University of California
Berkeley

College of Engineering
Department of Electrical Engineering
and Computer Sciences

Professors : N.Morgan / B.Gold
EE225D Spring,1999

Pitch Detection & Vocoders

Lecture 24

Major Question

How to make a “Perfect” Vocoder? (Can it be done?)

What limitations are encountered for low bit rate representation?

Today’s Topic

Traditional 2400bps systems [or at least in that range] and

Pitch &Voicing detection.

NEXT

Very low rate systems [600bps]

NEXT

Higher quality more rubust systems at 5-30Kbps

Difficulties Encountered in Pitch Detection

* Purpose of pitch detection is to automatically obtain a result that is in agreement with a psychoacoustic result for the same stimulus. And also to make a vocoder sound natural.

* Early researchers preferred to use the term “fundamental frequency estimator” but we saw in Chapter 16 that pitch would be “perceived” even if the stimulus was a harmonic for that frequency. (example - shift of virtual pitch)

* What we’re really after is the NATURE and quantitative description of the excitation function.

*** This means:**

1. Detection of the time when the vocal cords are vibrating in a [perhaps rapidly varying] quasi-periodic way and tracking the period.
2. Representation of the friction noise caused by a vocal tract constriction.
3. Representation of the transient excitation during plosive.
4. Representation of the noise for a whispered vowel
5. Representations of various combitnations of all the above.

Expamples of speech waveforms that makes the above analysis difficult.

* Dynamic range of quasi-periodic vocal cord vibrations

as low as 50Hz for some adults
as high as 800Hz for children —16:1 range

* Rapid variation in glottal period

* Sudden change in vocal treat shape [e.g. nasal]

* Transition from unvoiced to voiced.

* Environmental transmission problems.

With linear assumptions, the speech wave can be represented as the

Convolution of an excitation function with a vocal tract filter function.

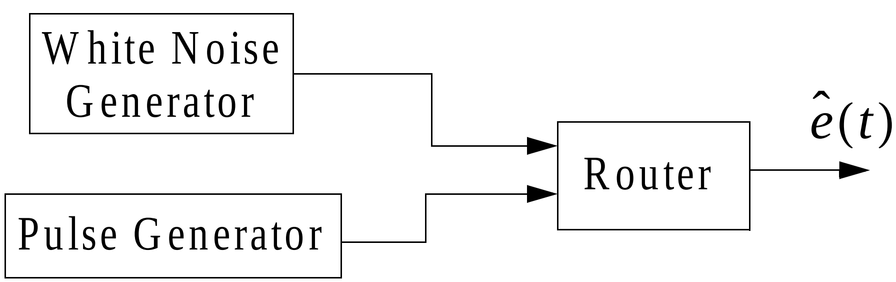
In spectral terms : $S(\omega) = E(\omega)H(\omega)$

* In a channel vocoder analyzer, measurements $S(\omega)$, $H(\omega)$ and $E(\omega)$ are NOT computed separately.

* In the channel vocoder synthesizer, the spectrum is obtained as follows.

$$S(\omega) = E(\omega)H(\omega)$$

* If $\hat{E}(\omega)$ is a FLAT SPECTRUM, $\hat{S}(\omega) \cong S(\omega)$, although the phases may be different.



* The situation is complicated by the fact that $E(\omega)$ and $H(\omega)$ are time-varying.

In LPC, we start off with $S(n) = Ex(\omega) \cdot H(\omega)$

$$s(\omega) = ex(n) \cdot h(n)$$

change of nomenclature $ex(n)$ is the model of the speech excitation signal.

LPC derives an all-pole model $\hat{\mathrm{H}}(\omega) \rightarrow \hat{h}(n)$

It would be nice if $\hat{\mathrm{H}}(\omega)$ was really a good representation *of* $H(\omega)$, the real vocal tract function.

Speech can be perfectly reconstructed by carvolving $\hat{h}(n)$ with the error signal $e(n)$.

$$s(n) = e(n) \times \hat{h}(n)$$

$$S(\omega) = E(\omega) \cdot \hat{H}(\omega) = Ex(\omega) \cdot H(\omega)$$

if $\hat{H}(\omega)$ differs greatly from $H(\omega)$, $E(\omega)$ will compensate by being correspondingly different than $Ex(\omega)$.

* Many LPC systems [multi pulse, celp, etc.] derive their power by searching for an error signal that compensates for $\hat{H}(\omega)$.

Homomorphic analysis has the hypothesis that source - filter separation is manifested as spectrum envelope - spectral fine structure separation. The model also assumes that these are multiplied in the spectral domain, so that taking the log turns the product into a sum. Finally, the model assumes that the two are separable with liftering. Given this separation, the excitation function and the vocal tract filter function can be represented and then Convolved to give the synthesized speech.

In order to achieve low transmission rates (e.g. 2400bps), all systems relay on the excitation model consisting of a noise source and a variable period pulse source

- Both sources are reasonable approximations to flat spectra and take few bits to transmit.

buzz-hiss switch - 1bit every 10msec. → 100bps
pitch tracker - 6bits every 10msec. → 600bps

Major Motivation for Dorry Research on Vocoders : Past, Present, Future.

Past - Secrecy - WWII - Data rates were limited. 2400bps became a standard.

Nearly all funding came from DOD to try to improve quality at 2400 bps.

Present - Modems are much better. As cellular phones proliferate,

date rate limitations still apply but 2400bps is no longer the sole criterian.

Main direction is still quality (robustness) - bit rate tradeoff.

Future - Greater robustness - efficient storage of

speech (and music) - coding -recognition tie-in.

Two Sides of the Coin

1

Basic Models for Analysis & Synthesis

- channel
- LPC
- Homomorphic

2

Wave form methods

PCM,APCM,ADPCM

Some predictive ability

Complete Channel Vocoder

Remember basic assumption for all vocoders.

* Synthetic speech is the convolution of an excitation function and a vocal tract filter function.

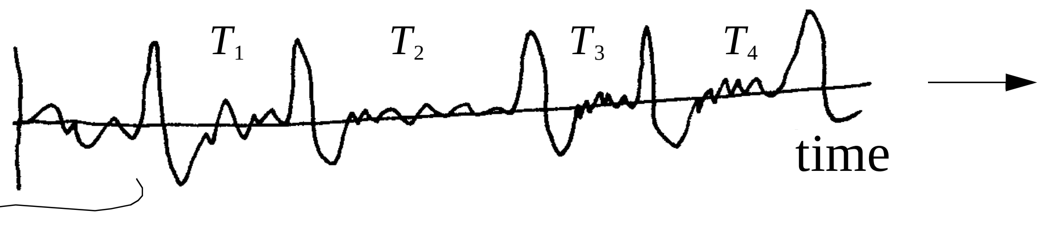
* Assumption : Synthesizer is a Time variable Linear System

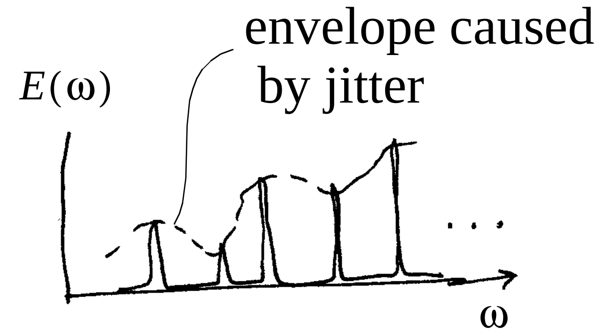
If this assumption was wrong and excitation and Vocal Tract Interacted in some Non-Linear Way, problem of implementing a “transparent” system probably becames intractable.

Working Hypothesis for 2400bps (and lower) systems.

Excitation is either buss [variable pulse generator] or hiss [white noise generator]

Consider the spectrum of a jittered pulse train.





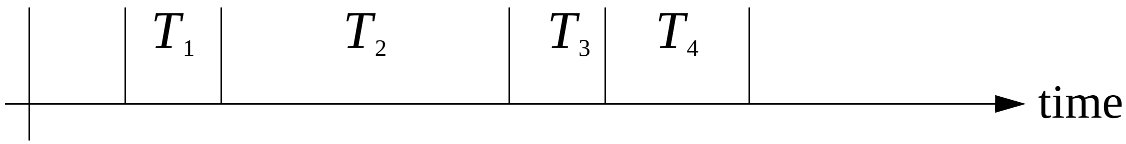
Most of the time, pitch dose NOT behave this badly.

Spectral distortion introduced by pitch jitter.

Now, assume that you have built a great pitch detector that tracks perfectly and records T_1, T_2, T_3, etc.

Now, this information is transmitted and the buzz generator at the synthesizer is forced to produce pulses based on the above measurements.

$S(\omega)$ is the product of the above $E(\omega)$ and $H(\omega)$.



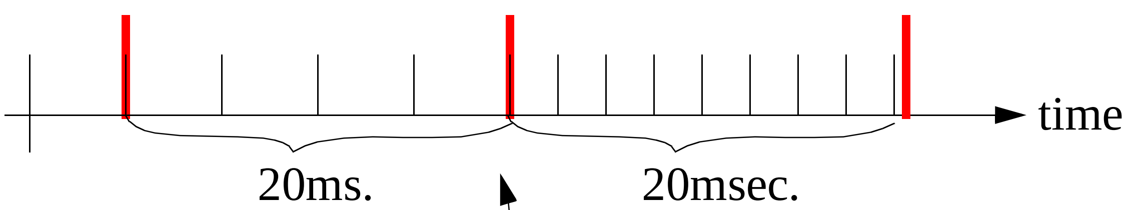
$$S(\omega) = E(\omega) \cdot H(\omega)$$

at the synthesizer

$$e(t) \rightarrow \boxed{s(t)} \rightarrow \hat{S}(\omega) = E^2(\omega) \cdot H(\omega)$$

Spectral distortion introduced by pitch jitter.

