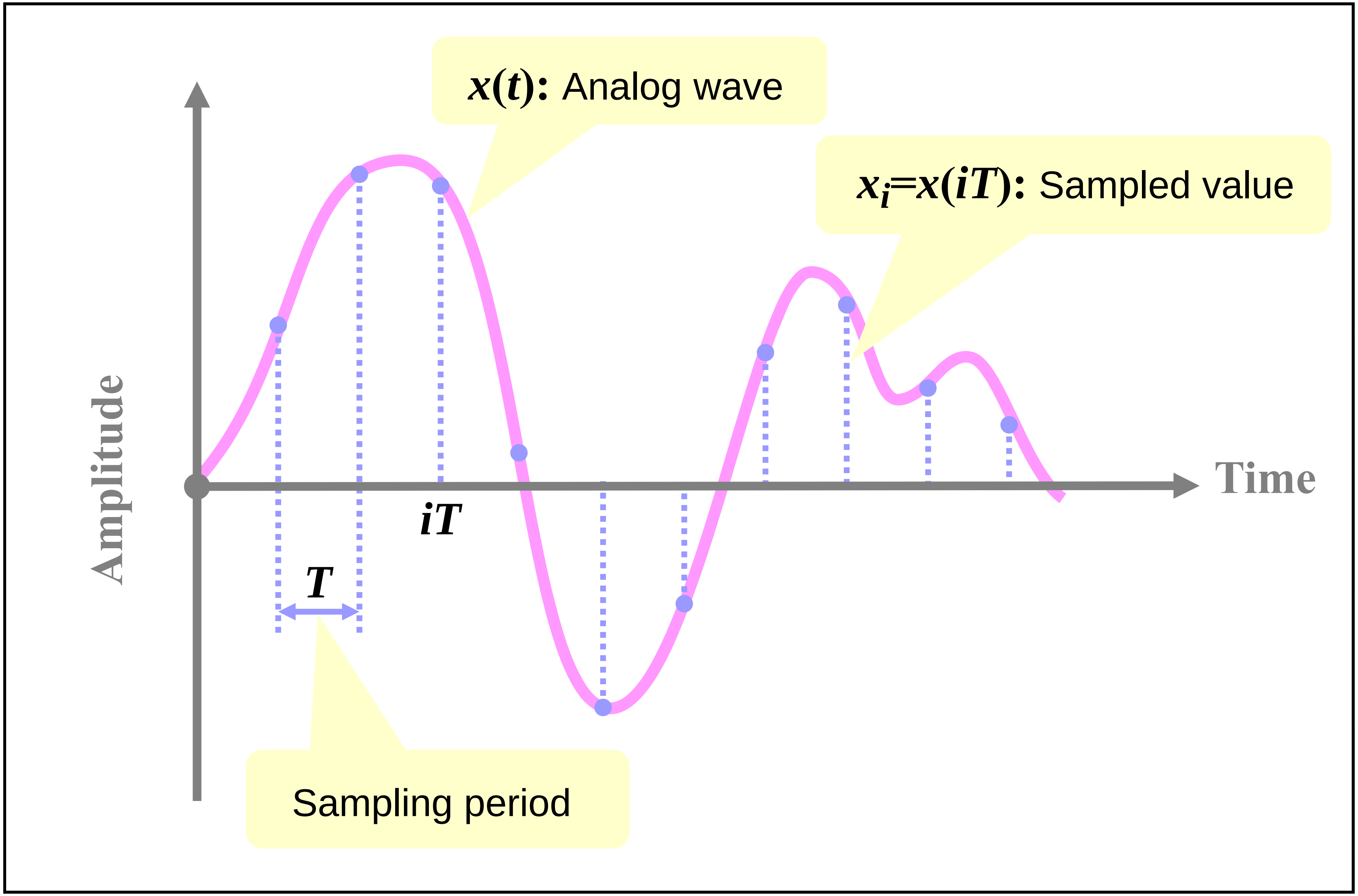
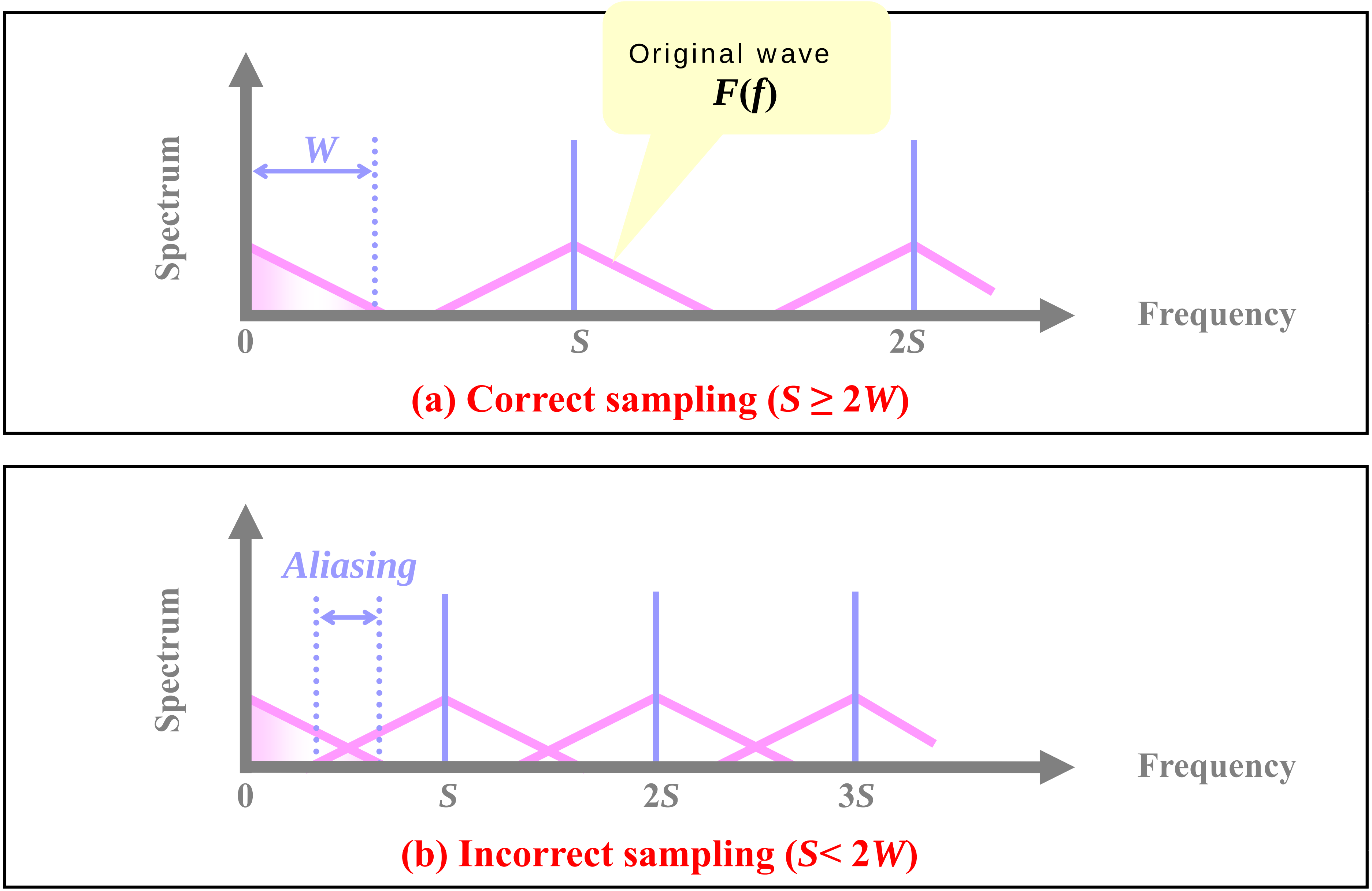
Speech Analysis

Sadaoki Furui
Tokyo Institute of Technology
Department of Computer Science
furui@cs.titech.ac.jp

