Speech Signal Representation

- Fourier Analysis
  - Discrete-time Fourier transform
  - Short-time Fourier transform
  - Discrete Fourier transform
- Cepstral Analysis
  - The complex cepstrum and the cepstrum
  - Computational considerations
  - Cepstral analysis of speech
  - Applications to speech recognition
  - Mel-Frequency cepstral representation
- Performance Comparison of Various Representations

Discrete-Time Fourier Transform

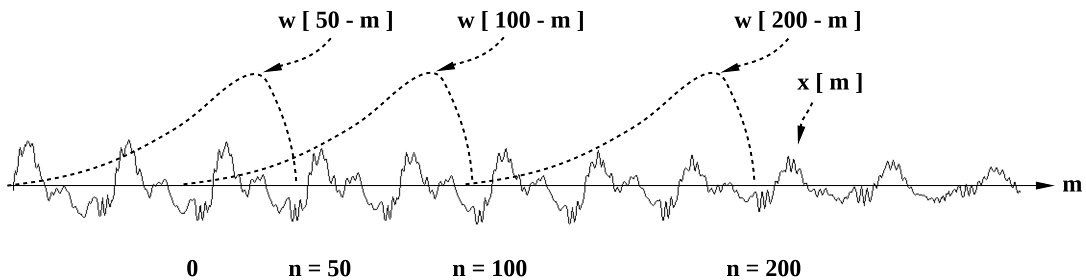
$$\begin{cases} X(e^{j\omega}) & = \sum_{n=-\infty}^{+\infty} x[n]e^{-j\omega n} \\ x[n] & = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega \end{cases}$$

- Sufficient condition for convergence: $\sum_{n=-\infty}^{+\infty} \left| x[n] \right| < +\infty$
- Although $x[n]$ is discrete, $X(e^{j\omega})$ is continuous and periodic with period 2π.
- Convolution/multiplication duality:

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

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

Short-Time Fourier Analysis (Time-Dependent Fourier Transform)



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

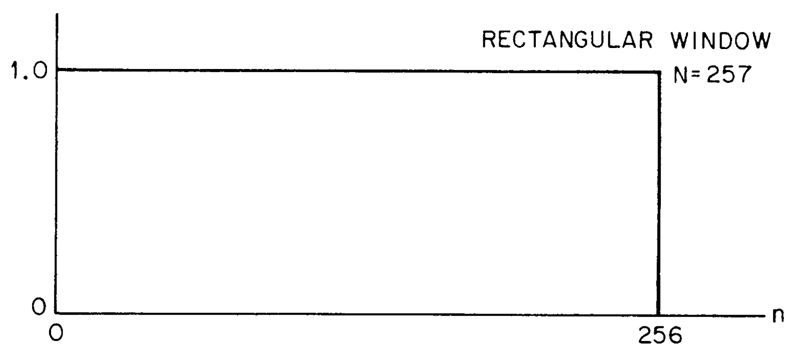
- If n is fixed, then it can be shown that:

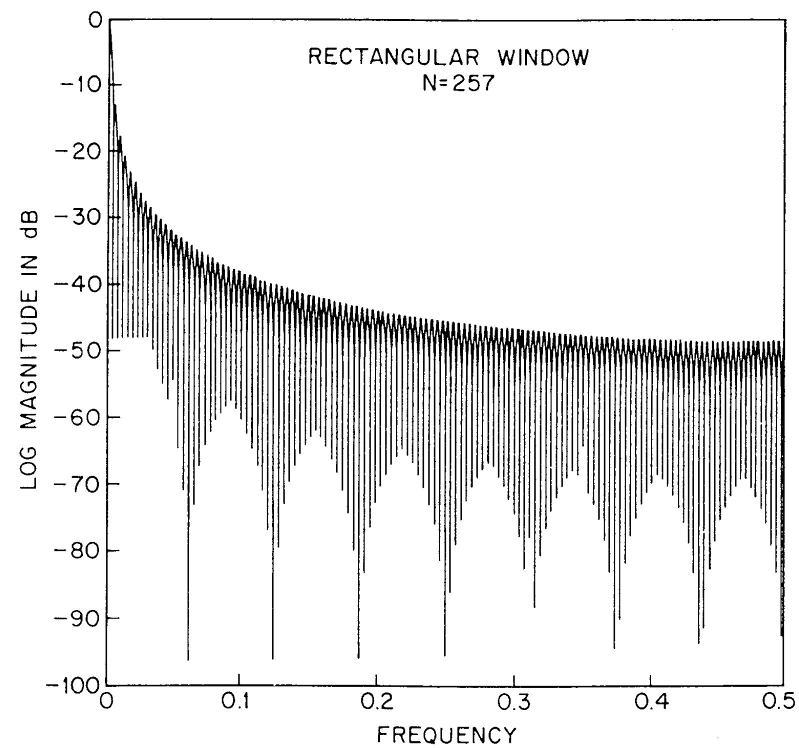
$$X_n(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{j\theta})e^{j\theta n}X(e^{j(\omega+\theta)})d\theta$$

- The above equation is meaningful only if we assume that $X(e^{j\omega})$ represents the Fourier transform of a signal whose properties continue outside the window, or simply that the signal is zero outside the window.
- In order for $X_n(e^{j\omega})$ to correspond to $X(e^{j\omega})$, $W(e^{j\omega})$ must resemble an impulse with respect to $X(e^{j\omega})$.