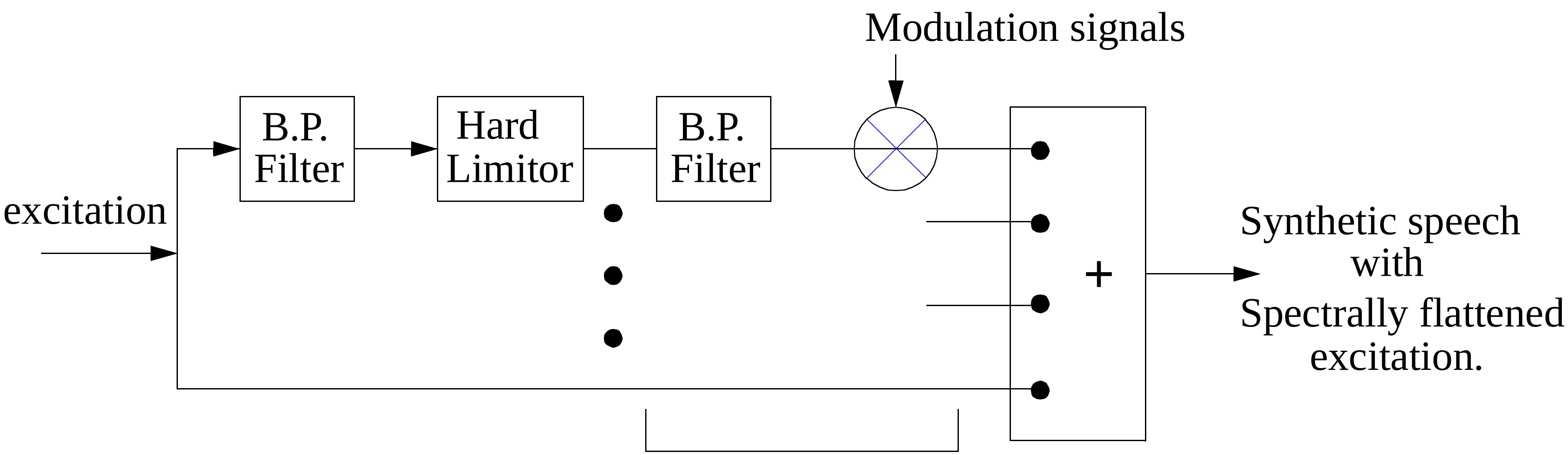
In real life, lets assume that analysis takes place every 20msec.. Analyzer generates a single pitch number, so at synthesizer, for a period of 20msec.



actual excitation during voicing. ~ not as bad as

Spectral Flattering

Turn the excitation signal into a white signal or white noise.



Model of $S(\omega) = E(\omega) \cdot H(\omega)$

Major Question

Does all-pole synthesizer model the Vocal tract envelope function

or the complete speech envelope function?

- if the former is true, excitation should NOT be spectrally flattered.
- if the latter is true, spectral flattering may help.

* Joe Tierrey and I did an informal experiment to determine perceived quality.

The result was ambiguous.

* In general, existing LPC systems (low rate) do NOT use spectral flattering

It may depend on the ORDER of the pedictor & synthesizer.

a 10th order predictor corresponds to five “formants”.

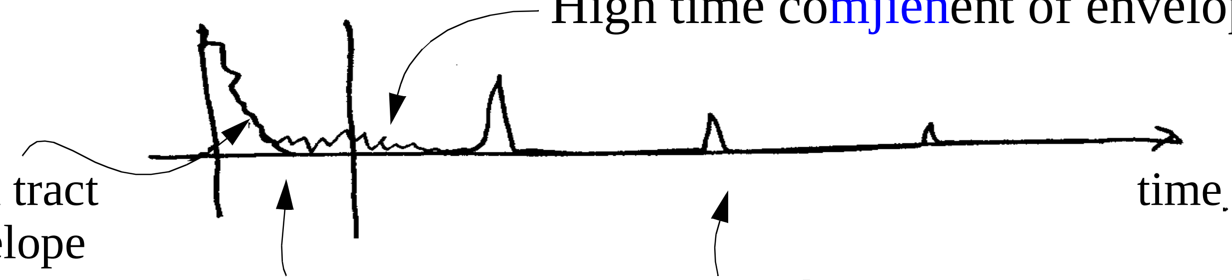
Homomorphic Vocoder

* Excitation is modelled in the same way as for channel vocoders &LPC.

Spectral flattering of the exciation signal has never [I think] been tried but it should work (in the same ballpark as channel &LPC).

Point C is Cepstrum.

High time comjienent of envelope



time

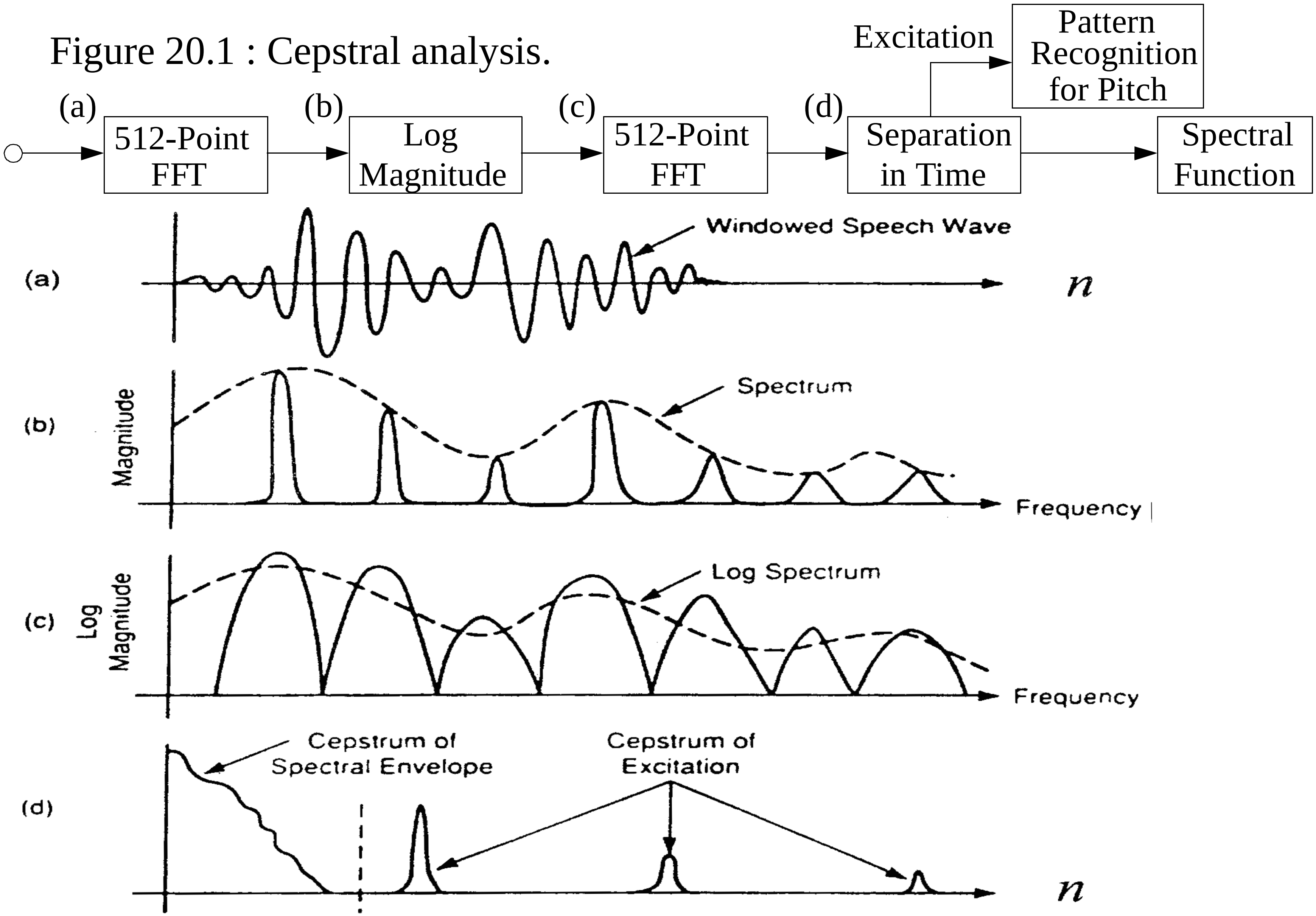
Vocal tract Envelope or Speech Spectrum Envelope

Low time cepstrum

High time cepstrum

Note - if excitation and envelope could be completely separated, synthesis ought to be perfect.

Figure 20.1 : Cepstral analysis.



