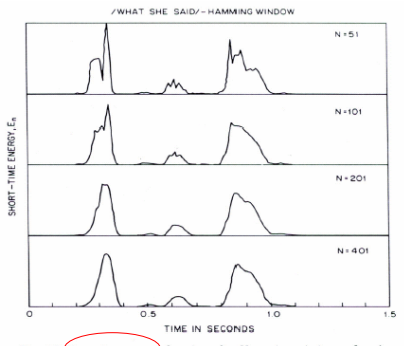


Fig. 4.7 Short-time energy functions for Hamming windows of various lengths.

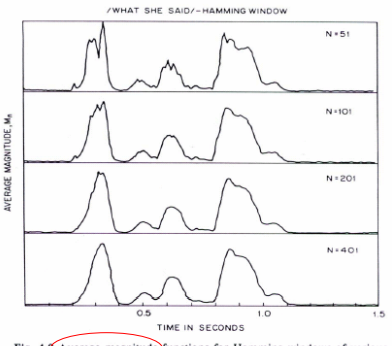
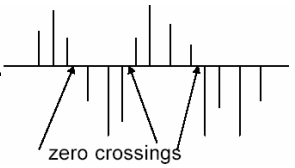


Fig. 4.9 Average magnitude functions for Hamming windows of various lengths.

Short-time average zero-crossing rate

- A zero-crossing occurs if successive samples have different algebraic signs.
- It is a measure of the frequency.
- Definition



$$Z_n = \sum_{m=-\infty}^{\infty} |\,\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\,|\, w(n-m)$$

where

$$\mathrm{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases}$$

and

$$w(n) = \begin{cases} \dfrac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & otherwise \end{cases}$$

Zero-crossing rate distributions

- A histogram of average zero-crossing rates (averaged over 10 msec) for both voiced and unvoiced speech
- Energe locates in different frequency bands

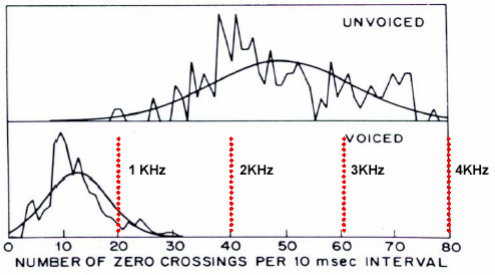


Fig. 4.11 Distribution of zero-crossings for unvoiced and voiced speech.

Example of zero-crossing rate

- Although the zero-crossing rate varies considerably, the voiced and unvoiced regions are quite prominent.

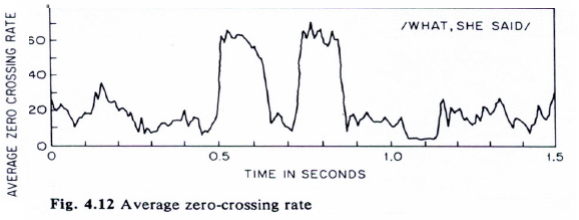


Fig. 4.12 Average zero-crossing rate

Short-time autocorrelation function

- The autocorrelation function

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k)$$

- The short-time autocorrelation function

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m+k)w(n-k-m)$$

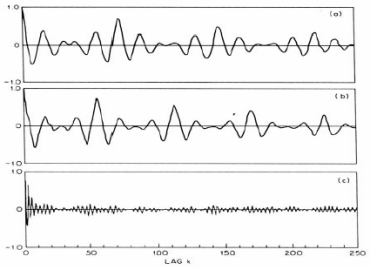


Fig. 4.24 Autocorrelation function for (a) and (b) voiced speech; and (c) unvoiced speech, using a rectangular window with $N = 401$.

Applications

- Boundary detection
  - short-time energy
  - zero crossing rate
- Pitch estimation
  - short-time autocorrelation function

Part III: Frequency-domain process.

- Short-time speech analysis
- Time-domain speech processing
- Frequency-domain (spectral) processing
- Linear predictive coding (LPC) analysis
- Cepstral analysis

Discrete-time Fourier transform

$$\begin{cases} X(e^{jw}) = \sum_{n=-\infty}^{+\infty} x[n]e^{-jwn} \\ x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{jw})e^{jwn}dw \end{cases}$$

Convolution and multiplication duality:

$$\begin{cases} y[n] = x[n] * h[n] \\ Y(e^{jw}) = X(e^{jw})H(e^{jw}) \end{cases}$$

$$\begin{cases} y[n] = x[n]w[n] \\ Y(e^{jw}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})X(e^{j(w-\theta)})d\theta \end{cases}$$

