Chapter 3

Frequency domain method for speech processing

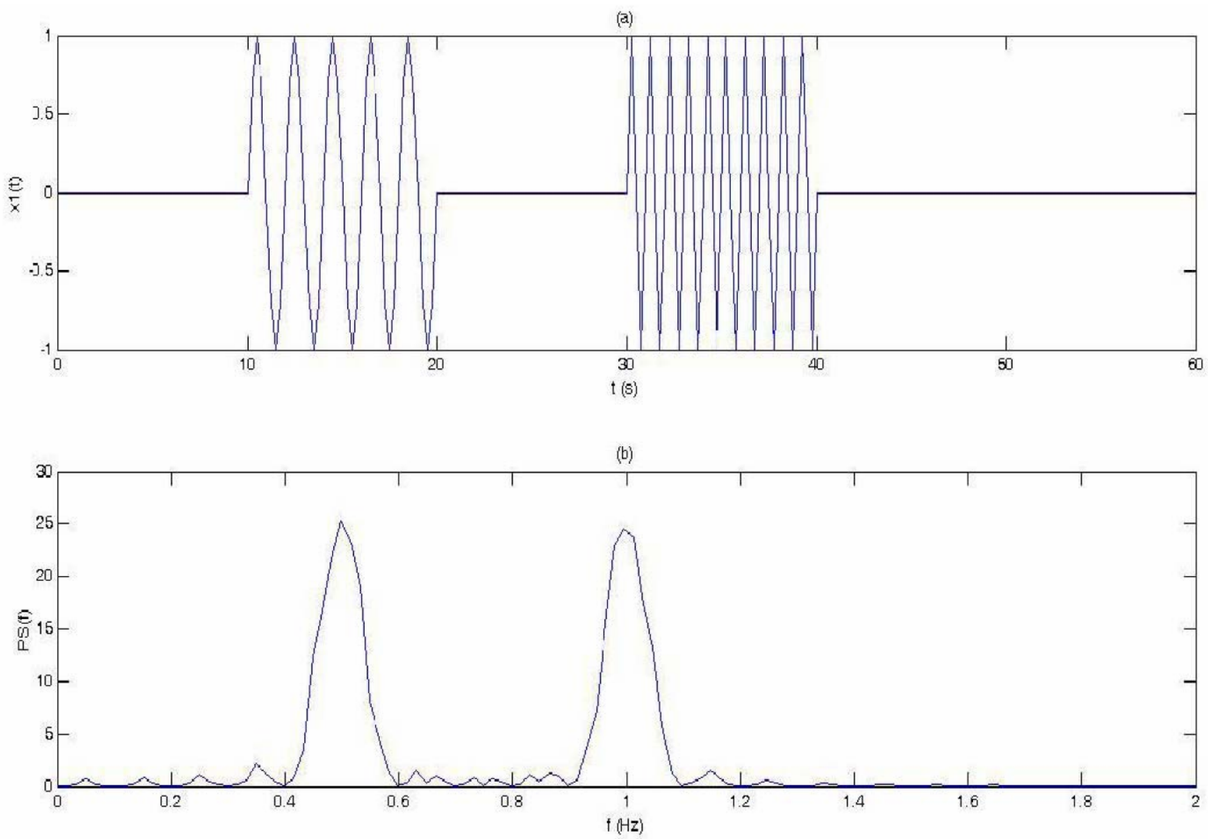
Basanta Joshi, PhD
basanta@ioe.edu.np
Lecture notes can be downloaded from
www.basantajoshi.com.np

Contents

- 3.1. Short Time Fourier analysis
  - 3.1.1. Fourier transform and linear filtering interpretations
  - 3.1.2. Sampling rates
- 3.2. Spectrographic displays
- 3.3. Pitch and formant extraction
- 3.4. Analysis by Synthesis
- 3.5. Analysis synthesis systems
  - 3.5.1. Phase vocoder
  - 3.5.2. Channel Vocoder
- 3.6. Homomorphic speech analysis
  - 3.6.1. Cepstral analysis of Speech
  - 3.6.2. Formant and PitchEstimation
  - 3.6.3. Homomorphic Vocoders

Conventional Fourier analysis

$$x_1(t) = \begin{cases} \sin(2\pi 0.5t) & 10 \le t \le 20 \\ \sin(2\pi t) & 30 \le t \le 40 \\ 0 & elsewhere \end{cases}$$



igure 2.1: (a) The non-stationary signal $x_1(t)$ in time domain. (b) The power spectrum of $x_1(t)$.

- Fourier transform of the entire time series does contain information about the spectral components in time series but
- The information about where in time these frequency components occur cannot be obtained.
- There are some leakages in the frequency domain but in reality they do not appear in the eqn.
- For stationary signal or periodic signal conventional Fourier Transform can do a good job but not a suitable analyzing tool for analyzing the non-stationary k and non-periodic signal. Since almost all the signals in real world are non-stationary and non-periodic, conventional Fourier Transform will not be very useful in real-world signal analysis. [3]

Short Time Fourier analysis

- STFT uses a fixed sliding window to mask the signal and then transforms it to the frequency domain.
- The two frequency components: 0.5 Hz and 1.0 Hz. 0.5 Hz appears around 10s~20s, and 1.0 Hz appears around 30s~40s which is same as eqn.
- Through the help of STFT, we can observe how the frequency of the signal changes with time

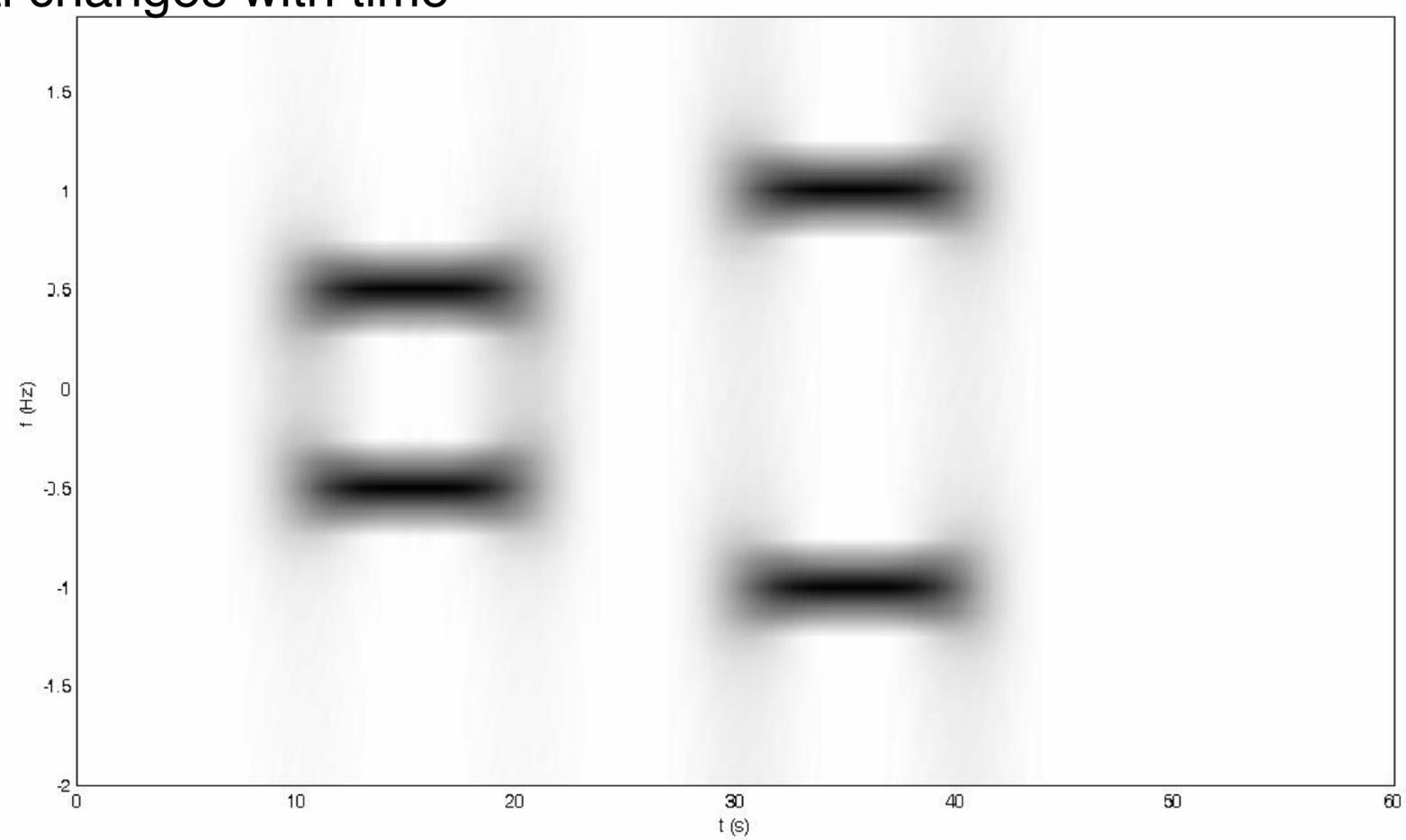


Figure 2.2: The STFT amplitude of x1 (t) as shown in (2.5) and Figure 2.1(a).

Why STFT for Speech Signals

- steady state sounds, like vowels, are produced by **periodic excitation of a linear system** => speech spectrum is the product of the excitation spectrum and the vocal tract frequency response
- speech is a **time-varying signal** => need more sophisticated analysis to reflect time varying properties
  - changes occur at syllabic rates (~10 times/sec)
  - over fixed time intervals of 10-30 msec, properties of most speech signals are relatively constant

Short-Time Fourier Analysis

- represent signal by **sum of sinusoids** or complex exponentials as it leads to convenient solutions to problems (formant estimation, pitch period estimation, analysis-by-synthesis methods), and insight into the signal itself
- such **Fourier representations** provide
  - convenient means to determine response to a sum of sinusoids for linear systems
  - clear evidence of signal properties that are obscured in the original signal

Frequency and the *DTFT*

- sinusoids

$$x(n) = \cos(\omega_0 n) = (e^{j\omega_0 n} + e^{-j\omega_0 n})/2$$

where ω_0 is the *frequency* (in radians) of the sinusoid

- the Discrete-Time Fourier Transform (*DTFT*)

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\omega n} = DTFT\{x(n)\}$$

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega = DTFT^{-1}\{X(e^{j\omega})\}$$

where ω is the *frequency variable* of $X(e^{j\omega})$

DTFT and DFT of Speech

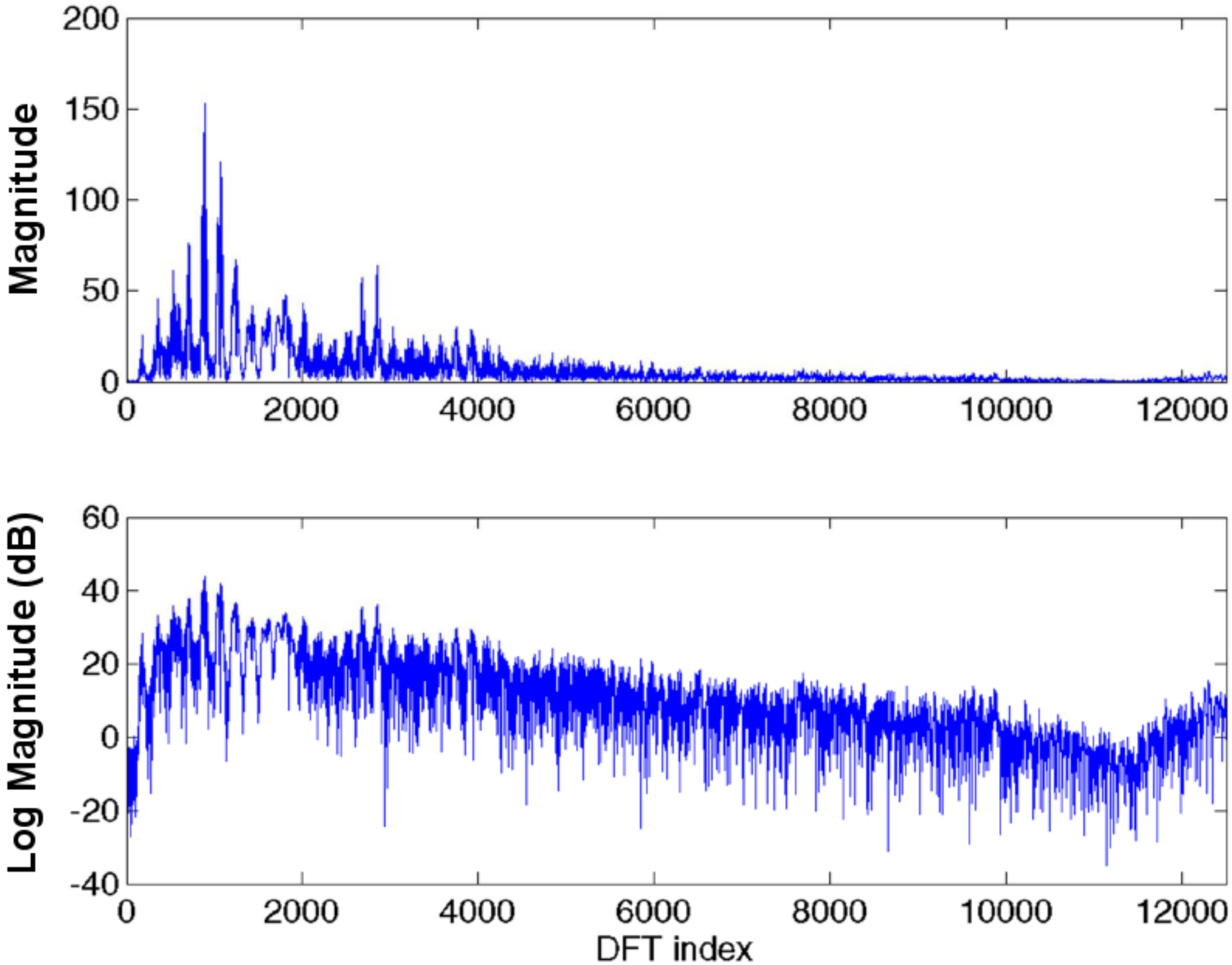
- The DTFT and the DFT for the infinite duration signal could be calculated (the DTFT) and approximated (the DFT) by the following:

$$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\omega m} \quad (DTFT)$$

$$X(k) = \sum_{m=0}^{L-1} x(m)w(m)e^{-j(2\pi/L)km}, \quad k = 0,1,\ldots,L-1$$

$$= X(e^{j\omega})\Big|_{\omega=(2\pi k/L)} \qquad (DFT)$$

DFT of Speech



Short-Time Fourier Transform (STFT)

Short-Time Fourier Transform

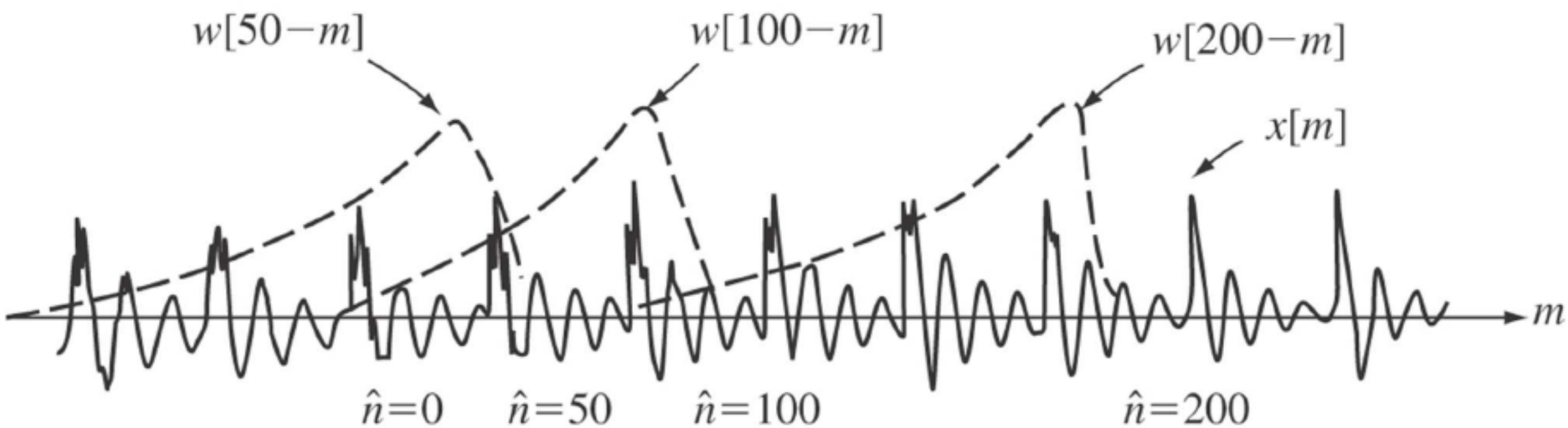
- speech is not a **stationary signal**, i.e., it has properties that **change with time**
- thus a **single representation** based on all the samples of a speech utterance, for the most part, has no meaning
- instead, we define a **time-dependent Fourier transform** (TDFT or STFT) of speech that changes periodically as the speech properties change over time

Definition of STFT

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

both $\hat{n}$ and $\hat{\omega}$ are variables

- $w(\hat{n}-m)$ is a real window which determines the portion of $x(\hat{n})$ that is used in the computation of $X_{\hat{n}}(e^{j\hat{\omega}})$

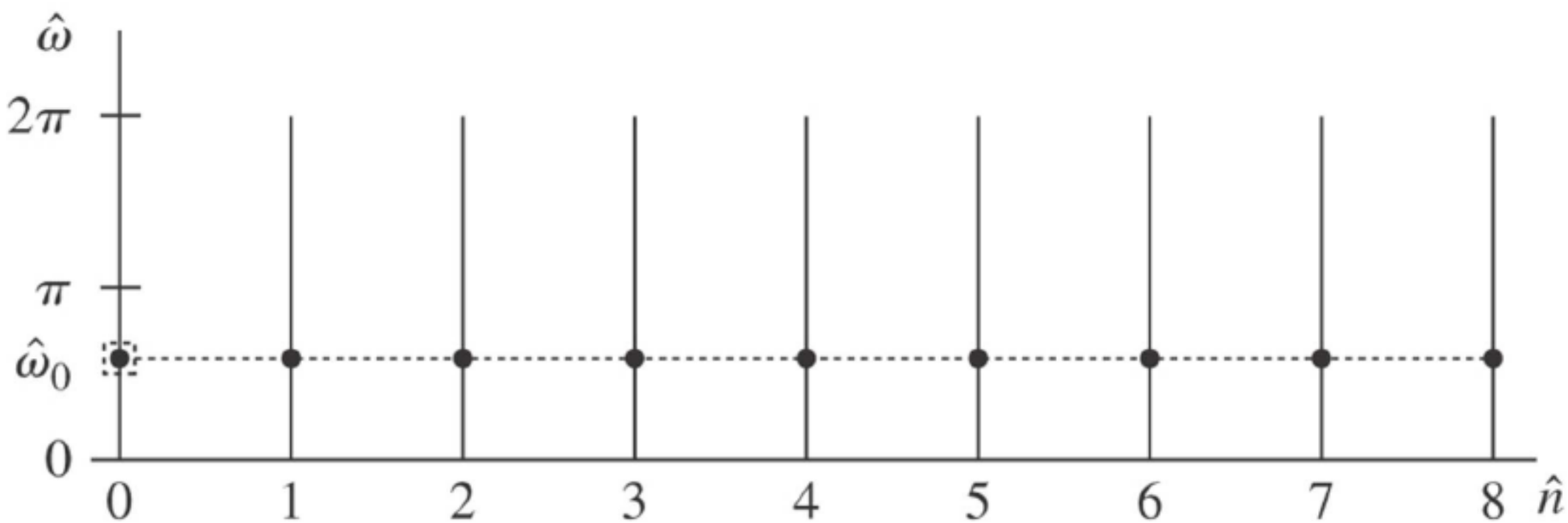


Short Time Fourier Transform

- STFT is a function of two variables, the time index, $\hat{n}$, which is discrete, and the frequency variable, $\hat{\omega}$, which is continuous

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

$$= DTFT\ (x(m)w(\hat{n}-m)) \Rightarrow \hat{n} \text{ fixed, } \hat{\omega} \text{ variable}$$



Short-Time Fourier Transform

- alternative form of STFT (based on change of variables) is

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} w(m)\, x(\hat{n}-m)\, e^{-j\hat{\omega}(\hat{n}-m)}$$

$$= e^{-j\hat{\omega}\hat{n}} \sum_{m=-\infty}^{\infty} x(\hat{n}-m) w(m) e^{j\hat{\omega}m}$$

- if we define

$$\boxed{\tilde{X}_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(\hat{n}-m) w(m) e^{j\hat{\omega}m}}$$

- then $X_{\hat{n}}(e^{j\hat{\omega}})$ can be expressed as (using $m' = -m$)

$$\boxed{X_{\hat{n}}(e^{j\hat{\omega}}) = e^{-j\hat{\omega}\hat{n}} \tilde{X}_n(e^{j\hat{\omega}}) = e^{-j\hat{\omega}\hat{n}}\, DTFT\left[x(\hat{n}+m) w(-m)\right]}$$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

Interpretations of STFT

- there are 2 distinct interpretations of $X_{\hat{n}}(e^{j\hat{\omega}})$

1. assume $\hat{n}$ is fixed, then $X_{\hat{n}}(e^{j\hat{\omega}})$ is simply the normal Fourier transform of the sequence $w(\hat{n}-m)x(m), -\infty < m < \infty$ => for <u>fixed $\hat{n}$</u>, $X_{\hat{n}}(e^{j\hat{\omega}})$ has the same properties as a normal Fourier transform
2. consider $X_{\hat{n}}(e^{j\hat{\omega}})$ as a function of the time index $\hat{n}$ with $\hat{\omega}$ fixed. Then $X_{\hat{n}}(e^{j\hat{\omega}})$ is in the form of a convolution of the signal $x(\hat{n})e^{-j\hat{\omega}\hat{n}}$ with the window $w(\hat{n})$. This leads to an interpretation in the form of linear filtering of the frequency modulated signal $x(\hat{n})e^{-j\hat{\omega}\hat{n}}$ by $w(\hat{n})$.

$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}w(n-m)$$

$$= w(n) * \left(x(n)e^{-j\hat{\omega}n}\right), \quad n \text{ variable, } \hat{\omega} \text{ fixed}$$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

Fourier Transform Interpretation

- consider $X_{\hat{n}}(e^{j\hat{\omega}})$ as the normal Fourier transform of the sequence $w(\hat{n}-m)\,x(m), -\infty < m < \infty$ for fixed $\hat{n}$.
- the window $w(\hat{n}-m)$ slides along the sequence $x(m)$ and defines a new STFT for <u>every</u> value of $\hat{n}$
- what are the conditions for the existence of the STFT
  - the sequence $w(\hat{n}-m)\,x(m)$ must be absolutely summable for all values of $\hat{n}$
    - since $|x(\hat{n})| \le L$ (32767 for 16-bit sampling)
    - since $|w(\hat{n})| \le 1$ (normalized window levels)
    - since window duration is usually finite
  - $w(\hat{n}-m)\,x(m)$ is absolutely summable for all $\hat{n}$

Frequencies for STFT

- the STFT is periodic in ω with period 2π, i.e.,

  $X_{\hat{n}}(e^{j\hat{\omega}}) = X_{\hat{n}}(e^{j(\hat{\omega}+2\pi k)}),\ \forall k$

- can use any of several frequency variables to express STFT, including

  --$\hat{\omega} = \hat{\Omega}T$ (where T is the sampling period for $x(m)$) to represent analog radian frequency, giving $X_{\hat{n}}(e^{j\hat{\Omega}T})$

  --$\hat{\omega} = 2\pi\hat{f}$ or $\hat{\omega} = 2\pi\hat{F}T$ to represent normalized frequency ($0 \le \hat{f} \le 1$) or analog frequency ($0 \le \hat{F} \le F_s = 1/T$), giving $X_{\hat{n}}(e^{j2\pi\hat{f}})$ or $X_{\hat{n}}(e^{j2\pi\hat{F}T})$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

Signal Recovery from STFT

- since for a given value of $\hat{n}$, $X_{\hat{n}}(e^{j\hat{\omega}})$ has the same properties as a normal Fourier transform, we can recover the input sequence <u>exactly</u>
- since $X_{\hat{n}}(e^{j\hat{\omega}})$ is the normal Fourier transform of the windowed sequence $w(\hat{n}-m)\,x(m)$, then

$$w(\hat{n}-m)\,x(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}m}d\hat{\omega}$$

- assuming the window satisfies the property that $w(0) \neq 0$ (a trivial requirement), then by evaluating the inverse Fourier transform when $m = \hat{n}$, we obtain

$$x(\hat{n}) = \frac{1}{2\pi\, w(0)}\int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}})e^{j\omega\hat{n}}d\hat{\omega}$$

Signal Recovery from STFT

$$x(\hat{n}) = \frac{1}{2\pi w(0)} \int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}}) e^{j\omega\hat{n}} d\hat{\omega}$$

- with the requirement that $w(0) \neq 0$, the sequence $x(\hat{n})$ can be recovered exactly from $X_{\hat{n}}(e^{j\hat{\omega}})$, if $X_{\hat{n}}(e^{j\hat{\omega}})$ is known for all values of $\hat{\omega}$ over one complete period
  - sample-by-sample recovery process
  - $X_{\hat{n}}(e^{j\hat{\omega}})$ must be known for every value of $\hat{n}$ and for all $\hat{\omega}$

Alternative Forms of STFT

Alternative forms of $X_{\hat{n}}(e^{j\hat{\omega}})$

1. real and imaginary parts

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \mathrm{Re}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right] + j\,\mathrm{Im}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right]$$

$$= a_{\hat{n}}(\hat{\omega}) - j\,b_{\hat{n}}(\hat{\omega})$$

$$a_{\hat{n}}(\hat{\omega}) = \mathrm{Re}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right]$$

$$b_{\hat{n}}(\hat{\omega}) = -\mathrm{Im}\left[X_{\hat{n}}(e^{j\hat{\omega}})\right]$$

- when $x(m)$ and $w(\hat{n}-m)$ are both real

2. magnitude and phase

$$X_{\hat{n}}(e^{j\hat{\omega}}) = |X_{\hat{n}}(e^{j\hat{\omega}})|e^{j\theta_{\hat{n}}(\hat{\omega})}$$

- can relate $|X_{\hat{n}}(e^{j\hat{\omega}})|$ and $\theta_{\hat{n}}(\hat{\omega})$ to $a_{\hat{n}}(\hat{\omega})$ and $b_{\hat{n}}(\hat{\omega})$

Role of Window in STFT

- The window $w(\hat{n}-m)$ does the following:
  1. chooses portion of $x(m)$ to be analyzed
  2. window shape determines the nature of $X_{\hat{n}}(e^{j\hat{\omega}})$
- Since $X_{\hat{n}}(e^{j\hat{\omega}})$ (for fixed $\hat{n}$) is the normal FT of $w(\hat{n}-m)\,x(m)$, then if we consider the normal FT's of both $x(n)$ and $w(n)$ individually, we get

$$X(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}$$

$$W(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} w(m)e^{-j\hat{\omega}m}$$

Role of Window in STFT

- then for fixed $\hat{n}$, the normal Fourier transform of the product $w(\hat{n}-m)\,x(m)$ is the convolution of the transforms of $w(\hat{n}-m)$ and $x(m)$
- for fixed $\hat{n}$, the FT of $w(\hat{n}-m)$ is $W(e^{-j\hat{\omega}})e^{-j\hat{\omega}\hat{n}}$--thus

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{-j\theta})e^{-j\theta\hat{n}}\,X(e^{j(\hat{\omega}-\theta)})\,d\theta$$

- and replacing θ by $-\theta$ gives

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})e^{j\theta\hat{n}}\,X(e^{j(\hat{\omega}+\theta)})\,d\theta$$

Interpretation of Role of Window

- $X_{\hat{n}}(e^{j\hat{\omega}})$ is the convolution of $X(e^{j\hat{\omega}})$ with the FT of the shifted window sequence $W(e^{-j\hat{\omega}})e^{-j\hat{\omega}\hat{n}}$
- $X(e^{j\hat{\omega}})$ really doesn't have meaning since $x(\hat{n})$ varies with time; consider $x(\hat{n})$ defined for window duration and extended for all time to have the <u>same</u> properties $\Rightarrow$ then $X(e^{j\hat{\omega}})$ does exist with properties that reflect the sound within the window (can also consider $x(\hat{n}) = 0$ outside the window and define $X(e^{j\hat{\omega}})$ appropriately--but this is another case)

<u>Bottom Line</u>: $X_{\hat{n}}(e^{j\hat{\omega}})$ is a smoothed version of the FT of the part of $x(\hat{n})$ that is within the window w.

