

CECS401
Fundamentals of Spoken Language Processing

Note-10
Thursday 9/30/99

F. Speech Feature Analysis

Cepstral analysis

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega}$$

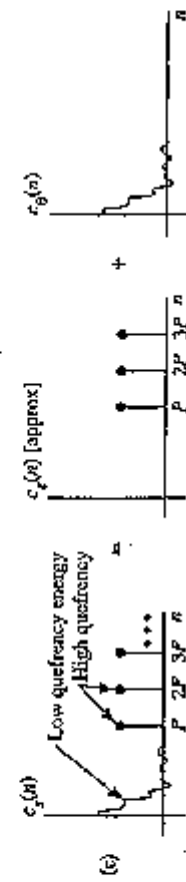
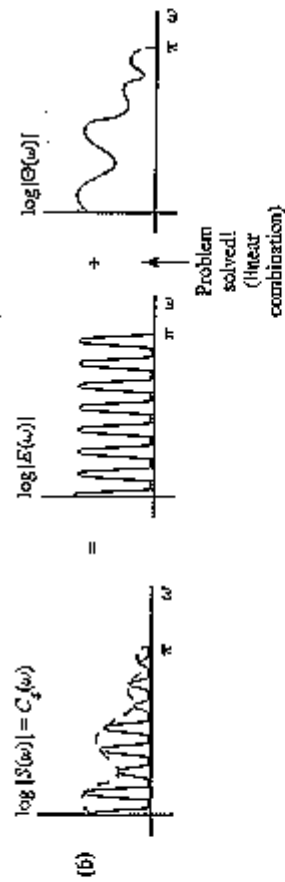
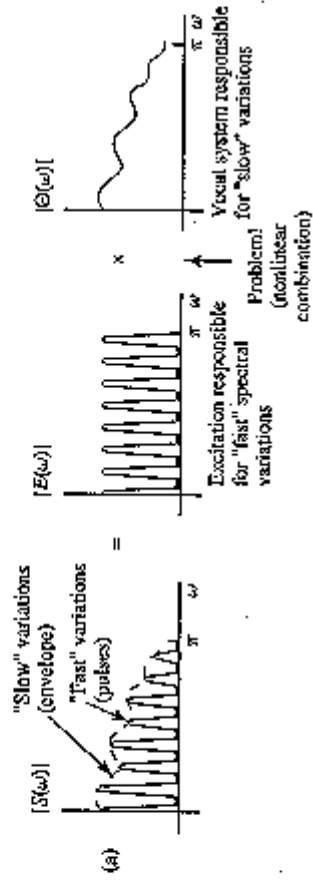
$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log S(\omega) e^{jn\omega} d\omega$$

$S(\omega)$ is power spectrum and c_n is cepstrum.

Advantages of cepstral analysis:

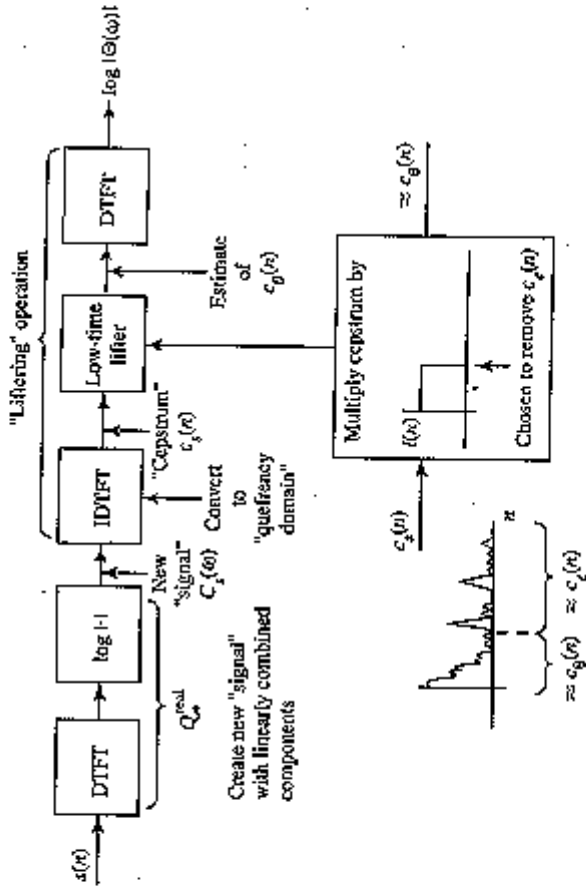
- source and channel separation in speech representation
- dynamic range compression in log power spectrum