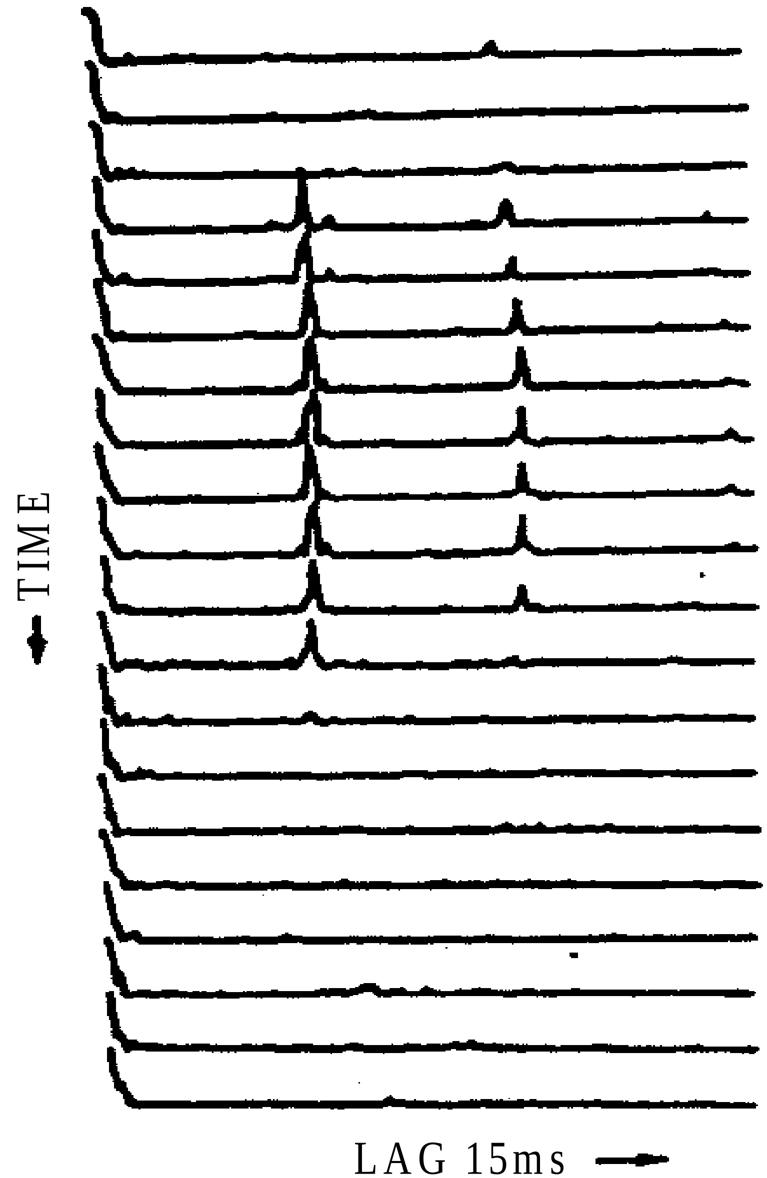
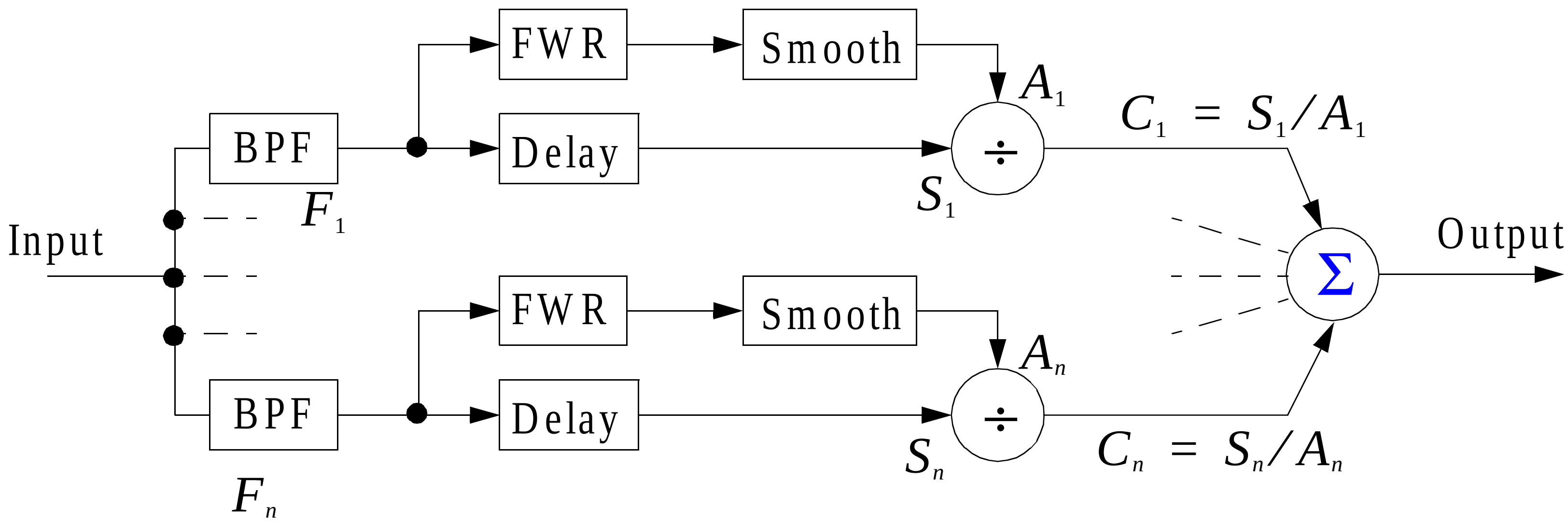


Figure 30.8 : Autocorrelation Function of Spectrally Flattened Speech. Successive 30ms sections with 15ms overlap.



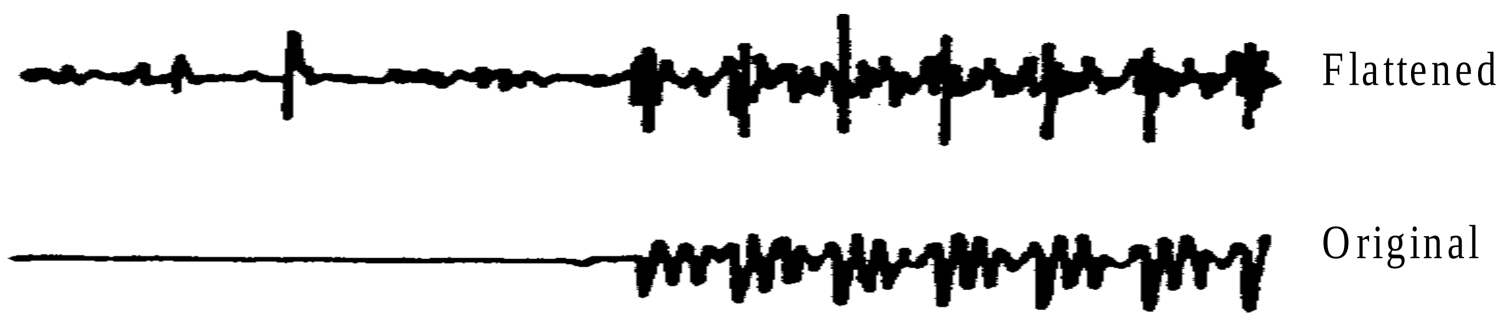


Figure 30.7 : Spectral Flattening and its Effect on the Speech Signal.

