ECE 8440 Unit 24

13.2 Applications (of Homomorphic Deconvolution) to Speech Processing

Speech production can be modeled as the convolution of an excitation signal with the unit sample response of a linear speech production system. There are two kinds of excitation signals:

1. pulse train (for "voiced" speech where the vocal cords are vibrating)
2. random noise (corresponding to "unvoiced" speech, such as the "s" sound")

This is summarized in the diagram below:

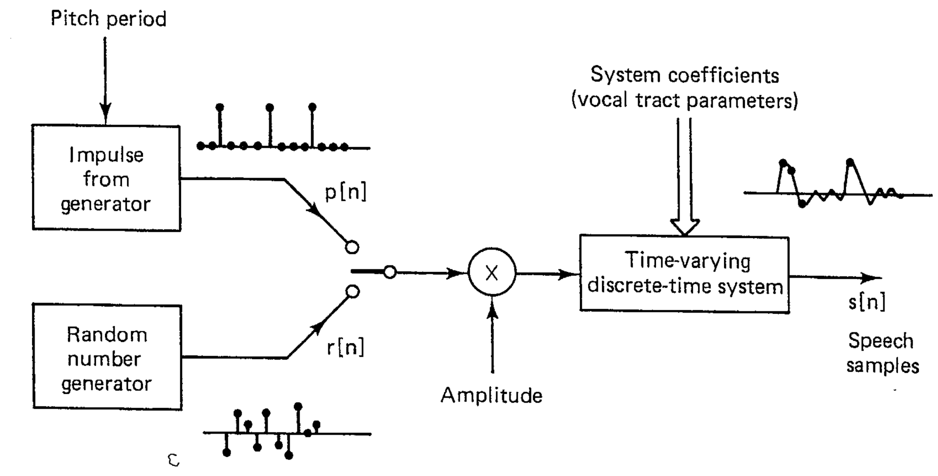


Figure 13.22 Discrete-time model of speech production

Vocal tract model:

$$V(z) = \frac{\sum_{k=0}^{K} b_k z^{-k}}{\sum_{k=0}^{P} a_k z^{-k}}$$ equation 13.118

$$= AZ^{-K_0} \frac{\prod_{k=1}^{K_i}(1-\alpha_k z^{-1})\prod_{k=1}^{K_o}(1-\beta_k z)}{\prod_{k=1}^{P/2}(1-r_k e^{j\theta_k} z^{-1})(1-r_k e^{-j\theta_k} z^{-1})}$$ equation 13.119

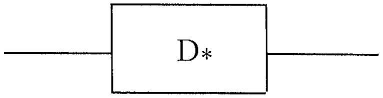
The output speech signal is represented by $s(n) = v(n) * x(n)$.

For voiced speech, $x(n) = p(n) = \sum_k \delta(n - kN_0)$.

For unvoiced speech, $x(n) = r(n)$ = random noise signal.

<u>Using windowing for "short-time" analysis</u>.

Let $x(n) = s(n)w(n)$ be the input to a homomorphic deconvolution system where s(n) is a speech signal and w(n) is a window function.

$x(n) = s(n)w(n)$ 

For voiced speech, $x(n) = [p(n) * v(n)] \cdot w(n)$.

If w(n) varies slowly relative to variations of v(n), we can approximate the above as

$x(n) \approx v(n) * p_w(n)$ equation 13.122

where $p_w(n) = w(n)p(n) = \sum_{k=0}^{M-1} w(kN_0)\delta(n - kN_0).$ equation 13.125

Complex cepstrum of x(n)

$\hat{x}(n) = \hat{v}(n) + \hat{p}_w(n)$ equation 13.126

To obtain $\hat{p}_w(n)$, define a sequence

$$w_{N_o}(k) = \begin{cases} w(kN_0), & k = 0,1,\cdots M\text{-}1 \\ 0; & \text{otherwise} \end{cases}$$ equation 13.127

The Fourier Transform of $p_w(n)$ is

$$P_w(e^{j\omega}) = \sum_{k=0}^{M-1} p_w(k)e^{-j\omega k} = \sum_{k=0}^{M-1}\left[\sum_{m=0}^{M-1} w(mN_0)\delta(k - mN_0)\right]e^{-j\omega k}$$

$$= \sum_{m=0}^{M-1} w(mN_0)\sum_{k=0}^{M-1}\delta(k - mN_0)e^{-j\omega k} = \sum_{m=0}^{M-1} w(mN_0)e^{-j\omega mN_0}$$

$$P_w(e^{j\omega}) = \sum_{m=0}^{M-1} w(mN_0)e^{-j\omega mN_0} = W_{N_0}(e^{j\omega N_0})$$ equation 13.128

Period: Set $\omega N_0 = 2\pi$ and solve for ω: $\omega = \dfrac{2\pi}{N_0} = \text{period}$

$$\log[P_w(e^{j\omega})] = \log[W_{N_0}(e^{j\omega N_0})]$$

Note that $\log[P_w(e^{j\omega})]$ also has the same period as $P_w(e^{j\omega})$.

Because of the relation between $\log[P_w(e^{j\omega})]$ and $\log[W_{N_0}(e^{j\omega N_0})]$, the relation between their inverse DTFT's is

$$\hat{p}_w(n) = \begin{cases} \hat{w}_{N_0}[n / N_0] & n = \pm N_0, \pm 2N_0, \cdots \\ 0 & \text{otherwise} \end{cases}$$ equation 13.129

(Recall this property from material covered in Chapter 4.)

Therefore, $\hat{p}_w(n)$ has the form:

$$\hat{p}_w(n) = \sum_i a_i \delta(n - iN_0)$$

As shown before, the complex cepstrum for a signal whose z-transform V(z) has the form shown in equation 13.119 (shown again below) is as follows:

$$= AZ^{-K_0} \frac{\prod_{k=1}^{K_i}(1-\alpha_k z^{-1})\prod_{k=1}^{K_o}(1-\beta_k z)}{\prod_{k=1}^{P/2}(1-r_k e^{j\theta_k} z^{-1})(1-r_k e^{-j\theta_k} z^{-1})}$$

(equation 13.119, shown again)

$$\hat{v}(n) = \begin{cases} -\sum_{k=1}^{K_0} \frac{\beta_k^{-n}}{n}, & n < 0 \\ \log|A|, & n = 0 \\ -\sum_{k=1}^{K_i} \frac{\alpha_k^{n}}{n} + \sum_{k=1}^{P/2} \frac{2r_k^n}{n}\cos(\theta_k n), & n > 0 \end{cases} .$$

(equation 13.130)

Note: The above expression assumes that the z^{-K_0} in equation 13.119 is removed before calculating the complex cepstrum $\hat{v}(n)$.