0110-02

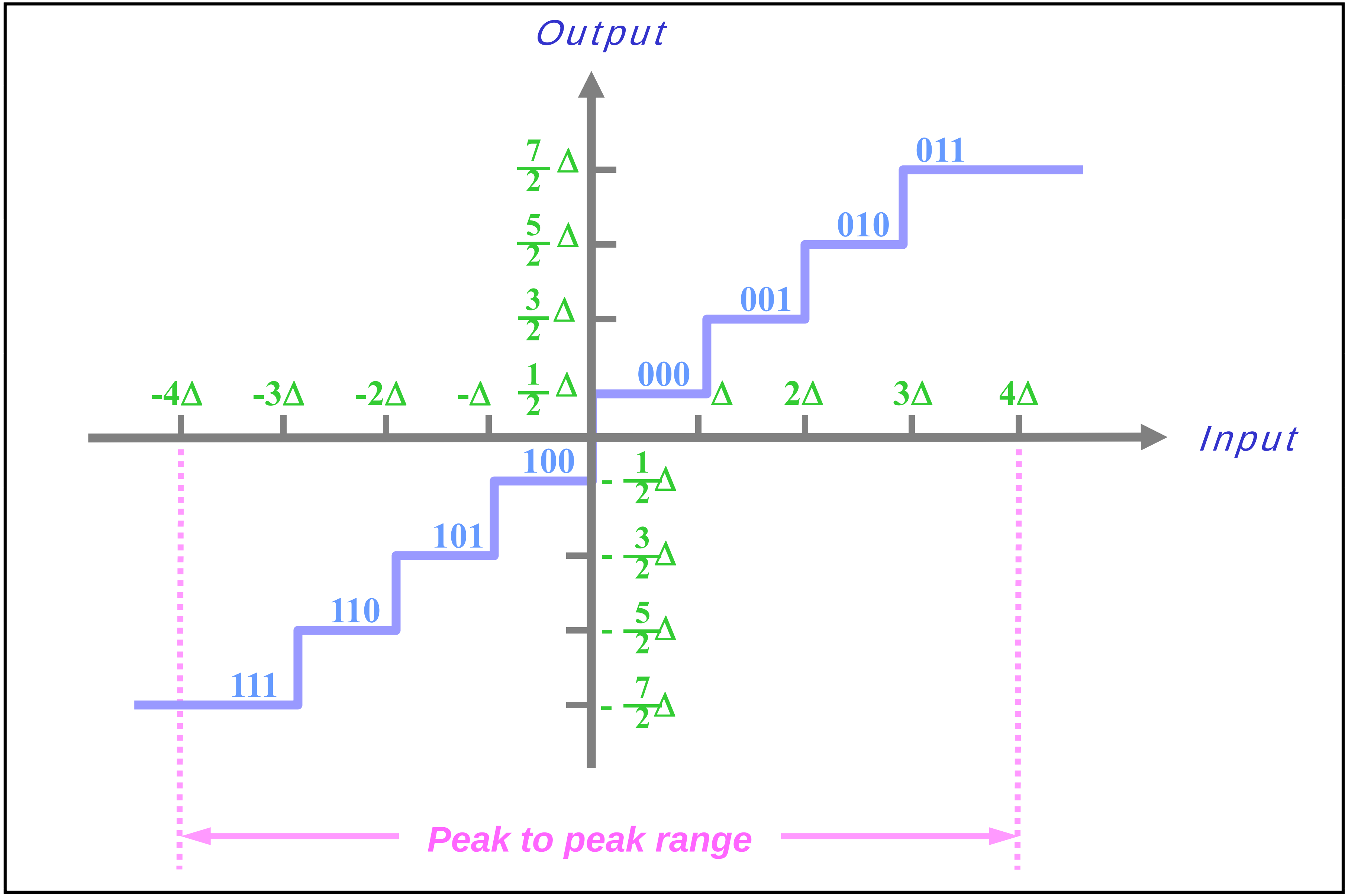


Sampling in the time domain

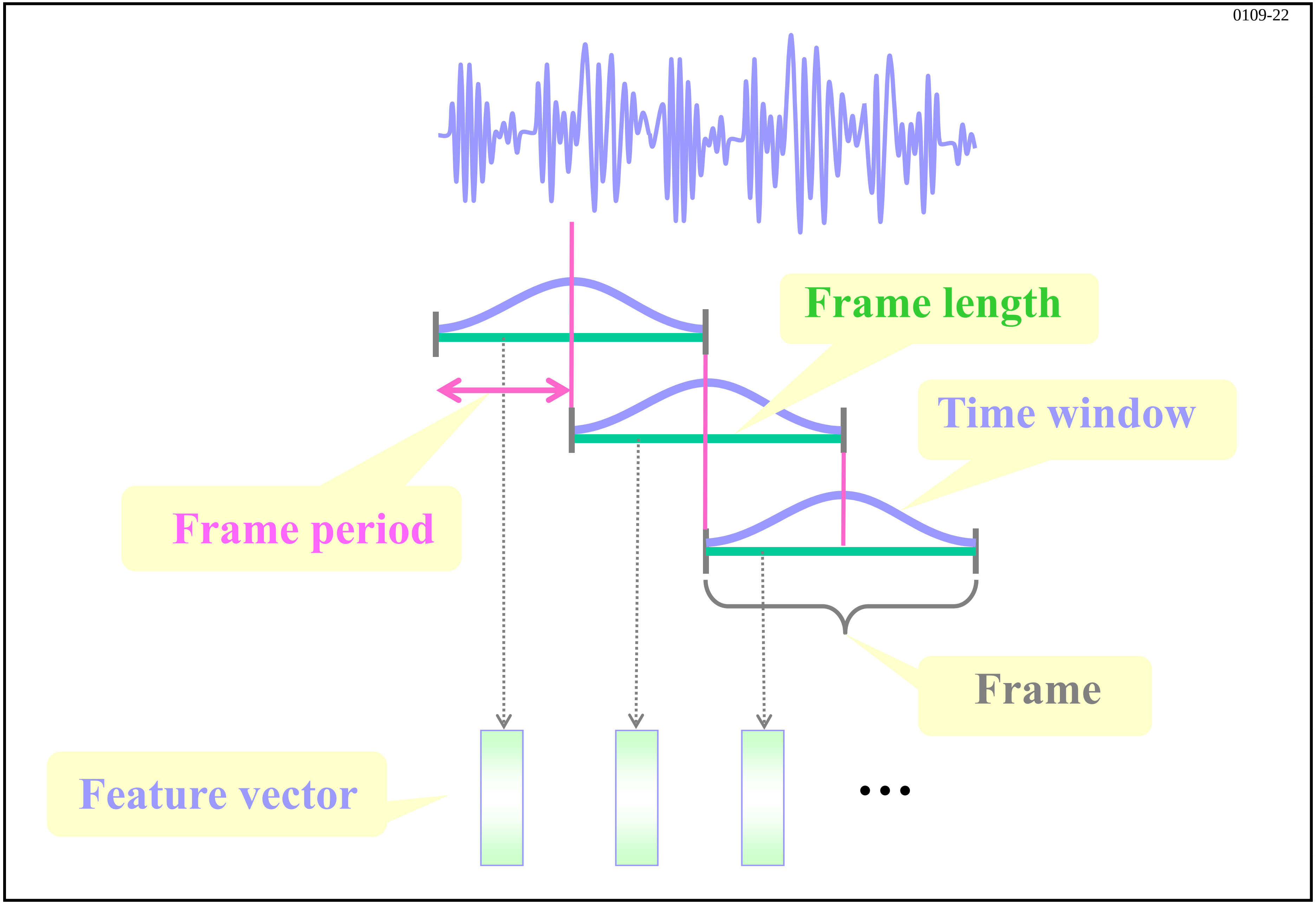


Sampling in the frequency domain

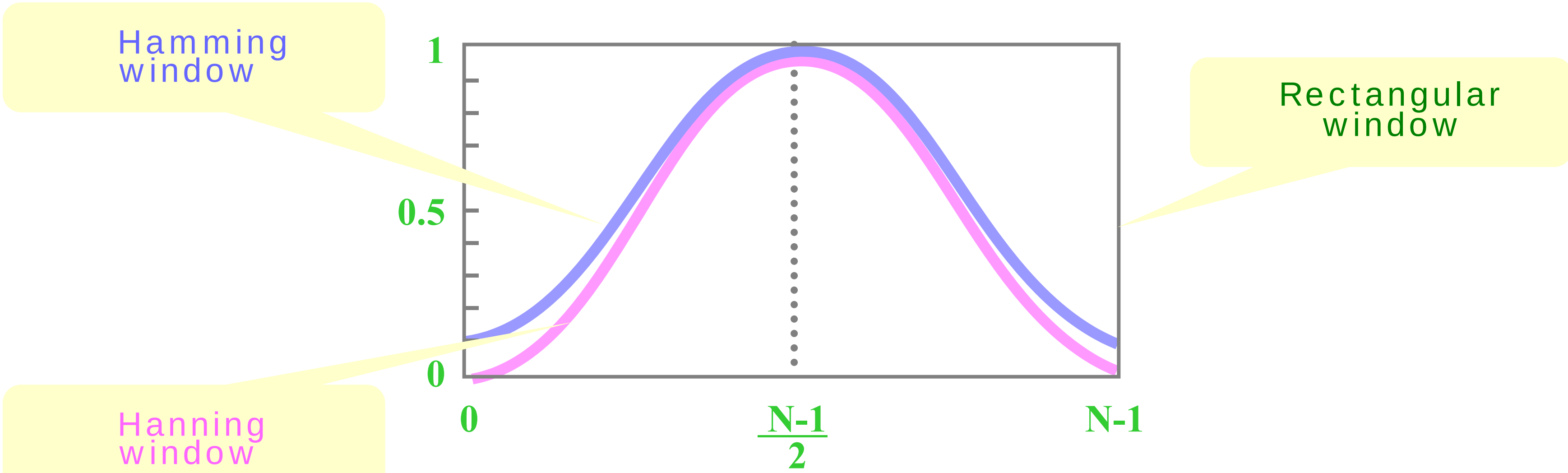
0106-1



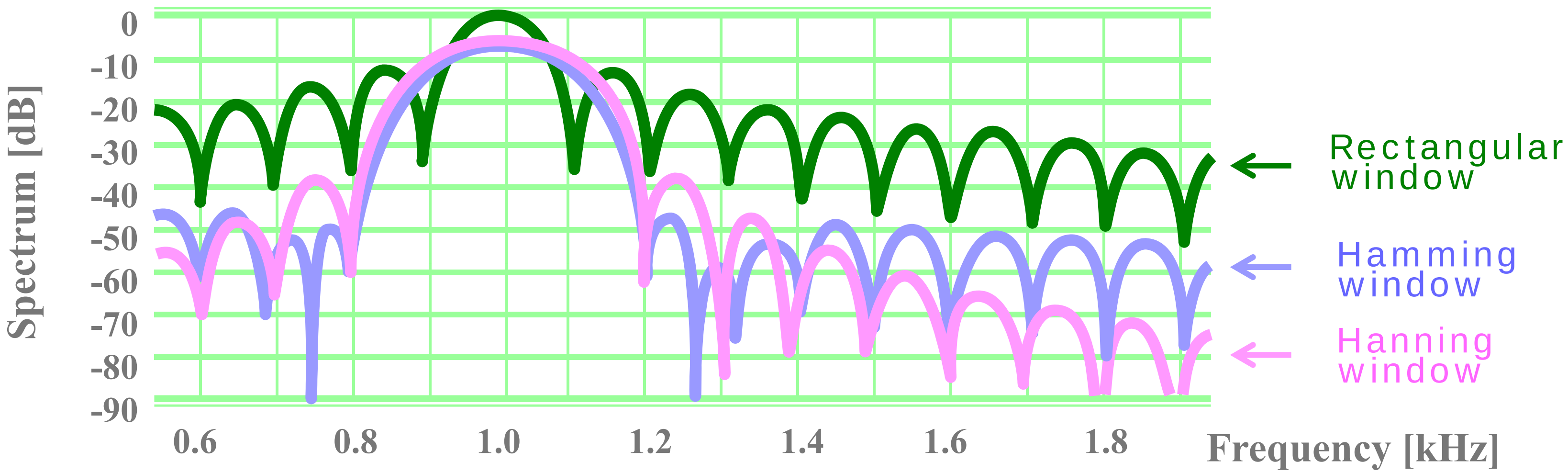
An example of the input-output characteristics of eight-level (3-bit) quantization



Feature vector (short-time spectrum) extraction from speech



(a) Major window functions

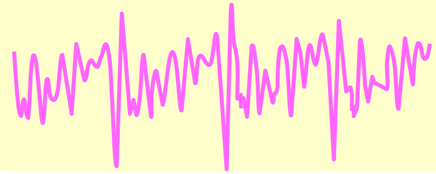


(b) The spectrum for the 10 periods of a 1-kHz sinusoidal wave extracted using each of the windows

Block diagram of a typical speech analysis procedure.

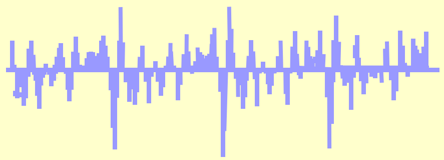
Low-pass filter

Cut off frequency = 8 kHz



A/D (Sampling and quantization)

Sampling frequency = 16 kHz
Quantization bit rate = 16 bit



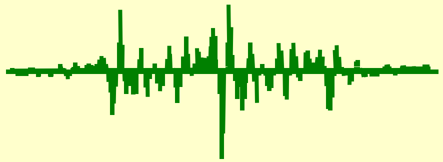
Analysis frame extraction

Frame length = 30 ms
Frame interval = 10 ms

30 ms

Windowing (Hamming, Hanning, etc.)

Window length = Frame length



Spectral analysis (FFT, LPC, etc.)

Feature extraction

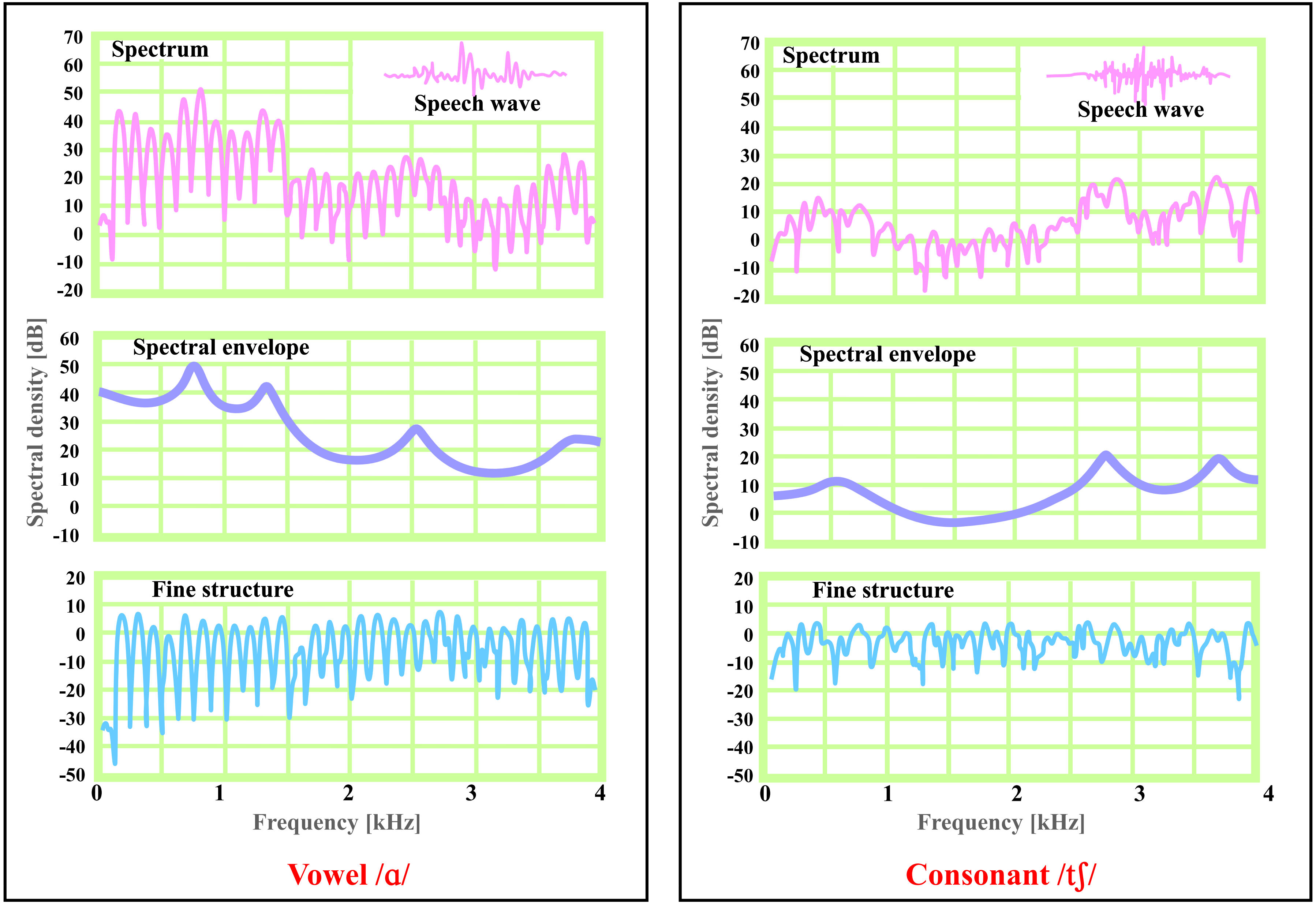
Parametric representation

- Excitation parameters
- Vocal tract parameters

Typical parameter values

Examples of speech waves at each stage

0110-04



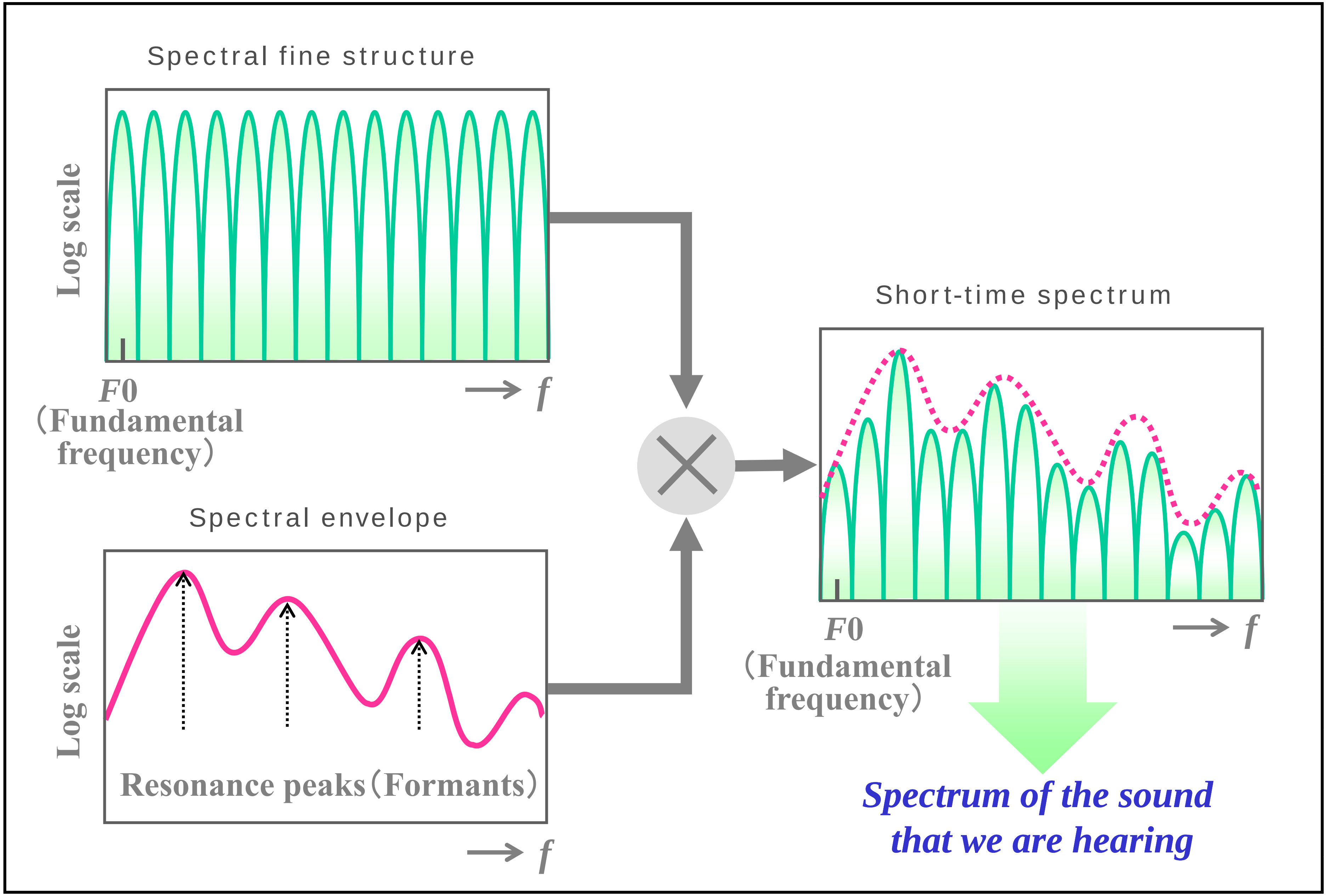
Structure of short-time speech spectra for male voices when uttering vowel /ɑ/ and consonant /tʃ/