Illustration of cepstral analysis (for $\log|S(\omega)|$)



- (a). In the speech magnitude spectrum, $|S(\omega)|$, two components can be identified: a “quickly varying” part due to the excitation, $|E(\omega)|$, and a “slowly varying” part due to the vocal system, $|\Theta(\omega)|$. The two components are combined by multiplication. Their time-domain counterparts, $e(n)$ and $\theta(n)$, are convolved.
- (b). In the log spectrum $C_s(\omega)$, the two components, $\log |E(\omega)|$ and $\log |\Theta(\omega)|$, become additive. The first term is fast varying and the second term is slowly varying.
- (c). In $c_s(n)$ ($= \text{IDFT}(C_s(\omega))$), the fast varying part yields a “**cepstral**” component, $c_e(n)$, at high “**quefrequencies**” (large n), and the slowly varying part yields a “**cepstral**” component, $c_\theta(n)$, at low “**quefrequencies**” (small n). $c_e(n)$ corresponds to the cepstrum of the excitation, and $c_\theta(n)$ corresponds to the cepstrum of the vocal system impulse response.

Illustration of low-pass "liftering" to remove source excitation $c_e(n)$ from cepstrum $c_s(n)$.



Mel-frequency cepstral coefficients (MFCC)

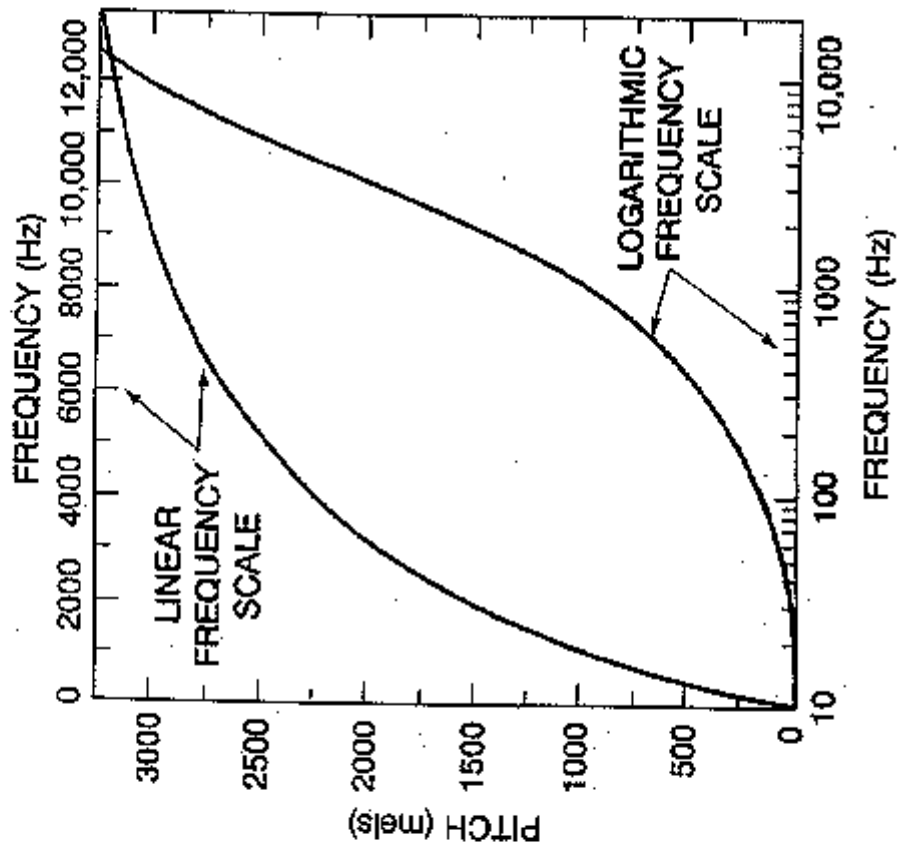
Incorporate nonlinear frequency resolution of human ear in cepstral analysis.

