Speech Signal Representations

Berlin Chen

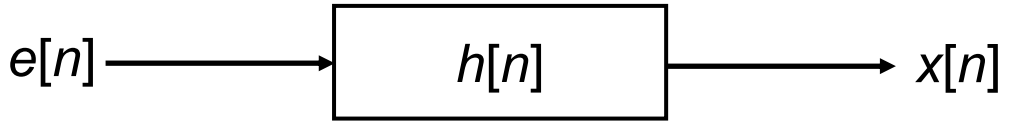
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

1. X. Huang et. al., *Spoken Language Processing*, Chapters 5, 6
2. J. R. Deller et. al., *Discrete-Time Processing of Speech Signals*, Chapters 4-6
3. J. W. Picone, "Signal modeling techniques in speech recognition," *proceedings of the IEEE*, September 1993, pp. 1215-1247
4. L. Rabiner and R.W. Schafer. Introduction to Digital Speech Processing, Chapters 4-6

Source-Filter model

- Source-Filter model: decomposition of speech signals
  - A source passed through a linear time-varying filter
    - But assume that the filter is short-time time-invariant
  - **Source** (excitation): the air flow at the vocal cord (聲帶)
  - **Filter**: the resonances (共鳴) of the vocal tract (聲道) which change over time



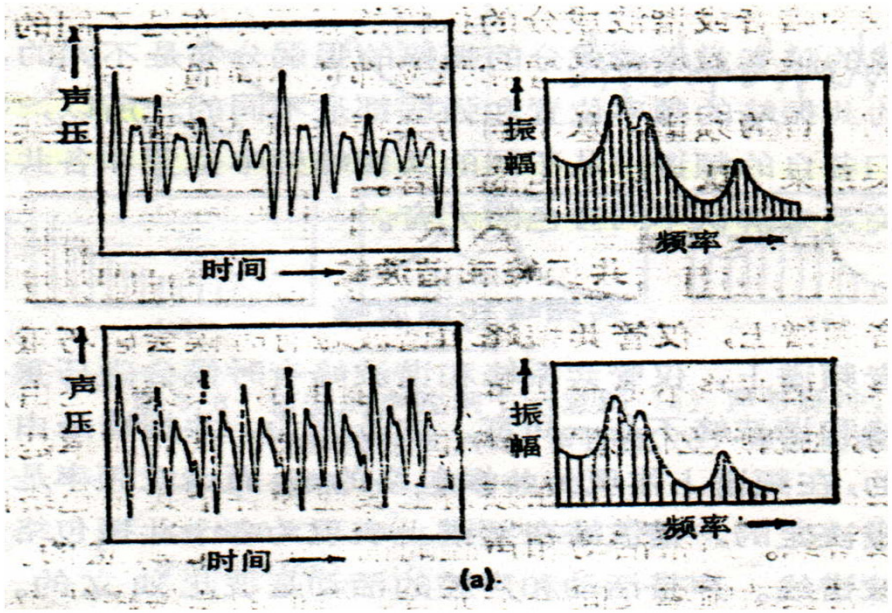
- Once the filter has been estimated, the source can be obtained by passing the speech signal through the inverse filter

Source-Filter model (cont.)

- Phone classification is mostly dependent on the characteristics of the filter (vocal tract)
  - **Speech recognizers** estimate the filter characteristics and ignore the source
    - **Speech Production Model**: *Linear Prediction Coding, Cepstral Analysis*
    - **Speech Perception Model**: *Mel-frequency Cepstrum*
  - **Speech synthesis techniques** use a source-filter model to allow flexibility in altering the pitch and filter
  - **Speech coders** use a source-filter model to allow a low bit rate

Characteristics of the Source-Filter Model

- The characteristics of the vocal tract define the current uttered phoneme
  - Such characteristics are evidenced in the frequency domain by the location of the formants
    - I.e., the peaks given by resonances of the vocal tract



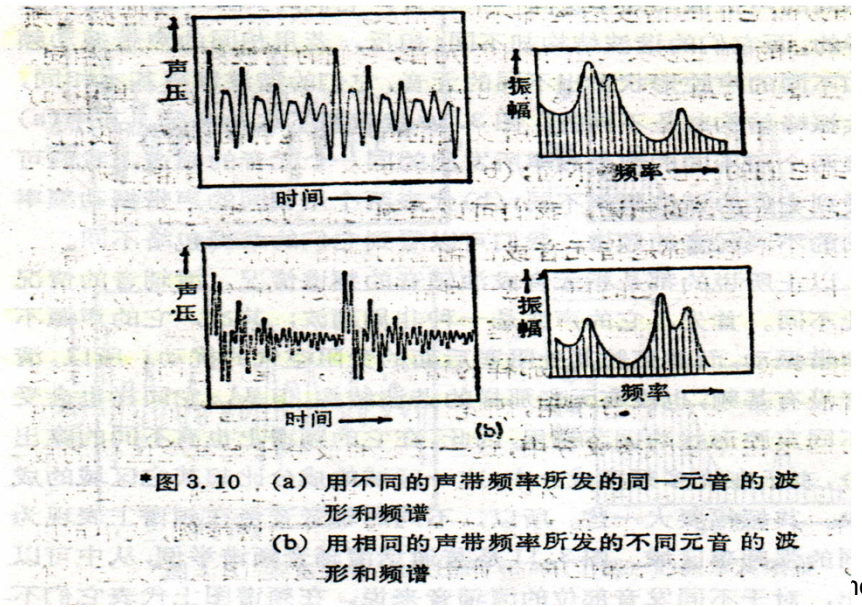
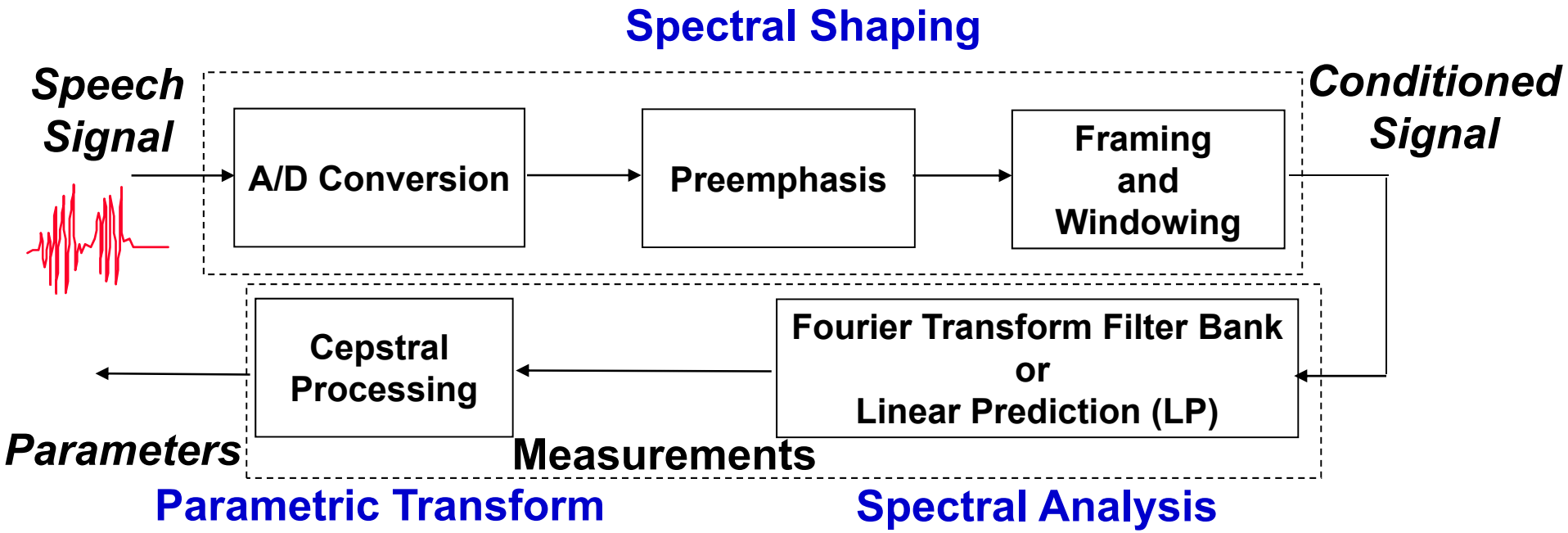


*图 3.10 (a) 用不同的声带频率所发的同一元音的波形和频谱
(b) 用相同的声带频率所发的不同元音的波形和频谱

Main Considerations in Feature Extraction

- **Perceptually Meaningful**
  - Parameters represent salient aspects of the speech signal
  - Parameters are analogous to those used by human auditory system (perceptually meaningful)
- **Robust Parameters**
  - Parameters are more robust to variations in environments such as channels, speakers and transducers
- **Time-Dynamic Parameters**
  - Parameters can capture spectral dynamics, or changes of spectra with time (temporal correlation)
  - Contextual information during articulation

Typical Procedures for Feature Extraction



Spectral Shaping

- A/D conversion
  - Convert the signal from a sound pressure wave to a digital signal
- Digital Filtering (e.g., “pre-emphasis”)
  - Emphasize important frequency components in the signal
- Framing and Windowing
  - Perform short-term (short-time) processing

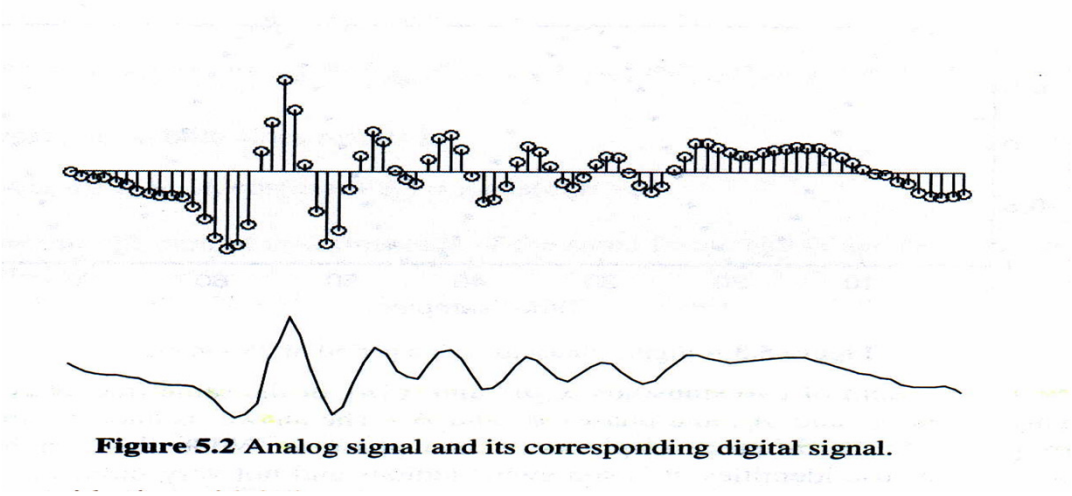


Figure 5.2 Analog signal and its corresponding digital signal.

Spectral Shaping (cont.)

- Sampling Rate/Frequency and Recognition Error Rate

Sampling Rate	Relative Error-Rate Reduction
8 kHz	Baseline
11 kHz	+10%
16 kHz	+10%
22 kHz	+0%

Table 9.1 Relative error rate reduction with different sampling rates. The reduction is relative to that of the preceding row.

E.g., Microphone Speech Mandarin Syllable Recognition
Accuracy: 67% (16KHz)
Accuracy: 63% (8KHz)
⇒Error rate reduction
4/37=10.8%

Spectral Shaping (cont.)

- Problems for A/D Converter
  - Frequency distortion (50-60-Hz hum)
  - Nonlinear input-output distortion
    - Example:
      - Frequency response of a typical telephone grade A/D converter
      - The sharp attenuation of low frequency and high frequency response causes problem for subsequent parametric spectral analysis algorithms
- The Most Popular Sampling Frequency
  - Telecommunication: 8KHz
  - Non-telecommunication: 10~16KHz

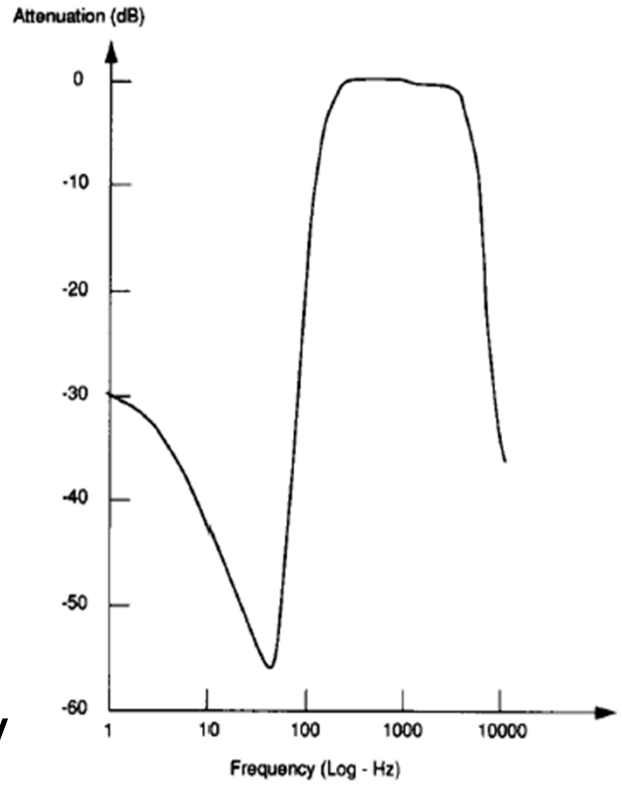


Fig. 3. The frequency response of a typical telephone grade A/D converter is shown.

