Mel-Generalized Cepstral Representation of Speech
—A Unified Approach to Speech Spectral Estimation

Keiichi Tokuda

Nagoya Institute of Technology

Carnegie Mellon University

Tamkang University
March 13, 2002

Conventional Speech Spectral Estimation

• Linear prediction (LPC)	Autoregressive (AR) model
• Cepstral analysis	Exponential (EX) model
• Subband filter bank	Nonparametric

Variations

- Model ⇒ Pole-zero (ARMA) model
- Analysis window ⇒ Adaptive analysis (sample by sample basis)
- Auditory characteristics ⇒ Warped LPC, PLP, etc.
  - Auditory frequency scales (mel, Bark)
  - Loudness scales (log, sone)

Structure of This Talk

1. Conventional cepstral analysis

2. Introduction of generalized logarithmic function

   $\Rightarrow$ Generalized cepstral analysis

3. Introduction of auditory frequency scale

   $\Rightarrow$ Mel-generalized cepstral analysis

4. Applications to speech recognition and coding

History of Cepstral Analysis

- *B.P. Bogert, M.J.R. Healy, J.W. Tukey (1963)*
  Analysis of seismic signals
  — decomposition into direct wave and echo
  ⇒ Cepstrum, Quefrency, Lifter

- *A.M. Noll (1964, 1967)*
  Pitch extraction based on cepstrum

- *A.V. Oppenheim (1966, 1968)*
  Homomorphic deconvolution
  — decomposition into source and vocal tract function
  ⇒ Complex cepstrum

Definition of Cepstrum

Fourier transform of signal $s(n)$

$$S(e^{j\omega}) = \mathcal{F}\,[\,s(n)\,]$$

Cepstrum

$$C(m) = \mathcal{F}^{-1}\left[\,\log|S(e^{j\omega})|^2\,\right] \quad (\textit{Bogert et al., Noll})$$

$$C(m) = \mathcal{F}^{-1}\left[\,\log|S(e^{j\omega})|\,\right] \quad (\textit{Oppenheim})$$

Complex Cepstrum

z-transform of signal $s(n)$

$$S(z) = \mathcal{Z}\left[\, s(n) \,\right]$$

Complex cepstrum

$$\begin{aligned} c(m) &= \mathcal{Z}^{-1}\left[\, \log S(z) \,\right] \\ &= \mathcal{F}^{-1}\left[\, \log S(e^{j\omega}) \,\right] \\ &= \mathcal{F}^{-1}\left[\, \log |S(e^{j\omega})| + j \arg S(e^{j\omega}) \,\right] \end{aligned}$$

Cepstrum and Complex Cepstrum

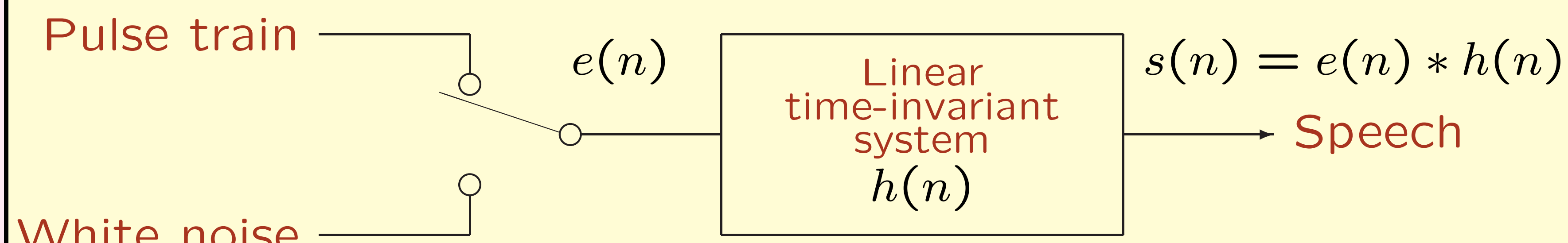
$$\log |S(e^{j\omega})| = \mathcal{F}[\,C(m)\,] = \text{Re}[\,\mathcal{F}[\,c(m)\,]\,]$$

$$\Downarrow$$

When it is minimum phase (all poles and zeros are located in the unit circle)

$$c(m) = \begin{cases} 0, & m < 0 \\ C(m), & m = 0 \\ 2C(m), & m > 0 \end{cases}$$

Homomorphic Deconvolution



$$s(n) = h(n) * e(n)$$

$$\downarrow \quad \mathcal{F}$$

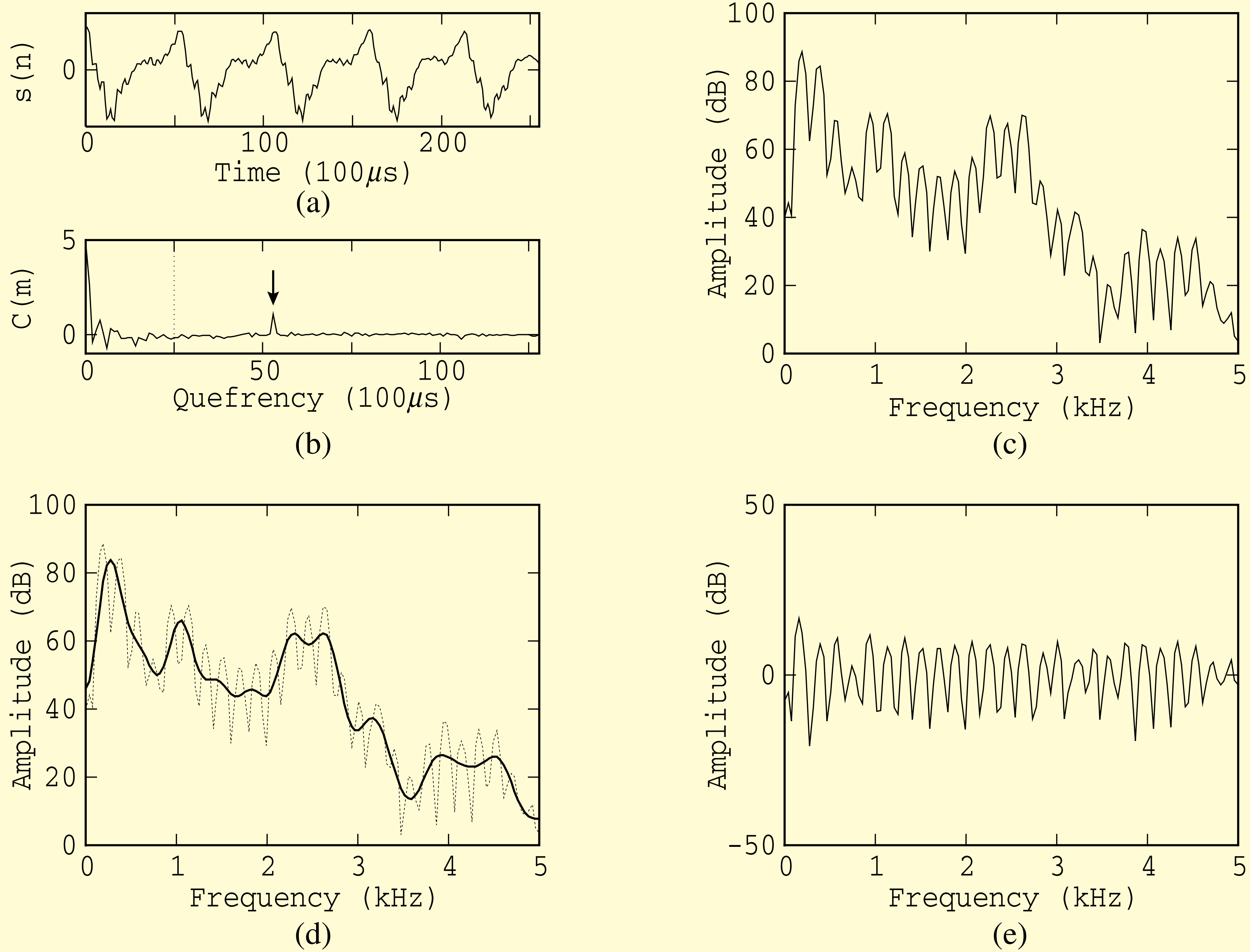
$$S(e^{j\omega}) = H(e^{j\omega})E(e^{j\omega})$$

$$\downarrow \quad \log|\cdot|$$

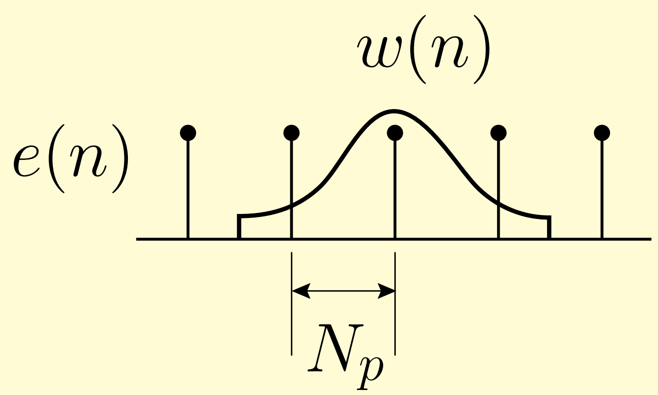
$$\log|S(e^{j\omega})| = \log|H(e^{j\omega})| + \log|E(e^{j\omega})|$$

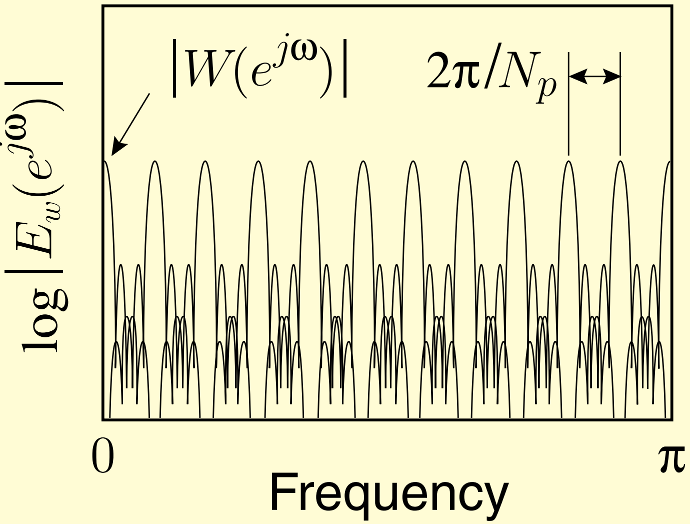
$$\downarrow \quad \mathcal{F}^{-1}$$

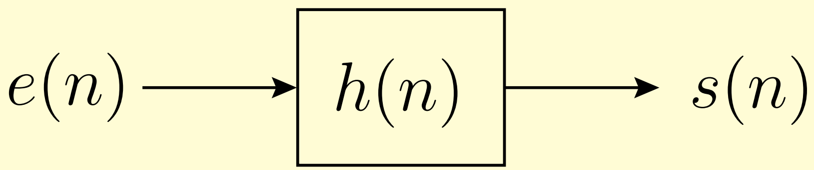
$$C(m) = C_h(m) + C_e(m)$$

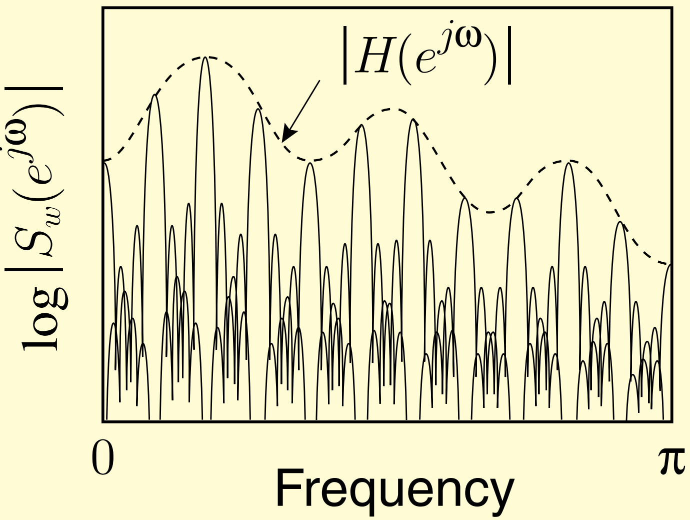

s(n)
0
0
100
200
Time (100μs)
(a)
5
C(m)
0
0
50
100
Quefrency (100μs)
(b)
100
80
60
40
20
0
Amplitude (dB)
0
1
2
3
4
5
Frequency (kHz)
(c)
100
80
60
40
20
0
Amplitude (dB)
0
1
2
3
4
5
Frequency (kHz)
(d)
50
0
−50
Amplitude (dB)
0
1
2
3
4
5
Frequency (kHz)
(e)

