8. Cepstral Analysis

(most slides taken from MIT course by Glass and Zue)

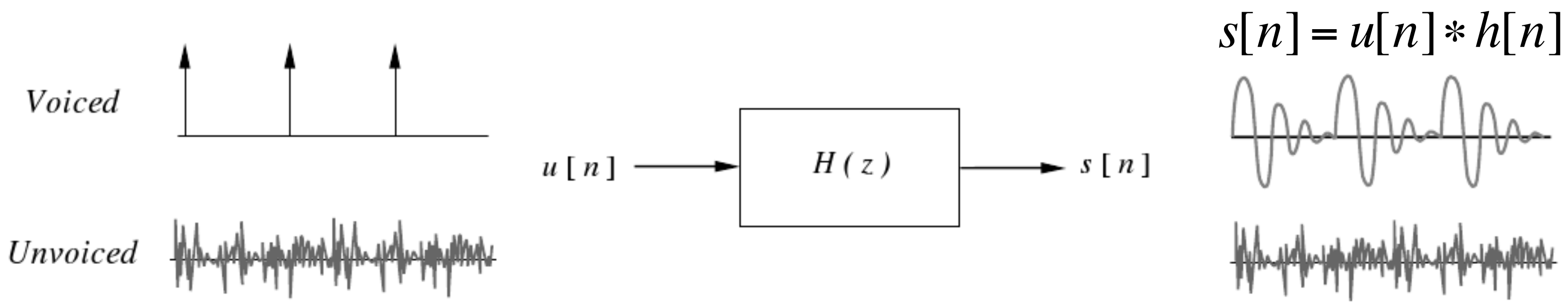
Homomorphic filtering

Homomorphic filtering/transformation is a nonlinear transformation $\hat{x}[n] = D(x[n])$ usually applied to image and speech processing used to convert a signal obtained from a convolution of two original signals into the sum of two signals.

$$x[n] = e[n] * h[n] \rightarrow \hat{x}[n] = \hat{e}[n] + \hat{h}[n]$$

In speech processing it can be applied to separate the filter from the excitation in the source-filter model

The *cepstrum* is one such homomorphic transformation that allows us to perform such separation.



It is an **alternative** option to linear prediction analysis seen before

Basics of cepstral analysis

Cepstral analysis is based on the observation that

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

by taking the log of X(z)

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

If the complex log is unique and the z transform is valid then, by applying Z^{-1}

$$\hat{x}[n]=\hat{x}_1[n]+\hat{x}_2[n]$$

the two convolved signals are now additive.

Basics of cepstral analysis (II)

Consider now that we restrict our signal x[n] to have poles and zeros only in the unit circle, i.e.:

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

then if $X(\omega)=X_1(\omega)X_2(\omega)$

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

This is the complex logarithm of X(w)

Definition of Cepstrum

The real cepstrum is defined as:

$$c_x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\left|X\left(e^{j\omega}\right)\right| e^{j\omega n} d\omega$$

Its magnitude is real and non-negative

And the complex cepstrum:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[X(e^{j\omega})] e^{j\omega n} \, d\omega$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} [\log\left|X(e^{j\omega})\right| + j \arg(X(e^{j\omega}))] e^{j\omega n} \, d\omega$$

Where arg() represents the phase. We call it complex because it uses the complex logarithm, not due to the sequence, which can also ne real. In fact, the complex cepstrum of a real sequence is also real

Definition of cesptrum(II)

It can be shown that the the real cepstrum is the even part of the complex cepstrum:

$$c_x[n] = \frac{\hat{x}[n] + \hat{x}^*[-n]}{2}$$

In speech processing we generally use the real cepstrum, which is obtained by applying an inverse Fourier Transform of the log-spectrum of the signal.

In fact, the name "cepstrum" comes from inverting the first syllable of the word "spectrum". Similarly, the variable "n" in $c_x[n]$ is called "quefrency", which is the inversion of "frequency"

Properties of cepstrums

From this we can derive the following general properties:

1) the complex cepstrum decays at least as fast as $1/|n|$

2) it has infinite duration, even if $x[n]$ has finite duration

3) it is real if $x[n]$ is real (poles and zeros are in complex conjugate pairs)

NOTE: from 2) and 3) we see why usually a finite number of cepstrums is used in speech processing (12-20 is sufficient), as very high order cepstrums have very small values.

An example

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

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

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

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

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

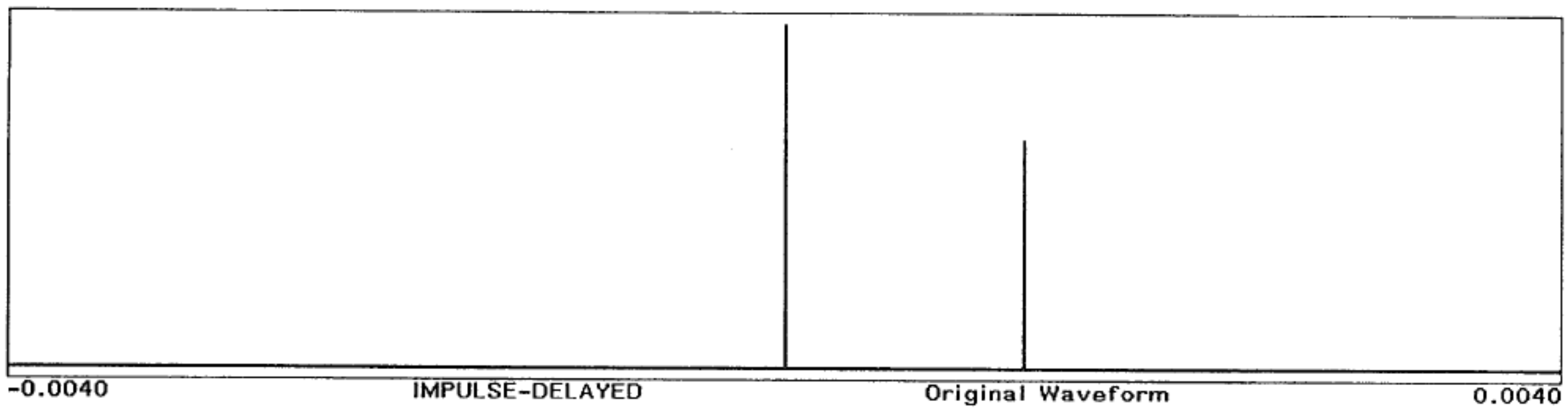
(Using Taylor series expansion and several tricks)