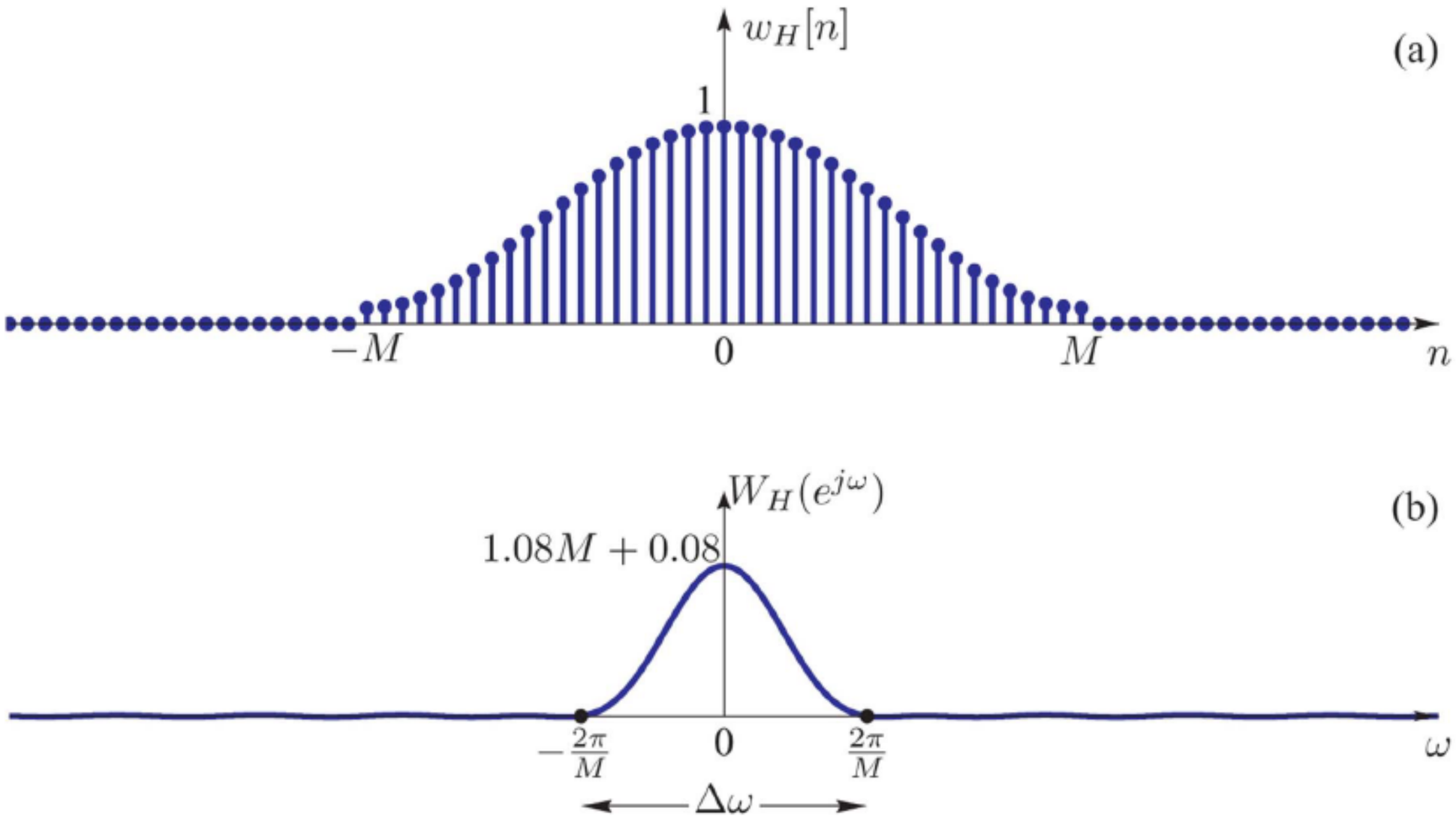
Windows in STFT

- for $X_{\hat{n}}(e^{j\hat{\omega}})$ to represent the short-time spectral properties of $x(\hat{n})$ inside the window $\Rightarrow$ $W(e^{j\theta})$ should be much narrower in frequency than significant spectral regions of $X(e^{j\hat{\omega}})$--i.e., almost an impulse in frequency
- consider rectangular and Hamming windows, where width of the main spectral lobe is inversely proportional to window length, and side lobe levels are essentially independent of window length

Rectangular Window: flat window of length L samples; first zero in frequency response occurs at F_S/L, with sidelobe levels of -14 dB or lower

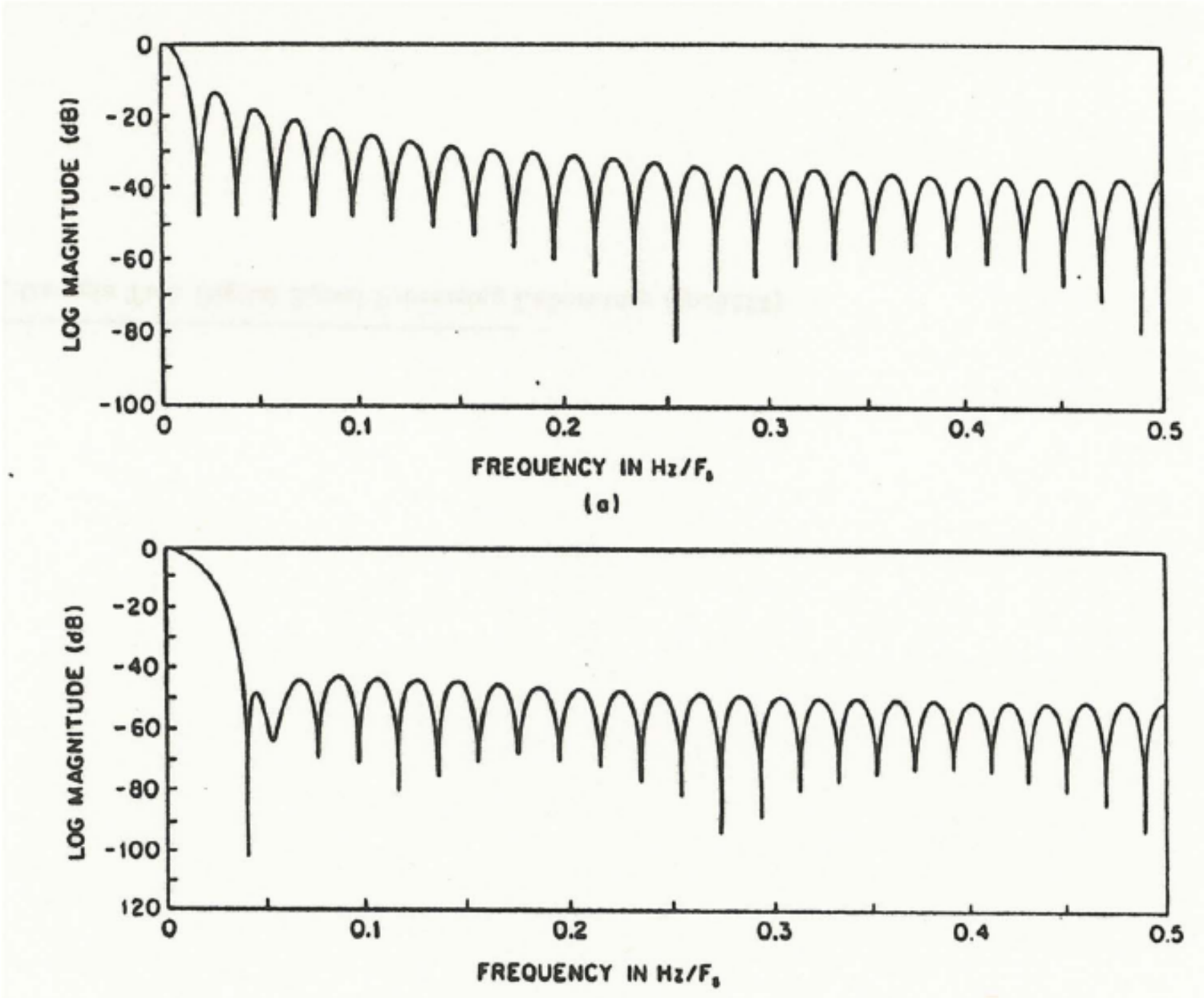
Hamming Window: raised cosine window of length L samples; first zero in frequency response occurs at $2F_S/L$, with sidelobe levels of -40 dB or lower

Windows

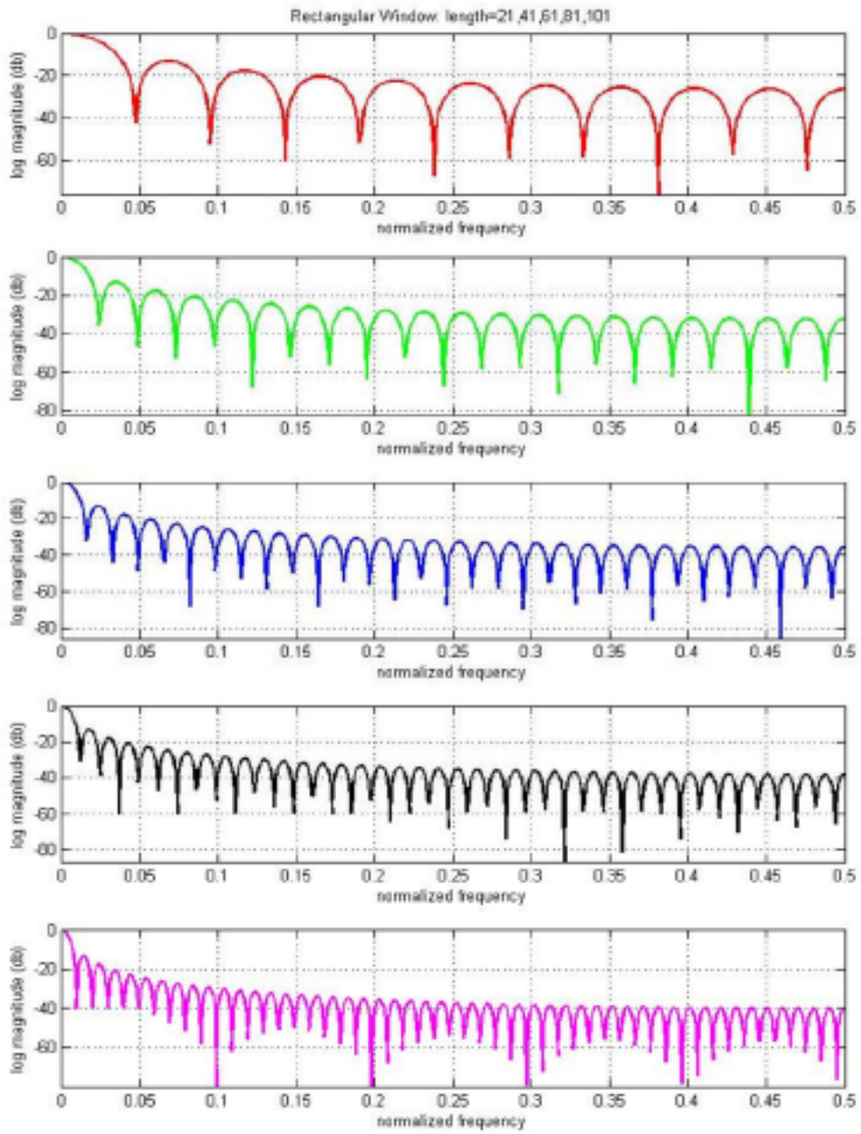


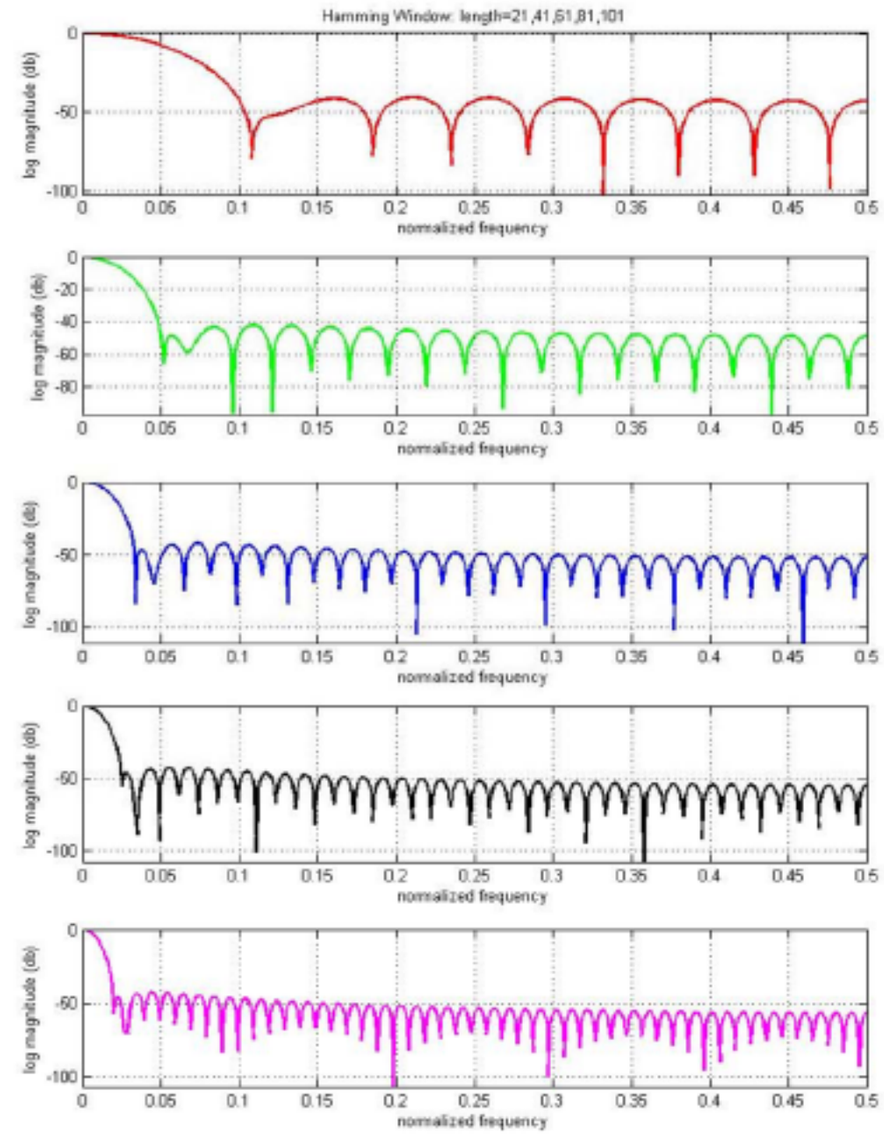
L=2M+1-point Hamming window and its corresponding DTFT

Frequency Responses of Windows



Window Frequency Responses

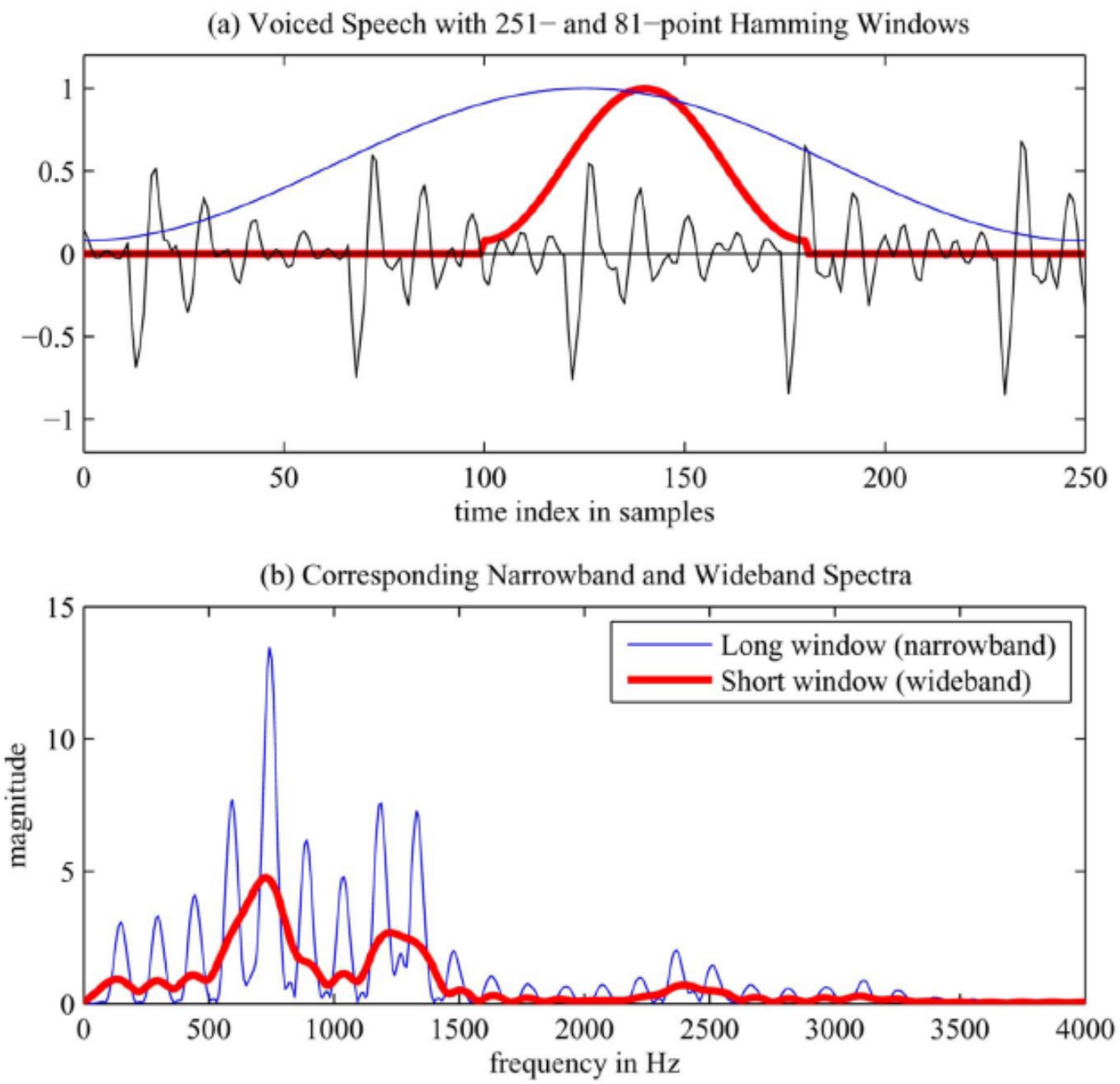




Rectangular Windows, L=21,41,61,81,101

Hamming Windows, L=21,41,61,81,101

Effect of Window Length-HW

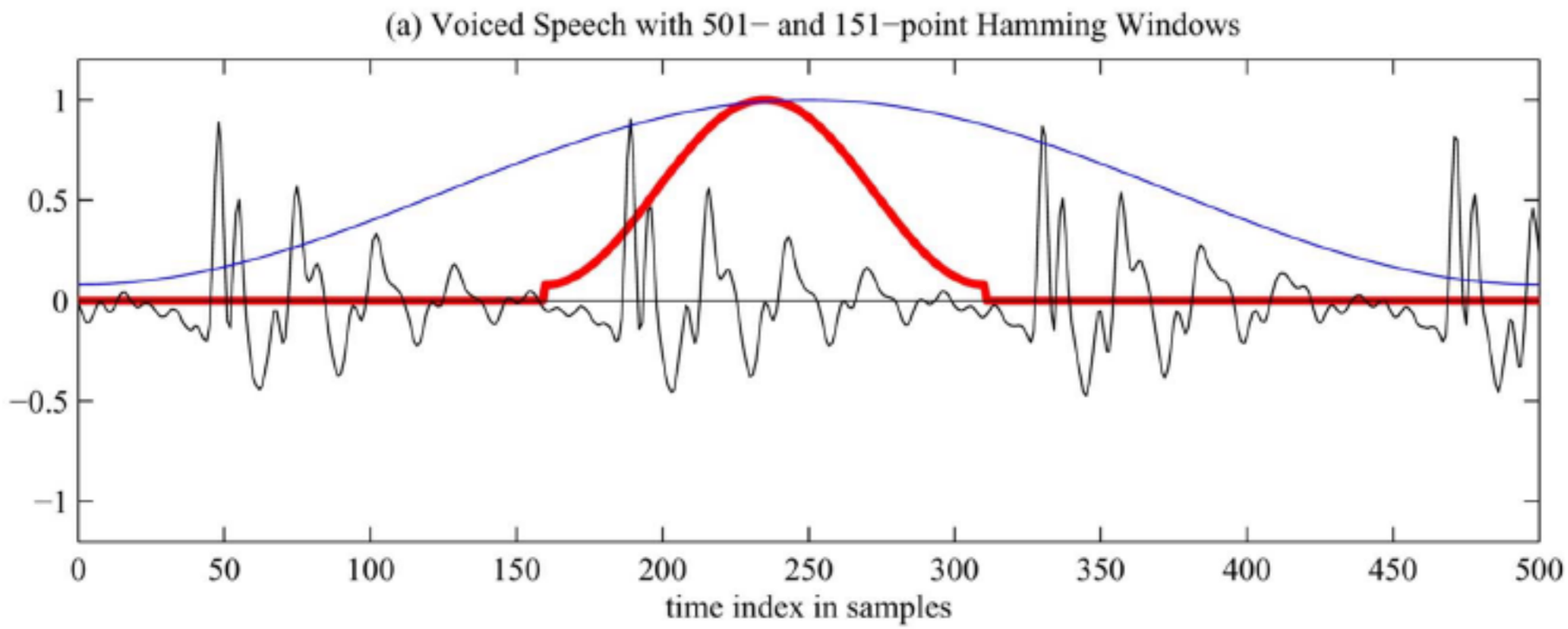


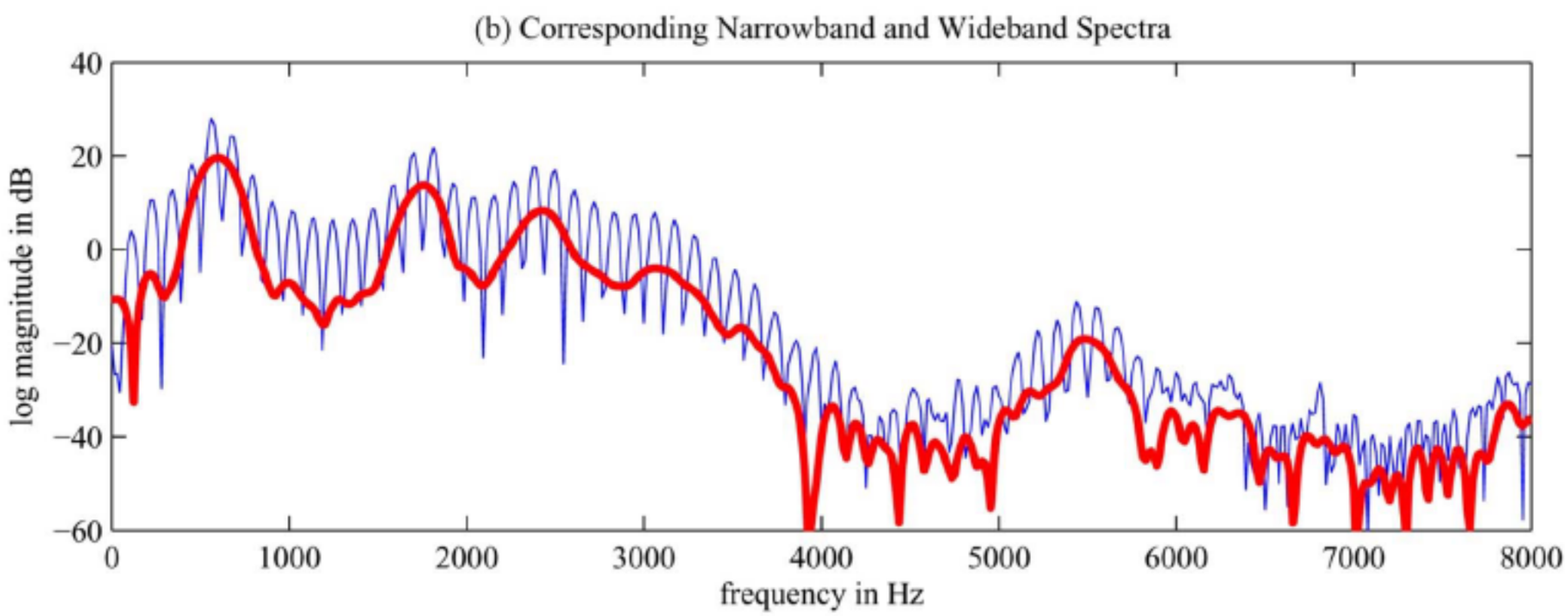
- The size of $w[n]$ affects the time-frequency resolution of STFT:

Window size of $w[n]$	Bandwidth of $W(\omega)$	Time resolution	Frequency resolution	Good for:
long	narrow	bad	good	sinusoidal components, (harmonic)
short	wide	good	bad	fast time-varying components, (rapid conversational speech)

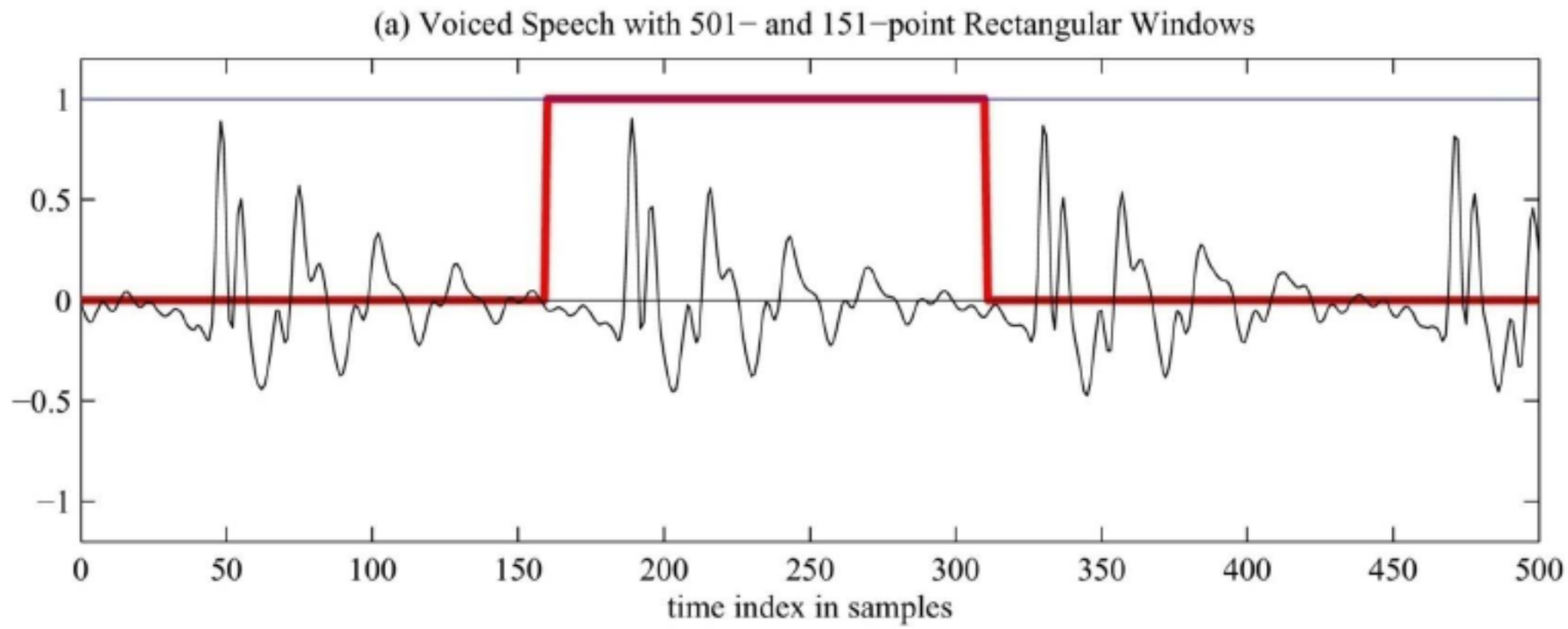
A fundamental problem of STFT and other time-frequency analysis techniques is the selection of the windows to achieve a good tradeoff between time and frequency resolution.

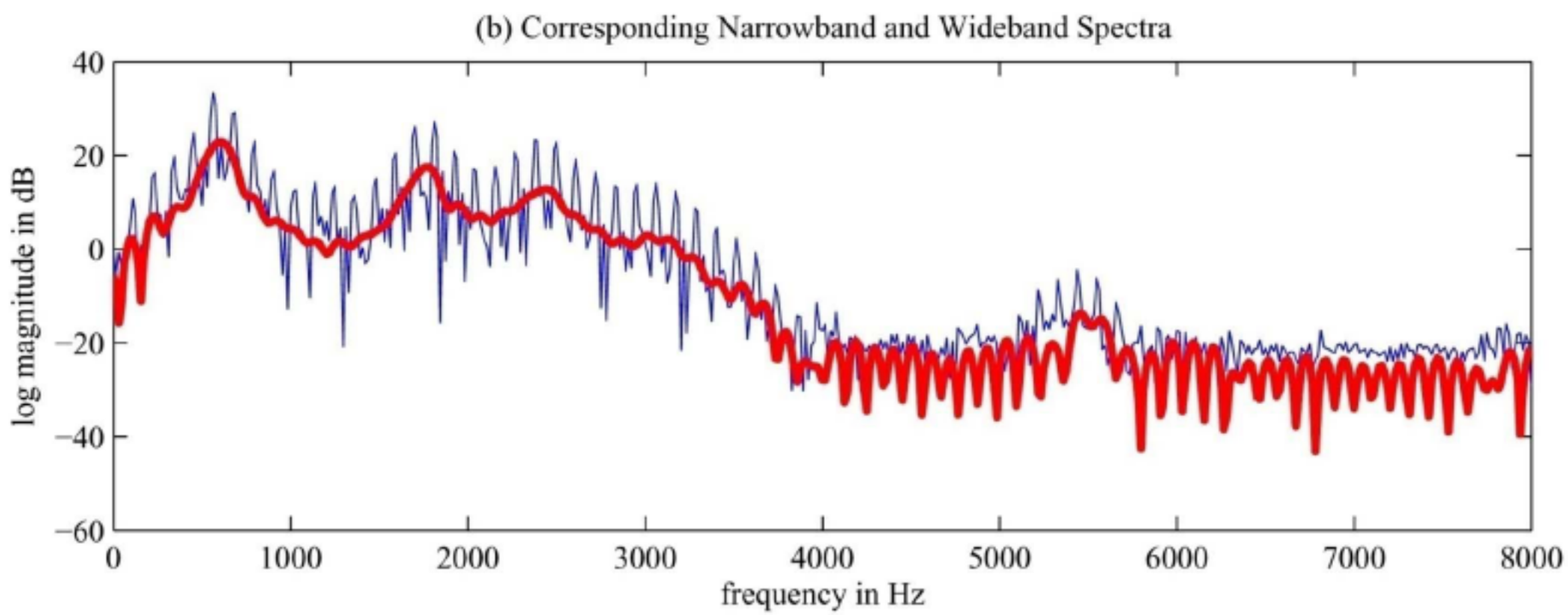
Effect of Window Length-HW



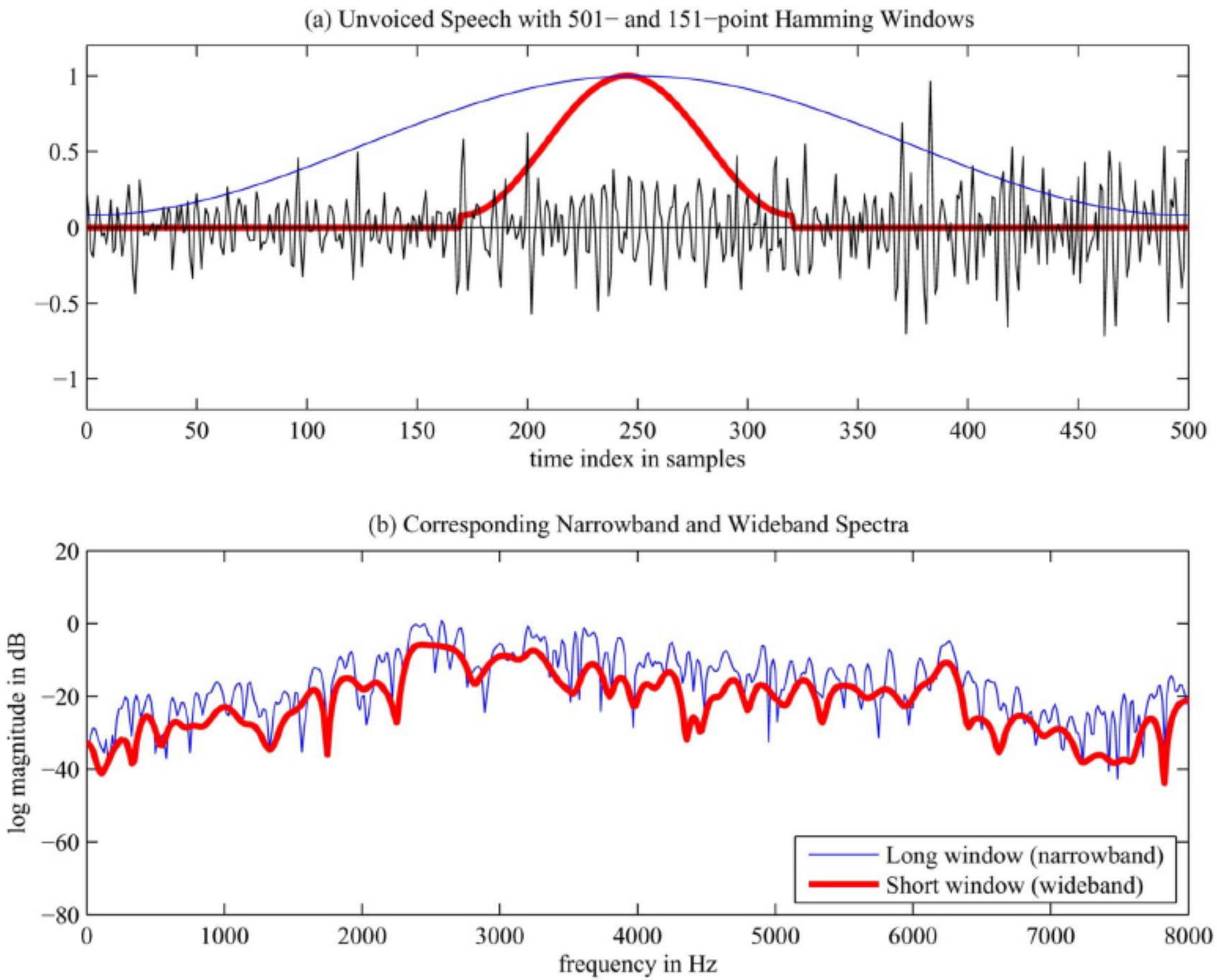


Effect of Window Length-RW





Effect of Window Length-HW



Relation to Short-Time Autocorrelation

- $X_{\hat{n}}(e^{j\hat{\omega}})$ is the discrete-time Fourier transform of $w[\hat{n}-m]x[m]$ for each value of $\hat{n}$, then it is seen that