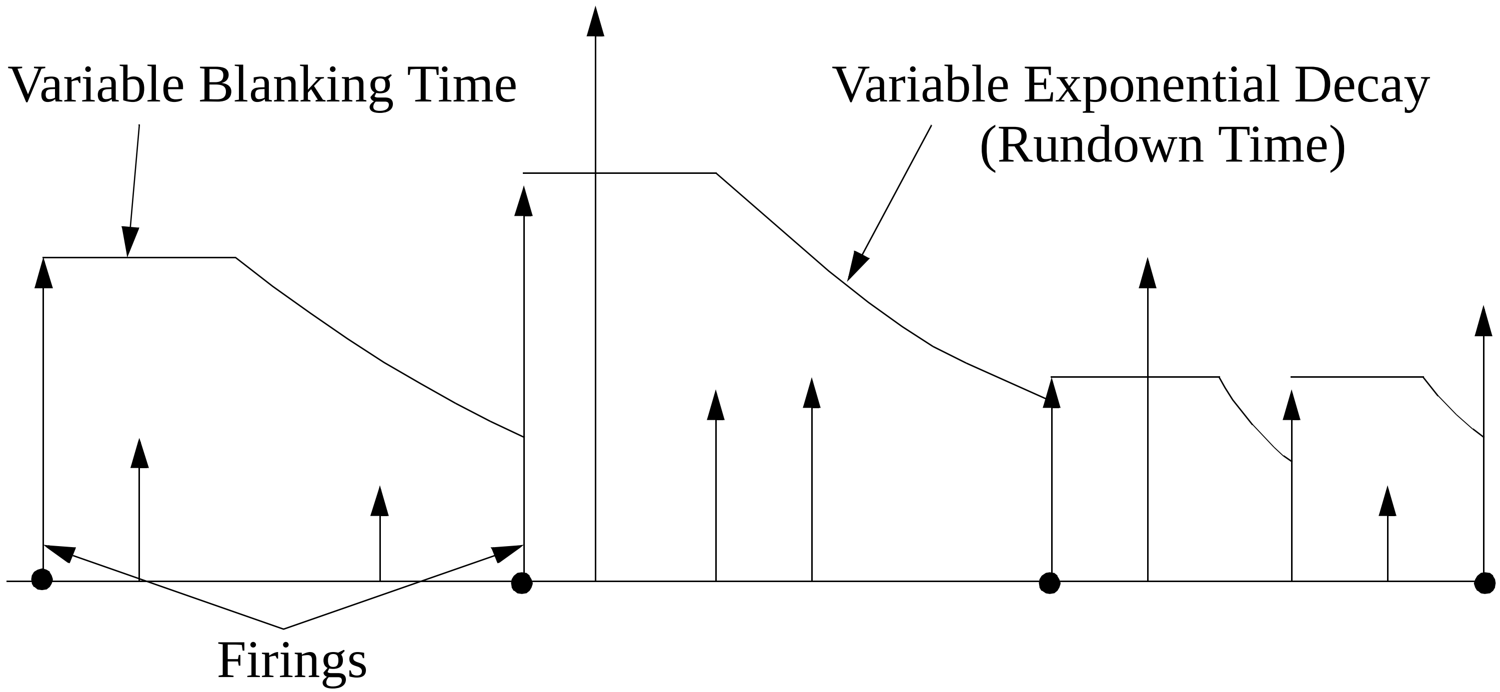


Figure 30.3 Extraction of the Period

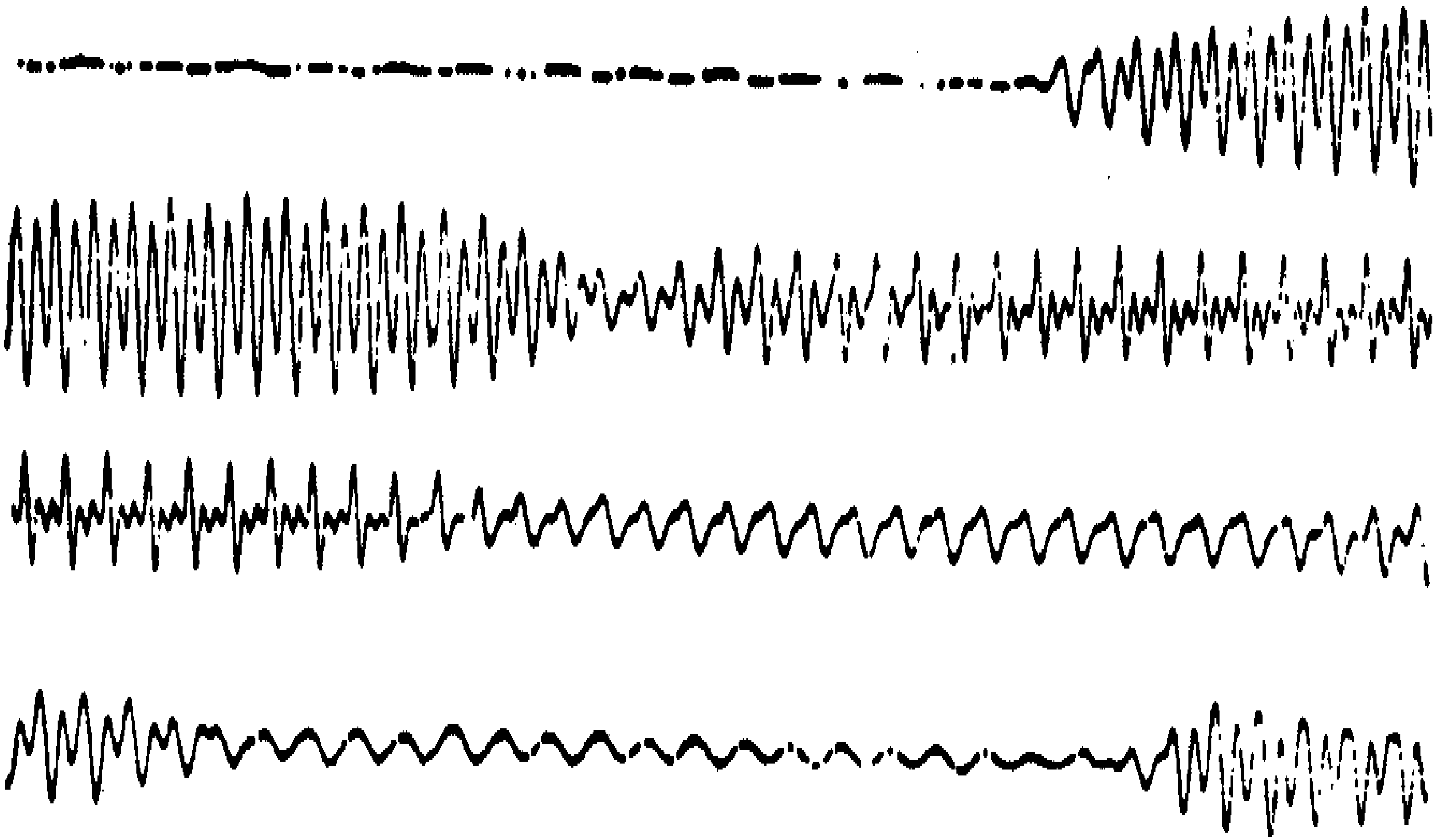
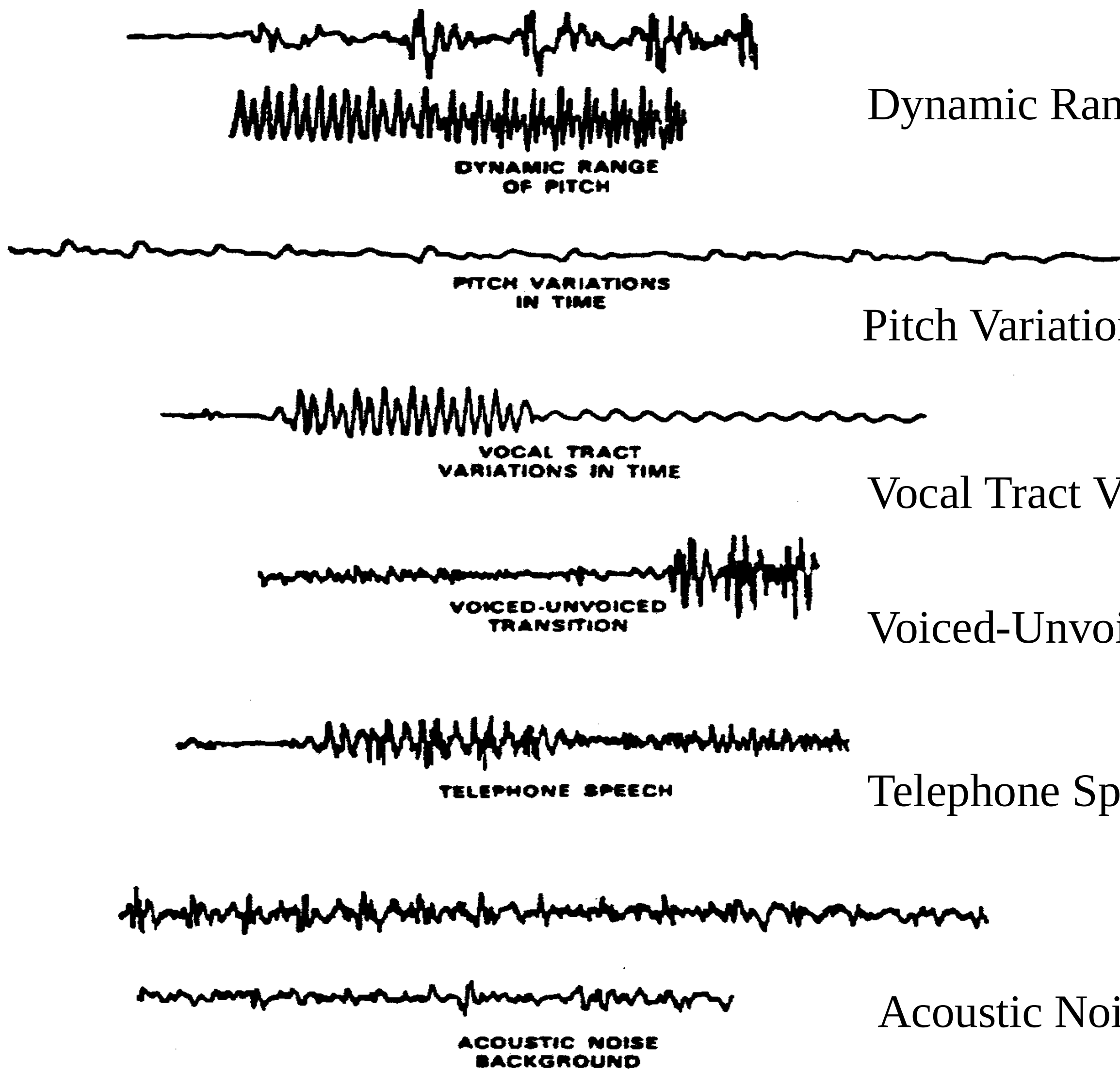


Figure 30.6 : Low-Pass filtered speech signal.



Dynamic Range of Pitch

Pitch Variations in Time

Vocal Tract Variations in Time

Voiced-Unvoiced Transition

Telephone Speech

Acoustic Noise Background

Figure 30.4 : Six Examples of Difficulties in Pitch Detection.

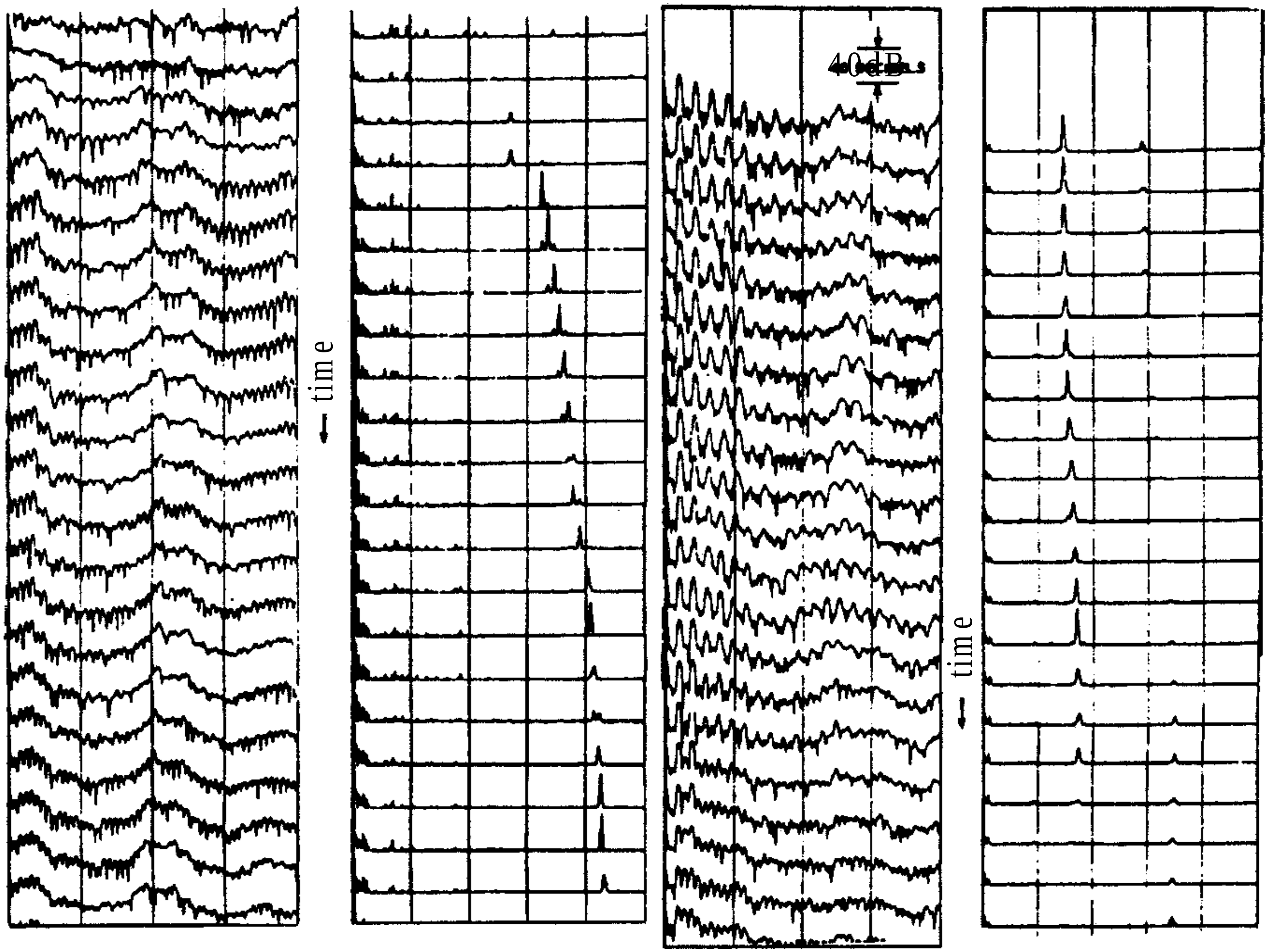


Figure 30.10 : Cepstral Analysis for Pitch Detection.