Spectrum of Periodic Signal









Problems of Homomorphic Processing (Cepstral Analysis)

Linear smoothing of log spectrum

- affected by fine structure of FFT spectrum
- results in a large bias and variance

Voiced speech (periodic)

- Envelope of peaks of spectral fine structure
  $\Rightarrow$ Improved cepstral analysis , PSE: Biased

Cost Function

$P(\omega)$: Estimate of Power Spectrum

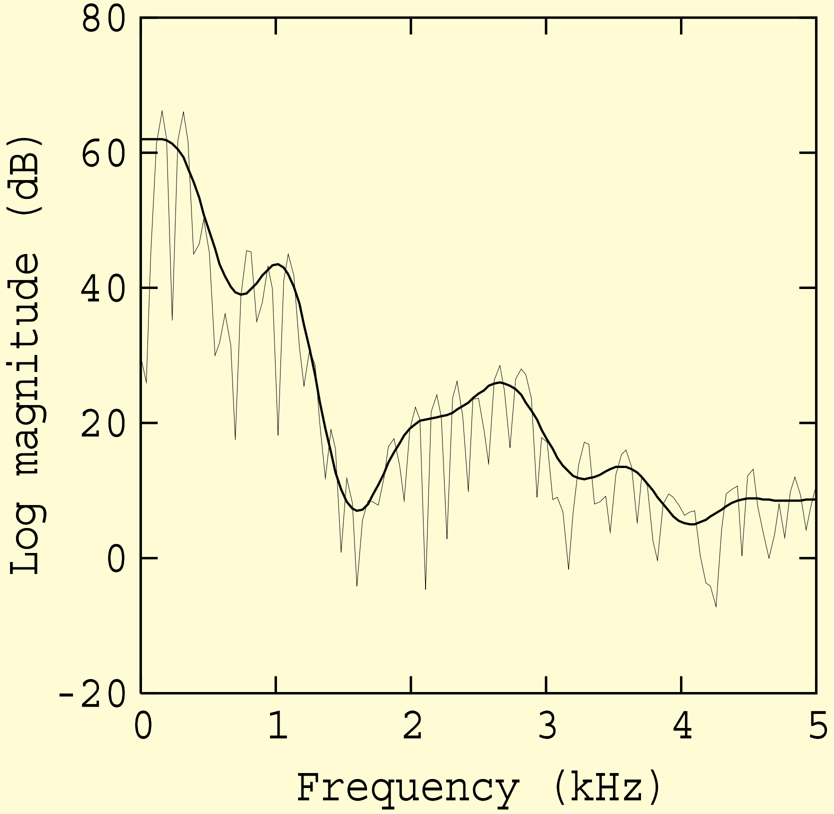
$I_N(\omega)$: Periodogram

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{\frac{I_N(\omega)}{P(\omega)} - \log\frac{I_N(\omega)}{P(\omega)} - 1\right\} d\omega \Rightarrow \min$$

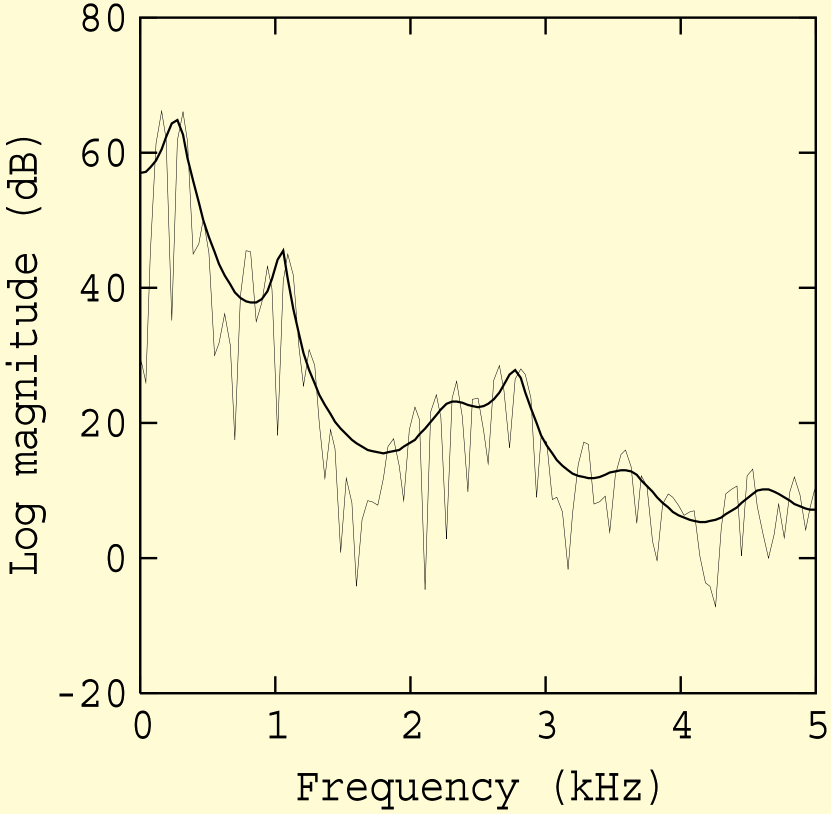
$\boldsymbol{x}$: Gaussian Process $\Rightarrow$ Maximizing $p(\boldsymbol{x}|\boldsymbol{c})$

- Unbiased estimation of log spectrum
- equivalent to one used in LPC
- Minimization of energy of inverse filter output

Analysis of Natural Speech



(a) Unbiased cepstral analysis



(b) Linear prediction

Generalized Cepstrum

Complex Cepstrum

$$\begin{aligned} c(m) &= \mathcal{Z}^{-1}[\log S(z)] \\ \log S(z) &= \mathcal{Z}[c(m)] \end{aligned}$$

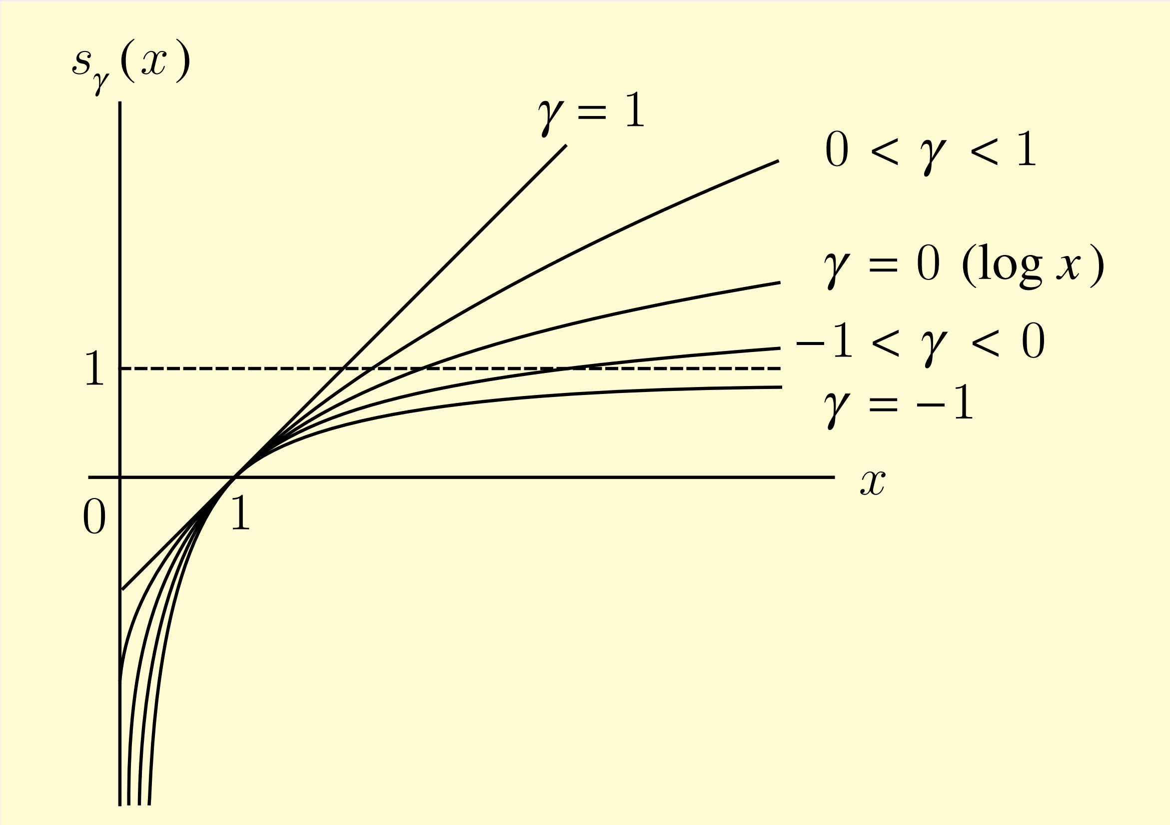
$\Downarrow$

Generalized Cepstrum

$$\begin{aligned} c_\gamma(m) &= \mathcal{Z}^{-1}[s_\gamma(S(z))] \\ s_\gamma(S(z)) &= \mathcal{Z}[c_\gamma(m)] \end{aligned}$$

Generalized logarithmic function

$$s_\gamma(w) = \begin{cases} (w^\gamma - 1)/\gamma, & 0 < |\gamma| \le 1 \\ \log w, & \gamma = 0 \end{cases}$$



Spectral Model

Generalized Cepstrum: $c_\gamma(m)$

$$H(z) = s_\gamma^{-1}\left(\sum_{m=0}^{M} c_\gamma(m)\, z^{-m}\right)$$

$$= \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_\gamma(m)\, z^{-m}\right)^{1/\gamma}, & 0 < |\gamma| \le 1 \\ \exp \sum_{m=0}^{M} c_\gamma(m)\, z^{-m}, & \gamma = 0 \end{cases}$$

Inverse function of Generalized logarithm

$$s_\gamma^{-1}(w) = \begin{cases} (1 + \gamma w)^{1/\gamma}, & 0 < |\gamma| \le 1 \\ \exp w, & \gamma = 0 \end{cases}$$

Cost Function

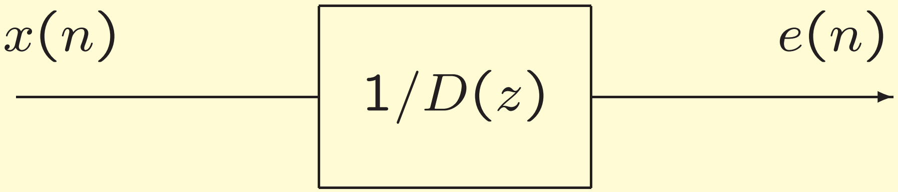
$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{\frac{I_N(\omega)}{P(\omega)} - \log\frac{I_N(\omega)}{P(\omega)} - 1\right\} d\omega \Rightarrow \min$$