Rectangular Window

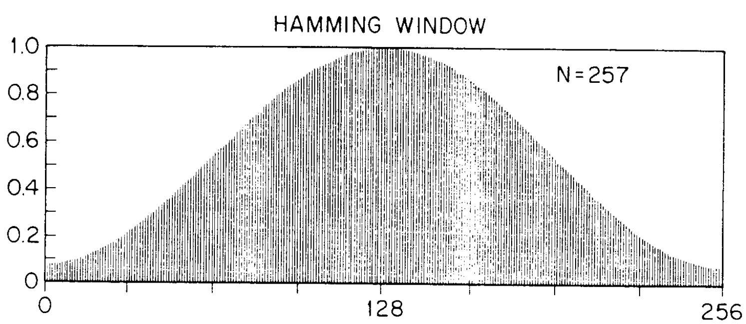
$$w[n] = 1, \qquad 0 \le n \le N-1$$

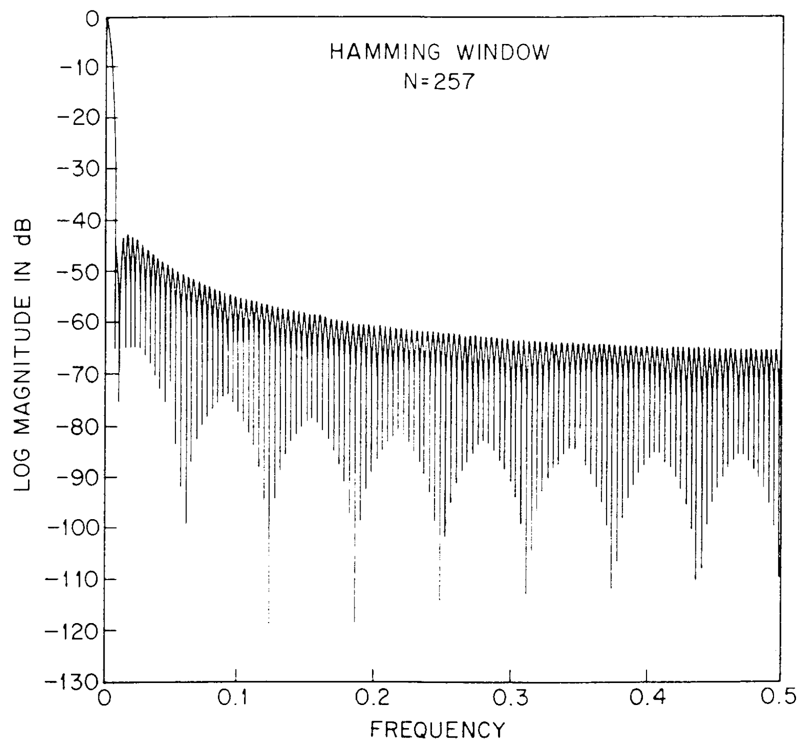




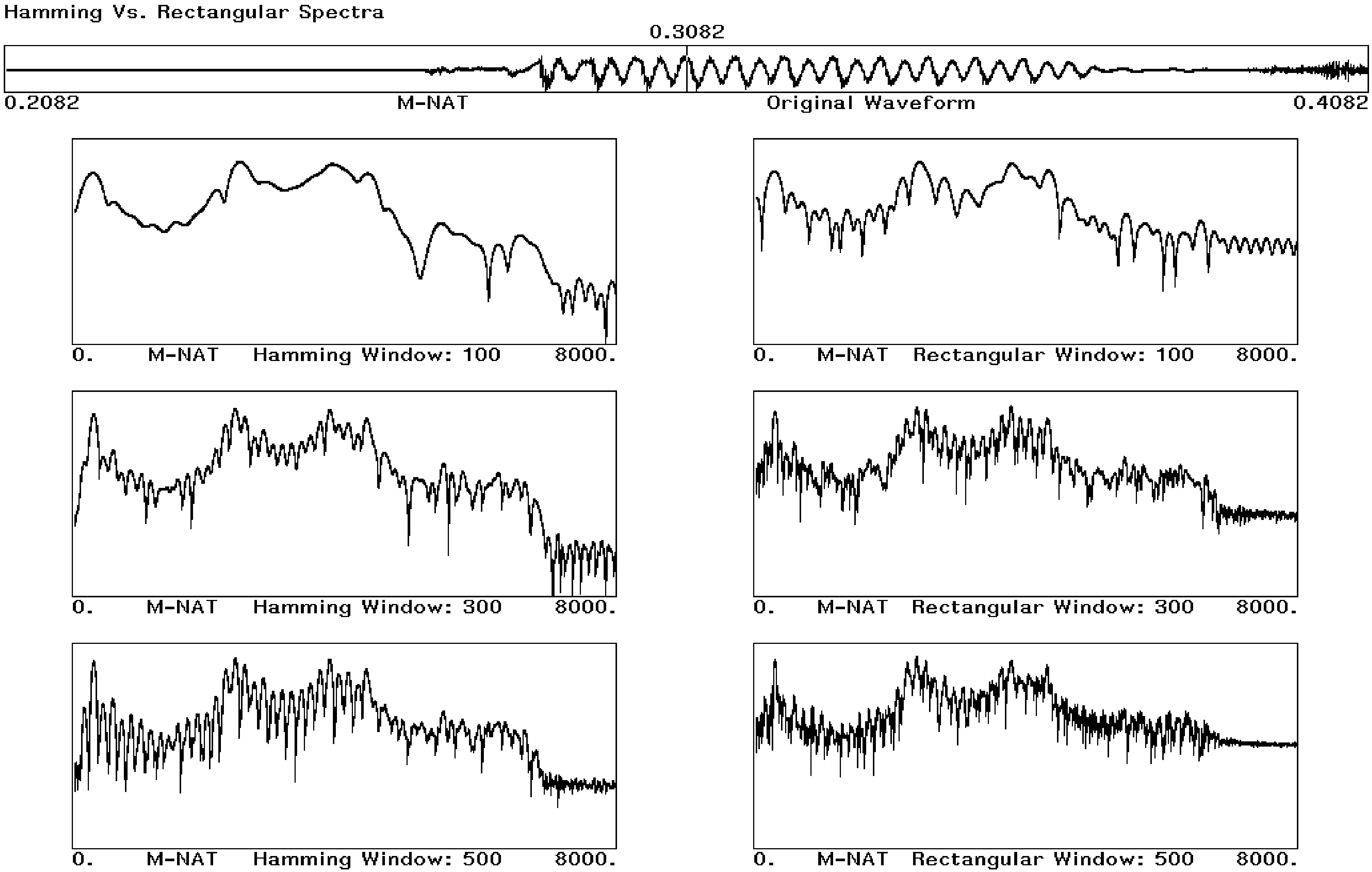
Hamming Window

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

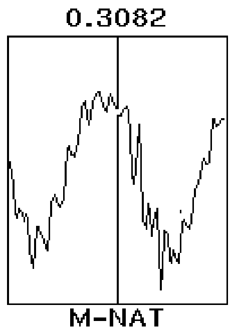


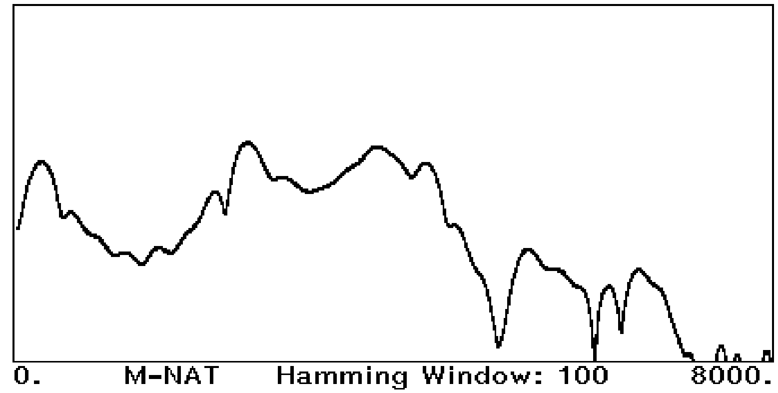


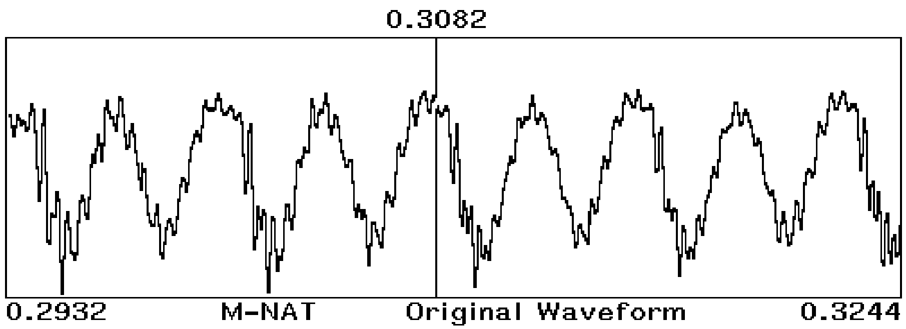
Comparison of Windows

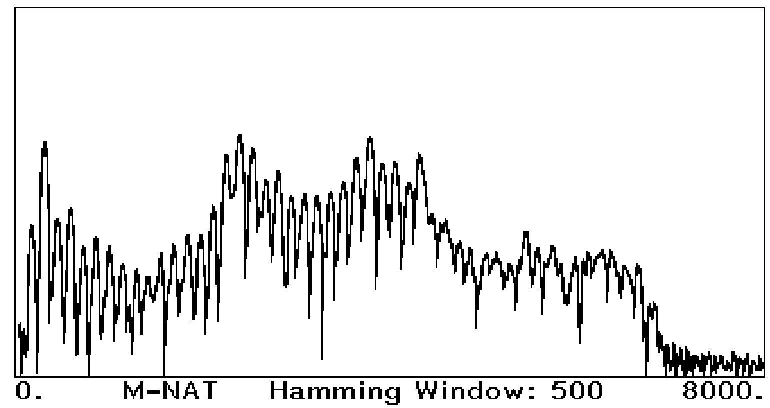


Comparison of Windows (cont'd)