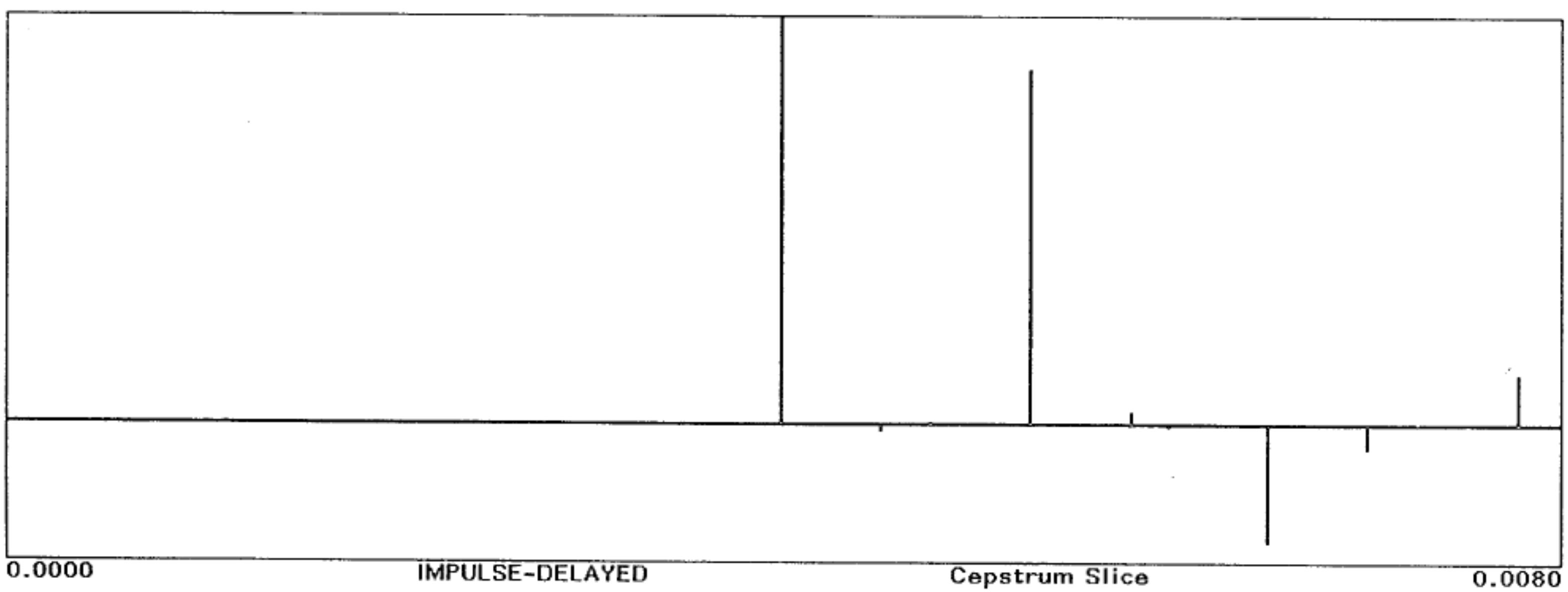
$$\hat{P}(z) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\alpha^n}{n} (z^{N})^{-n}$$

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

Given the “r” in the denominator, it is an infinite train of deltas that converges to 0

(...)


-0.0040
IMPULSE-DELAYED
Original Waveform
0.0040


0.0000
IMPULSE-DELAYED
Cepstrum Slice
0.0080

Computational considerations: using DFT

In digital signals we replace the Fourier Transform by the Discrete Fourier Transform

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

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

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

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

Aliasing by repetition of the cepstrums with period N

- Likewise:

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

where,

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

- To minimize aliasing, N must be large.

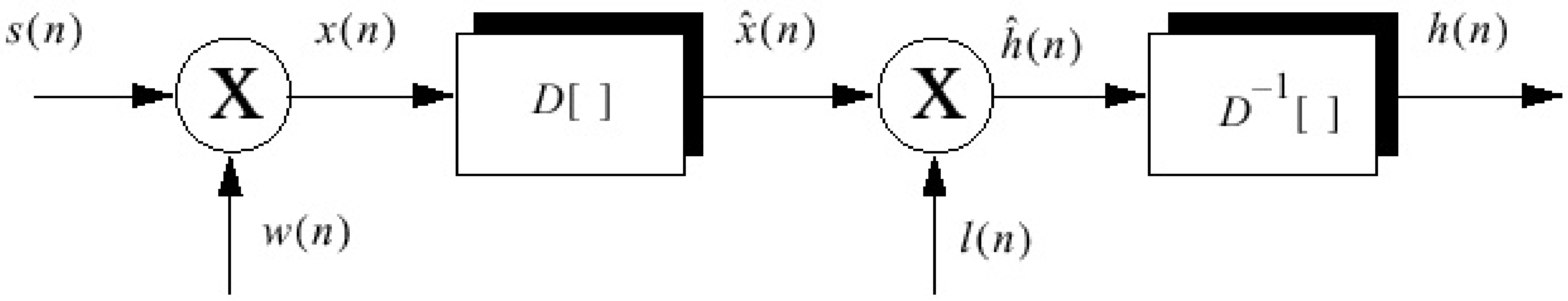
When N> the number of used cepstrums we do not have a problem (which is usually the case)

Cepstral analysis of speech

As pointed out at the beginning, we would like to separate the excitation from the vocal tract filter h(n) by using a homomorphic transformation.

We can do so easily as the filter parameters usually reside in the lower quefrencies, while the excitation parameters have higher quefrencies

Consider the problem of recovering a filter's response from a periodic signal (such as a voiced excitation):



The filter response can be recovered if we can separate the output of the homomorphic transformation using a simple filter:

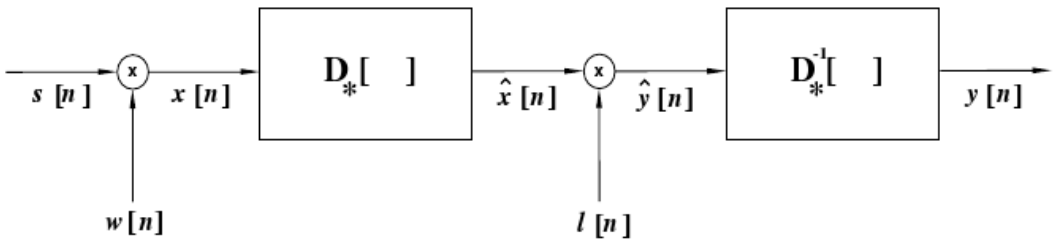
$$l(n) = \begin{cases} 1 & |n| < N \\ 0 & |n| \geq N \end{cases}$$

Cepstral analysis of speech

- For voiced speech:

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

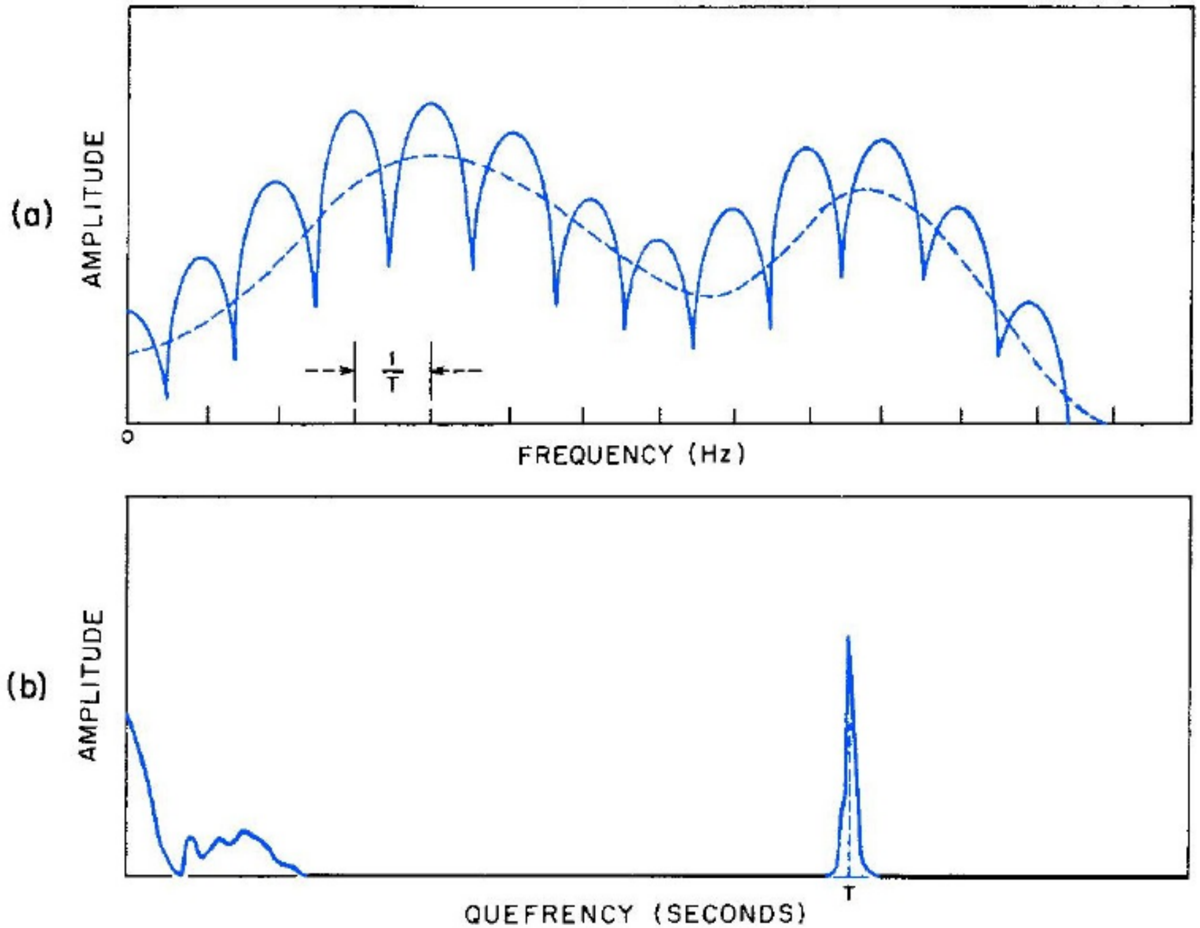
- For unvoiced speech: $s[n] = w[n] * v[n] * r[n] = w[n] * h_u[n]$.
- Contributions to the cepstrum due to periodic excitation will occur at integer multiples of the fundamental period.
- Contributions due to the glottal waveform (for voiced speech), vocal tract, and radiation will be concentrated in 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.