Estimate of Power Spectrum

$$P(\omega) = |H(e^{j\omega})|^2 = \sigma^2 |D(e^{j\omega})|^2$$

Interpretation in time-domain

$$\varepsilon = E\left[e^2(n)\right] \Rightarrow \min$$



Advantage

$-1 \leq \gamma \leq 0$:

- Convex function $\Rightarrow$ Global solutioncan easily be obtained
- The obtained system $H(z)$ is minimum phase, e.g., stable

- $\gamma = -1 \Rightarrow$ Linear Prediction

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_\gamma(m) z^{-m}}$$

- $\gamma = 0 \Rightarrow$ Cepstrum

$$H(z) = \exp \sum_{m=0}^{M} c_\gamma(m) z^{-m}$$

Prediction Gain

- $D(z)$ is minimum phase
- Gain of $D(z)$ is one

$\Rightarrow$

Predictor:

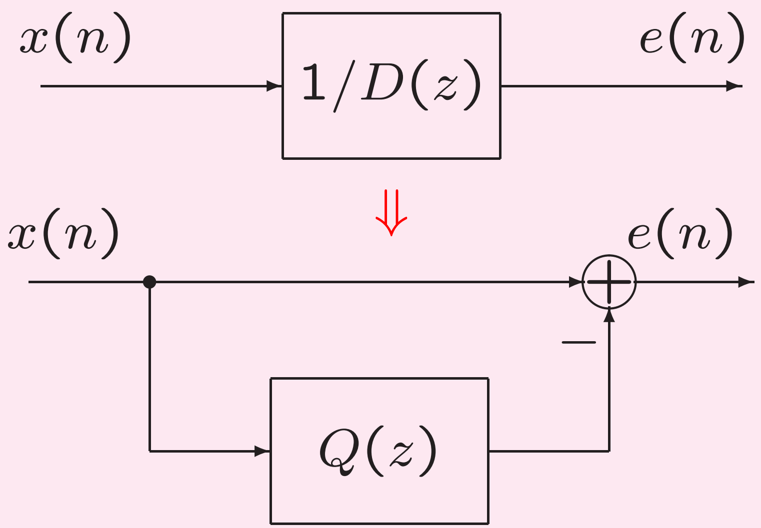
$$Q(z) = \sum_{k=1}^{\infty} a(k) z^{-k}$$

Cost Function:

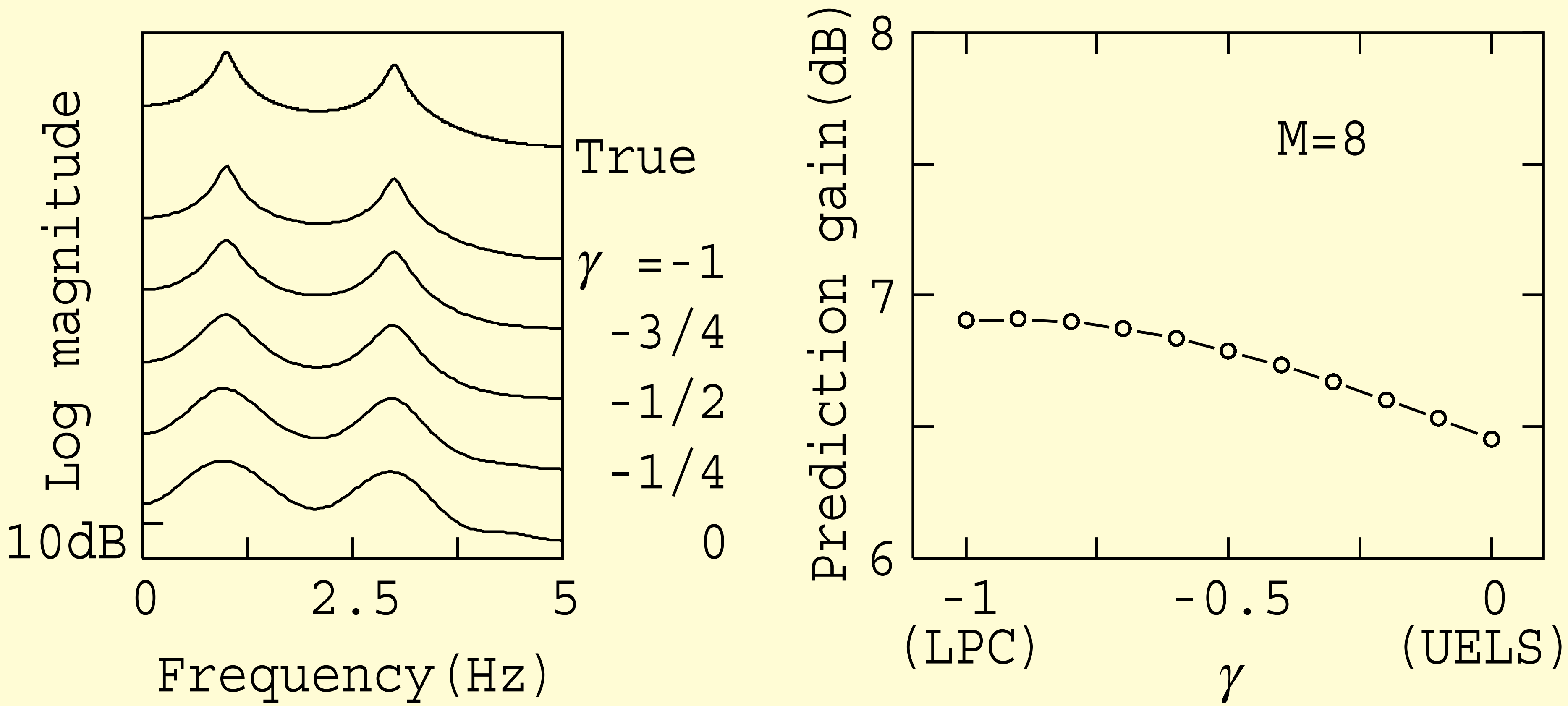
$$\varepsilon = E\left[e^2(n)\right]$$

$\Rightarrow$ Prediction Gain:

$$G = \frac{E\left[x^2(n)\right]}{E\left[e^2(n)\right]}$$

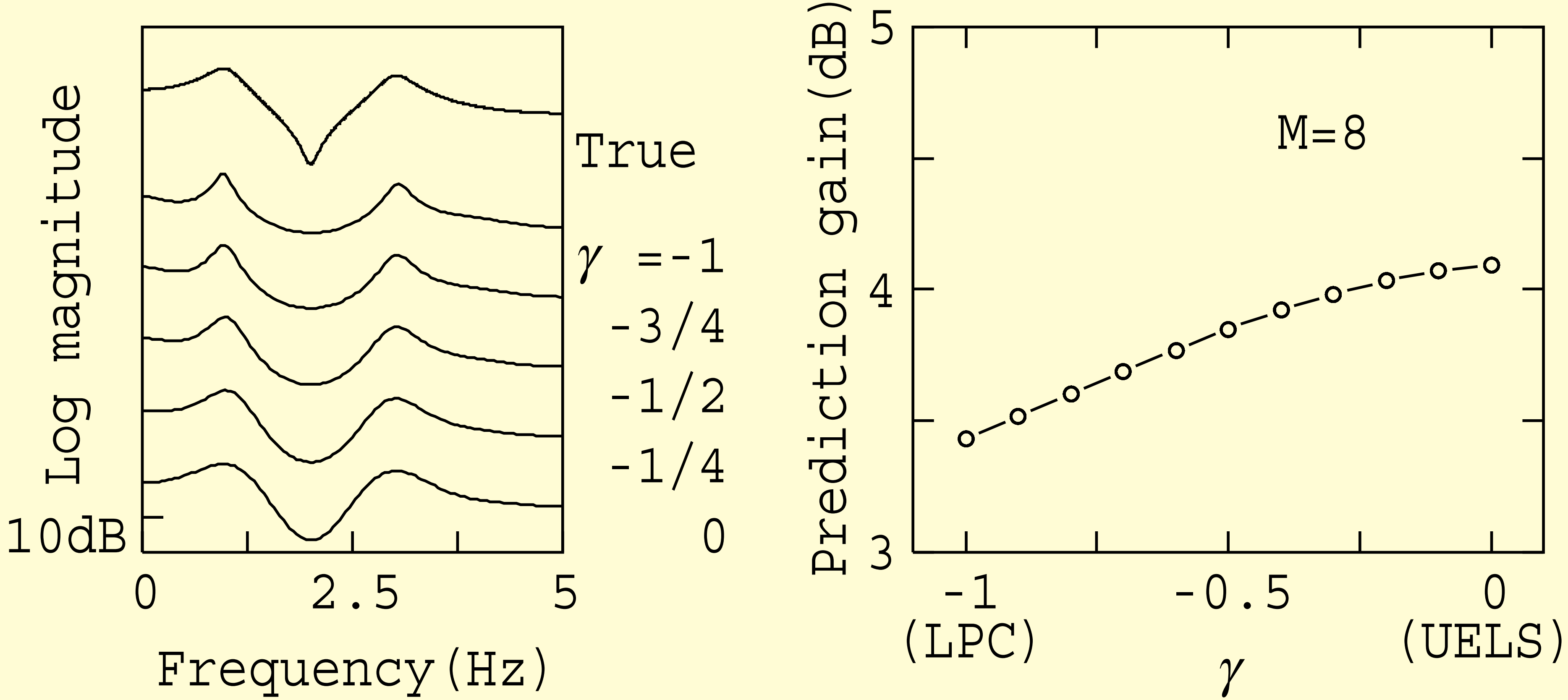


Analysis of synthetic signal (Generalized Cepstral Analysis)



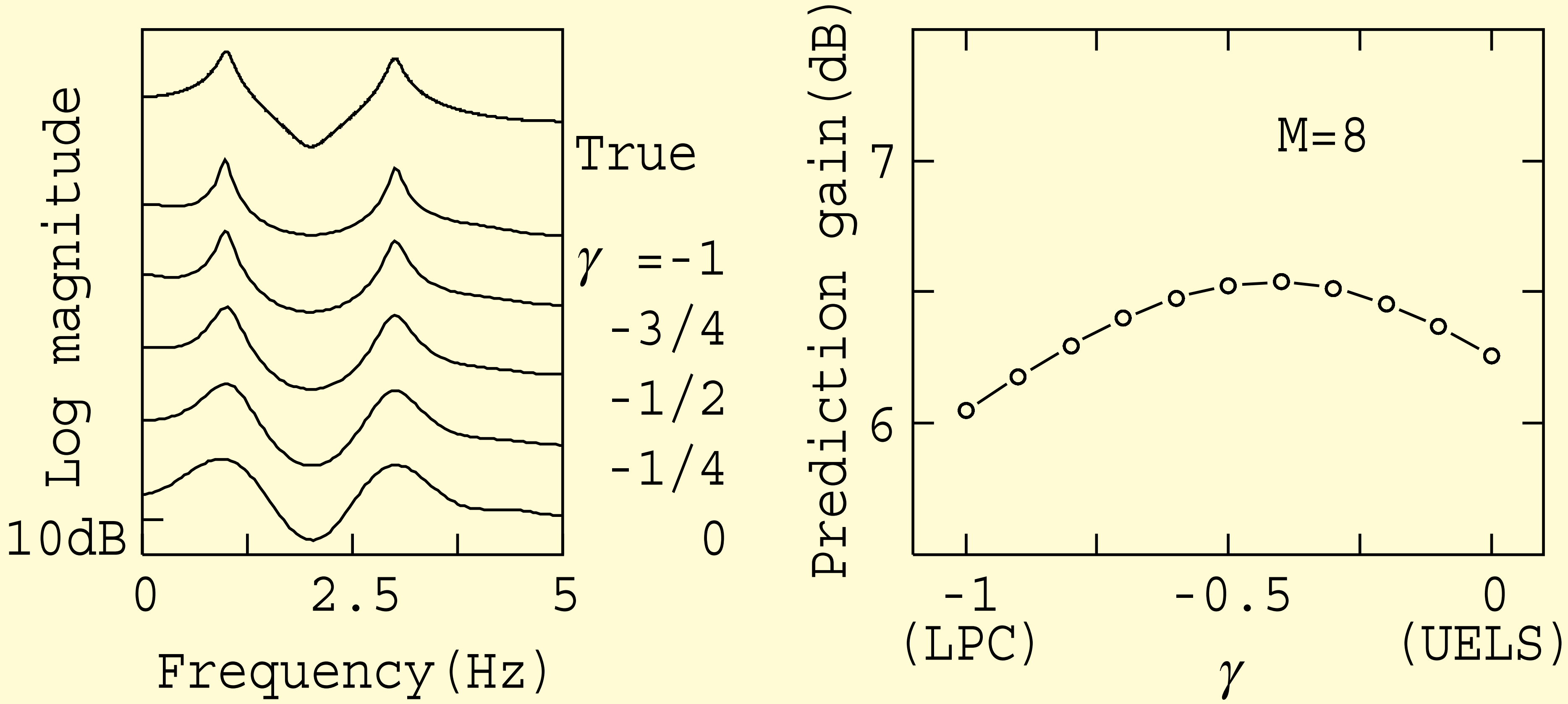
(a) Example 1

Analysis of synthetic signal (Generalized Cepstral Analysis)