Major methods for analyzing speech spectra and their principal features

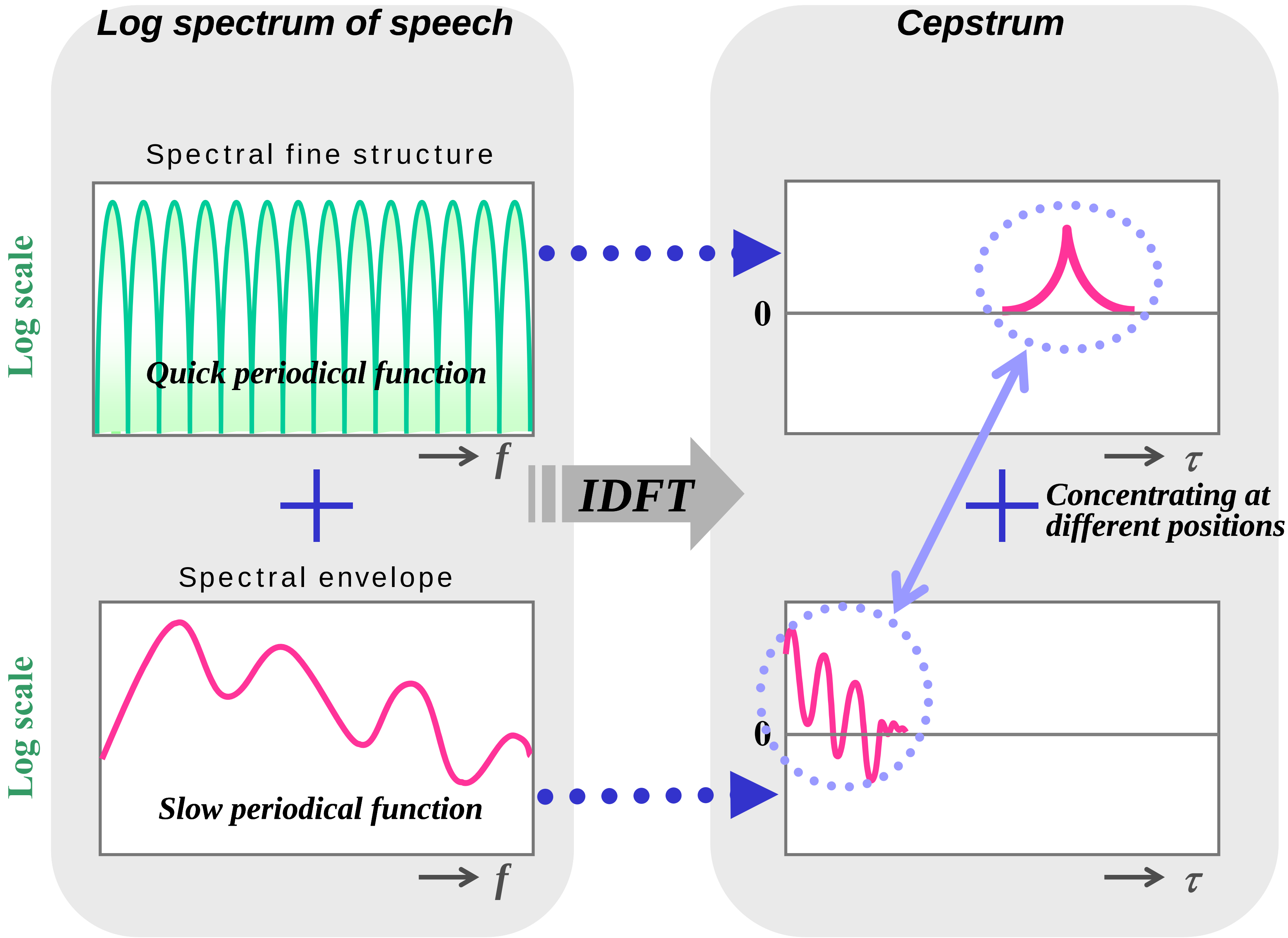
Type	Analysis methods	Parameters	Features
NPA	(i) Short-time autocorrelation	$\phi(m)$	Spectral envelope and fine structure are convoluted.
	(ii) Short-time spectrum	$S(\omega)$	Spectral envelope and fine structure are multiplied. Fast algorithm can be realized by FFT.
	(iii) Cepstrum	$c(\tau)$	Spectral envelope and fine structure can be Separated in quefrency domain. Two FFTs and log transform are necessary.
	(iv) Band-pass filter bank	rms of filter output	Global spectral envelope can be obtained.
	(v) Zero-crossing analysis	Zero-crossing rate	Formant freq. Can be obtained by Combination with (iv). Realized by simple hardware.
PA	(i) Analysis-by -synthesis	Formant, band-width, etc.	Precise modeling is possible. Accurate formant freq. can be obtained. Complicated iteration is necessary.
	(ii) Linear predictive coding		Simple all-pole spectrum modeling.Parameters can be estimated from auto-corr. or ovariance without iteration.

Major methods for analyzing speech spectra and their principal features (continued)

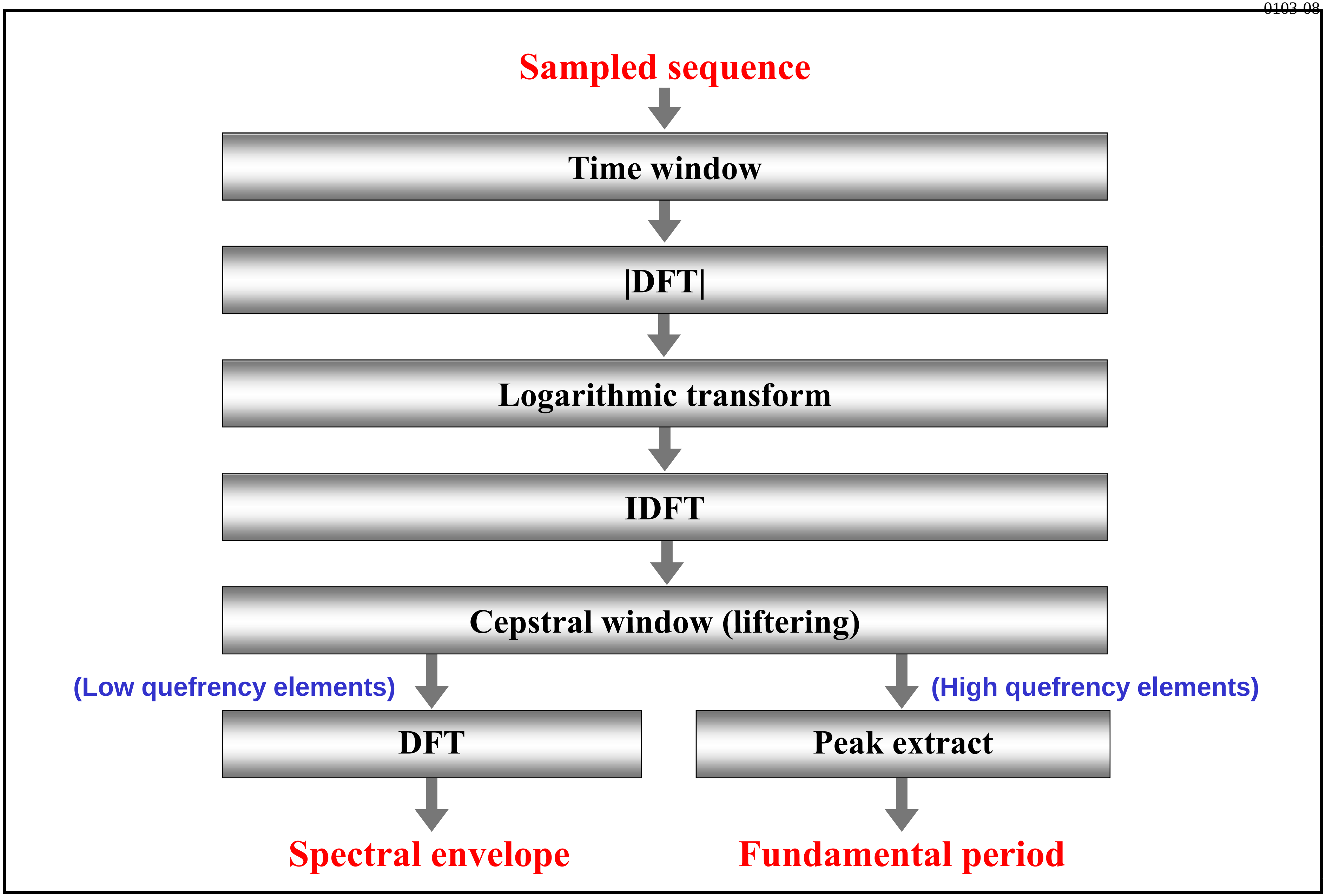
Type	Analysis methods	Parameters	Features
PA (cont.)	(ii-*a*) Maximum likelihood method	α_i	Stability of synthesis filter is guaranteed. Time window is necessary. Number of calculations $\propto p^2$
	(ii-*b*) Convariance method	α_i	Stability of synthesis filter is not guaranteed. Suitable for short-time analysis. Number of calculations $\propto p^3$
	(ii-*c*) PARCOR method	k_i	Normal equation can be solved by lattice filter. Equivalent to (*a*) and (*b*). Number of calculations $\propto p^2$
	(ii-*d*) LSP method	ω_i	Quantization and interpolation characteristics are good. Similar to formant. Number of calculation is slightly larger than for PARCOR.