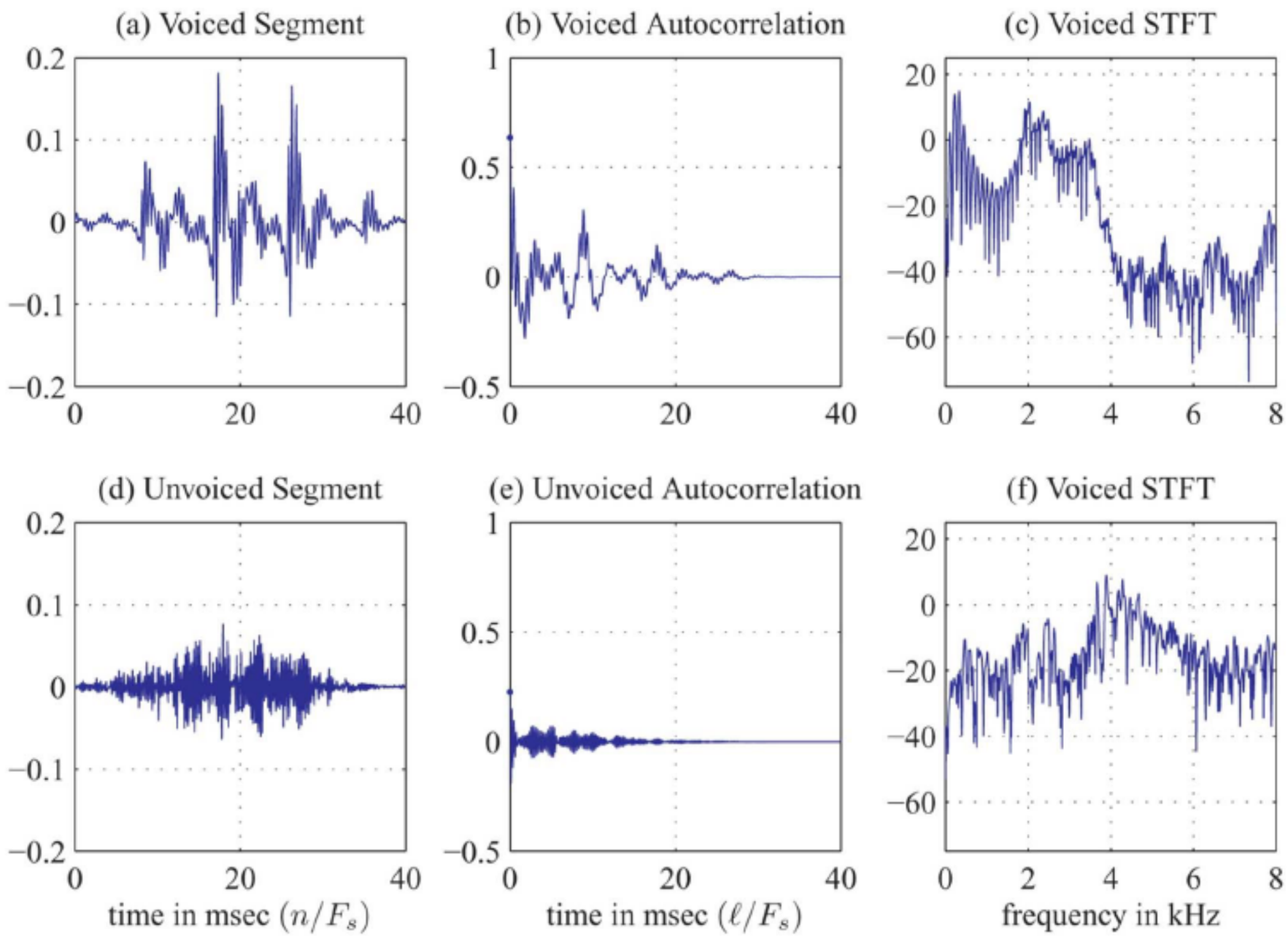
$$S_{\hat{n}}(e^{j\hat{\omega}}) = |X_{\hat{n}}(e^{j\hat{\omega}})|^2 = X_{\hat{n}}(e^{j\hat{\omega}})X_{\hat{n}}^*(e^{j\hat{\omega}})$$

is the Fourier transform of

$$R_{\hat{n}}(l) = \sum_{m=-\infty}^{\infty} w[\hat{n}-m]x[m]w[\hat{n}-l-m]x[m+l]$$

which is the short-time autocorrelation function of the previous chapter. Thus the above equations relate the short-time spectrum to the short-time autocorrelation,

Short-Time Autocorrelation and STFT



Summary of FT view of STFT

- interpret $X_{\hat{n}}(e^{j\omega})$ as the normal Fourier transform of the sequence $w(\hat{n}-m)\,x(m), -\infty < m < \infty$
- properties of this Fourier transform depend on the window
  - frequency resolution of $X_{\hat{n}}(e^{j\omega})$ varies inversely with the length of the window => want long windows for high resolution
  - want $x(n)$ to be relatively stationary (non-time-varying) during duration of window for most stable spectrum => want short windows

$\Leftrightarrow$ as usual in speech processing, there needs to be a compromise between good temporal resolution (short windows) and good frequency resolution (long windows)

Linear Filtering Interpretation of STFT

Linear Filtering Interpretation

1. modulation-lowpass filter form (n rather than $\hat{n}$)

$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}w(n-m)$$

$$= w(n) * \left(x(n)e^{-j\hat{\omega}n}\right), \quad n \text{ variable, } \hat{\omega} \text{ fixed}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})\,X(e^{j(\theta+\hat{\omega})})e^{j\theta n}d\theta$$

2. bandpass filter-demodulation

$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} w(m)\,x(n-m)e^{-j\hat{\omega}(n-m)}$$

$$= e^{-j\hat{\omega}n}\sum_{m=-\infty}^{\infty}(w(m)e^{j\hat{\omega}m})\,x(n-m)$$

$$= e^{-j\hat{\omega}n}[(w(n)e^{j\hat{\omega}n}) * x(n)], \quad n \text{ variable, } \hat{\omega} \text{ fixed}$$

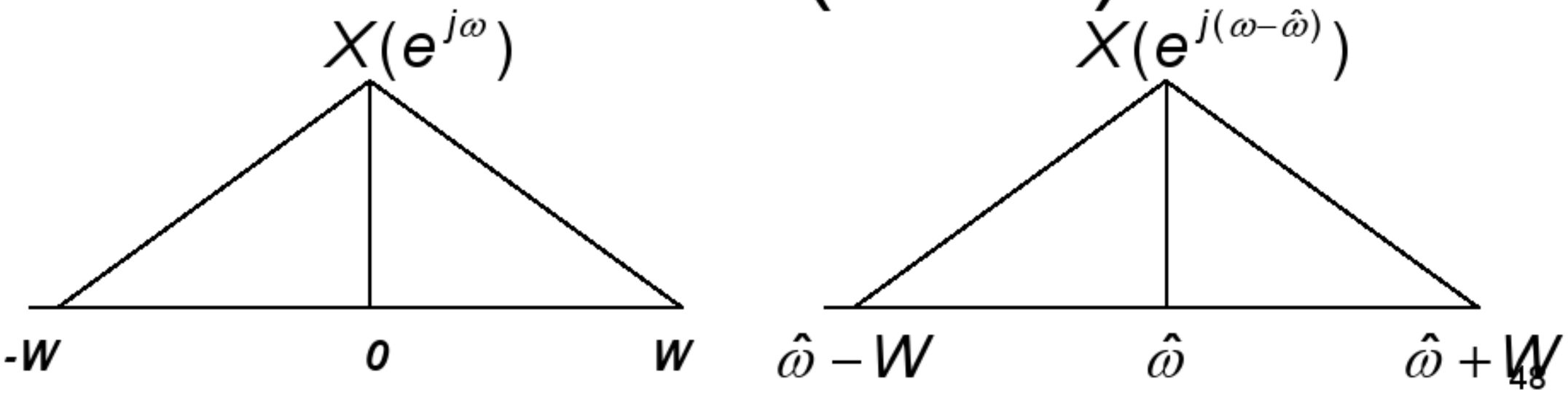
Modulation

Modulation

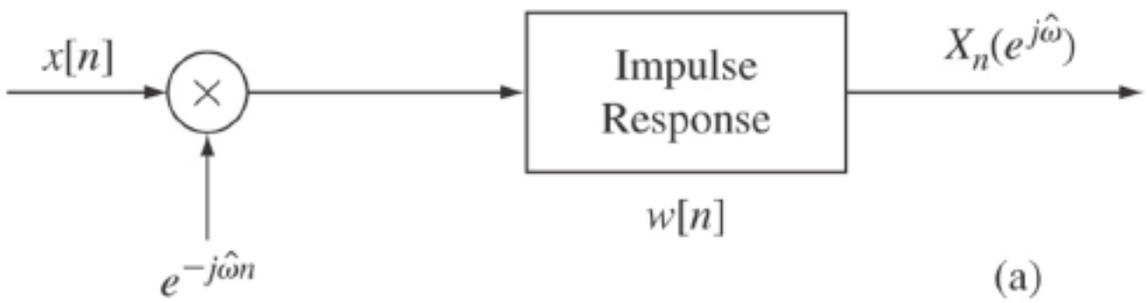
$$x[n] \cdot e^{j\hat{\omega}n} \leftrightarrow X(e^{j\omega}) * FT(e^{j\hat{\omega}n})$$

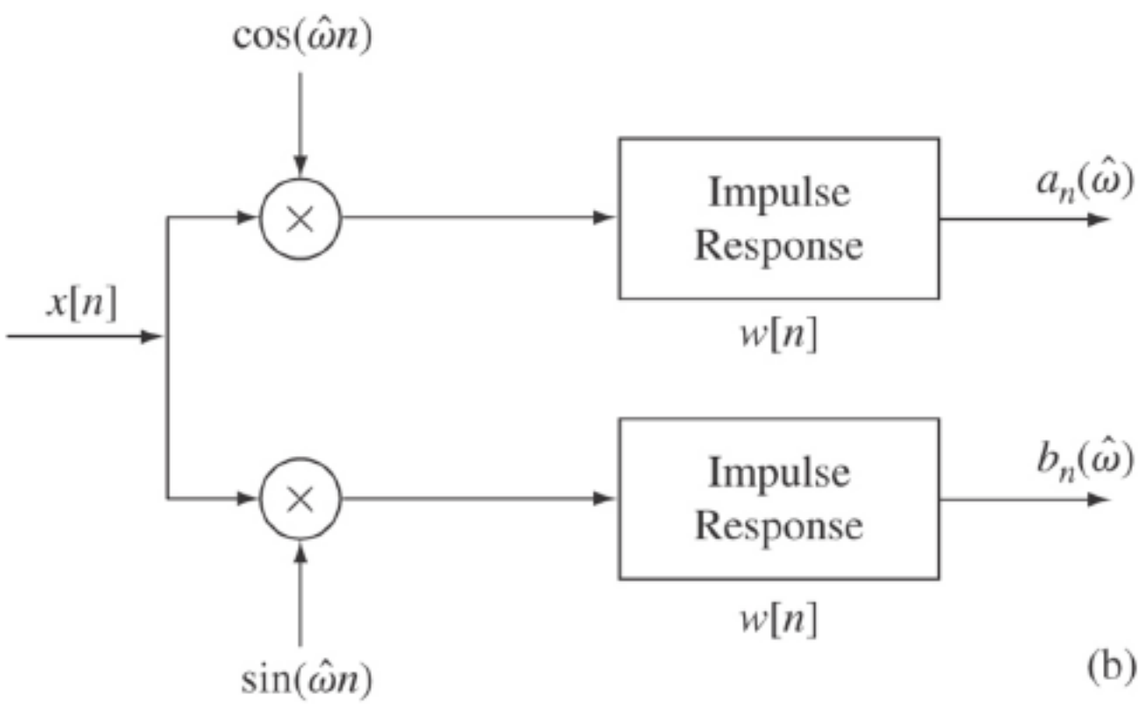
$$= X(e^{j\omega}) * \delta(\omega - \hat{\omega})$$

$$= X(e^{j(\omega-\hat{\omega})})$$



Linear Filtering Interpretation





1. modulation-lowpass filter form:

$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}w(n-m),$$

n variable, $\hat{\omega}$ fixed

$$= \left(x(n)e^{-j\hat{\omega}n}\right) * w(n)$$

$$= \left(x(n)\cos(\hat{\omega}n)\right) * w(n) -$$

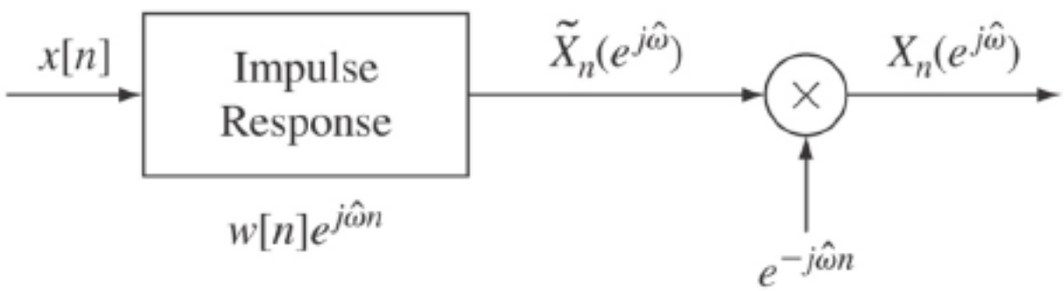
$$j\left(x(n)\sin(\hat{\omega}n)\right) * w(n)$$

$$= a_n(\hat{\omega}) - jb_n(\hat{\omega})$$

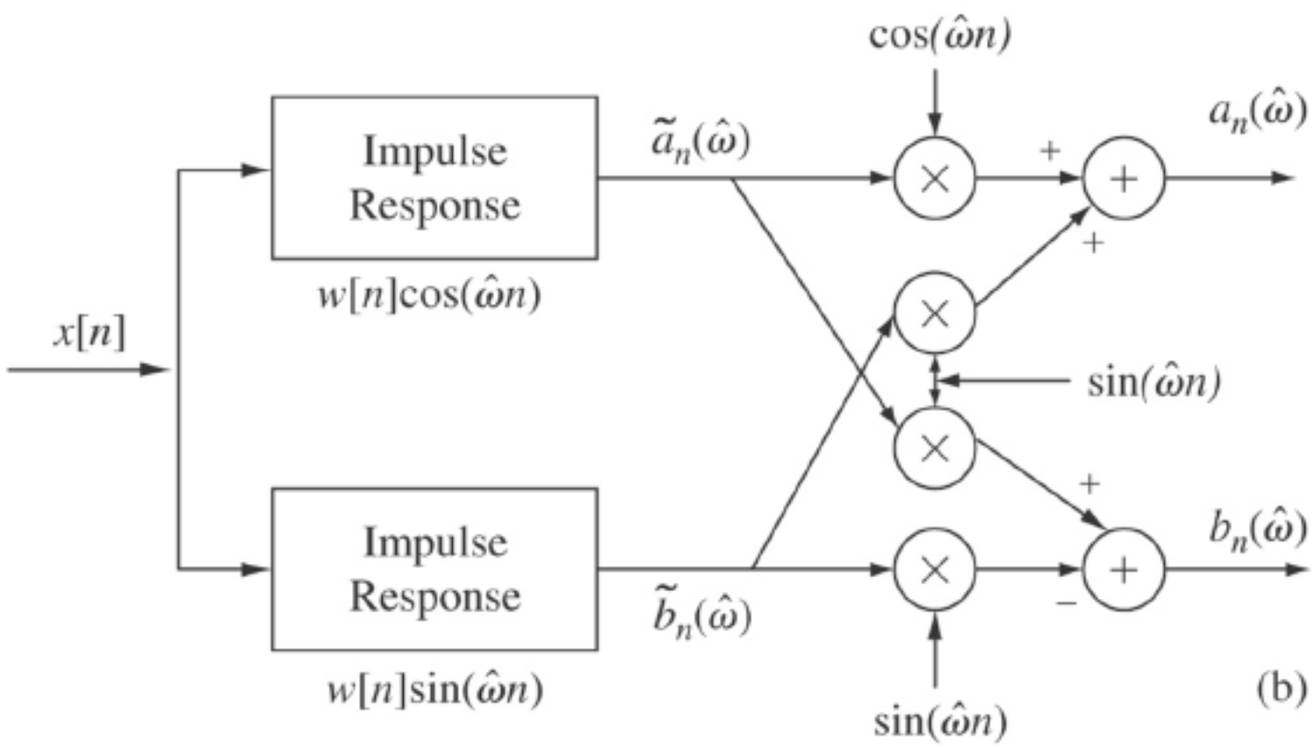
Linear Filtering Interpretation

2. bandpass filter-demodulation form

$$X_n(e^{j\hat{\omega}}) = e^{-j\hat{\omega}n}\left[\left(w(n)e^{j\hat{\omega}n}\right) * x(n)\right], \quad n \text{ variable}, \ \hat{\omega} \text{ fixed}$$



(a)

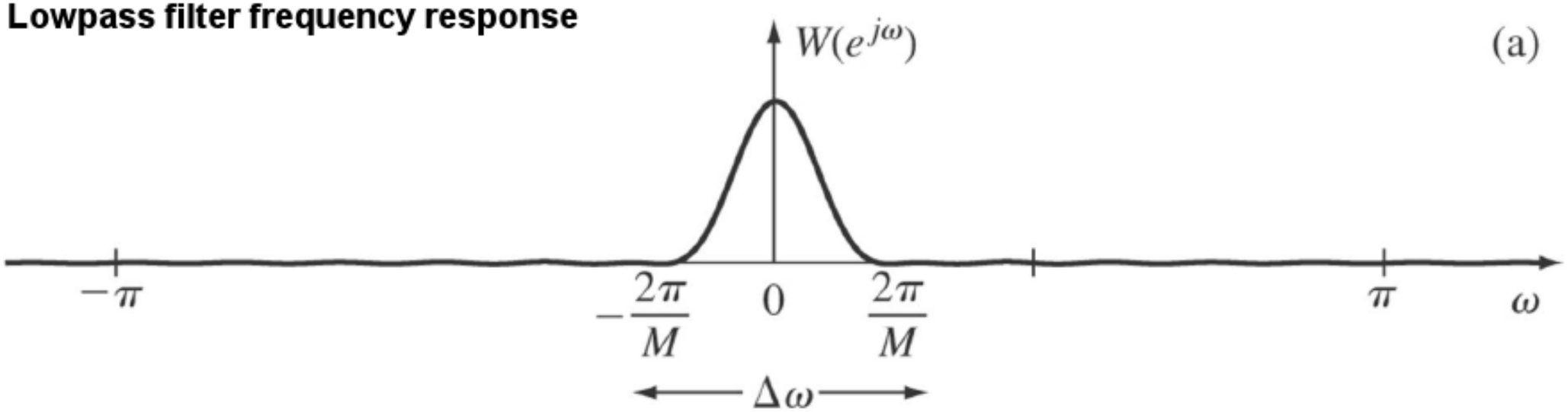


(b)

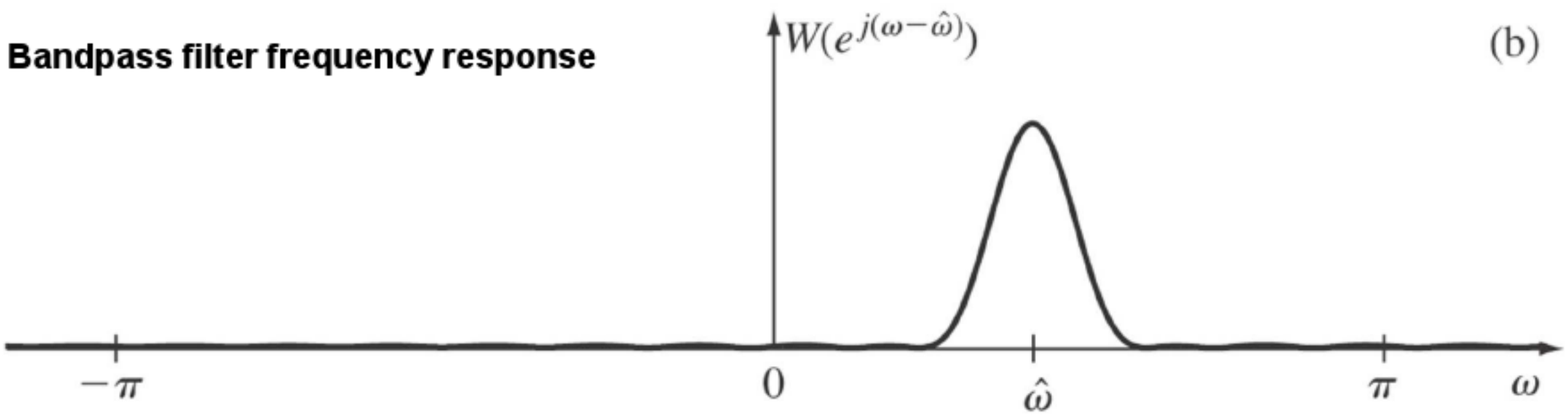
- complex bandpass filter output modulated by signal $e^{-j\hat{\omega}n}$
- if $W(e^{j\theta})$ is lowpass, then filter is bandpass around $\theta = \hat{\omega}$
- all real computation for lower half structure

Linear Filtering Interpretation

Lowpass filter frequency response



Bandpass filter frequency response



Linear Filtering Interpretation

- assume normal FT of $x(n)$ exists

  $x(n) \leftrightarrow X(e^{j\theta})$ (recall that $\hat{\omega}$ is a particular frequency)

  $x(n)e^{-j\hat{\omega}n} \leftrightarrow X(e^{j(\theta+\hat{\omega})})$

=> spectrum of $x(n)$ at frequency $\hat{\omega}$ is shifted to zero frequency;

- since the STFT is a convolution, the FT of the STFT is the product of the individual FT's, i.e.,

  $$X(e^{j(\theta+\hat{\omega})}) \cdot W(e^{j\theta})$$

- if $W(e^{j\theta})$ resembles a narrow band lowpass filter, i.e., $W(e^{j\theta}) = 1$ for small θ and is 0 otherwise, then

  $$X(e^{j(\theta+\hat{\omega})}) \cdot W(e^{j\theta}) \approx X(e^{j\hat{\omega}})$$

Summary-STFT

Short-Time Fourier Transform (STFT)

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m},$$

$$-\infty < \hat{n} < \infty,\ 0 \le \hat{\omega} < 2\pi$$

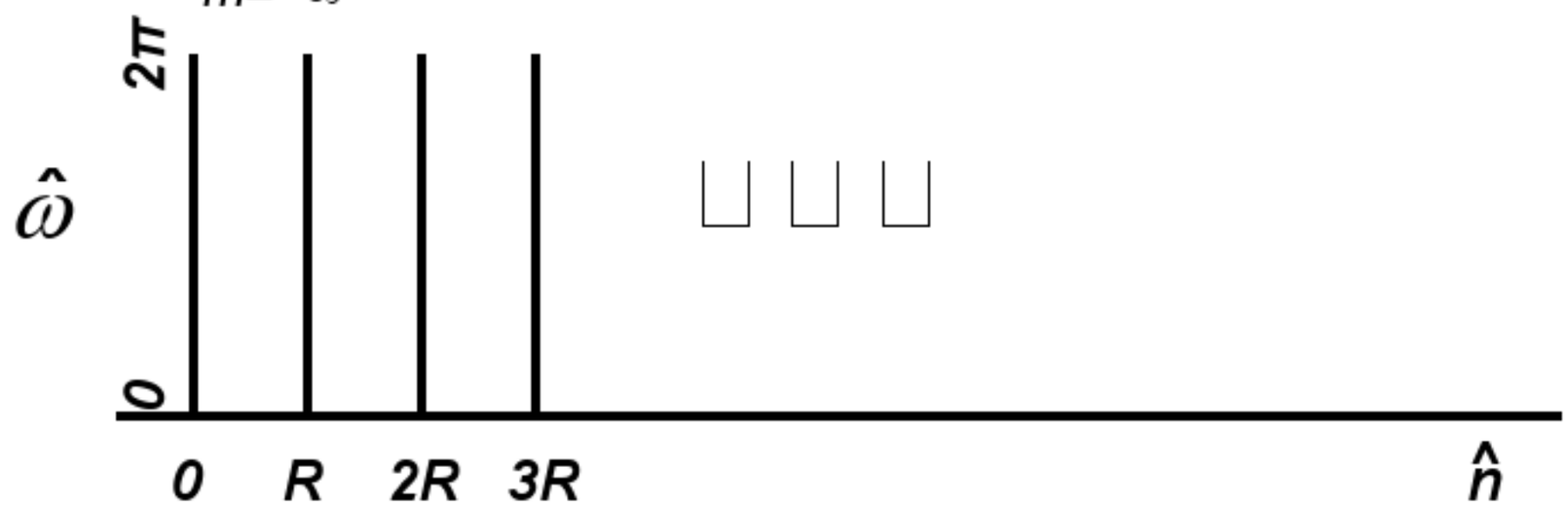
Fixed value of $\hat{n}$, varying $\hat{\omega}$ -- DFT Interpretation

Fixed value of $\hat{\omega}$, varying $\hat{n}$ -- Filter Bank Interpretation

Summary

Short-Time Fourier Transform (STFT)

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m}, \; -\infty < \hat{n} < \infty, \; 0 \le \hat{\omega} < 2\pi$$



DFT: $$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=\hat{n}-L+1}^{\hat{n}} \left(x[m]w[\hat{n}-m]\right)e^{-j\hat{\omega}m}$$

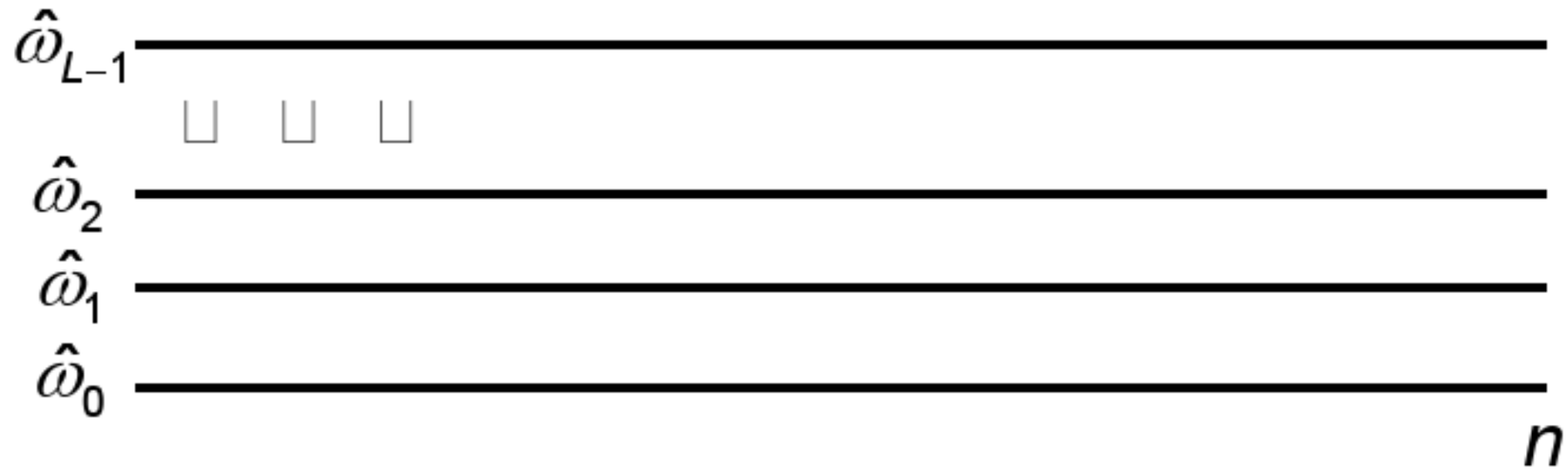
$$X_{\hat{n}}(e^{j\hat{\omega}}) = \mathrm{DFT}\left(x[m]w[\hat{n}-m]\right)$$

$$0 \le \hat{\omega} < 2\pi, \quad \hat{n} = 0, R, 2R, \ldots$$

Summary – Modulation/Lowpass Filter

Short-Time Fourier Transform (STFT)

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m}, \; -\infty < \hat{n} < \infty, \; 0 \le \hat{\omega} < 2\pi$$



Filter Bank: $X_n(e^{j\hat{\omega}}) = \sum_{m=n-L+1}^{n} \left(x[m]e^{-j\hat{\omega}m} \right) w[n-m]$

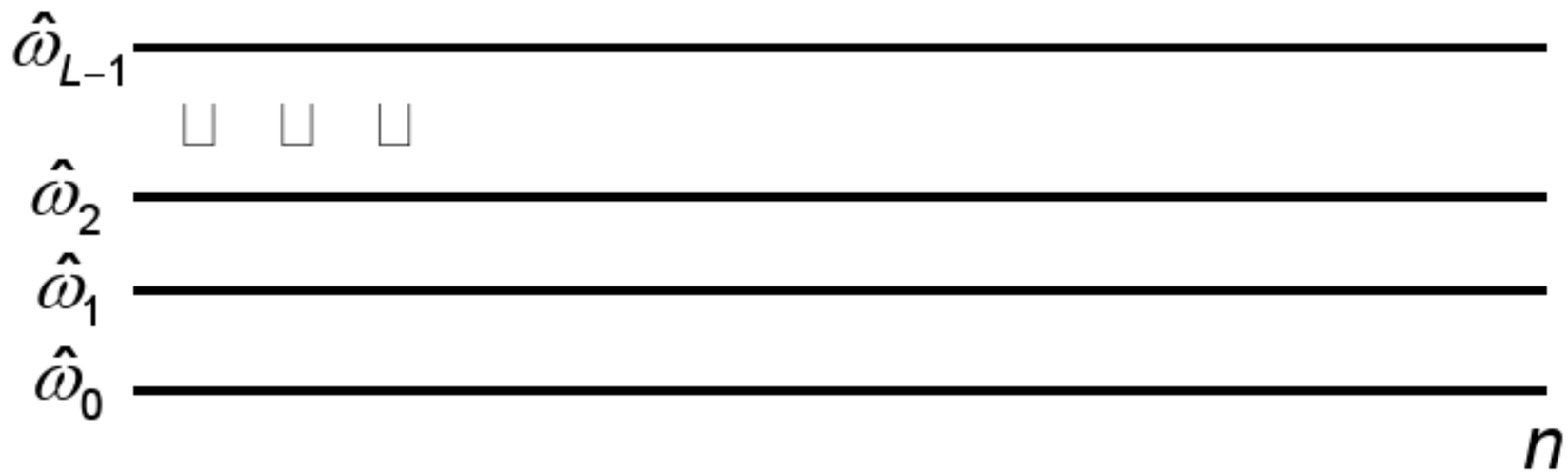
$$X_n(e^{j\hat{\omega}}) = \left(x[n]e^{-j\hat{\omega}n} \right) w[n-m]$$

$$= \left(x[n]e^{-j\hat{\omega}n} \right) * w[n]$$

Summary – Bandpass Filter/Demodulation

Short-Time Fourier Transform (STFT)

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m}, \; -\infty < \hat{n} < \infty, \; 0 \le \hat{\omega} < 2\pi$$

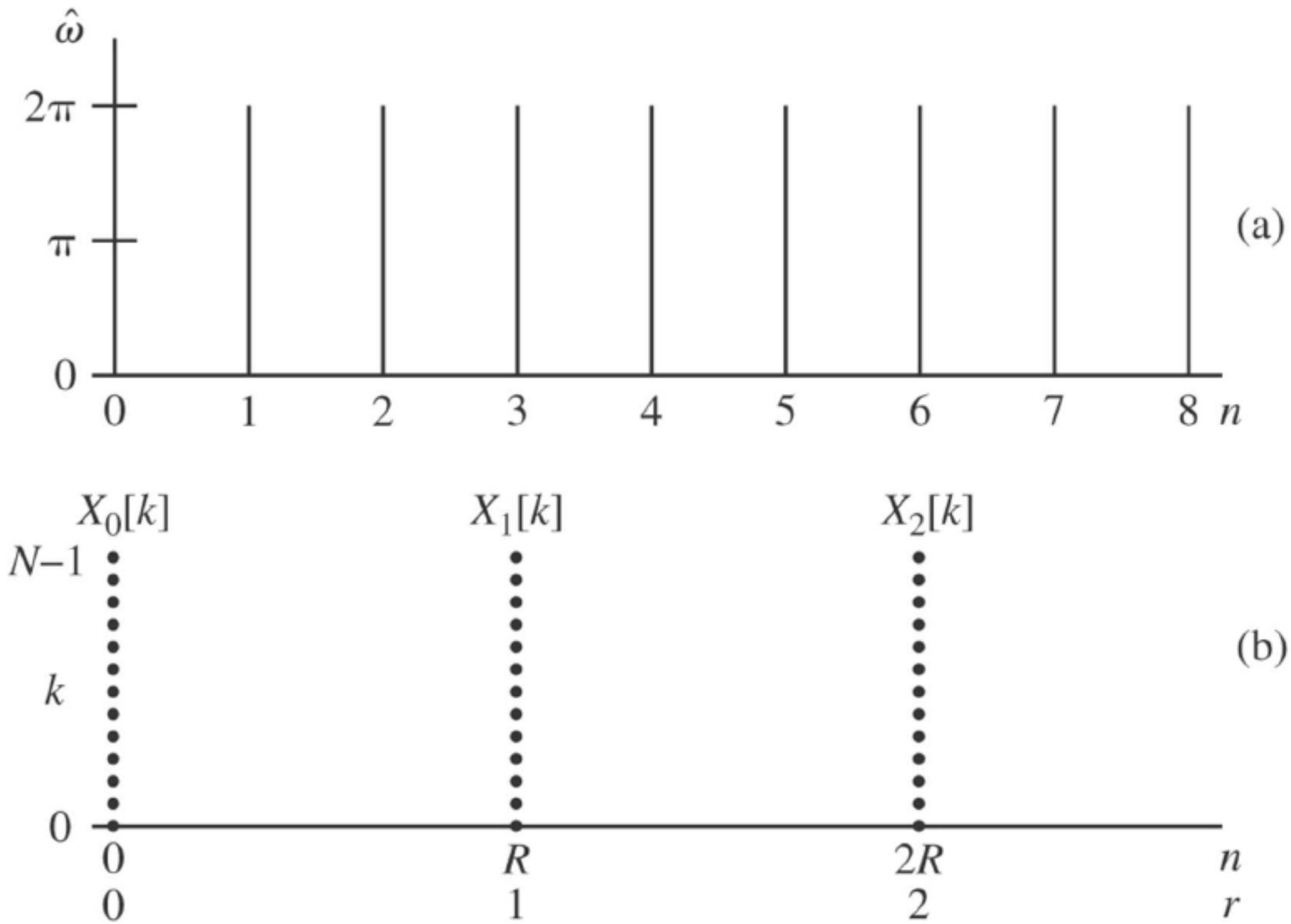


Filter Bank: $X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} \left(x[n-m]e^{-j\hat{\omega}(n-m)} \right) w[m]$

$$X_n(e^{j\hat{\omega}}) = e^{-j\hat{\omega}n} \left[(w[n]e^{j\hat{\omega}n}) * x[n] \right]$$