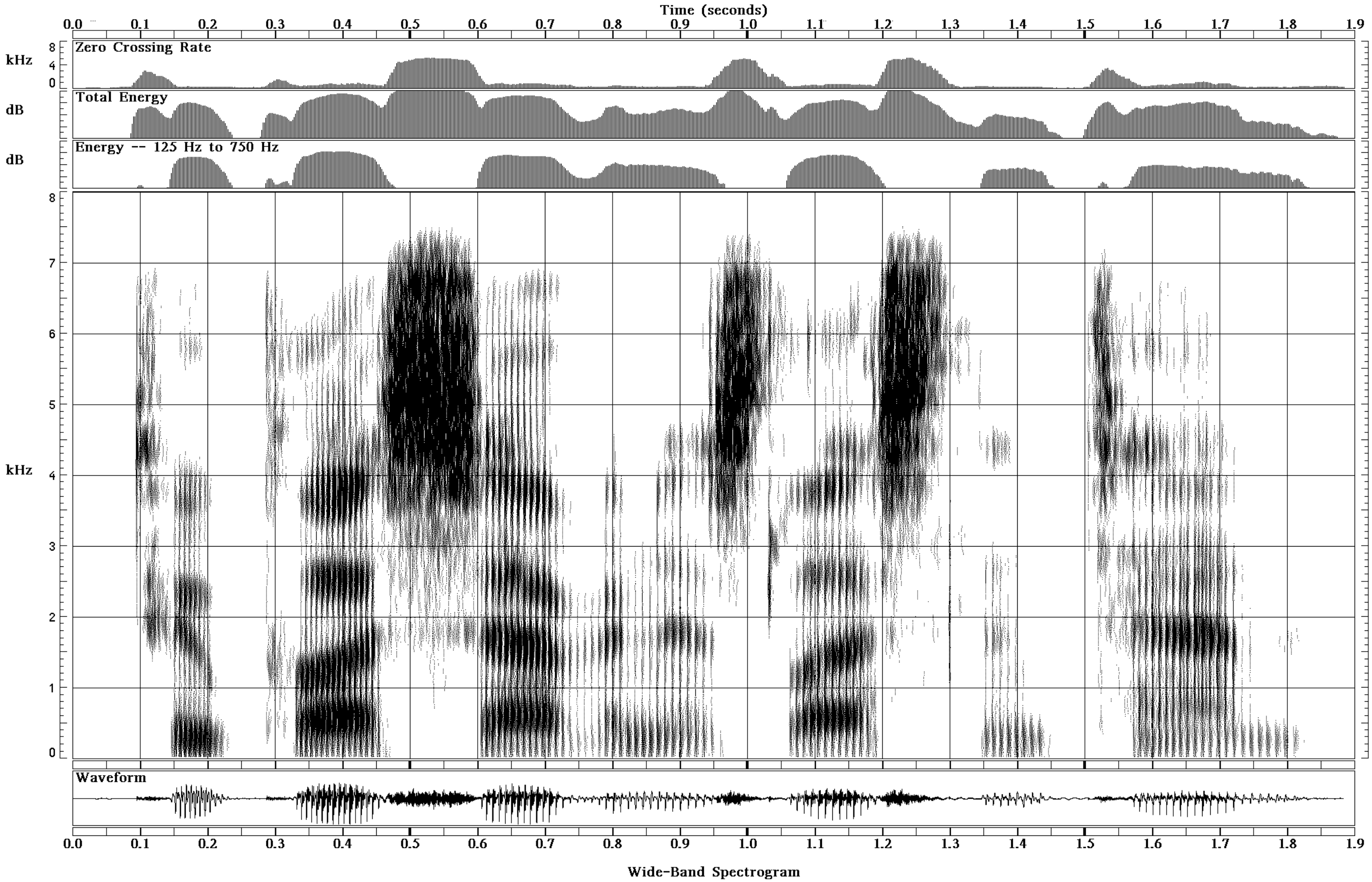






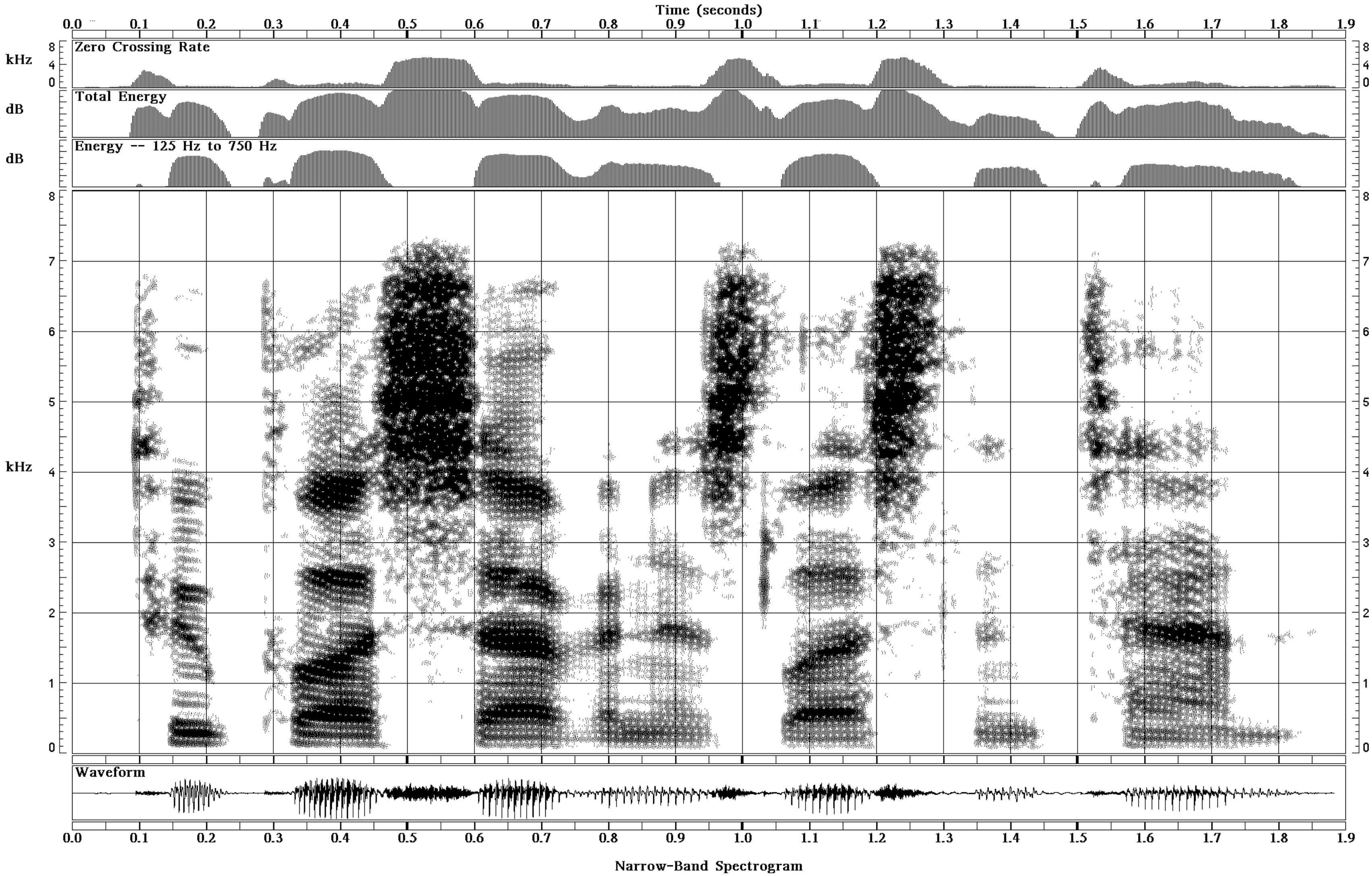


A Wideband Spectrogram



Two plus seven is less than ten

A Narrowband Spectrogram



Two plus seven is less than ten

Discrete Fourier Transform

$$x[n] \quad \Longleftrightarrow \quad X[k] = X(z)\,\big|_{z=e^{j\frac{2\pi k}{M}n}}$$

