7.0 Speech Signals and Front-end Processing

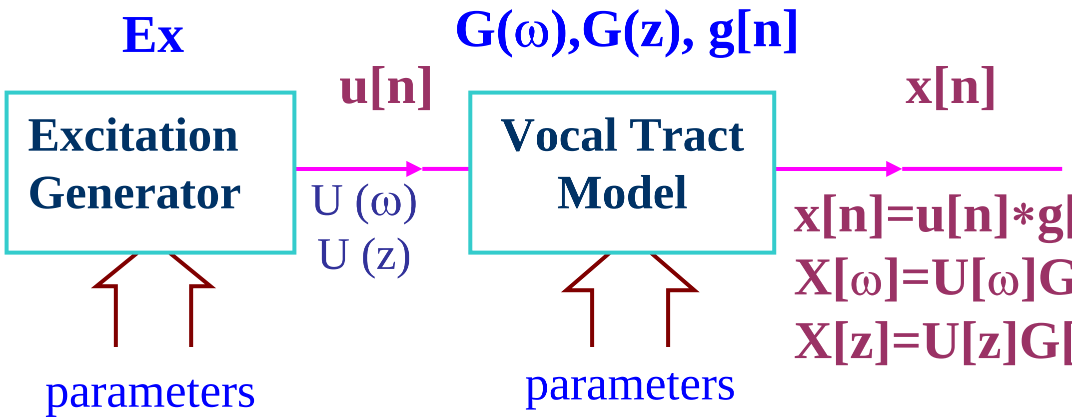
References: 1. 3.3, 3.4 of Becchetti

2. 2.2, 2.3, 3.3.1 ~ 3.3.6 of Rabiner& Juang

3. 9.3 of Huang

Speech Signals

- Voiced/unvoiced 濁音、清音
- Pitch/tone 音高、聲調
- Vocal tract 聲道
- Frequency domain/formant frequency
- Spectrogram representation
- Speech Source Model

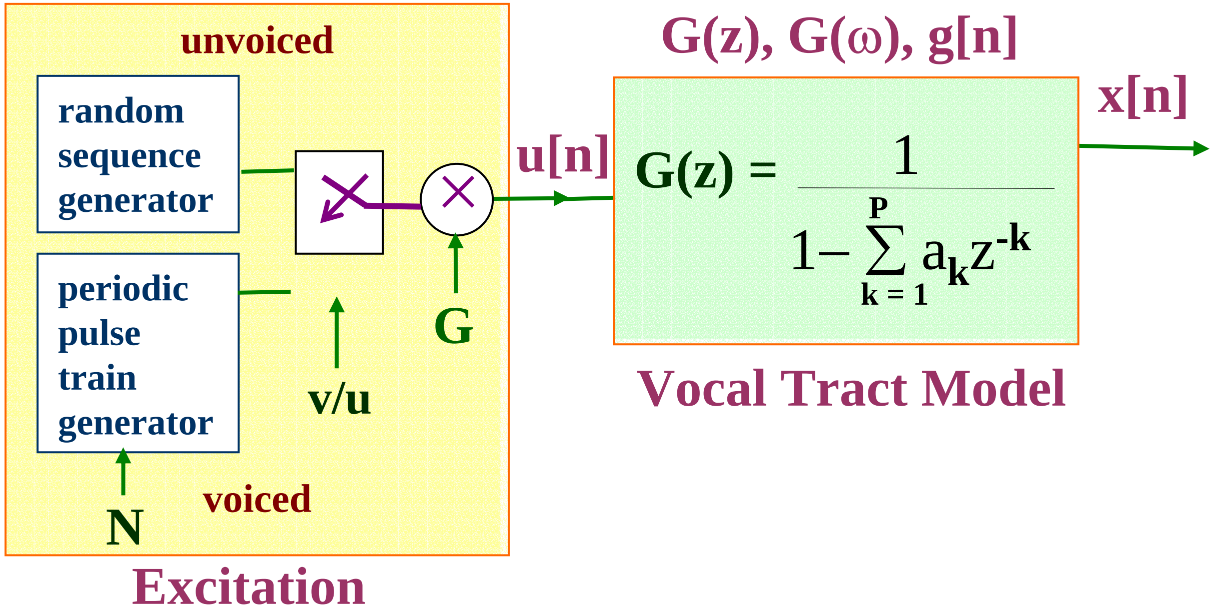


$x[n]=u[n]*g[n]$
$X[\omega]=U[\omega]G[\omega]$
$X[z]=U[z]G[z]$

  - digitization and transmission of the parameters will be adequate
  - at receiver the parameters can produce x[n] with the model
  - much less parameters with much slower variation in time lead to much less bits required
  - the key for low bit rate speech coding

Speech Signals

- **Speech Source Model**



– Excitation parameters

v/u : voiced/ unvoiced

N : pitch for voiced

G : signal gain

→ excitation signal u[n]