Sampling Rates of STFT

Sampling the STFT



$$X_n(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}w(n-m)$$
$$= w(n) * \left(x(n)e^{-j\hat{\omega}n}\right), \quad n \text{ variable, } \hat{\omega} \text{ fixed}$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})X(e^{j(\theta+\hat{\omega})})e^{j\theta n}d\theta$$

Sampling Rate in Time

- to determine the sampling rate in time, we take a linear filtering view
  1. $X_n(e^{j\hat{\omega}})$ is the output of a filter with impulse response $\tilde{w}(n)$
  2. $W(e^{j\hat{\omega}})$ is a lowpass response with effective bandwidth of B Hertz
- thus the effective bandwidth of $X_n(e^{j\hat{\omega}})$ is B Hertz => $X_n(e^{j\hat{\omega}})$ has to be sampled at a rate of $2B$ samples/second to avoid aliasing

<u>Example</u>: Hamming Window

$$w(n) = 0.54 - 0.46\cos(2\pi n/(L-1)) \quad 0 \le n \le L-1$$
$$= 0 \quad \text{otherwise}$$

$\Rightarrow B \approx \frac{2F_s}{L}$(Hz); for $L = 400$, $F_s = 10{,}000$ Hz => $B = 50$ Hz => need rate of 100/sec (<u>every 100 samples</u>) for sampling rate in time

Sampling Rate in Frequency

- since $X_n(e^{j\hat{\omega}})$ is periodic in $\hat{\omega}$ with period 2π, it is only necessary to sample over an interval of length 2π
- need to determine an appropriate finite set of frequencies, $\hat{\omega}_k = 2\pi k / N,\ k = 0,1,\ldots,N-1$ at which $X_n(e^{j\hat{\omega}})$ must be specified to exactly recover $x(n)$
- use the Fourier transform interpretation of $X_n(e^{j\hat{\omega}})$
  1. if the window $w(n)$ is time-limited, then the inverse transform of $X_n(e^{j\hat{\omega}})$ is time-limited
  2. the sampling theorem requres that we sample $X_n(e^{j\hat{\omega}})$ in the frequency dimension at a rate of at least twice its ('symmetric') "time width"
  3. since the inverse Fourier transform of $X_n(e^{j\hat{\omega}})$ is the signal $x(m)w(n-m)$ and this signal is of duration L samples (the duration of $w(n)$), then according to the sampling theorem $X_n(e^{j\hat{\omega}})$ must be sampled (in frequency) at the set of frequencies

     $\hat{\omega}_k = \dfrac{2\pi k}{L},\ k = 0,1,\ldots,L-1$ (where $L/2$ is the effective width of the window)

     in order to exactly recover $x(n)$ from $X_n(e^{j\hat{\omega}_k})$

- thus for a Hamming window of duration L=400 samples, we require that the STFT be evaluated at at least 400 uniformly spaced frequencies around the unit circle

"Total" Sampling Rate of STFT

- the "total" sampling rate for the STFT is the product of the sampling rates in time and frequency, i.e.,

  $SR = SR(\text{time}) \times SR(\text{frequency})$

  $= 2B \times L$ samples/sec

  B = frequency bandwidth of window (Hz)

  L = time width of window (samples)
- for most windows of interest, B is a multiple of F_S/L, i.e.,

  $B = C\, F_S/L$ (Hz), C=1 for Rectangular Window

  C=2 for Hamming Window

  $SR = 2C\, F_S$ samples/second
- can define an 'oversampling rate' of

  $SR/F_S = 2C$ = oversampling rate of STFT as compared to conventional sampling representation of $x(n)$

  for RW, $2C$=2; for HW $2C$=4 => range of oversampling is 2-4

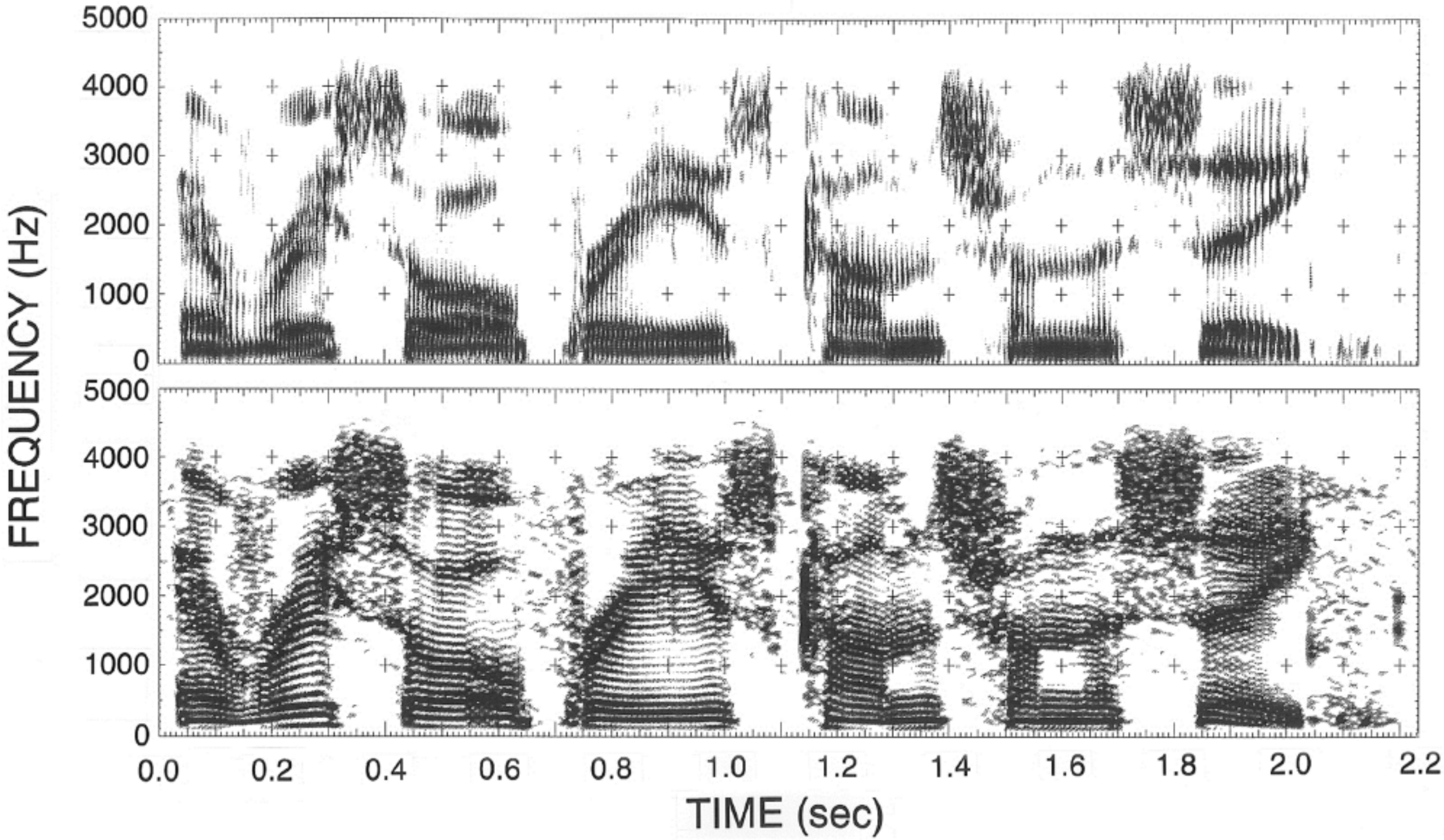
  this <u>*oversampling*</u> gives a <u>*very flexible representation*</u> of the speech signal

Spectrographic Displays

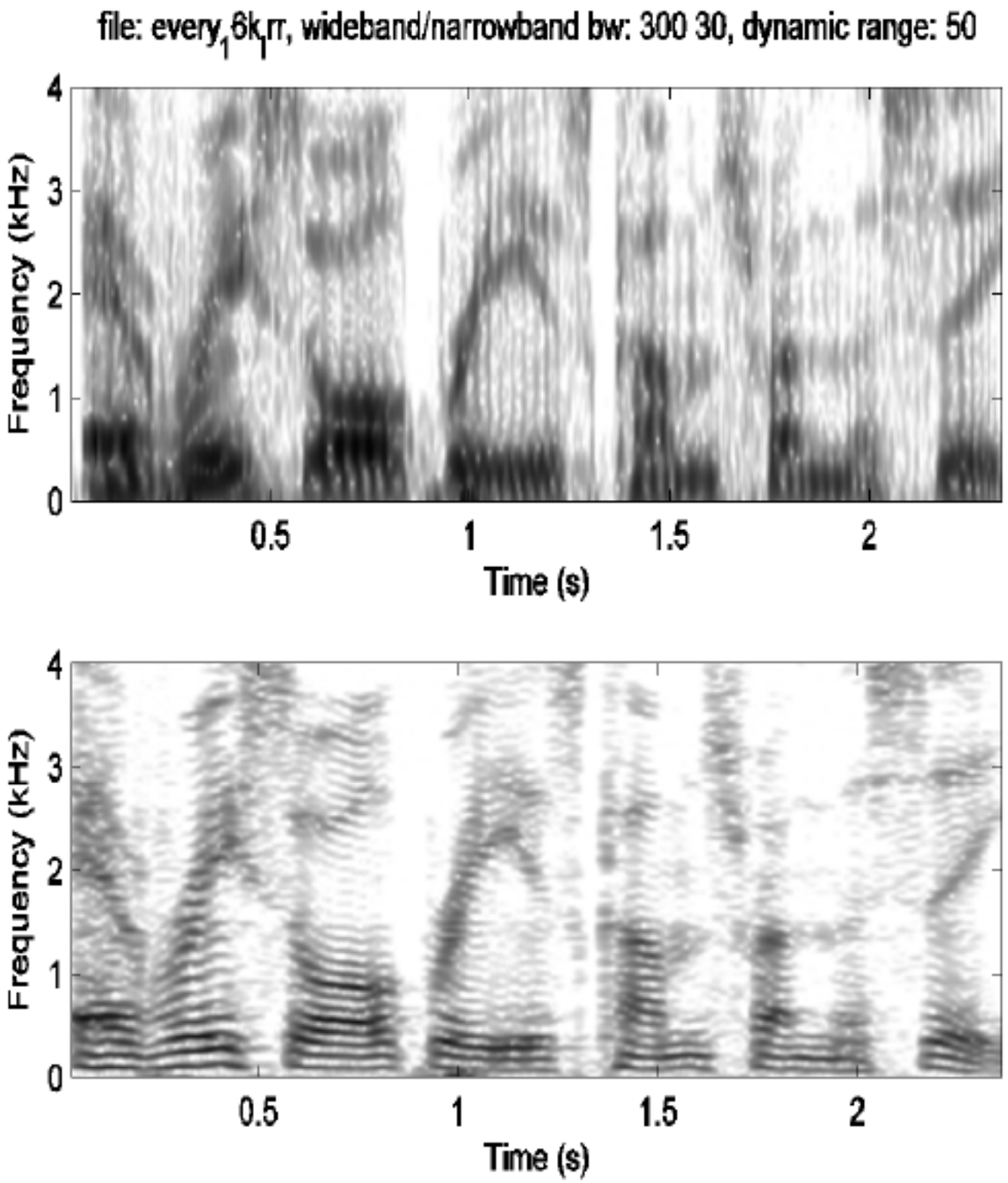
Spectrographic Displays

- ***Sound Spectrograph***-one of the earliest embodiments of the time-dependent spectrum analysis techniques
  - 2-second utterance repeatedly modulates a variable frequency oscillator, then bandpass filtered, and the average energy at a given time and frequency is measured and used as a crude measure of the STFT
  - thus energy is recorded by an ingenious electro-mechanical system on special electrostatic paper called teledeltos paper
  - result is a two-dimensional representation of the time-dependent spectrum-with vertical intensity being spectrum level at a given frequency, and horizontal intensity being spectral level at a given time-with spectrum magnitude being represented by the darkness of the marking
  - wide bandpass filters (300 Hz bandwidth) provide good temporal resolution and poor frequency resolution (resolve pitch pulses in time but not in frequency)—called wideband spectrogram
  - narrow bandpass filters (45 Hz bandwidth) provide good frequency resolution and poor time resolution (resolve pitch pulses in frequency, but not in time)—called narrowband spectrogram

Conventional Spectrogram (Every salt breeze comes from the sea)



Digital Speech Spectrograms



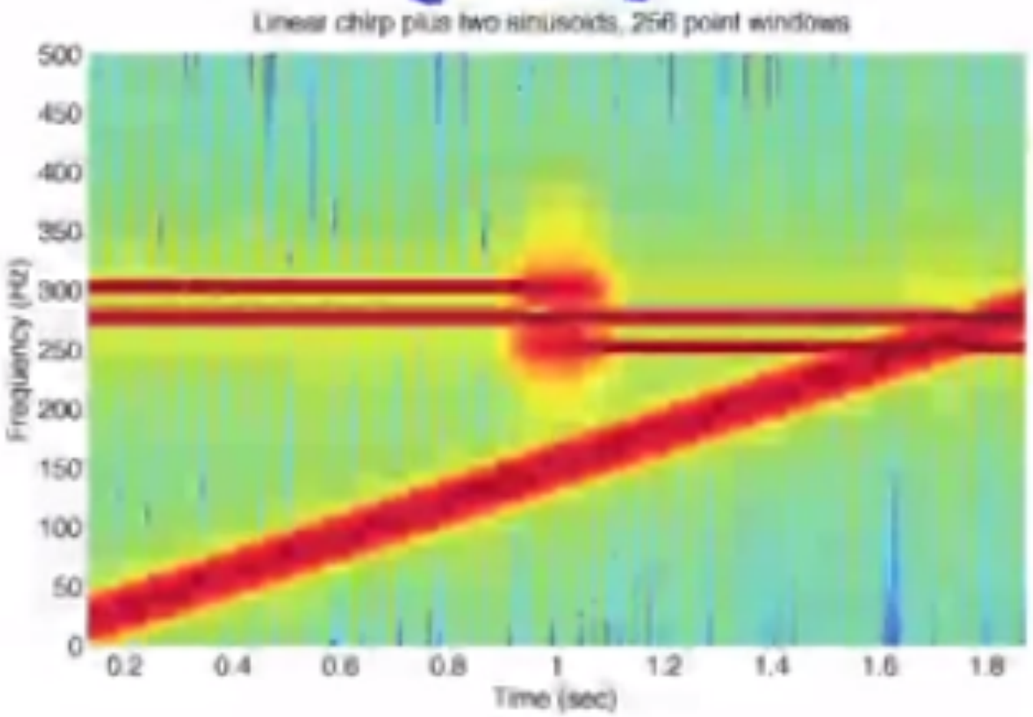
- **wideband spectrogram**
  - follows broad spectral peaks (formants) over time
  - resolves most individual pitch periods as vertical striations since the IR of the analyzing filter is comparable in duration to a pitch period
  - what happens for low pitch males—high pitch females
  - for unvoiced speech there are no vertical pitch striations
- **narrowband spectrogram**
  - individual harmonics are resolved in voiced regions
  - formant frequencies are still in evidence
  - usually can see fundamental frequency
  - unvoiced regions show no strong structure

Digital Speech Spectrograms

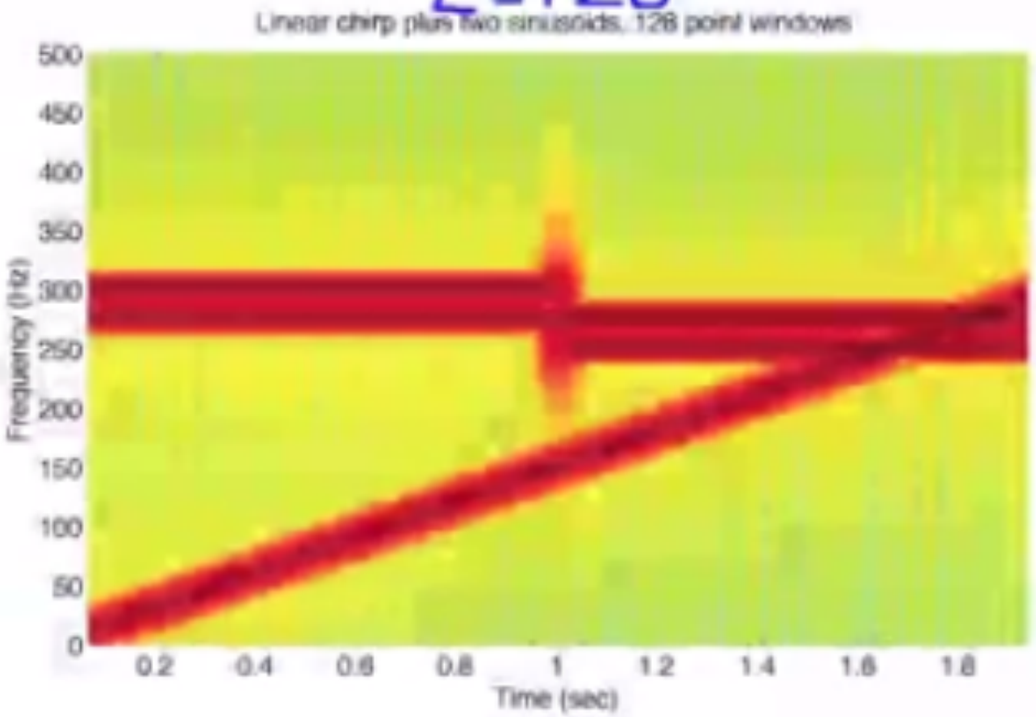
L trades temporal resolution for frequency resolution 4

Example: linear chirp 0-300 Hz; sin #1 : 275 Hz ;
sin #2 : 300 Hz 0-1s, 250 Hz 1-2s ; f_s = 1000 Hz

L = 256



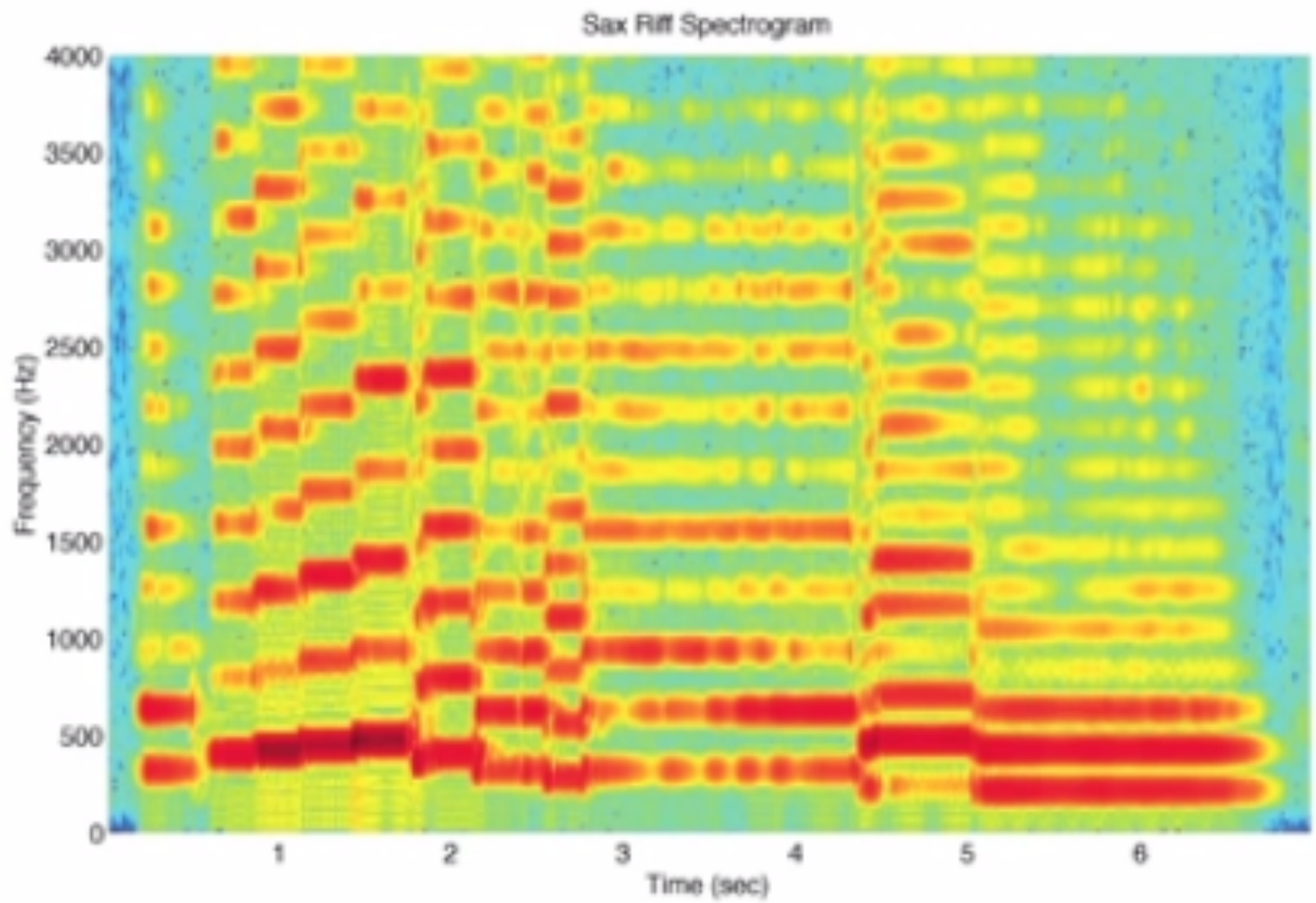
L = 128



Digital Speech Spectrograms

Saxophone Riff

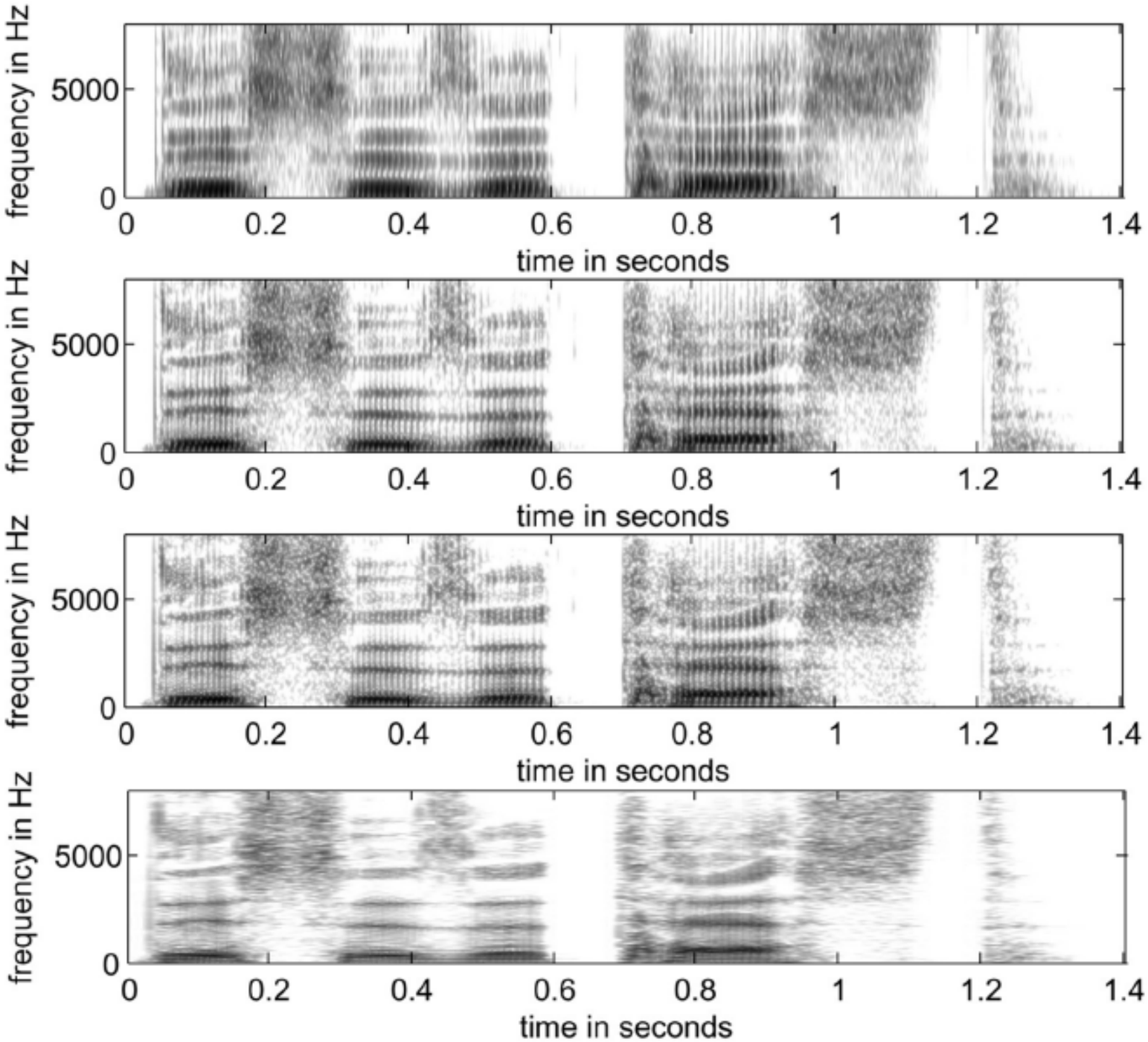
- $f_s = 44{,}100$ kHz
- $L = 812$ pts/segment
- $N = 2048$ freq. samples
- 192 point overlap



Digital Speech Spectrograms

- Speech Parameters ("This is a test"):
  - sampling rate: 16 kHz
  - speech duration: 1.406 seconds
  - speaker: male
- Wideband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 6 msec (96 samples)
  - analysis window shift: 0.625 msec (10 samples)
  - FFT size: 512
  - dynamic range of spectral log magnitudes: 40 dB
- Narrowband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 60 msec (960 samples)
  - analysis window shift: 6 msec (96 samples)
  - FFT size: 1024
  - dynamic range of spectral log magnitudes: 40 dB

Digital Speech Spectrograms



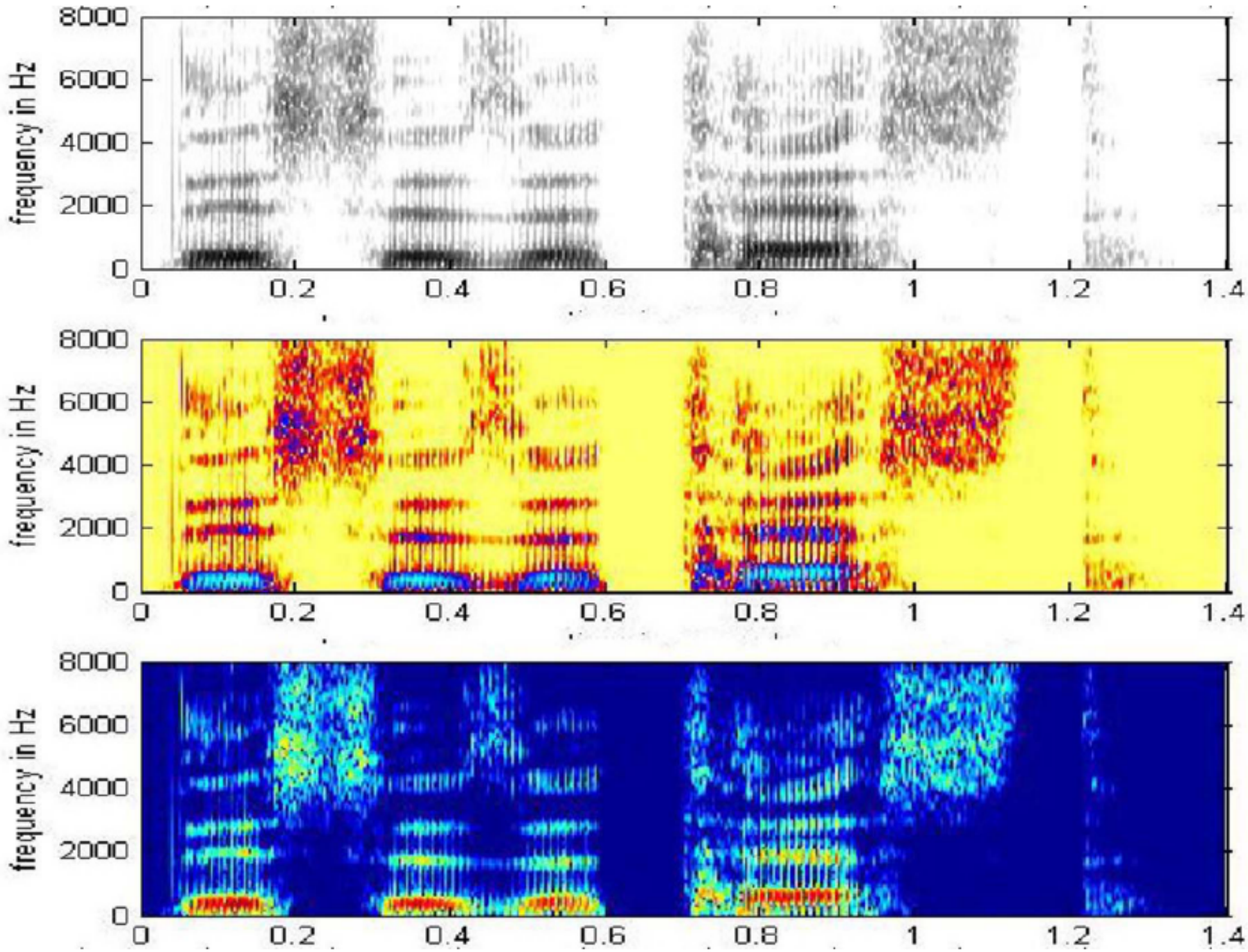
Top Panel:
3 msec (48 samples) window

Second Panel:
6 msec (96 samples) window

Third Panel:
9 msec (144 sample) window

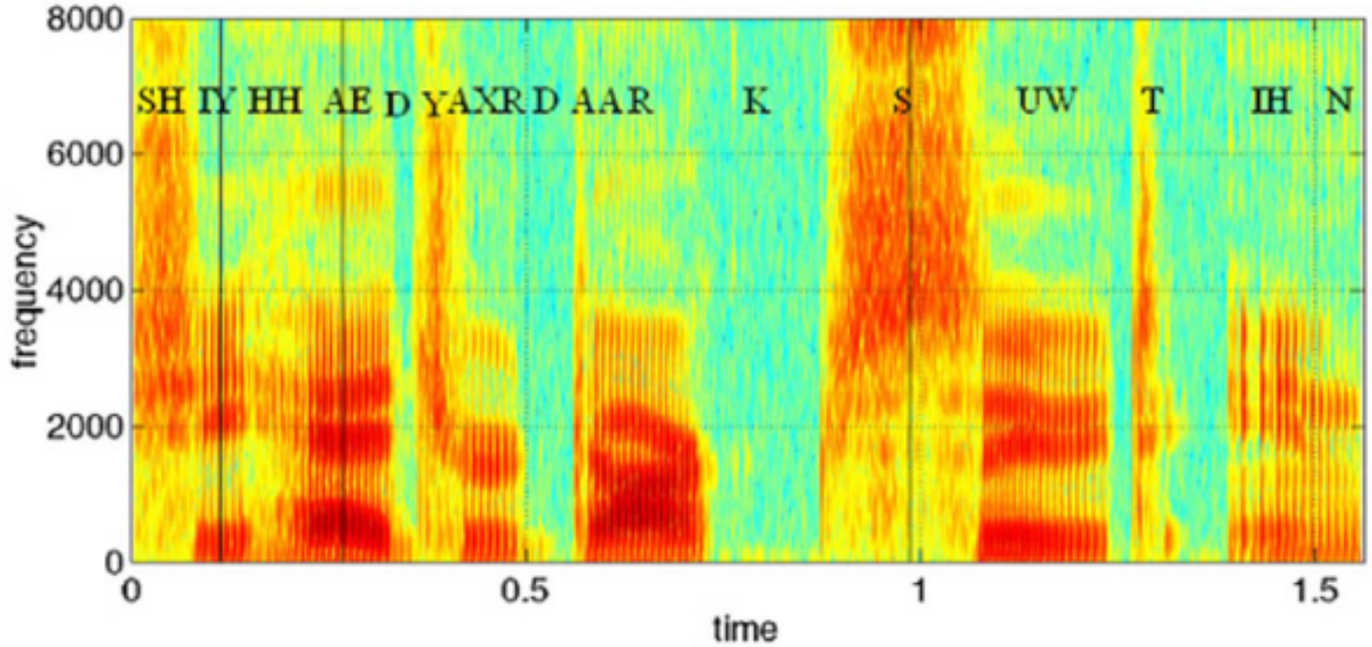
Fourth Panel:
30 msec (480 sample) window