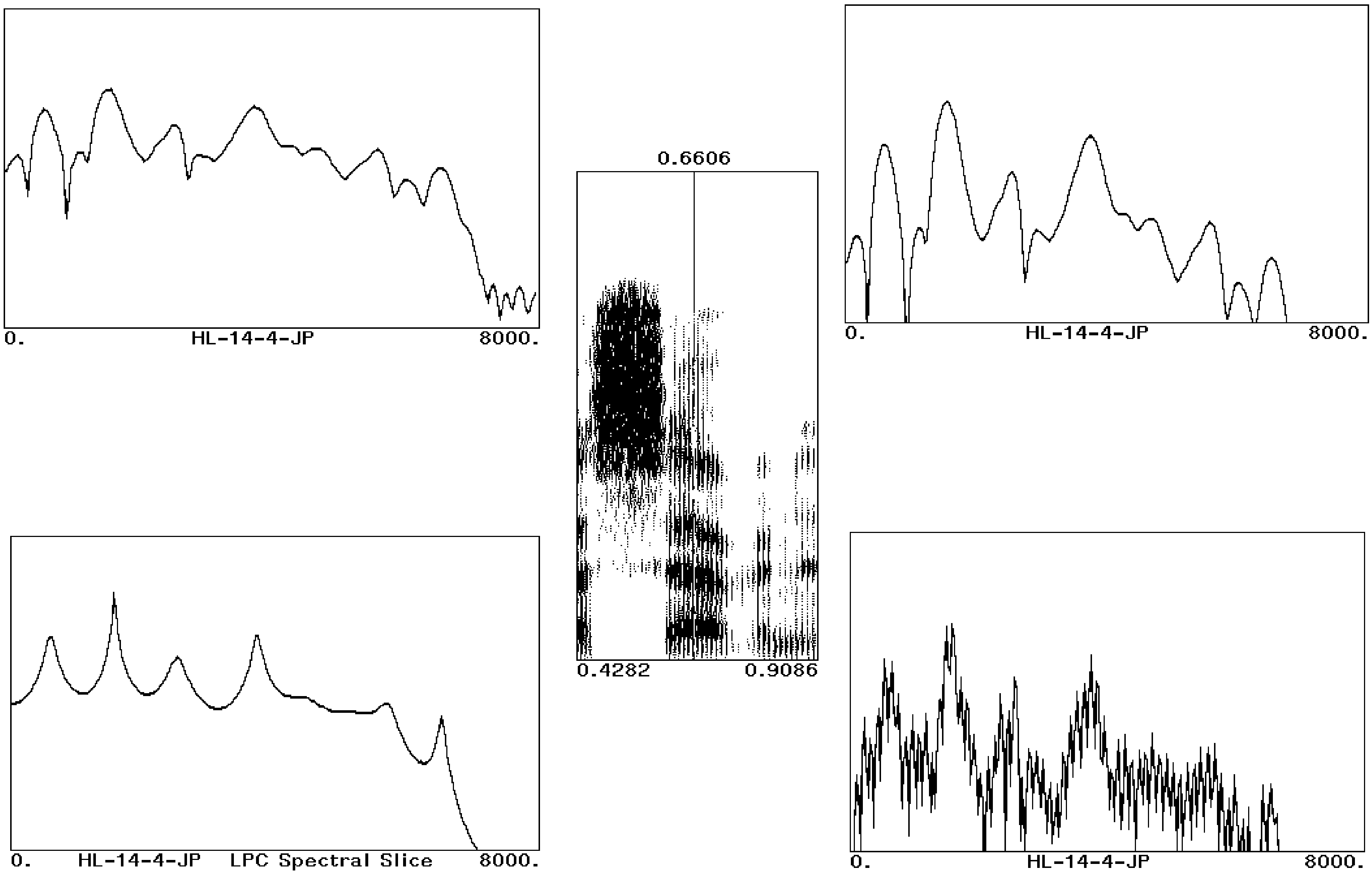
$$N points \qquad\qquad M points$$

$$\begin{cases} X[k] & = \displaystyle\sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi k}{M}n} \\ x[n] & = \displaystyle\frac{1}{M}\sum_{k=0}^{M-1} X[k] e^{j\frac{2\pi k}{M}n} \end{cases}$$

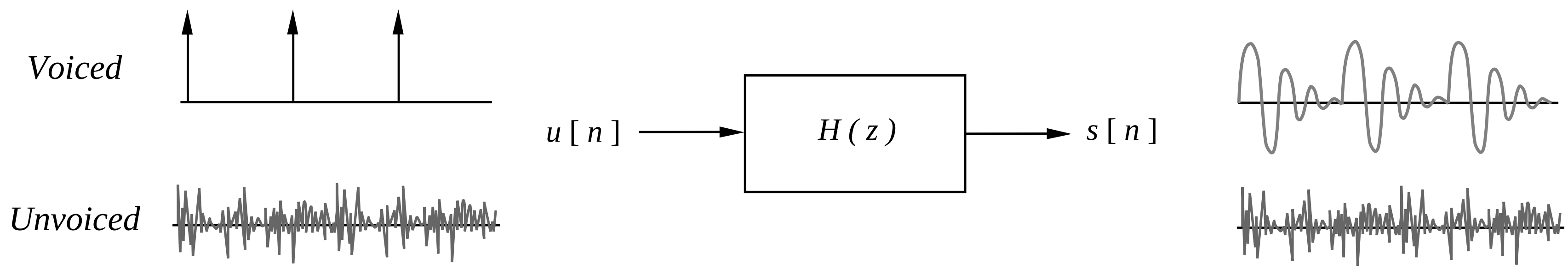
In general, the number of input points, N, and the number of frequency samples, M, need not be the same.

- If $M > N$, we must zero-pad the signal
- If $M < N$, we must time-alias the signal

Examples of Various Spectral Representations



Cepstral Analysis of Speech



- The speech signal is often assumed to be the output of an LTI system; i.e., it is the convolution of the input and the impulse response.
- If we are interested in characterizing the signal in terms of the parameters of such a model, we must go through the process of de-convolution.
- Cepstral, analysis is a common procedure used for such de-convolution.

Cepstral Analysis

- Cepstral analysis for convolution is based on the observation that:

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

  By taking the *complex* logarithm of $X(z)$, then

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

- If the complex logarithm is unique, and if $\hat{X}(z)$ is a valid z-transform, then

$$\hat{x}(n) = \hat{x}_1(n) + \hat{x}_2(n)$$

  The two convolved signals will be additive in this new, cepstral domain.

- If we restrict ourselves to the unit circle, $z = e^{j\omega}$, then:

$$\hat{X}(e^{j\omega}) = \log|X(e^{j\omega})| + j\ \arg\{X(e^{j\omega})\}$$

  It can be shown that one approach to dealing with the problem of uniqueness is to require that $\arg\{X(e^{j\omega})\}$ be a continuous, odd, periodic function of ω.

Cepstral Analysis (cont'd)

- To the extent that $\hat{X}(z) = \log\{X(z)\}$ is valid,

$$\begin{cases} \hat{x}[n] & = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \hat{X}(e^{j\omega})\, e^{j\omega n} d\omega \\ & = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log\{X(e^{j\omega})\}\, e^{j\omega n} d\omega \qquad \text{complex cepstrum} \\ c[n] & = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log|X(e^{j\omega})|\, e^{j\omega n} d\omega \qquad \text{cepstrum} \end{cases}$$

- It can easily be shown that $c[n]$ is the even part of $\hat{x}[n]$.
- If $\hat{x}[n]$ is real and causal, then $\hat{x}[n]$ be recovered from $c[n]$. This is known as the Minimum Phase condition.

An Example

$$p[n] = \delta[n] + \alpha\delta[n-N] \qquad 0 < \alpha < 1$$

$$P(z) = 1 + \alpha z^{-N}$$

$$\hat{P}(z) = \log\left[P(z)\right] = \log\left[1 + \alpha z^{-N}\right]$$

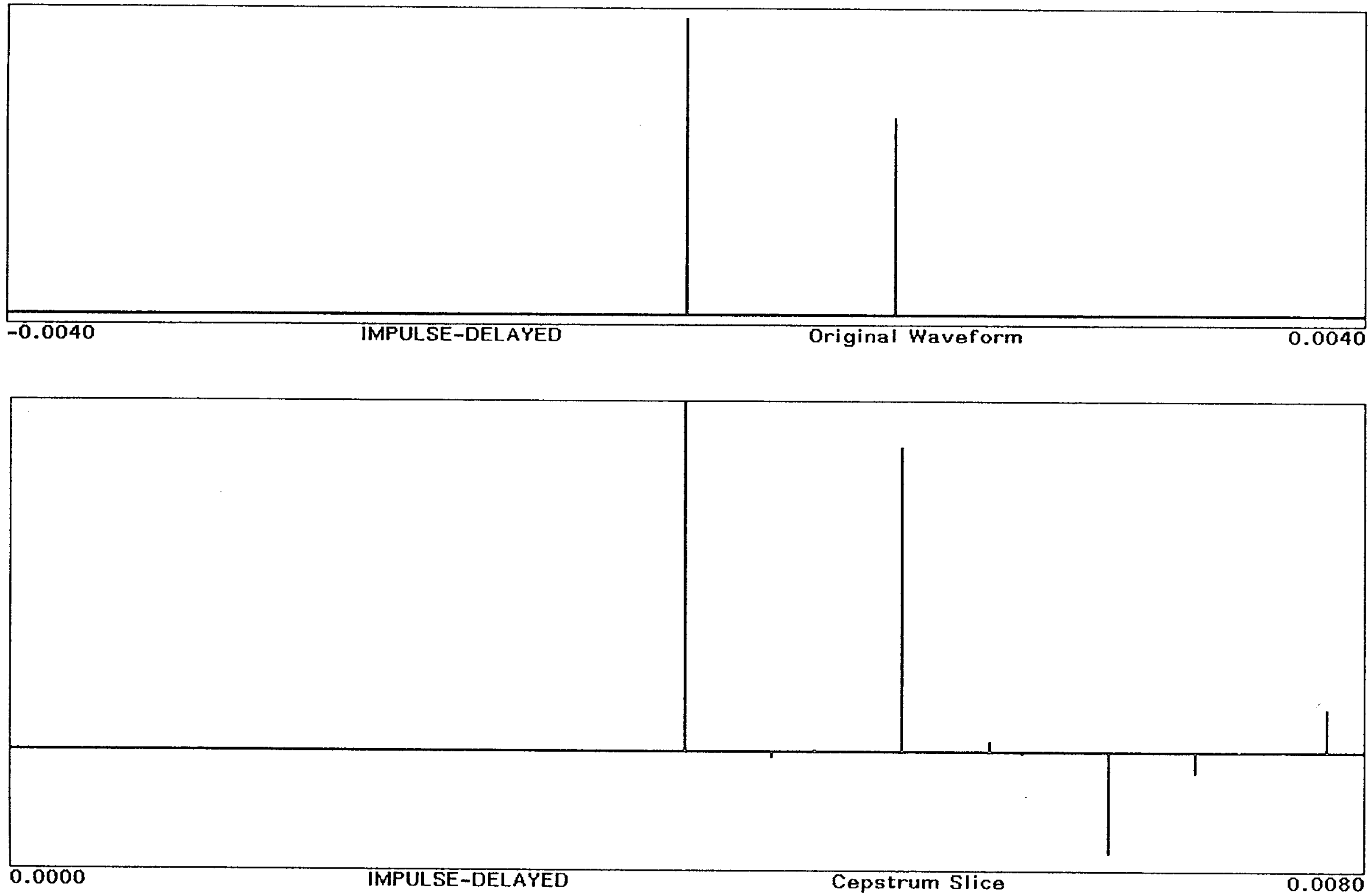
$$= \log\left[1 - (-\alpha)(z^{N})^{-1}\right]$$