Pre-emphasis

- A high-pass filter is used
  - Most often executed by using Finite Impulse Response filters (FIRs)
  - Normally an one-coefficient digital filter (called **pre-emphasis filter**) is used

$$H_{pre}(z) = \sum_{k=0}^{N_{pre}} -a_{pre}(k) z^{-k} \qquad (1)$$

$$H_{pre}(z) = 1 - a_{pre} z^{-1} \qquad (2)$$

Z-transform Representation

$$H(z) = \frac{Y(z)}{X(z)} = 1 - az^{-1}$$

$$\Rightarrow Y(z) = X(z) - az^{-1} X(z)$$

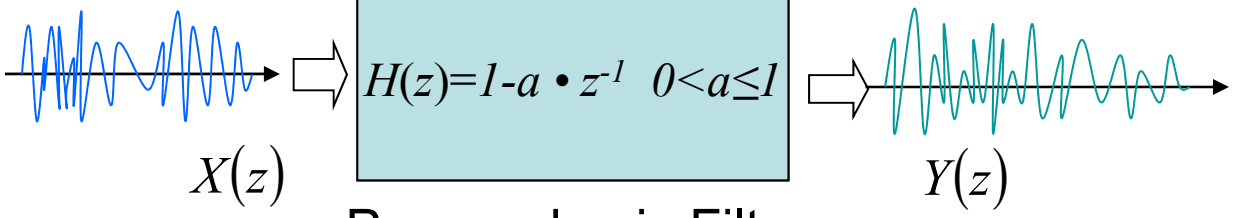
Notice that the Z transform of $ax[n-1]$

$$= \sum_{n=-\infty}^{n=\infty} ax[n-1] z^{-n} = \sum_{n'=-\infty}^{n'=\infty} ax[n'] z^{-(n'+1)}$$

$$= az^{-1} \sum_{n'=-\infty}^{n'=\infty} x[n'] z^{-n'} = az^{-1} X(z)$$

$$\Rightarrow y[n] = x[n] - ax[n-1]$$

Speech signal $x[n]$ $h[n]$ $y[n] = x'[n] = x[n] - ax[n-1]$

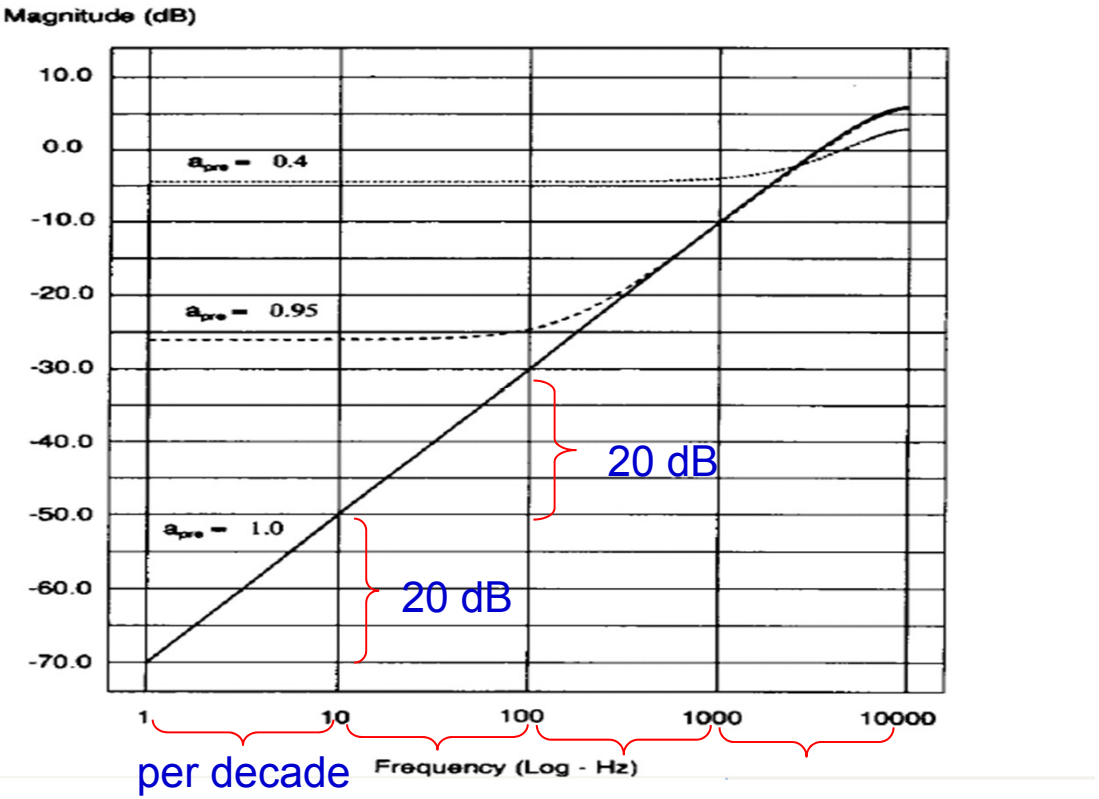


$X(z)$ $Y(z)$

Pre-emphasis Filter

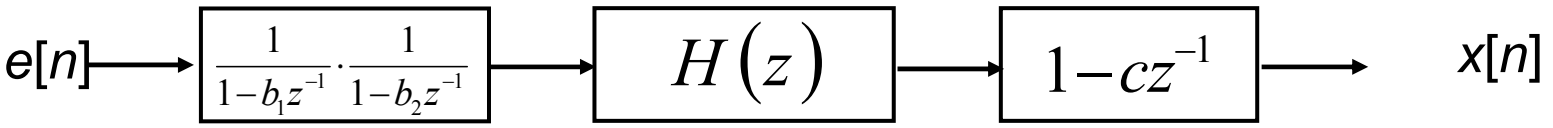
Pre-emphasis (cont.)

- Implementation and the corresponding effect
  - Values close to 1.0 that can be efficiently implemented in fixed point hardware are most common (most common is around 0.95)
  - Boost the spectrum about 20 dB per decade



Pre-emphasis: Why?

- Reason 1: Physiological Characteristics
  - The component of the glottal signal can be modeled by a simple two-real-pole filter whose poles are near z=1
  - The lip radiation characteristic, with its zero near z=1, tends to cancel the spectral effects of one of the glottal pole
    - By introducing a second zero near z=1 (pre-emphasis), we can eliminate effectively the larynx and lips spectral contributions
  - Analysis can be asserted to be seeking the parameters corresponding to the vocal tract only



glottal signal/ larynx

vocal tract

lips

$$\frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}$$

Pre-emphasis: Why? (cont.)

- Reason 2: Prevent Numerical Instability
  - If the speech signal is dominated by low frequencies, it is highly predictable and a large LP model will result in an ill-conditioned autocorrelation matrix
- Reason 3 : Physiological Characteristics Again
  - Voiced sections of the speech signal naturally have a negative spectral slope (attenuation) of approximately 20 dB per decade due to physiological characteristics of the speech production system
  - High frequency formants have small amplitude with respect to low frequency formants. A pre-emphasis of high frequencies is therefore required to obtain similar amplitude for all formants

Pre-emphasis: Why? (cont.)

- Reason 4 :
  - Hearing is more sensitive above the 1 kHz region of the spectrum

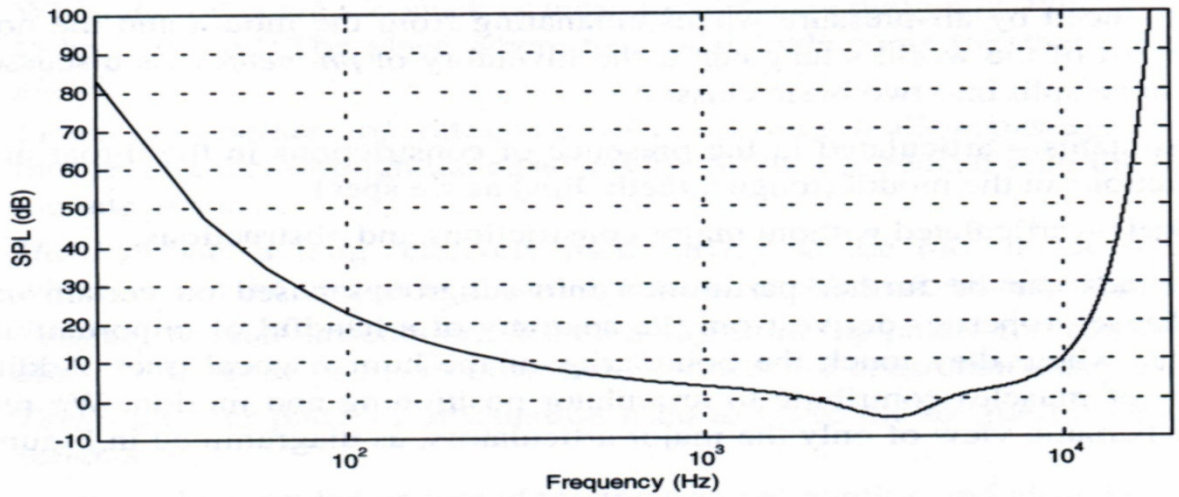
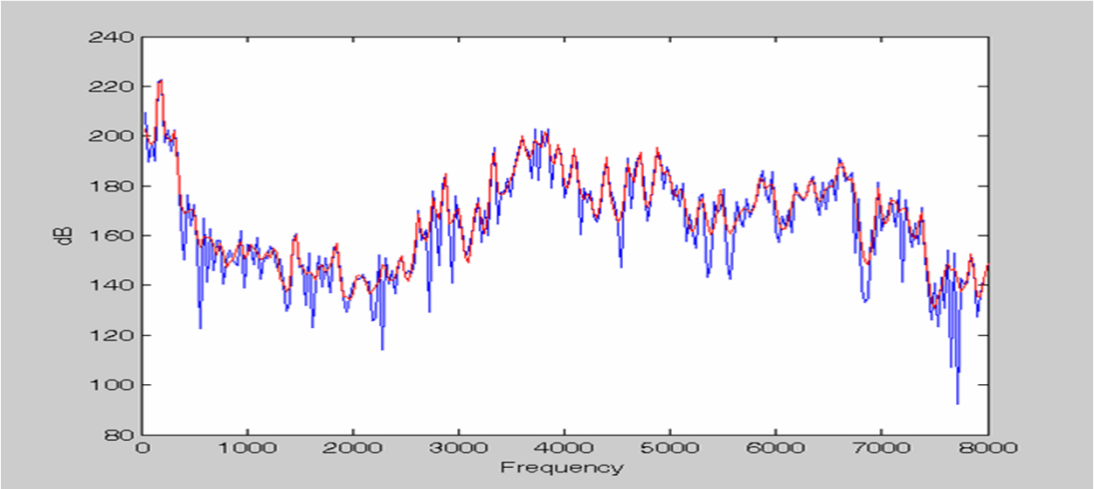
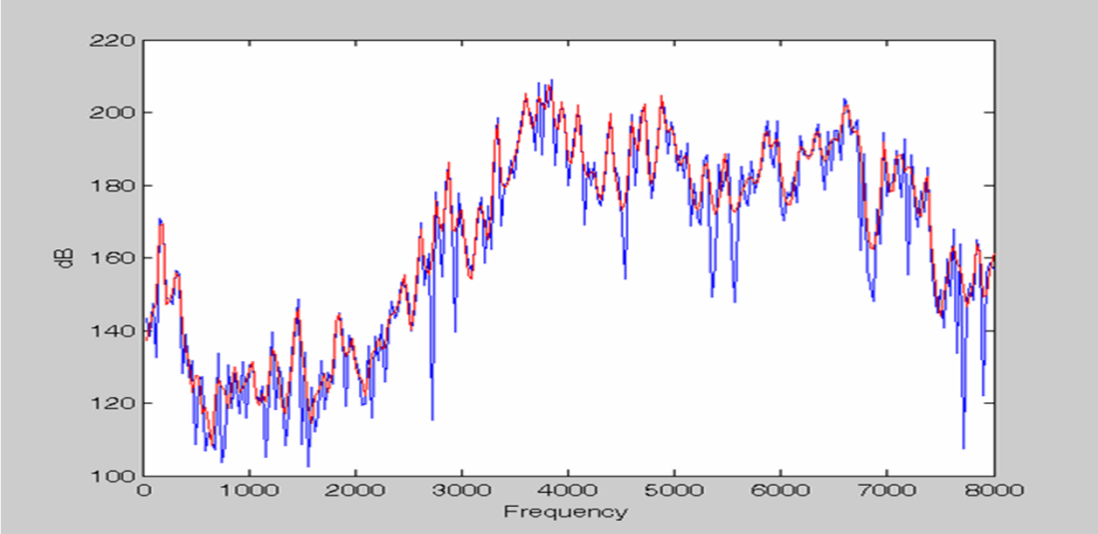


Figure 2.3 The sound pressure level (SPL) level in dB of the absolute threshold of hearing as a function of frequency. Sounds below this level are inaudible. Note that below 100 Hz and above 10 kHz this level rises very rapidly. Frequency goes from 20 Hz to 20 kHz and is plotted in a logarithmic scale from Eq. (2.3).