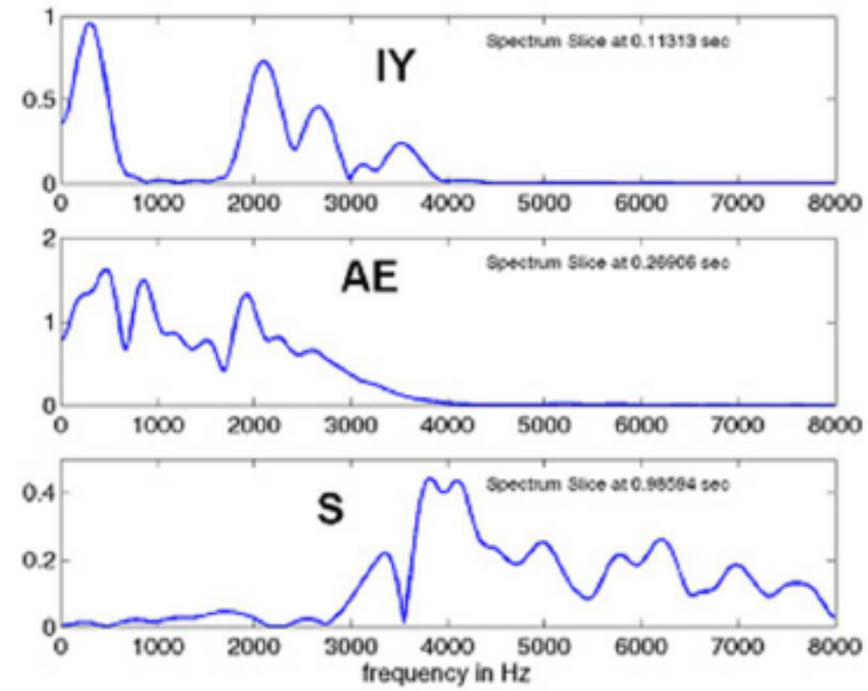
Spectrogram Comparisons

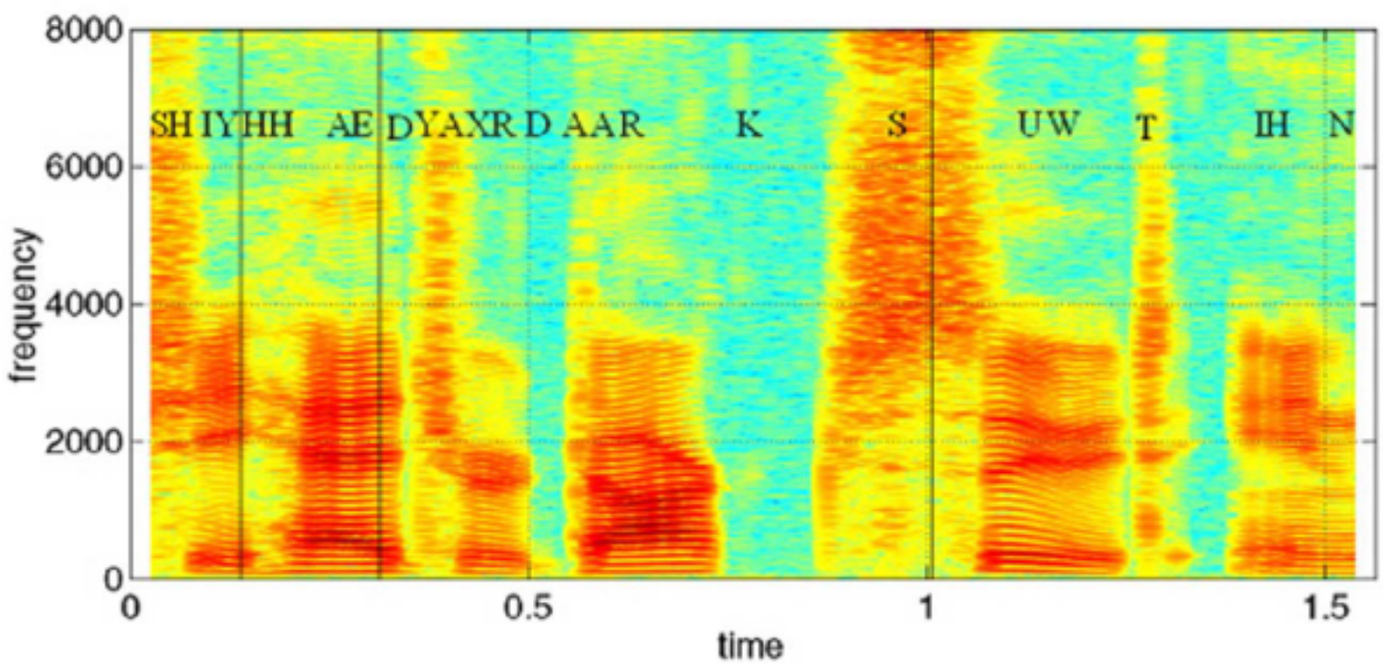


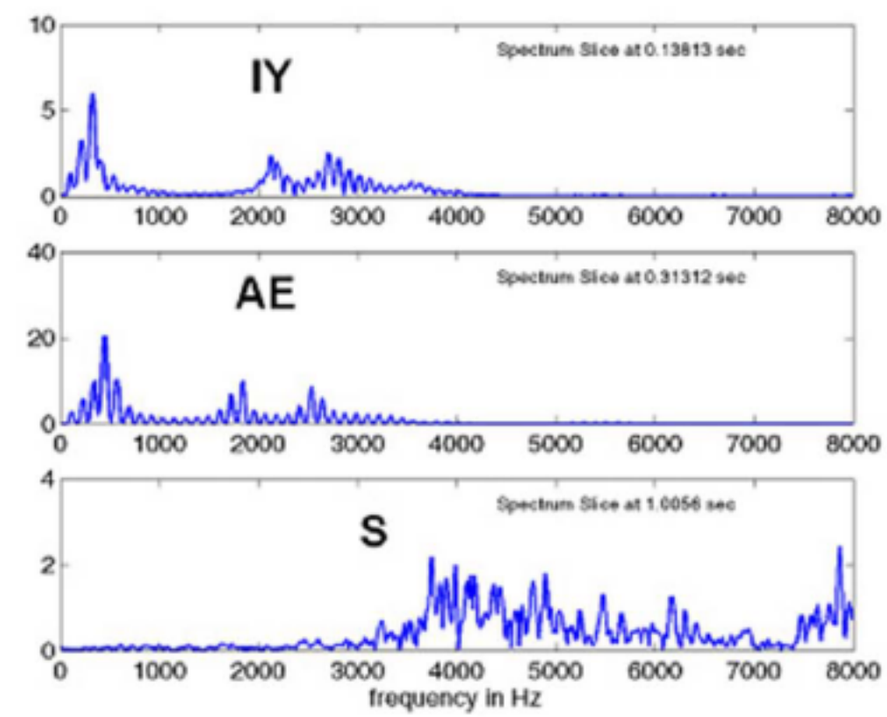
Spectrogram - Male

nfft $= 1024, \quad L = 80, \quad Overlap = 75$



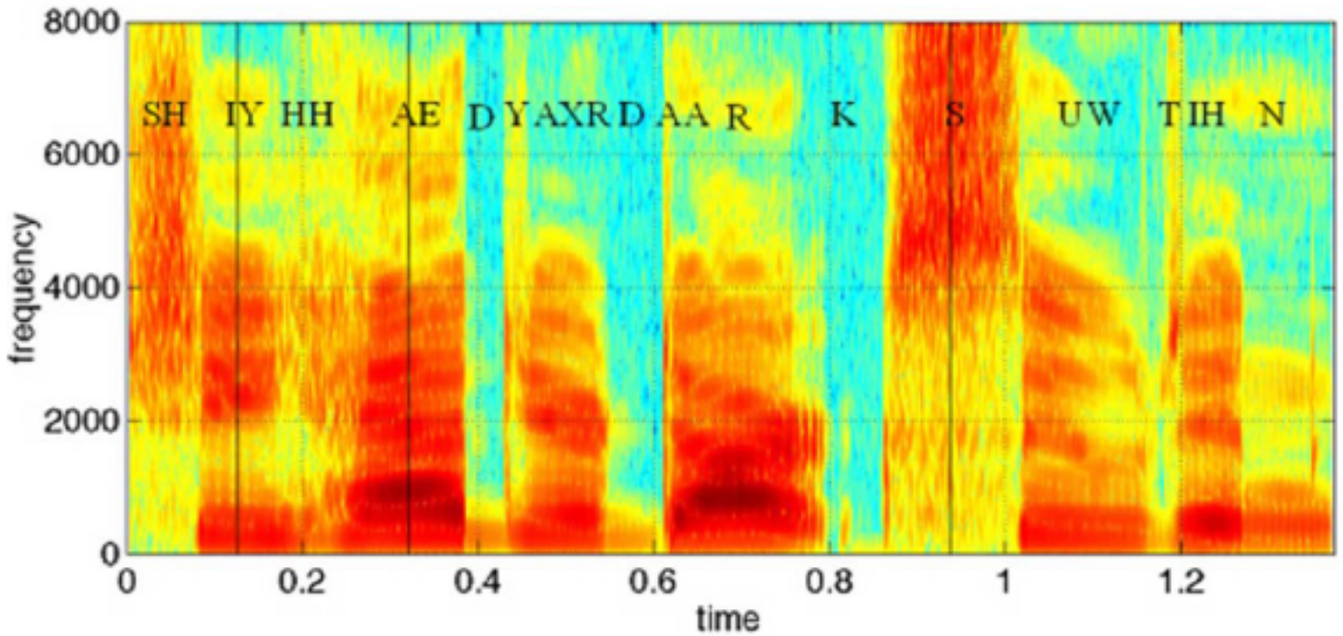


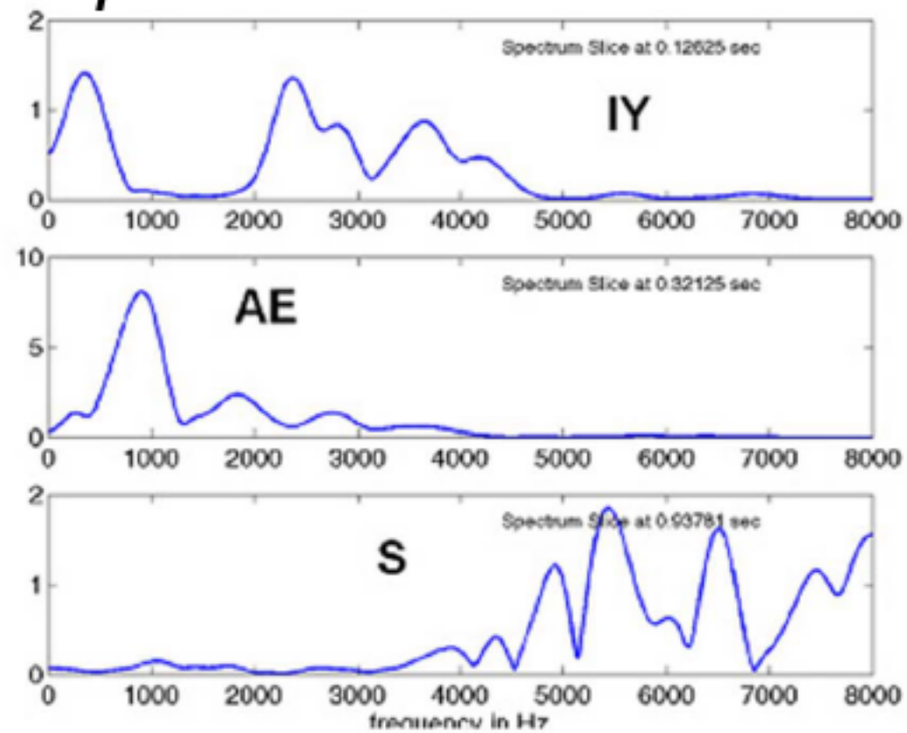


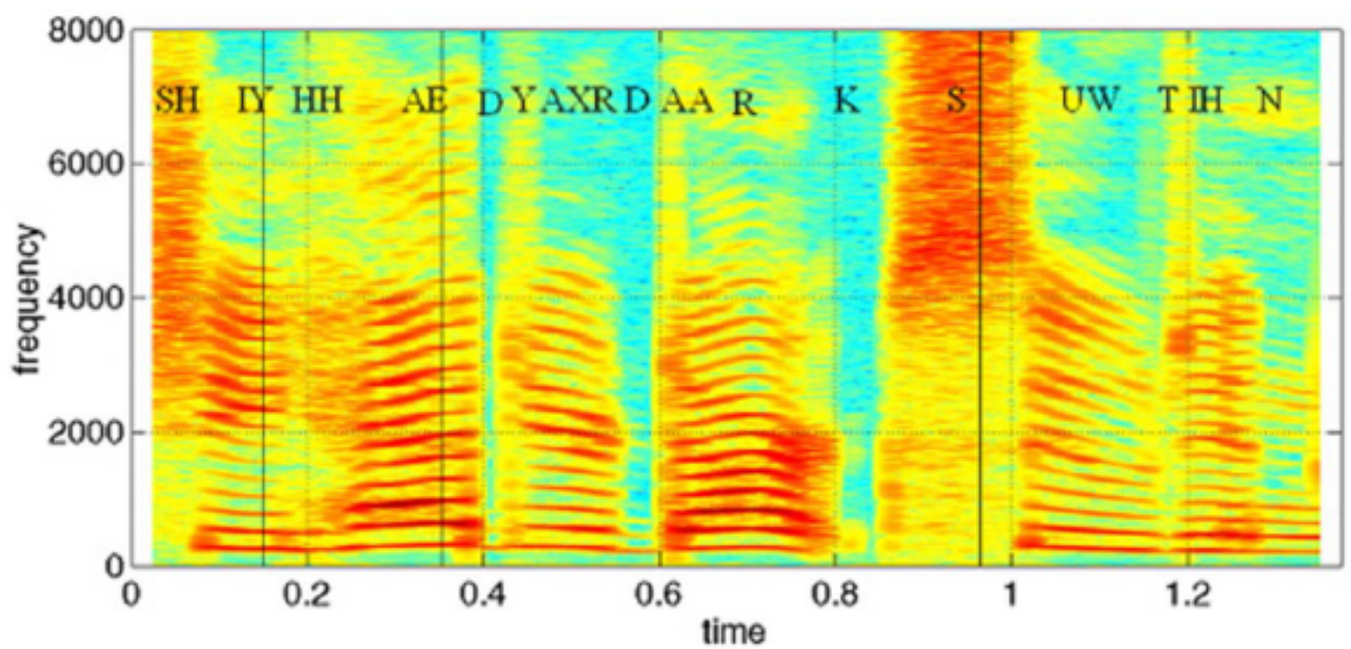


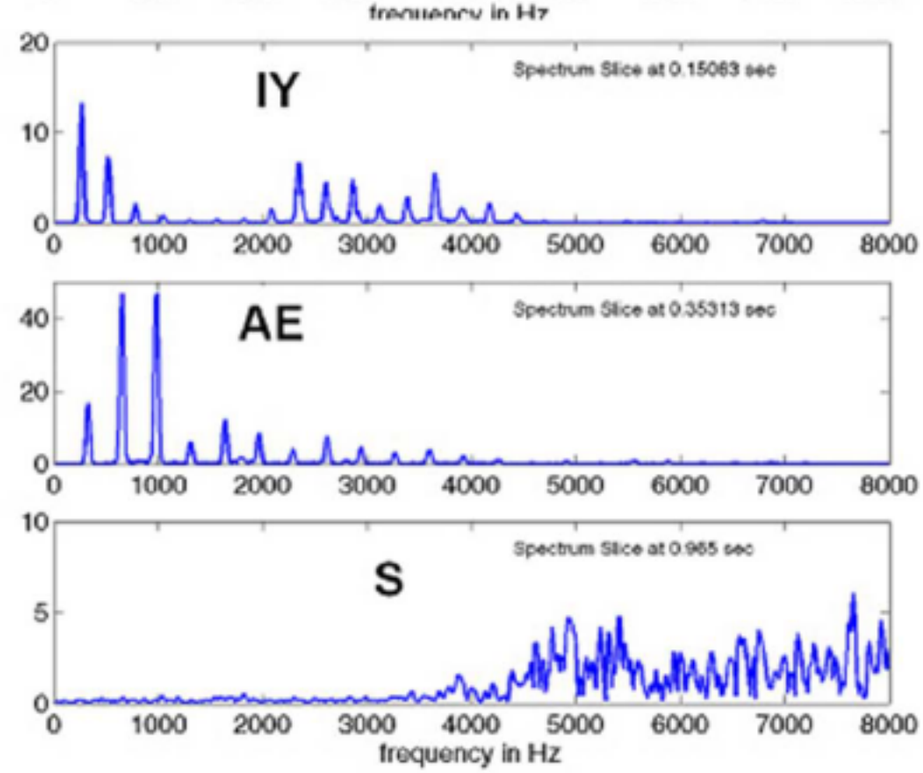
Spectrogram - Female

nfft $= 1024, \quad L = 80, \quad Overlap = 75$



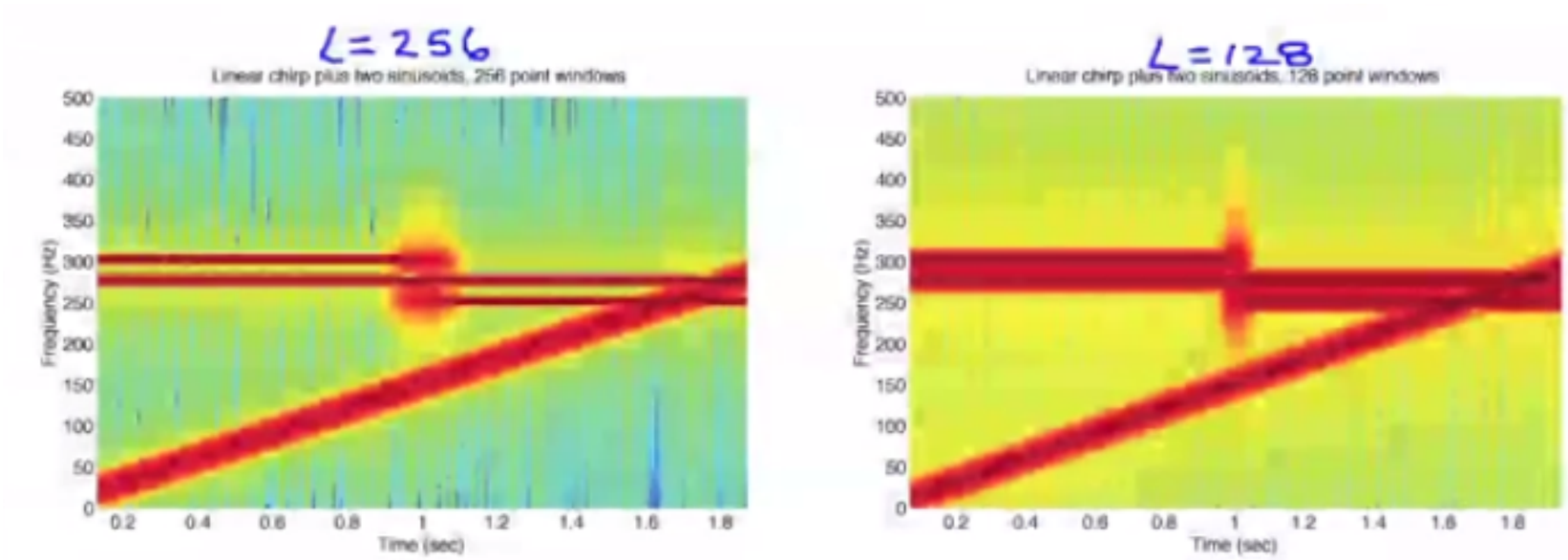


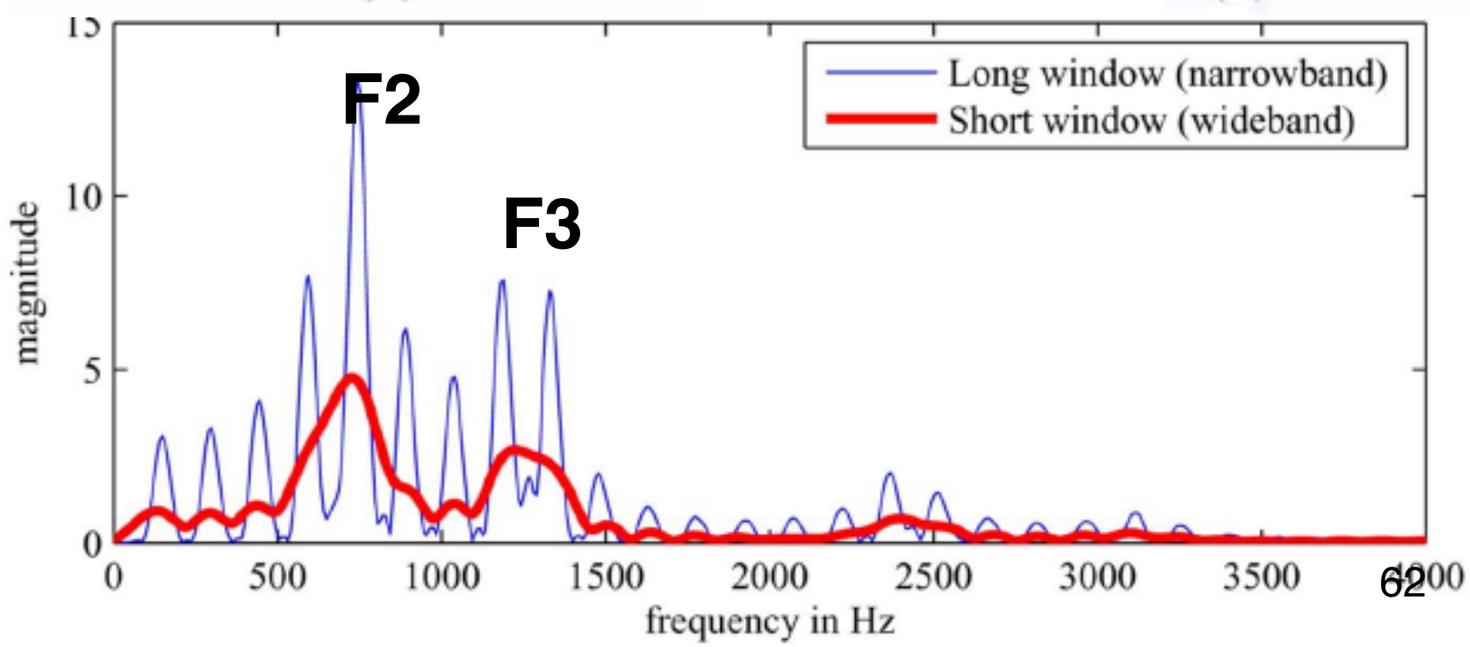




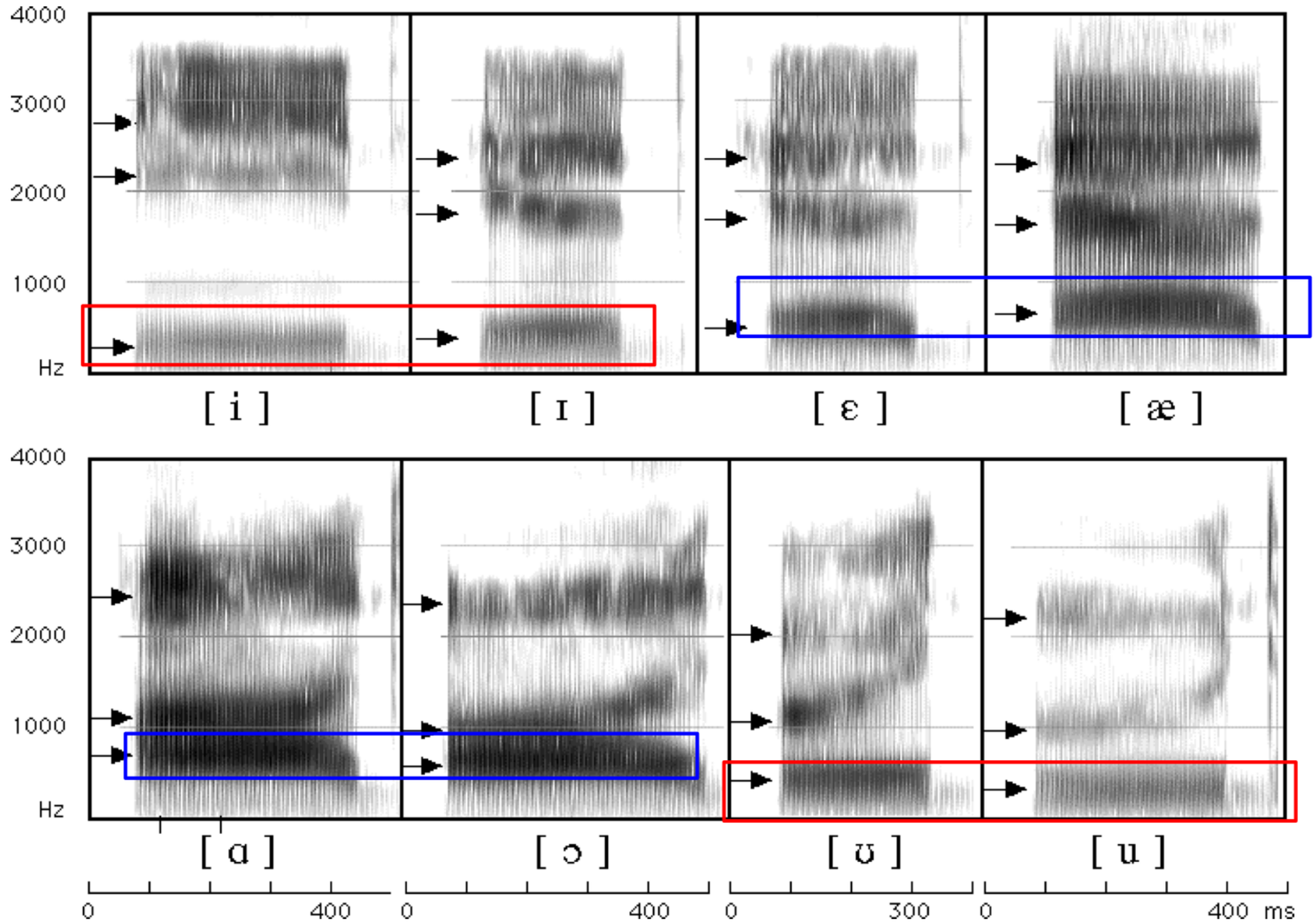
Formant extraction

- In speech, the resonant frequencies of the vocal tract (that is the frequencies that resonate the loudest) are called formants.
- Can be easily extracted from wideband spectrogram



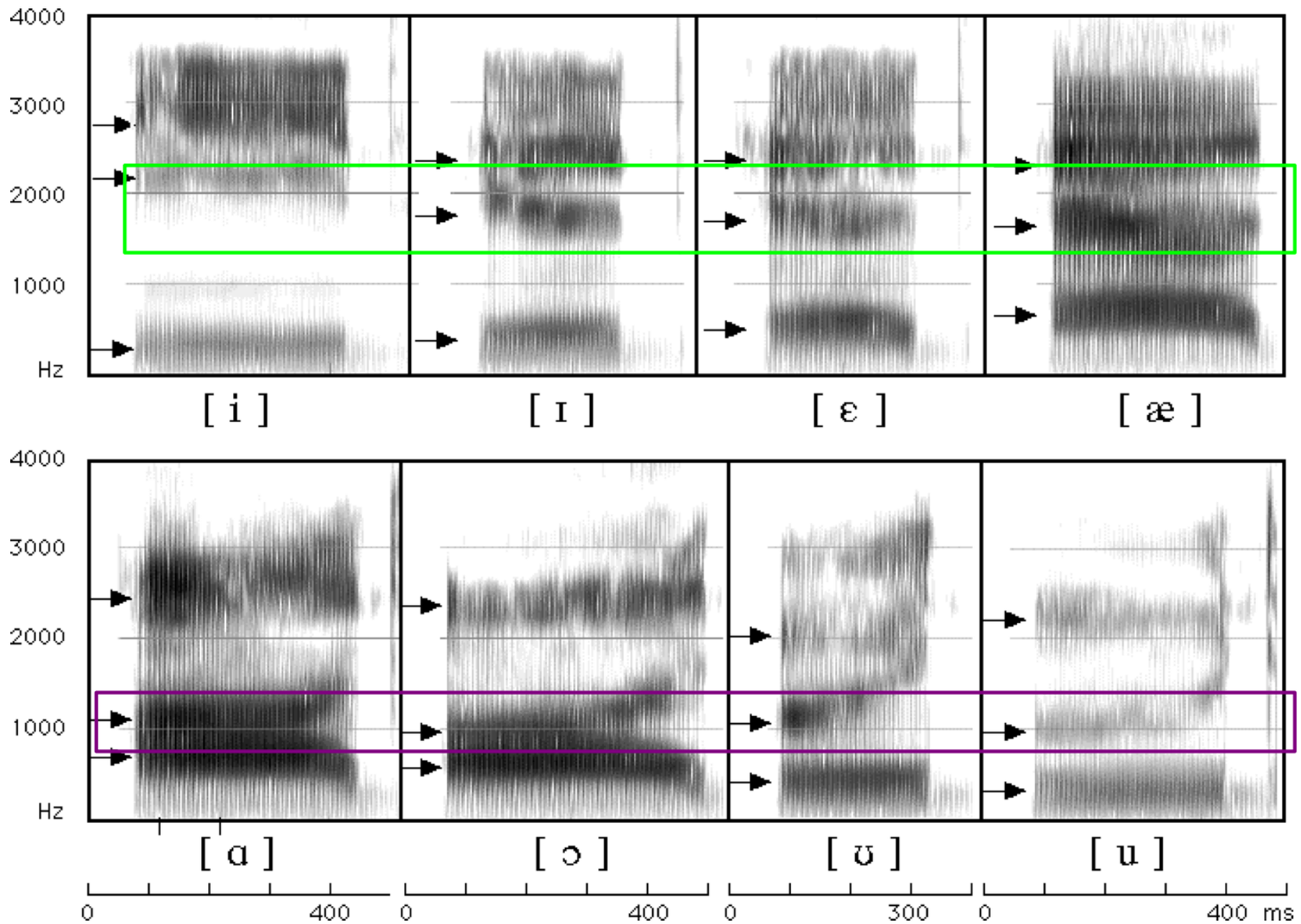


Formant extraction



First formant frequency F1

Formant extraction



First formant frequency F2

Pitch detection

Harmonic product spectrum

$$P_n(e^{j\omega}) = \prod_{r=1}^{K} |X_n(e^{j\omega r})|^2$$

Log Harmonic product spectrum

$$\hat{P}_n(e^{j\omega}) = 2 \sum_{r=1}^{K} \log|X_n(e^{j\omega r})|$$

- Voiced excitation can be seen at sharp picks that occur at integer multiple of fundamental frequency.
- For voiced speech, compressing the frequency scale by integer factors cause harmonics of the fundamental frequency to coincide at fundamental frequency.
- For the continuous function $|X_n(e^{j2\pi FTr})|$, the peak becomes sharper as r increases.
- This technique is simple and resistant to additive noise

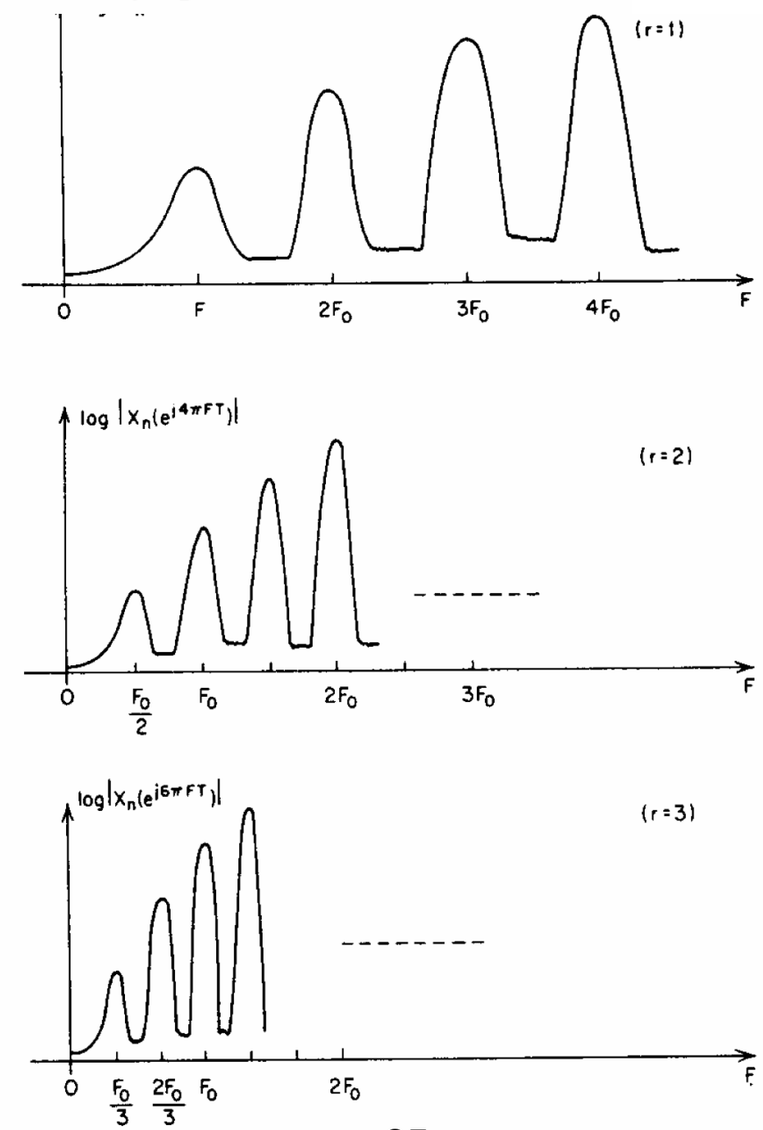
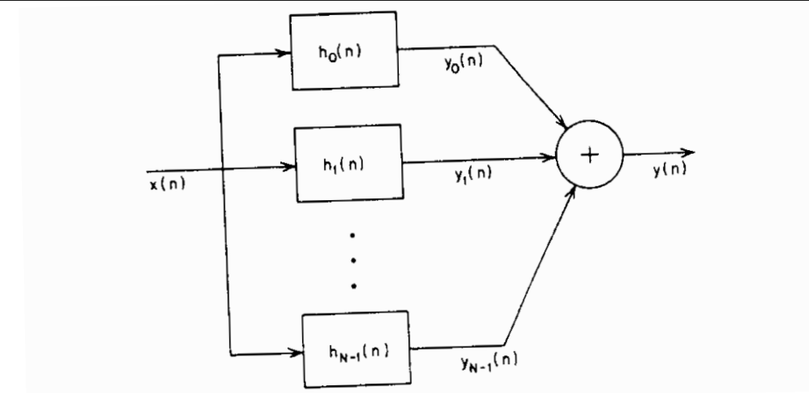


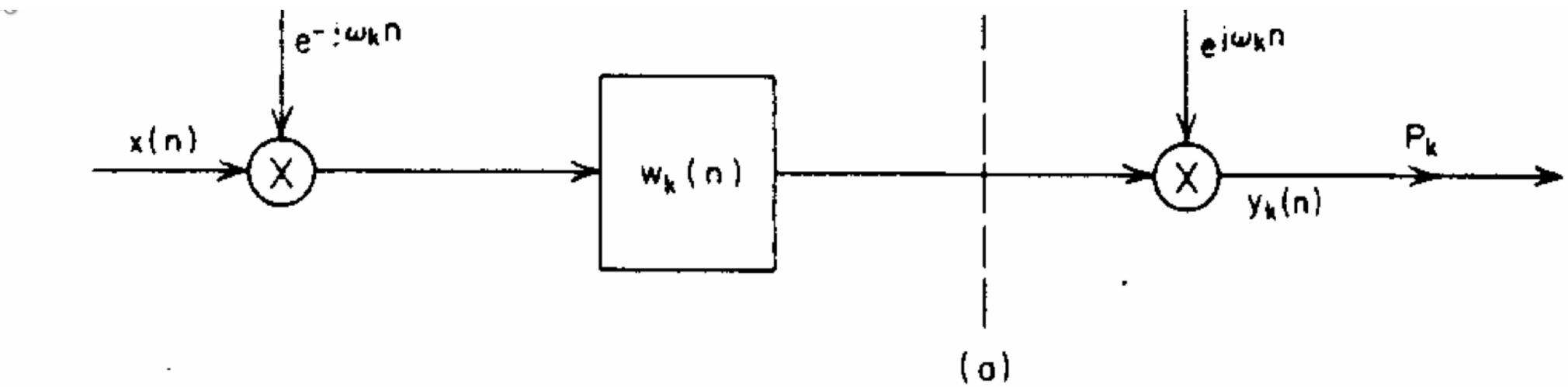
Fig. 6.41 Representations of terms in the log harmonic product spectrum.