$$= \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\alpha^n}{n} z^{-nN}$$

$$\hat{P}(z) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\alpha^n}{n} (z^{N})^{-n}$$

$$\hat{p}[n] = \sum_{r=1}^{\infty} (-1)^{r+1} \frac{\alpha^r}{r} \delta[n - rN]$$

An Example (cont'd)



Computational Considerations

- We now replace the Fourier transform expressions by the discrete Fourier transform expressions :

$$\begin{cases} X_p[k] & = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} & 0 \leq k \leq N-1 \\ \hat{X}_p[k] & = \log\{X_p[k]\} & 0 \leq k \leq N-1 \\ \hat{x}_p[n] & = \frac{1}{N}\sum_{k=0}^{N-1} \hat{X}_p[k]\, e^{j\frac{2\pi}{N}kn} & 0 \leq n \leq N-1 \end{cases}$$

- $\hat{X}_p[k]$ is a sampled version of $\hat{X}(e^{j\omega})$. Therefore,

$$\hat{x}_p[n] = \sum_{r=-\infty}^{\infty} \hat{x}[n + rN]$$

- Likewise:

$$c_p[n] = \sum_{r=-\infty}^{\infty} c[n + rN]$$

  where,

$$c_p[n] = \frac{1}{N}\sum_{k=0}^{N-1} \log|X_p[k]|\, e^{j\frac{2\pi}{N}kn} \quad 0 \leq n \leq N-1$$

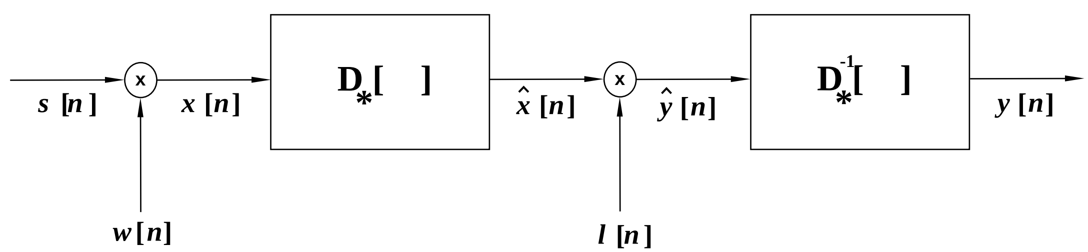
- To minimize aliasing, N must be large.

Cepstral Analysis of Speech

- For voiced speech:

  $$s[n] = p[n] * g[n] * v[n] * r[n] = p[n] * h_v[n] = \sum_{r=-\infty}^{\infty} h_v[n - rN_p].$$

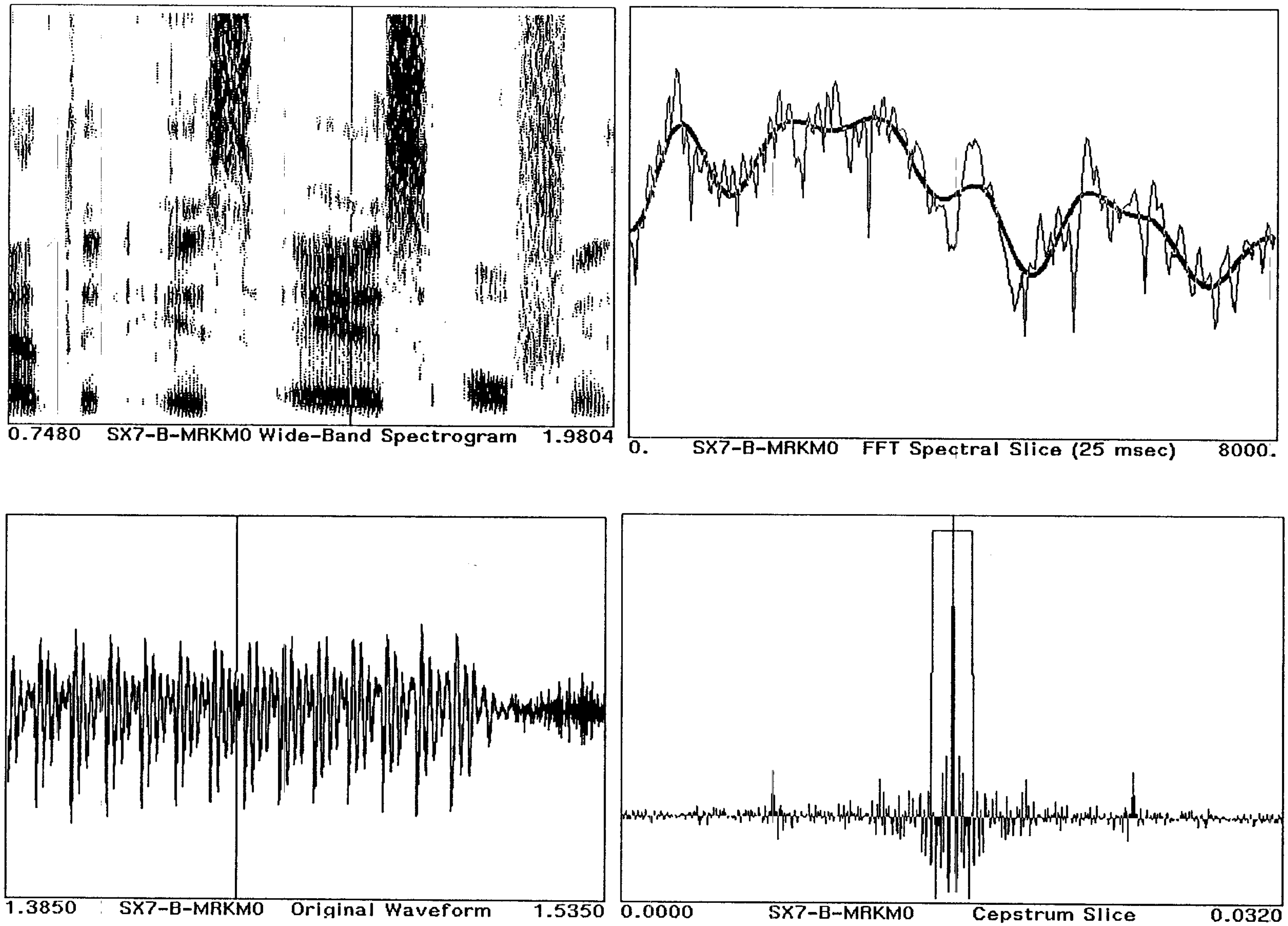
- For unvoiced speech: $s[n] = w[n] * v[n] * r[n] = w[n] * h_u[n]$.
- Contributions to the cepstrum due to periodic excitation will occur at integer multiples of the fundamental period.
- Contributions due to the glottal waveform (for voiced speech), vocal tract, and radiation will be concentrated in the low *quefrency* region, and will decay rapidly with n.
- Deconvolution can be achieved by multiplying the cepstrum with an appropriate window, $l[n]$.