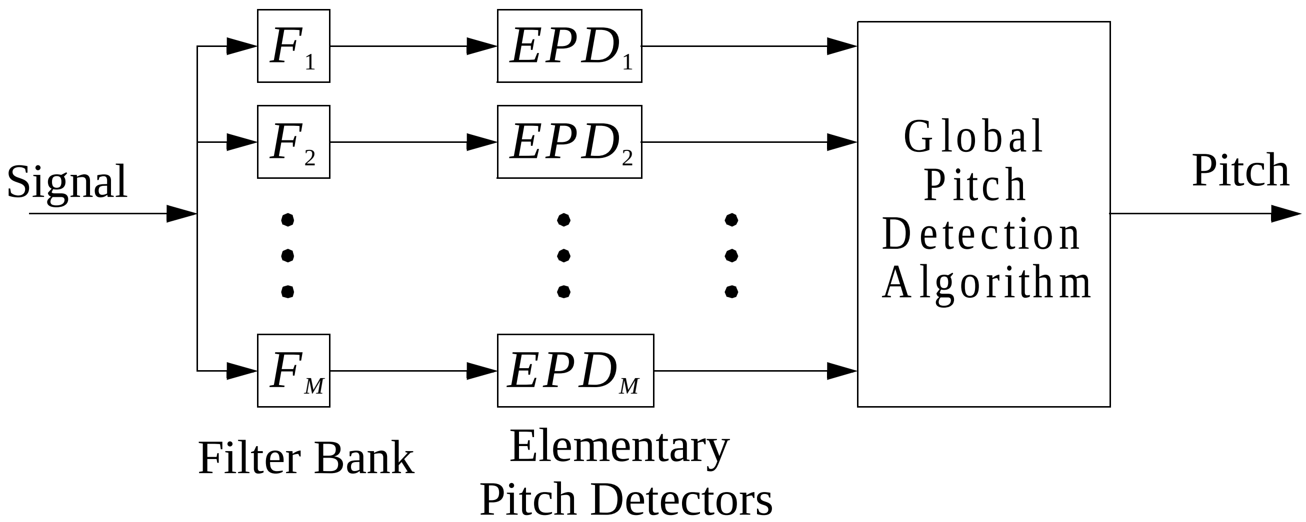


Figure 16.9 : Block Diagram of the Periodicity Model.

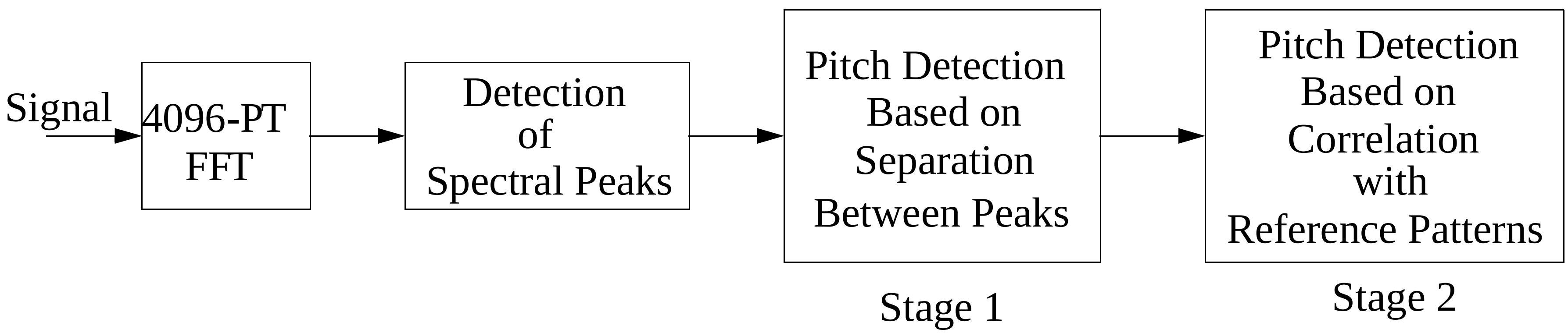
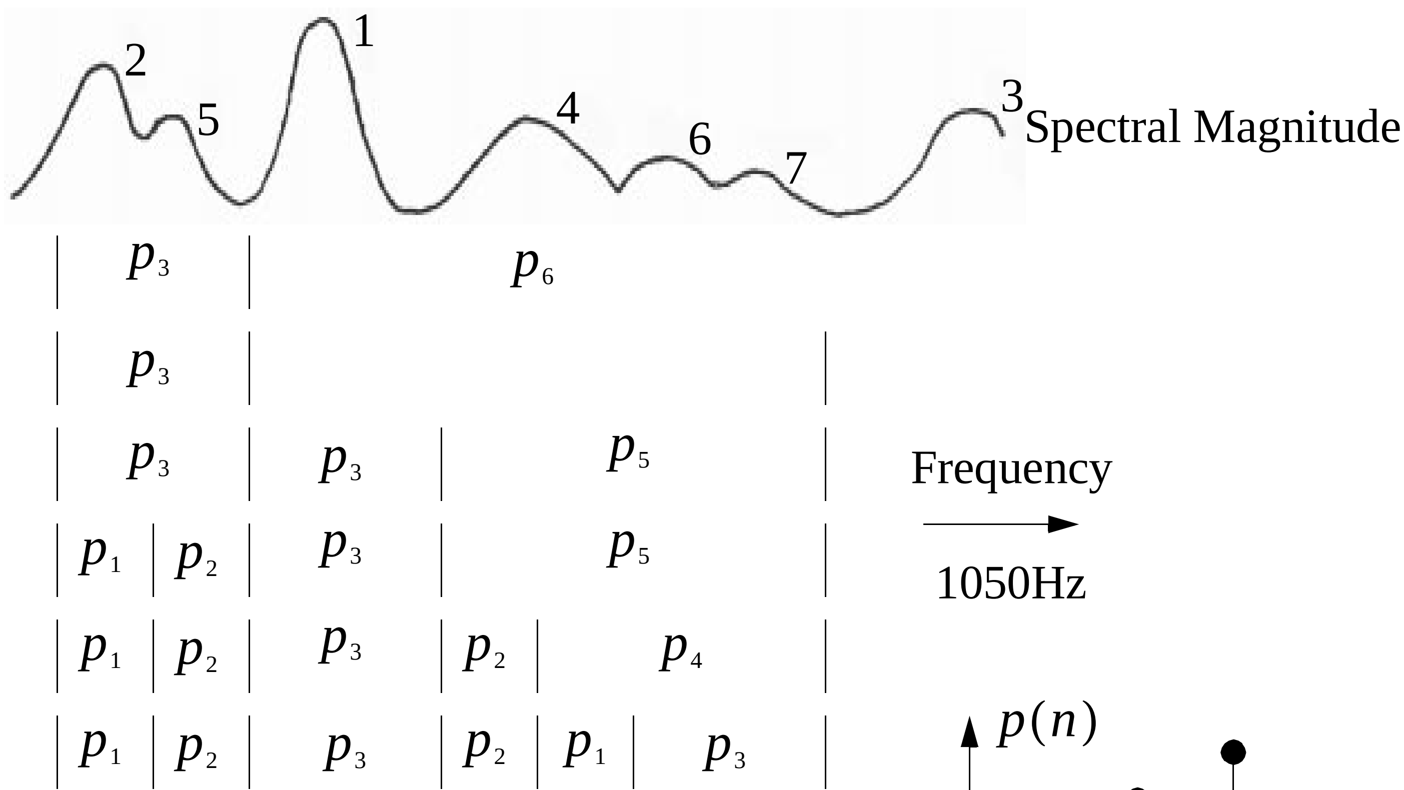


Figure 16.10 : Block Diagram of the Place Model.



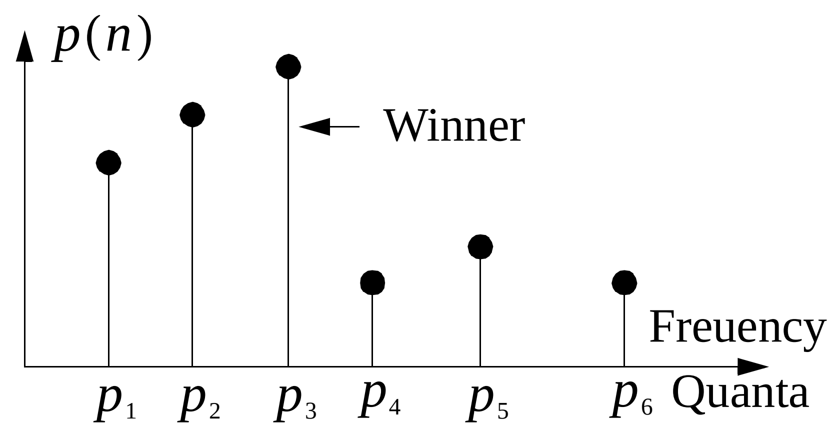


Figure 30.13 :