Short-time Fourier transform

- It is motivated by the need for a spectral representation to reflect the time-varying properties of the speech waveform

$$X_n(e^{jw}) = \sum_{m=-\infty}^{+\infty} w[n-m]x[m]e^{-jwm}$$

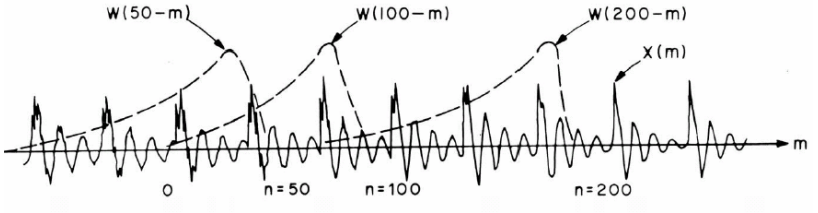
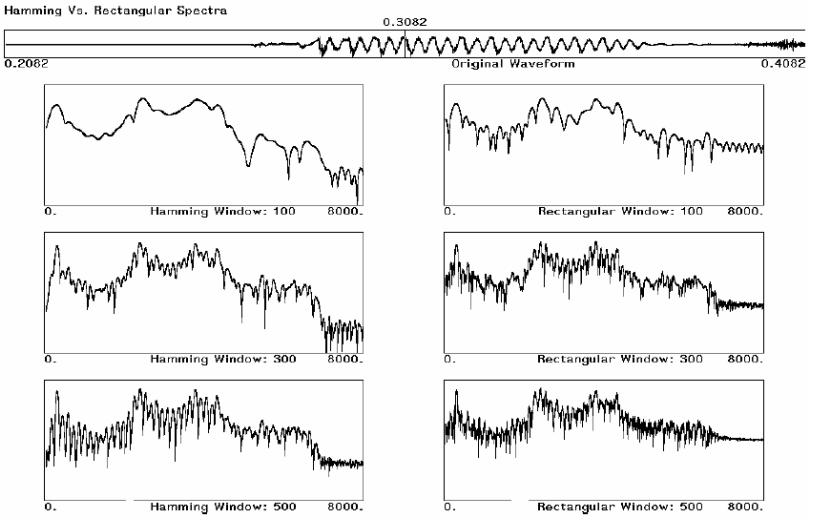


Fig. 6.1 Sketches of $x(m)$ and $w(n-m)$ for several values of n.

Spectra



Spectra of voiced/unvoiced sounds

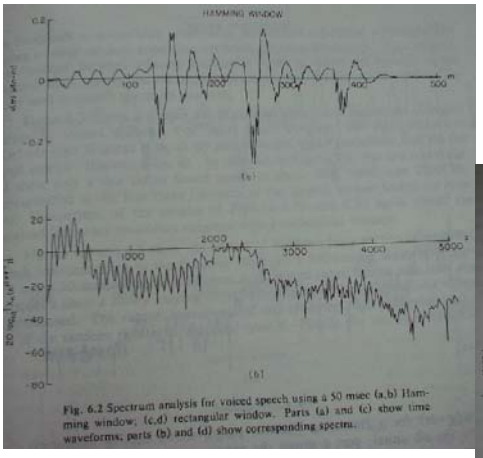


Fig. 6.2 Spectrum analysis for voiced speech using a 50 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

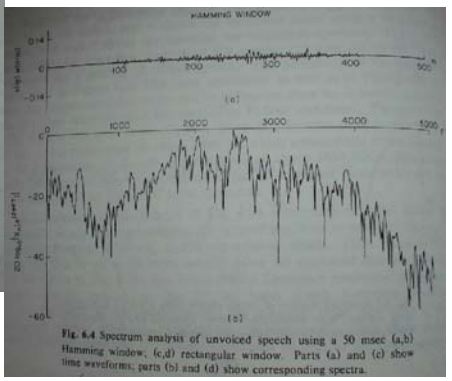
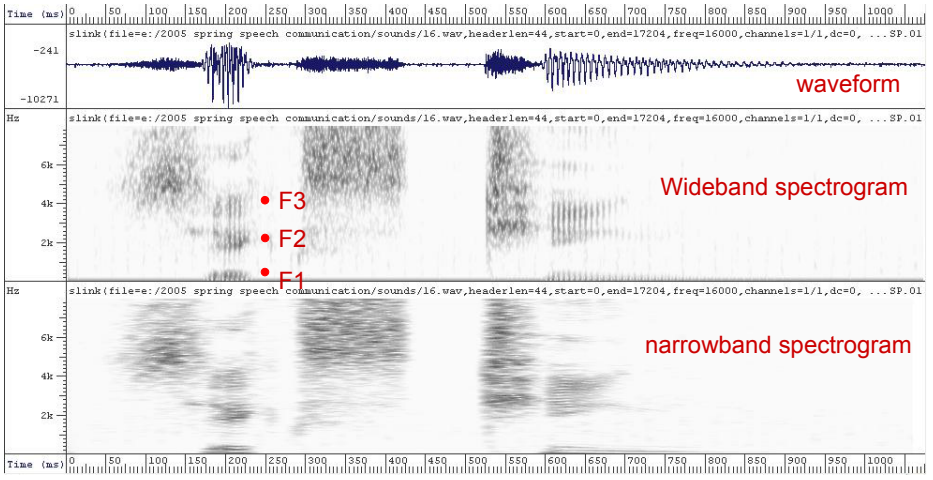