Analysis synthesis system

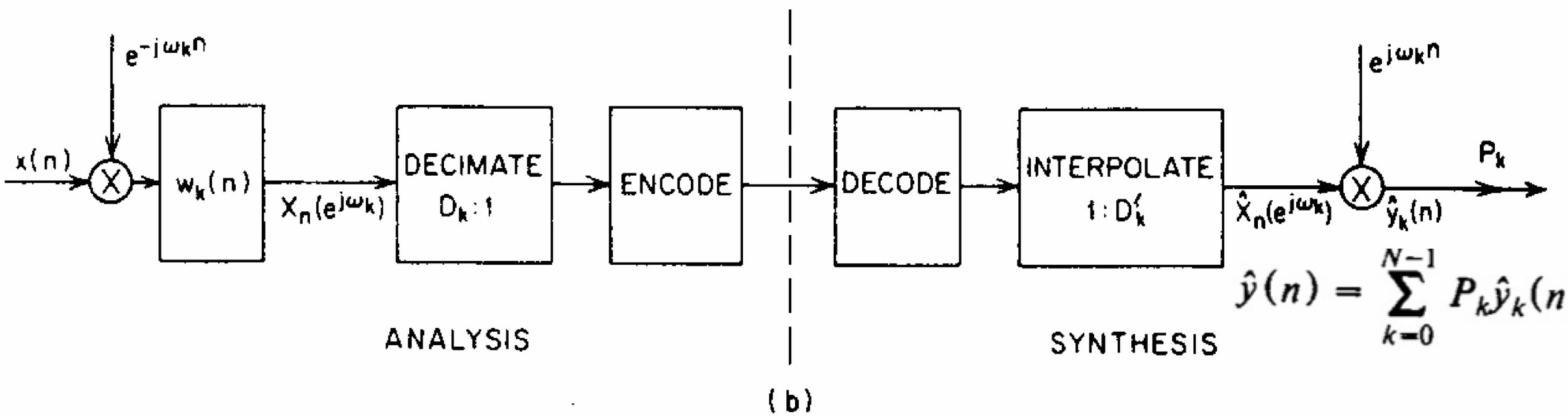
- The time-dependent Fourier transform is useful tool for estimating the parameters of the model for speech production.
- The speech signal can be exactly recovered from the time-dependent Fourier representation and is basis for vocoder systems.
- The main purpose of vocoder systems is to obtain a digital representation of speech at a much lower bit rate than is possible with waveform coding schemes.
- Vocoder includes the removal of additive noise or effect of reverberation and the modification of basic speech parameters to permit alteration of the time and frequency dimensions of a speech.



Analysis synthesis system



$$P_k = |P_k| e^{j\phi_k}$$

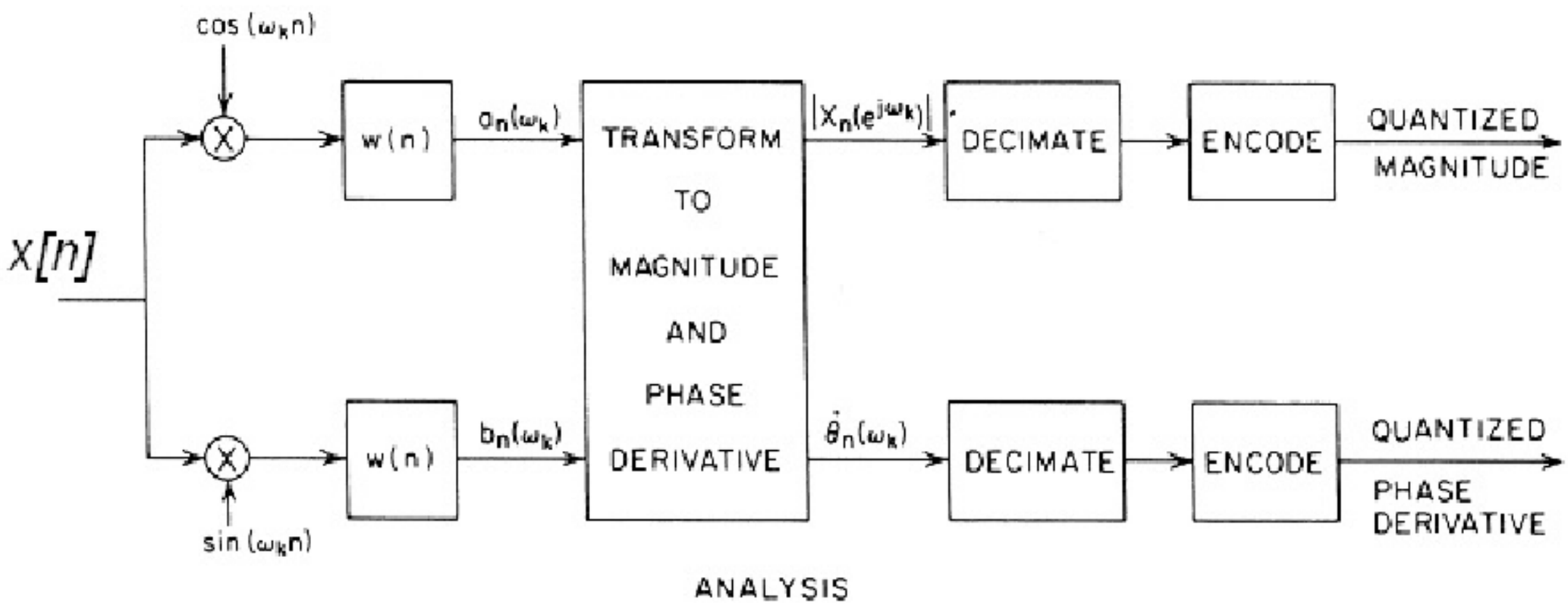


$$\hat{y}(n) = \sum_{k=0}^{N-1} P_k \hat{y}_k(n)$$

Fig. 6.45 Block diagram of the coding for a single analysis-synthesis channel

Phase Vocoder

- A phase vocoder is a type of vocoder which can scale both the frequency and time domains of audio signals by using phase information.



Complete single channel for a phase vocoder analyzer

Phase Vocoder

$$X_a(t, \Omega_k) = |X_a(t, \Omega_k)| e^{j\theta_a(t, \Omega_k)}$$
$$= a_a(t, \Omega_k) - jb_a(t, \Omega_k)$$

where

$$|X_a(t, \Omega_k)| = [a_a^2(t, \Omega_k) + b_a^2(t, \Omega_k)]^{1/2}$$

$$\dot{\theta}_a(t, \Omega_k) = \frac{d\theta_a(t, \Omega_k)}{dt}$$

and

$$\theta_a(t, \Omega_k) = -\tan^{-1}\left[\frac{b_a(t, \Omega_k)}{a_a(t, \Omega_k)}\right]$$

$$\dot{\theta}_a(t, \Omega_k) = \frac{b_a(t, \Omega_k)\dot{a}_a(t, \Omega_k) - a_a(t, \Omega_k)\dot{b}_a(t, \Omega_k)}{a_a^2(t, \Omega_k) + b_a^2(t, \Omega_k)}$$

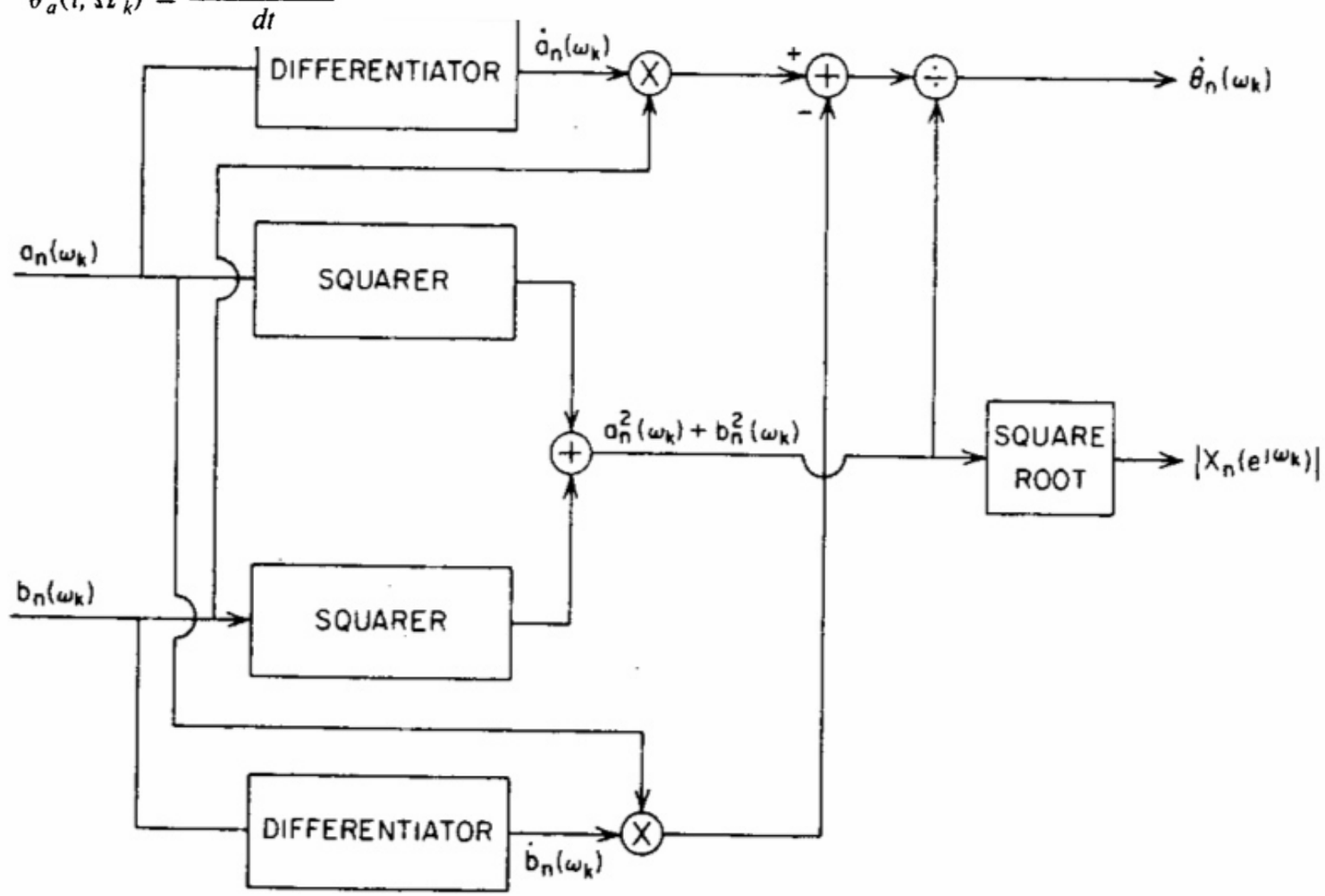
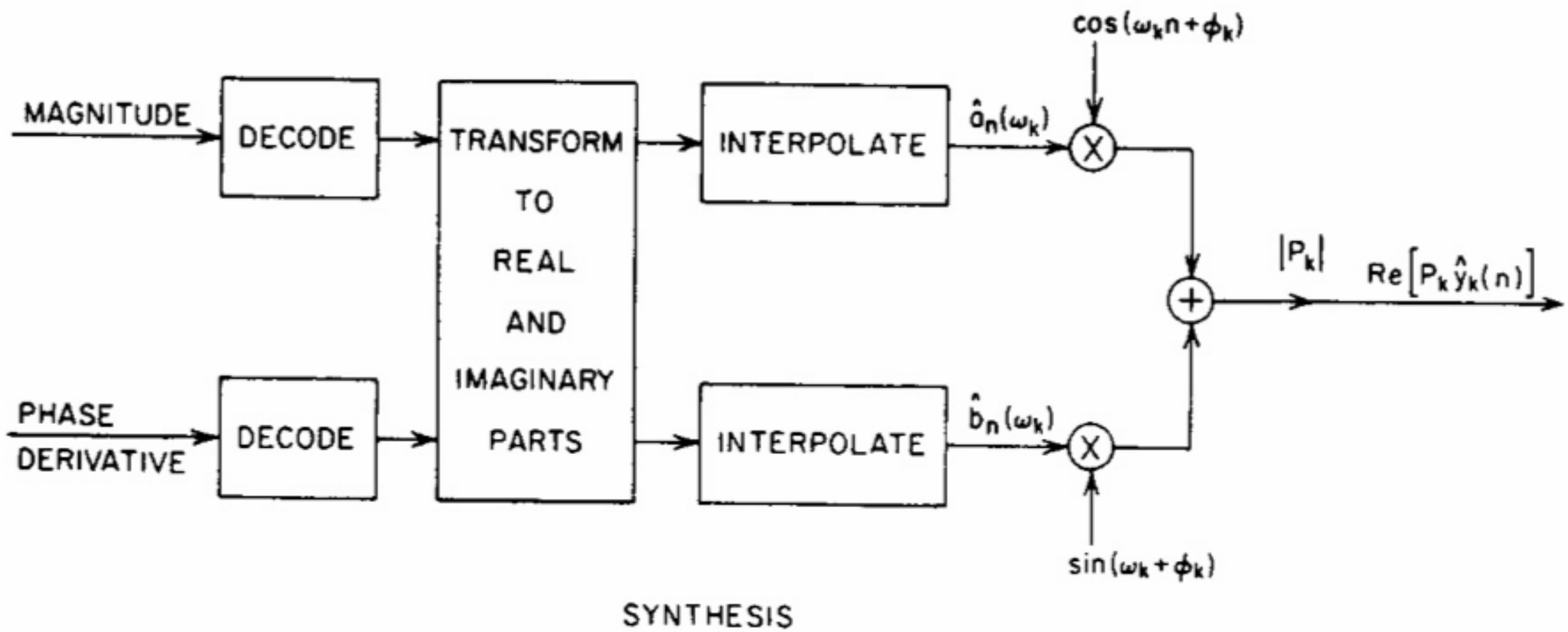


Fig. 6.55 Conversion from a and b to $\dot{\theta}$ and $|X(e^{j\omega})|$.

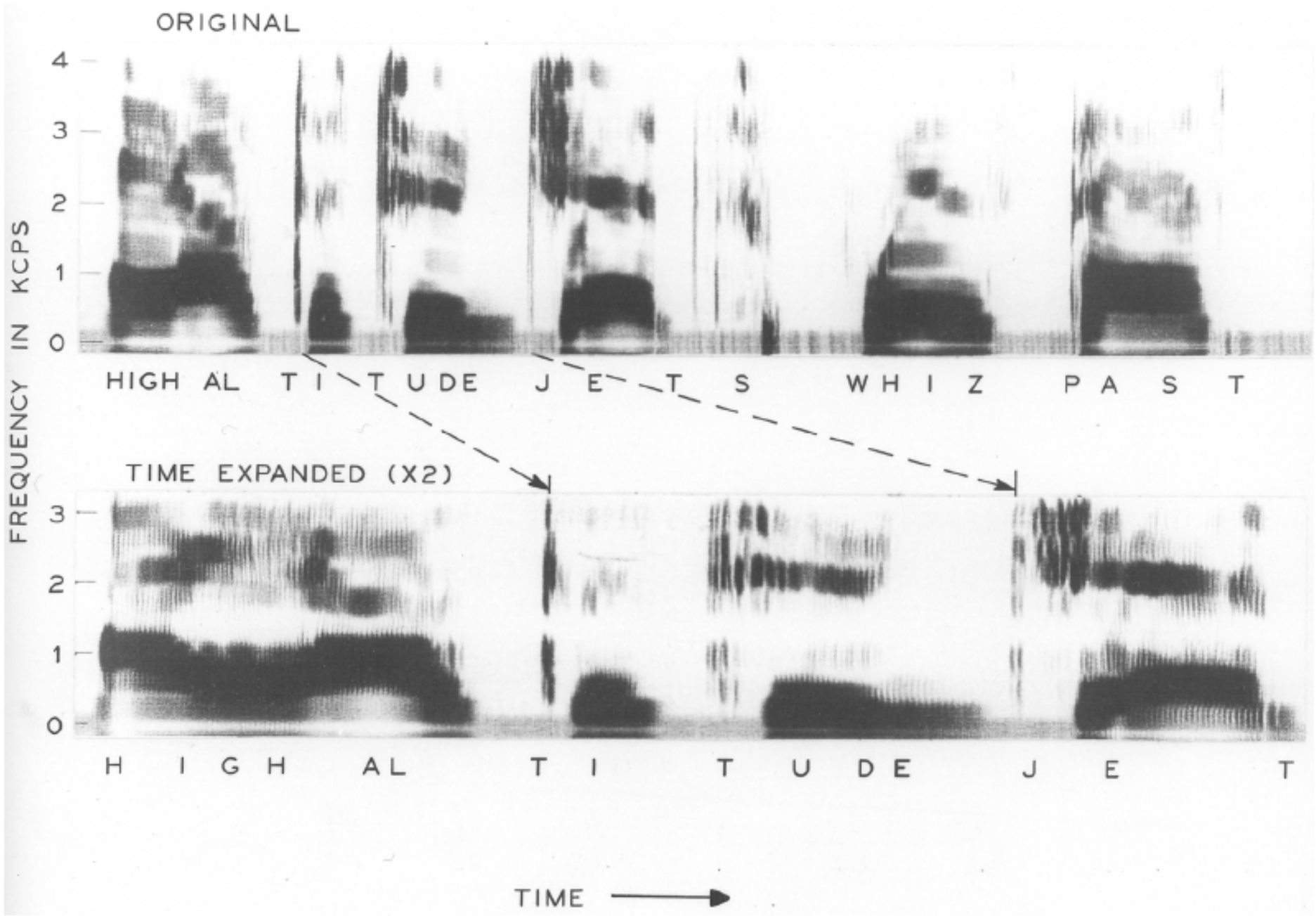
Phase Vocoder



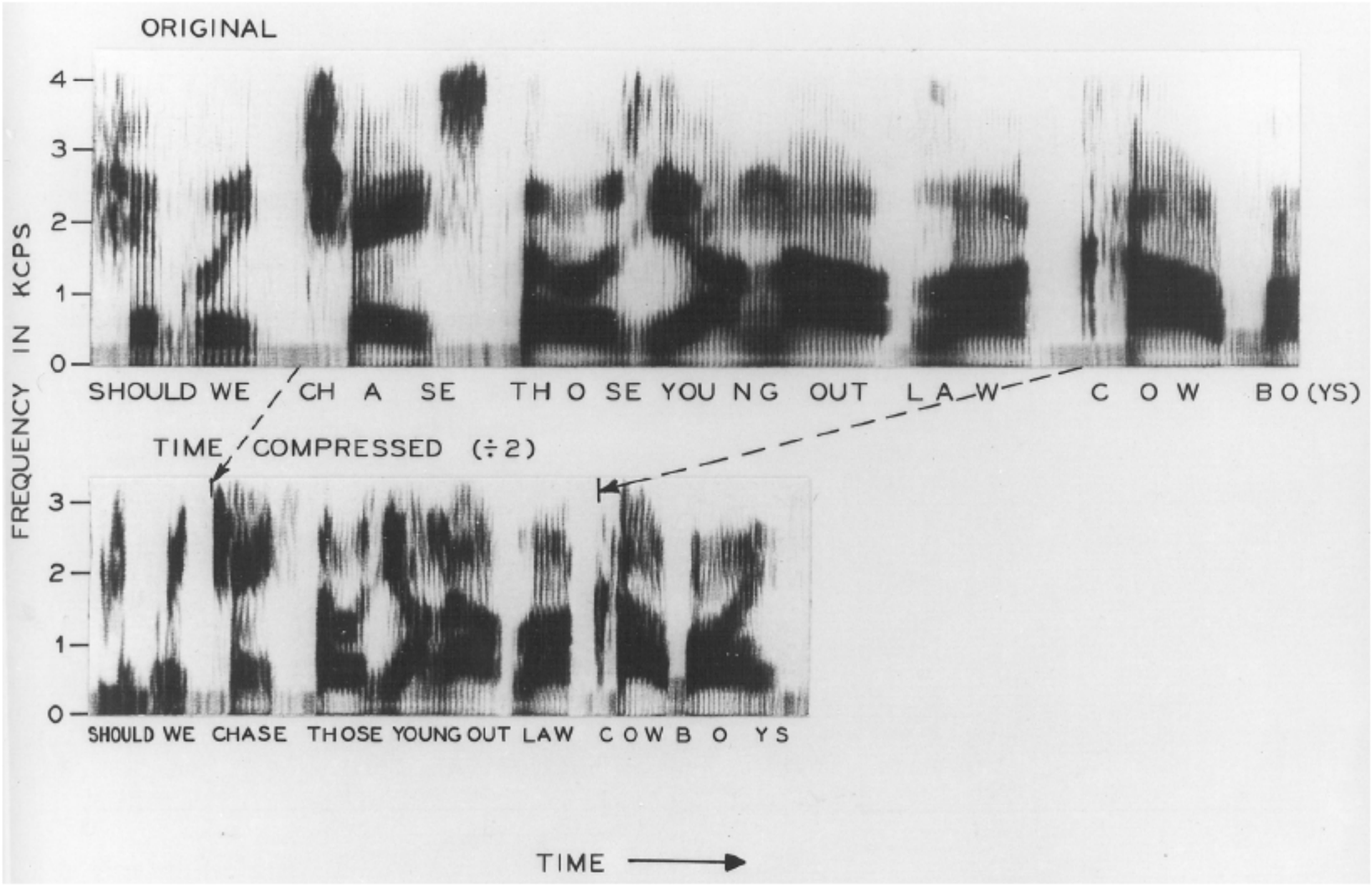
Complete single channel for a phase vocoder synthesizer

$$R_e[P_k y_k(n)] = |P_k||X_n(e^{j\omega_k})|\cos[\omega_k n + \theta_n(\omega_k) + \phi_k]$$

Phase Vocoder Time Expanded



Phase Vocoder Time Compressed

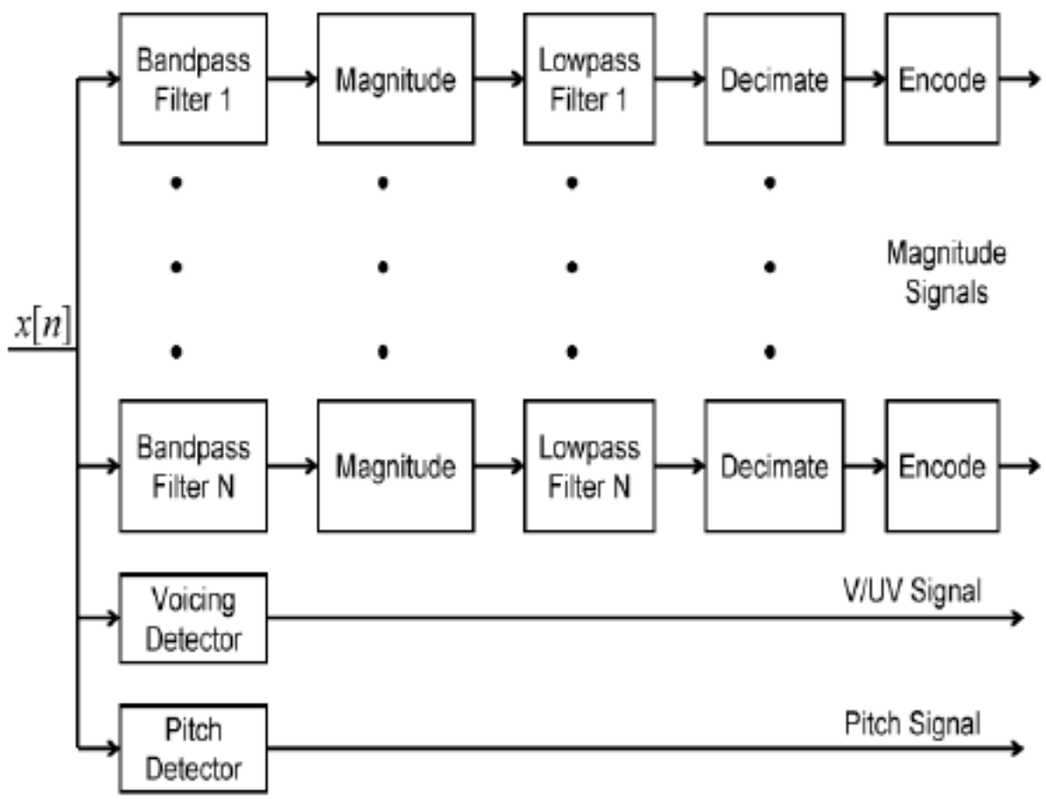


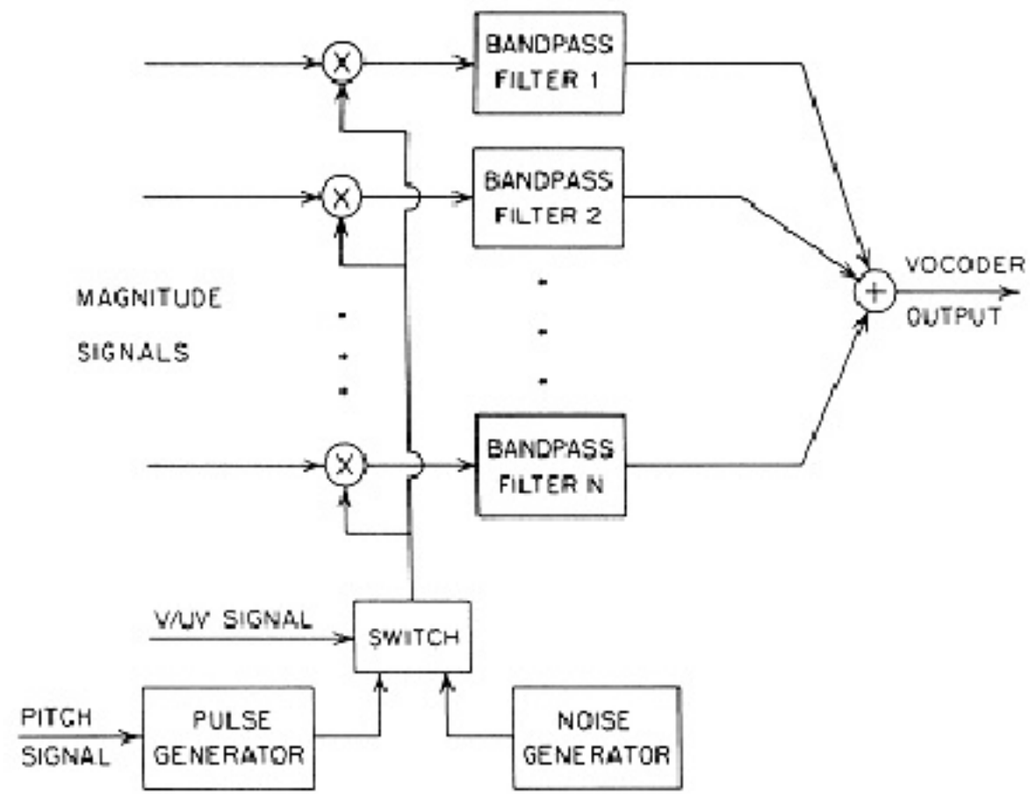
Channel Vocoder

- interpret STFT so that each channel can be thought of as a a bandpass filter with center frequency ω_k
- magnitude of STFT can be approximated by envelope detection on the BPF output (FWR and LPF)
- ***analyzer***-bank of channels; need excitation info (the phase component) => V/UV detector, pitch detector
- ***synthesizer***-channel signal control channel amplitude; excitation signals control detailed structure of output for a given channel; V/UV choice of excitation source

=> highly reverberant speech because of total lack of control of composite filter bank response