13.10.2 Example of Homomorphic Deconvolution of Speech

- Analysis of a section of voiced speech
- Sampling rate - 8,000 samples per second
- Hamming window of length 401 used (50 ms window length)

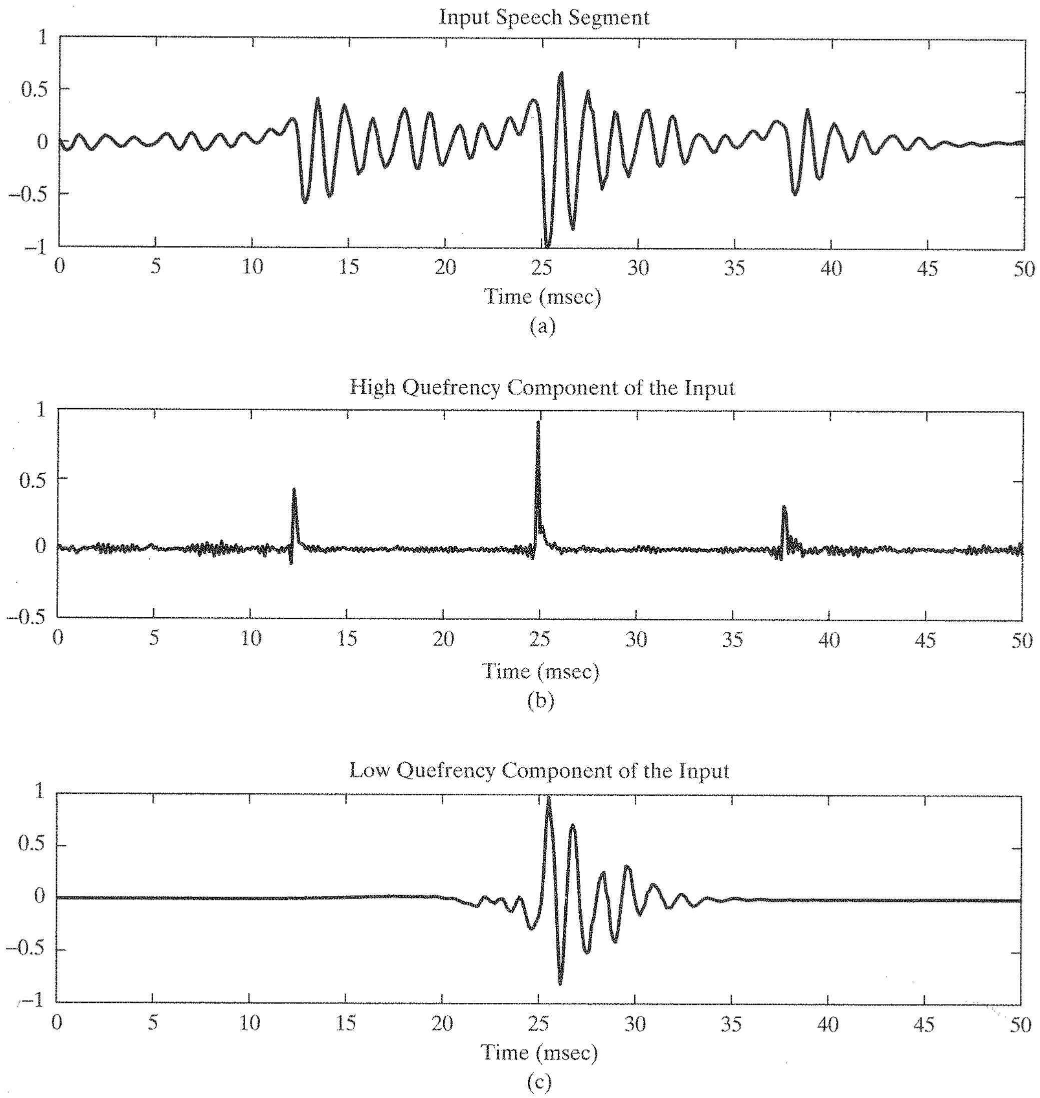


Figure 13.23 Homomorphic deconvolution of speech. (a) Segment of speech weighted by a Hamming window. (b) High quefrency component of the signal in (a). (c) Low quefrency component of the signal in (a).