Spectral structure of speech

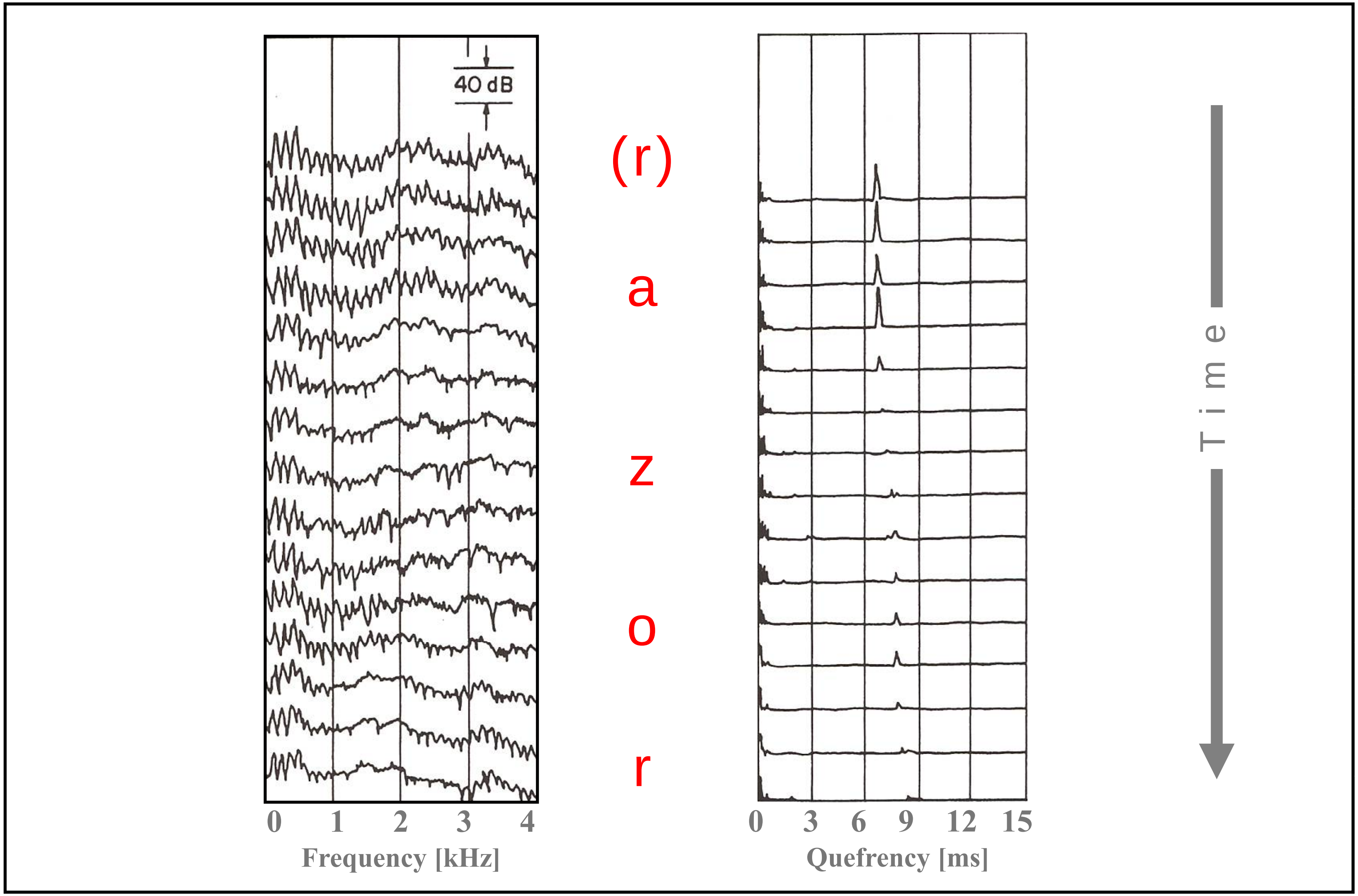


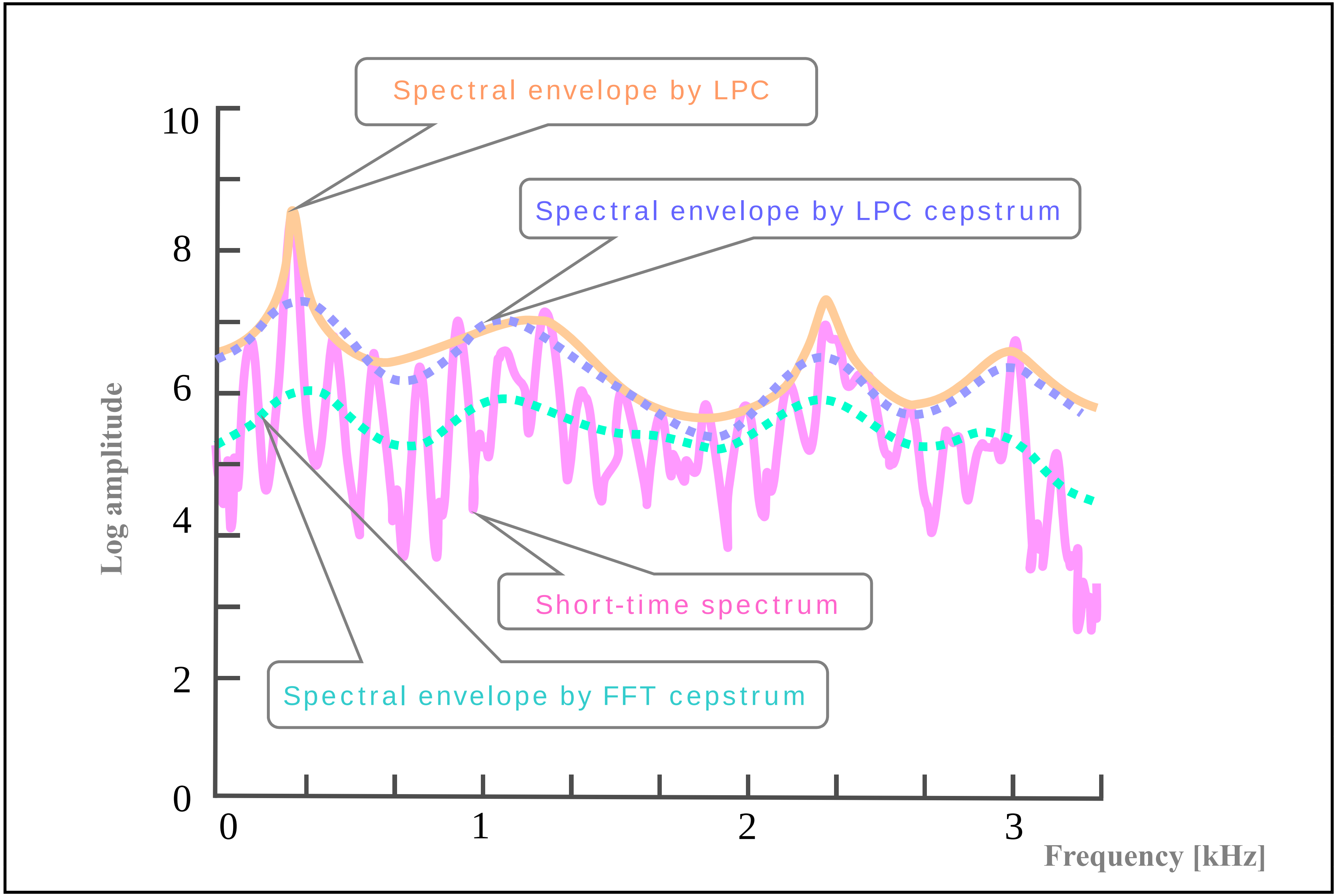
Log spectrum and Cepstrum



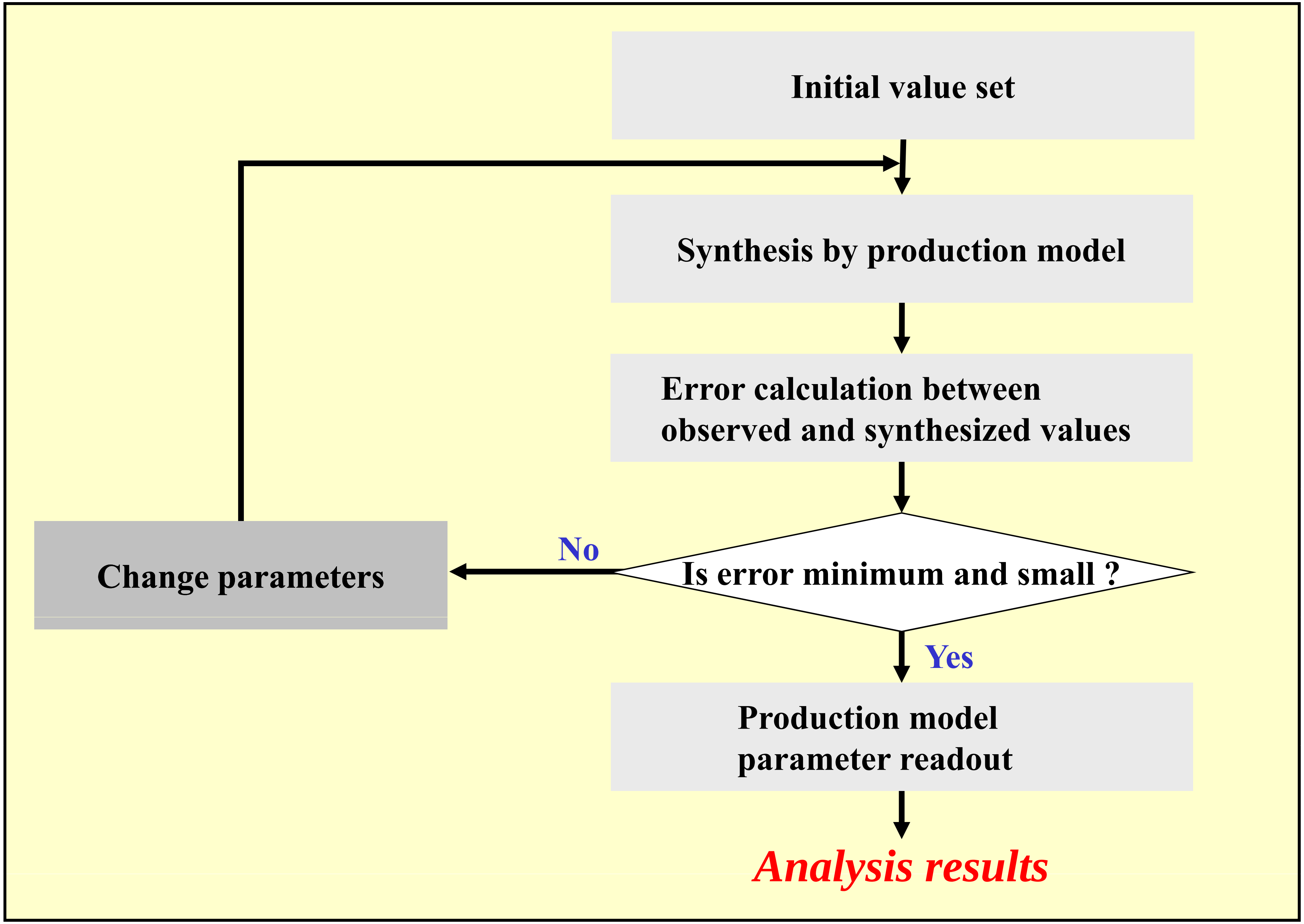
Block diagram of cepstrum analysis for extracting spectral envelope and fundamental period

Examples of short-time spectra (left) and cepstra (right) for male voice when uttering "(r)azor". Sampling frequency 10kHz; Hamming window length 40ms; frame interval 10ms.





Comparison of spectral envelopes by LPC, LPC cepstrum, and FFT cepstrum methods



Principle of analysis-by-synthesis (A-b-S) method

Linear prediction

(Time-domain representation)

$$x_t + \underbrace{\sum_{i=1}^{p} \alpha_i x_{t-i}}_{-\hat{x}_t} = \varepsilon_t$$

$$x_t - \hat{x}_t = \varepsilon_t$$

All-pole/Auto-regression model

(Spectral-domain representation)

$$H(z) = \frac{1}{1+\sum_{i=1}^{p} \alpha_i z^{-i}} = \frac{1}{\prod_{i=1}^{p}\left(1-\frac{z}{z_i}\right)}$$

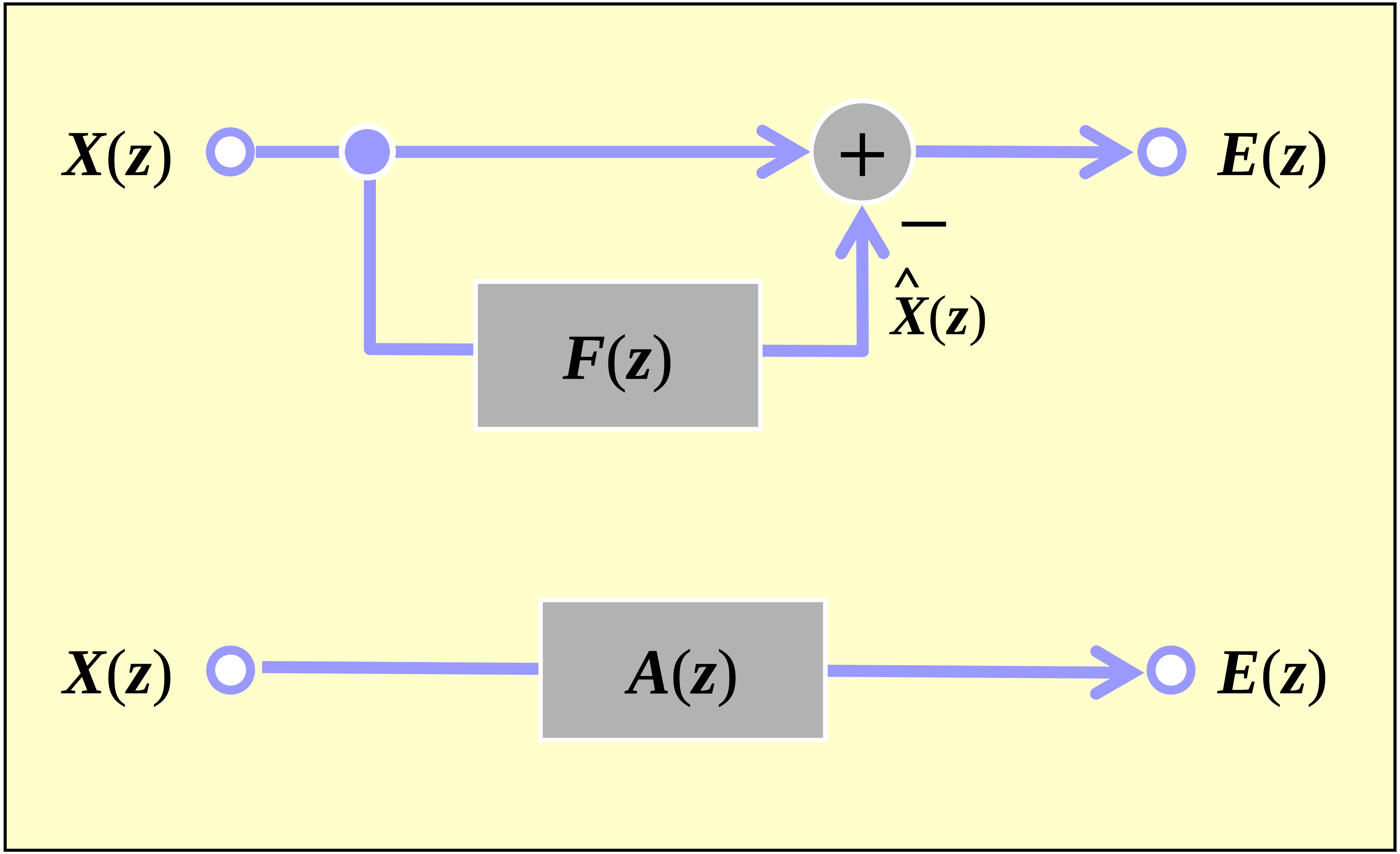
LMS estimation

Maximum-likelihood estimation

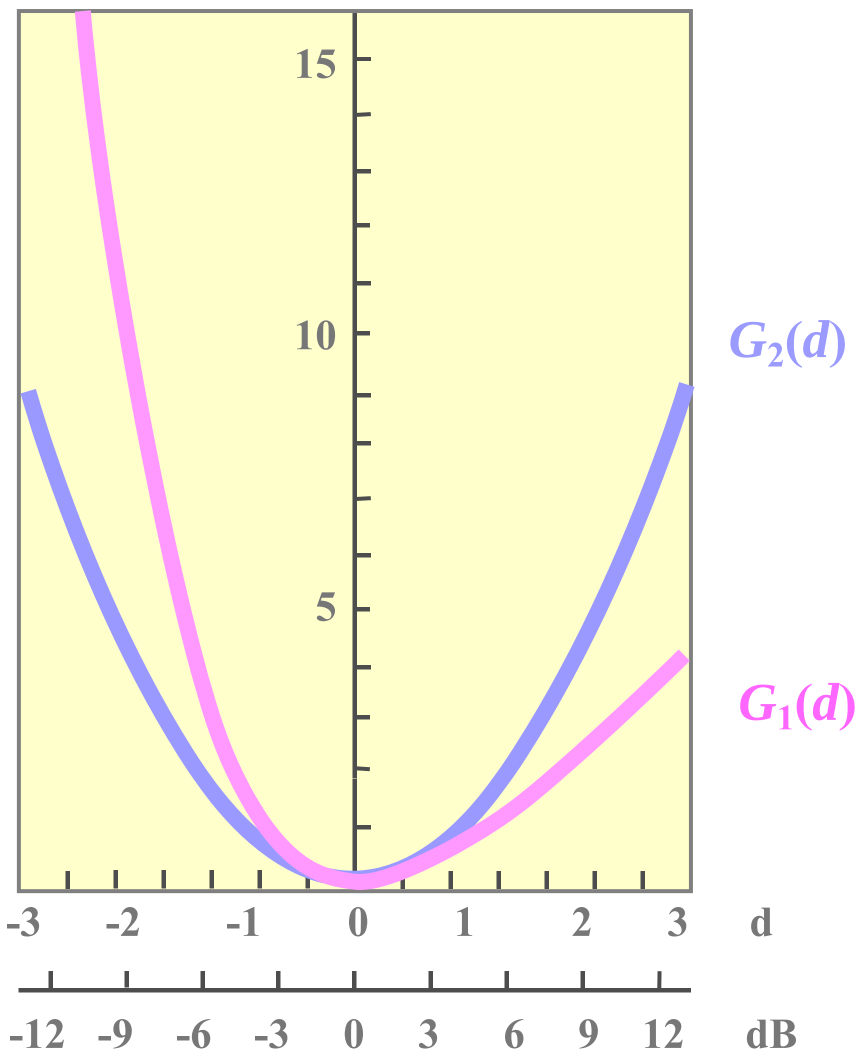
Normal equation

$$\begin{pmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \ddots & \vdots \\ \vdots & \ddots & r(0) & r(1) \\ r(p-1) & \cdots & r(1) & r(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} = - \begin{pmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{pmatrix}$$