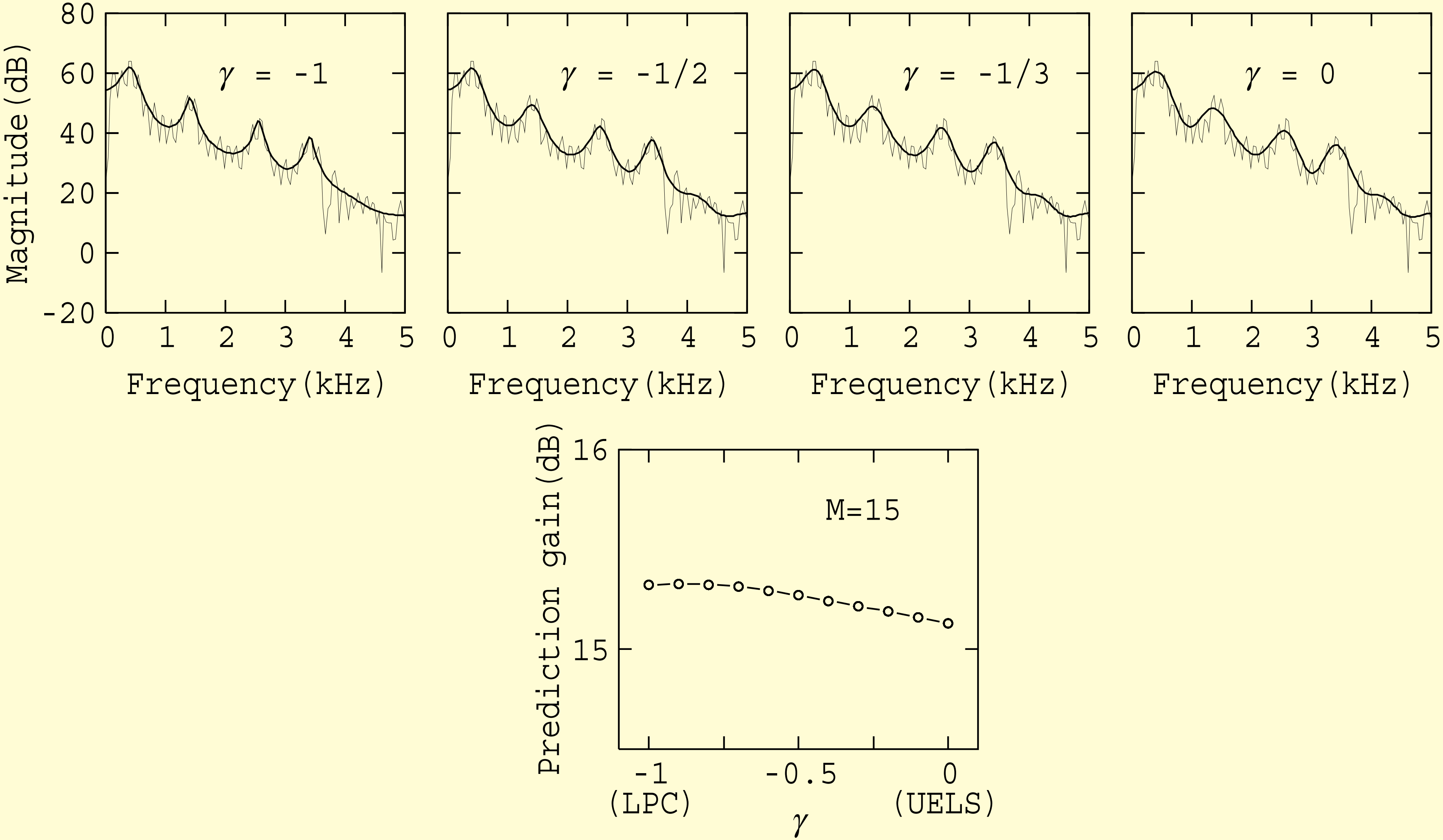
(c) Example 3

Analysis of synthetic signal (Generalized Cepstral Analysis)



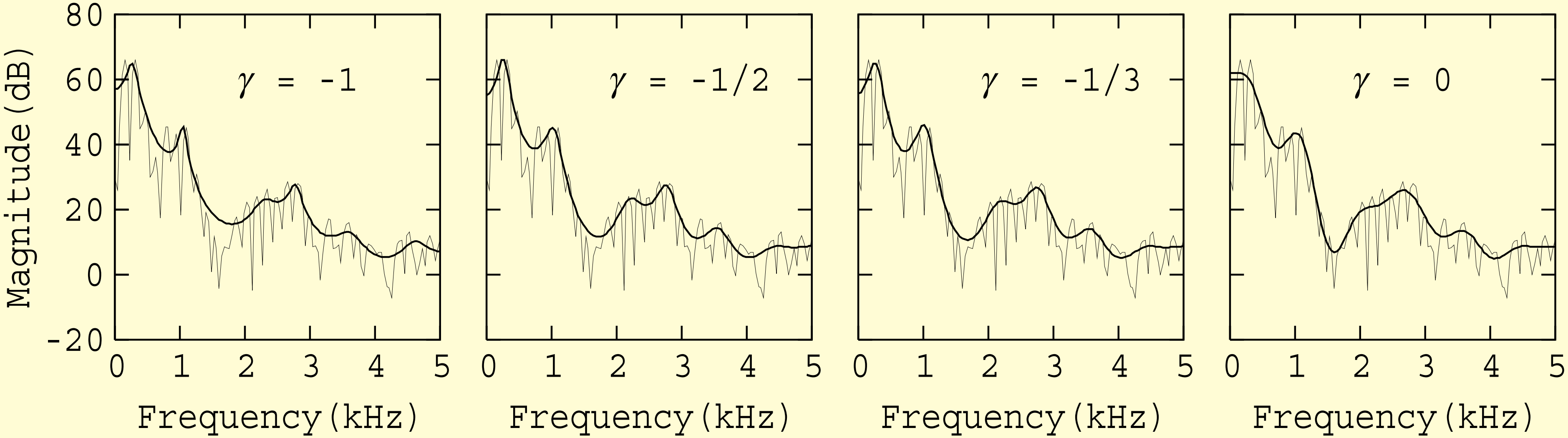
(b) Example 2

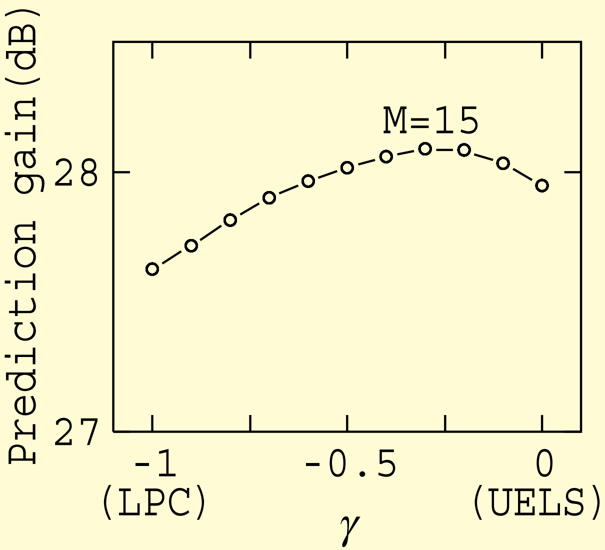
Analysis of natural speech (Generalized Cepstral Analysis) /e/



(a) male /e/

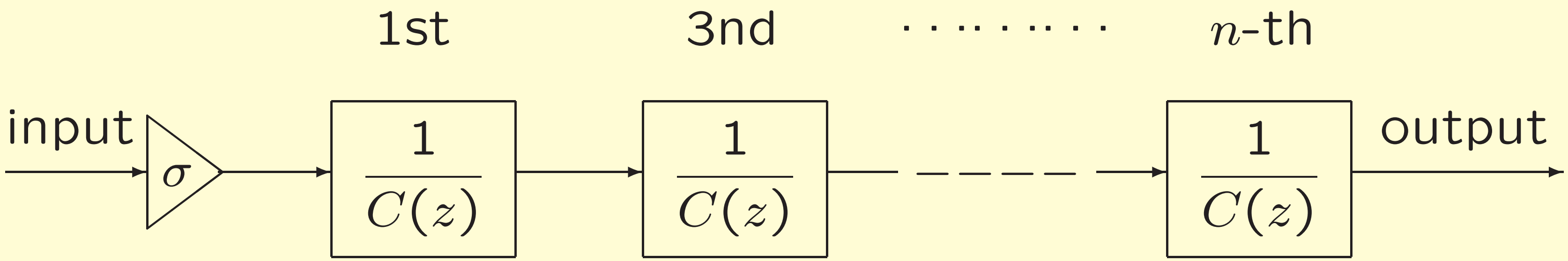
Analysis of natural speech (Generalized Cepstral Analysis) /N/