Armonic Pitch Detection Algorithm

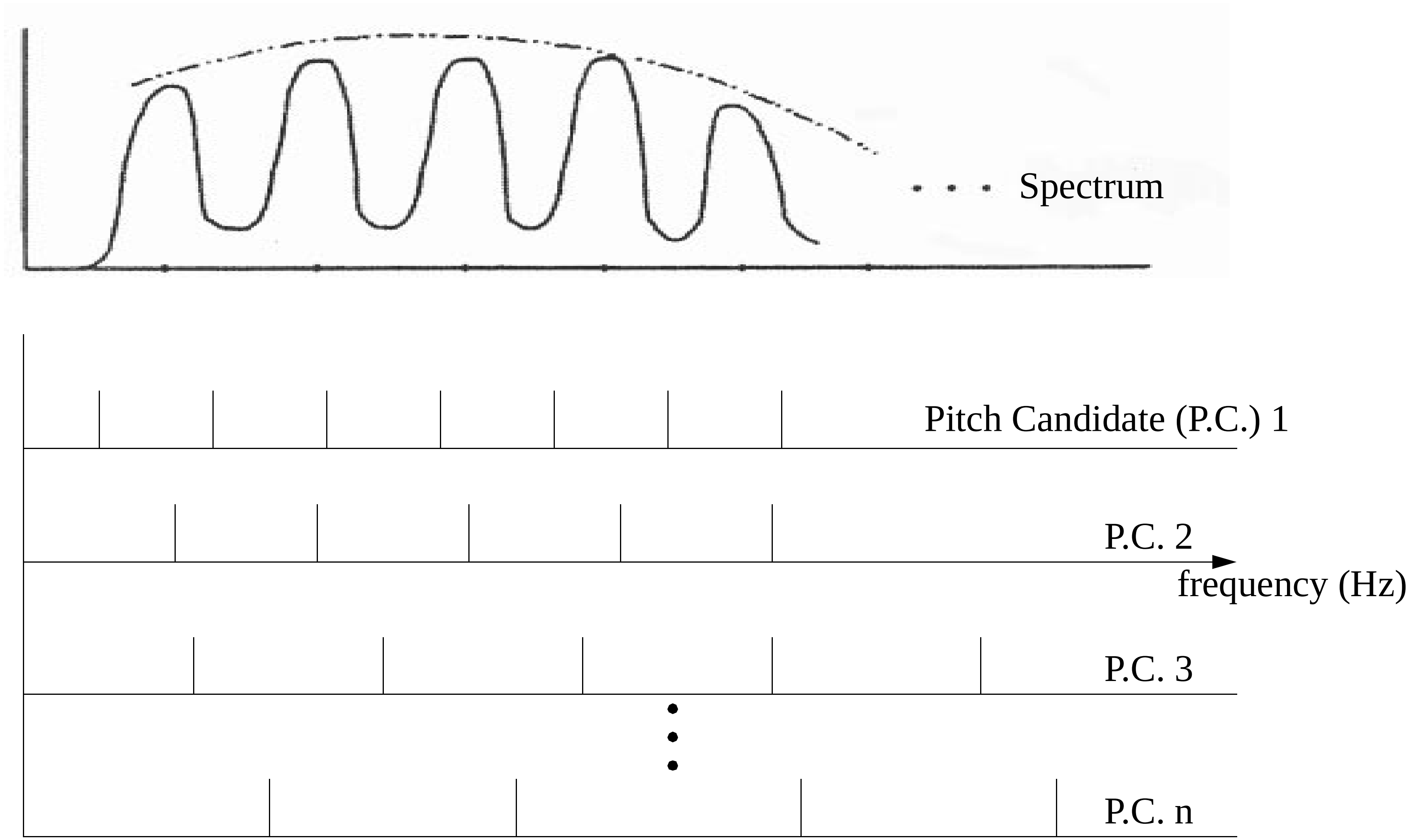


Figure 30.14 : Goldstein-Duifhuis Optimum Processor Algorithm

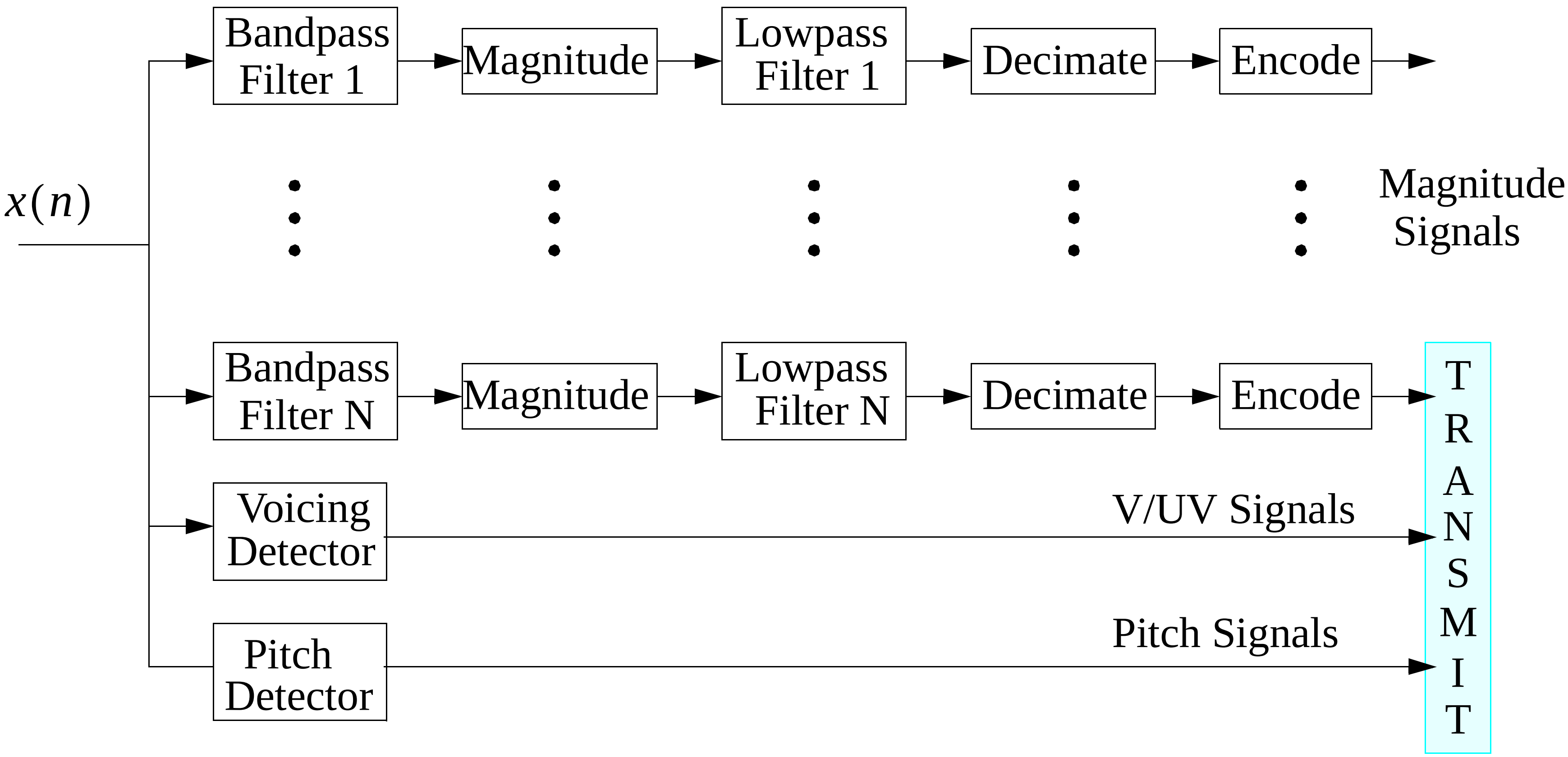


Figure 31.2 : Channel Vocoder Analyzer and Synthesizer

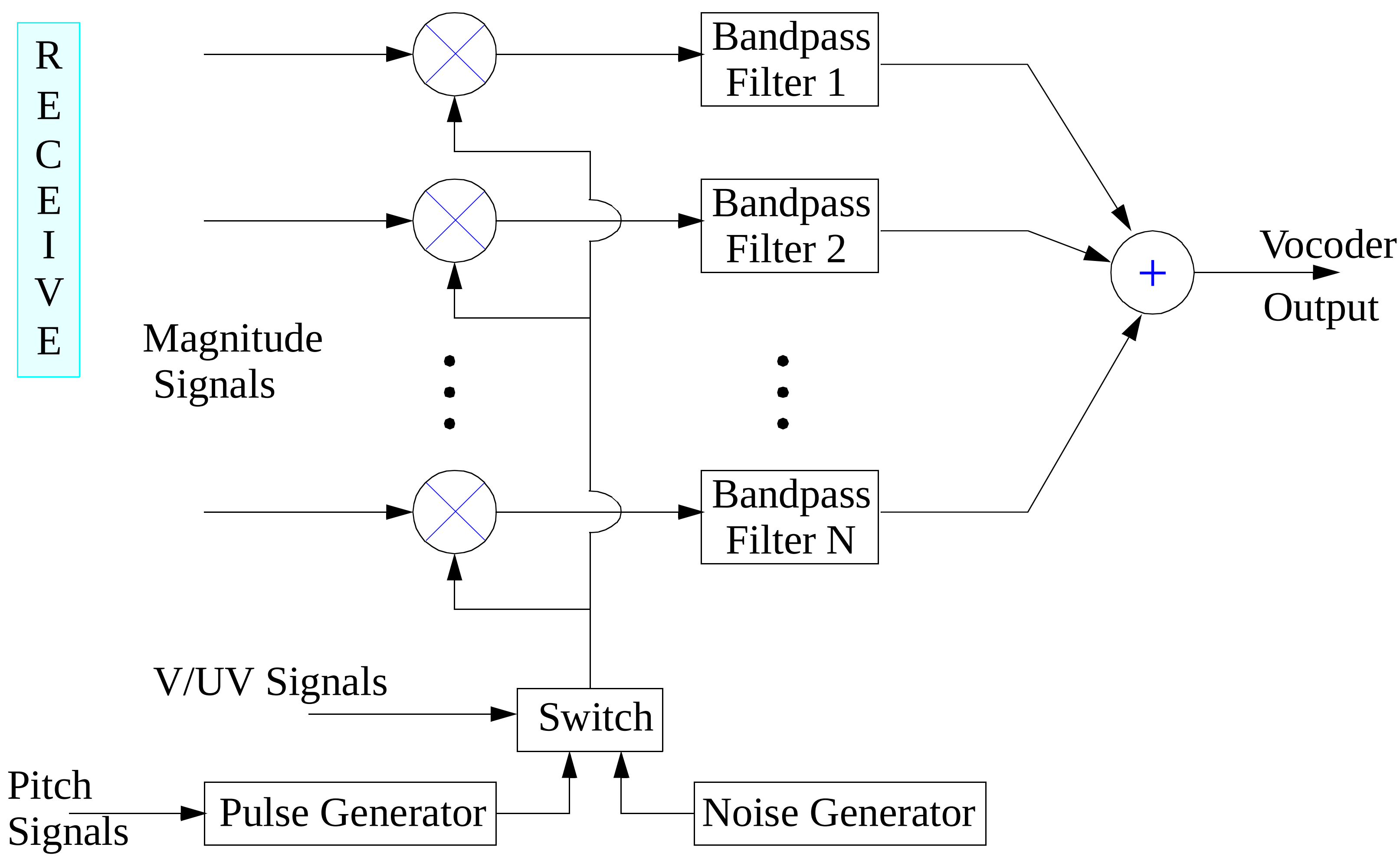


Figure 31.2 : Channel Vocoder Analyzer and Synthesizer

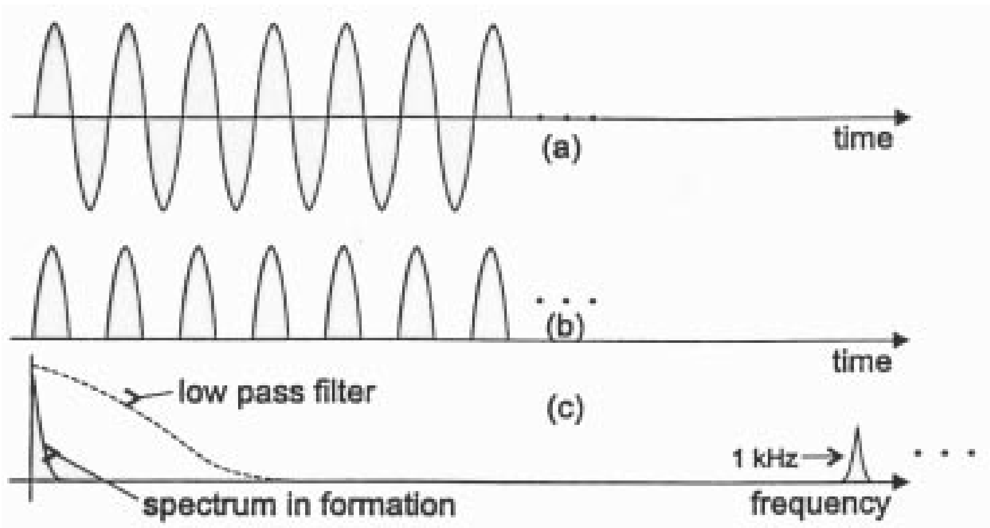


Figure 31.3 : Example of Energy Measurement With a Half-Wave Rectifier.