Fig. 6.4 Spectrum analysis of unvoiced speech using a 50 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

Spectrogram

- Spectrogram
  - two-dimensional waveform (amplitude/time) is converted into a three-dimensional pattern (amplitude/frequency/time)
  - Wideband spectrogram: analyzed on 15ms sections of waveform with a step of 1ms
    - voiced regions with vertical striations due to the periodicity of the time waveform (each vertical line represents a pulse of vocal folds) while unvoiced regions are solid/random, or 'snowy'
  - Narrowband spectrogram: on 50ms
    - pitch for voiced intervals in horizontal lines

Wide- and narrow-band spectrograms



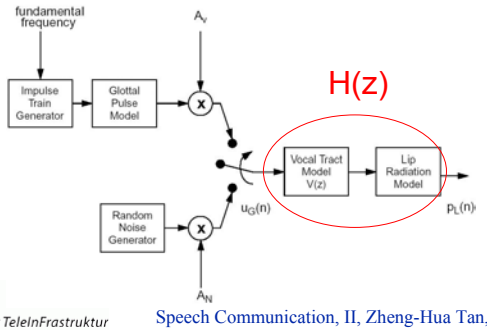
Part IV: LPC analysis

- Short-time speech analysis
- Time-domain speech processing
- Frequency-domain (spectral) processing
- Linear predictive coding (LPC) analysis
- Cepstral analysis

Discrete-time filter model for speech

Its philosophy is related to the speech model in which speech is modelled as the output of a linear, time-varying system excited by either quasi-periodic pulses or random noise.

The LPC provides a robust and accurate method for estimating the parameters of the time-varying system.



LPC analysis

- For efficient coding, speech signals are often modelled using parameters of the vocal tract shape that generates them.
- Pole-zero model (ideal during a stationary frame)

$$\hat{H}(z) = \frac{\hat{S}(z)}{\hat{U}(z)} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

- All-pole model (simple): a matter of analytical necessity

$$\hat{H}(z) = \frac{\hat{S}(z)}{\hat{U}(z)} = G \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

All-pole model – the LPC model

$$\hat{H}(z) = \frac{\hat{S}(z)}{\hat{U}(z)} = G \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \rightarrow \hat{S}(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \hat{U}(z)$$

$$\rightarrow \quad \hat{S}(z) = \hat{S}(z) \sum_{k=1}^{p} a_k z^{-k} + G\hat{U}(z)$$

$$\rightarrow \quad s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n)$$

where u(n) is a normalised excitation and G is the gain of the excitation

The LPC model

After excluding the excitation term, a given speech sample at time *n*, *s*(*n*), can be approximated as a linear combination of the past *p* speech samples:

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p)$$

where the coefficients $a_1, a_2, \ldots, a_p$ are assumed constant over the speech frame.

LPC analysis equations

Windowed speech: $x(n) = s(n)w(n)$

Error of linear predictor $e(n) = s(n) - \hat{s}(n)$

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

Error

$$e(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k)$$

Error energy

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} [x(n) - \sum_{k=1}^{p} a_k x(n-k)]$$

LPC analysis equations (cont'd)

Find a_k such that E is minimal

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} [x(n) - \sum_{k=1}^{p} a_k x(n-k)]$$

$$\frac{\partial E}{\partial a_k} = 0 \text{ for } k = 1,2,...,p$$

giving $$\sum_{n=-\infty}^{\infty} x(n-i)x(n) = \sum_{k=1}^{p} \hat{a}_k \sum_{n=-\infty}^{\infty} x(n-i)x(n-k)$$

given covariance $$\phi(i,k) = \sum_{n=-\infty}^{\infty} x(n-i)x(n-k)$$

so, $$\phi(i,0) = \sum_{k=1}^{p} \hat{a}_k \phi(i,k) \quad i = 1,2,...,p$$

Short-time LP analysis

- To solve the following equation for the optimum predictor coefficients (the $\hat{a}_k$ s)

$$\phi(i,0) = \sum_{k=1}^{p} \hat{a}_k \phi(i,k) \quad i = 1,2,...,p$$

we have to compute $\phi(i,k)$ and then solve the resulting set of p equations.