Pre-emphasis: An Example



No Pre-emphasis



Pre-emphasis

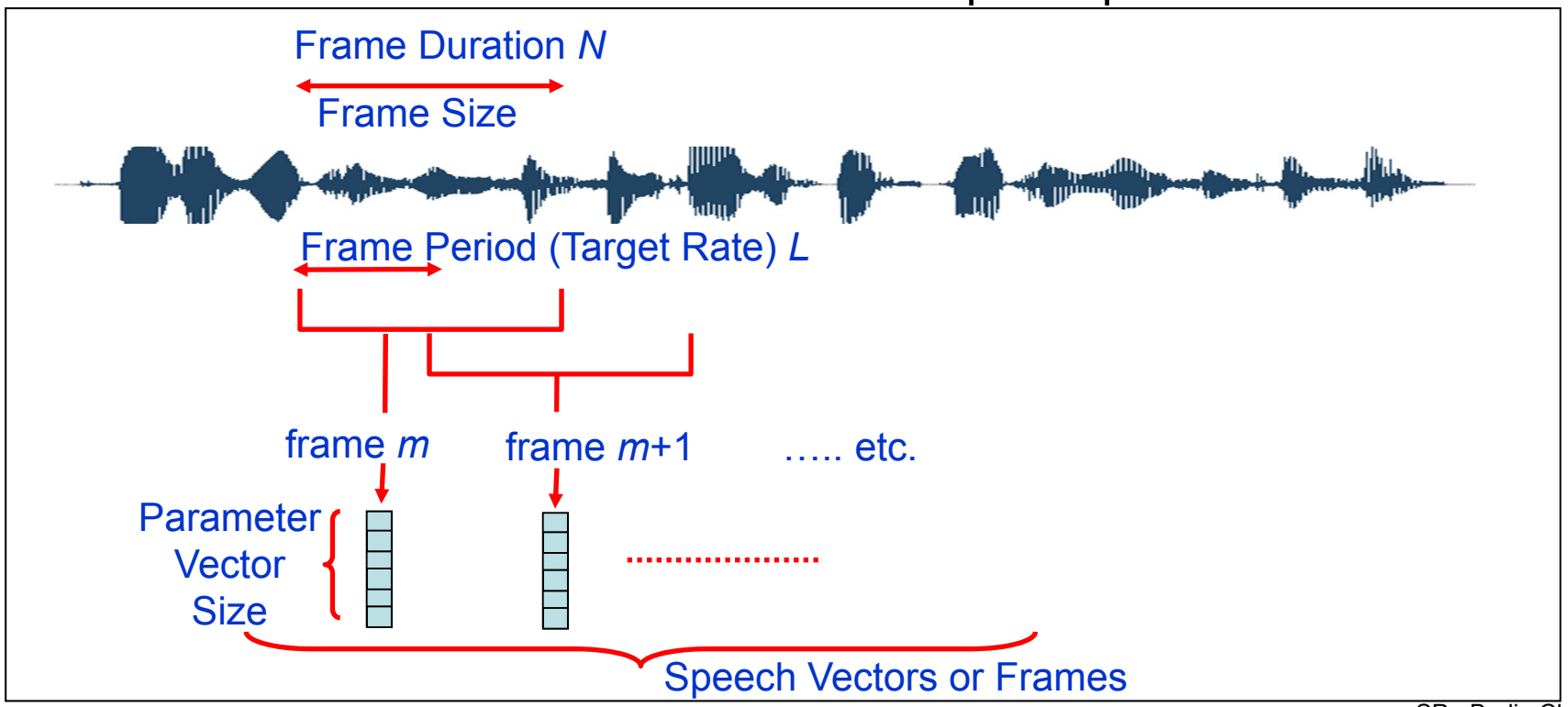
$a_{pre}=0.975$

Framing and Windowing

- Framing: decompose the speech signal into a series of overlapping frames
  - Traditional methods for spectral evaluation are reliable in the case of a **stationary signal** (i.e., a signal whose statistical characteristics are invariant with respect to time)
    - Imply that the region is short enough for the behavior (periodicity or noise-like appearance) of the signal to be approximately constant
    - Phrased another way, the speech region has to be short enough so that it can reasonably be assumed to be stationary
    - ***stationary*** in that region: i.e., the signal characteristics (whether periodicity or noise-like appearance) are uniform in that region

Framing and Windowing (cont.)

- Terminology Used in Framing
  - Frame Duration (*N*): the length of time over which a set of parameters is valid. Frame duration ranges between 10 ~ 25 ms
  - Frame Period (*L*): the length of time between successive parameter calculations ("Target Rate" used in HTK)
  - Frame Rate: the number of frames computed per second



Framing and Windowing (cont.)

- Windowing : a window, say $w[n]$, is a real, finite length sequence used to select a desired frame of the original signal, say $x_m[n]$
  - Most commonly used windows are symmetric about the time *(N-1)/2*
    N is the window duration

$$\tilde{x}_m[n] = x[m \cdot L + n], \qquad n = 0,1,...,N-1, \quad m = 0,1,...,M-1$$

Framed signal

$$x_m[n] = \tilde{x}_m[n] w[n], \qquad 0 \le n \le N-1$$

Multiplied with the window function

  - Frequency response:

$$X_m(k) = \tilde{X}_m(k) * W(k), \quad *: \text{convolution}$$

Frequency Response

  - Ideally, $w[n]=1$ for all n, whose frequency response is just an impulse
    - This is invalid since the speech signal is stationary only within short time intervals

Framing and Windowing (cont.)

- Windowing (Cont.)
  - Rectangular window ($w[n]$=1 for $0 \le n \le N-1$):
    - Just extract the frame part of signal without further processing
    - Whose frequency response has high side lobes
  - **Main lobe**: spreads out in a wider frequency range in the narrow band power of the signal, and thus reduces the local frequency resolution
  - **Side lobe**: swaps energy from different and distant frequencies of $x_m[n]$, which is called *leakage* or *spectral leakage*

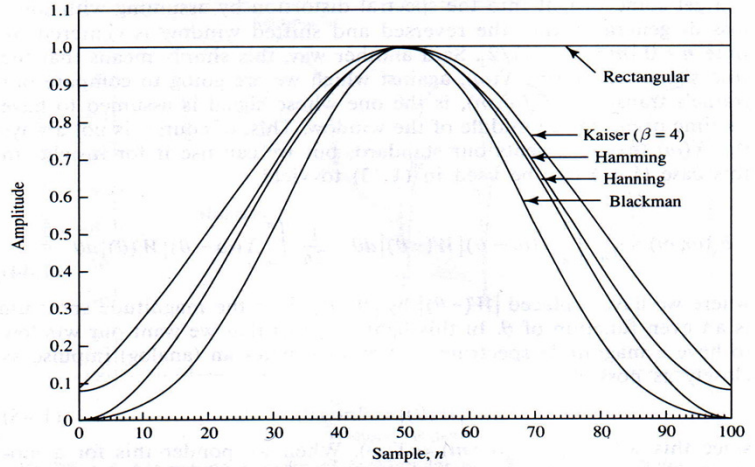


FIGURE 1.3. Definitions and example time plots for the rectangular, Kaiser, Hamming, Hanning, and Blackman windows. All plots are for window lengths $N = 101$, and for the Kaiser window, $\beta = 4$.

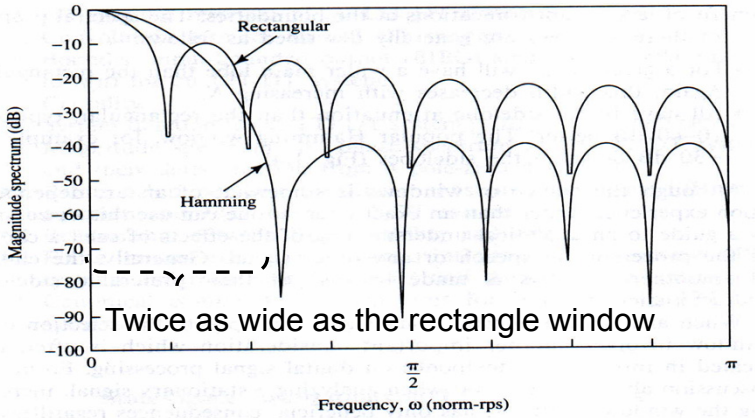
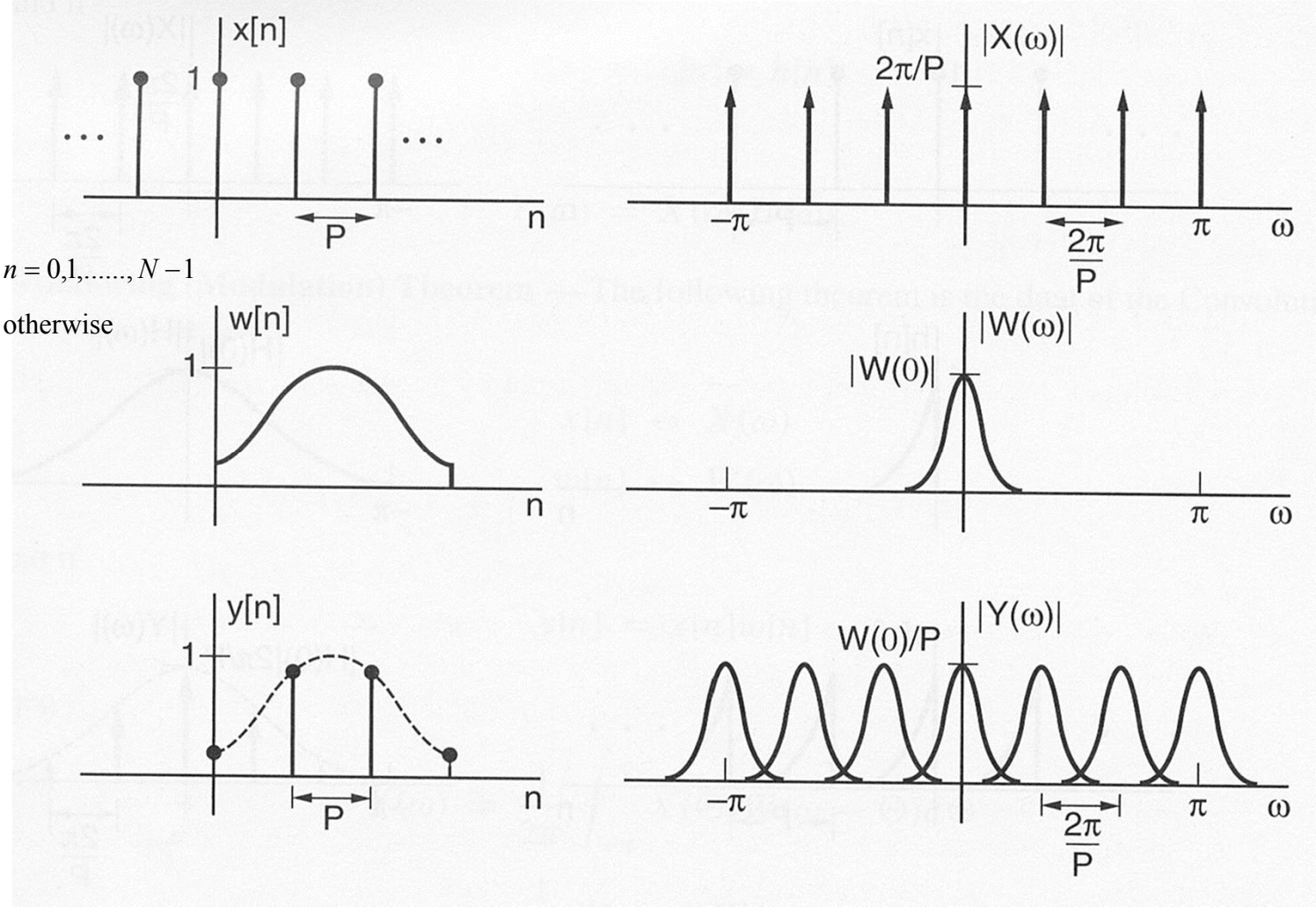


FIGURE 1.4. Magnitude spectra of rectangular and Hamming windows. Window length $N = 16$ is used in each case for clarity. Note that the nominal "bandwidth" (width of main lobe) is $2\pi/N = \pi/8$ for the rectangular case and about twice that for the Hamming. The sidelobe attenuation for the Hamming, however, is 20 dB better outside the passband.

Framing and Windowing (cont.)

$$x[n] = \sum_{k=-\infty}^{\infty} \delta[n - kP]$$

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & n = 0,1,.....,N-1 \\ 0 & \text{otherwise} \end{cases}$$



Framing and Windowing (cont.)

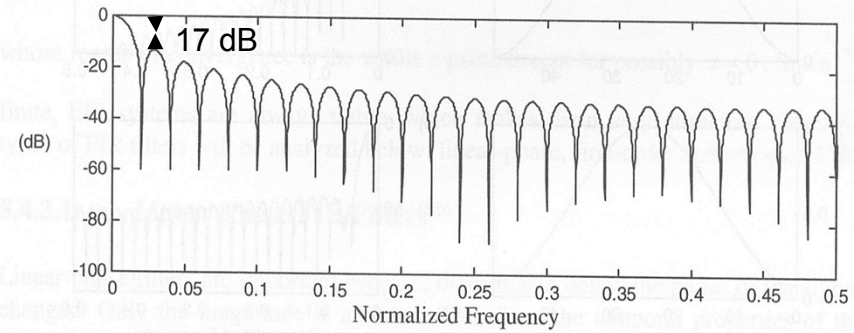


Figure 5.19 Frequency response (magnitude in dB) of the rectangular window with $N = 50$, which is a digital sinc function.