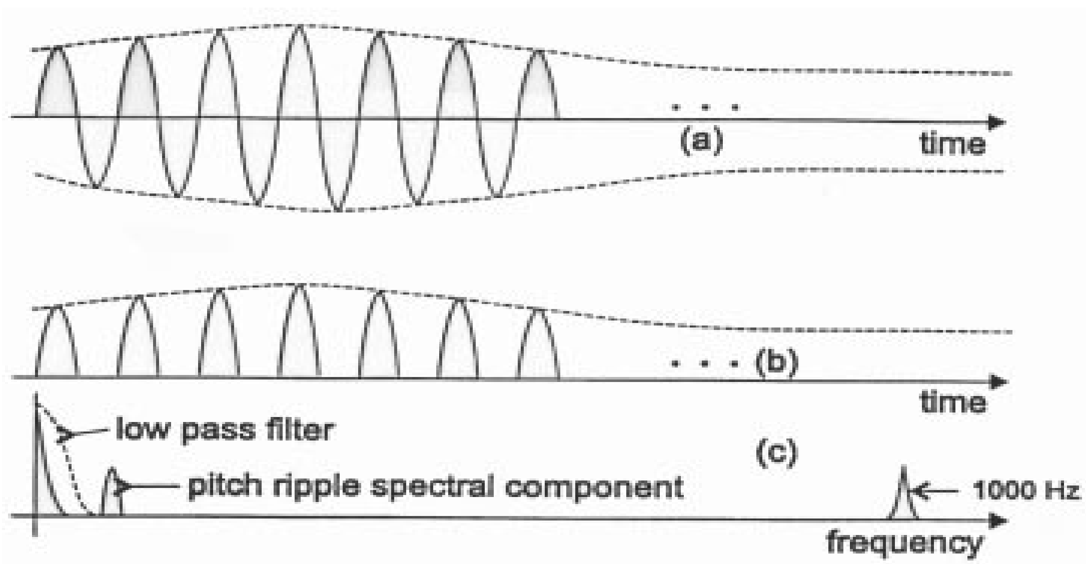


Figure 31.4 : Effect of Pitch Ripple in a Spectral Estimate

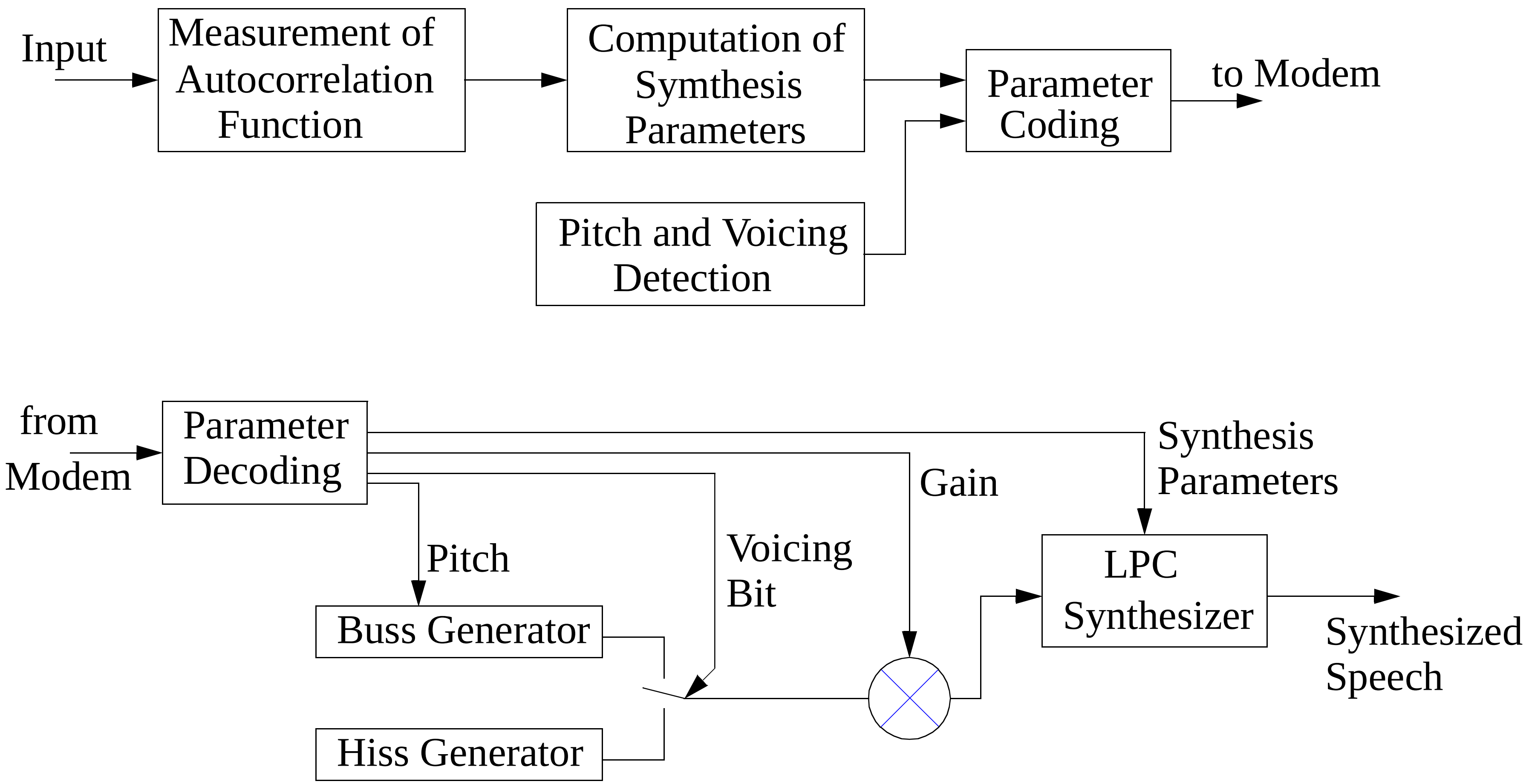


Figure 31.11 : Block Diagram of the LPC Algorithm

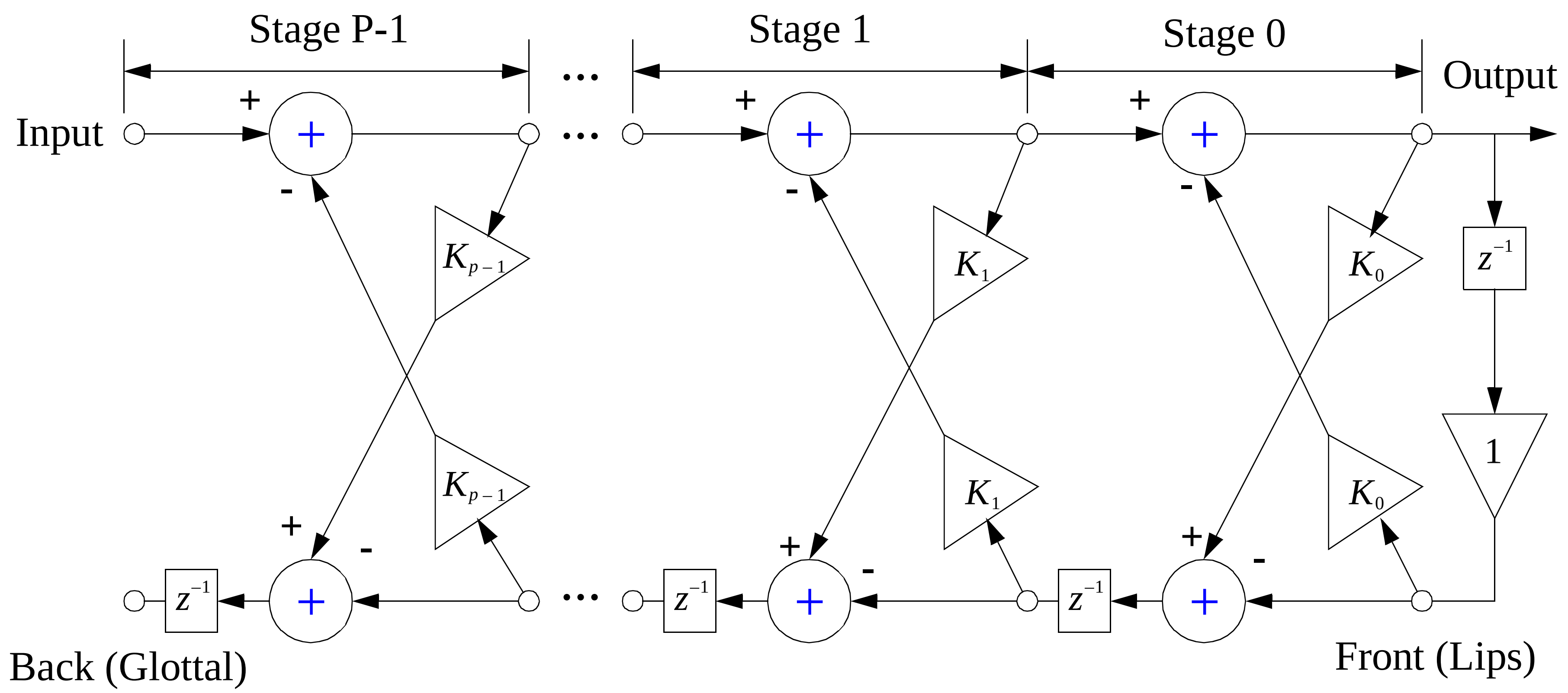
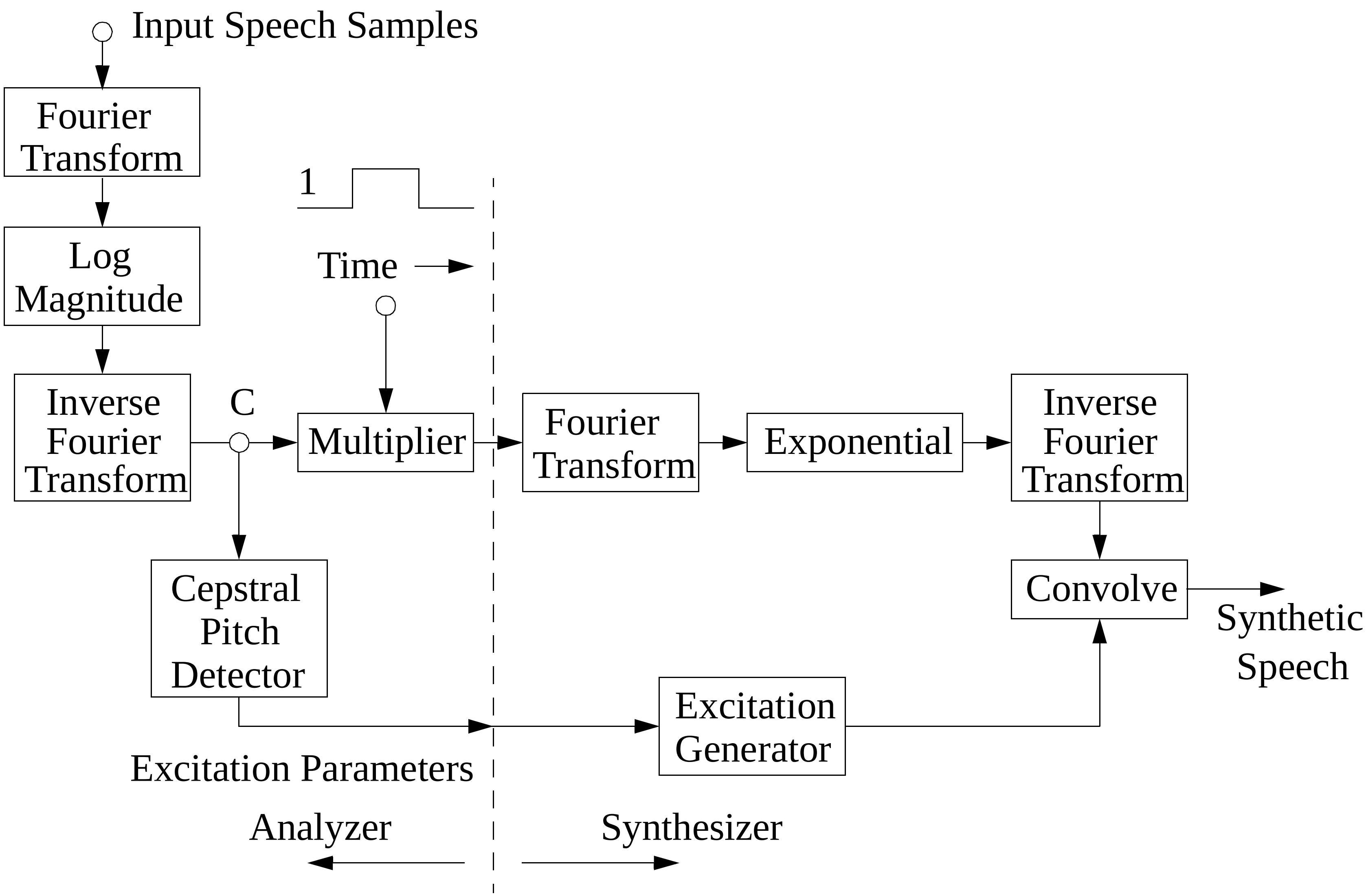