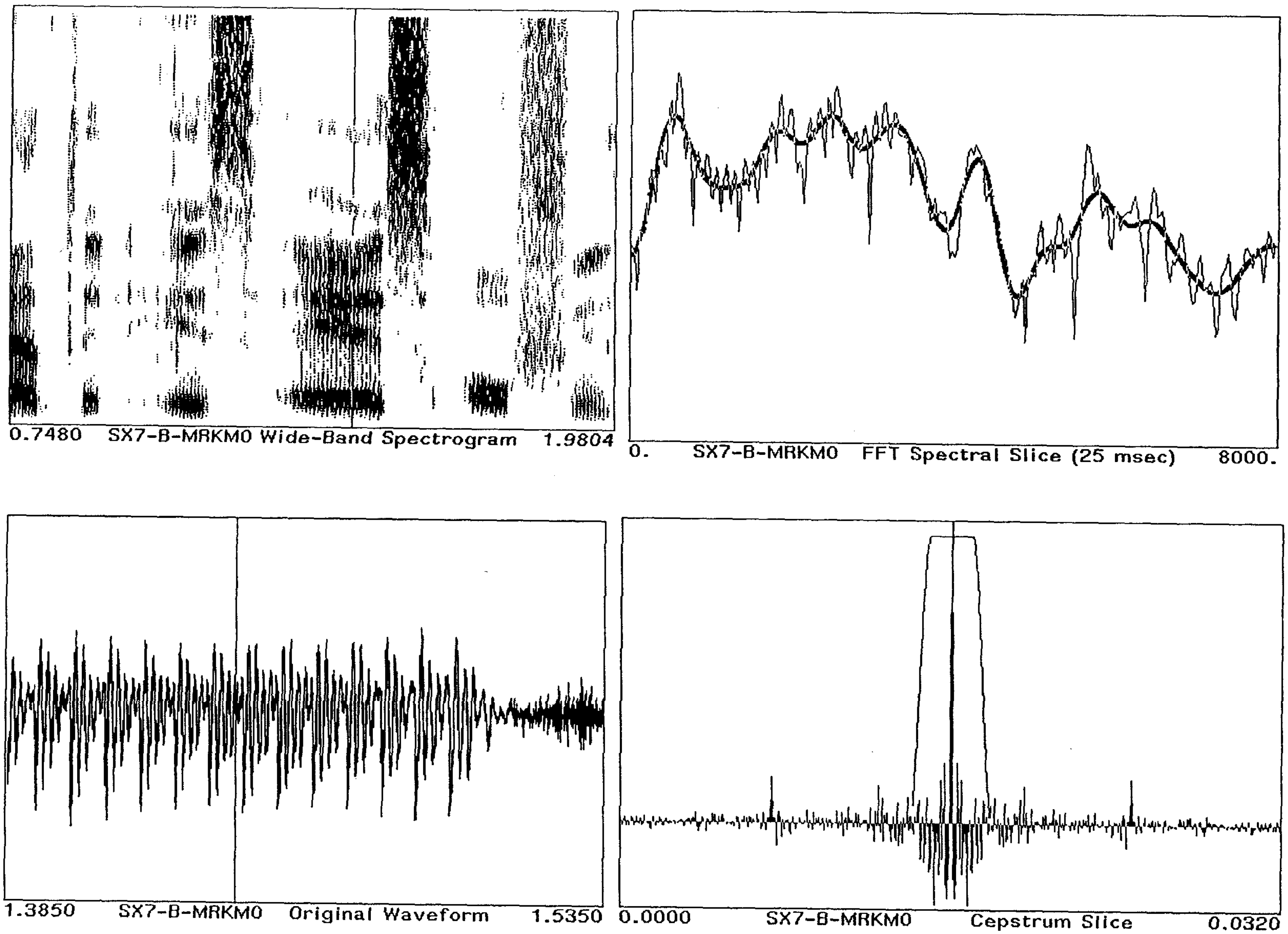
  where D_* is the characteristic system that converts convolution into addition.

- Thus cepstral analysis can be used for pitch extraction and formant tracking.

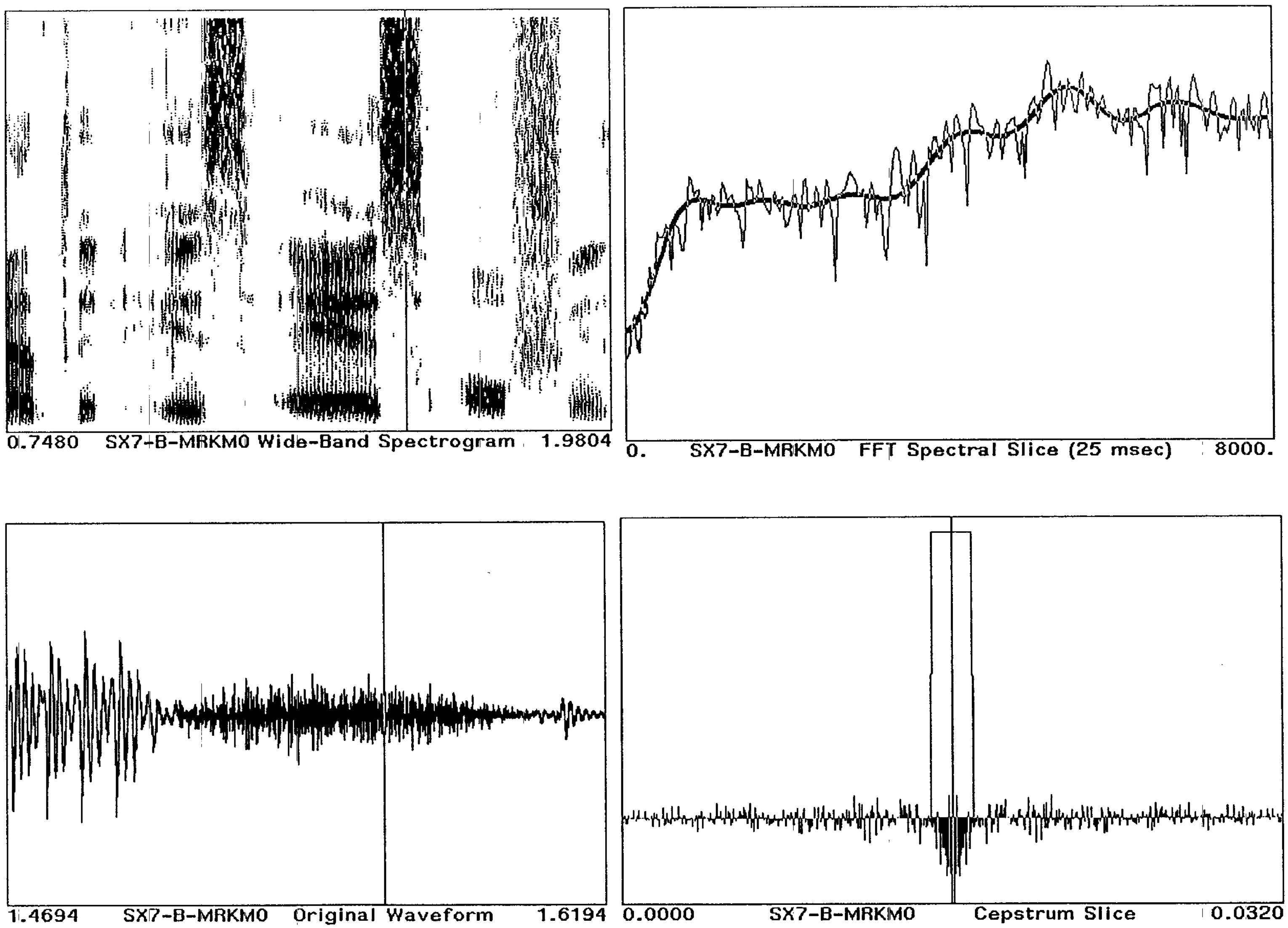
Example of Cepstral Analysis of Vowel
(Rectangular Window)