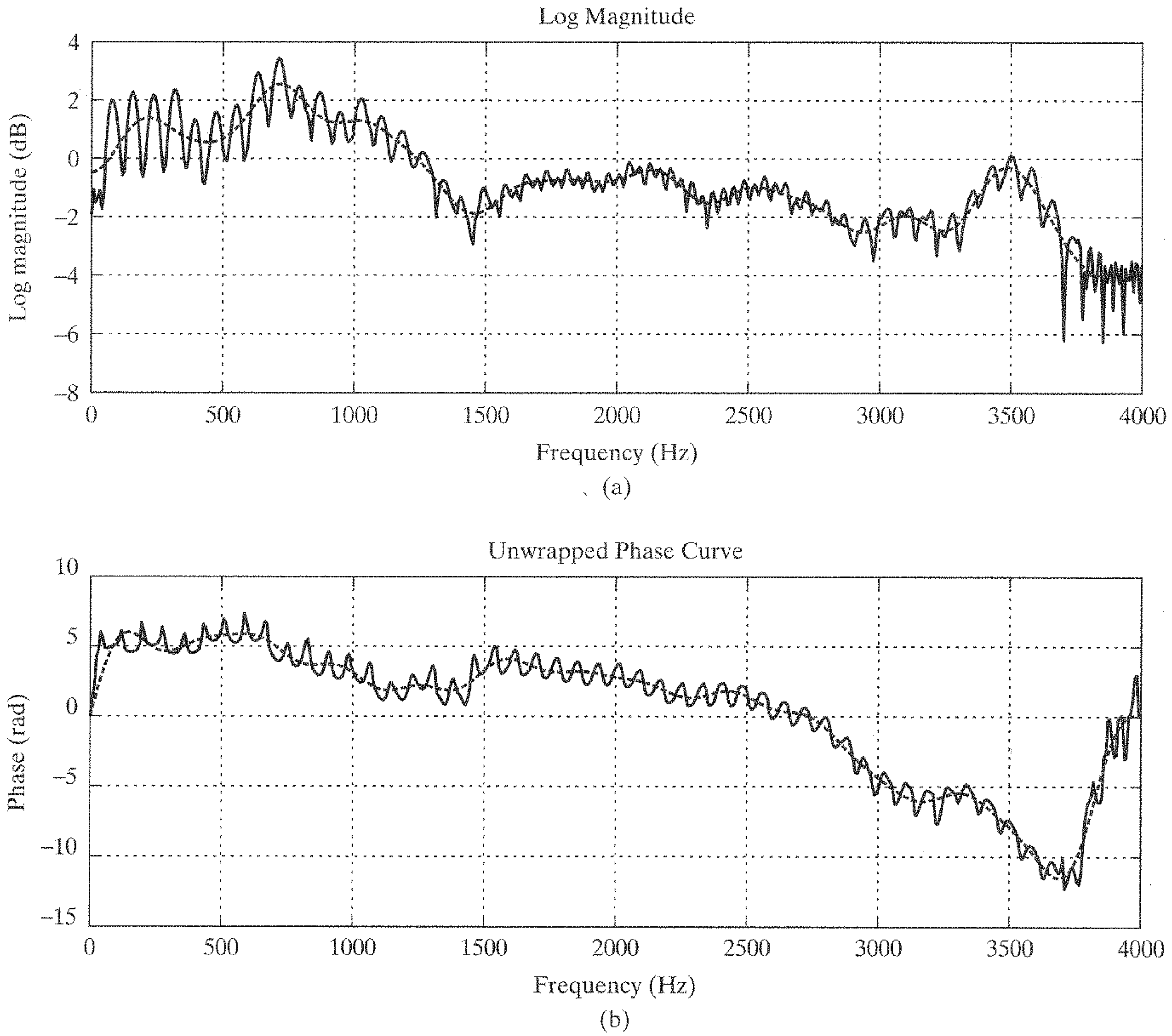


Figure 13.24 Complex logarithm of the signal of Figure 13.23(a): (a) Log magnitude. (b) Unwrapped phase.

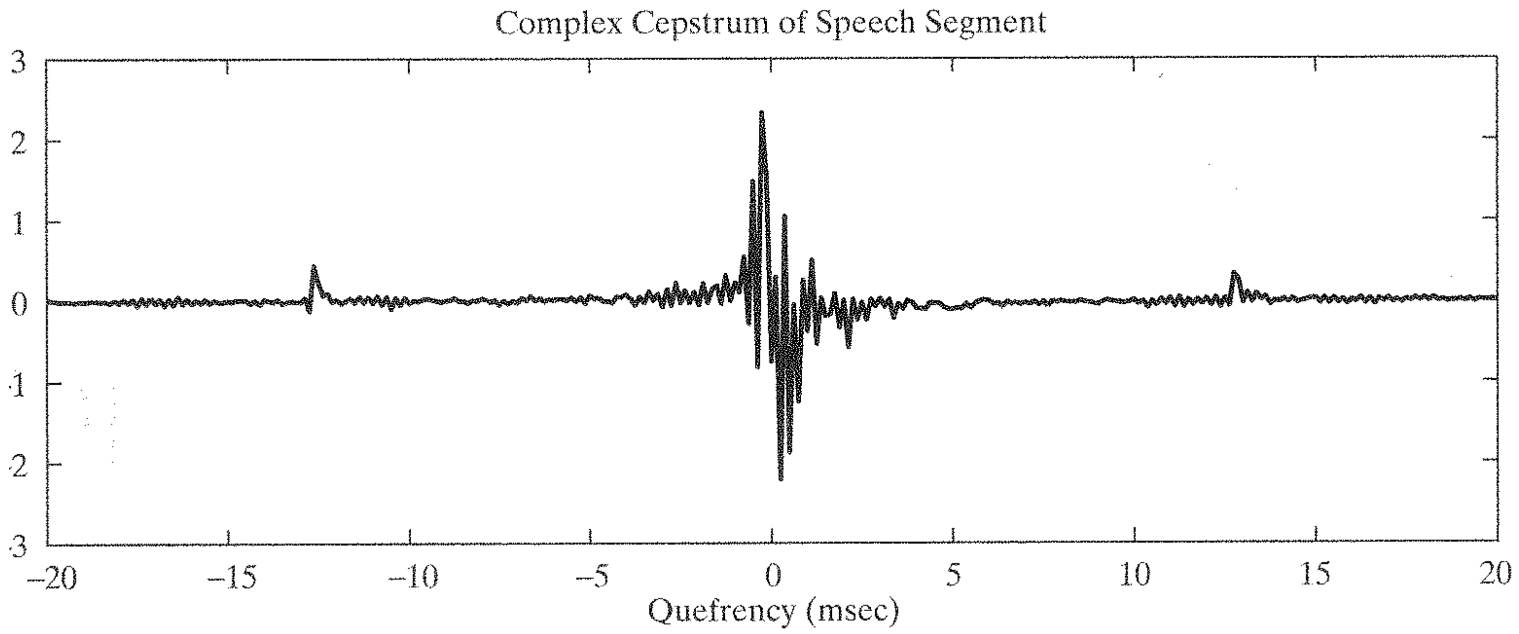


Figure 13.25 Complex cepstrum of the signal in Figure 13.23(a) (inverse DTFT of the complex logarithm in Figure 13.24).

Note the two components: One due to v(n) which decreases as 1/n, and the other which is a pulse train due to p(n).