– Vocal Tract parameters

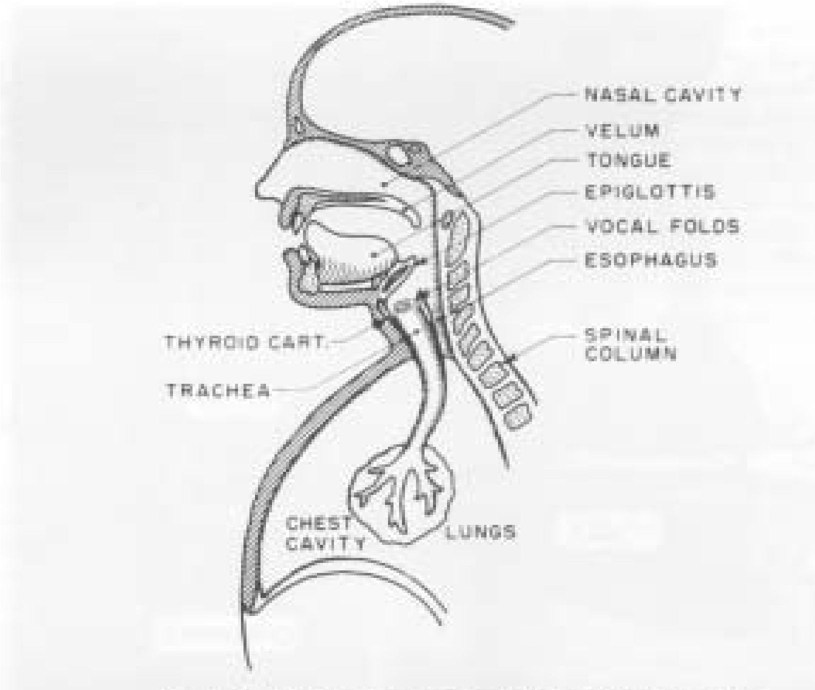
$\{a_k\}$: LPC coefficients

→formant structure of speech signals

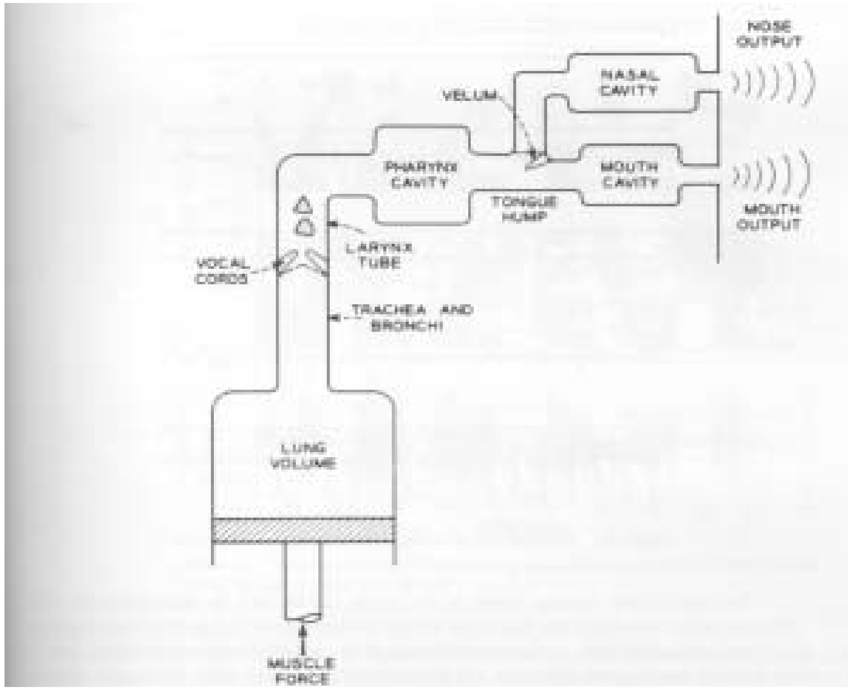
– A good approximation, though not precise enough

Speech Production

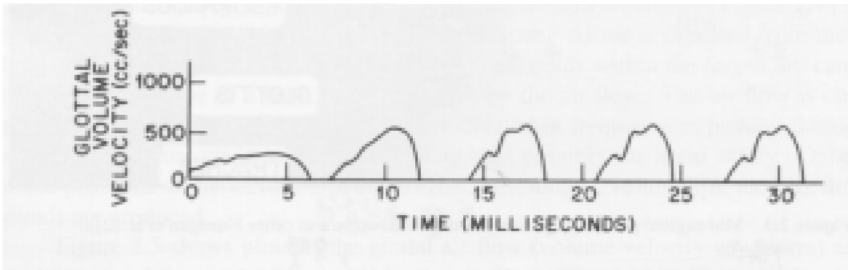
- Schematic view of the human vocal mechanism



- Schematic representation of the complete physiological mechanism of speech production