Cepstrum of a generic voiced signal

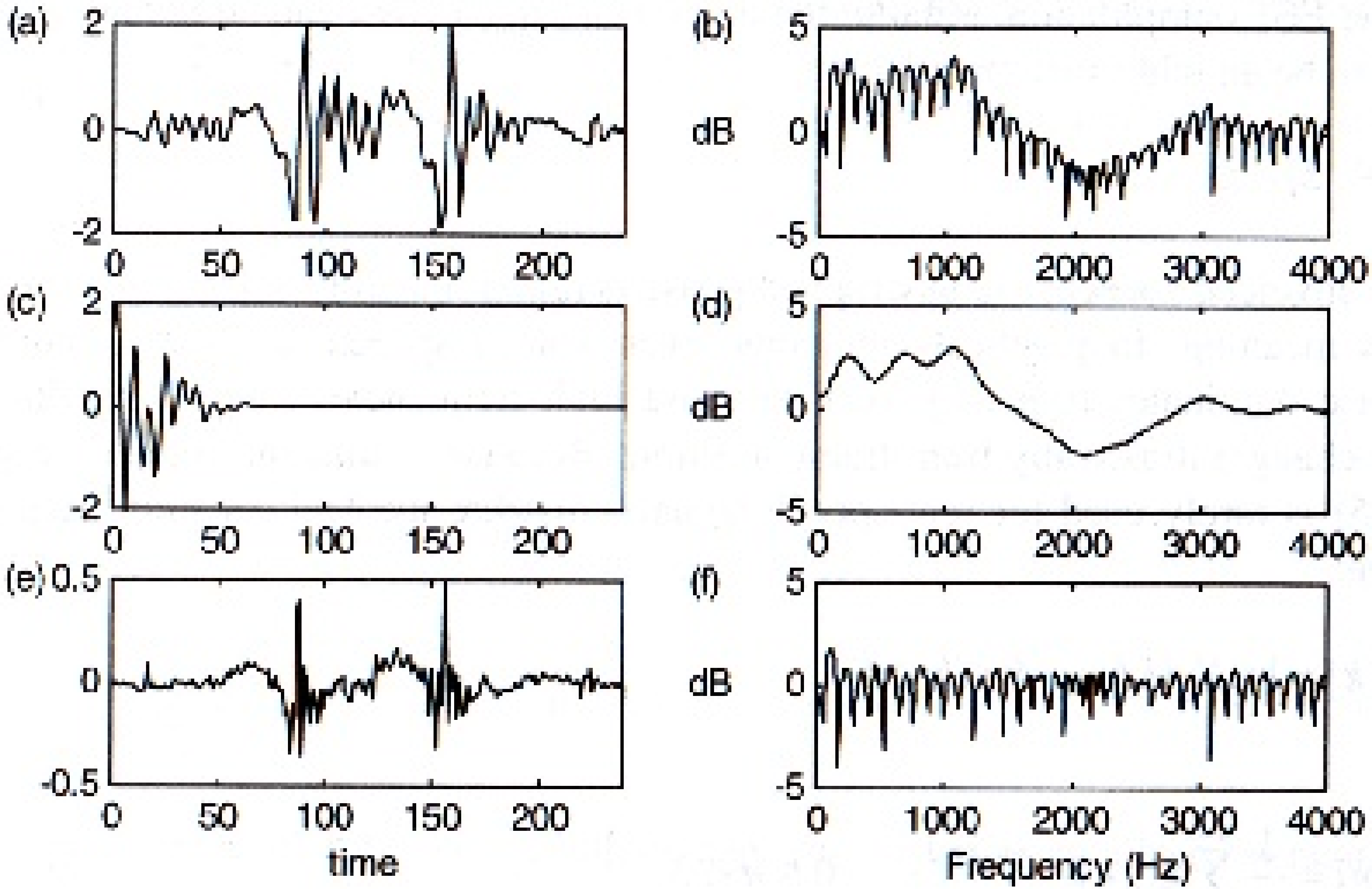


magnitude spectrum

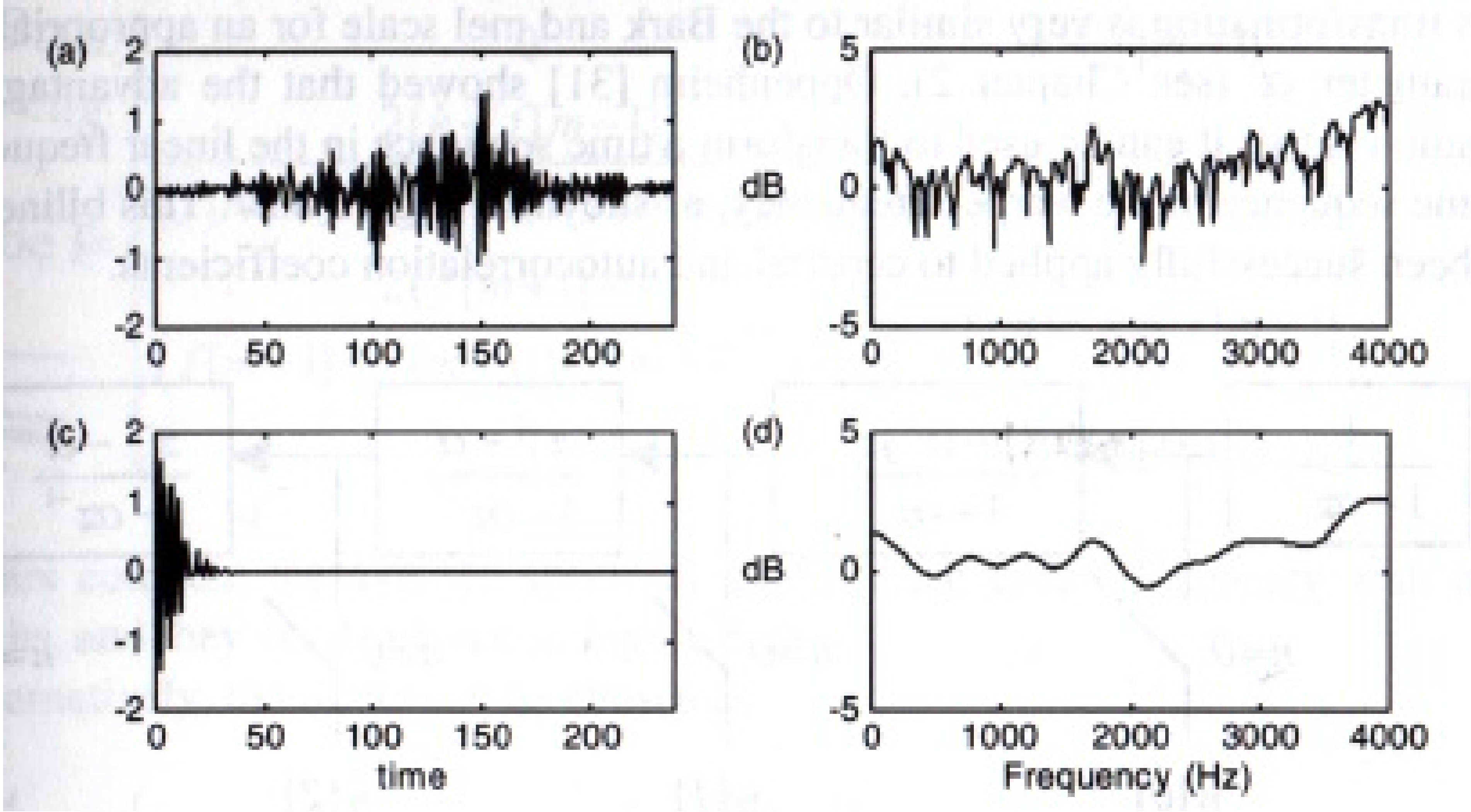
cepstrum

- Contributions to the cepstrum due to periodic excitation will occur at integer multiples of the fundamental period. **NOTE that for children and high-pitch women we might have a problem**
- Contributions due to parameters usually modeled by the filter will concentrate in the low quefrency region and will decay quickly with n

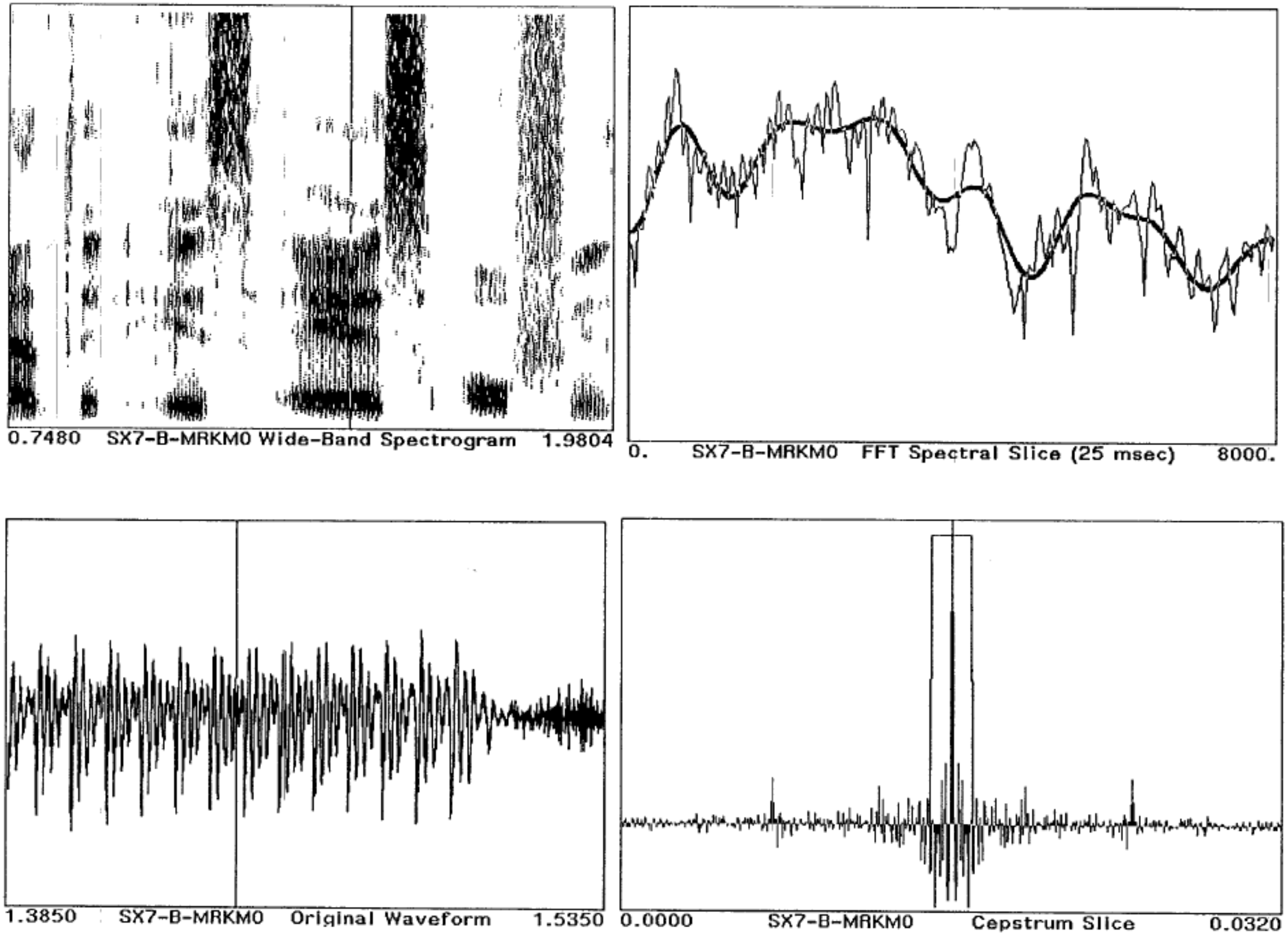
Cepstral analysis of speech (voiced signals)