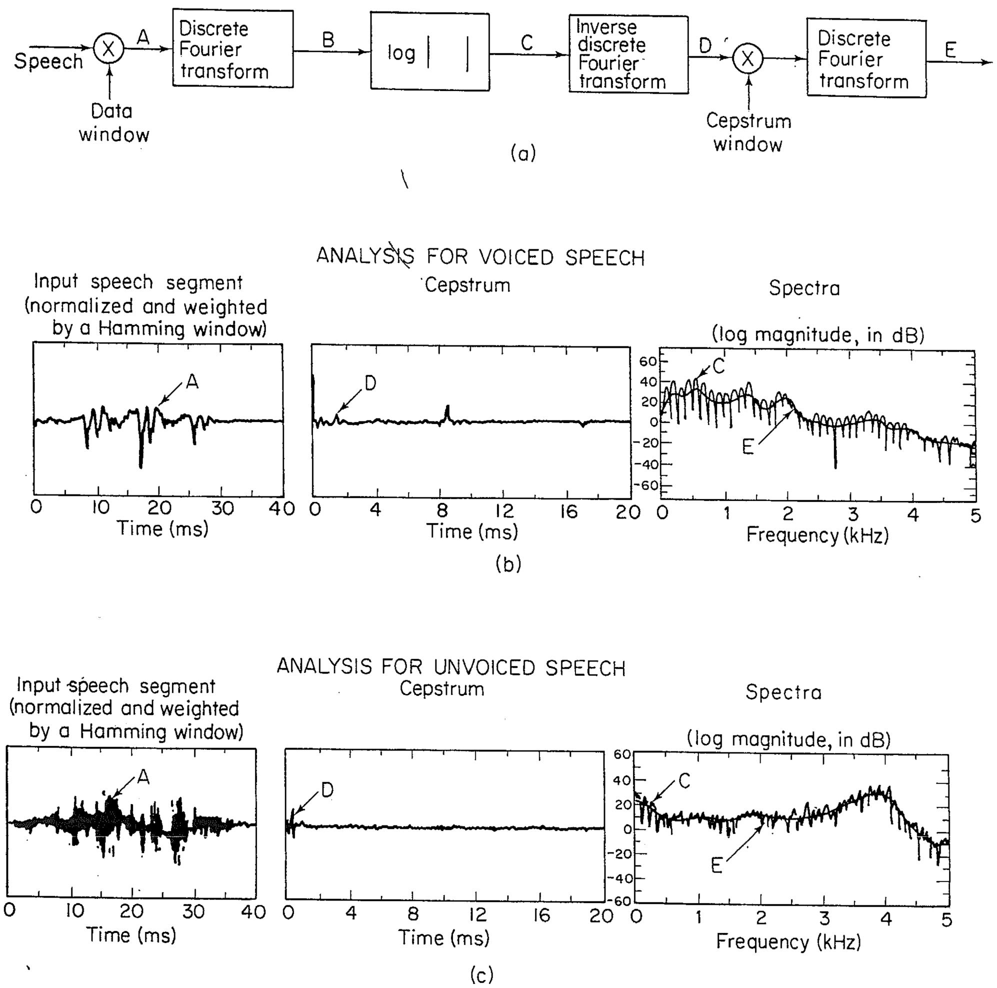

Speech
A
Data window
Discrete Fourier transform
B
log | |
C
Inverse discrete Fourier transform
D
Cepstrum window
Discrete Fourier transform
E
(a)
ANALYSIS FOR VOICED SPEECH
Input speech segment (normalized and weighted by a Hamming window)
A
0 10 20 30 40
Time (ms)
Cepstrum
D
0 4 8 12 16 20
Time (ms)
Spectra (log magnitude, in dB)
60 40 20 0 -20 -40 -60
C
E
0 1 2 3 4 5
Frequency (kHz)
(b)
ANALYSIS FOR UNVOICED SPEECH
Input speech segment (normalized and weighted by a Hamming window)
A
0 10 20 30 40
Time (ms)
Cepstrum
D
0 4 8 12 16 20
Time (ms)
Spectra (log magnitude, in dB)
60 40 20 0 -20 -40 -60
C
E
0 1 2 3 4 5
Frequency (kHz)
(c)

Figure 13.26 (a) System for cepstrum analysis of speech signals. (b) Analysis for voiced speech. (c) Analysis for unvoiced speech.