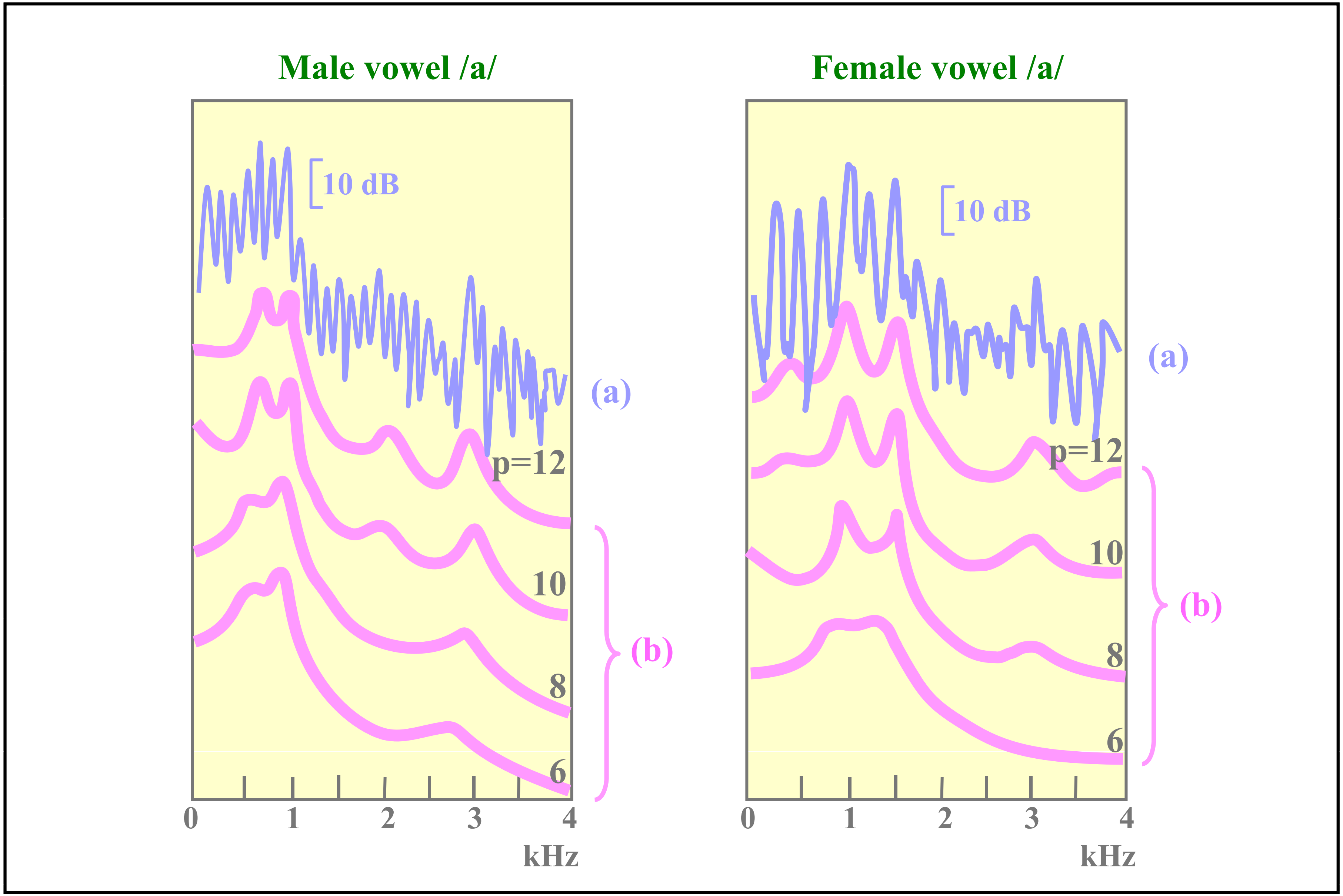
Principles of LPC analysis



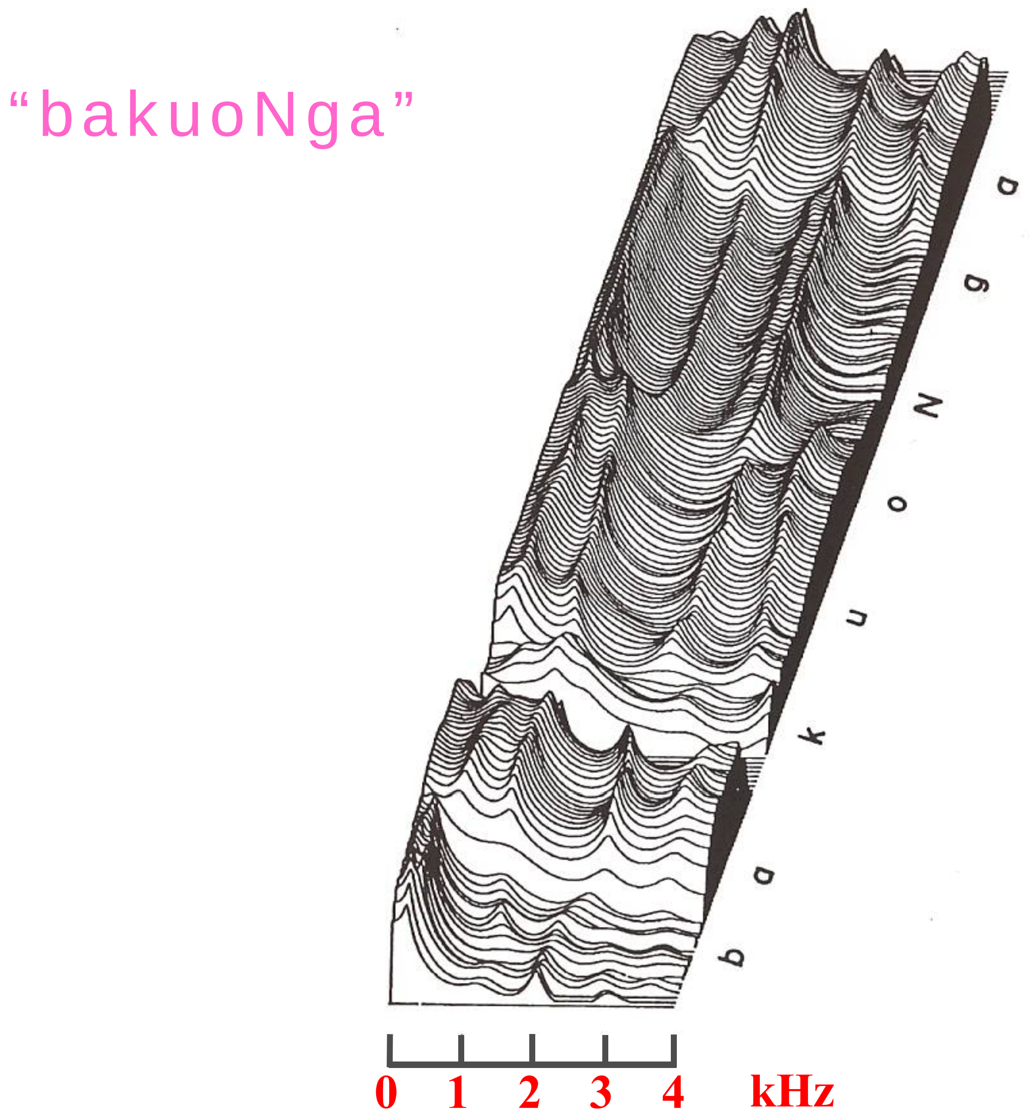
Linear prediction model block diagram



Comparison of matching error measure in maximum likelihood method, $G_1(d)$, with that in analysis-by-synthesis (A-b-S) method, $G_2(d)$.
$d = \log\{f(\lambda) / \hat{f}(\lambda)\}$; $f(\lambda)$ = model spectrum; $\hat{f}(\lambda)$ = short-term spectrum.

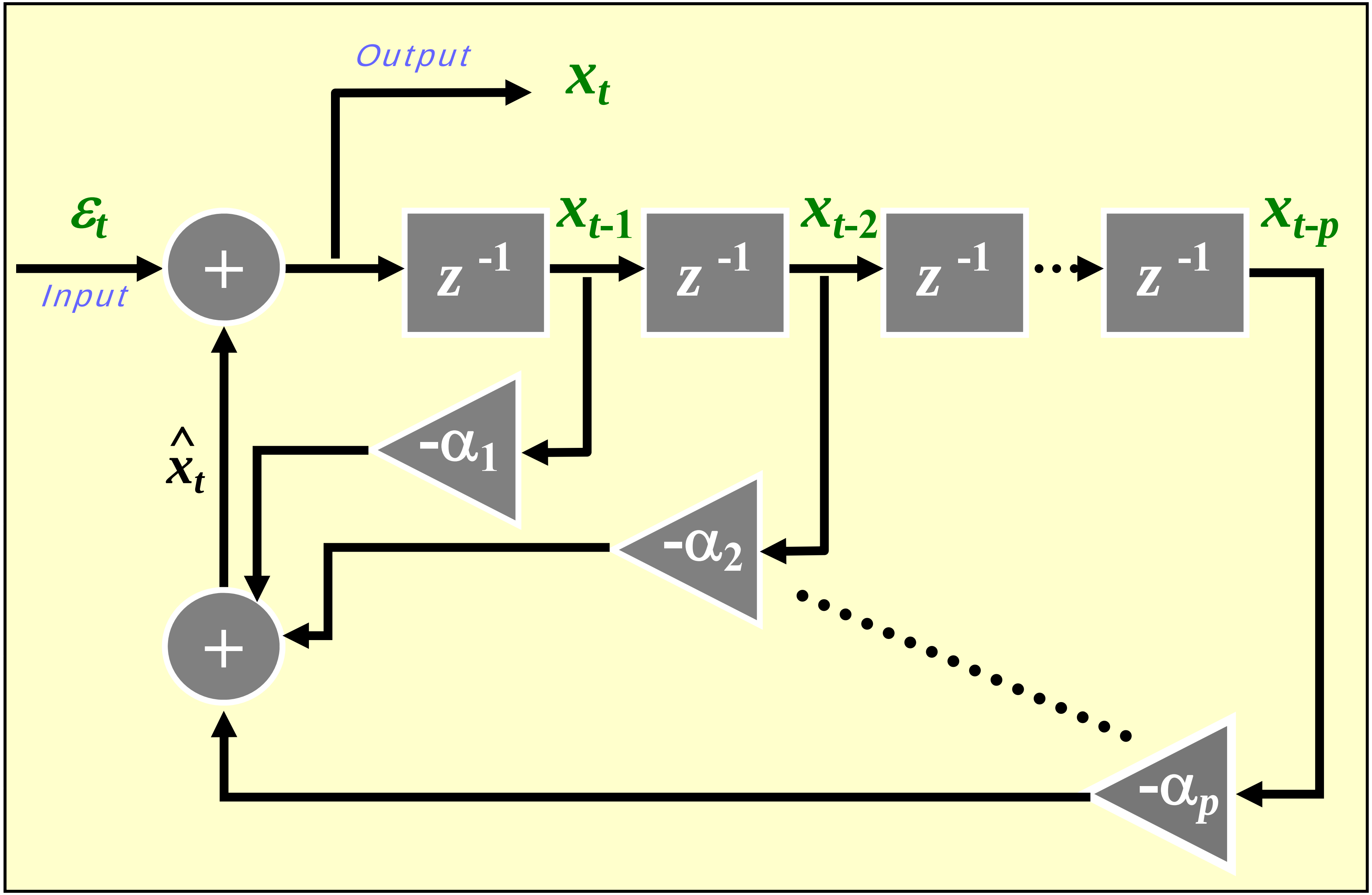


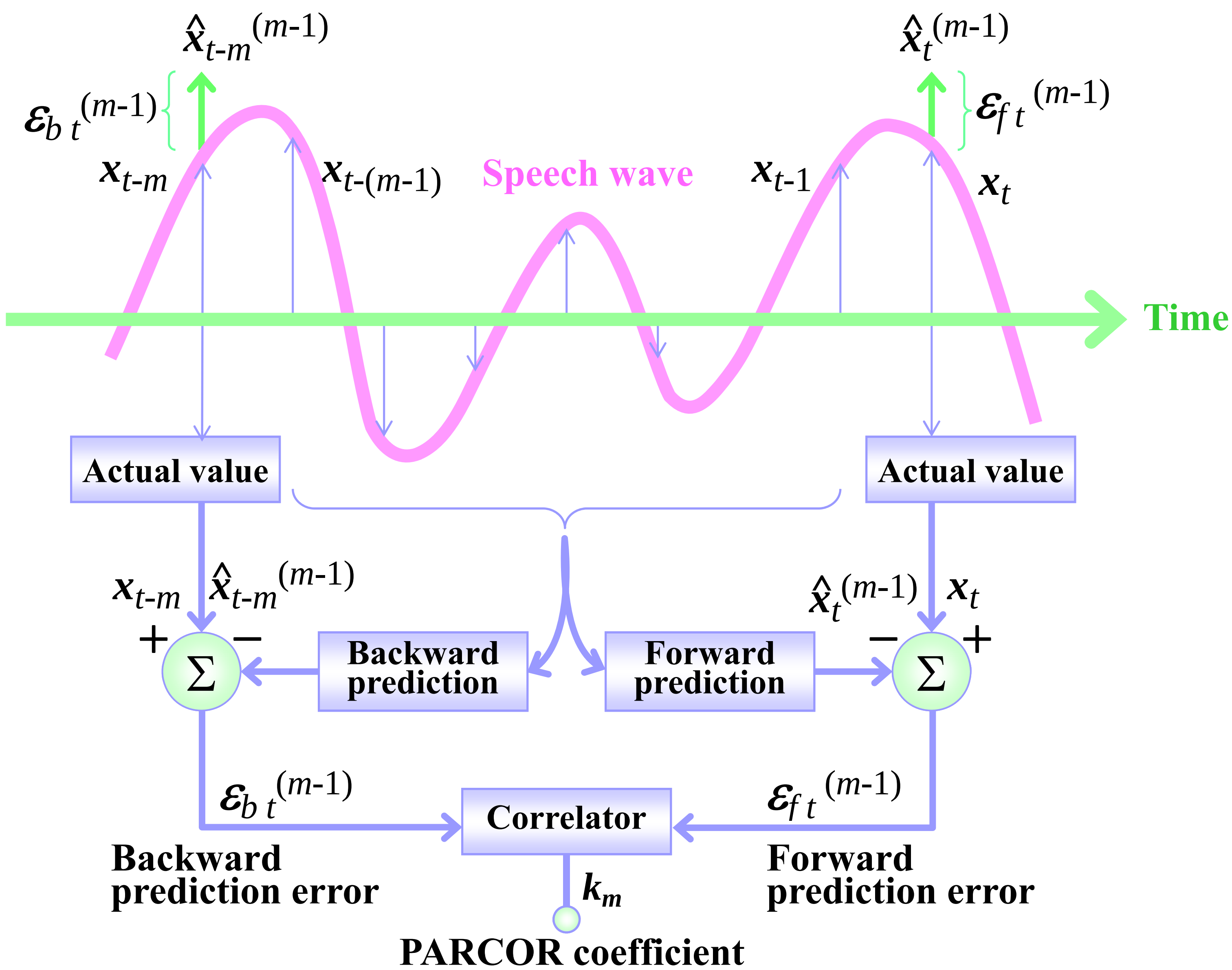
Comparison of (a) short-term spectra and (b) spectral envelopes obtained by the maximum likelihood method



Time function of spectral envelopes for the Japanese phrase /bakuoNga/ uttered by a male speaker