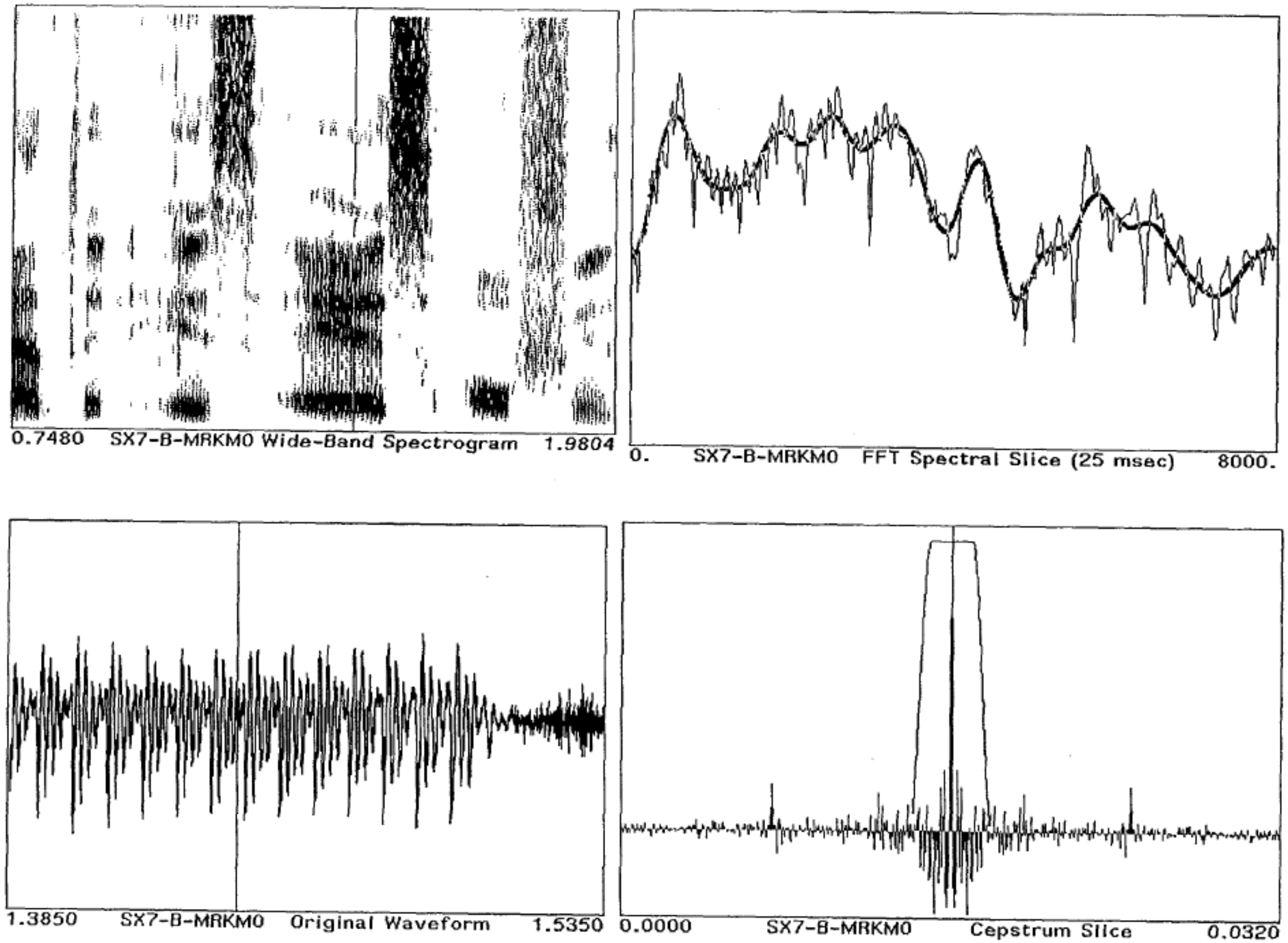
Cepstral analysis of speech (unvoiced signals)



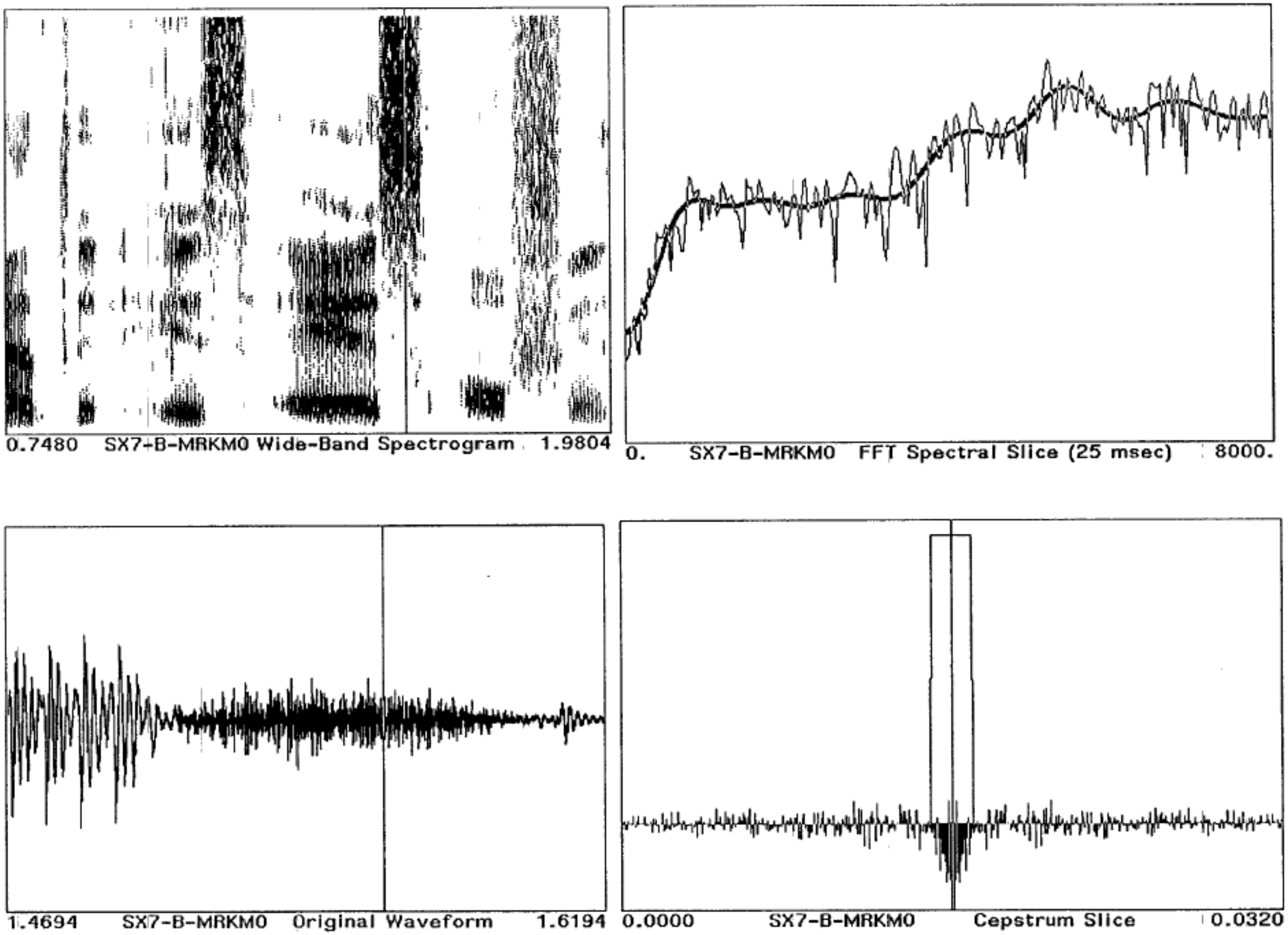
Cepstral analysis of vowel (rectangular window)