Channel Vocoder

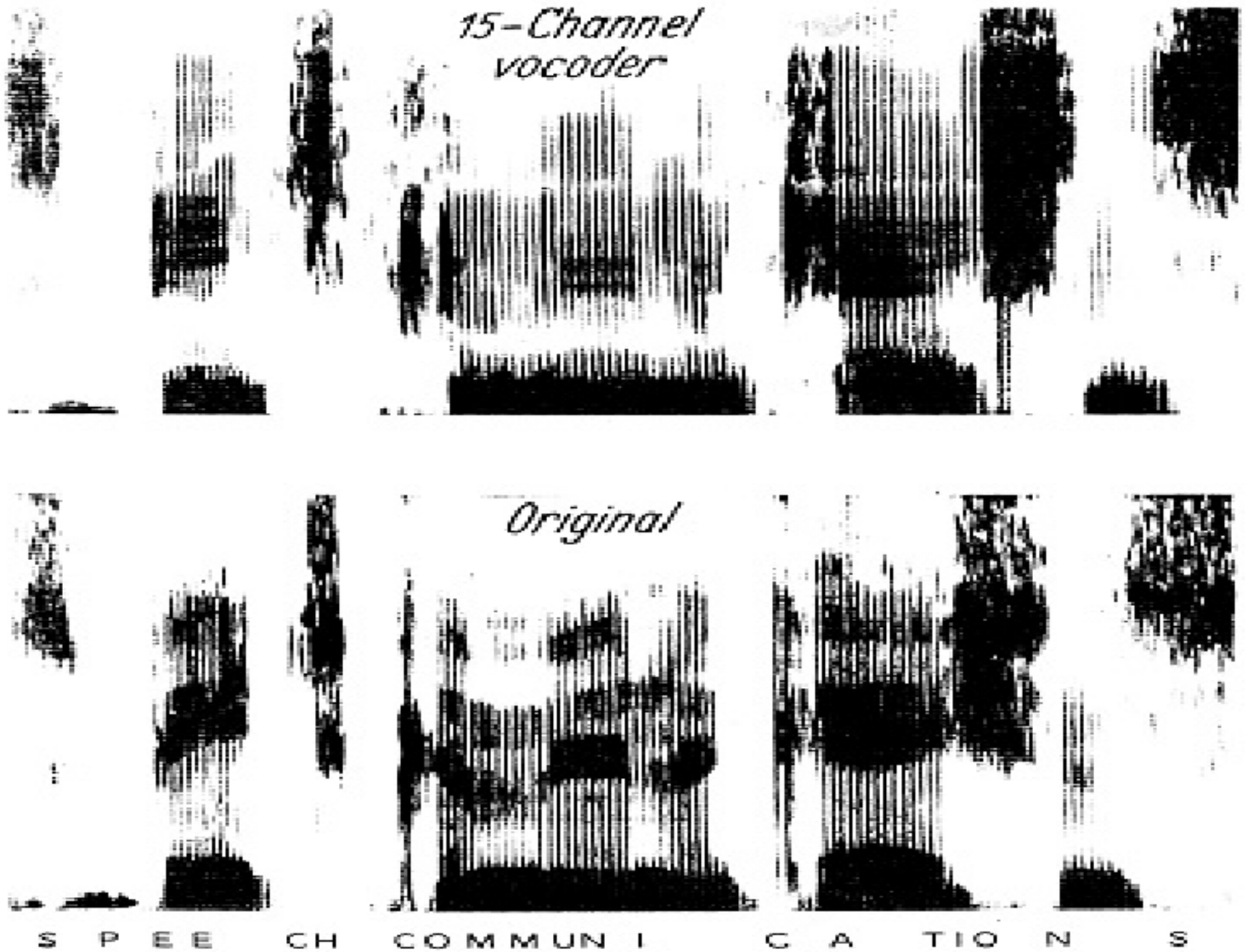




- 1200-9600 bps
- 600 bps for pitch and V/UV
- easy to modify pitch, timing

Channel Vocoder- 2.4 kbps

Channel Vocoder

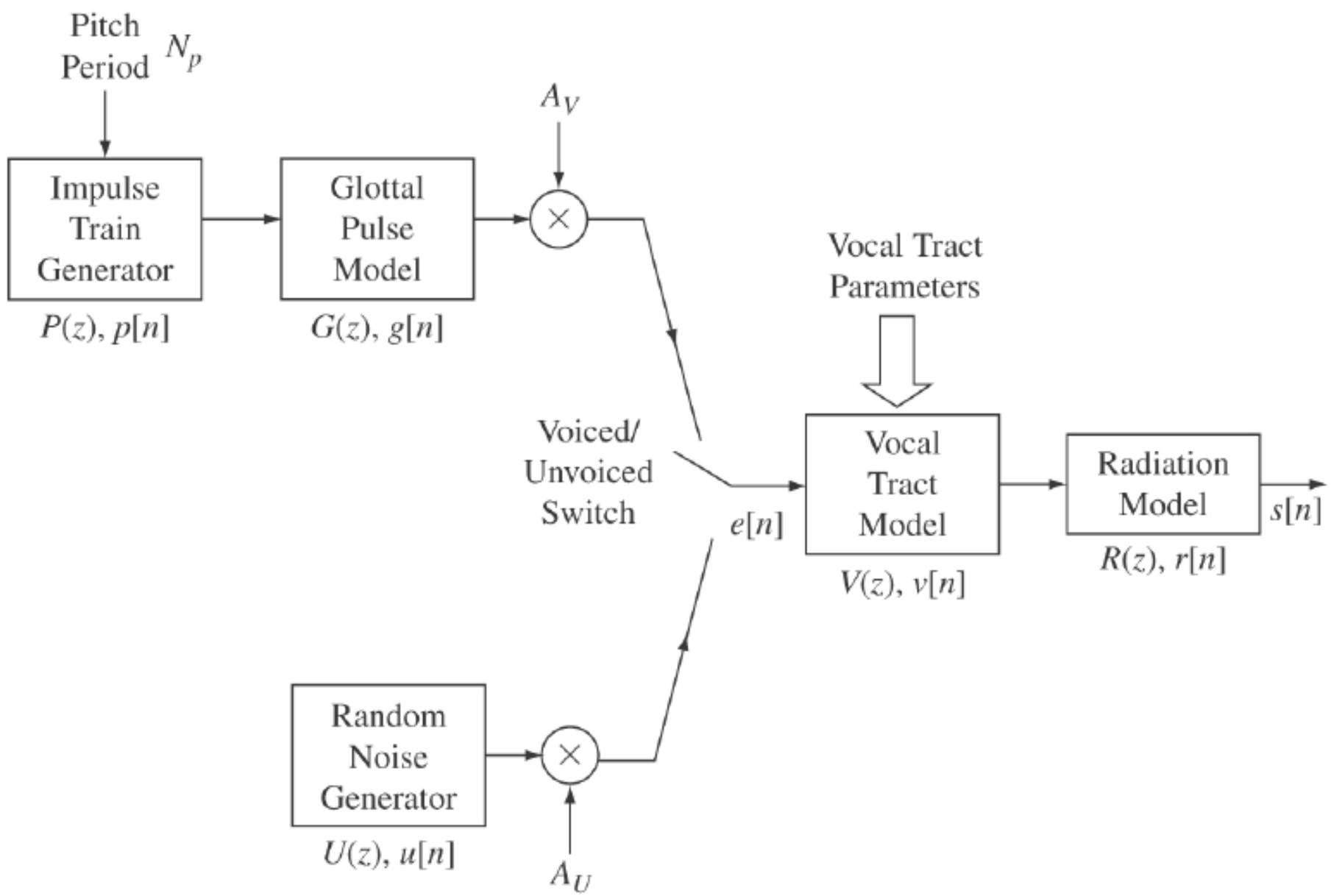


Homomorphic Analysis of speech

- short segment of speech can be modeled as having been generated by exciting an LTI system either by a quasi-periodic impulse train, or a random noise signal
- speech analysis => estimate parameters of the speech model, measure their variations (and perhaps even their statistical variabilites-for quantization) with time
- speech = excitation * system response

=> want to deconvolve speech into excitation and system response => do this using homomorphi filtering methods

Homomorphic Analysis of Speech Model



Homomorphic Analysis of Speech Model

- the transfer function for voiced speech is of the form

$$H_V(z) = A_V \cdot G(z)V(z)R(z)$$

- with effective impulse response for voiced speech

$$h_V[n] = A_V \cdot g[n] * v[n] * r[n]$$

- similarly for unvoiced speech we have

$$H_U(z) = A_U \cdot V(z)R(z)$$

- with effective impulse response for unvoiced speech

$$h_U[n] = A_U \cdot v[n] * r[n]$$

Complex Cepstrum for Speech

- the models for the speech components are as follows:

1. vocal tract: $V(z) = \dfrac{Az^{-M}\prod_{k=1}^{M_i}(1-a_k z^{-1})\prod_{k=1}^{M_0}(1-b_k z)}{\prod_{k=1}^{N_i}(1-c_k z^{-1})}$

--for voiced speech, only poles => $a_k = b_k = 0$, all k

--unvoiced speech and nasals, need pole-zero model but all poles are inside the unit circle => $c_k < 1$

--all speech has complex poles and zeros that occur in complex conjugate pairs

2. radiation model: $R(z) \approx 1 - z^{-1}$ (high frequency emphasis)

3. glottal pulse model: finite duration pulse with transform

$$G(z) = B\prod_{k=1}^{L_i}(1-\alpha_k z^{-1})\prod_{k=1}^{L_0}(1-\beta_k z)$$

with zeros both inside and outside the unit circle

Complex Cepstrum of Speech

- model of speech:
  - voiced speech produced by a quasi-periodic pulse train exciting slowly time-varying linear system => $p[n]$ convolved with $h_v[n]$
  - unvoiced speech produced by random noise exciting slowly time-varying linear system => $u[n]$ convolved with $h_v[n]$

Complex Cepstrum for Voiced Speech

- combination of vocal tract, glottal pulse and radiation will be non-minimum phase => complex cepstrum exists for all values of *n*
- the complex cepstrum will decay rapidly for large *n* (due to polynomial terms in expansion of complex cepstrum)
- effect of the voiced source is a periodic pulse train for multiples of the pitch period

Simplified Speech Model

- short-time speech model

$$x[n] = w[n] \cdot \left[p[n] * g[n] * v[n] * r[n] \right]$$

$$\approx p_w[n] * h_v[n]$$

- short-time complex cepstrum

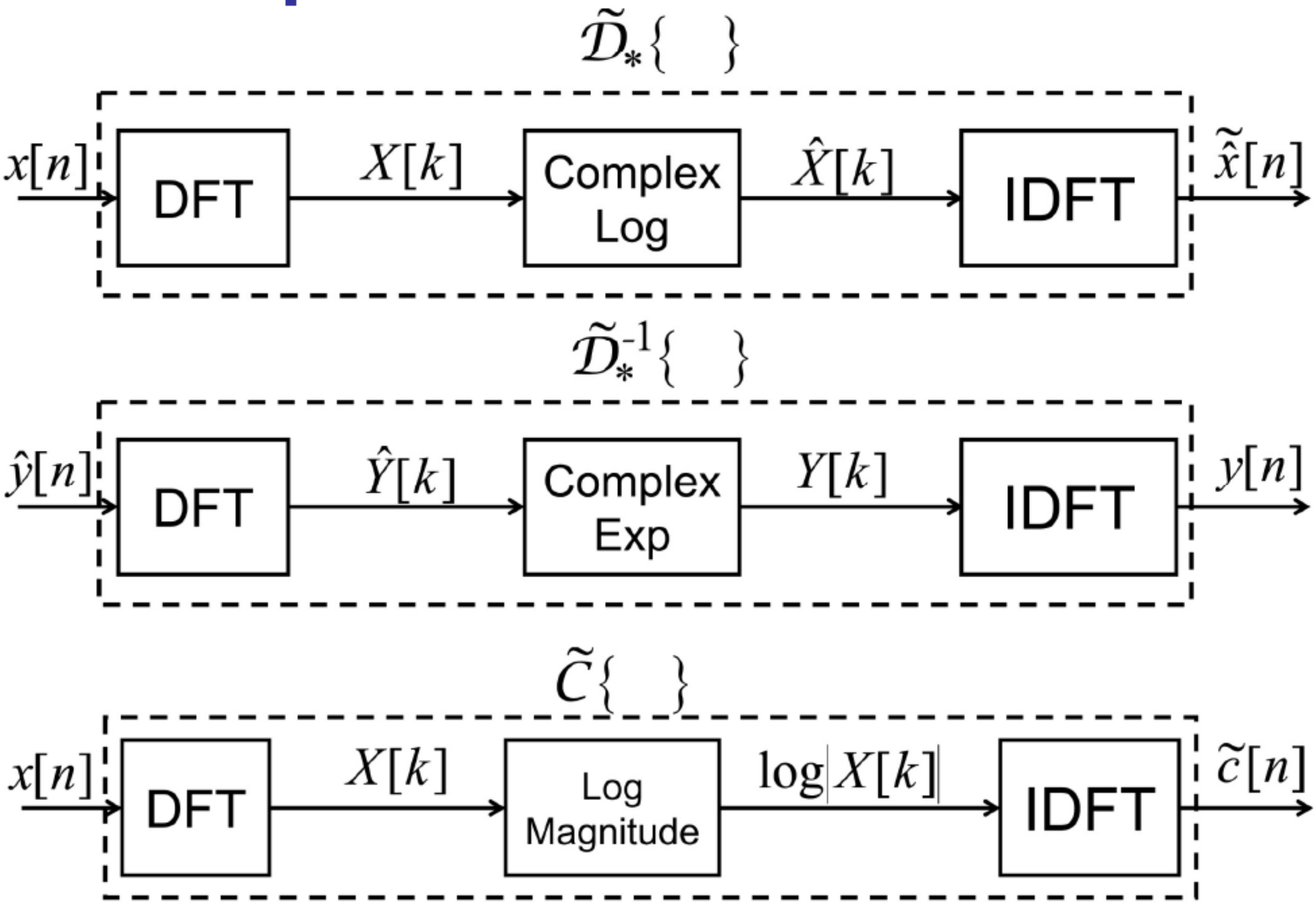
$$\hat{x}[n] = \hat{p}_w[n] + \hat{g}[n] + \hat{v}[n] + \hat{r}[n]$$

Computing Short-Time Cepstrums from Speech Using the DFT

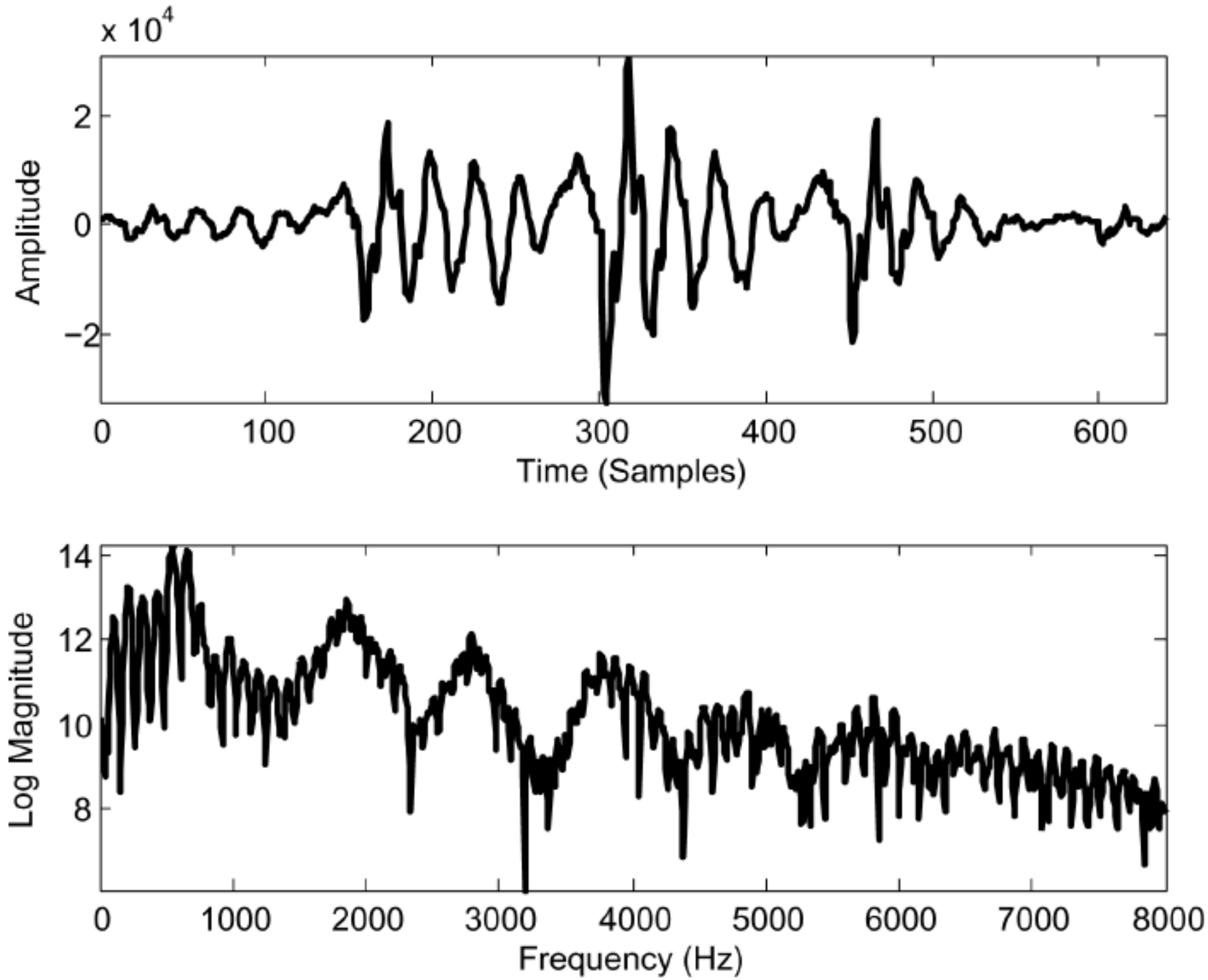
Practical Considerations

- window to define short-time analysis
- window duration (should be several pitch periods long)
- size of FFT (to minimize aliasing)
- elimination of linear phase components (positioning signals within frames)
- cutoff quefrency of lifter
- type of lifter (low/high quefrency)

Computational Considerations



Voiced Speech Example

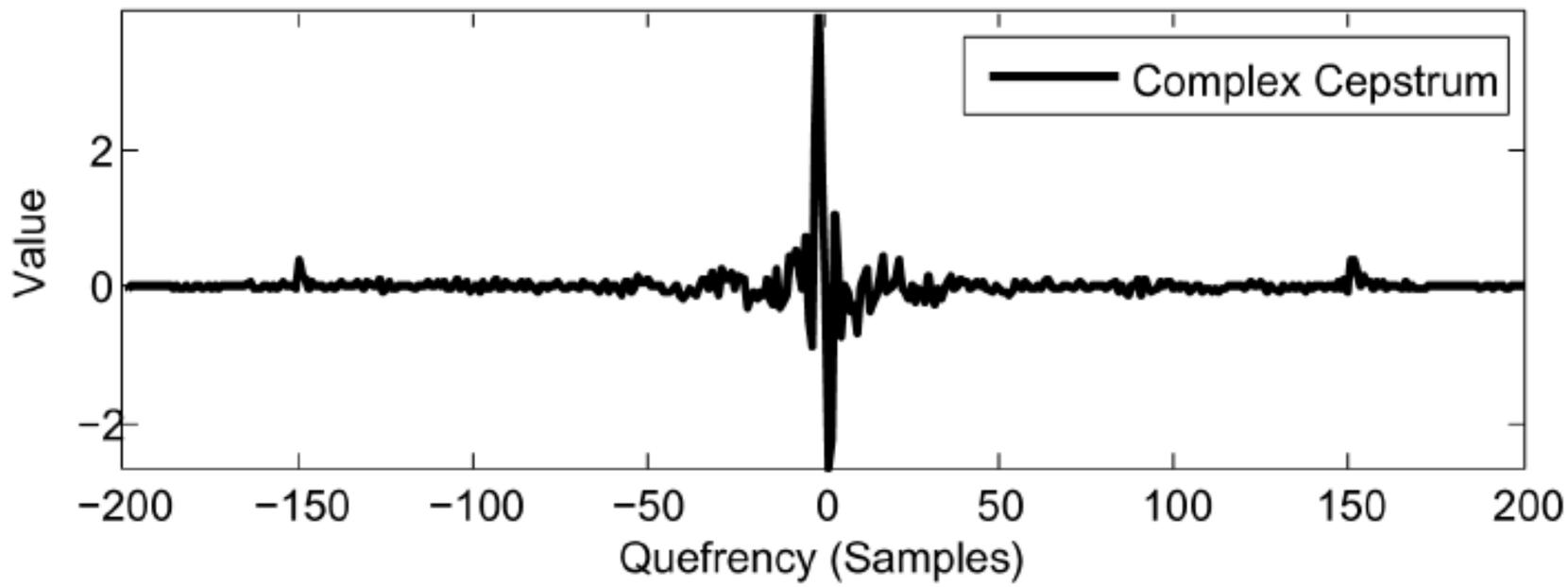


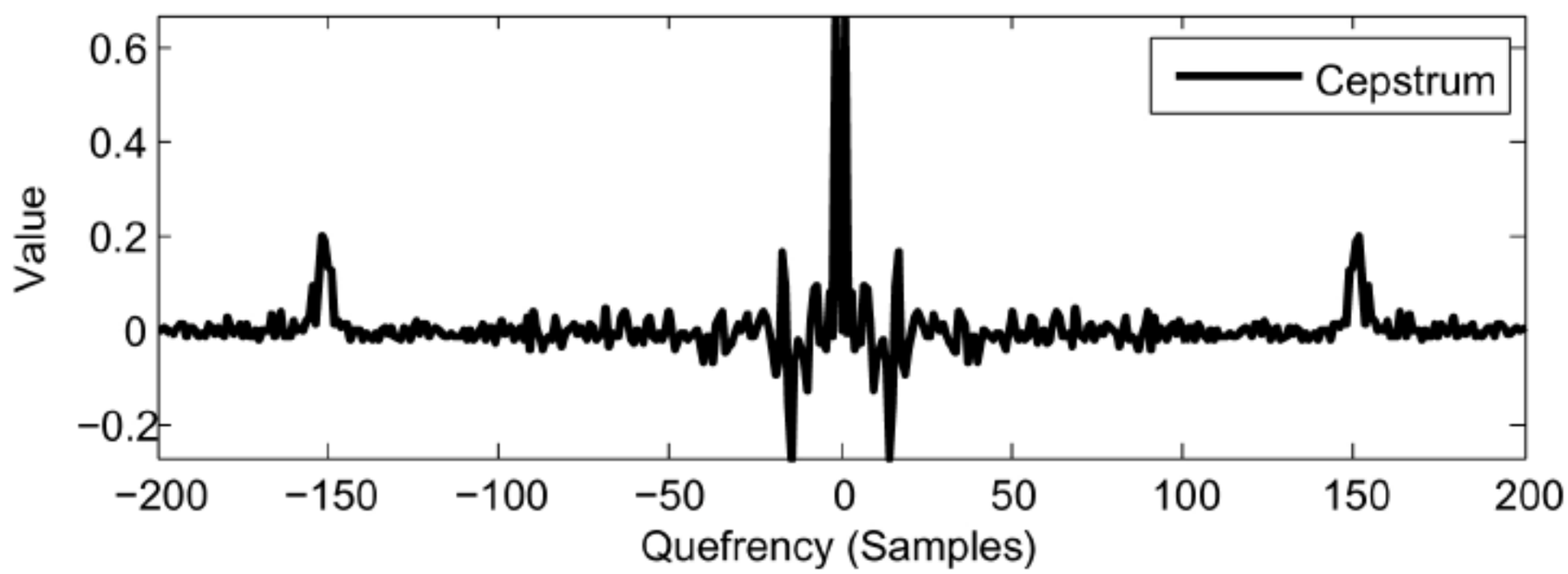
Hamming window

40 msec duration

(section beginning at sample 13000 in file test_16k.wav)

Voiced Speech Example



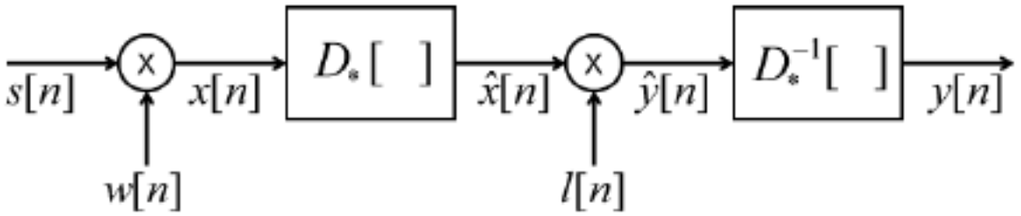


Characteristic System for Homomorphic Convolution

- still need to define (and design) the *L* operator part (the linear system component) of the system to completely define the characteristic system for homomorphic convolution for speech
  - to do this properly and correctly, need to look at the properties of the complex cepstrum for speech signals

Homomorphic Filtering of Voiced Speech

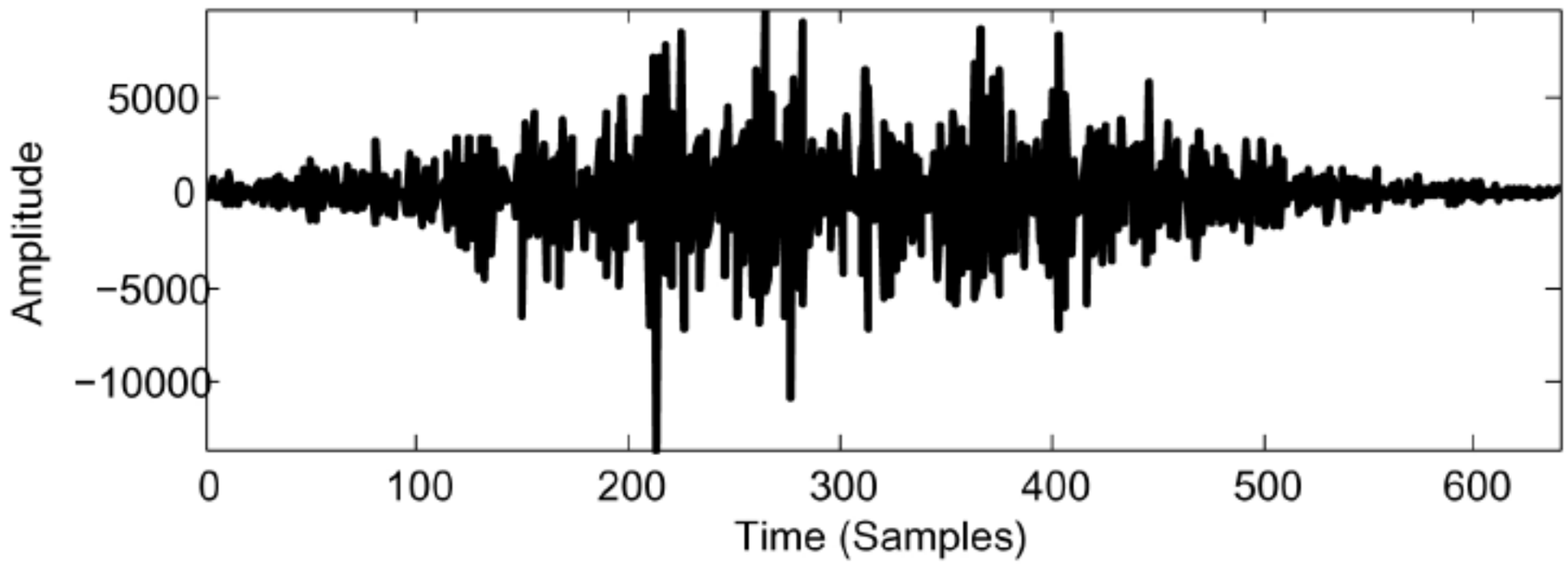
- goal is to separate out the excitation impulses from the remaining components of the complex cepstrum
- use cepstral window, *l(n),* to separate excitation pulses from combined vocal tract
  - $l(n)=1$ for $|n|<n_0<N_p$
  - $l(n)=0$ for $|n|\geq n_0$
  - this window removes excitation pulses
  - $l(n)=0$ for $|n|<n_0<N_p$
  - $l(n)=1$ for $|n|\geq n_0$
  - this window removes combined vocal tract
- the filtered signal is processed by the inverse characteristic system to recover the combined vocal tract component



$$\hat{y}(n) = \ell(n) \cdot \hat{x}(n)$$

$$\hat{Y}(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}(e^{j\theta}) L(e^{j(\omega-\theta)}) d\theta$$

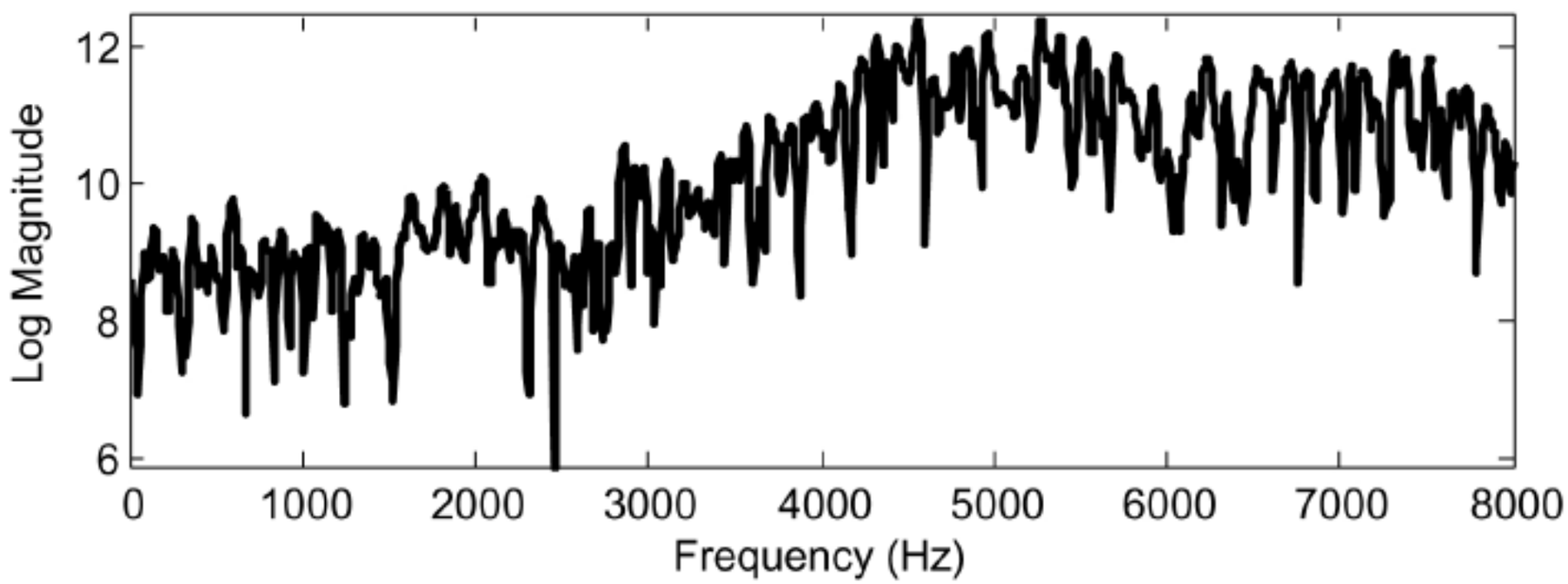
Unvoiced Speech Example



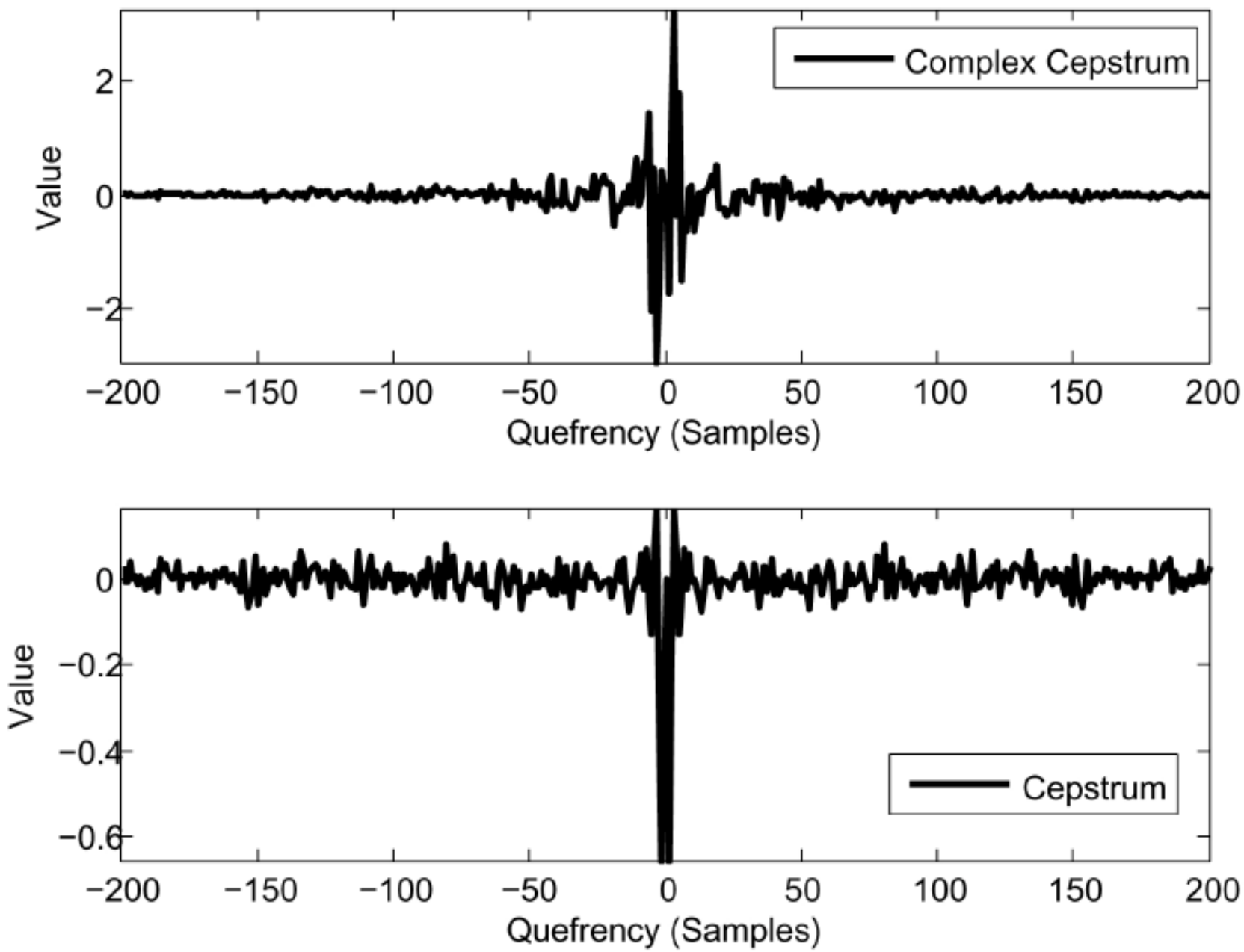
Hamming window

40 msec duration

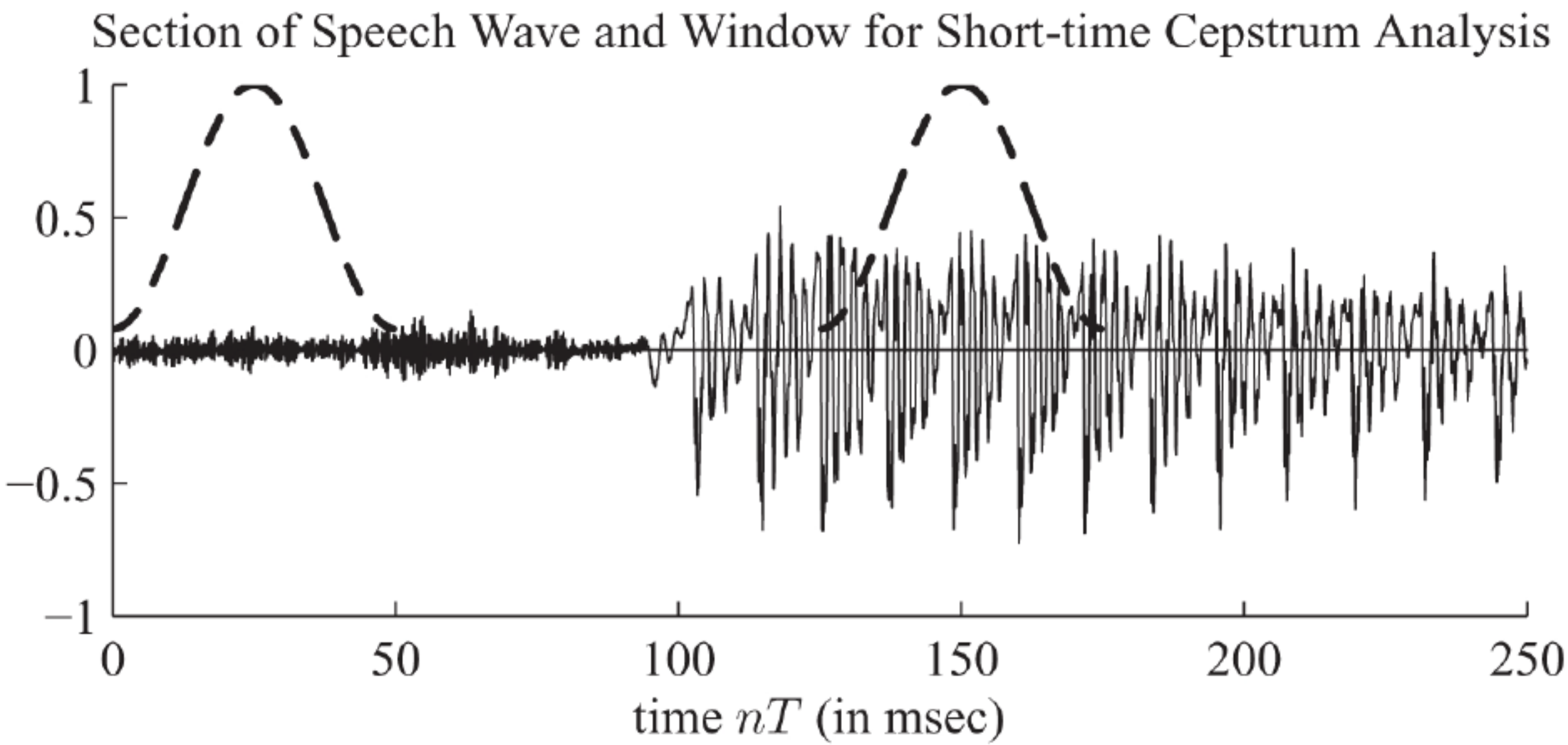
(section beginning at sample 3200 in file test_16k.wav)