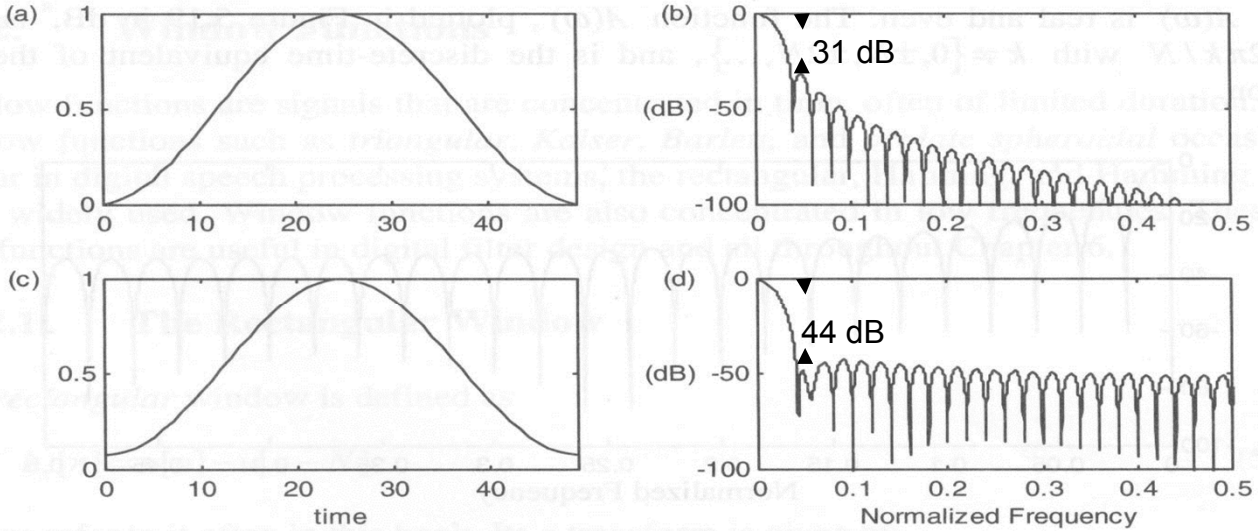


Figure 5.20 (a) Hanning window and (b) the magnitude of its frequency response in dB; (c) Hamming window and (d) the magnitude of its frequency response in dB for $N = 50$.

Framing and Windowing (cont.)

- For a designed window, we wish that
  - A narrow bandwidth main lobe
  - Large attenuation in the magnitudes of the sidelobes

  However, this is a trade-off!

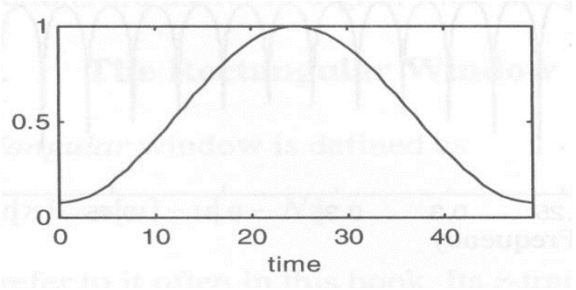
Notice that:

1. A narrow main lobe will resolve the sharp details of $\widetilde{X}_m(k)$ (the frequency response of the framed signal) as the convolution proceeds in frequency domain
2. The attenuated sidelobes prevents "noise" from other parts of the spectrum from corrupting the true spectrum at a given frequency

Framing and Windowing (cont.)

- The most-used window shape is the Hamming window, whose impulse response is a raised cosine impulse

$$w[n]=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & n=0,1,.....,N-1\\ 0 & \text{otherwise}\end{cases}$$



$$w[n]=\begin{cases}(1-\alpha)-\alpha\cos\left(\dfrac{2\pi n}{N-1}\right), & n=0,1,.....,N-1\\ 0 & \text{otherwise}\end{cases}$$

Generalized Hamming Window

Framing and Windowing (cont.)

- Male Voiced Speech

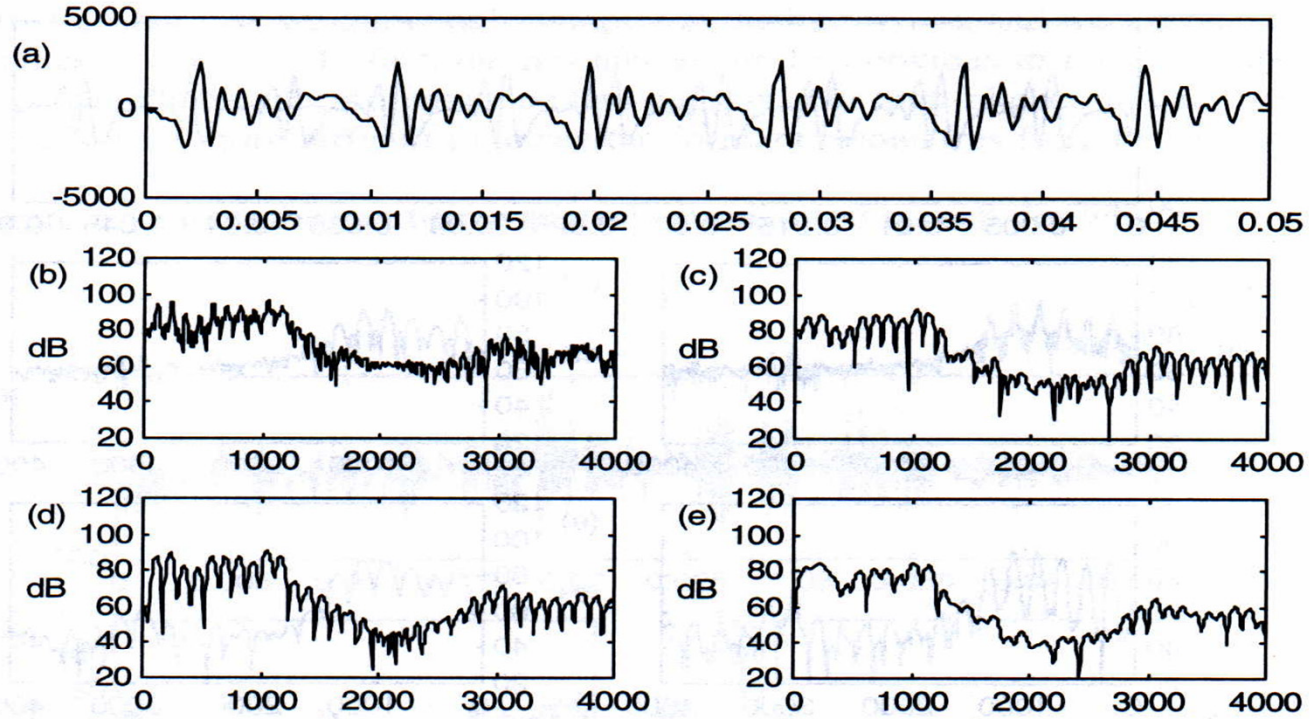


Figure 6.3 Short-time spectrum of male voiced speech (vowel /ah/ with local pitch of 110Hz): (a) time signal, spectra obtained with (b) 30 ms rectangular window and (c) 15 ms rectangular window, (d) 30 ms Hamming window, (e) 15 ms Hamming window. The window lobes are not visible in (e), since the window is shorter than 2 times the pitch period. Note the spectral leakage present in (b).

Note: The longer the window during the finer local frequency resolution !

Framing and Windowing (cont.)

- Female Voiced Speech

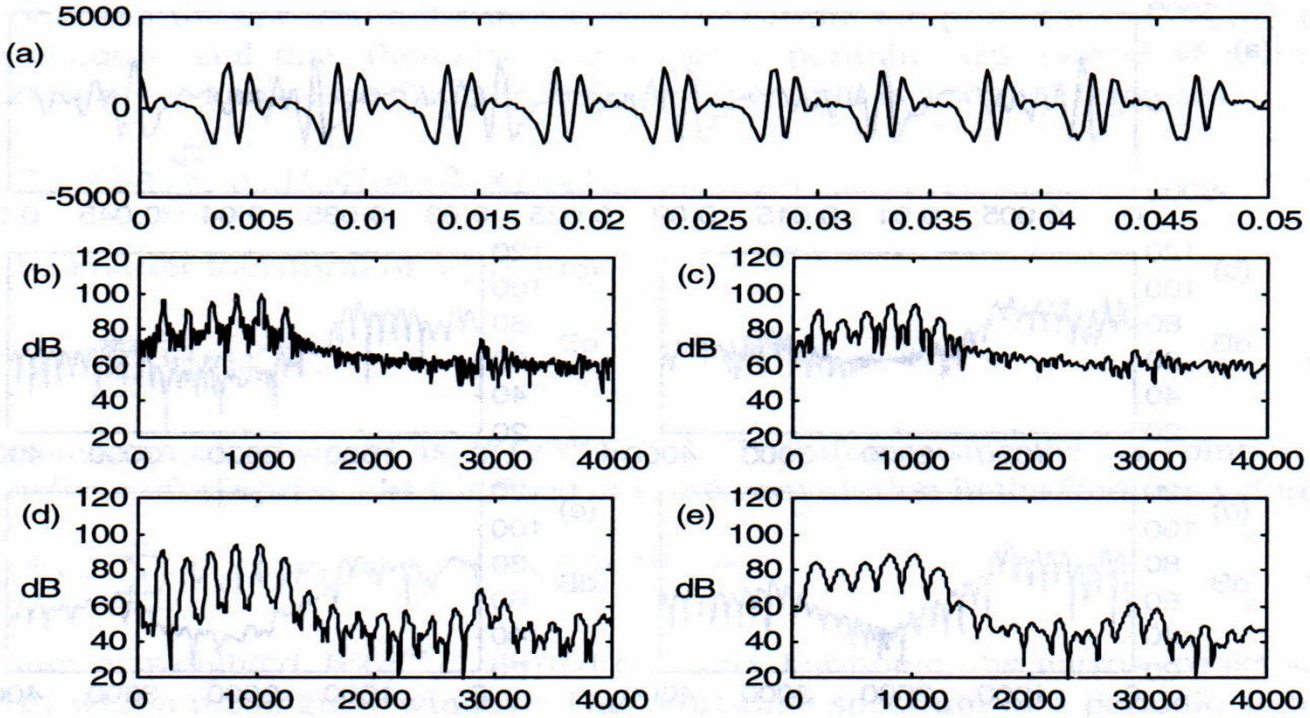


Figure 6.4 Short-time spectrum of female voiced speech (vowel /aa/ with local pitch of 200Hz): (a) time signal, spectra obtained with (b) 30 ms rectangular window and (c) 15 ms rectangular window, (d) 30 ms Hamming window, (e) 15 ms Hamming window. In all cases the window lobes are visible, since the window is longer than 2 times the pitch period. Note the spectral leakage present in (b) and (c).

Framing and Windowing (cont.)

- Unvoiced Speech

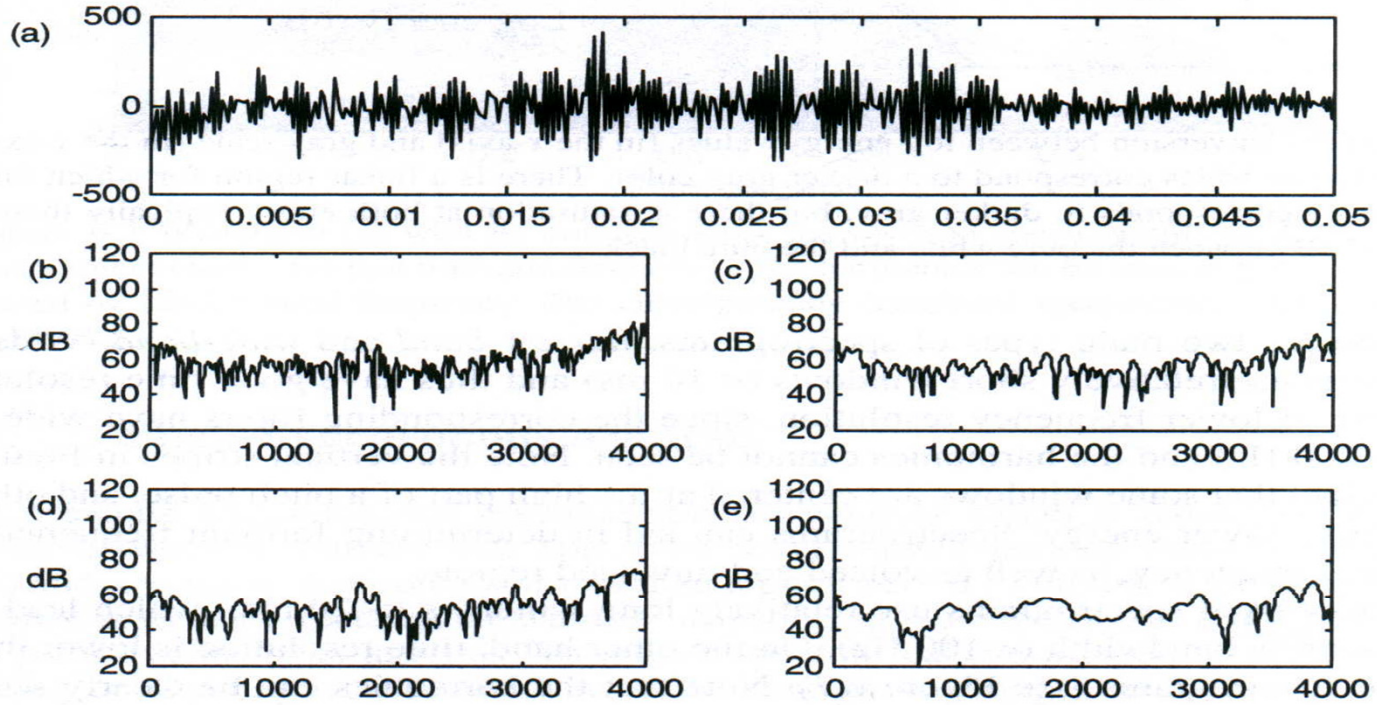


Figure 6.5 Short-time spectrum of unvoiced speech: (a) time signal, (b) 30 ms rectangular window, (c) 15 ms rectangular window, (d) 30 ms Hamming window, (e) 15 ms Hamming window.