(b) male /N/

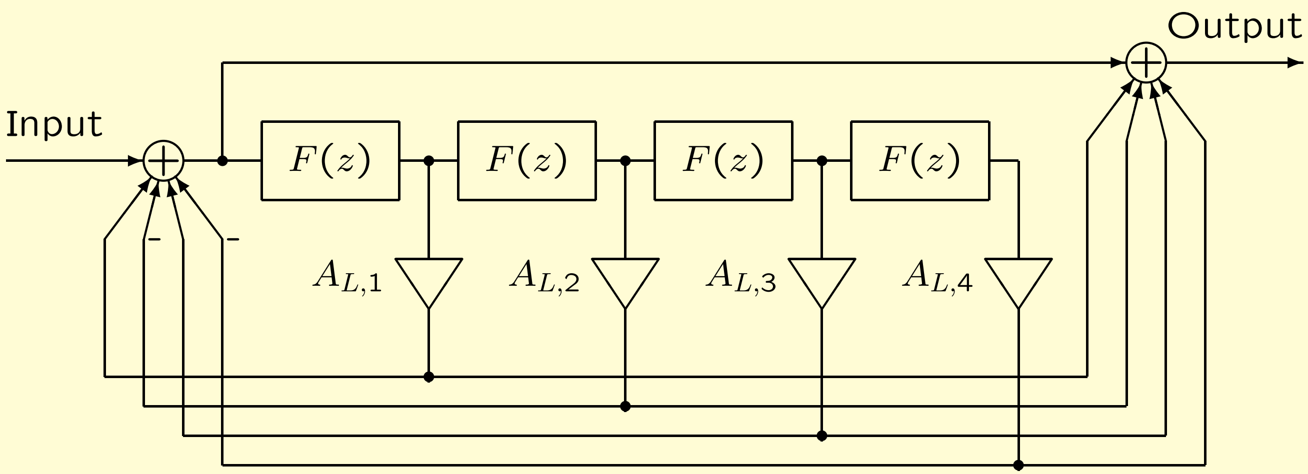
Structure of synthesis filter $H(z)$ ($\gamma = -1/n$)



$$H(z) = \sigma D(z) = \sigma \left\{ \frac{1}{C(z)} \right\}^{n}$$

$$C(z) = \left(1 + \gamma \sum_{m=0}^{M} c'_{\gamma}(m)\, z^{-m} \right)$$

Structure of synthesis filter $H(z)$ ($\gamma = 0$) —LMA filter



$$D(z) = \exp F(z) \simeq R_L(F(z)) = \frac{1 + \sum_{l=1}^{L} A_{L,l} \{F(z)\}^l}{1 + \sum_{l=1}^{L} A_{L,l} \{-F(z)\}^l}$$

$$F(z) = \sum_{m=1}^{M} c_\gamma(m)\, z^{-m}$$

Introduction of auditory frequency scale

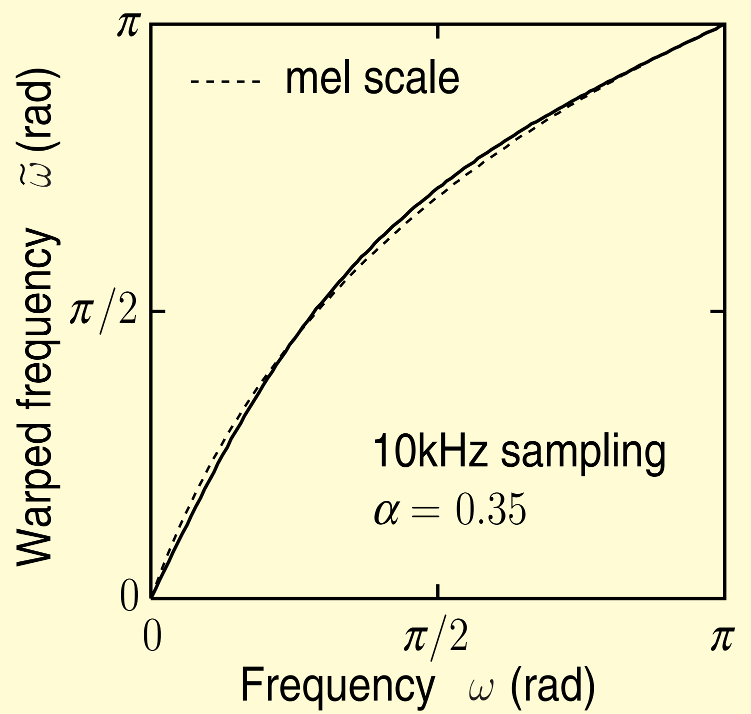
First-order all-pass function:

$$z_\alpha^{-1} = \Psi(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}$$

Phase Characteristics can be used for Frequency Transformation:

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}$$

where $\Psi(e^{j\omega}) = e^{-j\tilde{\omega}}$



Mel-Generalized Cepstral Analysis

Mel-generalized cepstrum: $c_{\alpha,\gamma}(m)$

$$H(z) = s_\gamma^{-1}\left(\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_\alpha^{-m}\right)$$

$$= \begin{cases} \left(1+\gamma\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_\alpha^{-m}\right)^{1/\gamma}, & 0<|\gamma|\le 1 \\ \exp\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_\alpha^{-m}, & \gamma = 0 \end{cases}$$

$$z_\alpha^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}$$

- $(\alpha, \gamma) = (0, 0) \Rightarrow$ Cepstral model:

$$H(z) = \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z^{-m}$$

- $(\alpha, \gamma) = (0, -1) \Rightarrow$ AR model:

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z^{-m}}$$

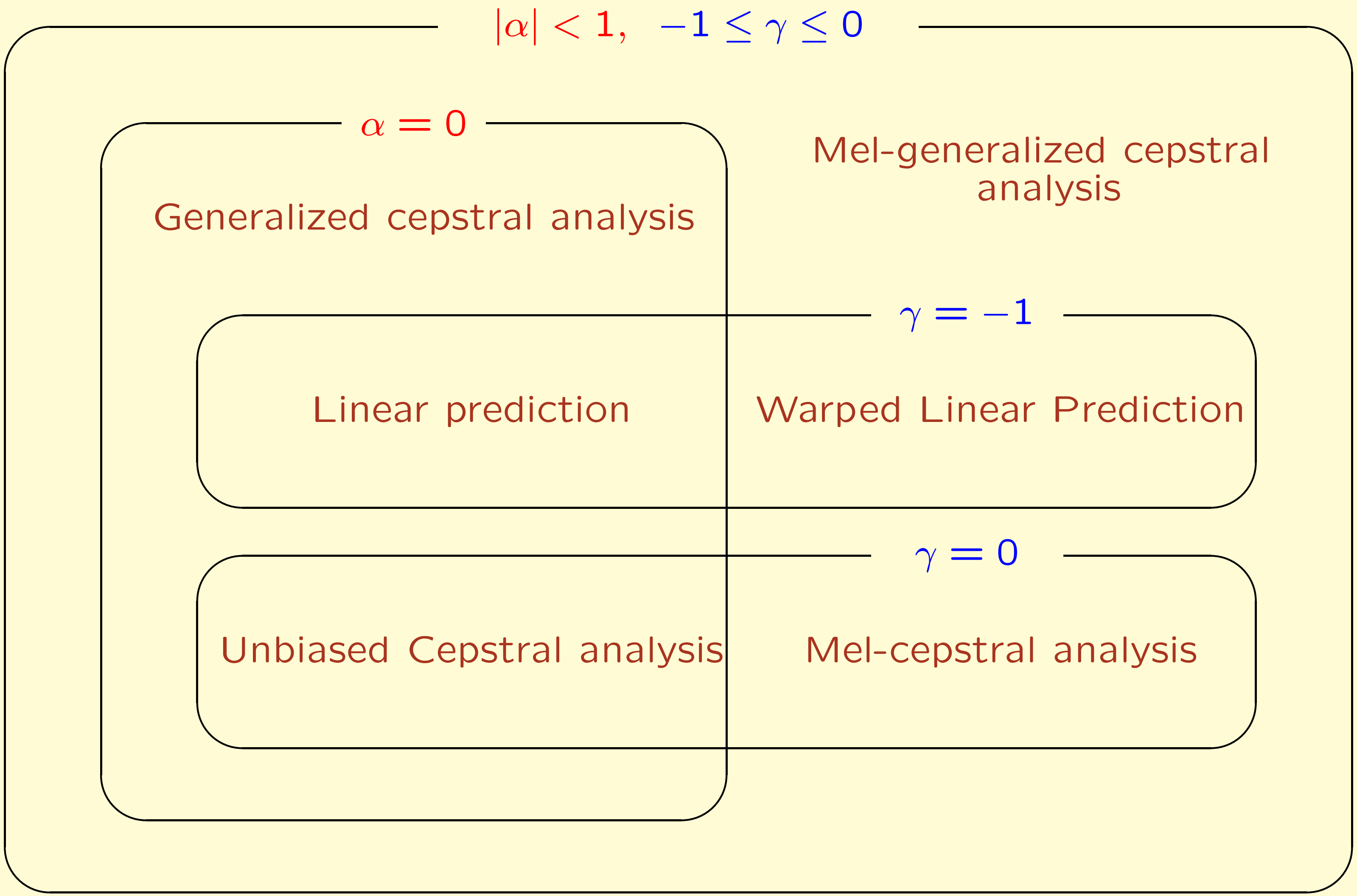
- $(\alpha, \gamma) = (0.35, 0) \Rightarrow$ Mel-cepstral model:

$$H(z) = \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_{\alpha}^{-m}$$

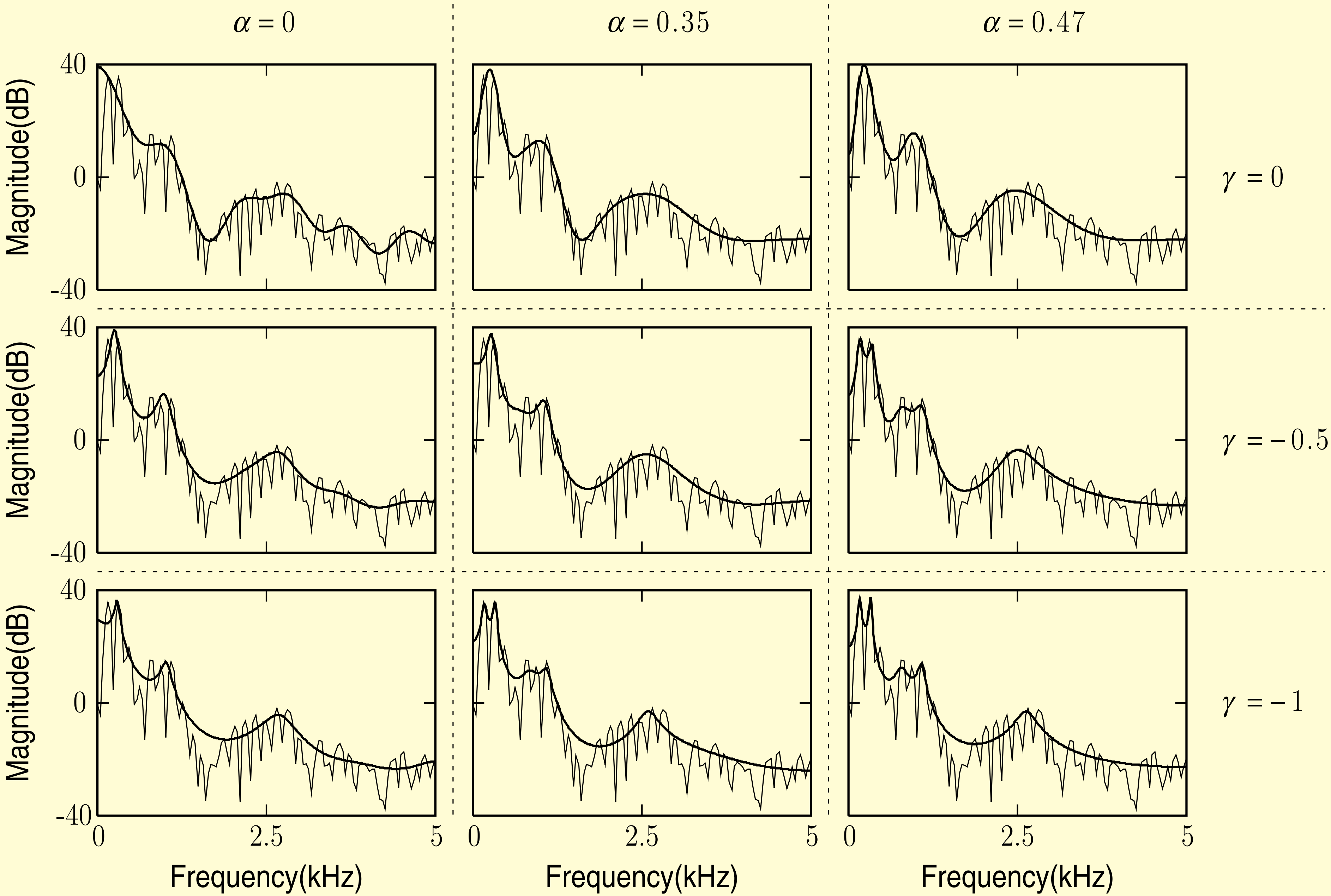
- $(\alpha, \gamma) = (0.47, -1) \Rightarrow$ Warped AR model:

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_{\alpha}^{-m}}$$

A Unified Approach to Speech Spectral Estimation



Mel-generalized analysis of natural speech /N/ $M = 12$



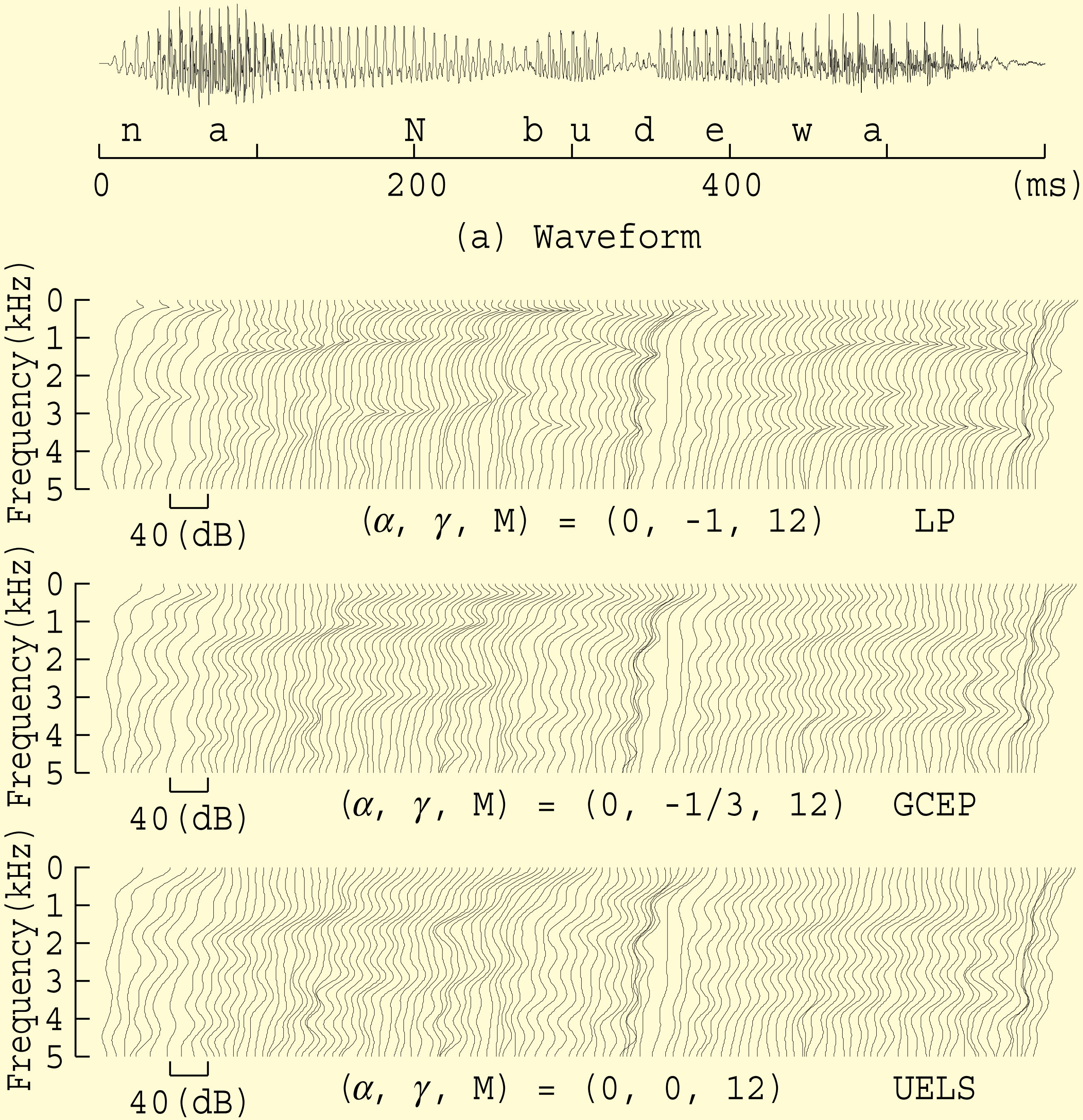
Example

$\alpha = 0$

$\gamma = -1$

$\gamma = -1/3$

$\gamma = 0$



(b) Spectral estimates (α = 0)

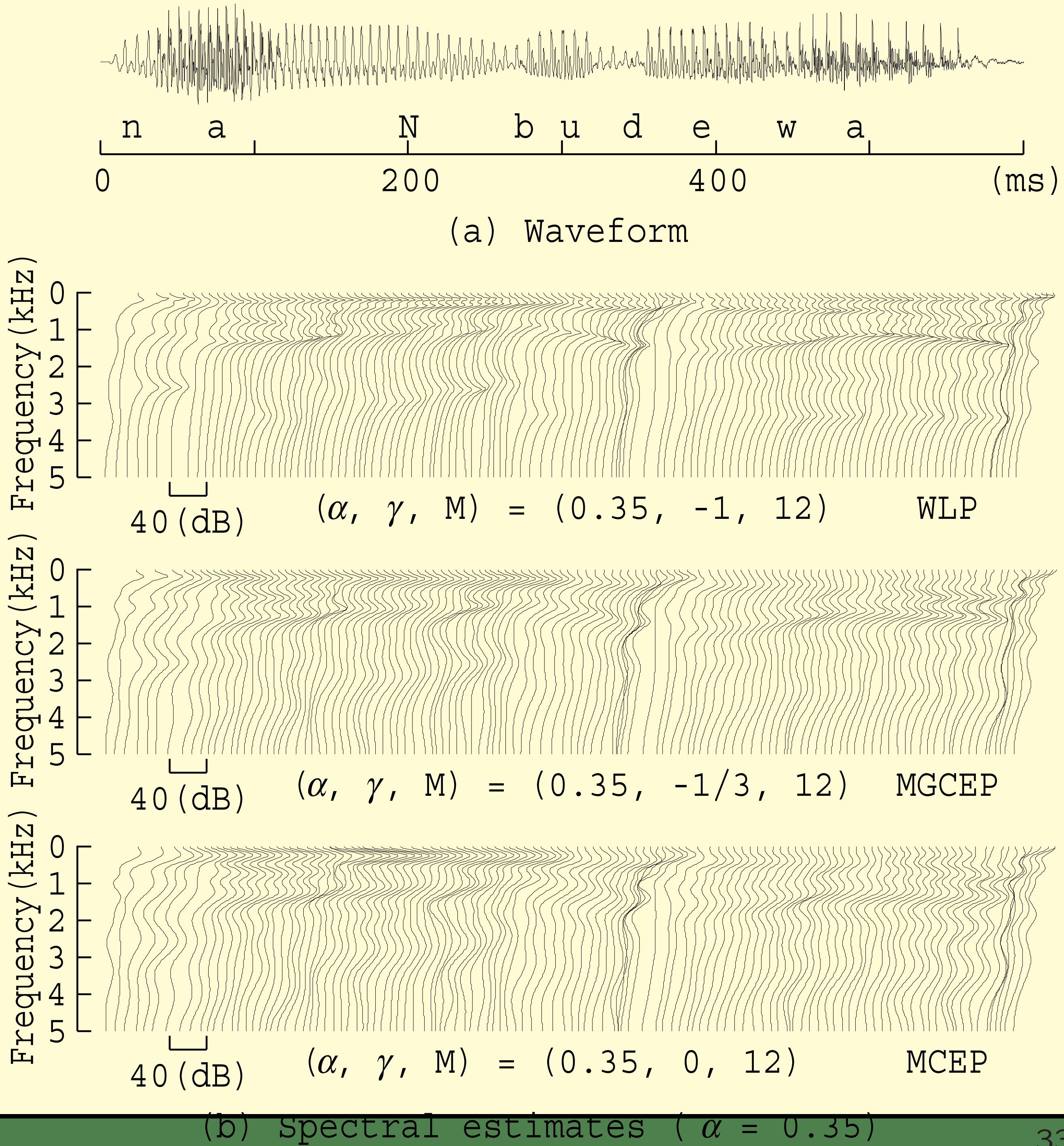
Example

$\alpha = 0.35$

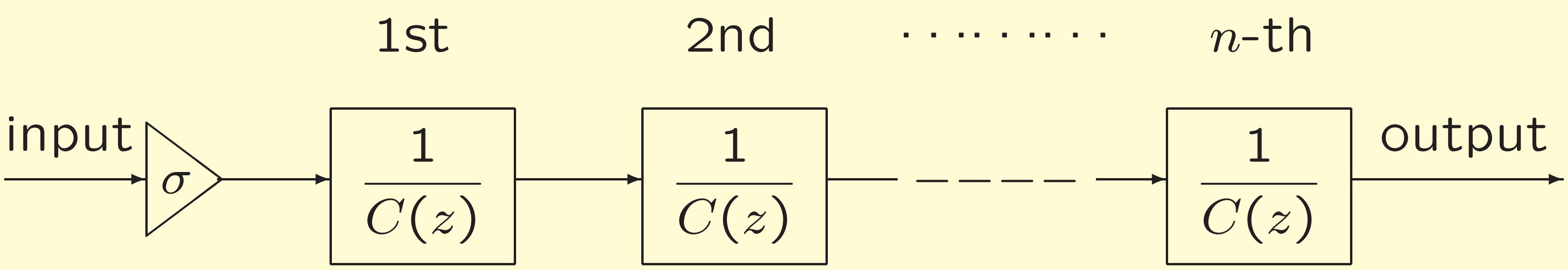
$\gamma = -1$

$\gamma = -1/3$

$\gamma = 0$



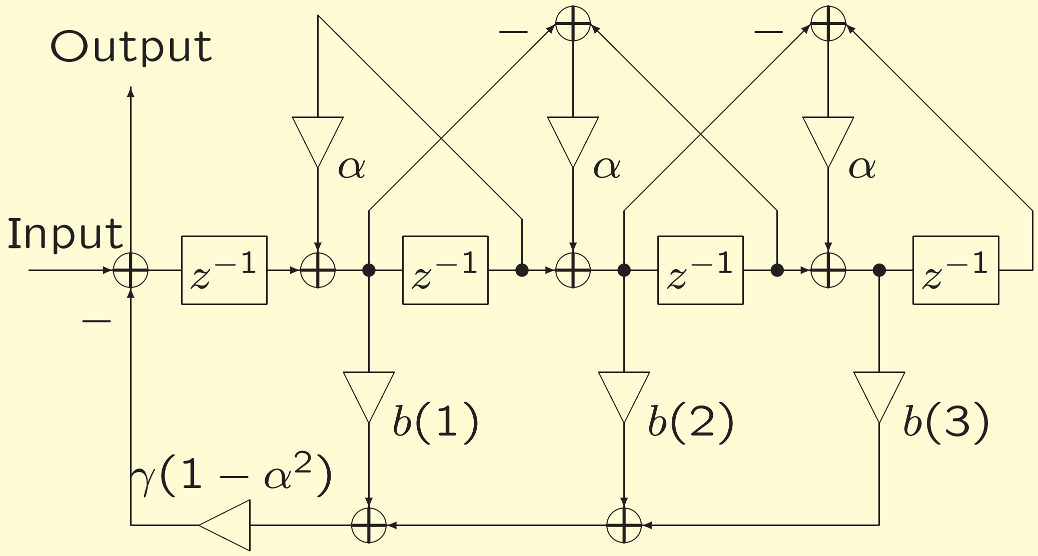
Structure of synthesis filter $H(z)$ ($\gamma = -1/n$)