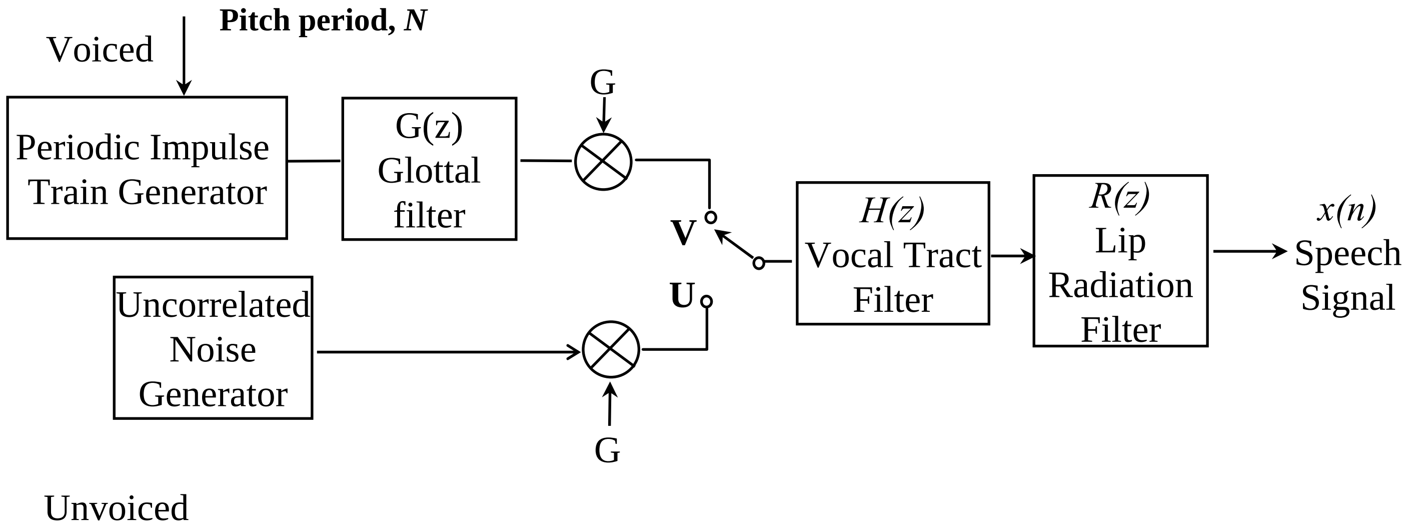


- Glottal volume velocity (excitation)