Short-Time Fourier Analysis

- Spectral Analysis
  - Notice that the response for each frequency is not completely uncorrelated due to the windowing operation

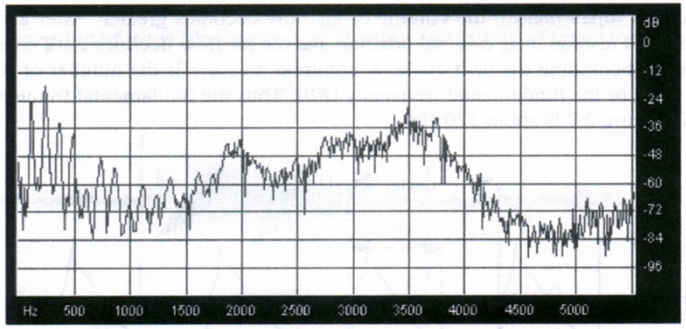


Figure 2.8 A spectral analysis of the vowel */iy/*, showing characteristically uneven distribution of energy at different frequencies.

- Spectrogram Representation
  - A spectrogram of a time signal is a two-dimensional representation that displays *time* in its horizontal axis and *frequency* in its vertical axis
  - A gray scale is typically used to indicate the energy at each point (*t,f*)
    - "white": low energy, "black": high energy

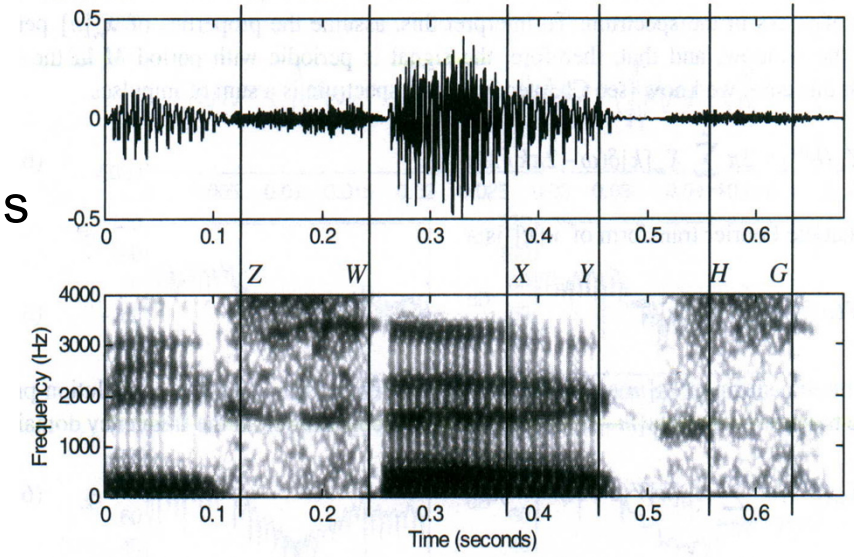
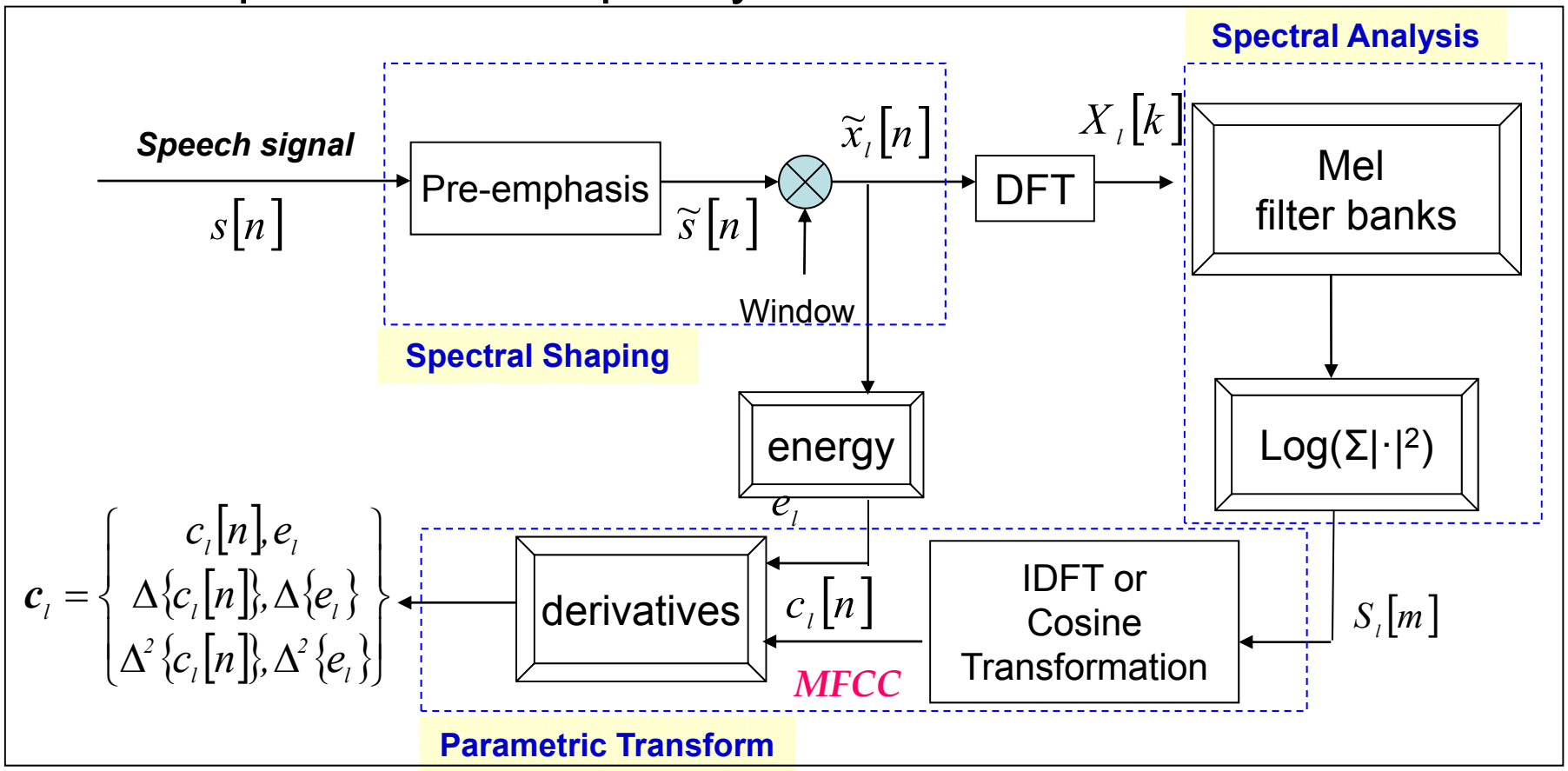


Figure 6.2 (a) Waveform with (b) its corresponding wideband spectrogram. Darker areas mean higher energy for that time and frequency. Note the vertical lines spaced by pitch periods.

Mel-Frequency Cepstral Coefficients (MFCC)

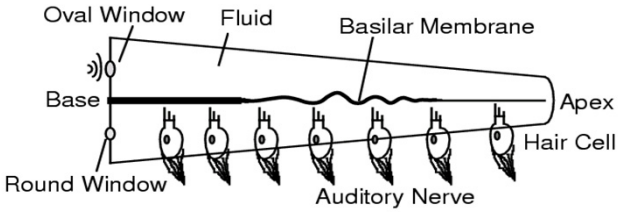
- Most widely used in the speech recognition
- Has generally obtained a better accuracy and a minor computational complexity



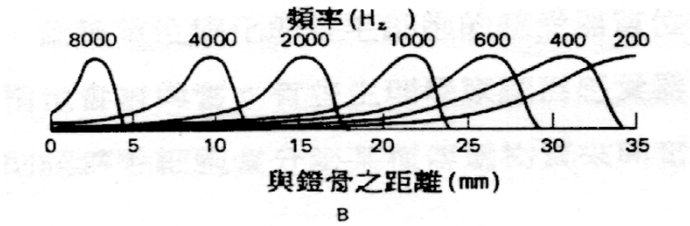
S.B. Davis, P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. on Acoustics, Speech & Signal Processing 28(4), 1980

Mel-Frequency Cepstral Coefficients (cont.)

- Characteristics of MFCC
  - Auditory-like frequency
    - Mel spectrum
  - Filter (critical)-band soothing
    - Sum of weighted frequency bins
  - Amplitude warping
    - Logarithmic representation of filter bank outputs
  - Feature decorrelation and dimensionality reduction
    - Projection on the cosine basis

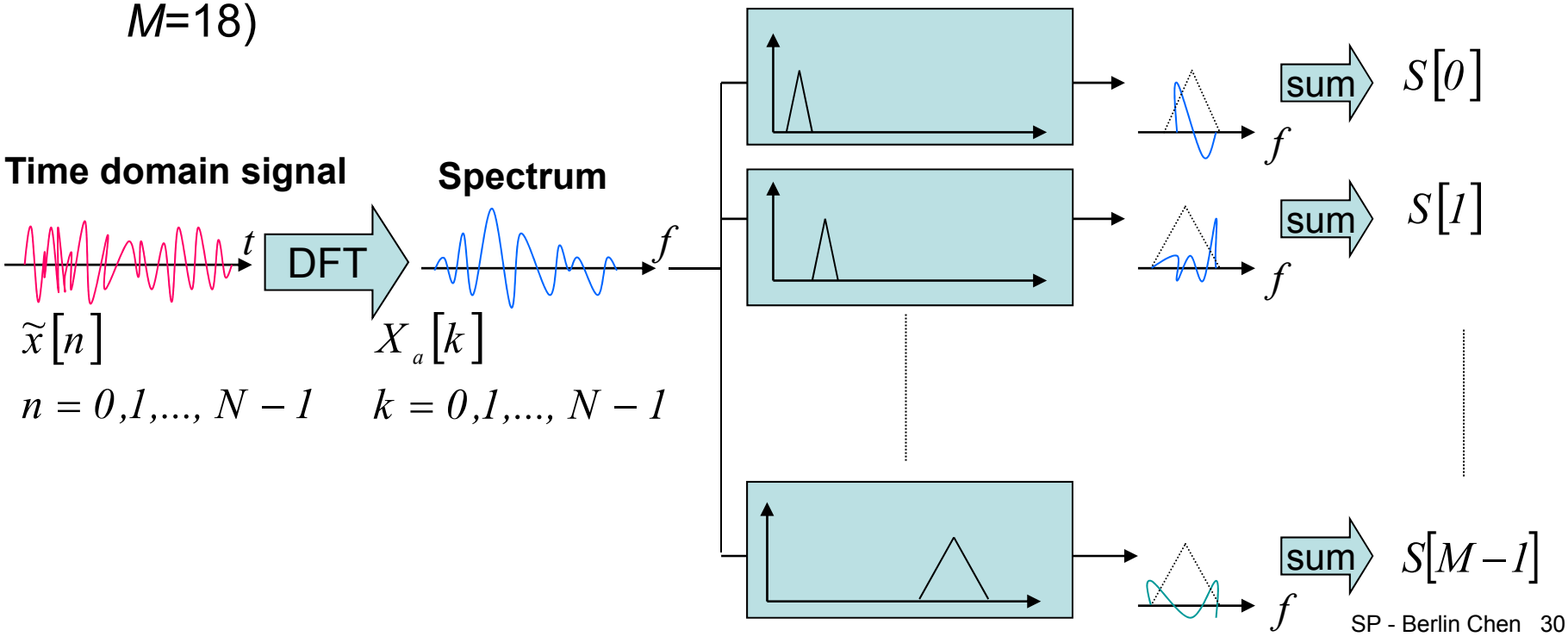


Adopted from Kumar's Ph.D. Thesis



DFT and Mel-filter-bank Processing

- For each frame of signal (*N* points, e.g., *N*=512)
  - The Discrete Fourier Transform (DFT) is first performed to obtain its spectrum (*N* points, for example *N*=512)
  - The spectrum is then processed by a bank of filters according to Mel scale, and the each filter output is the sum of its filtered spectral components (*M* filters, and thus *M* points, for example *M*=18)

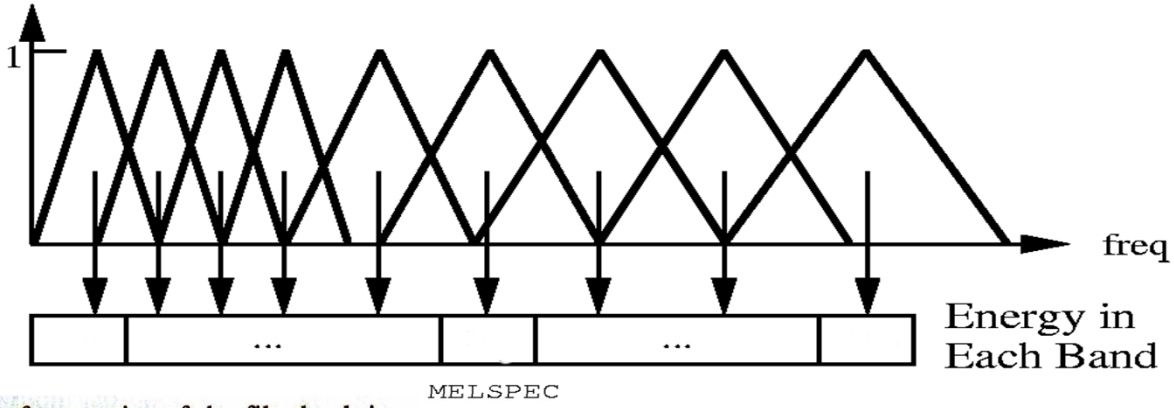


Filter-bank Processing

- Mel-filter-bank

$$H_{m-1}[f_k] = \frac{f[m] - f_k}{f[m] - f[m-1]}$$

$$H_m[f_k] = \frac{f_k - f[m-1]}{f[m] - f[m-1]}$$



Let's define f_l and f_h to be the lowest and highest frequencies of the filterbank in Hz, F_s the sampling frequency in Hz, M the number of filters, and N the size of the FFT. The boundary points $f[m]$ are uniformly spaced in the mel-scale:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \qquad (6.142)$$

where the mel-scale B is given by Eq. (2.6), and B^{-1} is its inverse

$$B^{-1}(b) = 700\left(\exp(b/1125) - 1\right) \qquad (6.143)$$

Mel-Scale Filter Bank

$$B(f) = 1125\ln(1 + f/700)$$

$$\sum_{p=0}^{M-1} H'_p[k] = 1$$

A filterbank with M filters

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], \quad 0 < m \le M$$

approximate homomorphic transform (more robust to noise and spectral estimation errors)

or $S[m] = \sum_{k=0}^{N-1} \ln\left(|X_a[k]|^2 H_m[k]\right) \quad 0 < m \le M$

homomorphic transform

HTK use such a configuration

Filter-bank Processing (cont.)