Structure of $H(z)$

$$H(z) = \sigma D(z) = \sigma \left\{ \frac{1}{C(z)} \right\}^n$$

$$C(z) = \left(1 + \gamma \sum_{m=0}^{M} c'_{\alpha,\gamma}(m)\, z_\alpha^{-m} \right)$$



Structure of $C(z)$ ($M = 3$)

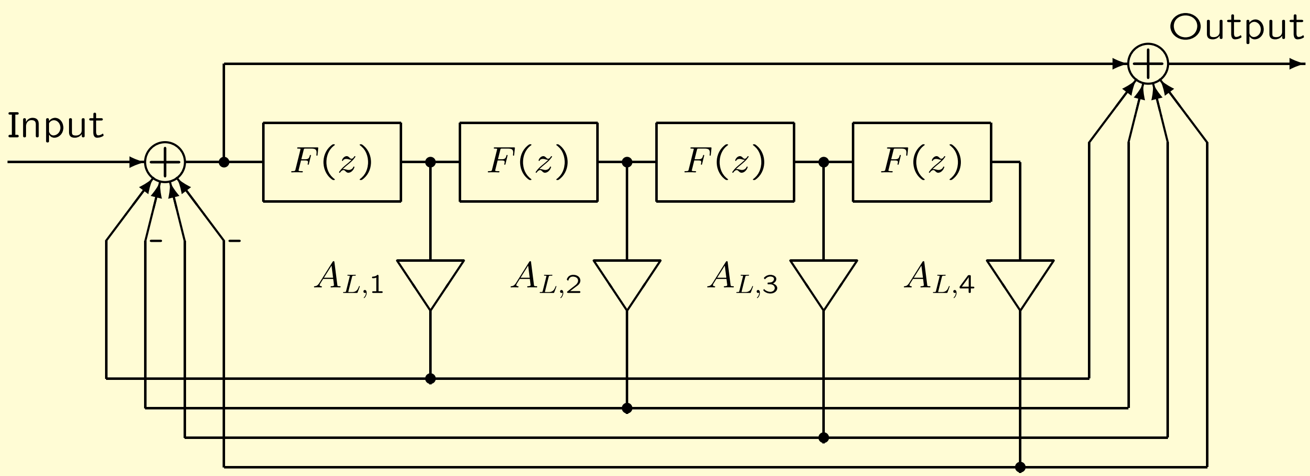
Structure of synthesis filter $H(z)$ ($\gamma = 0$) —MLSA filter

$$D(z) = \exp F(z) \simeq R_L(F(z))$$

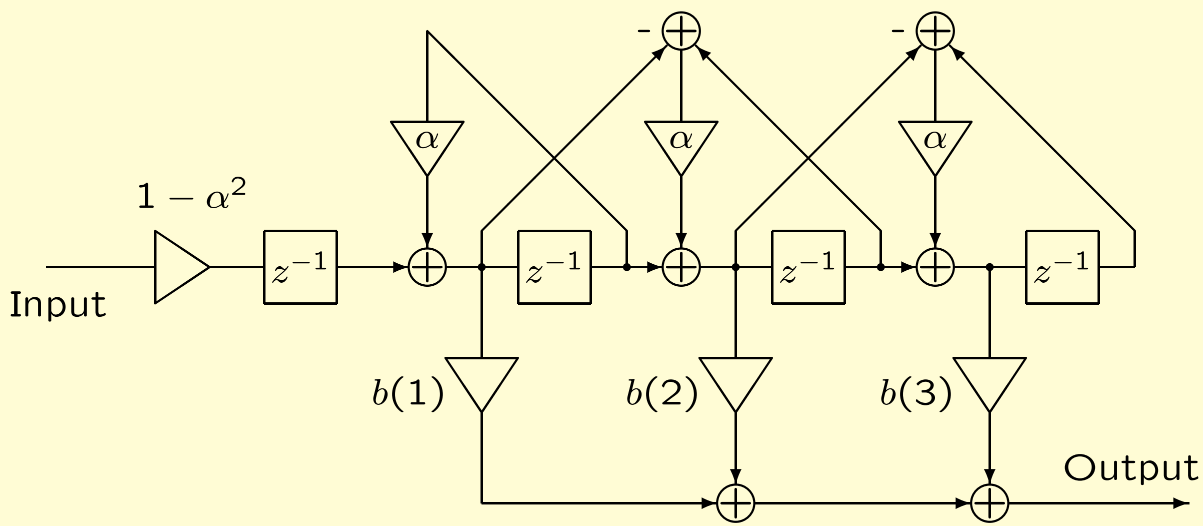
$$F(z) = \sum_{m=0}^{M} c'_{\alpha,\gamma}(m)\, z_{\alpha}^{-m}$$

- sufficient accuracy: maximum spectral error 0.24dB
- $O(8M)$ multiply-add operations a sample
- guaranteed stability
- M multiply-add operations for filter coefficients calculation

Structure of MLSA filter



$$R_L(F(z)) \simeq \exp F(z) = D(z), \; L = 4$$



Basic filter $F(z)$, $M = 3$

The choice of α, γ for speech analysis/synthesis

Analysis/synthesis system with fixed α and γ

- speech quality change with γ

  $\gamma \to -1$ Clear

  $\gamma \to 0$ Smooth

- When $\gamma = 0$, speech quality with $(\alpha, M) = (0.35, 15)$ is almost equivalent to that with $(\alpha, M) = (0, 30)$.

- When the analysis order is high enough, the difference becomes small.

Feature of Unified Approach

- Linear prediction analysis, Cepstral analysis are the special cases.
- Mathematically well-defined
- Physical interpretation

  $\Rightarrow$ Minimization of energy of inverse filter output

  $\Rightarrow$ $\boldsymbol{x}$ is Gaussian $\Rightarrow$ Minimization of $p(\boldsymbol{x}|\boldsymbol{c})$ (ML estimation)
- Global solution, stability of the system function
- Synthesis filter for direct synthesis from the estimated coefficients

  $\Rightarrow$ LMA/MLSA/GMSLA filter
- Extension to adaptive analysis (sample by sample basis)
- Parameter transformation for speech recognition

Word Recognition based on HMM

Spectral Analysis:

1. $(\alpha_1, \gamma_1, M_1) = (0, -1, 12) \Rightarrow$ Linear Prediction
2. $(\alpha_1, \gamma_1, M_1) = (0.35, -1/3, 12) \Rightarrow$ Mel-generalized cepstral analysis
3. $(\alpha_1, \gamma_1, M_1) = (0.35, 0, 12) \Rightarrow$ Mel-cepstral analysis

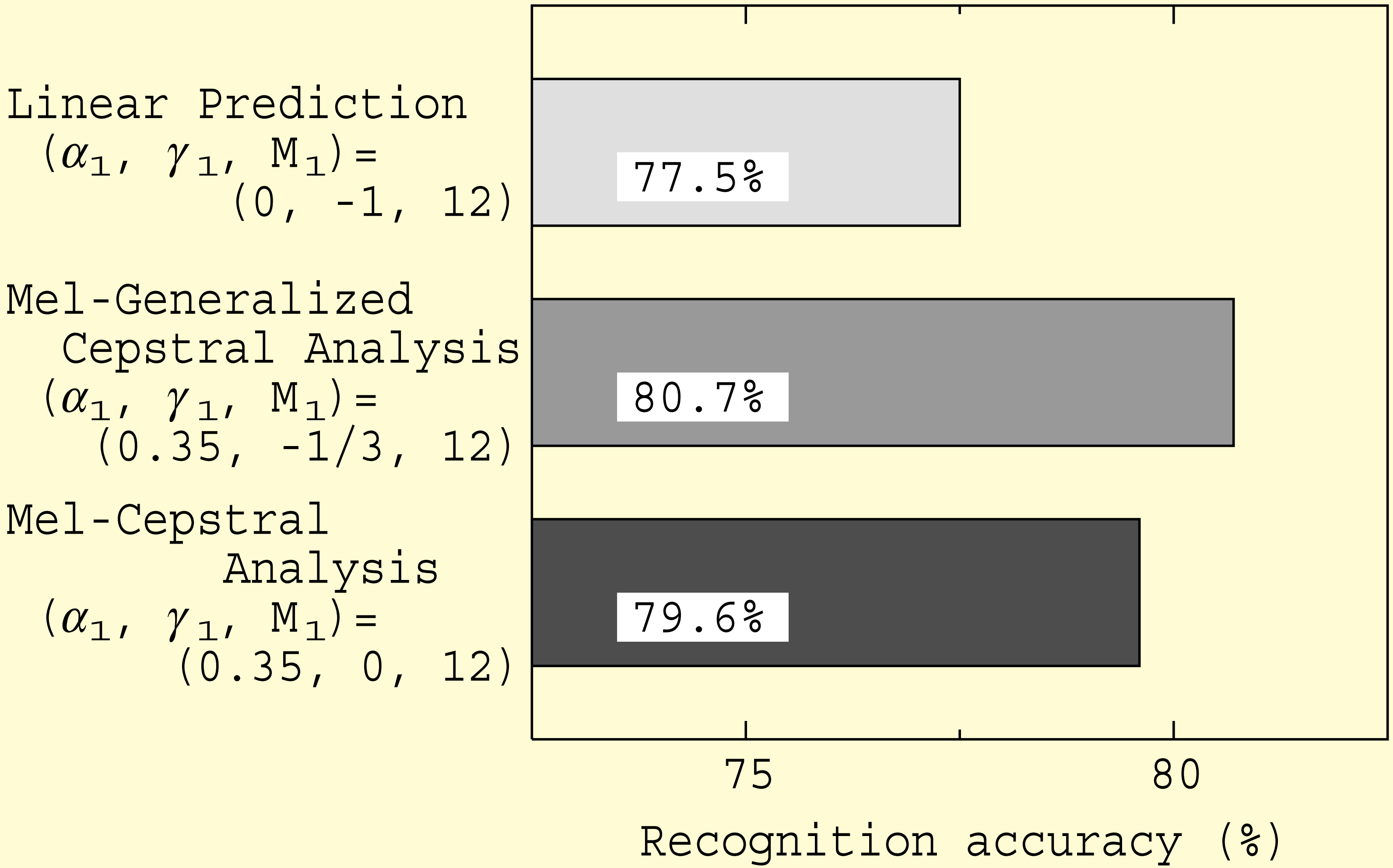
Output vector of HMM:

$(\alpha_2, \gamma_2, M_2) = (0.35, 0, 12)$
Mel-cepstral coefficients
and Δ (dynamic coefficients)

$$H(z) = s_{\gamma_1}^{-1}\left(\sum_{m=0}^{M_1} c_{\alpha_1,\gamma_1}(m)\, z_{\alpha_1}^{-m}\right) = s_{\gamma_2}^{-1}\left(\sum_{m=0}^{\infty} c_{\alpha_2,\gamma_2}(m)\, z_{\alpha_2}^{-m}\right)$$