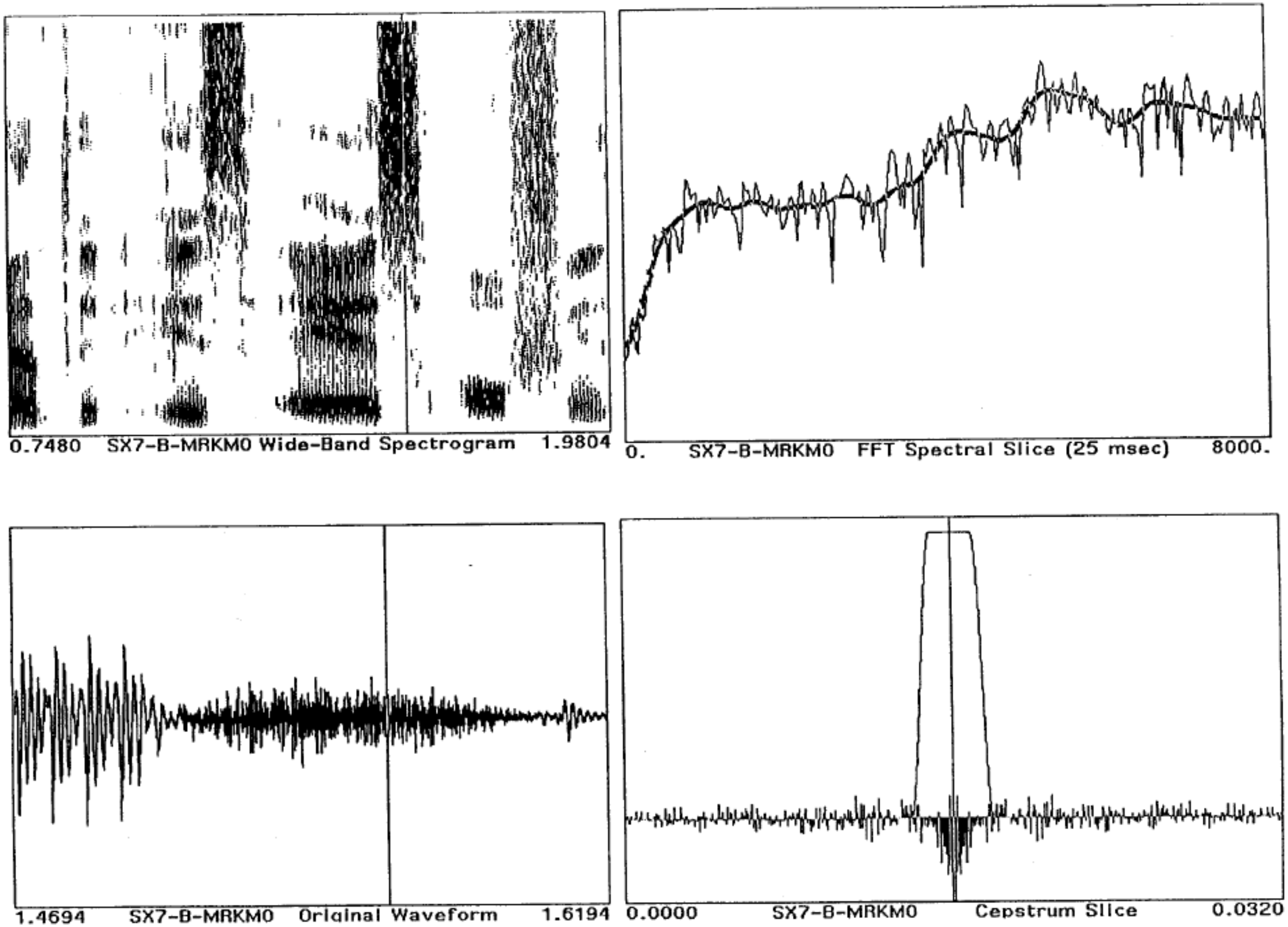
Cepstral analysis of vowel (tapering window)



Cepstral analysis of fricative (rectangular window)

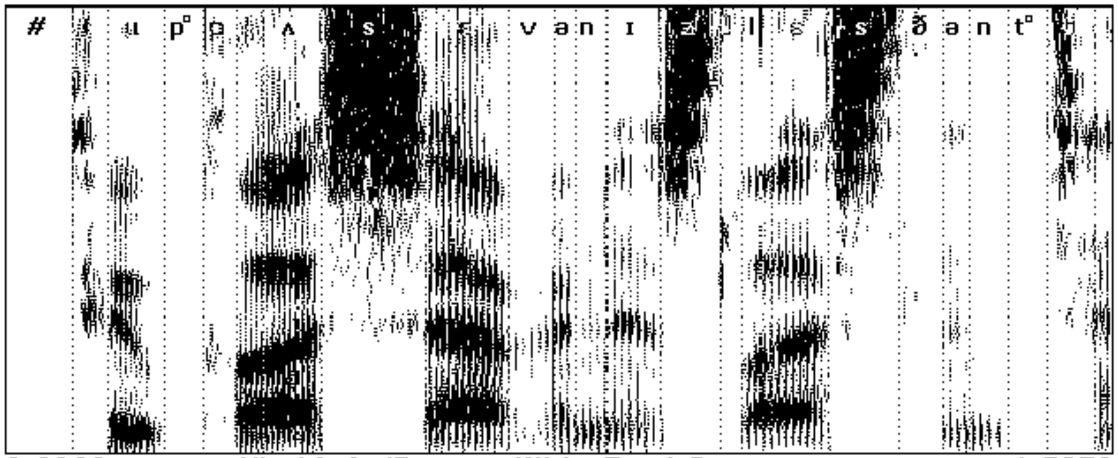


Cepstral analysis of fricative (tapering window)