- An Example

Original $\log_{10}(50 + 50 + 50) = 2.1761$

Corrupted (I) $\log_{10}(1000 + 50 + 50) = 3.0414$

Corrupted (II) $\log_{10}(0.1 + 50 + 50) = 2.0004$

Original $\log_{10}(50) + \log_{10}(50) + \log_{10}(50) = 5.0969$

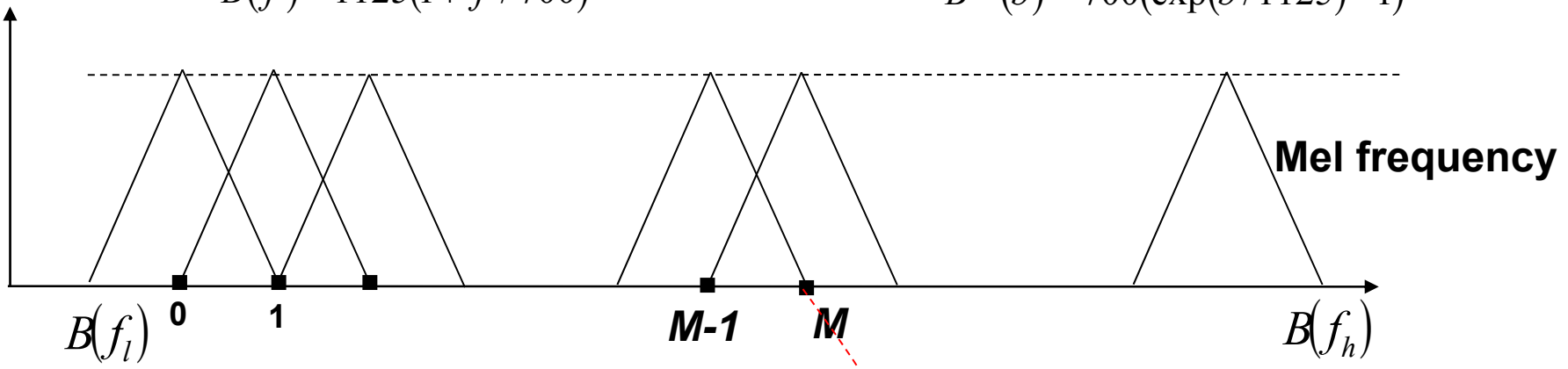
Corrupted (I) $\log_{10}(1000) + \log_{10}(50) + \log_{10}(50) = 6.3979$

Corrupted (II) $\log 10(0.1) + \log_{10}(50) + \log_{10}(50) = 2.3979$

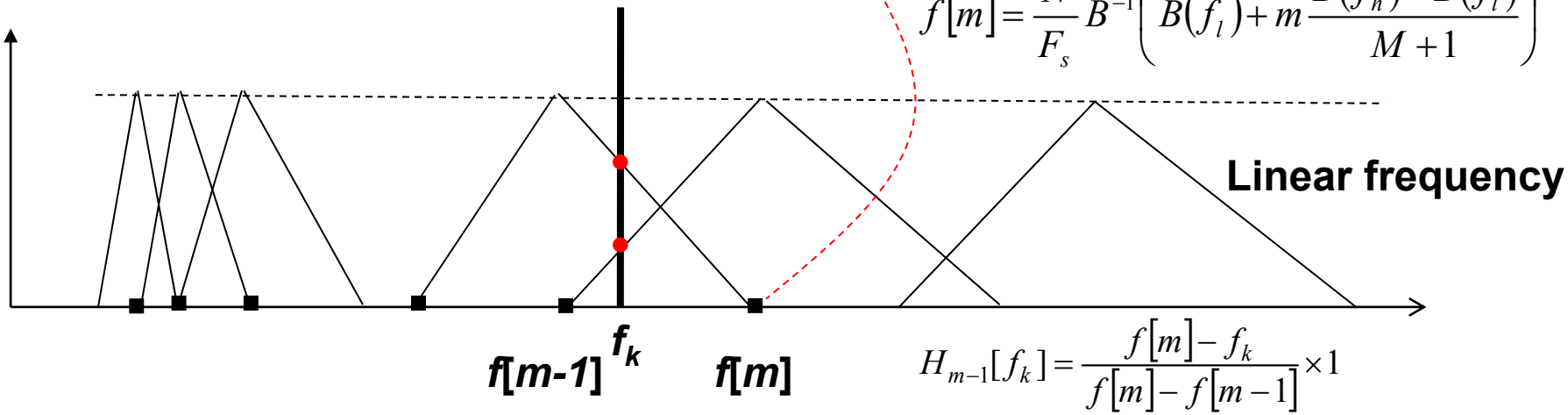
Filter-bank Processing (cont.)

$$B(f) = 1125(1 + f/700)$$

$$B^{-1}(b) = 700(\exp(b/1125) - 1)$$



$$f[m] = \frac{N}{F_s} B^{-1}\left(B(f_l) + m \frac{B(f_h) - B(f_l)}{M+1} \right)$$



$$H_{m-1}[f_k] = \frac{f[m] - f_k}{f[m] - f[m-1]} \times 1$$

$$H_m[f_k] = \frac{f_k - f[m-1]}{f[m] - f[m-1]} \times 1$$

Filter-bank Processing: Why?

- The filter-bank processing simulates human ear processing
  - Center frequency of each filter
    - The position of maximum displacement along the basilar membrane for stimuli such as pure tone is proportional to the *logarithm* of the frequency of the tone
  - Bandwidth
    - Frequencies of a complex sound within a certain bandwidth of some nominal frequency cannot be individually identified
    - When one of the components of this sound falls outside this bandwidth, it can be individually distinguished
    - This bandwidth is referred to as the critical bandwidth
    - A critical bandwidth is nominally 10% to 20% of the center frequency of the sound

Filter-bank Processing: Why? (cont.)

- For speech recognition purpose :
  - Filters are non-uniformly spaced along the frequency axis
  - The part of the spectrum below 1kHz is processed by more filter banks
    - This part contains more information on the vocal tract such as the first formant
  - Non-linear frequency analysis is also used to achieve frequency/time resolution
    - Narrow band-pass filters at low frequencies enables harmonics to be detected
    - Longer bandwidth at higher frequencies allows for higher temporal resolution of bursts (?)

Filter-bank Processing: Why? (cont.)

- The most-used two warped frequency scale : Bark scale and Mel scale

Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

Homomorphic Transformation
Cepstral Processing

- A homomorphic transform $D(\cdot)$ is a transform that converts a convolution into a sum

$$x[n] = e[n] * h[n]$$
$$\hat{x}[n] = D(x[n]) = \hat{e}[n] + \hat{h}[n]$$

$$\hat{h}[n] \approx 0 \quad \text{for} \quad n \geq L$$
$$\hat{e}[n] \approx 0 \quad \text{for} \quad n < L$$

$x(n)=e(n)*h(n)$ ➔$X(\omega)=E(\omega)H(\omega)$
➔$|X(\omega)|=|E(\omega)||H(\omega)|$➔$\log|X(\omega)|=\log|E(\omega)|+\log|H(\omega)|$

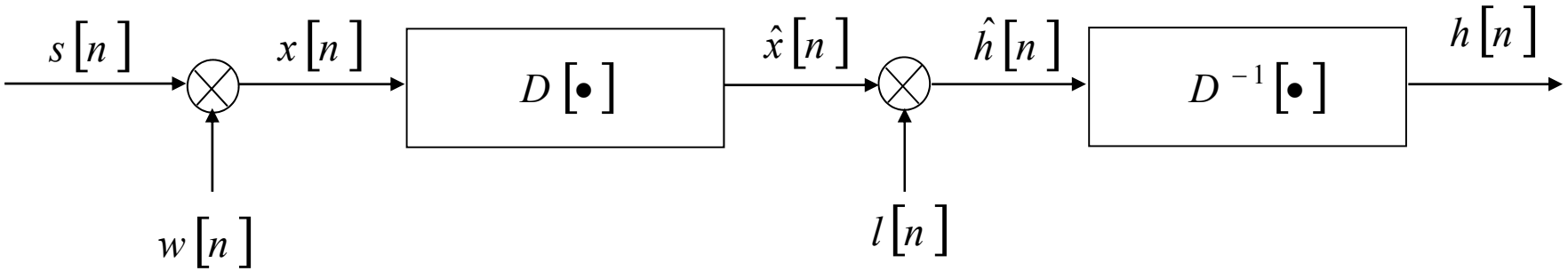
- Cepstrum is regarded as one homomorphic function (filter) that allow us to separate the source (excitation) from the filter for speech signal processing
  - We can find a value L such that
    - The cepstrum of the filter
    - The cepstrum of the excitation

could be separated

Cepstrum is an anagram (回文構詞) of spectrum

Homomorphic Transformation
Cepstral Processing (cont.)

$$s[n] = e[n] * h[n]$$



$$l[n] = \begin{cases} 1 & |n| < N \\ 0 & |n| \geq N \end{cases}$$

liftering operation

Source-Filter Separation via Cepstrum (1/3)

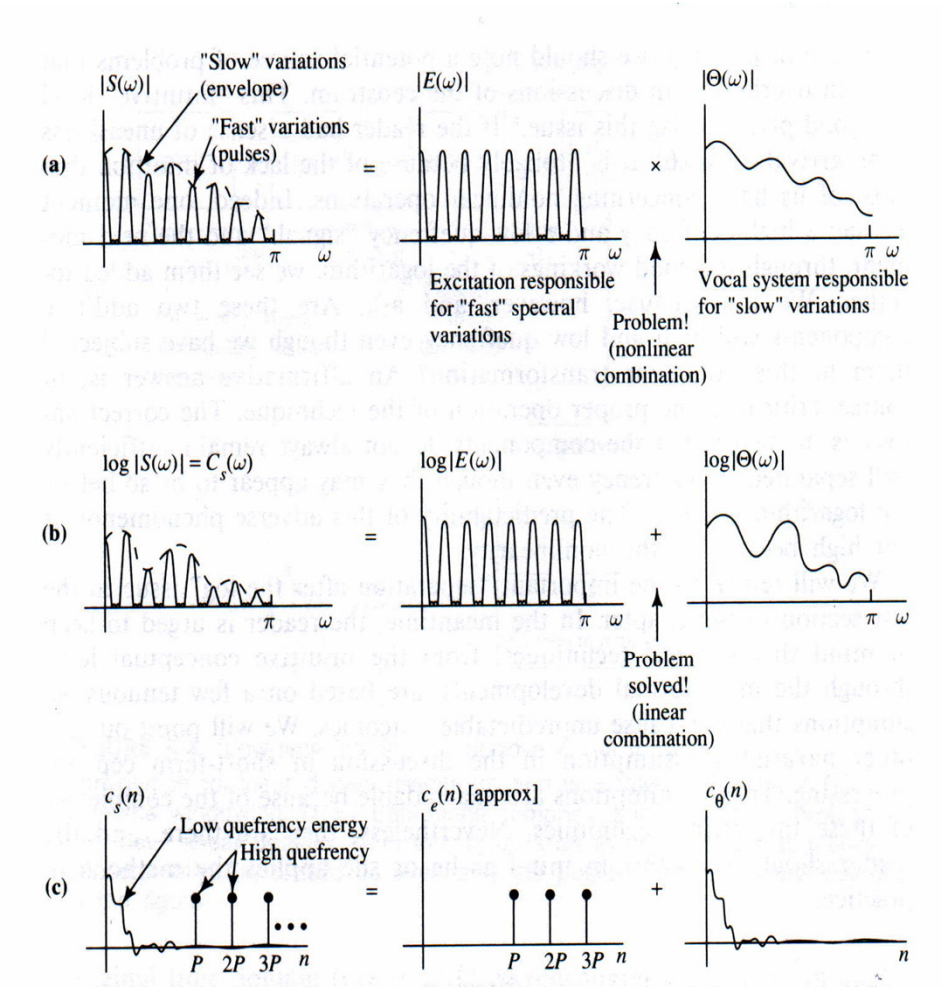


FIGURE 6.3. The motivation behind the RC, and some of the accompanying vocabulary. (a) In the speech magnitude spectrum, $|S(\omega)|$, two components can be identified: a "slowly varying" part (envelope) due to the speech system, $|\Theta(\omega)|$, and a "quickly varying" part due to the excitation, $|E(\omega)|$. These components are combined by addition. Their time domain counterparts, $\theta(n)$ and $e(n)$, are convolved. (b) Once the logarithm of the spectral magnitude is taken, the two convolved signal components, $\theta(n)$ and $e(n)$, have additive correlates in the new "signal," $C_s(\omega)$. The former corresponds to a slowly varying ("low-quefrency") component of $C_s(\omega)$, and the latter to a quickly varying ("high-quefrency") component. (c) When the IDTFT is taken, the slowly varying part yields a "cepstral" component at low quefrencies (smaller values on the time axis), and the component with fast variations results in a "cepstral" component at high quefrencies (larger values on the time axis). The low-quefrency part of the cepstrum therefore represents an approximation to the cepstrum of the vocal system impulse response, $c_\theta(n)$, and the high-quefrency part corresponds to the cepstrum of the excitation, $c_e(n)$.