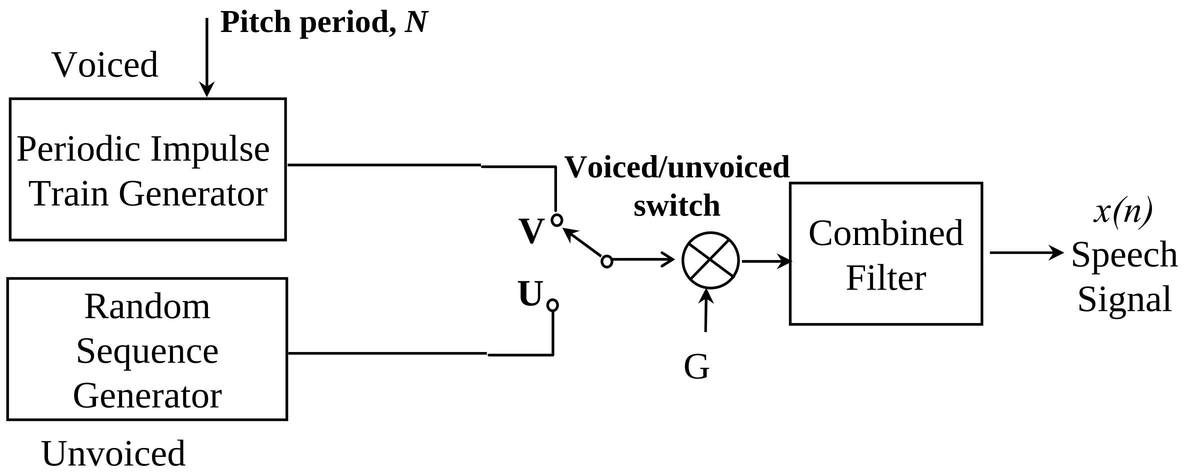


Speech Production

- **Sophisticated model for speech production**



- **Simplified model for speech production**



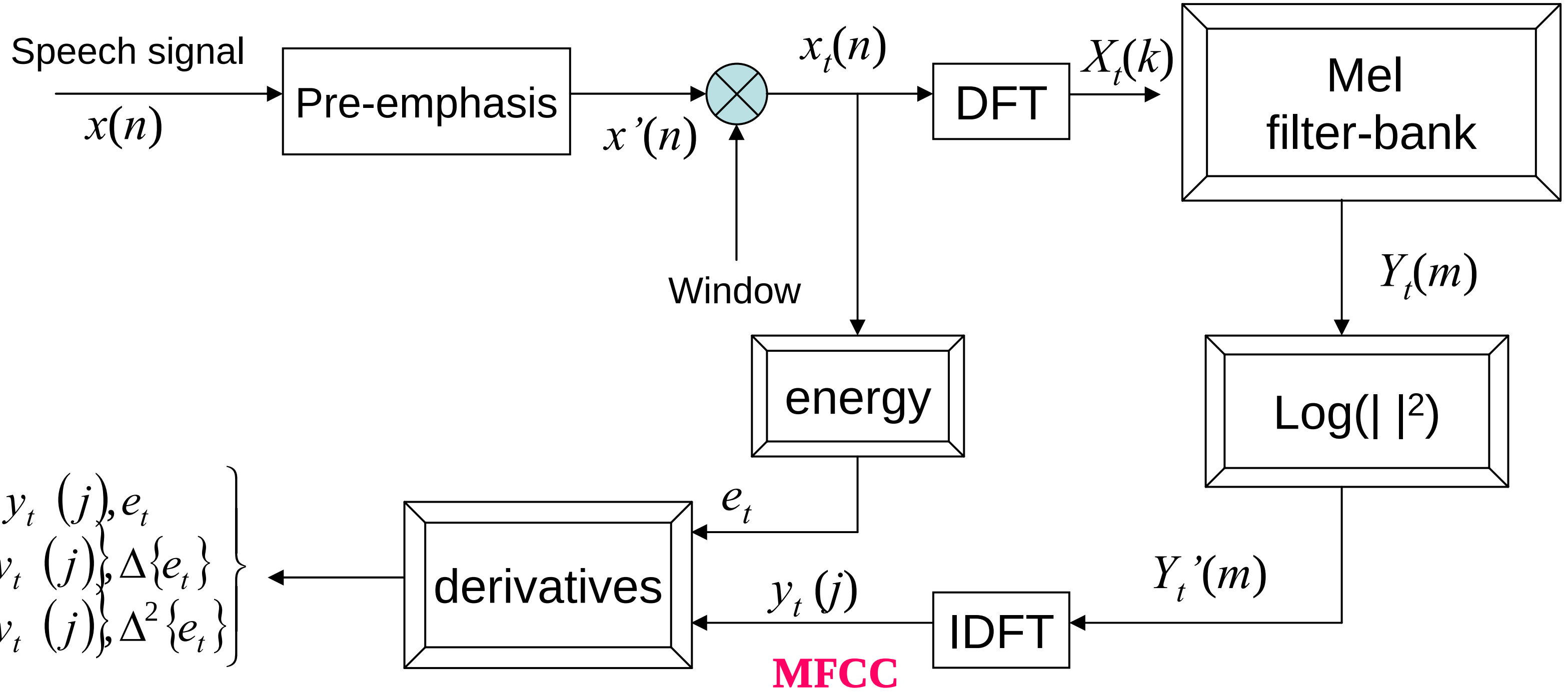
Feature Extraction

Major Considerations

- **Perceptually Meaningful**
  - Parameters representing salient aspects of the speech signal
  - parameters analogous to those used by human auditory system – perceptually meaningful
- **Robustness**
  - Parameters more robust to variations in environments, noise, channel, speaker, and transducer
- **Dynamic Characteristics**
  - Parameters capturing spectral dynamics, or changes of the spectrum with time

MFCC

- **Mel-Frequency Cepstral Coefficients (MFCC)**
  - Most widely used in the speech recognition
  - Has generally obtained a better accuracy at relatively low computational complexity
  - The process of MFCC extraction :

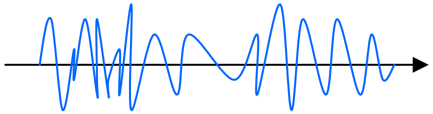


$$\mathbf{y}_t = \begin{Bmatrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{Bmatrix}$$

Pre-emphasis

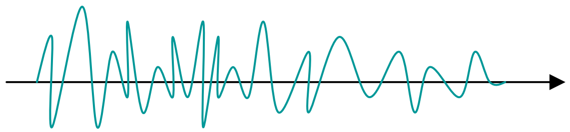
- The process of Pre-emphasis :
  - a high-pass filter

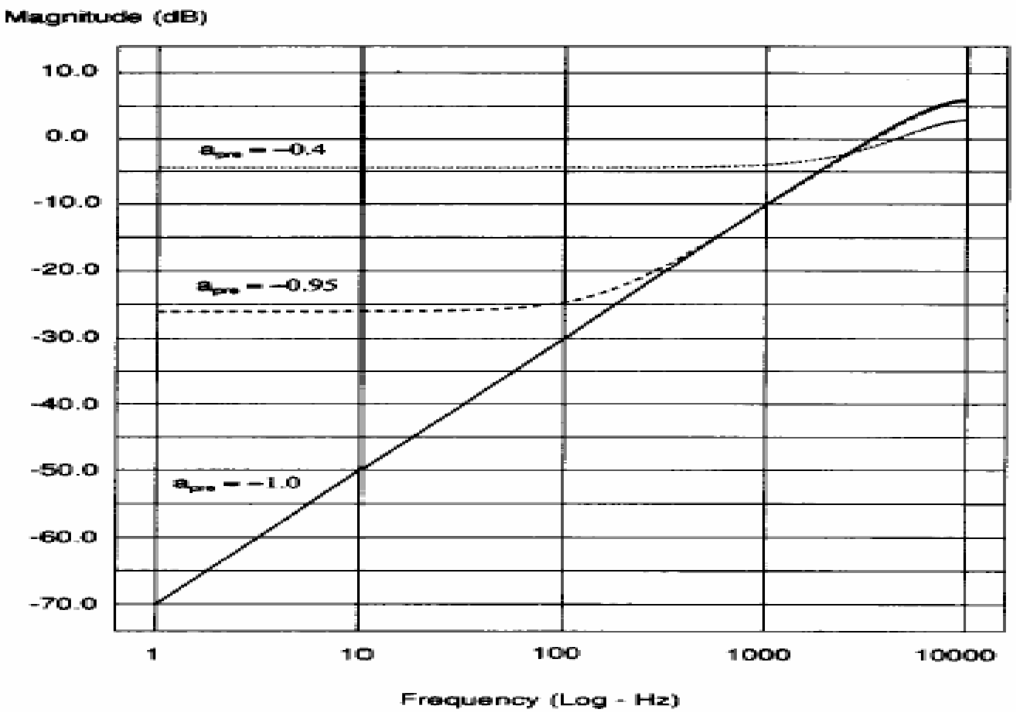
Speech signal $x(n)$



$$H(z)=1-a \bullet z^{-1} \quad 0<a\leq 1$$

$x'(n)=x(n)-ax(n-1)$





Why pre-emphasis?