- Two standard methods: (Rabiner and Juang, pp103-107)
  - Autocorrelation method
  - Covariance method

Part V: Cepstral analysis

- Short-time speech analysis
- Time-domain speech processing
- Frequency-domain (spectral) processing
- Linear predictive coding (LPC) analysis
- Cepstral analysis

Homomorphic speech processing

- Again, speech is modelled as the output of a linear, time-varying system (linear time-invariant (LTI) in short seg.) excited by either quasi-periodic pulses or random noise.
- The problem of speech analysis is to estimate the parameters of the speech model and to measure their variations with time.
- Since the excitation and impulse response of a LTI system are combined in a convolutional manner, the problem of speech analysis can also been viewed as a problem in separating the components of a convolution, called ”deconvolution”.

$$y[n] = x[n] * h[n]$$

Homomorphic systems for convolution

- The principle of superposition for conventional linear systems:

$$\begin{cases} L[x(n)] = L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)] \\ \qquad\qquad = y_1(n) + y_2(n) = y(n) \\ L[ax(n)] = aL[x(n)] = ay(n) \end{cases}$$

- Homomorphic systems for convolution

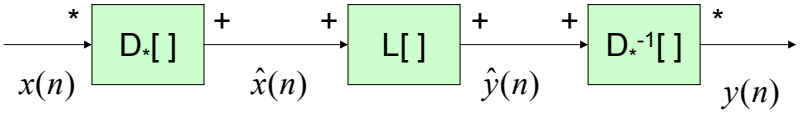
$$H[x(n)] = H[x_1(n) * x_2(n)] = H[x_1(n)] * H[x_2(n)]$$
$$= y_1(n) * y_2(n) = y(n)$$

Homomorphic deconvolution

- Converts a convolution into a sum

$$\begin{cases} y(n) = x(n) * h(n) \\ \hat{y}(n) = \hat{x}(n) + \hat{h}(n) \end{cases}$$

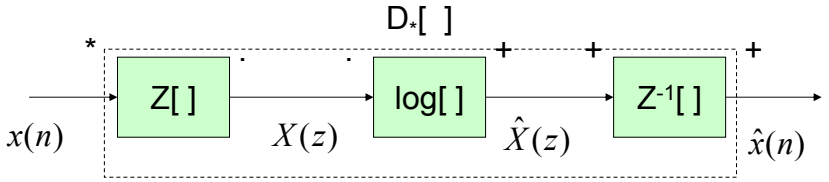
- Canonic form for system for homomorphic deconvolution



$x_1(n) * x_2(n)$ $\hat{x}_1(n) + \hat{x}_2(n)$ $\hat{y}_1(n) + \hat{y}_2(n)$ $y_1(n) + y_2(n)$

The characteristic system

- The characteristic system for homomorphic deconvolution



Cepstral analysis

Observation:

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

taking logarithm of X(z), then

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

$$\text{i.e.,}\ \ \hat{X}(z) = \hat{X}_1(z) + \hat{X}_2(z)$$

←→ $\hat{x}[n] = \hat{x}_1[n] + \hat{x}_2[n]$ in the cepstral domain

So, the two convolved signals are additive in the cepstral domain

Complex cepstrum and real cepstrum

Real cepstrum $c[n]$ is the even part of $\hat{x}[n]$

$$\begin{cases} \hat{x}[n] = \dfrac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{jw})e^{jwn}dw \\ \qquad = \dfrac{1}{2\pi}\int_{-\pi}^{\pi} \log\{X(e^{jw})\}e^{jwn}dw \qquad \text{complex cepstrum} \\ c[n] = \dfrac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{jw})|e^{jwn}dw \qquad \text{cepstrum} \end{cases}$$

- *cep*strum was coined by reversing the first syllable in the word *spec*trum.

Summary

- Short-time speech analysis
- Time-domain processing
- Frequency-domain (spectral) processing
- Linear predictive coding (LPC) analysis
- Cepstral analysis

- Next lecture: Speech Coding and Synthesis