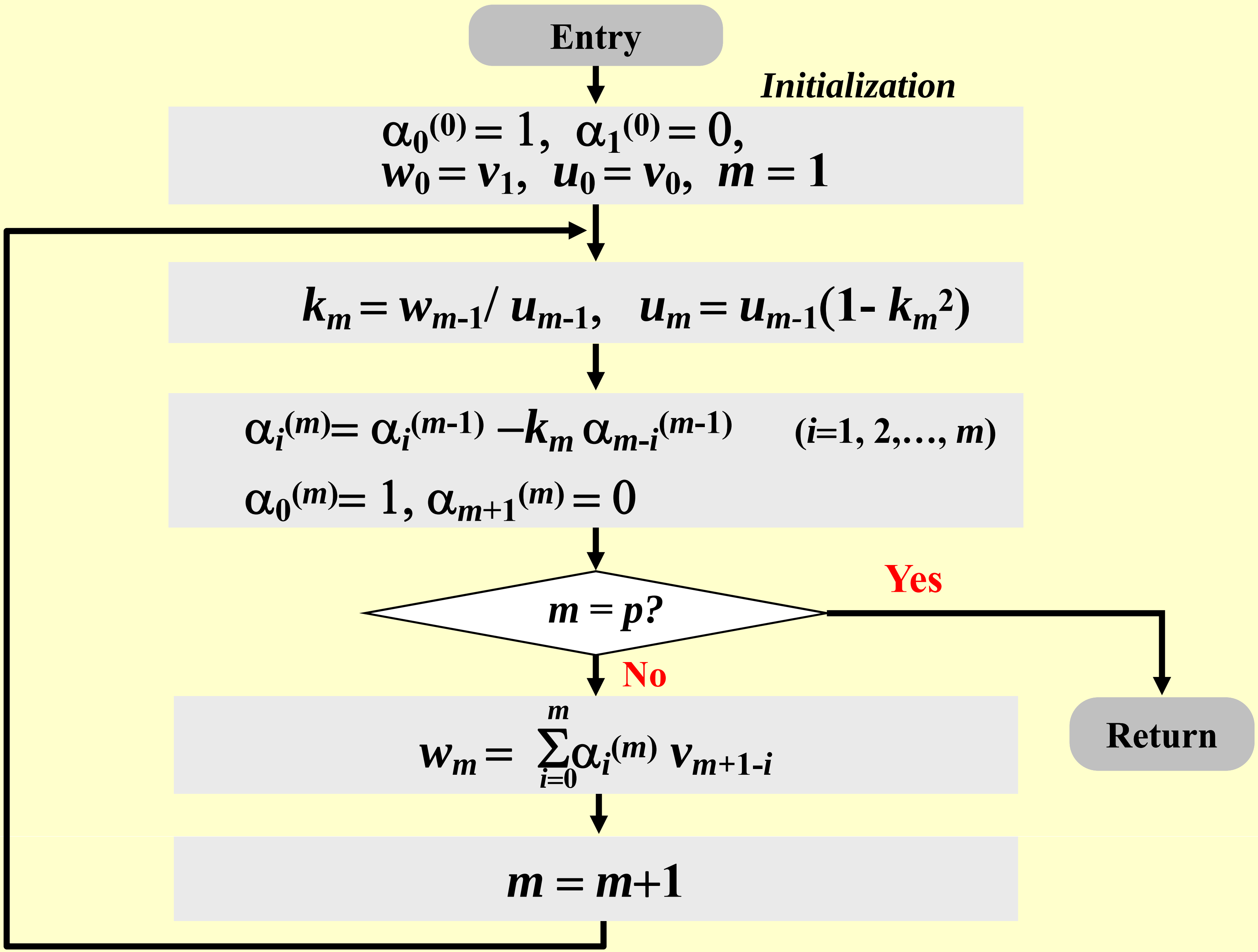
Speech synthesis circuit based on linear predictive analysis method

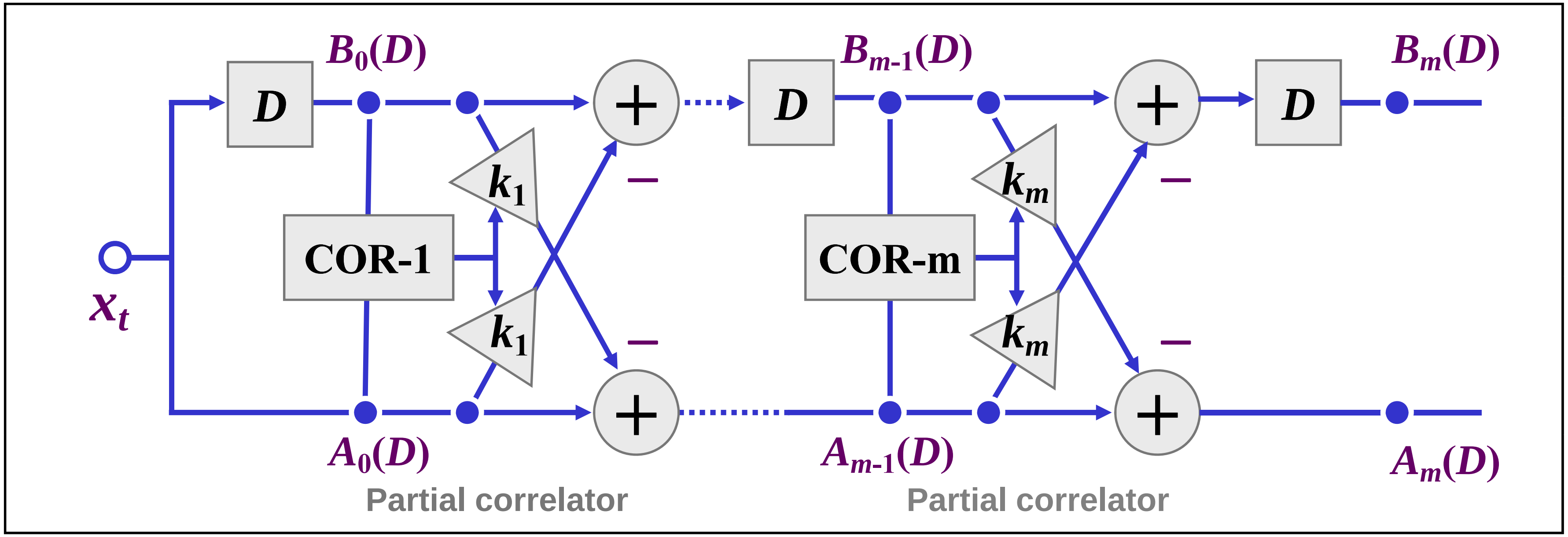




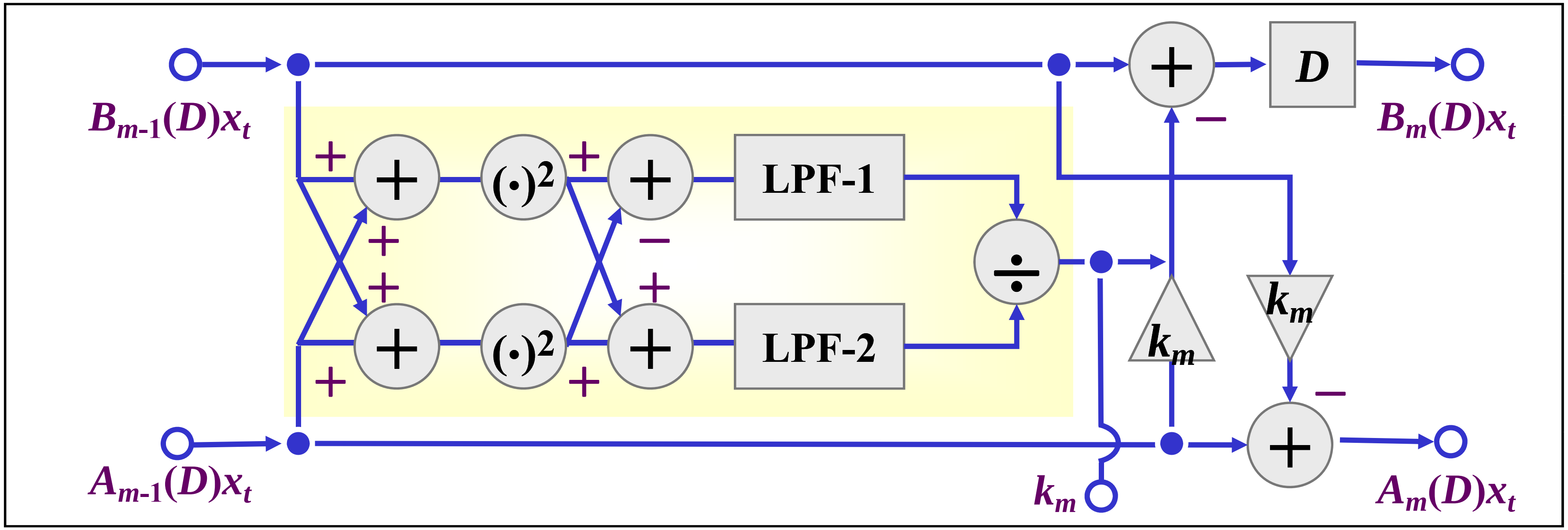
Definition of PARCOR coefficients

Flowchart for calculating $\{k_m\}_{i=0}^{p}$ and $\{\alpha_i\}_{i=0}^{p}$ from $\{v_i\}_{i=0}^{p}$