- **Reason 1 :**
  - Voiced sections of the speech signal naturally have a negative spectral slope (attenuation) of approximately 20 dB per decade due to the physiological characteristics of the speech production system
  - High frequency formants have small amplitude with respect to low frequency formants. A pre-emphasis of high frequencies is therefore helpful to obtain similar amplitude for all formants

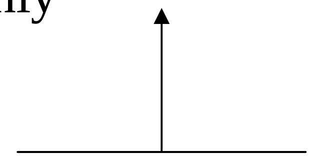
- **Reason 2：**
  - Hearing is more sensitive above the 1 kHz region of the spectrum

Why Windowing? (1)

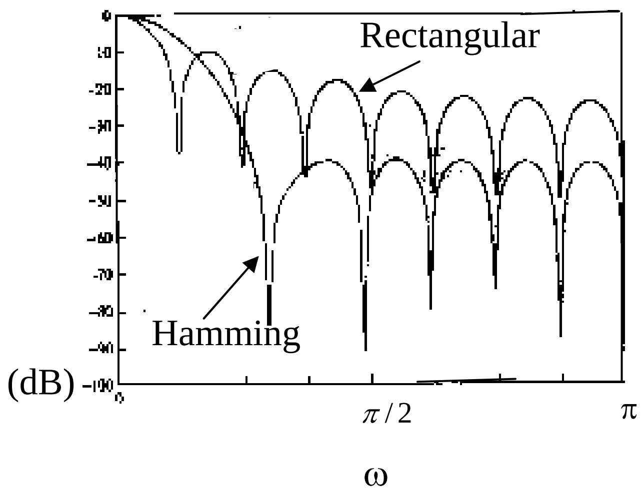
- **Why dividing the speech signal into successive and overlapping frames?**
  - Theoretical spectral evaluation approaches are in general for **stationary signals** (i.e., a signal whose statistical characteristics are invariant with respect to time)
    - For voice, this holds only within the short time intervals (short-time stationary, short-time Fourier analysis)
- **Frames**
  - **Frame Length :** the length of time over which a set of parameters is valid. Frame length ranges between **20 ~ 10** ms
  - **Frame Shift:** the length of time between successive parameter calculations
  - **Frame Rate:** number of frames per second

Why Windowing? (2)

- **Windowing :**
  - $x_t(n)=w(n)\cdot x'(n)$, $w(n)$: the shape of the window
    - Frequency response : $X_t(\omega)=W(\omega)*X'(\omega)$, *: convolution
  - Without windowing, $w(n)=1$ for all n , whose frequency response is just an impulse
    - This can't be used since the speech signal is stationary only within short-time intervals



  - Rectangular window ($w(n)=1$ for $0 \le n \le L-1$):
    - simply extract a segment of the signal
    - whose frequency response has high side lobes
  - *Main lobe* : spreads out in a wider frequency range the narrow band power of the signal, and thus reduces the local frequency resolution in formant allocation
  - *Side lobe* : swap energy from different and distant frequencies of $x'(n)$