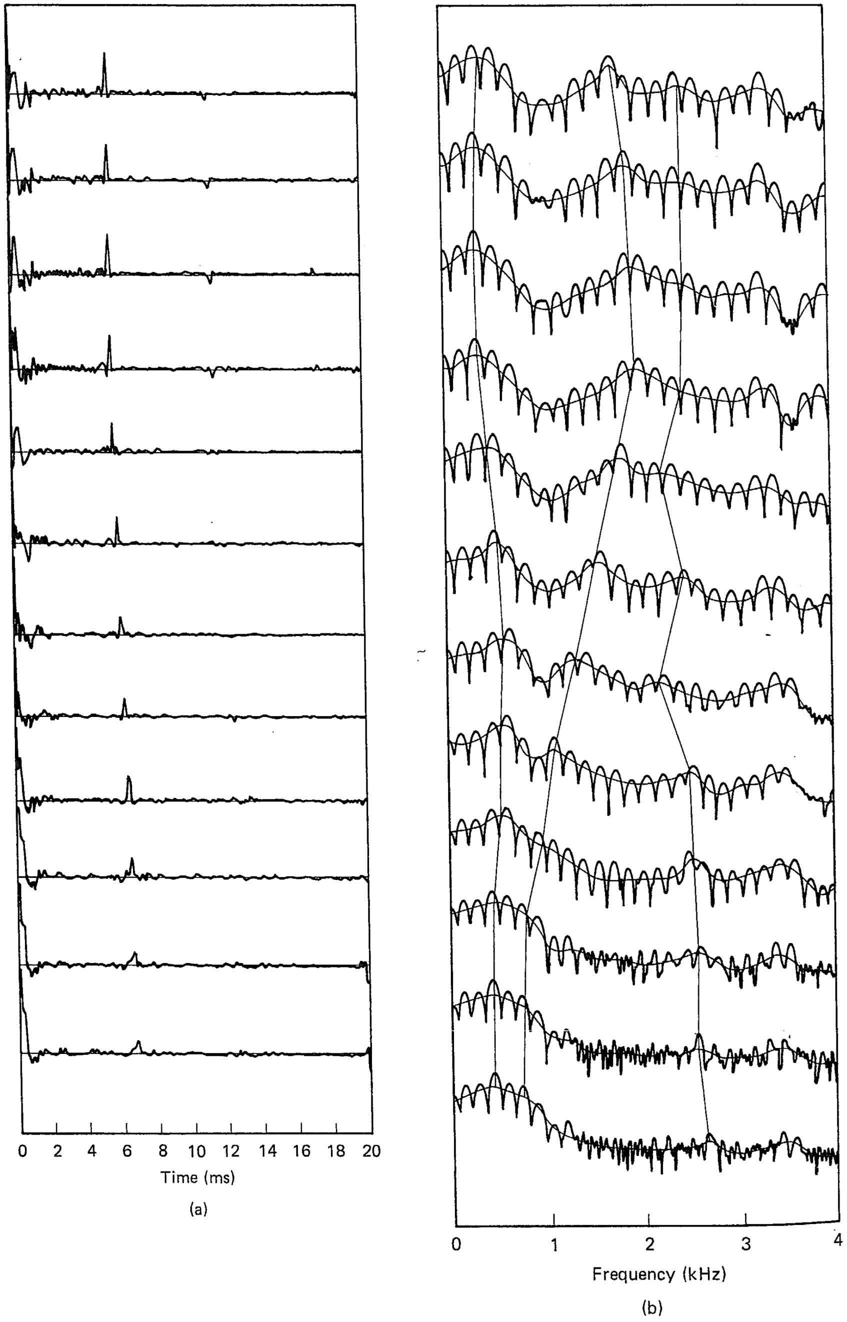


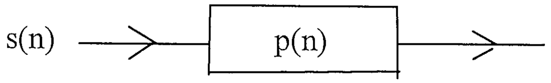
Figure 13.27 (a) Cepstra and (b) log specta for sequential segments of voiced speech.

Homomorphic Deconvolution of Seismic Signals

Two types of seismic signals:

1. Signal generated by an underground explosion and measured after passing through a portion of the earth's crust. (Used to detect underground mineral deposits.)
2. Signal generated by a natural underground event (e.g.,, earthquake)

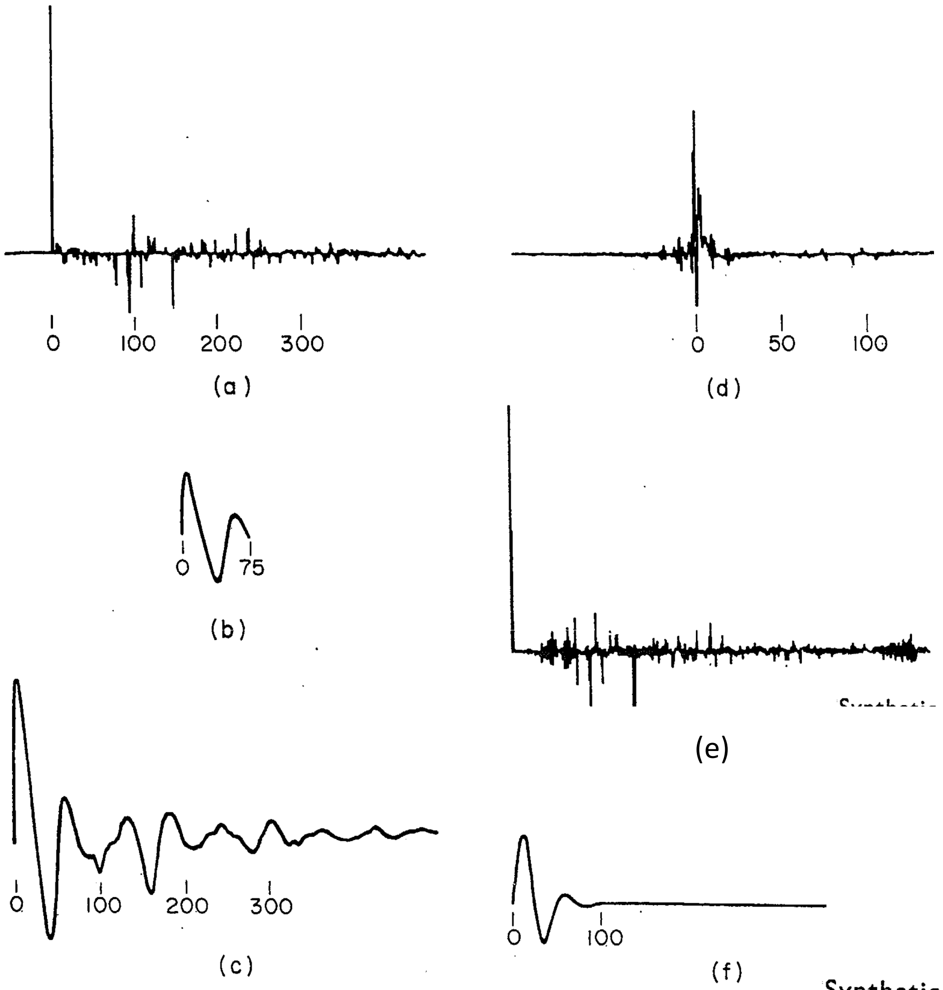
Generation of the first type of seismic signal can be modeled as shown below:



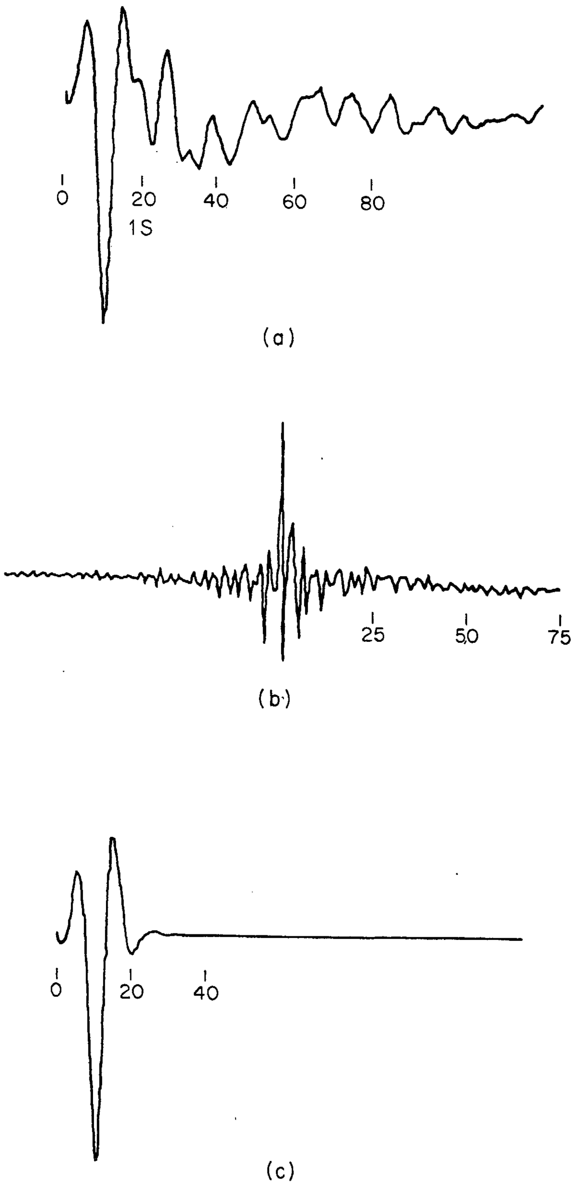
$$x(n) = s(n) * p(n)$$

where s(n) is a seismic wavelet that depends on the nature of the excitation and p(n) is the "impulse response" of the portion of the earth's crust between the excitation and the point of detection.

Cepstral analysis can be used to deconvolve s(n) and p(n). Typical signals are shown in the figures below:



Synthetic example of homomorphic deconvolution of seismic signals: (a) theoretical impulse response of crust near Leduc, Alberta (after O. Jensen); (b) assumed seismic wavelet; (c) synthetic seismogram; (d) complex cepstrum of the trace of (c) exponentially weighted with $\alpha = 0.985$; (e) highpass output; (f) lowpass output. (After Ulrych [15].)



Example of homomorphic deconvolution of an actual teleseismic event: (a) teleseismic event recorded in 1968 at Leduc, Alberta, and originating in Venezuela; (b) complex cepstrum of (a) after exponential weighting with $\alpha = 0.985$, (c) estimate of the seismic wavelet obtained by the low-time filtering (b). (After Ulrych [15].)

Restoration of Acoustic Recordings

A old (low quality) audio recording can be represented as a convolution of the true signal with a unit sample response which represented the "distorting system," as shown in the figure below:

$$s(n) \longrightarrow \boxed{h(n)} \longrightarrow x(n) = s(n) * h(n)$$

Processing approach: estimate h(n) so that its effect can be compensated by inverse filtering.

First, section the available signal into M sections:

$x_m(n) = x(n + mN)$ $\quad n = 0,1, \ldots, N\text{-}1$ (index within section)

$\quad m = 0,1, \ldots, M\text{-}1$ (section index)

Assume that $x_m(n) \approx s_m(n) * h(n)$.

Therefore, $X_m(e^{j\omega}) \simeq S_m(e^{j\omega})H(e^{j\omega})$

and $\log|X_m(e^{j\omega})| \simeq \log|S_m(e^{j\omega})| + \log|H(e^{j\omega})|$.

We can obtain an estimate $H_e(e^{j\omega})$ of the frequency response of the distorting system $H(e^{j\omega})$ by obtaining averages of $\log|X_m(e^{j\omega})|$ and of $\log|S_m(e^{j\omega})|$:

$$\log[H_e(e^{j\omega})] = \frac{1}{M}\sum_{m=0}^{M-1} \log|X_m(e^{j\omega})| - \underbrace{\frac{1}{M}\sum_{m=0}^{M-1} \log|S_m(e^{j\omega})|}_{\overline{\log[S(e^{j\omega})]}}.$$

We can obtain $\frac{1}{M}\sum_{m=0}^{M-1}\log|X_m(e^{j\omega})|$ by averaging $\log|X_m(e^{j\omega})|$ over M data segments.

We can obtain $\frac{1}{M}\sum_{m=0}^{M-1}\log|S_m(e^{j\omega})|$, which is an estimate of the "long-time" power spectrum of high quality music or singing, from highly quality recordings of music or singing. This averaging is performed on music of the same music type as the music being processed.

The desired inverse filter to be used for removing the recording distortion is then

$$\begin{aligned} H_e^{-1}(e^{j\omega}) &= \frac{1}{H_e(e^{j\omega})}, && |\omega| < \omega_p \\ &= 0, && \omega_s < |\omega| < \pi \end{aligned}$$

After estimating $H_e^{-1}(e^{j\omega})$ as shown above, the remaining processing steps are:

1. Use IDFT of samples of $H_e^{-1}(e^{j\omega})$ to obtain $h_e^{-1}(n)$.
2. Convolve $h_e^{-1}(n)$ with x(n) to "undo" the distortion introduced by the low-quality recording system.

Homomorphic Image Processing

We can represent the generation of an image using the following relation:

$$f(u,v) = f_i(u,v)f_r(u,v)$$

where $f(u,v)$ is the image,

$f_i(u,v)$ is an illumination function,

and $f_r(u,v)$ is a reflection function which satisfies $0 < f_r(u,v) < 1$.

Therefore, $0 < f(u,v) < f_i(u,v) < \infty$.

Digital images: sampled versions of $f(u,v)$:

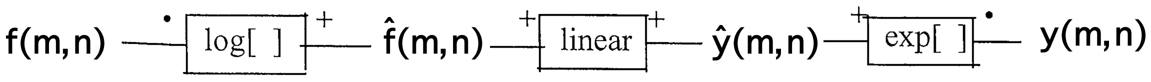
$$f(m,n) = f_i(m,n)f_r(m,n)$$

Example of processing goals:

Use homomorphic system for multiplication to:

(a) reduce the dynamic range (for communications and/or storage)

(b) enhance contrast (sharpen edges) in the image.