Source-Filter Separation via Cepstrum (2/3)

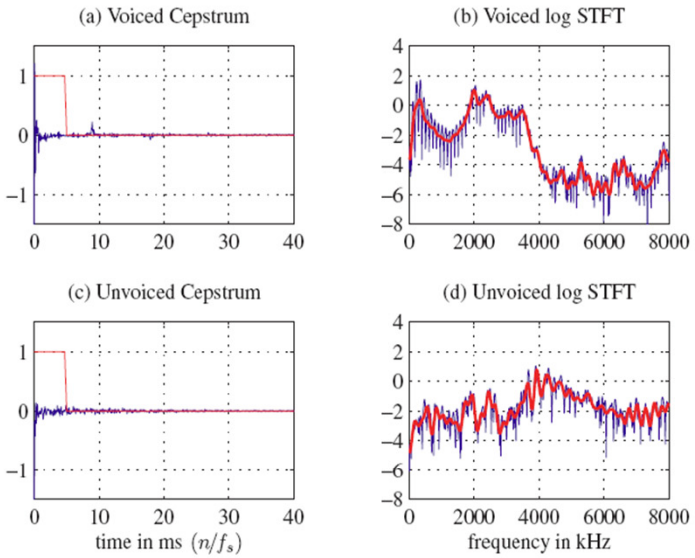


Fig. 5.5 Short-time cepstra and corresponding STFTs and homomorphically-smoothed spectra.

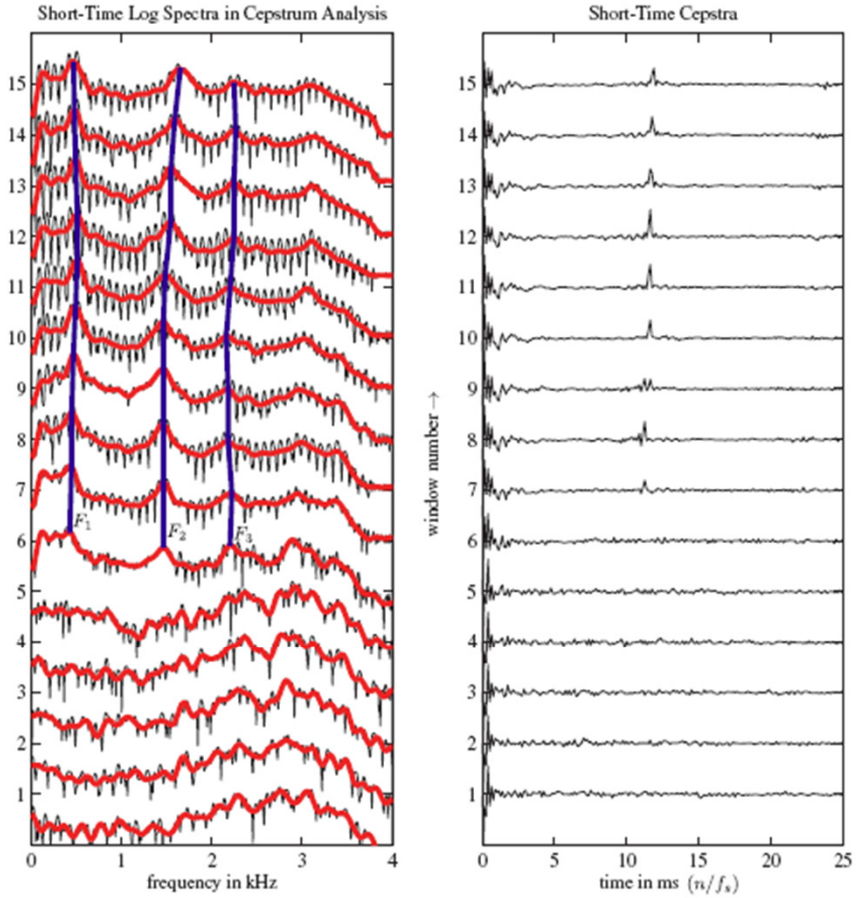
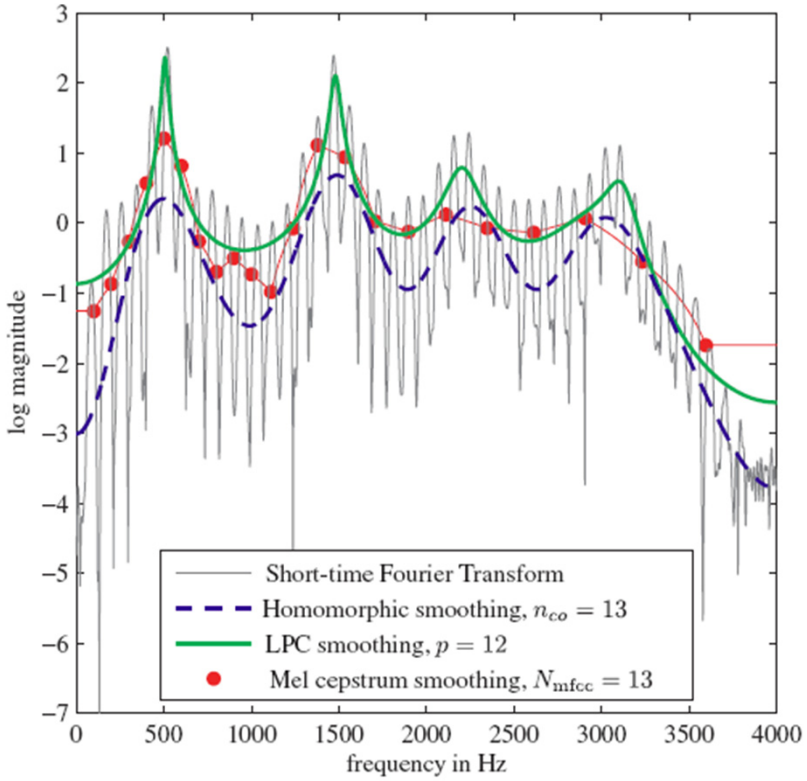


Fig. 5.6 Short-time cepstra and corresponding STFTs and homomorphically-smoothed spectra.

Source-Filter Separation via Cepstrum (2/3)

- The Result of MFCC analysis intrinsically represents a smoothed spectrum
  - Removal of the excitation/harmonics component



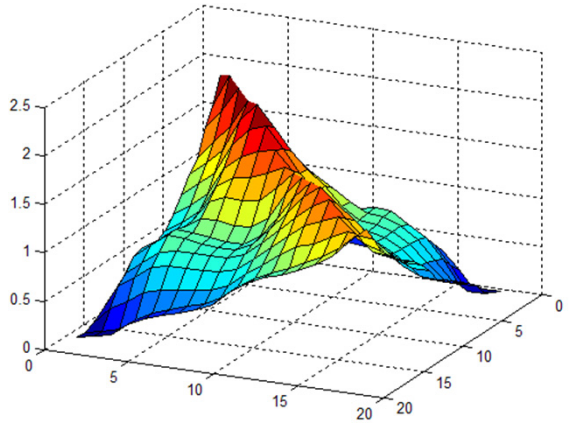
Cepstral Analysis

- Ideal case
  - Preserve the variance introduced by phonemes
  - Suppress the variances introduced by source likes coarticulation, channel, and speaker
  - Reduce the feature dimensionality

Cepstral Analysis (cont.)

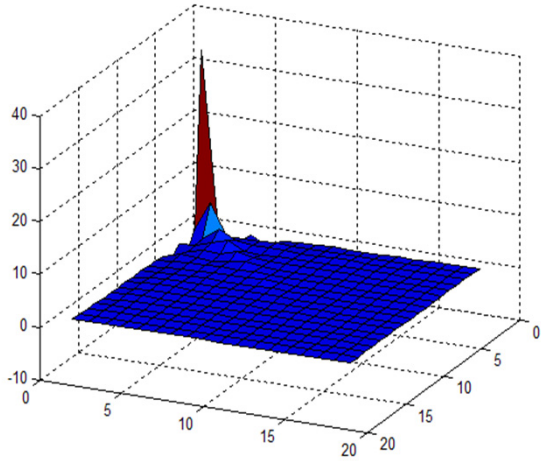
- Project the logarithmic power spectrum (most often modified by auditory-like processing) on the Cosine basis
  - The Cosine basis are used to project the feature space on directions of maximum global (overall) variability
    - Rotation and dimensionality reduction
  - Also partially decorrelates the log-spectral features

Covariance Matrix of the 18-Mel-filter-bank vectors



Calculated using 5471 files (Year 1999 BN)

Covariance Matrix of the 18-cepstral vectors

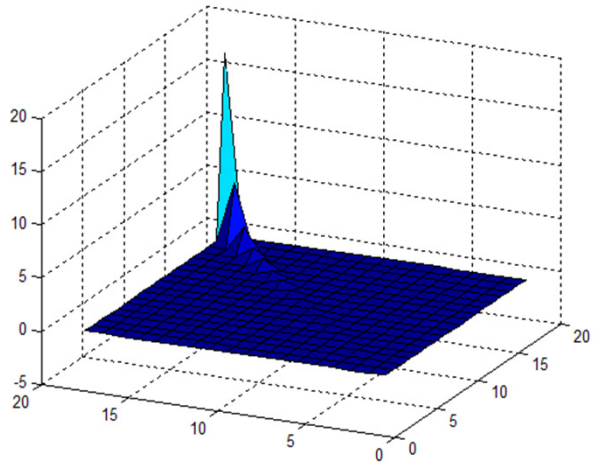


Calculated using 5471 files (Year 1999 BN)

Cepstral Analysis (cont.)

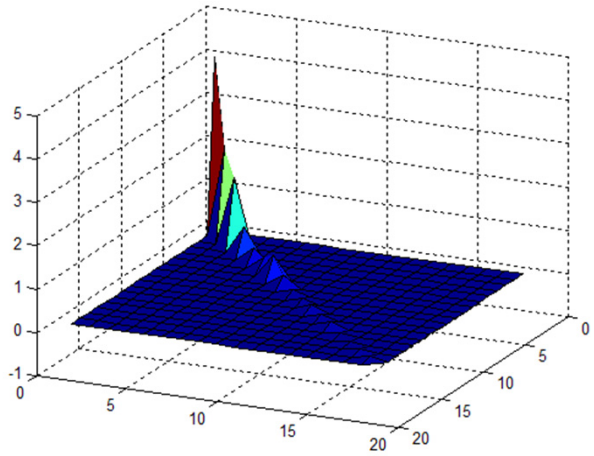
- PCA and LDA also can be used as the basis functions
  - PCA can completely decorrelate the log-spectral features
  - PCA-derived spectral basis projects the feature space on directions of maximum global (overall) variability
  - LDA-derived spectral basis projects the feature space on directions of maximum phoneme separability

Covariance Matrix of the 18-PCA-cepstral vectors



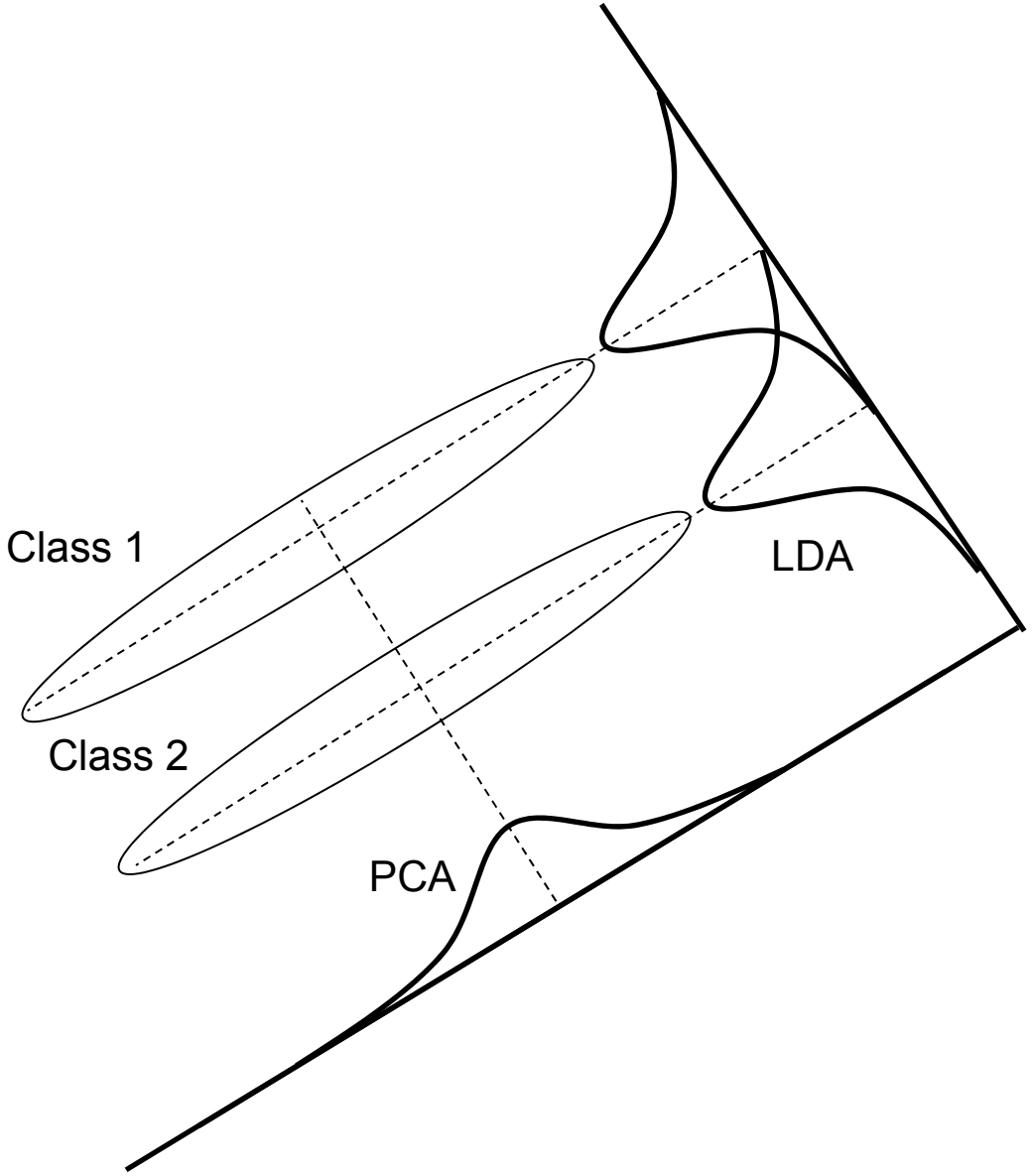
Calculated using 5471 files (Year 1999 BN)