Mel-frequency:

- Mel-frequency measures subjective pitch as a function of frequency.
- A reference point is defined as 1000 mels for a 1 KHz tone (pitch), at the level of 40 dB above perceptual hearing threshold.
- Other subjective pitch values are obtained by adjusting the frequency of a tone such that it is half or twice the perceived pitch of a reference tone (with a known mel frequency).

Two-section model for mel-frequency:

- Below 1000 Hz, mel-frequency is linear with Hz frequency.
- Beyond 1000 Hz, mel-frequency is linear with \log Hz frequency.

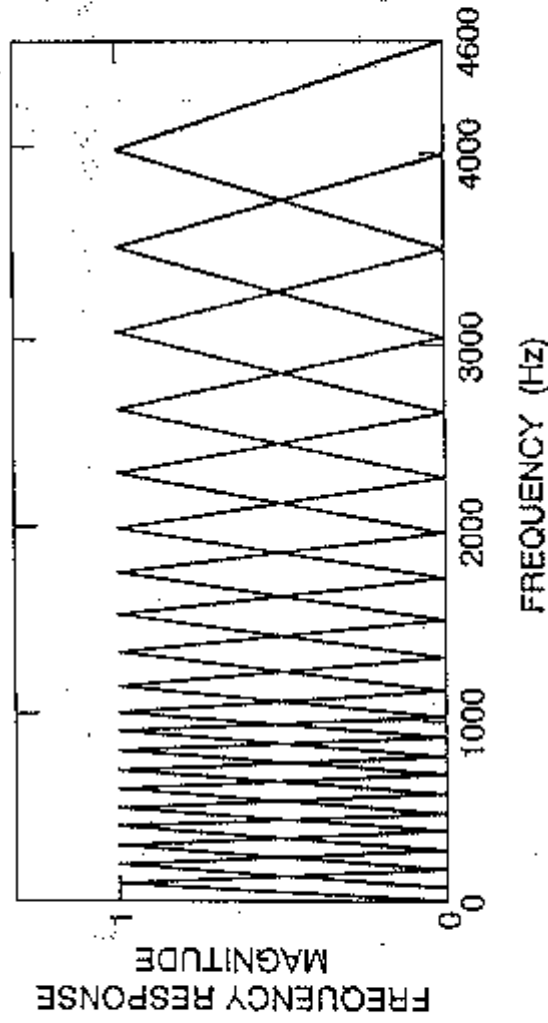


Mel-scale filter bank:

Place a bank of filters, $V_1(\omega)$, $V_2(\omega)$, \dots , $V_K(\omega)$, uniformly on the mel scale (nonuniformly on Hz frequency scale).

A commonly used mel-scale filter bank:

Width of each filter is 300 mels, spacing between successive filters is 150 mels, total filter is 20 in the frequency range of 0–4600 Hz.



The output energy of the k th filter is

$$\tilde{S}_k = \sum_{\omega=\omega_{kl}}^{\omega_{kh}} V_k(\omega) S(\omega)$$

MFCC \tilde{c}_i , $1 \leq i \leq L$, are defined as

$$\tilde{c}_i = \sum_{k=1}^K \log \tilde{S}_k \cos \left[i \left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \quad 1 \leq i \leq K$$

Derivation is made by using the definition

$$\tilde{c}_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \tilde{S}(\omega) e^{j i \omega} d\omega$$

and by the approximation

$$\log \tilde{S}(\omega) = \sum_{k=-(K-1)}^K \log \tilde{S}_k \cdot \delta \left(\omega - \frac{\pi}{K} \left(k - \frac{1}{2} \right) \right)$$

Perceptually-based linear-predictive analysis (PLP)

Several well-known properties of hearing are simulated by engineering approximations.

The resulting auditory-like spectrum of speech is approximated by an all-pole autoregressive model.

PLP yields perceptual LPC parameters or perceptual LPC cepstral parameters.