(Recall that $f(m,n) > 0$ so that real log can be used.)

$$\hat{f}(m,n) = \log[f_i(m,n)f_r(m,n] = \log[f_i(m,n)] + \log[f_r(m,n)]$$

$$= \hat{f}_i(m,n) + \hat{f}_r(m,n)$$

The output of the middle box (linear system) is

$$\hat{y}(m,n) = L[\hat{f}(m,n)]$$

$$= L[\hat{f}_i(m,n) + \hat{f}_r(m,n)] = L[\hat{f}_i(m,n)] + L[\hat{f}_r(m,n)]$$

and the output of the final box (exponentiation) is therefore

$$y(m,n) = \exp[\hat{y}(m,n)]$$

$$= \exp\left\{L[\hat{f}_i(m,n)] + L[\hat{f}_r(m,n)]\right\}$$

$$= \exp\left\{L[\hat{f}_i(m,n)]\right\} \bullet \exp\left\{L[\hat{f}_r(m,n)]\right\}.$$

<u>Basis for choosing the linear system</u>

- <u>Illumination usually varies "slowly"</u> (gradually) across a scene (although it may vary a large amount over the entire scene, and therefore have a large overall dynamic range).
- The <u>reflective component may vary "rapidly</u>," due to sharp edges and changes of texture.

Therefore, consider the following linear system for the middle box in the overall system:

$\hat{y}(m,n) = \gamma \cdot \hat{f}(m,n)$

Then the overall output will be

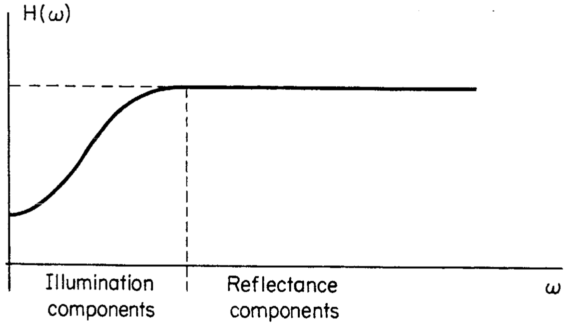
$$
\begin{aligned}
y(m,n) &= \exp[\hat{y}(m,n)] = \exp[\gamma \cdot \hat{f}(m,n)] \\
&= \exp[\gamma \cdot (\hat{f}_i(m,n) + \hat{f}_r(m,n))] \\
&= \exp[\gamma \cdot (\hat{f}_i(m,n)]\exp[\gamma \cdot \hat{f}_r m,n)]
\end{aligned}
$$

$$y(m,n) = [f_i(m,n)]^{\gamma}[f_r(m,n)]^{\gamma}.$$

<u>To reduce the dynamic range</u>, we need $\gamma < 1$.

<u>To increase the contrast</u> of the image (increase the ratio between two different intensities), we need $\gamma > 1$.

However, since $f_i(m,n)$ is a "low-frequency" signal and $f_r(m,n)$ is a "high frequency" signal, we can use the following linear system to target both goals:



Cross section of a circularly symmetric frequency response to be used for the linear part of a homomorphic image processor to achieve simultaneous contrast enhancement and dynamic range compression.

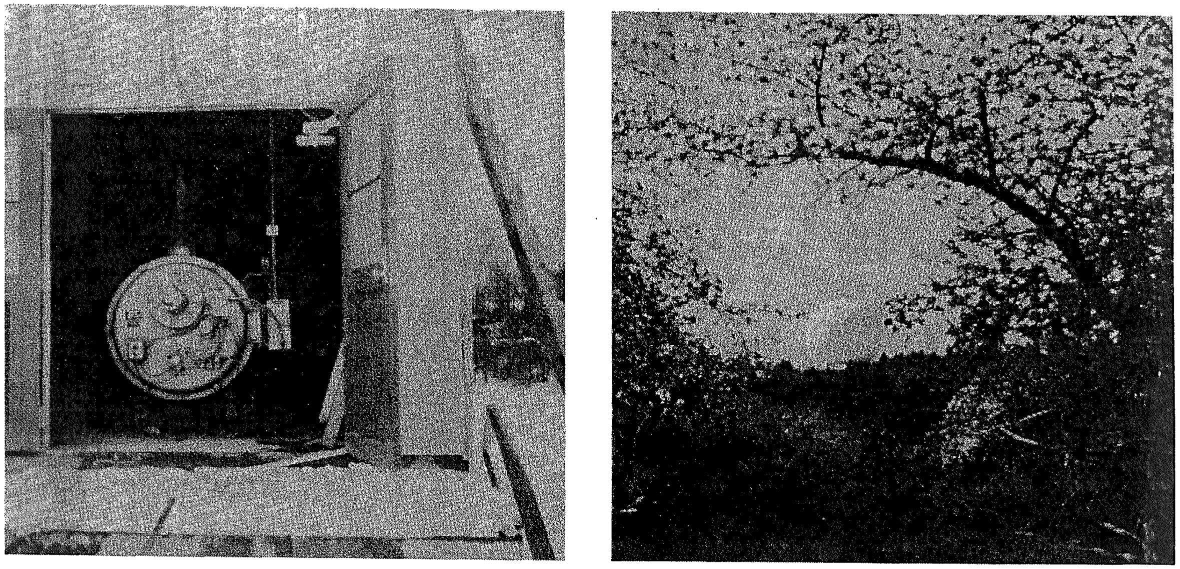


Fig. 10.7 Two original images.

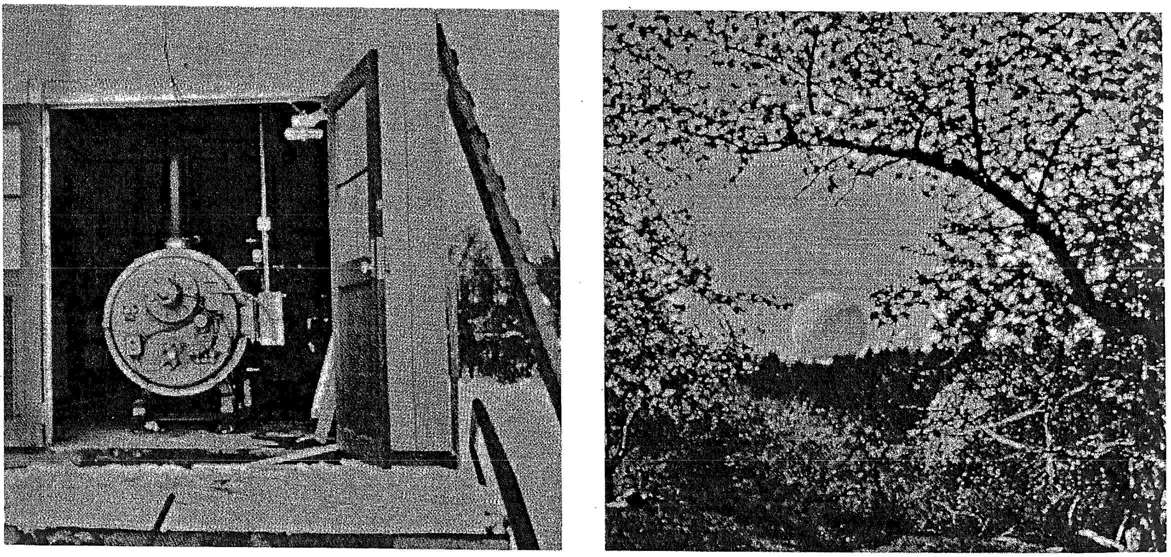


Fig. 10.8 Images of Fig. 10.7 after processing to achieve simultaneous dynamic range compression and contrast enhancement.