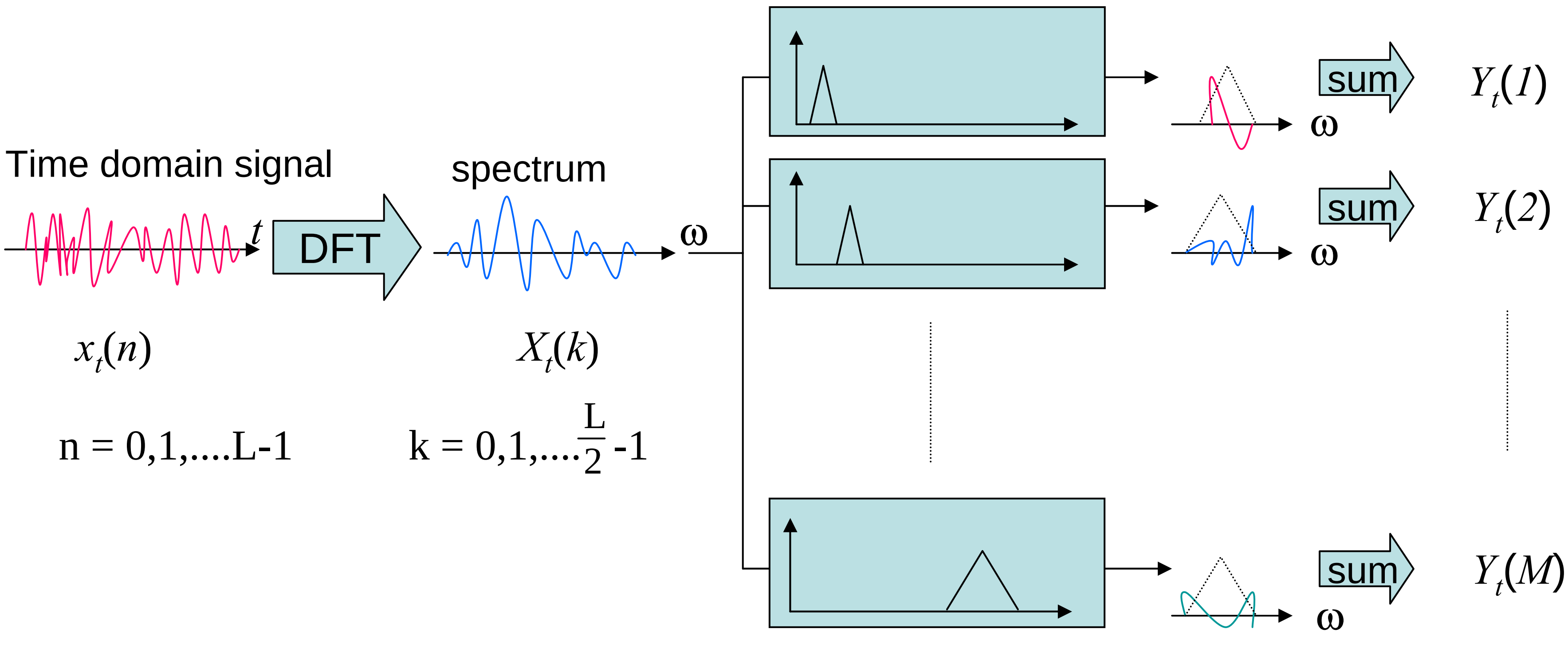
Why Windowing? (3)

- **Windowing (Cont.):**
  - For a designed window, we wish that
    - the main lobe is as narrow as possible
    - the side lobe is as low as possible
      - However, this is a trade-off
  - The most widely used window shape is the Hamming window, whose impulse response is a raised cosine impulse:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{L-1}\right), & n = 0,1,......,L-1 \\ 0 & \text{otherwise} \end{cases}$$

DFT and Mel-filter-bank Processing

- **For each frame of signal (*L* points, e.g., L=512),**
  - the Discrete Fourier Transform (DFT) is first performed to obtain its spectrum (*L* points, for example *L*=512)
  - The bank of filters according to Mel scale is then performed, and each filter output is the sum of its filtered spectral components (*M* filters, and thus *M* outputs, for example *M*=24)