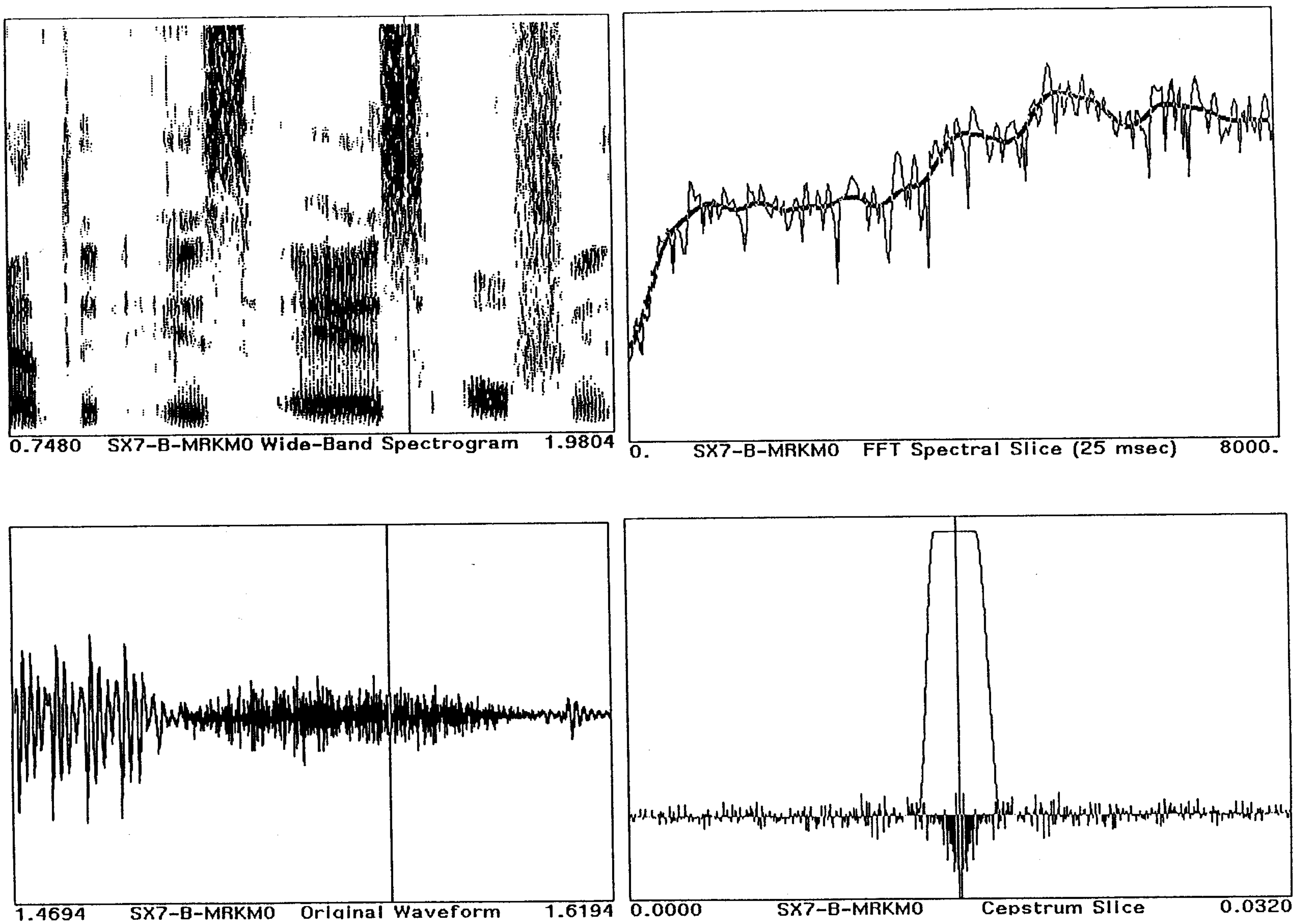
Example of Cepstral Analysis of Vowel (Tapering Window)



Example of Cepstral Analysis of Fricative (Rectangular Window)

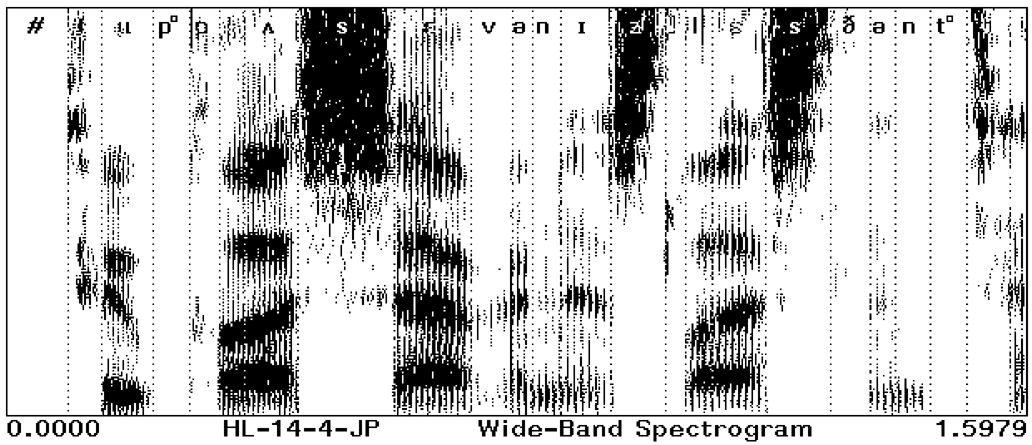


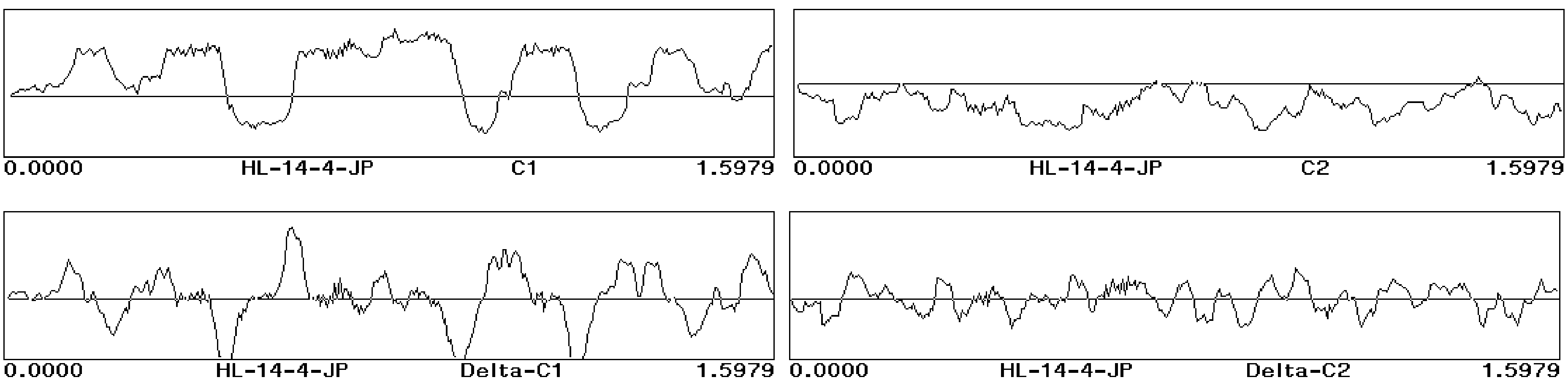
Example of Cepstral Analysis of Fricative (Tapering Window)



The Use of Cepstrum for Speech Recognition

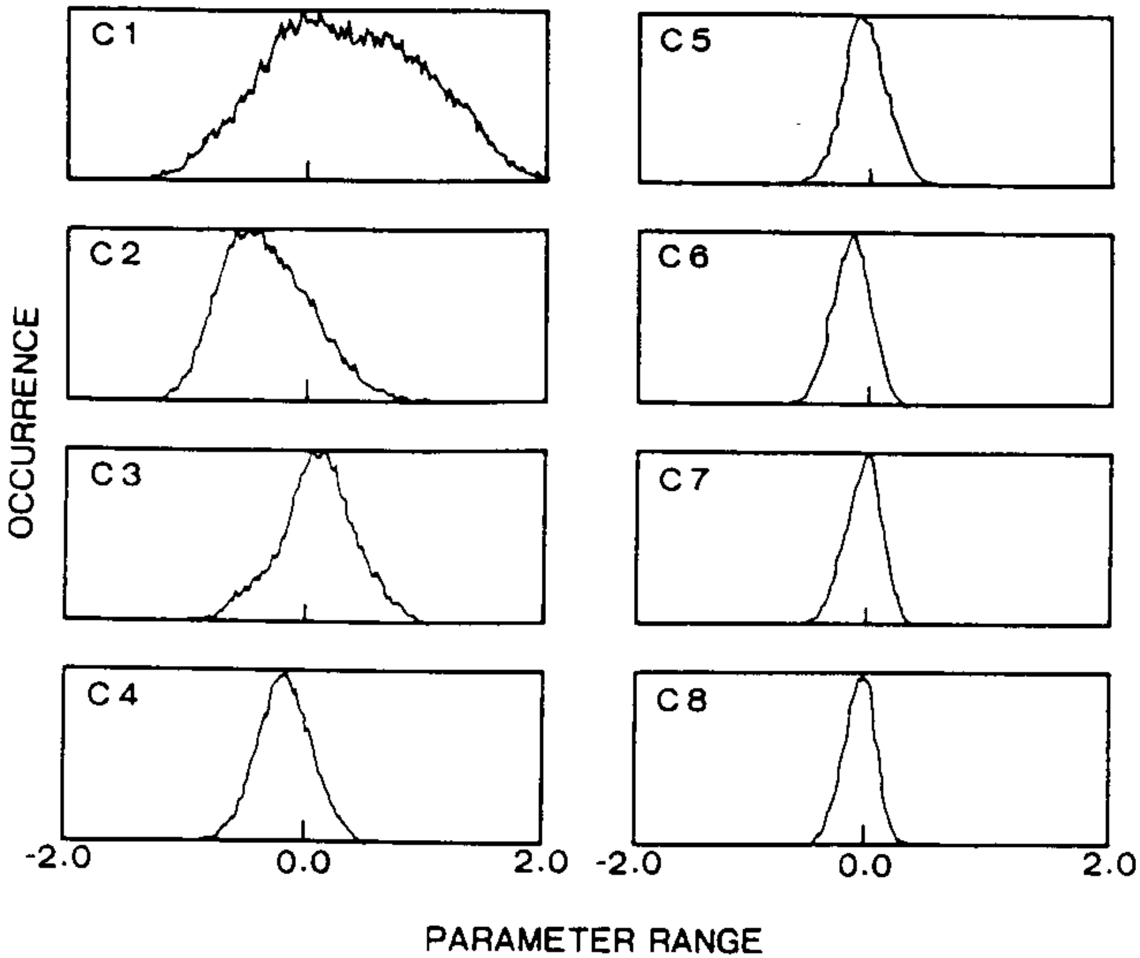
Many current speech recognition systems represent the speech signal as a set of cepstral coefficients, computed at a fixed frame rate. In addition, the time derivatives of the cepstral coefficients have also been used.