Covariance Matrix of the 18-LDA-cepstral vectors



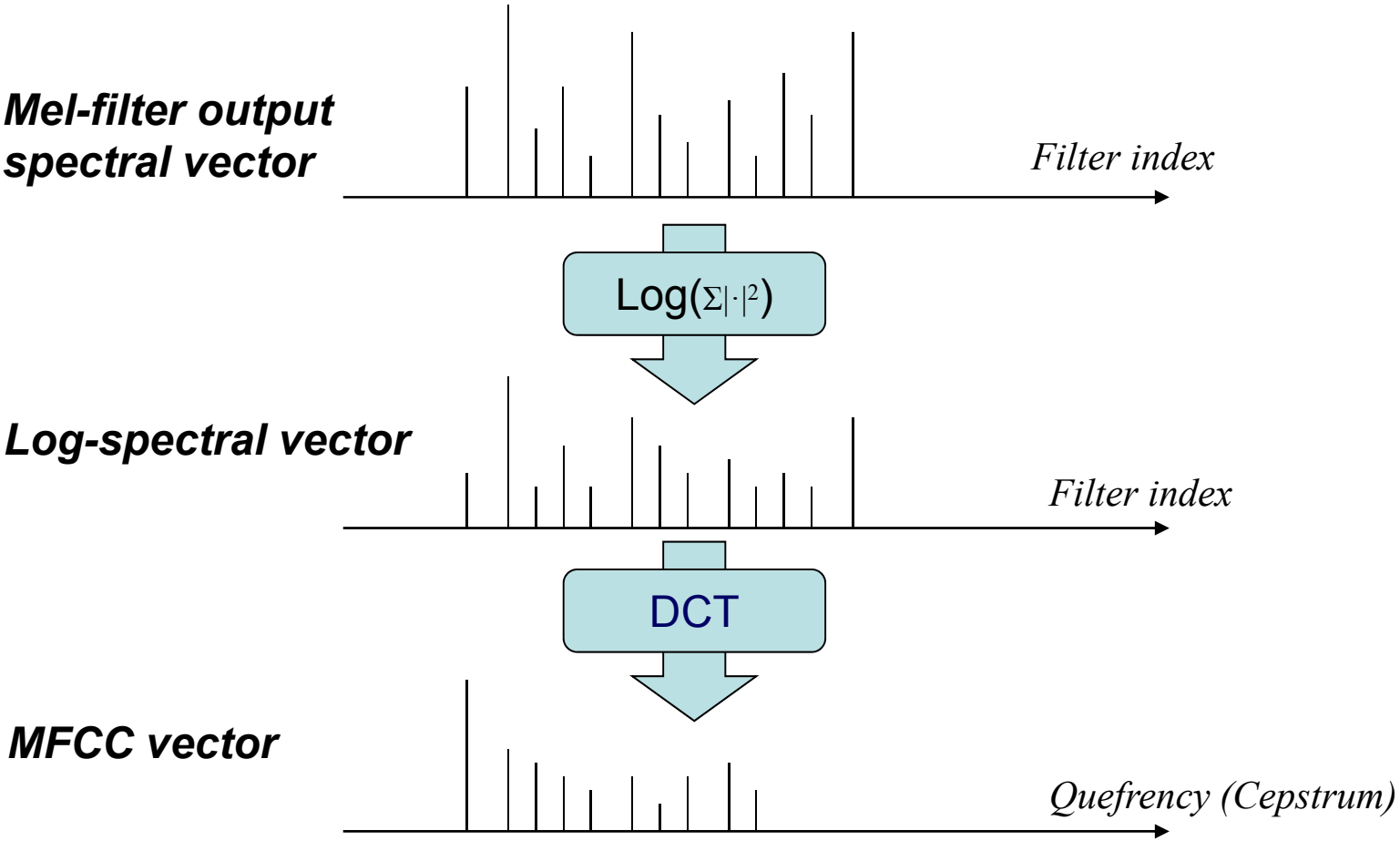
Calculated using 5471 files (Year 1999 BN)

Cepstral Analysis (cont.)



Logarithmic Operation and DCT in MFCC

- The final process of MFCC construction: logarithmic operation and DCT (Discrete Cosine Transform)



Log Energy Operation: Why ?

- Use the magnitude (power) only to discard phase information
  - Phase information is useless in speech recognition
    - Humans are phase-deaf
    - Replacing the phase part of the original speech signal with a continuous random phase won't be perceived by human ear
- Use the logarithmic operation to compress the component amplitudes at every frequency
  - The characteristic of the human hearing system
  - The dynamic compression makes feature extraction less sensitive to variations in dynamics
  - In order to separate more easily the excitation (source) produced by the vocal cords and the the filter that represents the vocal tract

Discrete Cosine Transform

- Final procedure for MFCC : perform inverse DFT on the log-spectral power

- Discrete Cosine Transform (DCT)
  - Since the log-power spectrum is real and symmetric, the inverse DFT reduces to a Discrete Cosine Transform (DCT). The DCT has the property to produce more highly uncorrelated features
    - Partial De-correlation

$$c_l[n] = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} S_l[m] \cos\left[\frac{n\pi}{M}\left(m - \frac{1}{2}\right)\right], \quad n = 0,1,\ldots,L < M$$

    - When *n*=0

$$c_l[0] = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} S_l[m]$$ (relative to the energy of spectrum/filter bank outputs)

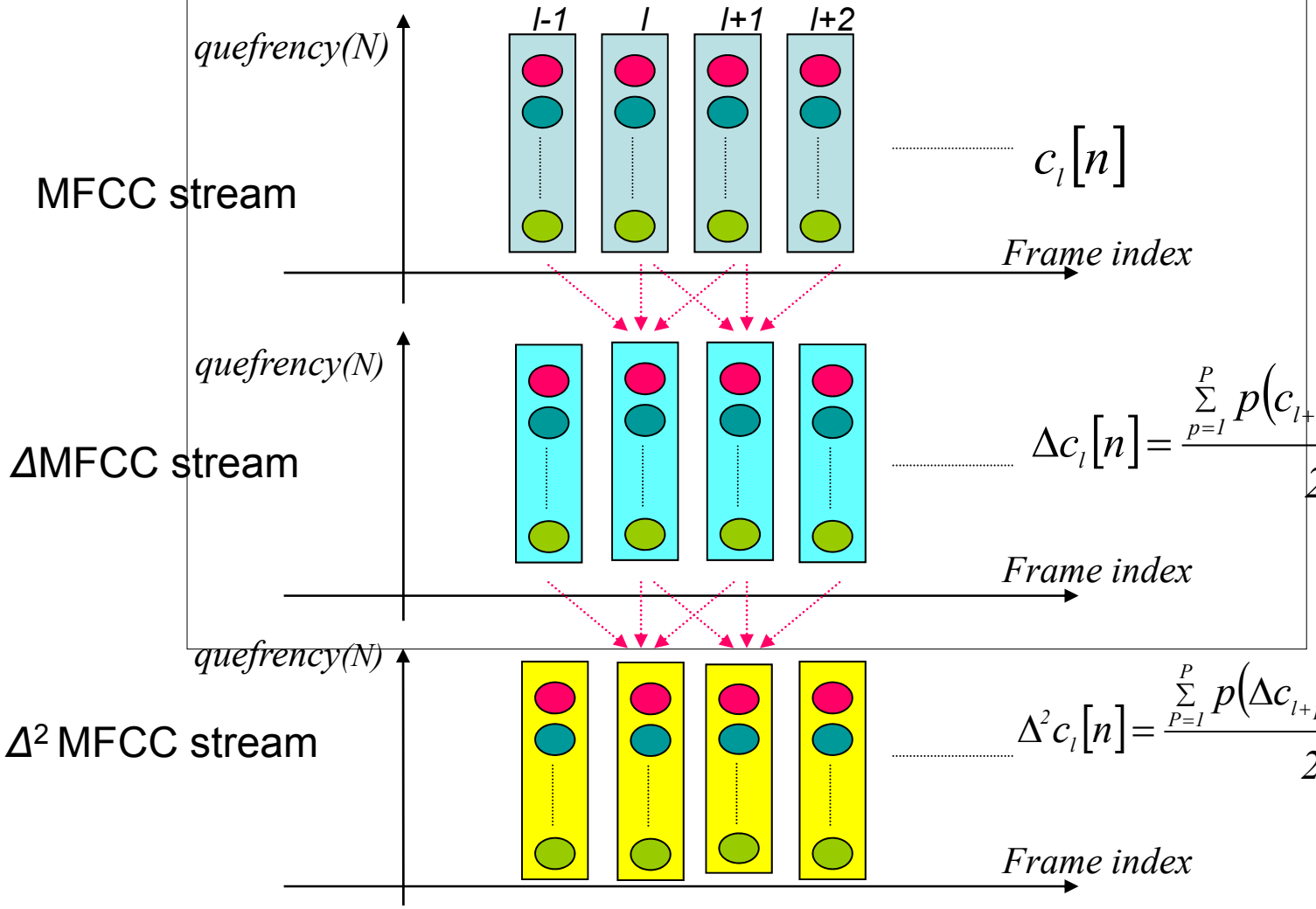
Discrete Cosine Transform: Why?

- Cepstral coefficients are more compact since they are sorted in variance order
  - Can be truncated to retain the highest energy coefficients, which represents an implicit liftering operation with a rectangular window

- Successfully separate the vocal tract and the excitation
  - The envelope of the vocal tract changes slowly, and thus at low quefrencies (lower order cepstrum), while the periodic excitation are at high quefrencies (higher order cepstrum)

Derivatives (1/2)

- Derivative operation : to obtain the temporal information of the static feature vector



Derivatives (2/2)

- The derivative (as that defined in the previous slide) can be obtained by "polynomial fits" to cepstrum sequences to extract simple representations of the temporal variation
  - Furui first noted that such temporal information could be of value for a speaker verification system

S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. on Acoustics, Speech & Signal Processing 29(2), 1981

Derivatives: Why?

- To capture the dynamic evolution of the speech signal
  - Such information carries relevant information for speech recognition
  - The distance (the value of *p*) should be taken into account
    - Too low distance may imply too correlated frames and therefore the dynamic cannot be caught
    - Too high values may imply frames describing too different states
- To cancel the DC part (channel effect) of the MFCC features
  - For example, for clean speech, the MFCC stream is

    while for a channel-distorted speech, the MFCC stream is
    $\{\ldots\ldots\ \boldsymbol{c}_{l-2}, \boldsymbol{c}_{l-1}, \boldsymbol{c}_{l}, \boldsymbol{c}_{l+1}, \boldsymbol{c}_{l+2} \ldots\ldots\ \}$
  - the channel effect *h* is eliminated in the delta (difference) coefficients

    $\{\ldots\ldots \boldsymbol{c}_{l-2}+\boldsymbol{h}, \boldsymbol{c}_{l-1}+\boldsymbol{h}, \boldsymbol{c}_{l}+\boldsymbol{h}, \boldsymbol{c}_{l+1}+\boldsymbol{h}, \boldsymbol{c}_{l+2}+\boldsymbol{h} \ldots\ldots\}$

MFCC v.s LDA

- Tested on Mandarin broadcast news speech
- Large vocabulary continuous speech recognition (LVCSR)
- For each speech frame
  - MFCC uses a set of 13 cepstral coefficients and its first and second time derivatives as the feature vector (39 dimensions)
  - LDA-1 uses a set of 13 cepstral coefficients as the basic vector
  - LDA-2 uses a set of 18 filter-bank outputs as the basic vector

  (Basic vectors from successive nine frames spliced together to form the supervector and then transformed to form a reduced vector with 39 dimensions)

	Character Error Rate	
	TC	WG
MFCC	26.32	22.71
LDA-1	23.12	20.17
LDA-2	23.11	20.11

The character error rates (%) achieved with respective to different feature extraction approaches.