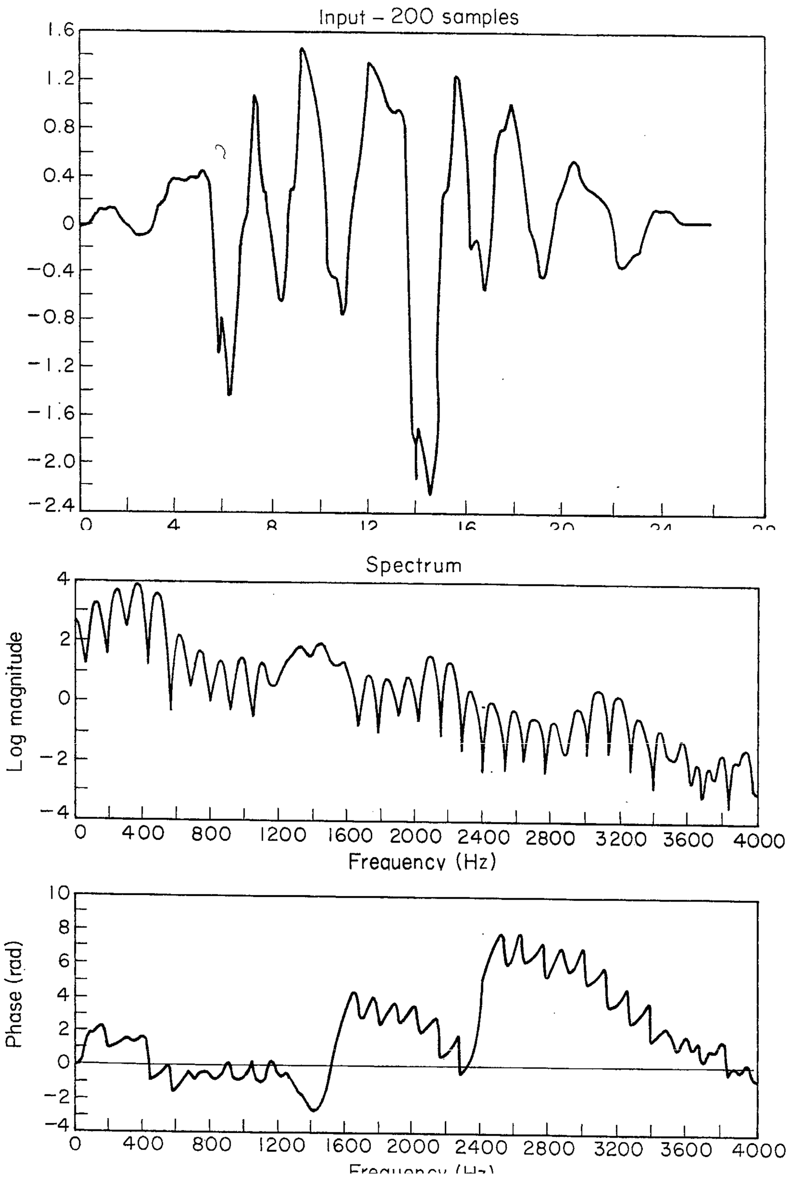


Figure 12.25 (a) Segment of speech weighted by a Hamming window. (b) Complex logarithm of the discrete Fourier transform of part (a).

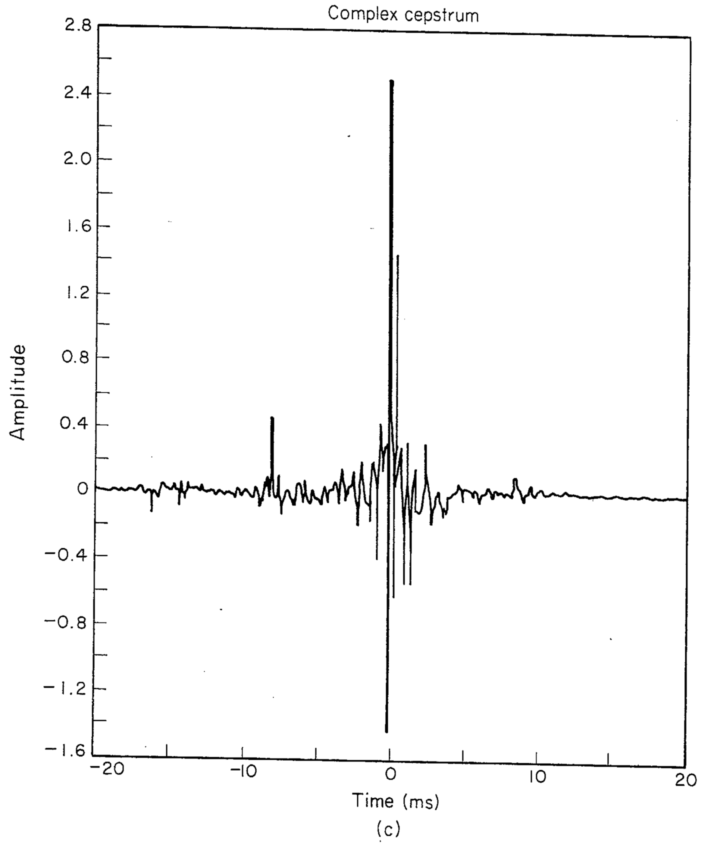


Figure 12.25 (*continued*) (c) Complex cepstrum of part (a).