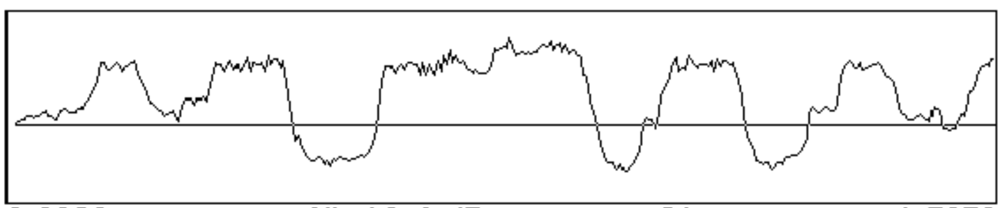


Use in Speech Recognition

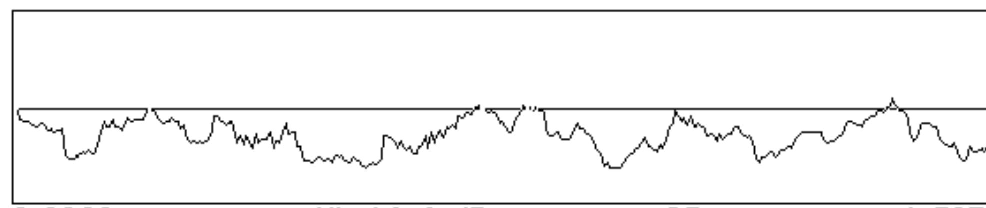
Many current speech recognition systems represent the speech signal as a set of cepstral coefficients, computed at a fixed frame rate. In addition, the time derivatives of the cepstral coefficients have also been used.



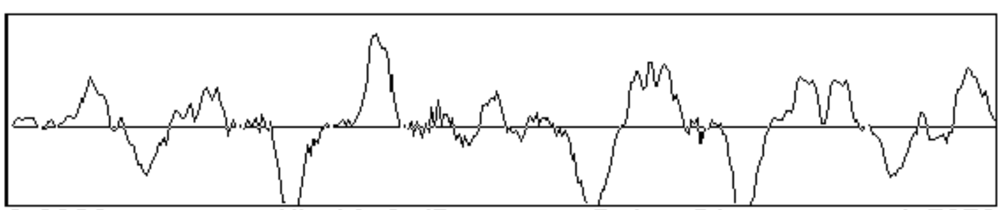
0.0000 HL-14-4-JP Wide-Band Spectrogram 1.5979