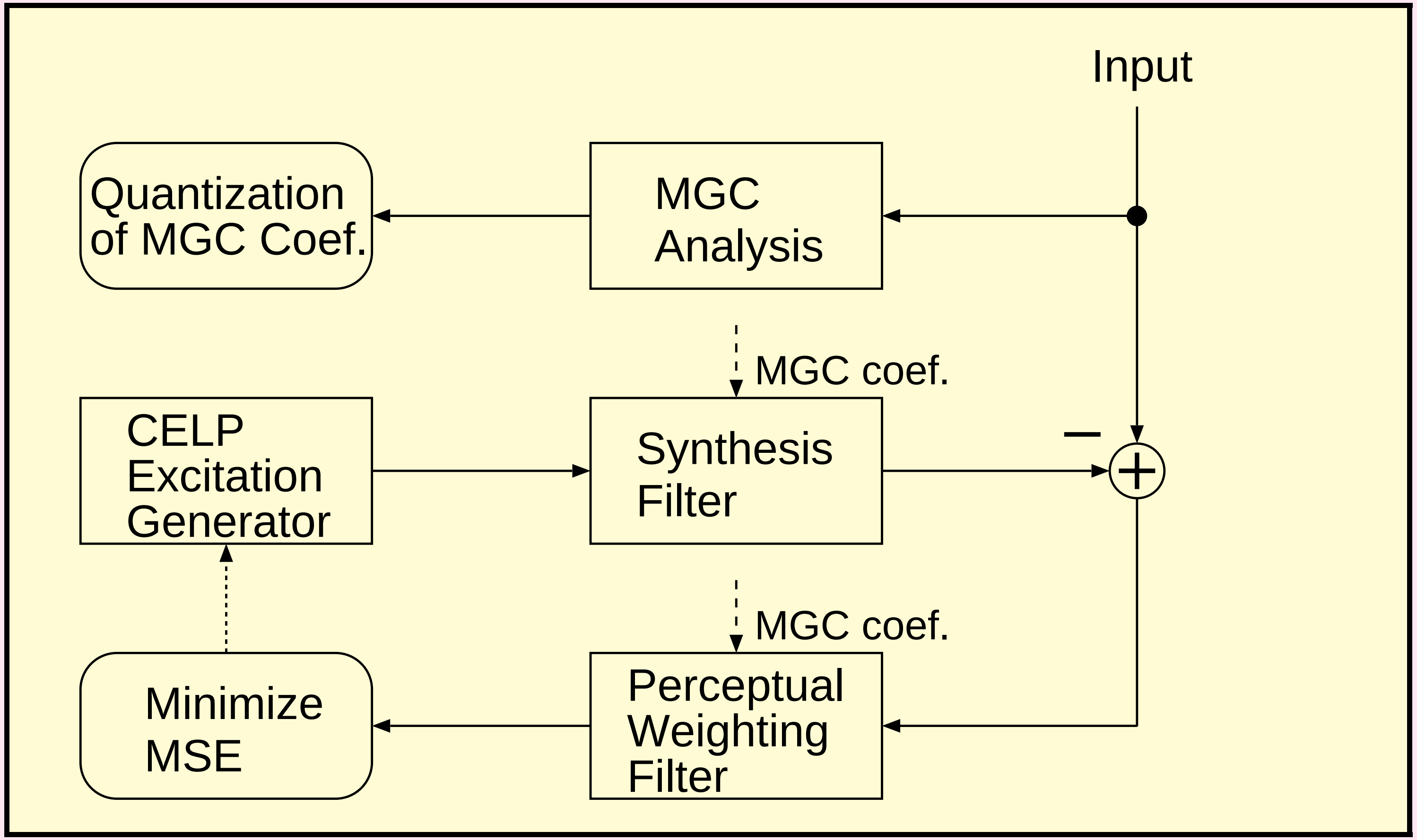
Recognition Accuracy

(Continuous HMM, 33 phoneme models, 2618 words)

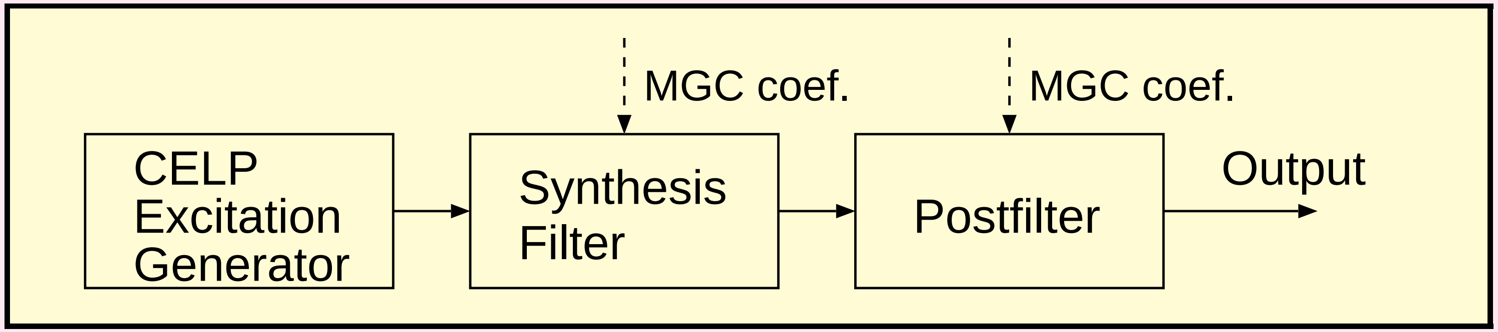


Application to 16kb/s wideband CELP coder

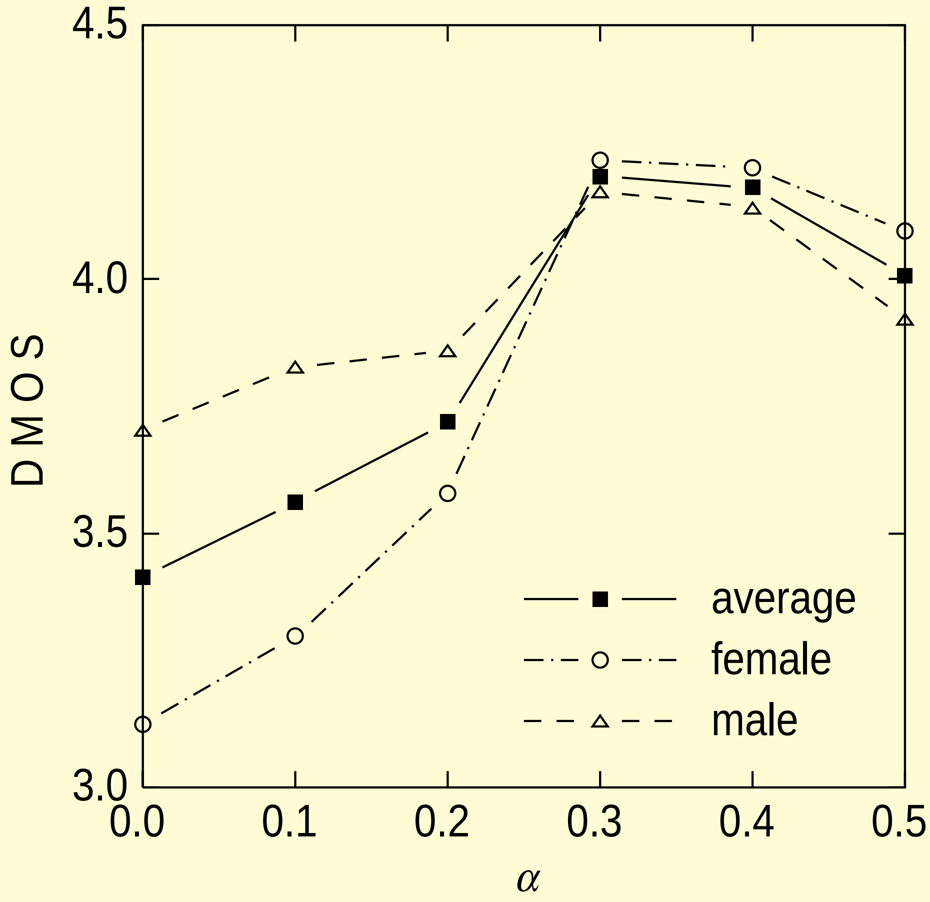
encoder



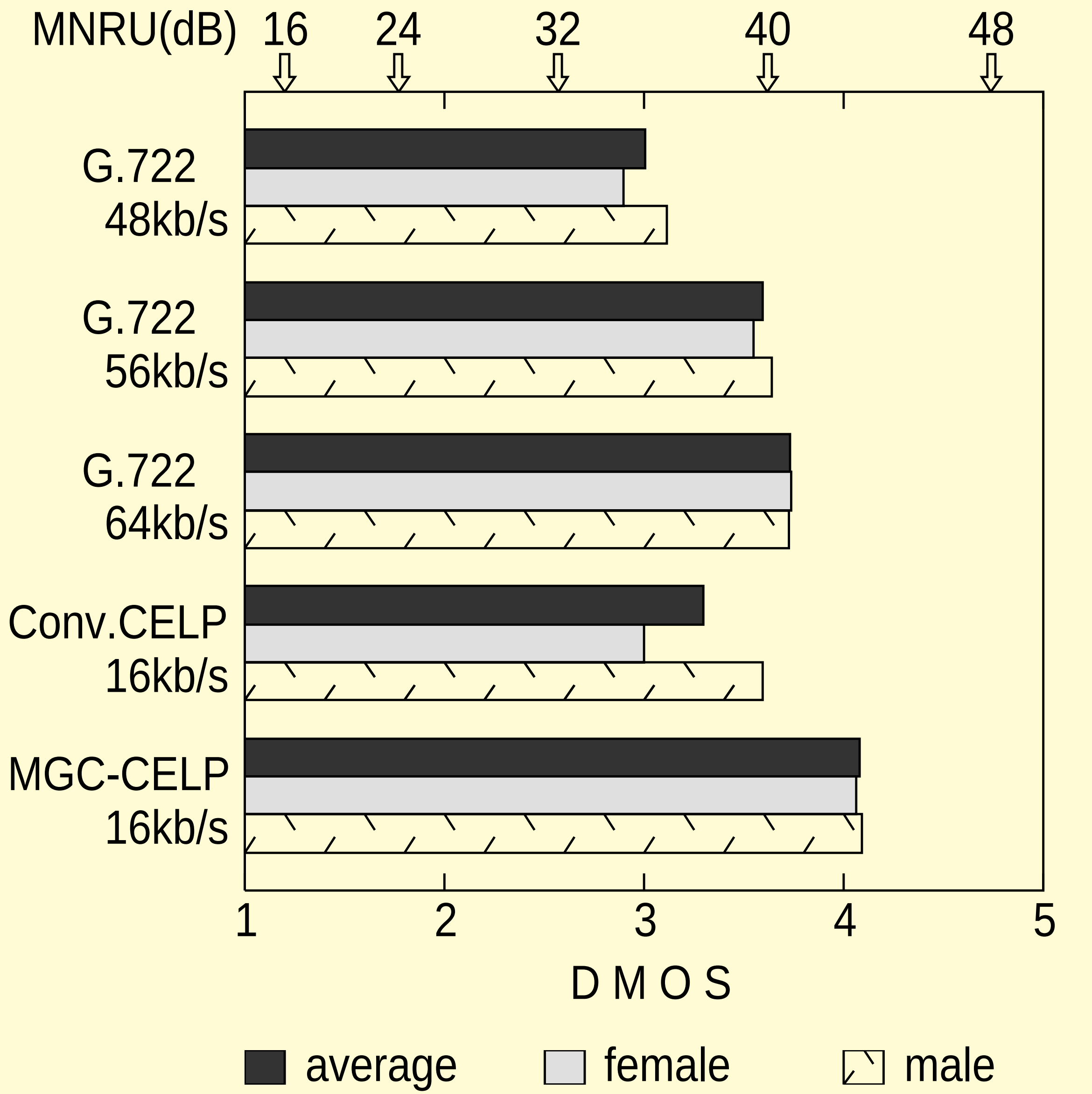
decoder



Speech quality as a function of α ($\gamma = -1/2$)



Subjective Evaluation



Summary

A unified approach to speech spectral estimation

- A unified approach to
  Linear predicton and Cepstral analysis
- Introduction of auditory frequency scale
- Efficint representation of speech spectrum with an appropriate choice of α and γ
- Application to speech anaylysis/synthesis, speech coding, speech recognition

Future work: Optimal α and γ
(Phoneme/Speaker dependent?)

Speech Signal Processing Toolkit:
`http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/`