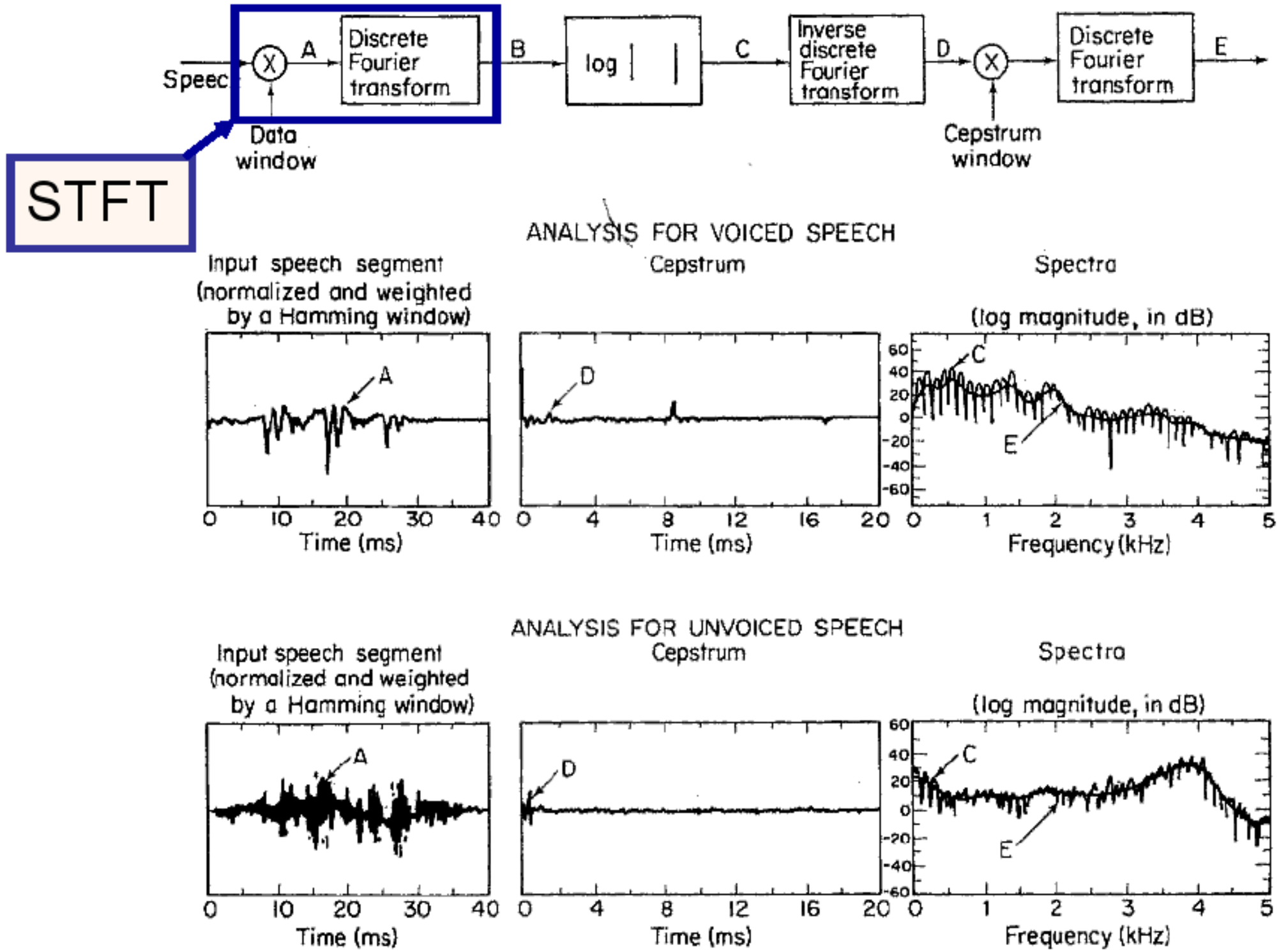
Unvoiced Speech Example



Running Cepstrum



Short-Time Homomorphic Analysis



Formant estimation

Liftering

Liftering operation is similar to filtering operation in the frequency domain where a desired quefrency region for analysis is selected by multiplying the whole cepstrum by a rectangular window at the desired position. There are two types of liftering performed, low-time liftering and high-time liftering. Low-time liftering operation is performed to extract the vocal tract characteristics in the quefrency domain and high-time liftering is performed to get the excitation characteristics of the analysis speech frame.

Low-time liftering for Formant estimation

Low-time liftering is used for estimating slow varying vocal tract characteristics from the computed cepstrum of the given speech sequence. The low-time liftering window used for extracting vocal tract characteristics can be represented as follows,

$$w_e[n] = \begin{cases} 1, 0 \le n \le L_c \\ 0, L_c \le n \le \frac{N}{2} \end{cases} \qquad (6)$$

where L_c is the cut off length of the liftering window and N/2 is half the total length of the cepstrum. Usually L_c is used as 15 or 20. The vocal tract characteristics can be obtained by multiplying the cepstrum c(n) with the low-time liftering window as indicated in Eqn. (7).

$$c_e(n) = w_e[n].\, c(n) \qquad (7)$$

The extraction of vocal tract characteristics is illustrated in Figure 8.

Applying DFT on the low-time liftered sequence takes to its log magnitude spectrum which is the vocal tract spectrum of the given short term speech as given in Eqn. (8).

$$\text{Log}\,[\,|H(w)|\,] = \text{DFT}\,[c_e(n)] \qquad (8)$$

Formant estimation

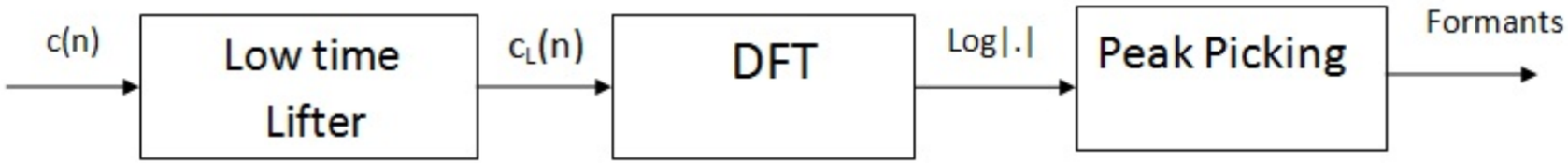


Figure 4: Block diagram representing low-time liftering

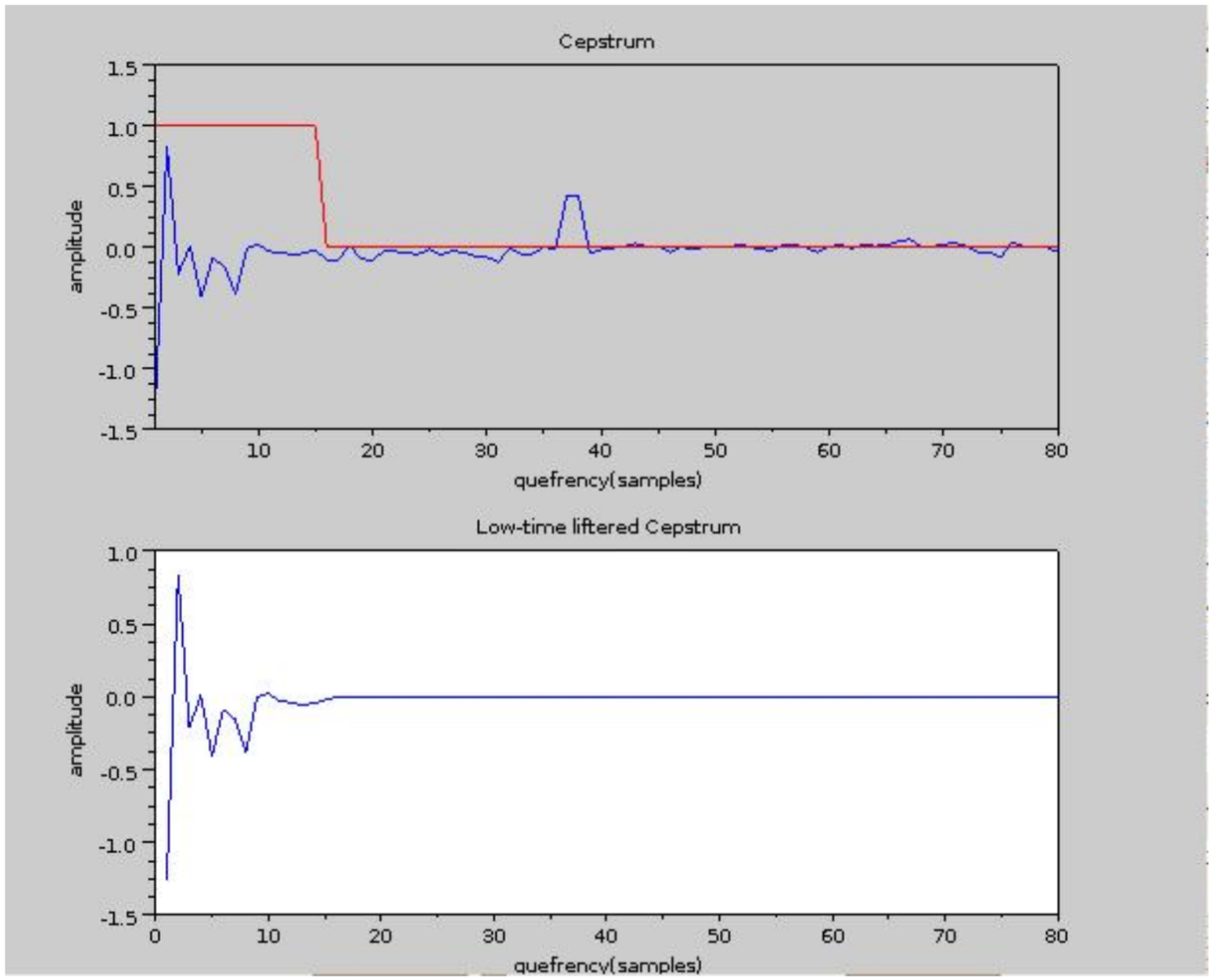


Figure 5: Low-time liftering: Cepstrum of a voiced segment and low-time liftering window (in red color) and vocal tract characteristics of the cepstrum obtained through the low-time liftering

Formant estimation

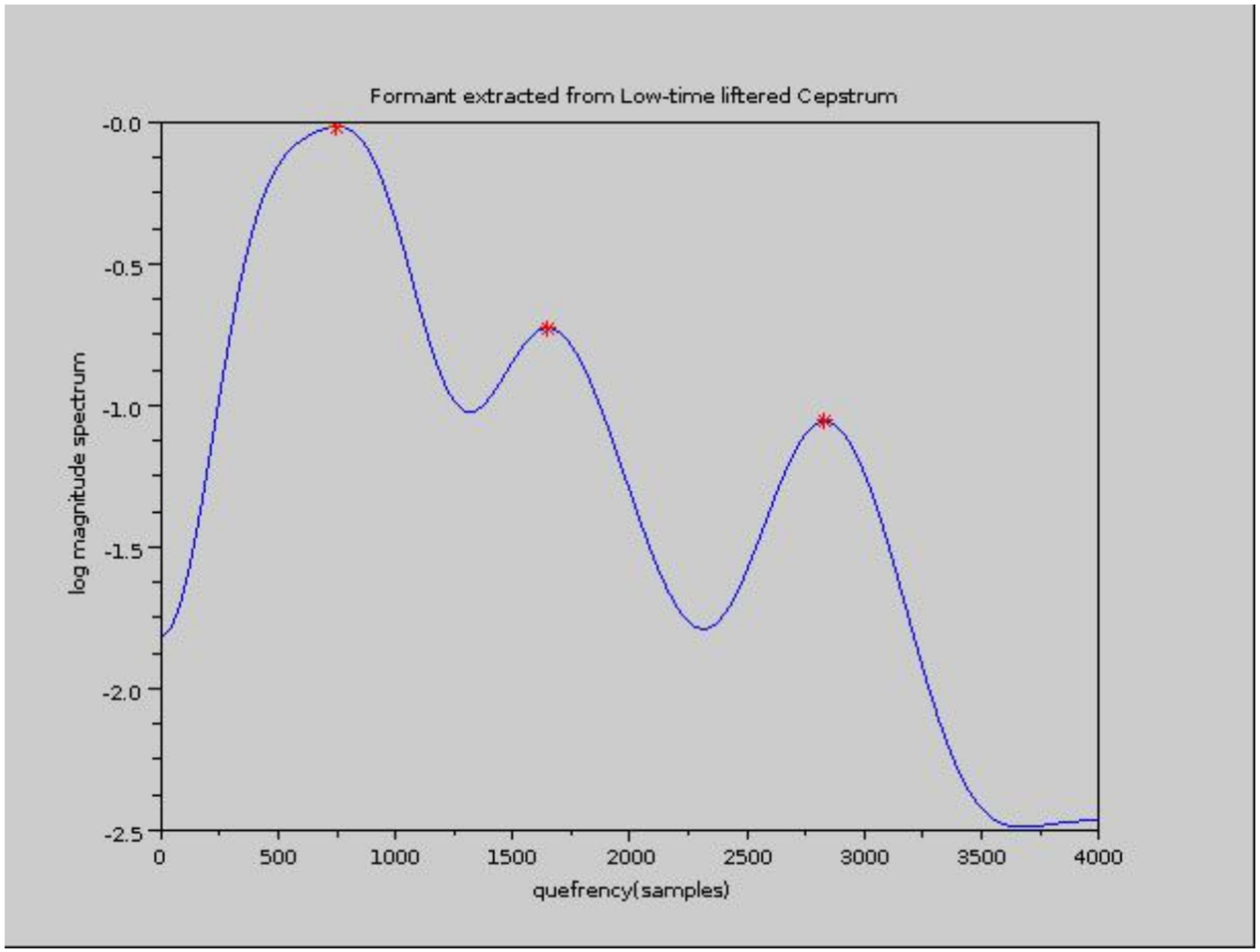


Figure 6: Formant locations from vocal tract spectrum

Pitch detection

High-time liftering for pitch estimation

As the cepstrum computed from the analysis speech sequence is symmetric, half the length of the cepstrum is considered for the liftering. The excitation characteristic are obtained through a high time liftering operation using the following window,

$$w_h[n] = \begin{cases} 1, L_c \leq n < \frac{N}{2} \\ 0, else \end{cases} \tag{9}$$

where $\boldsymbol{L_c}$ is the cut off length of the liftering window and $\boldsymbol{N/2}$ is the half the total length of the cepstrum. Usually $\boldsymbol{L_c}$ is used as 15 or 20. The excitation characteristics are obtained by multiplying high time liftering window with the cepstrum obtained as given in Eqn. (7).

$$c_h(n) = w_h(n) * c(n) \tag{10}$$

The block diagram given in Figure 7 indicates the high-time liftering process for pitch estimation. The computation of high-time liftered cepstrum from the cepstrum using high-time liftering window is given in Figure 8. Pitch can be estimated as the instant corresponds to the highest peak in the high-time liftered cepstrum. In the Figure 8, pitch period is the time instant corresponding to the largest peak in the high-time liftered cepstrum. The reciprocal of the pitch interval multiplied by the sampling frequency gives the pitch frequency of the analysis speech frame.

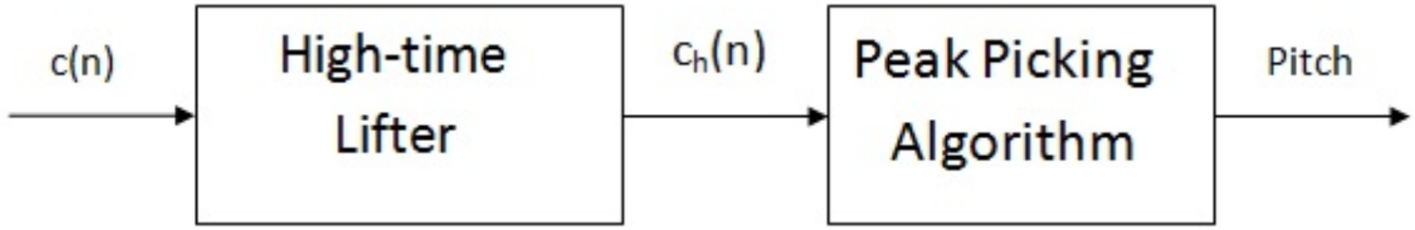


Figure 7: Block diagram representing high-time liftering

Pitch detection

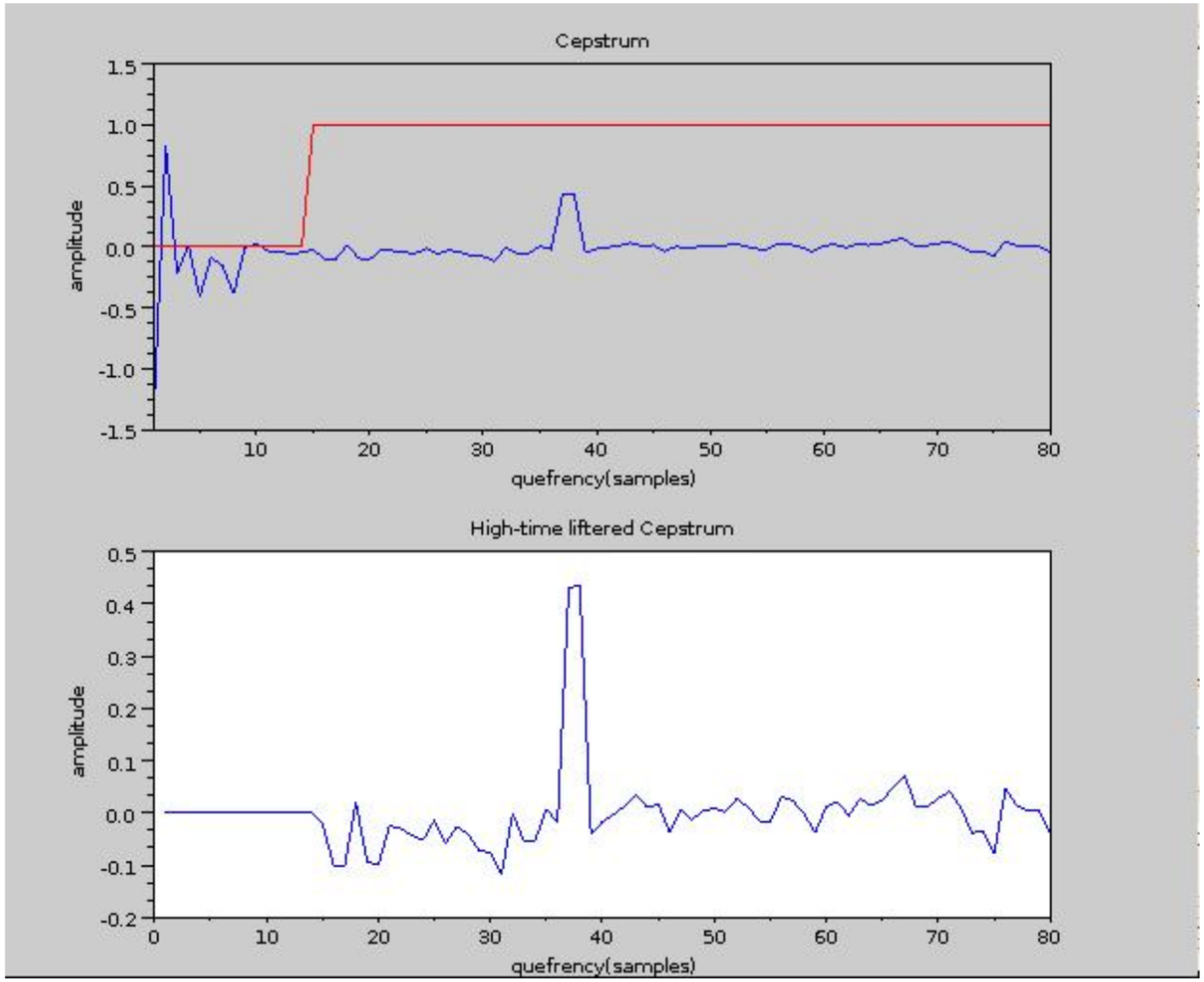
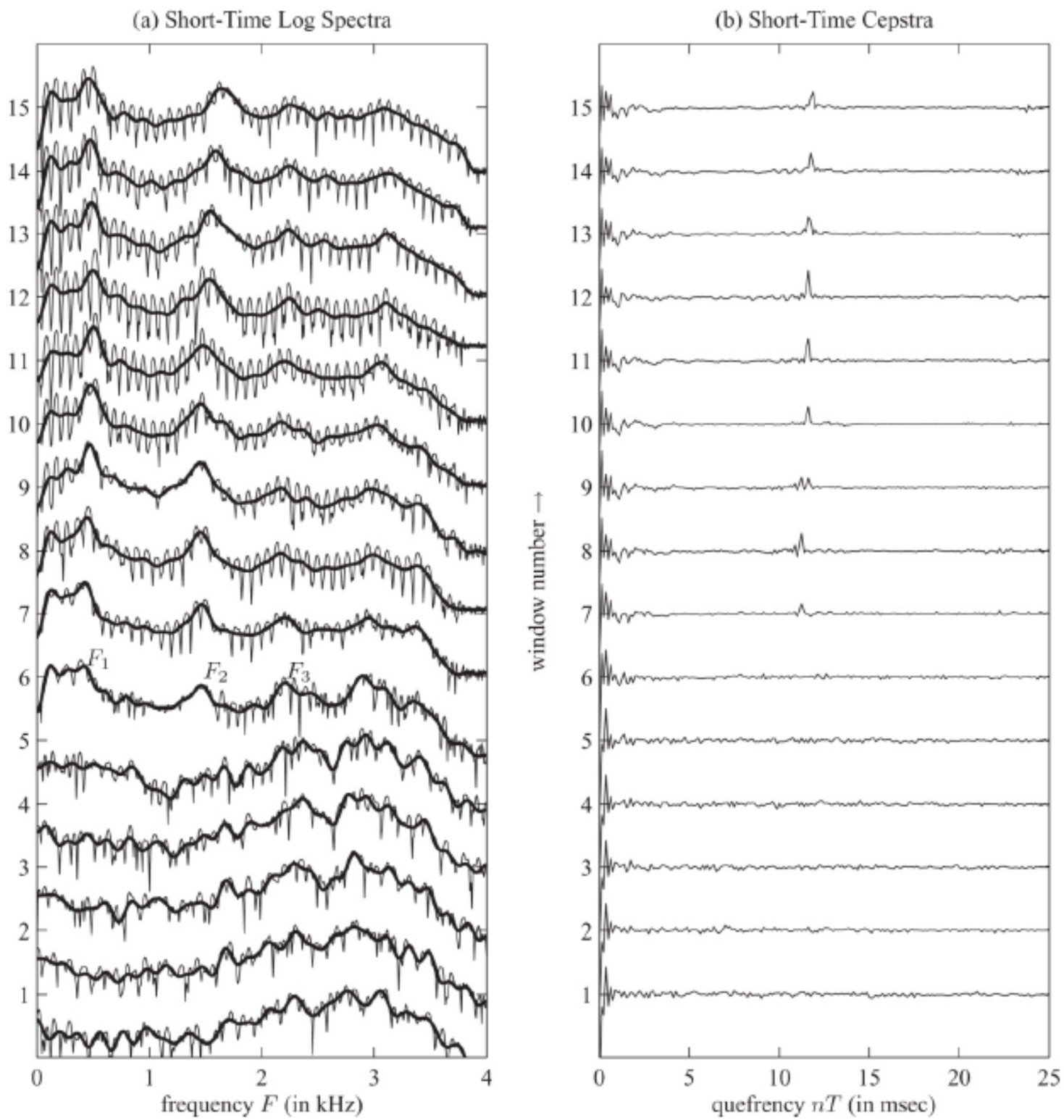


Figure 8: High-time liftering: Cepstrum of a voiced segment and liftering window (in red color) and vocal tract part of the cepstrum obtained through the high time liftering

Pitch detection



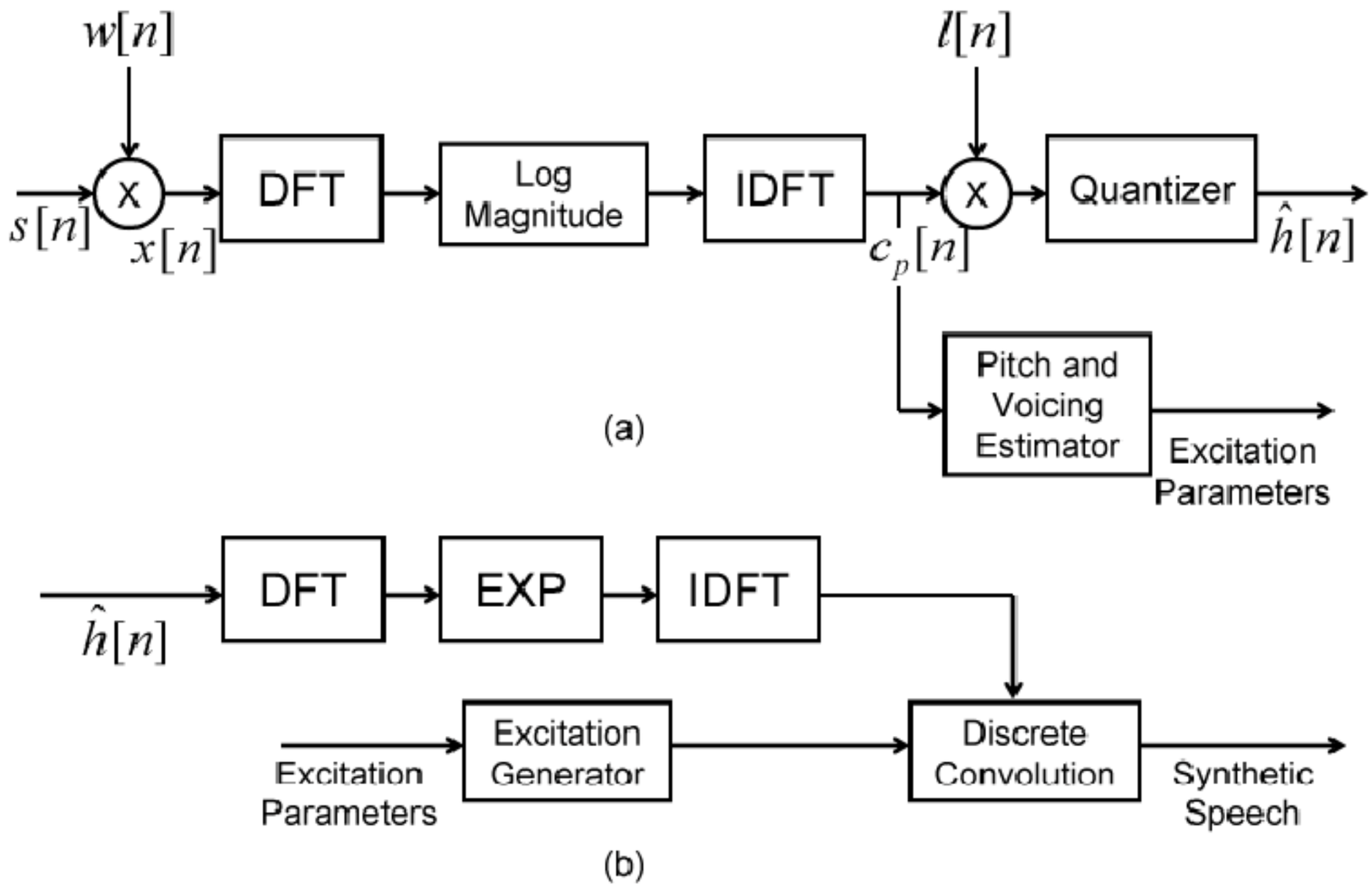
Homomorphic Vocoder

- time-dependent complex cepstrum retains all the information of the time-dependent Fourier transform => exact representation of speech
- time dependent real cepstrum loses phase information -> not an exact representation of speech
- quantization of cepstral parameters also loses information
- cepstrum gives good estimates of pitch, voicing, formants => can build homomorphic vocoder

Homomorphic Vocoder

1. compute cepstrum every 10-20 msec
2. estimate pitch period and voiced/unvoiced decision
3. quantize and encode low-time cepstral values
4. at synthesizer-get approximation to $h_v(n)$ or $h_u(n)$ from low time quantized cepstral values
5. convolve $h_v(n)$ or $h_u(n)$ with excitation created from pitch, voiced/unvoiced, and amplitude information

Homomorphic Vocoder



- *l(n)* is cepstrum window that selects low-time values and is of length 26 samples

homomorphic vocoder