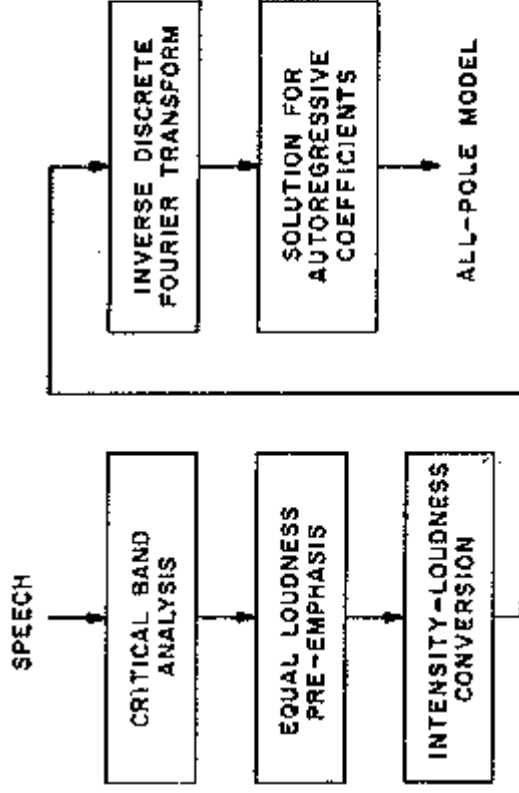
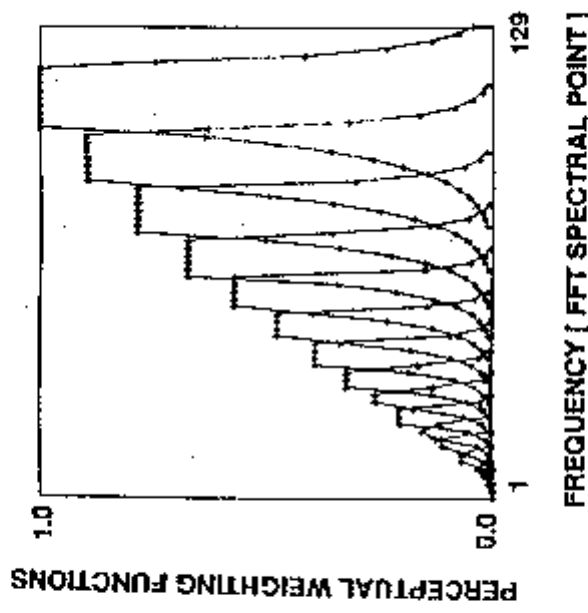


FIG. 1. Block diagram of perceptual linear predictive (PLP) speech analysis.

Integration of critical-band analysis and equal-loudness pre-emphasis ($S(\omega) \rightarrow \Xi(\Omega(\omega))$)



Intensity to loudness conversion:

$$\Phi(\Omega) = \Xi(\Omega)^{1/3}$$

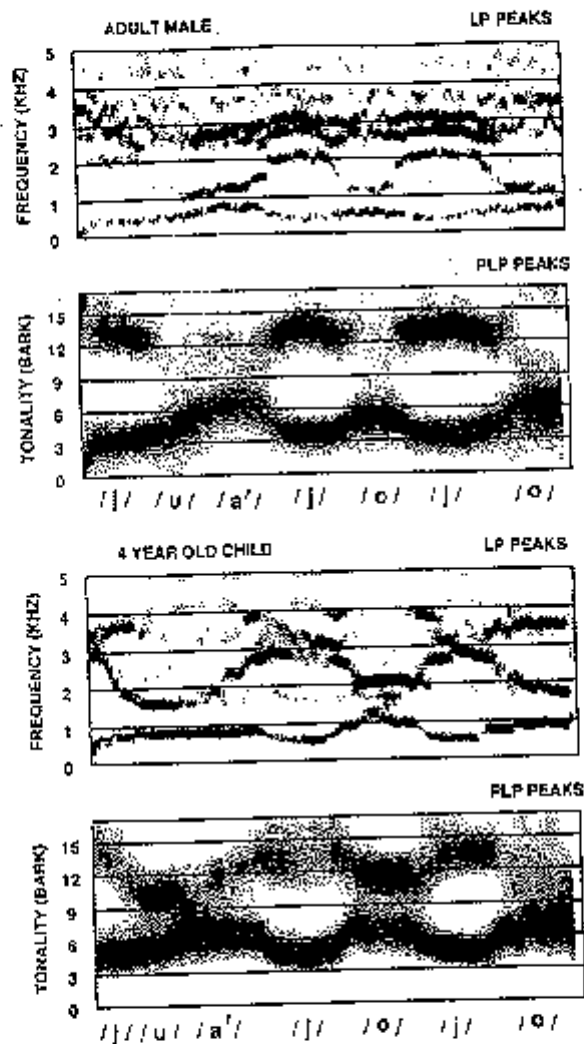


FIG. 13. Peak-enhanced spectrograms from conventional LP and PLP analyses of the utterance "You are yo-yo" uttered by adult male and a 4-year-old child. While for these two speakers LP analysis finds a different number of formants at different positions, PLP analysis finds the same number of peaks at similar positions.

Dynamic spectral or cepstral features

Motivation:

- Short-time spectral features represent only locally stationary spectral property of speech.
- Spectral transitions over time are important for human perception of sound classes.
- Spectral transitions can be modeled by time-derivatives of short-time log spectral sequence.

(a). Direct differentiation:

$$\log S(\omega, t) = \sum_{n=-\infty}^{\infty} c_n(t) e^{-jn\omega}$$

$$\frac{d}{dt} \log S(\omega, t) = \sum_{n=-\infty}^{\infty} \frac{d}{dt} c_n(t) e^{-jn\omega}$$

$$\frac{d^2}{dt^2} \log S(\omega, t) = \sum_{n=-\infty}^{\infty} \frac{d^2}{dt^2} c_n(t) e^{-jn\omega}$$

The differentiations can be approximated by

$$\begin{aligned} \frac{d}{dt} c_n(t) &\approx \frac{c_n(t + \Delta t) - c_n(t - \Delta t)}{2\Delta t} \triangleq \Delta c_n(t) \\ \frac{d^2}{dt^2} c_n(t) &\approx \frac{\Delta c_n(t + \Delta t) - \Delta c_n(t - \Delta t)}{2\Delta t} \triangleq \Delta \Delta c_n(t) \end{aligned}$$

(b). Orthogonal polynomial fitting

The temporal dynamics of $c_n(t)$ can be fitted by finite-order polynomials in the neighborhood of t .

Example

In the time interval $[-M, M]$ take the orthogonal polynomials

$$f_0(t) = a_{0,0}$$

$$f_1(t) = a_{1,0} + a_{1,1}t$$

$$f_2(t) = a_{2,0} + a_{2,1}t + a_{2,2}t^2$$

$$\sum_{t=-M}^M f_i(t)f_j(t) = 0 \quad i \neq j$$

Choose $a_{0,0} = a_{1,1} = a_{2,2} = 1$, then

$$a_{1,1} = 0, \quad a_{2,1} = 0, \quad a_{2,0} = -\frac{1}{2M+1} \sum_{t=-M}^M t^2$$

For $M = 4$,

$$f_0(t) = 1, f_1(t) = t, f_2(t) = t^2 - \frac{20}{3}$$

Then $c_n(t)$ is approximated as

$$c_n(t) \simeq \lambda_{n,0}f_0(t) + \lambda_{n,1}f_1(t) + \lambda_{n,2}f_2(t)$$

with

$$\lambda_{n,i} = \frac{\sum_{t=-M}^M c_n(t)f_i(t)}{\sum_{t=-M}^M f_i^2(t)}$$

Mean:

$$\lambda_{n,0} = \frac{\sum_{t=-M}^M c_n(t)}{2M+1}$$

Slope:

$$\lambda_{n,1} = \left. \frac{d}{dt} c_n(t) \right|_{t=0}$$

Curvature:

$$\lambda_{n,2} = \left. \frac{d^2}{dt^2} c_n(t) \right|_{t=0}$$

The $\lambda_{n,i}$'s are also called temporal regression coefficients. The fitting interval is often chosen as 50 ms.

The temporal regression coefficients have less spurious variations than the dynamic features from direct differentiation.

The incorporation of dynamic features in speech recognition accounted for significant improvement in system performance.