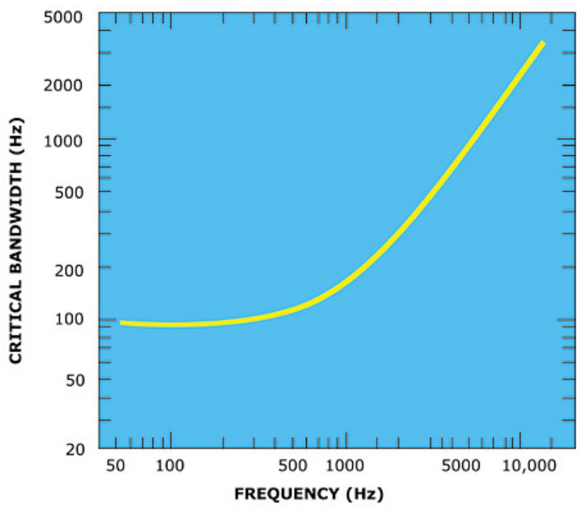
Statistical Properties of Cepstral Coefficients (Tohkura, 1987)

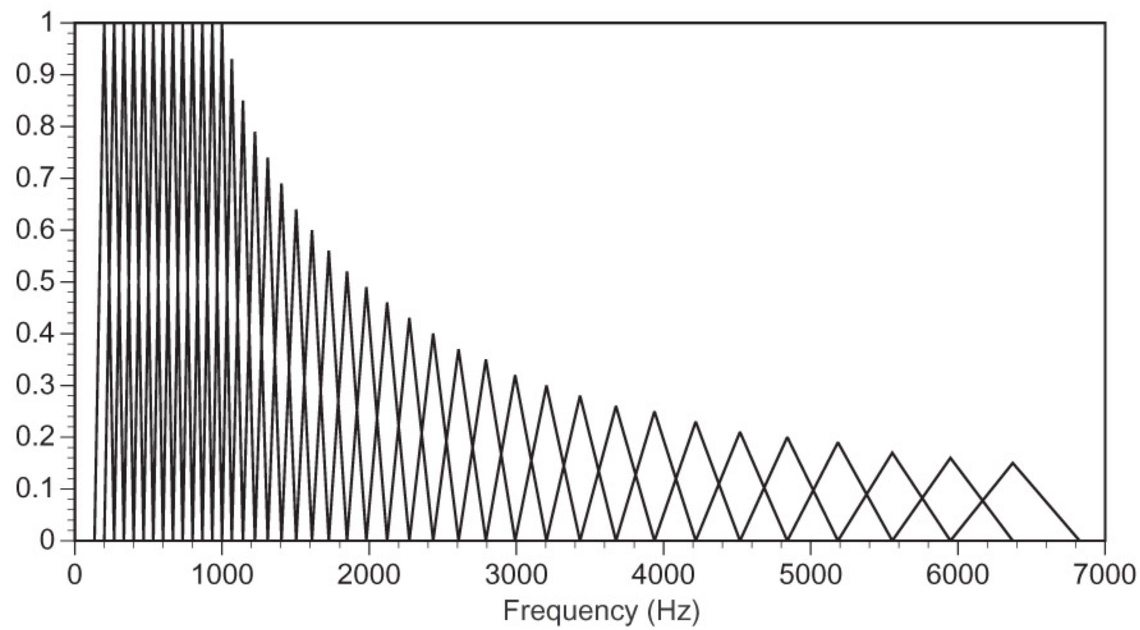
From a digit database (100 speakers) over dial-up telephone lines.



Mel-Frequency Cepstral Representation (Mermelstein & Davis, 1980)

Some recognition systems use Mel-scale cepstral coefficients to mimic auditory processing. (Mel frequency scale is linear up to 1000 Hz and logarithmic thereafter.) This is done by multiplying the magnitude (or log magnitude) of $S(e^{j\omega})$ with a set of filter weights as shown below:

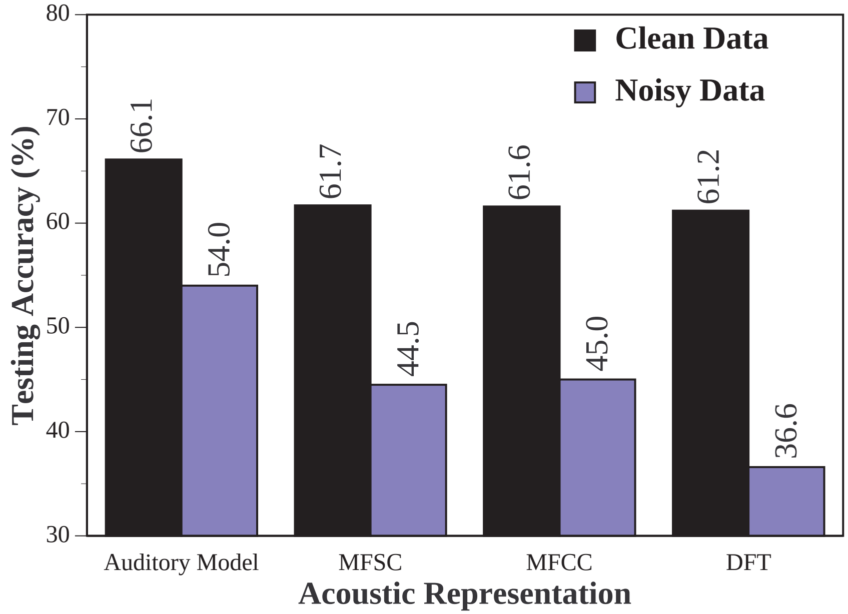




Signal Representation Comparisons

- Many researchers have compared cepstral representations with Fourier-, LPC-, and auditory-based representations.
- Cepstral representation typically out-performs Fourier- and LPC-based representations.

Example: Classification of 16 vowels using ANN (Meng, 1991)



- Performance of various signal representations cannot be compared without considering how the features will be used, i.e., the pattern classification techniques used. (Leung, et al., 1993).

Things to Ponder...

- Are there other spectral representations that we should consider (e.g., models of the human auditory system)?
- What about representing the speech signal in terms of phonetically motivated attributes (e.g., formants, durations, fundamental frequency contours)?
- How do we make use of these (sometimes heterogeneous) features for recognition (i.e., what are the appropriate methods for modelling them)?

References

1. Tohkura, Y., “A Weighted Cepstral Distance Measure for Speech Recognition," *IEEE Trans. ASSP*, Vol. ASSP-35, No. 10, 1414-1422, 1987.
2. Mermelstein, P. and Davis, S., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. ASSP*, Vol. ASSP-28, No. 4, 357-366, 1980.
3. Meng, H., *The Use of Distinctive Features for Automatic Speech Recognition*, SM Thesis, MIT EECS, 1991.
4. Leung, H., Chigier, B., and Glass, J., “A Comparative Study of Signal Represention and Classification Techniques for Speech Recognition," *Proc. ICASSP*, Vol. II, 680-683, 1993.