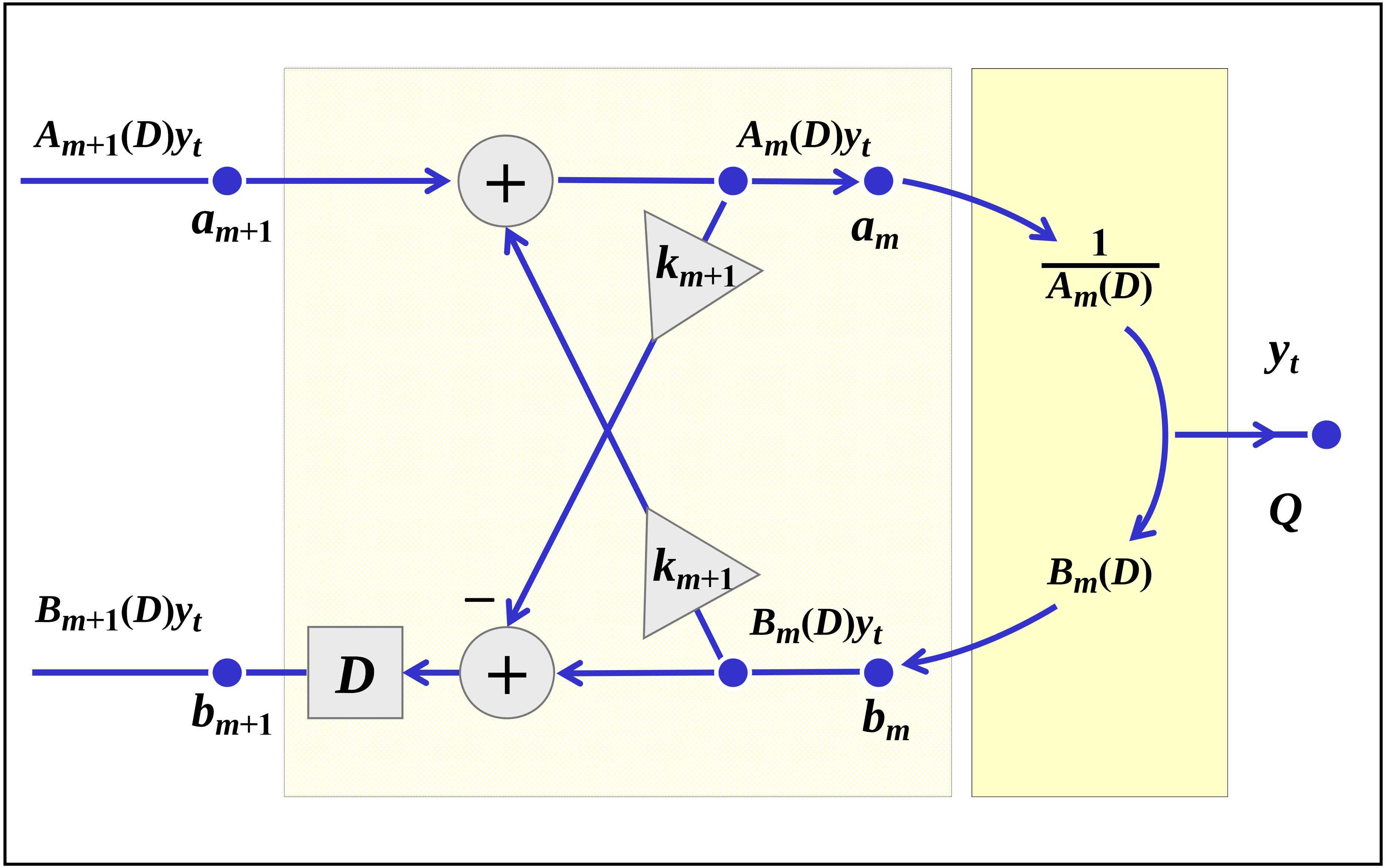




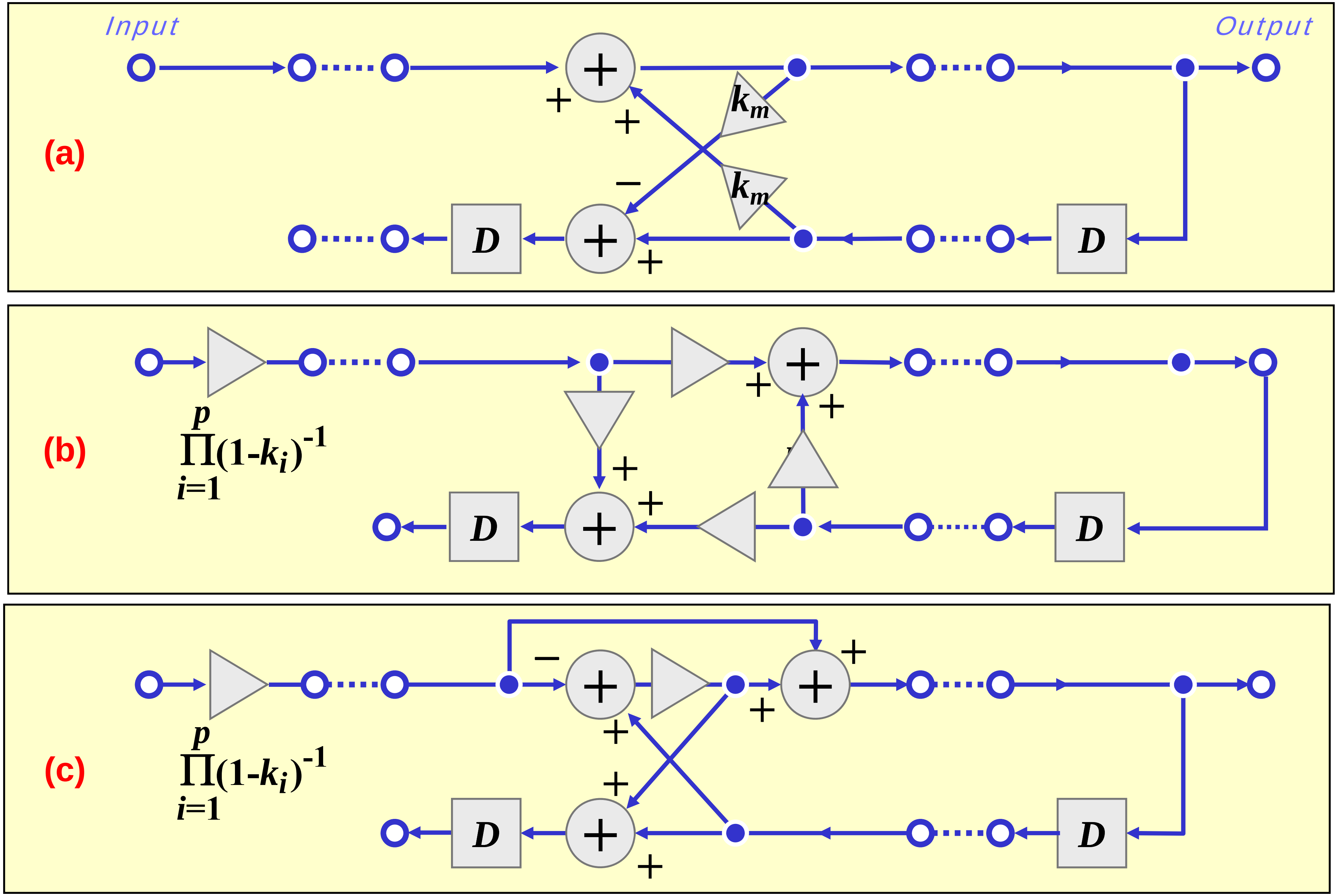
(a) PARCOR coefficient extraction circuit constructed by cascade connection of partial autocorrelators