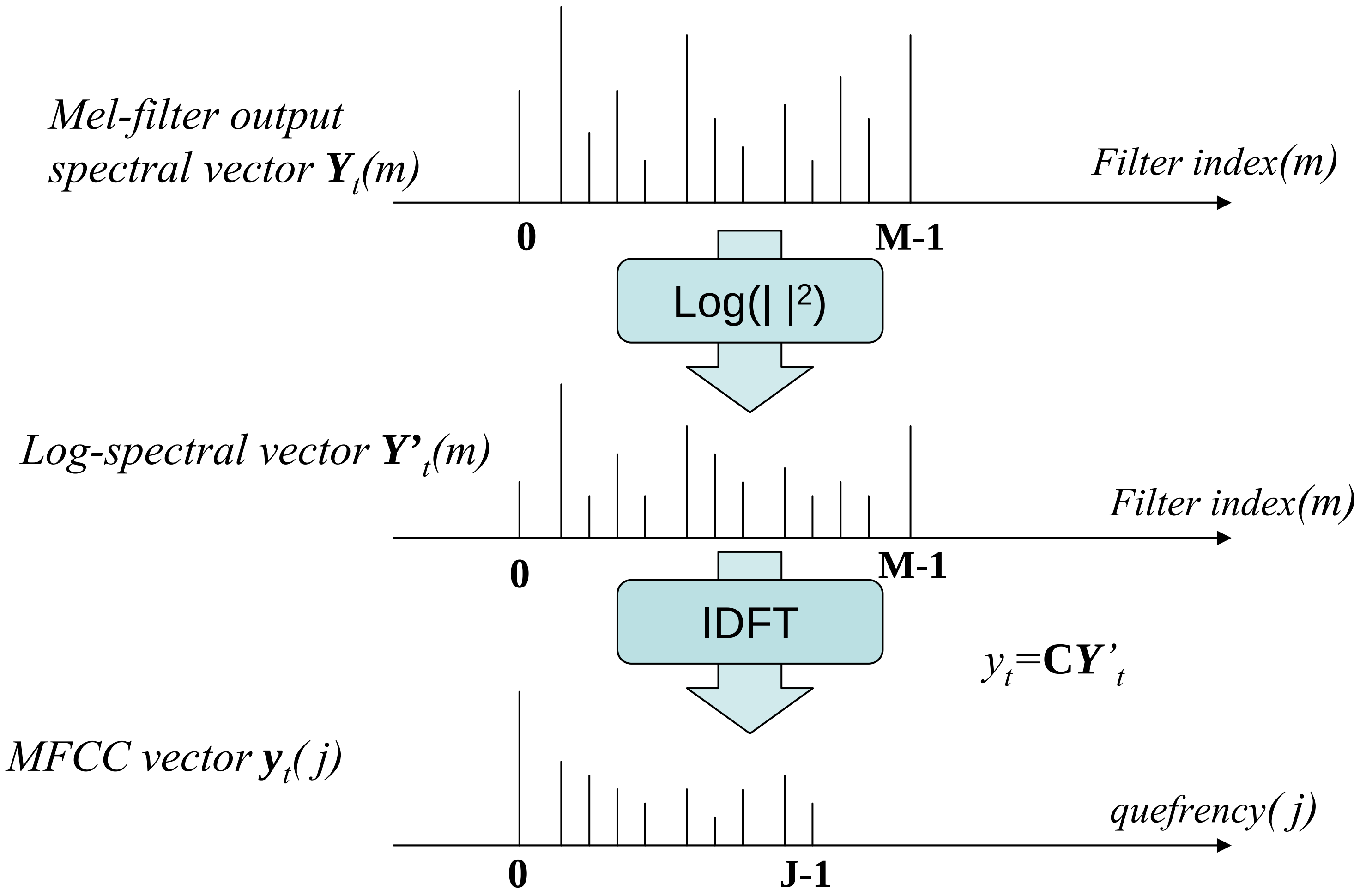


Why Filter-bank Processing?

- **The filter-bank processing simulates human ear perception**
  - *Center frequency of each filter* :
    - Human perception for pitch of signals is proportional to the *logarithm* of the frequencies
    - Lower frequencies (say, below 1 KHz) play important roles in human ear perception
  - *Bandwidth* :
    - Frequencies of a complex sound within a certain bandwidth around some center frequency cannot be individually identified.
    - When one of the components of this sound falls outside this bandwidth, it can be individually distinguished.
    - This bandwidth is referred to as the critical band.
    - The width of a critical band is roughly 10% to 20% of the center frequency of the sound

Logarithmic Operation and IDFT

- **The final process of MFCC evaluation : logarithm operation and IDFT**



Why Log Energy Computation?

- **Using the magnitude (energy) only**
  - Phase information is not very helpful in speech recognition
    - Replacing the phase part of the original speech signal with continuous random phase won't be perceived by human ear
    - Human perception sensitivity is proportional to signal energy
- **Using the Logarithmic operation**
  - The logarithm compresses the dynamic range of values, which is a characteristic of the human hearing system
  - The dynamic compression also makes feature extraction less sensitive to variations in signal dynamics
  - To make a convolved noisy process additive
    - Speech signal $x(n)$, excitation $u(n)$ and the impulse response of vocal tract $g(n)$

$$x(n)=u(n)*g(n) \Rightarrow X(\omega)=U(\omega)G(\omega)$$

$$\Rightarrow |X(\omega)|=|U(\omega)||G(\omega)| \Rightarrow \log|X(\omega)|=\log|U(\omega)|+\log|G(\omega)|$$

Why Inverse DFT?

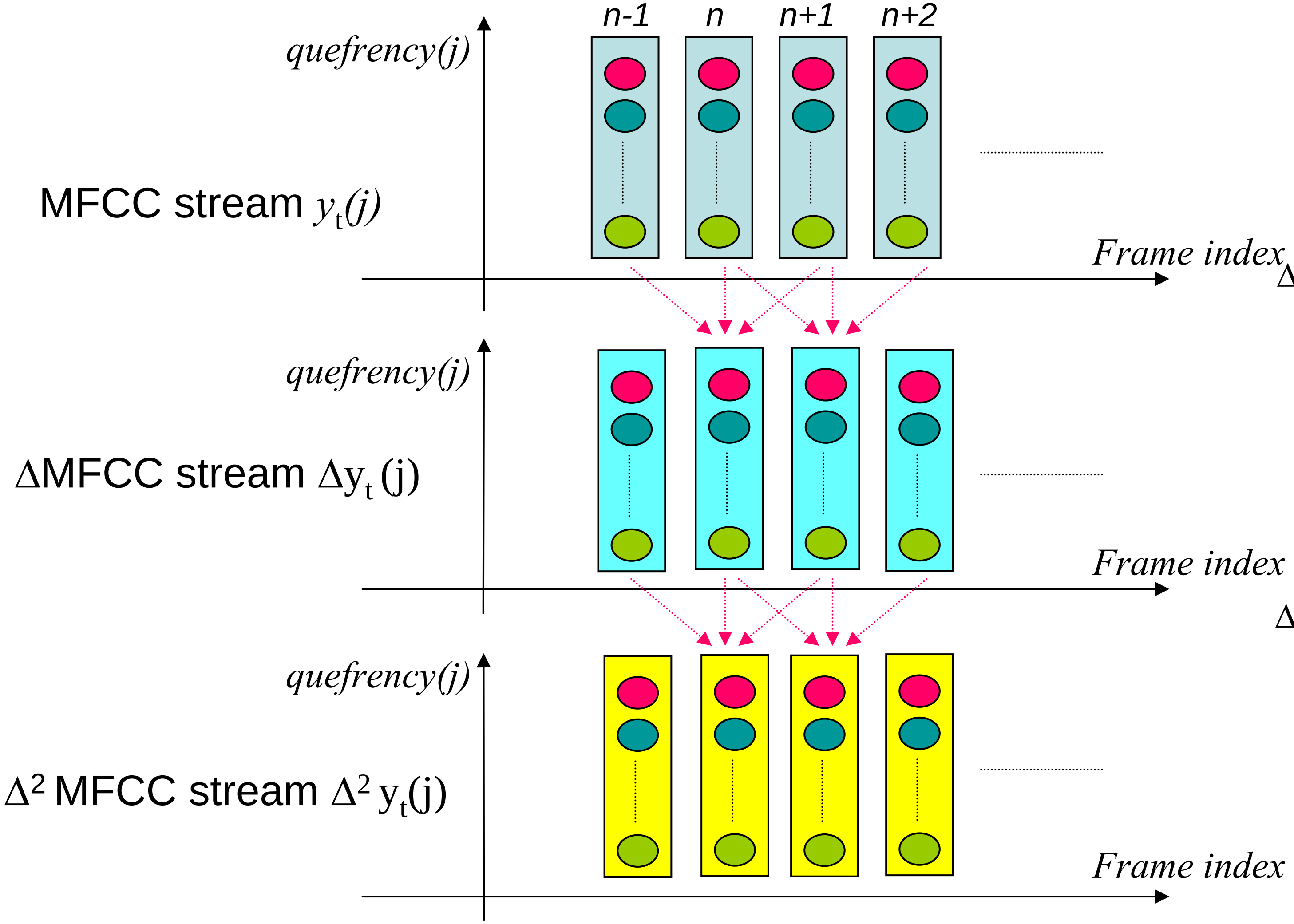
- **Final procedure for MFCC : performing the inverse DFT on the log-spectral power**