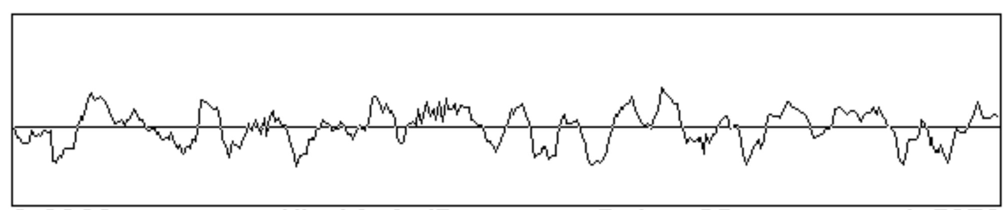
0.0000 HL-14-4-JP C1 1.5979



0.0000 HL-14-4-JP C2 1.5979



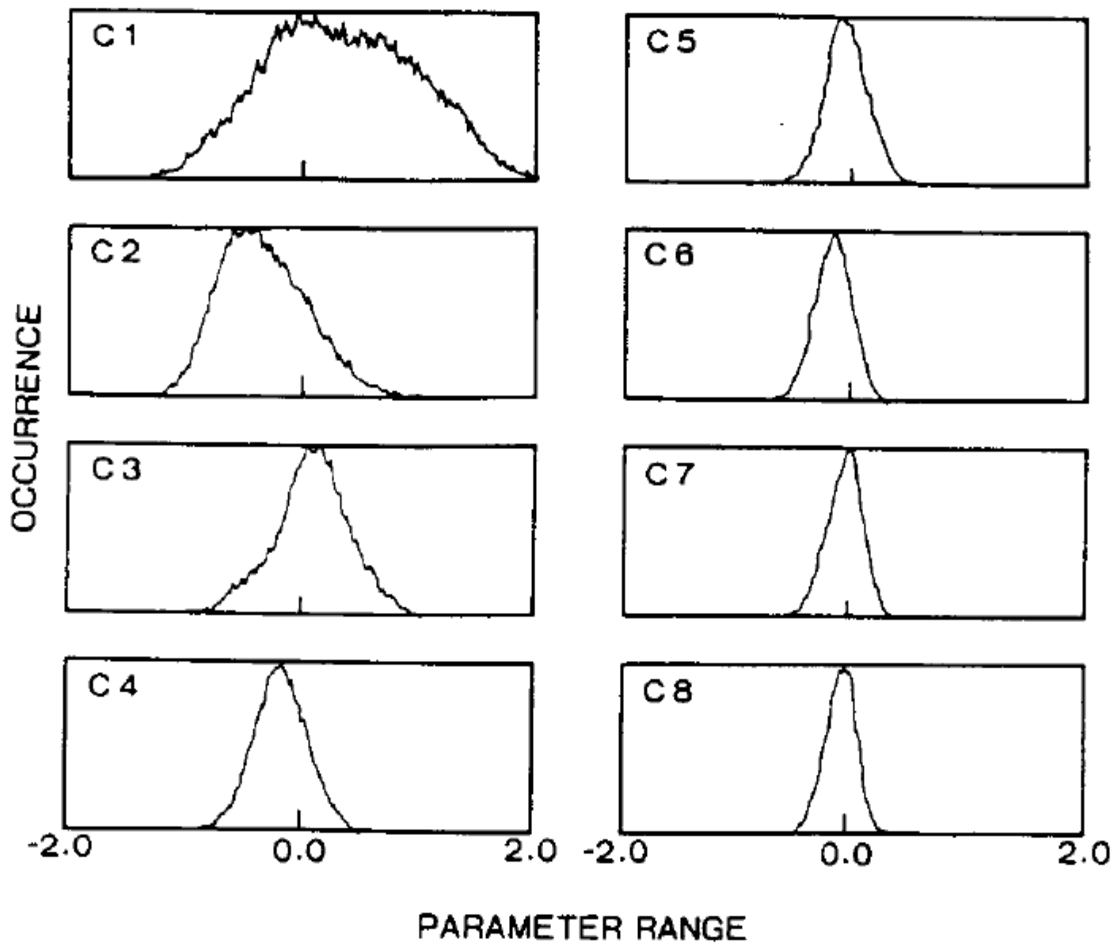
0.0000 HL-14-4-JP Delta-C1 1.5979



0.0000 HL-14-4-JP Delta-C2 1.5979

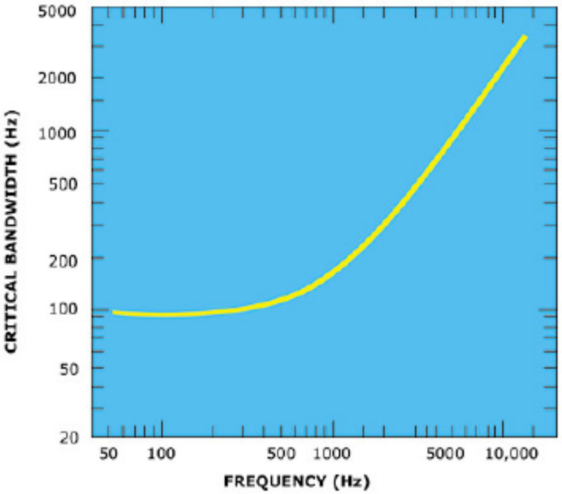
Statistical properties of cepstral coefficients

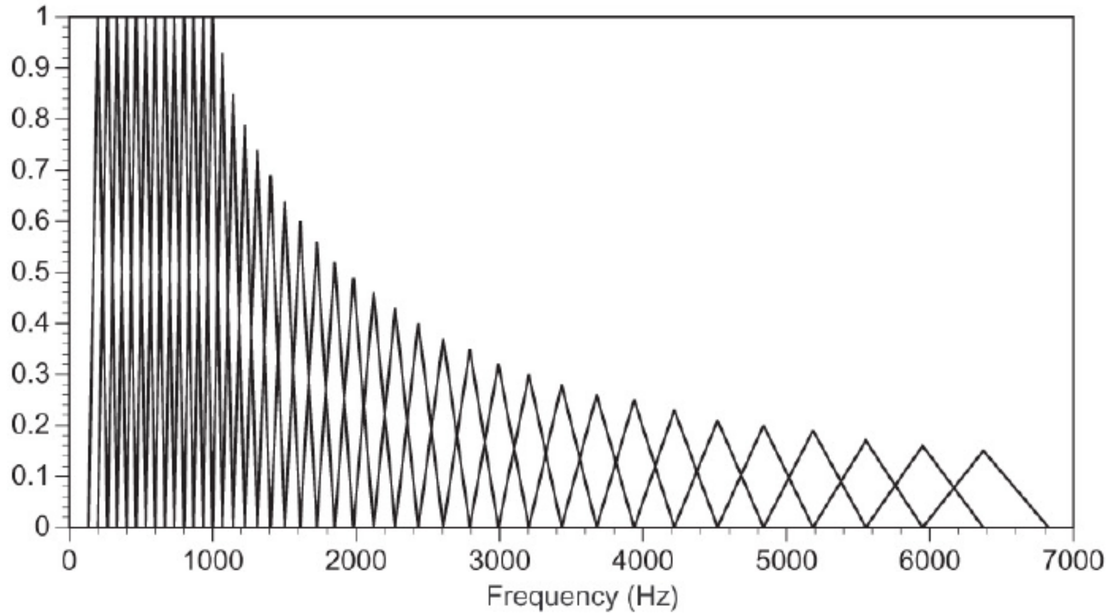
From a digit database (100 speakers) over dial-up telephone lines.



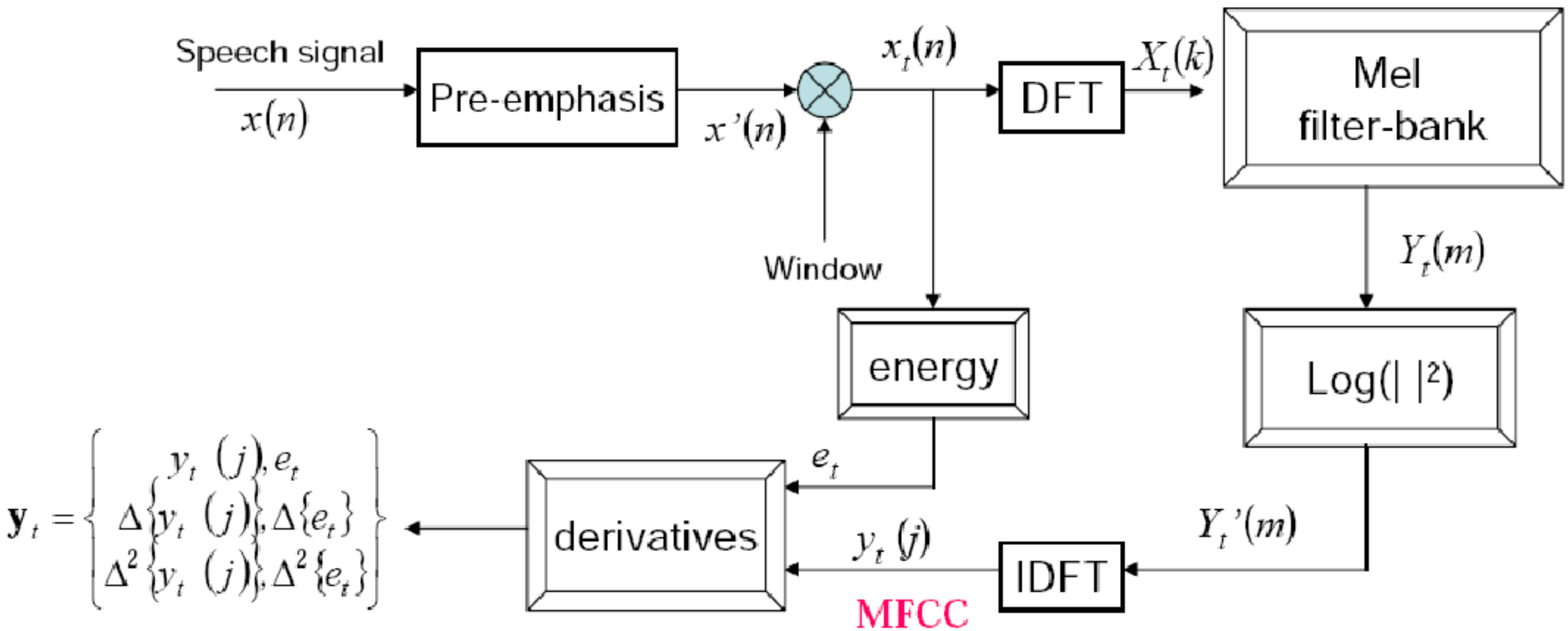
Mel-frequency cepstral representation

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:





MFCC computation diagram



Mel-filter bank processing