(b) Construction of each partial autocorrelator

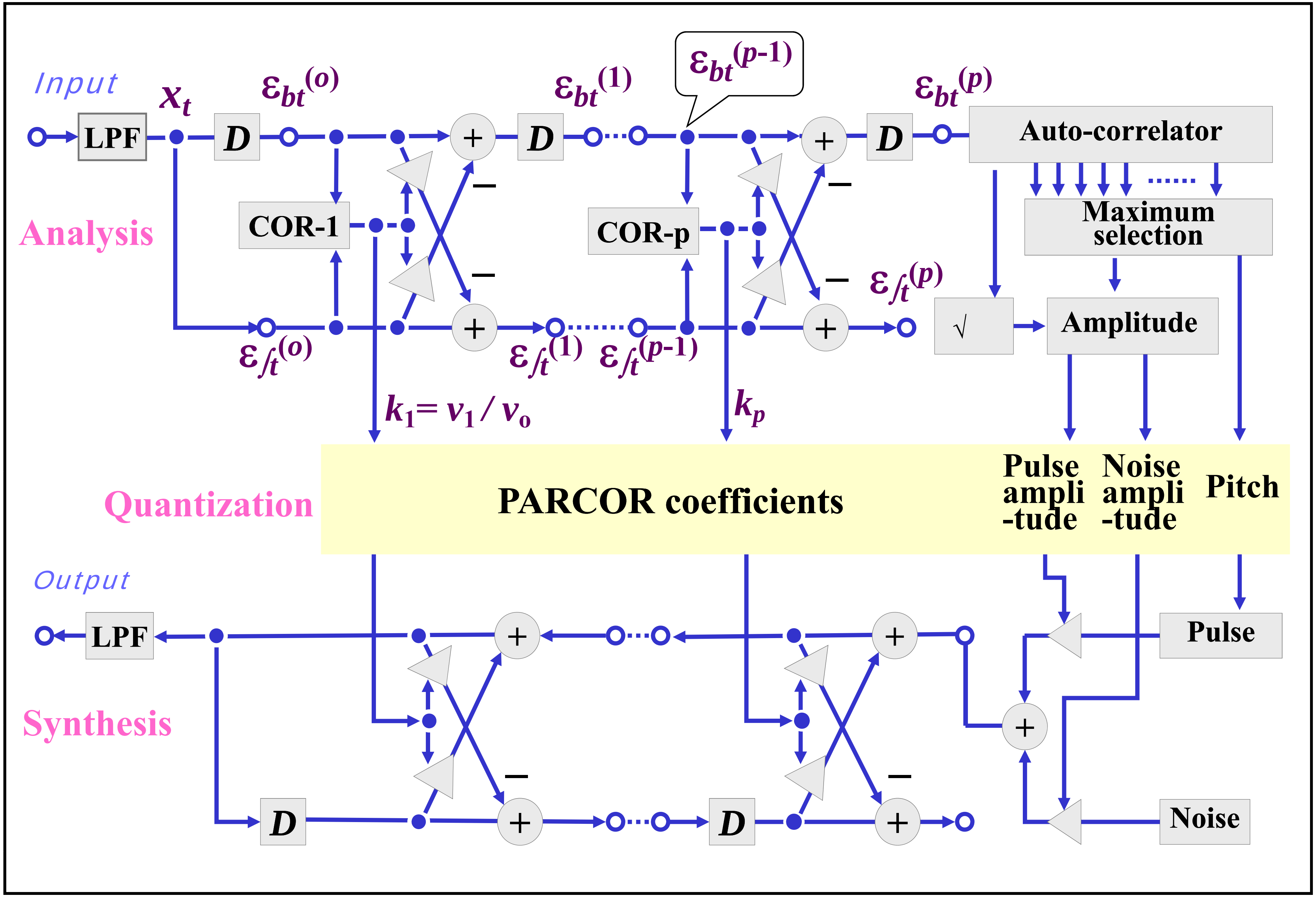


Principal construction features of synthesis filter using PARCOR coefficients

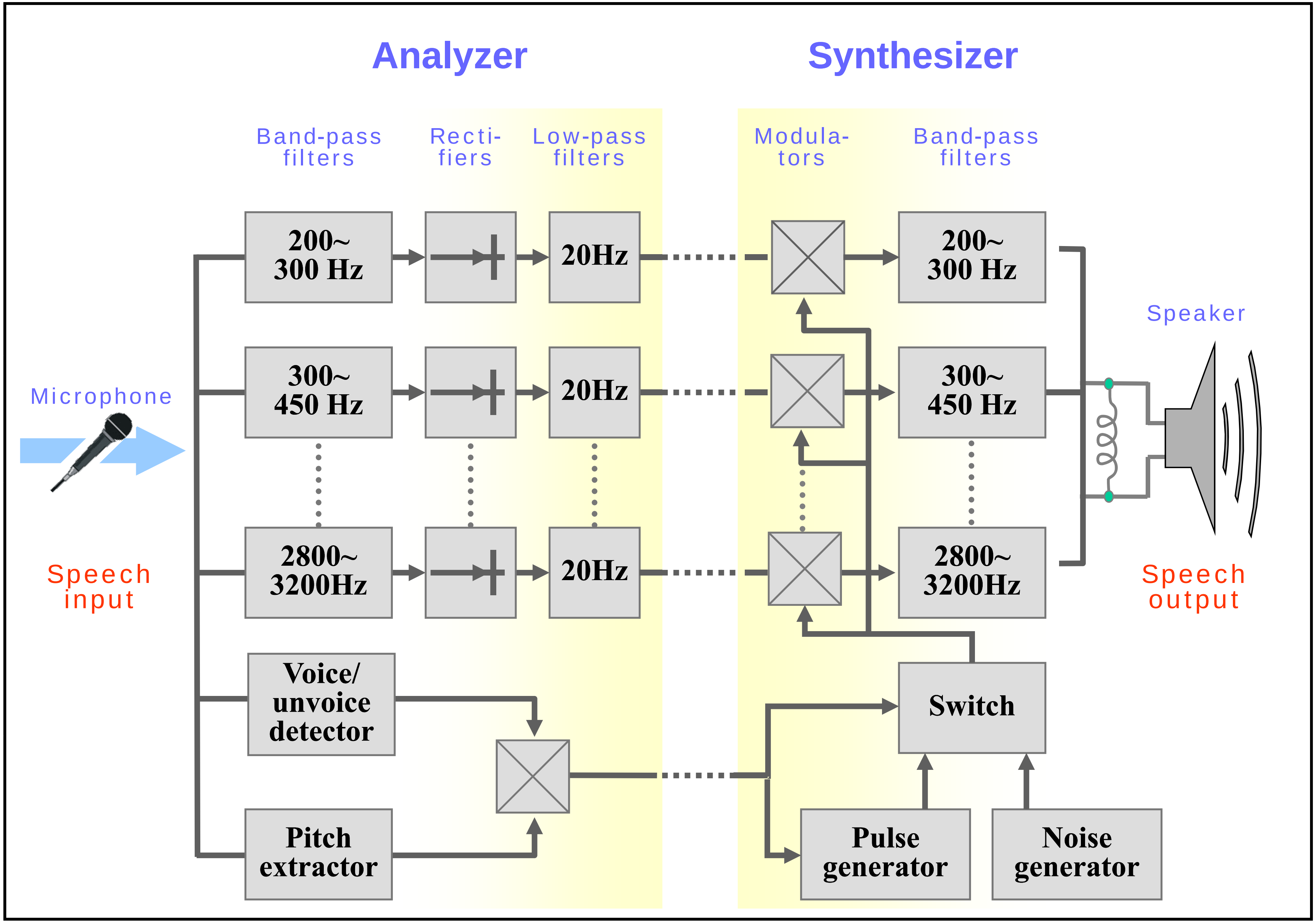


Equivalent transformations for lattice-type digital filter

0106-15

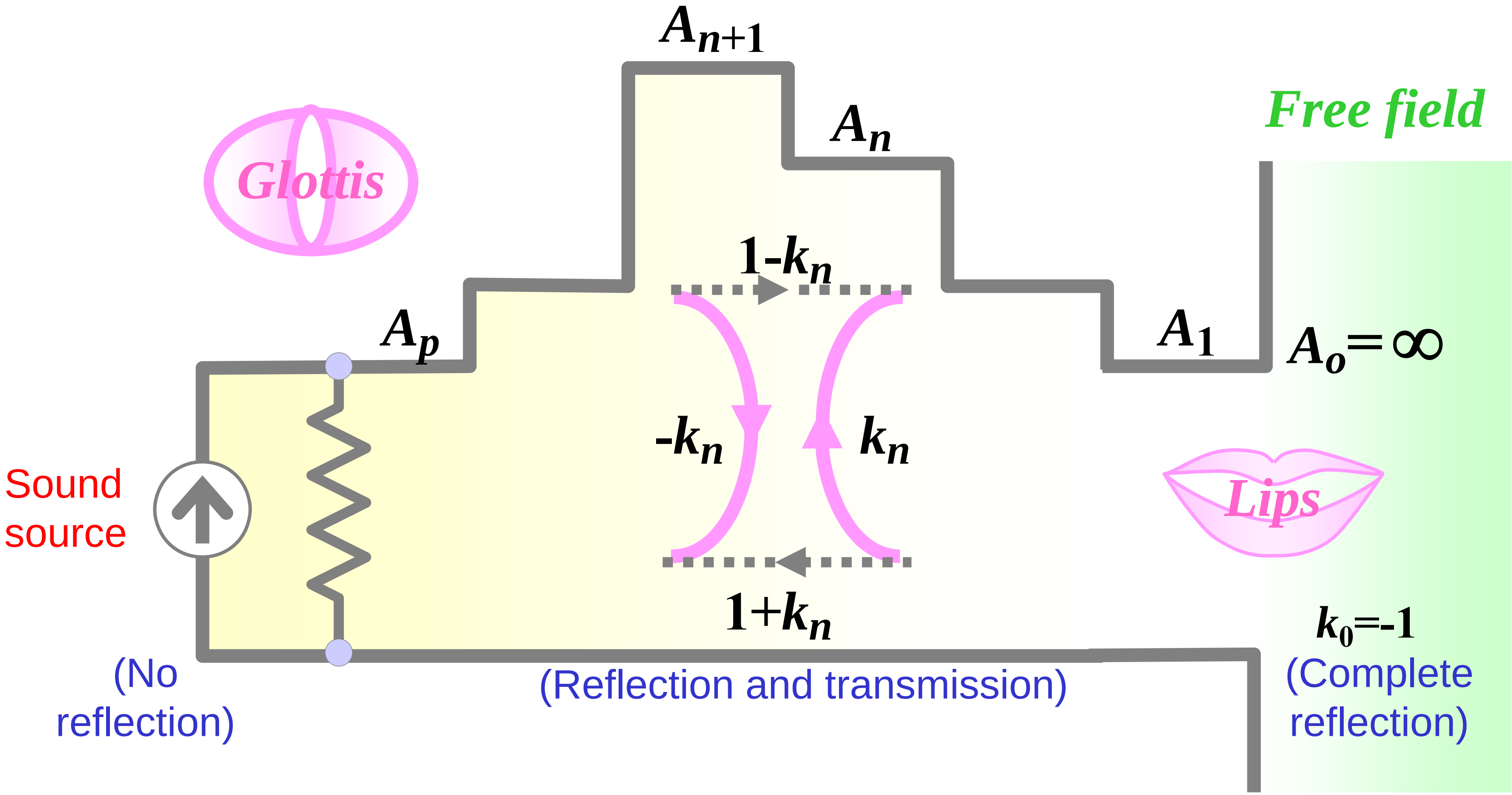


Structure of PARCOR analysis-synthesis system



Structure of the (channel) vocoder

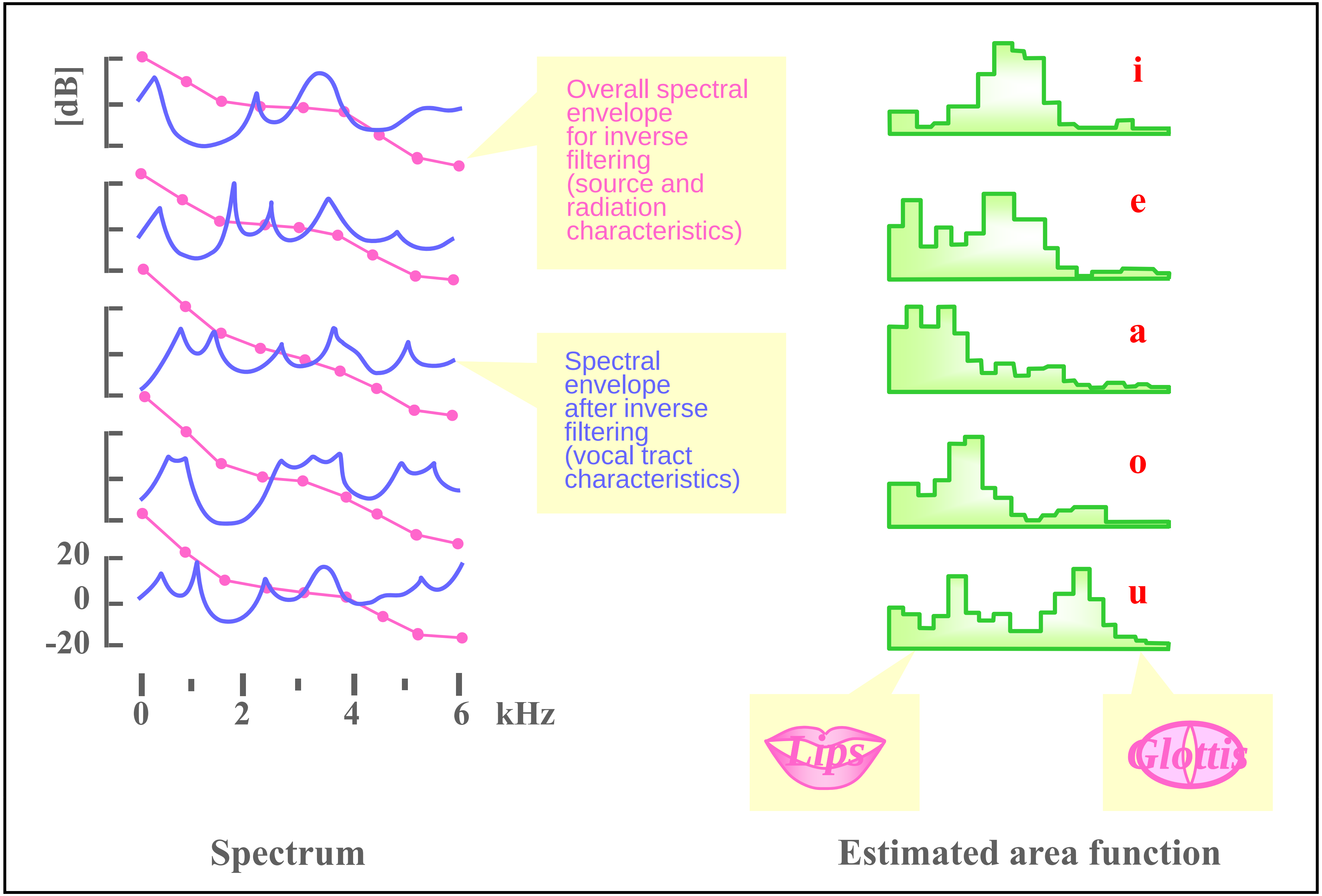
Relationship between PARCOR coefficients $\{k_n\}$ and vocal tract area function $\{A_n\}$

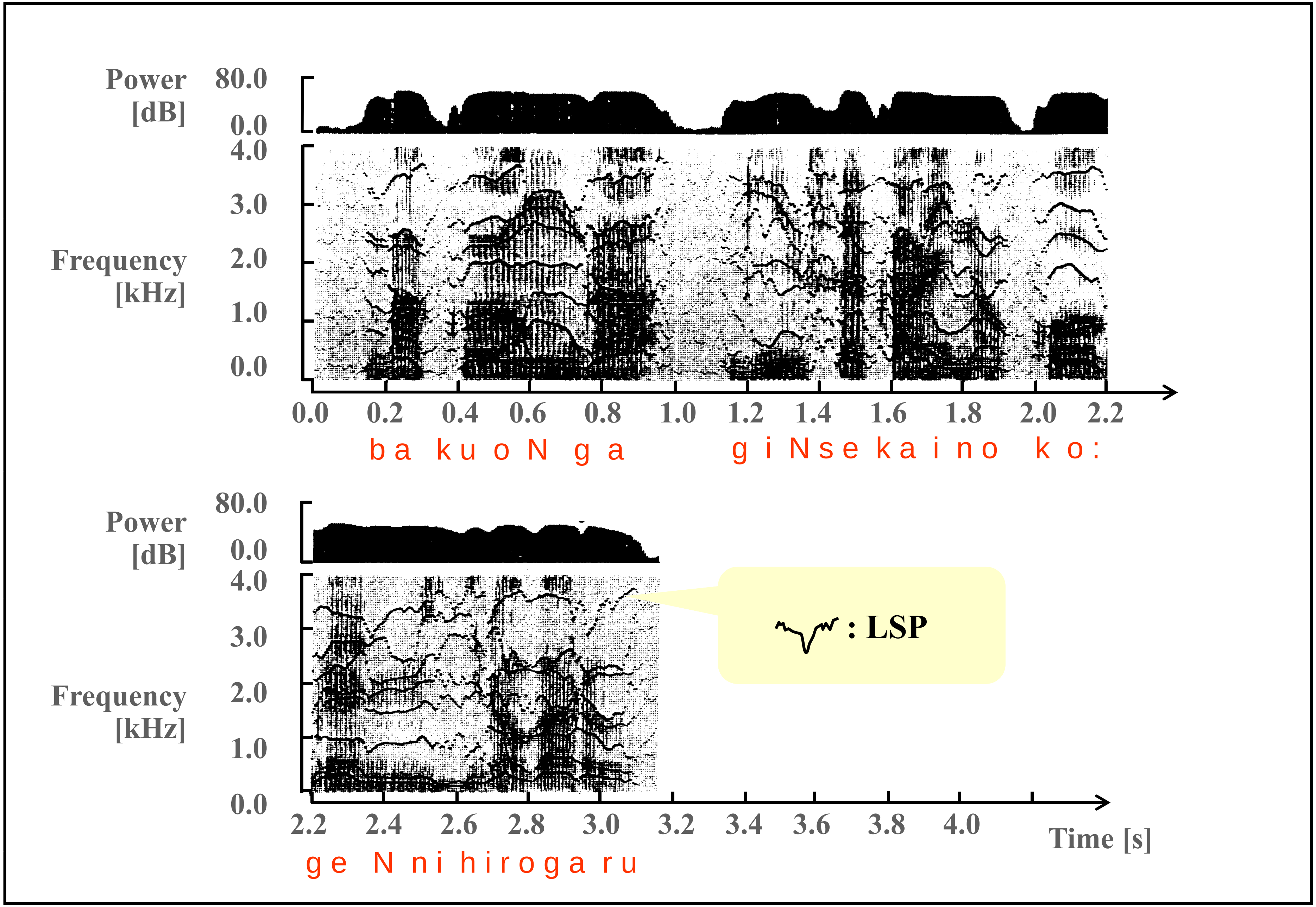


$$k_n = \frac{A_{n+1} - A_n}{A_{n+1} + A_n}$$

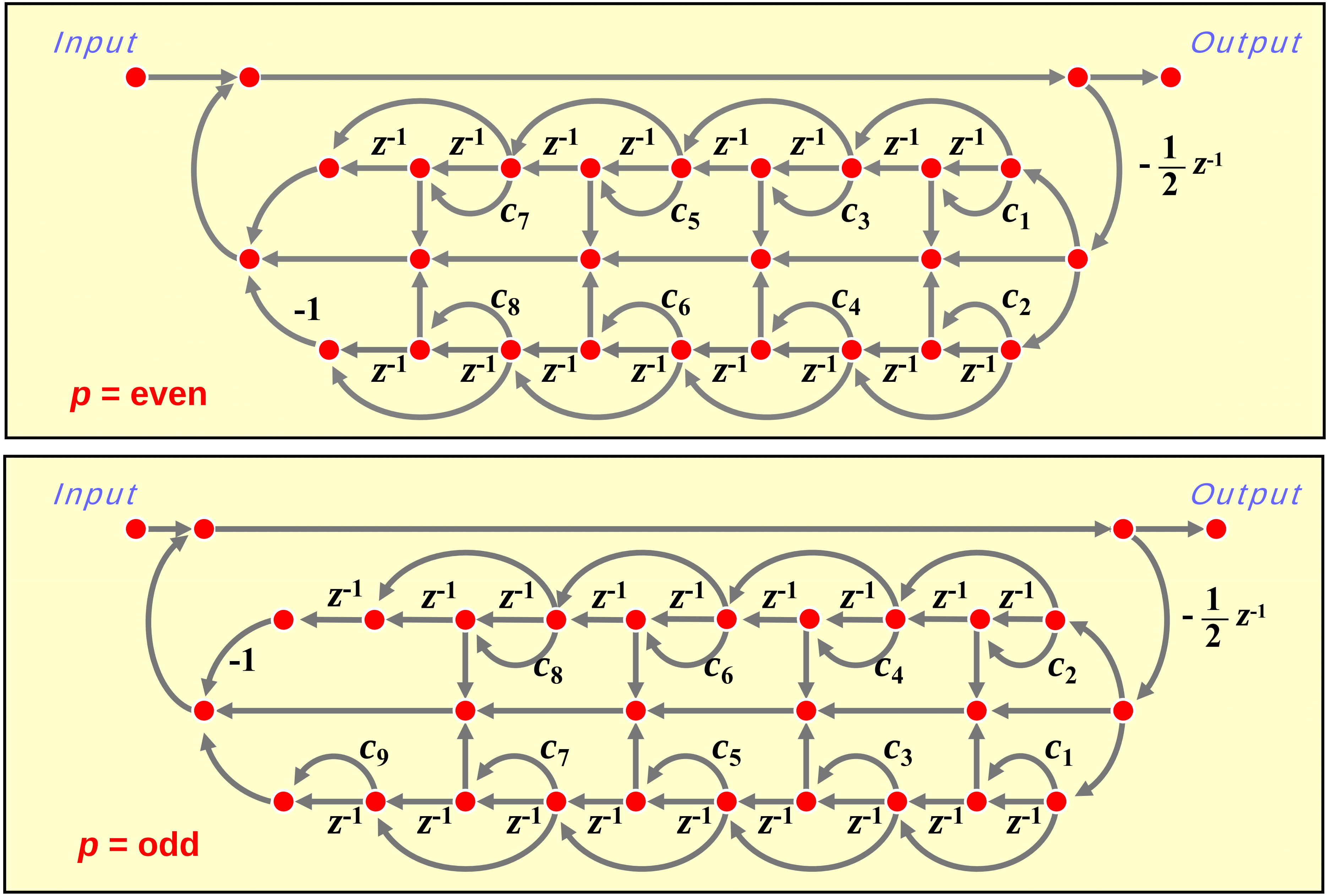
Examples of spectral envelopes and estimated area function for five vowels

0106-17

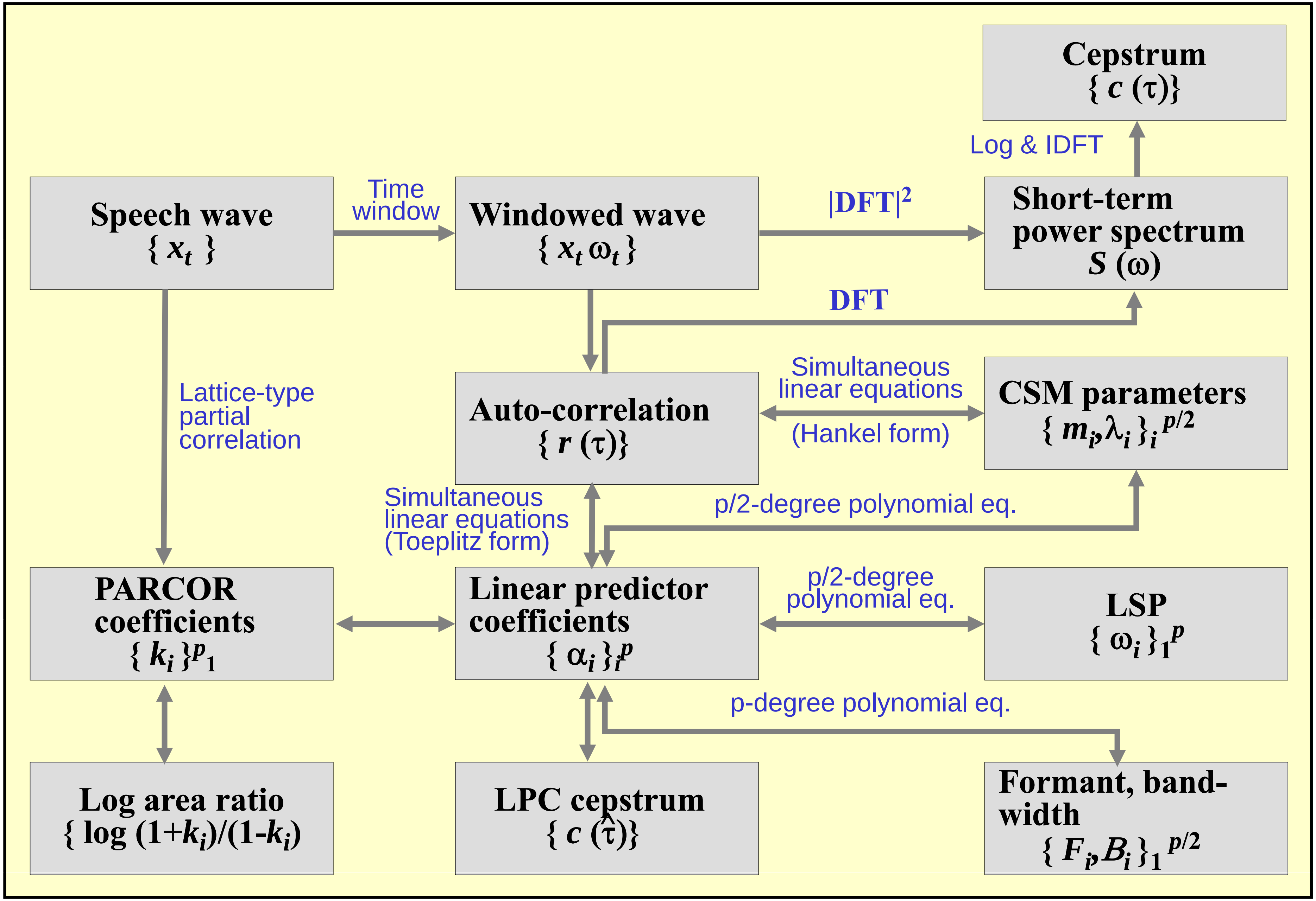




An example of LSP analysis



Signal flow graph of LSP synthesis filter



Mutual relationships between parameters based on all-pole spectrum modeling