Figure 31.12 : Lattice Synthesizer for LPC



Schematic of the Homomorphic Vocoder.

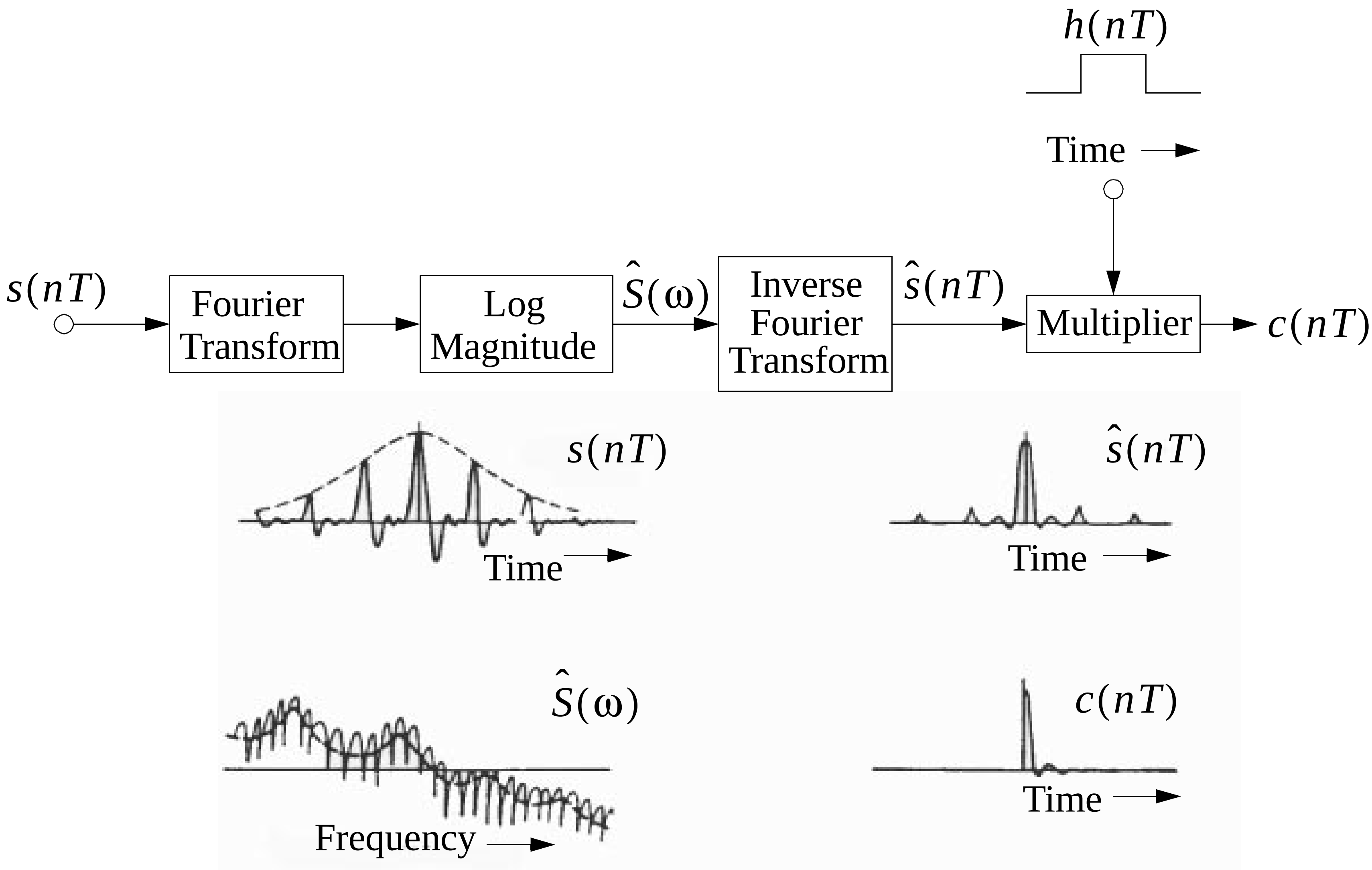


Figure 31.13 : Cepstral Vocoder Analysis

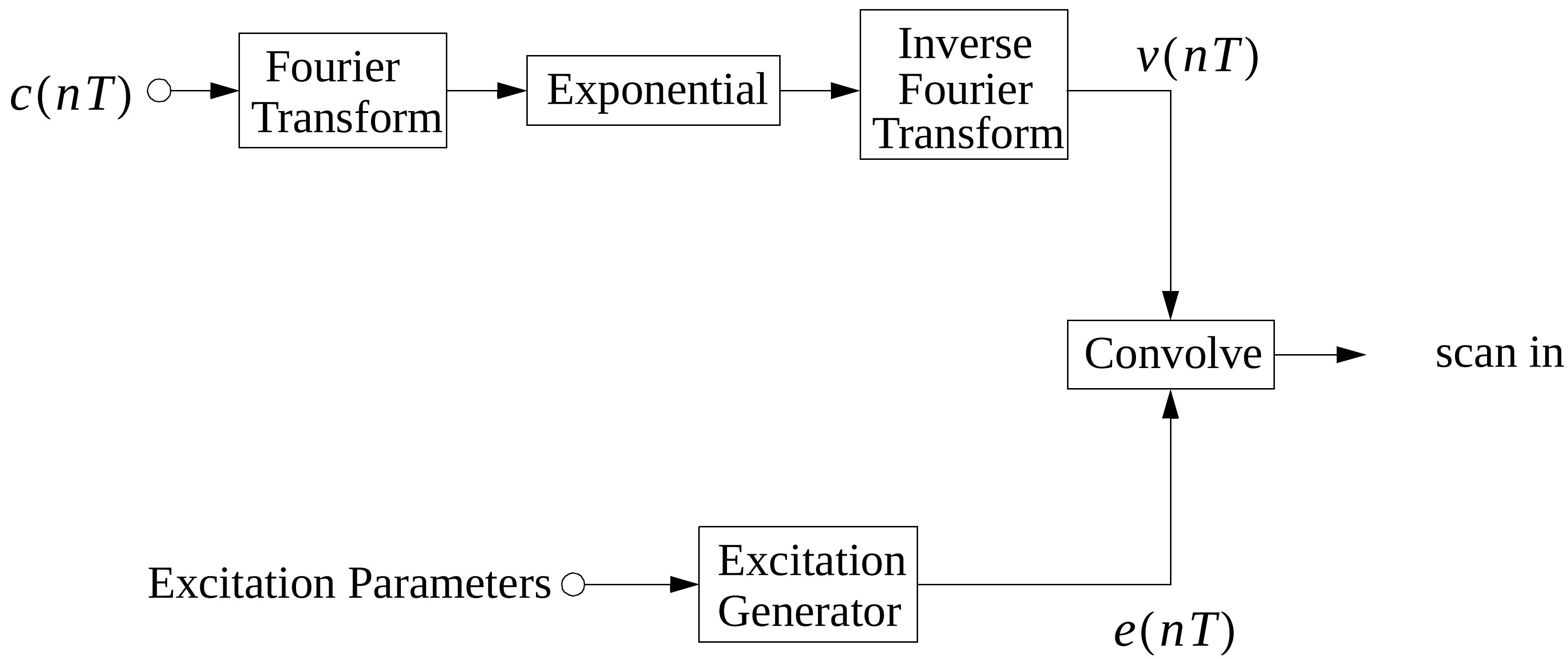


Figure 31.14 : Cepstral Vocoder Synthesizer