$$y_t(j)=\sum_{m=0}^{M-1}\log\left(\left|Y_t(m)\right|\right)\cos\left[j\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right],\quad j=0,1,\ldots,J-1<M$$

- **Advantages :**
  - Since the log-power spectrum is real and symmetric, the inverse DFT reduces to a Discrete Cosine Transform (DCT). The DCT has the property to produce highly uncorrelated features $\boldsymbol{y}_t$
    - diagonal rather than full covariance matrices can be used in the Gaussian distributions in many cases
  - Easier to remove the interference of excitation on formant structures
    - the envelope of the vocal tract changes slowly, while the excitation changes much faster

Derivatives

- **Derivative operation : to obtain the temporal information (change of the feature vectors with time)**



$$\Delta y_t(j) = \frac{\sum_{m=-p}^{p} m \bullet y_{t-m}(j)}{\sum_{m=-p}^{p} m^2}$$

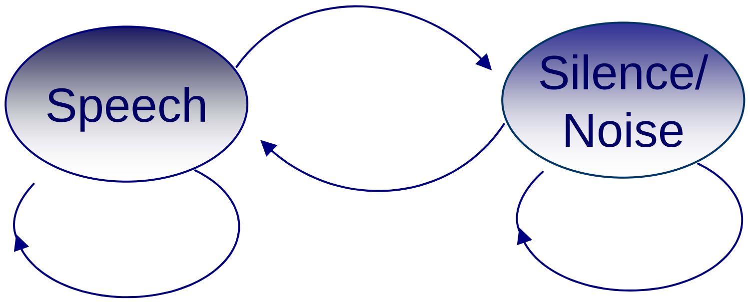
$$\Delta^2 y_t(j) = \frac{\sum_{m=-p}^{p} m \bullet \Delta y_{t-m}(j)}{\sum_{m=-p}^{p} m^2}$$

Why Delta Coefficients?

- **To capture the dynamic characters of the speech signal**
  - Such information carries relevant information for speech recognition
  - The value of p should be properly chosen
    - Too small P may imply too close frames and therefore the dynamic characters may not be properly extracted
    - Too large P may imply frames describing too different states
- **To cancel the DC part (convolutional noise) of the MFCC features**
  - For example, for clean speech, the MFCC stream is
    $\{\mathbf{y}(t\text{-}N), \mathbf{y}(t\text{-}N+1), \ldots\ldots, \mathbf{y}(t), \mathbf{y}(t+1), \mathbf{y}(t+2), \ldots\ldots\}$,
    while for a channel-distorted speech, the MFCC stream is
    $\{\mathbf{y}(t\text{-}N)+h, \mathbf{y}(t\text{-}N+1)+h, \ldots\ldots, \mathbf{y}(t)+h, \mathbf{y}(t+1)+h, \mathbf{y}(t+2)+h, \ldots\ldots\}$
    the channel effect h is eliminated in the delta (difference) coefficients

End-point Detection

- **Push (and Hold) to Talk/Continuously Listening**
- **Adaptive Energy Threshold**
- **Low Rejection Rate**
  – false acceptance may be rescued
- **Vocabulary Words Preceded and Followed by a Silence/Noise Model**
- **Two-class Pattern Classifier**



  – Gaussian density functions used to model the two classes
  – log-energy, delta log-energy as the feature parameters
  – dynamically adapted parameters