Speech Signal Representation

송치성

Intro

- The central theme is the decomposition of the speech signal as a source passed through a linear time-varying filter

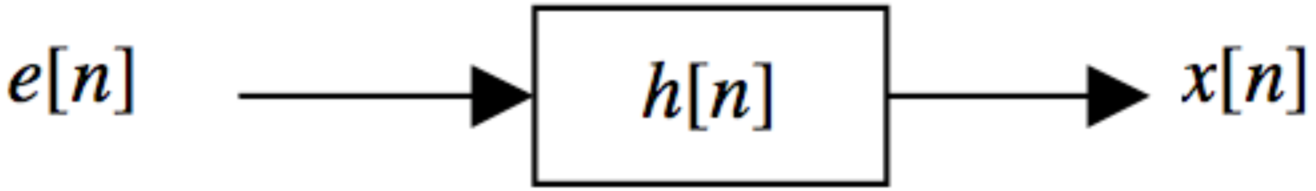


Figure 6.1 Basic source-filter model for speech signals.

- e[n] : the source or excitation
- h[n] : the filter
- x[n] : the speech signal

Contents

1. Short-Time Fourier Analysis
2. Acounstical Model of Speech Production
3. Linear Predictive Coding
4. Cepstral Processing
5. The Role of Pitch

Short-Time Fourier Analysis

- A spectrogram of a time signal is a special two-dimensional representation that displays time in its horizontal axis and frequency in its vertical axis.
- The idea behind a spectrogram is to compute a Fourier transform every 5 milliseconds or so, displaying the energy at each time/frequency point.
- (Z, W) , (H, G) : not periodic, looks like random noise.
  (The signal in (Z, W) appears to have different noisy characteristics than those of segment (H, G))

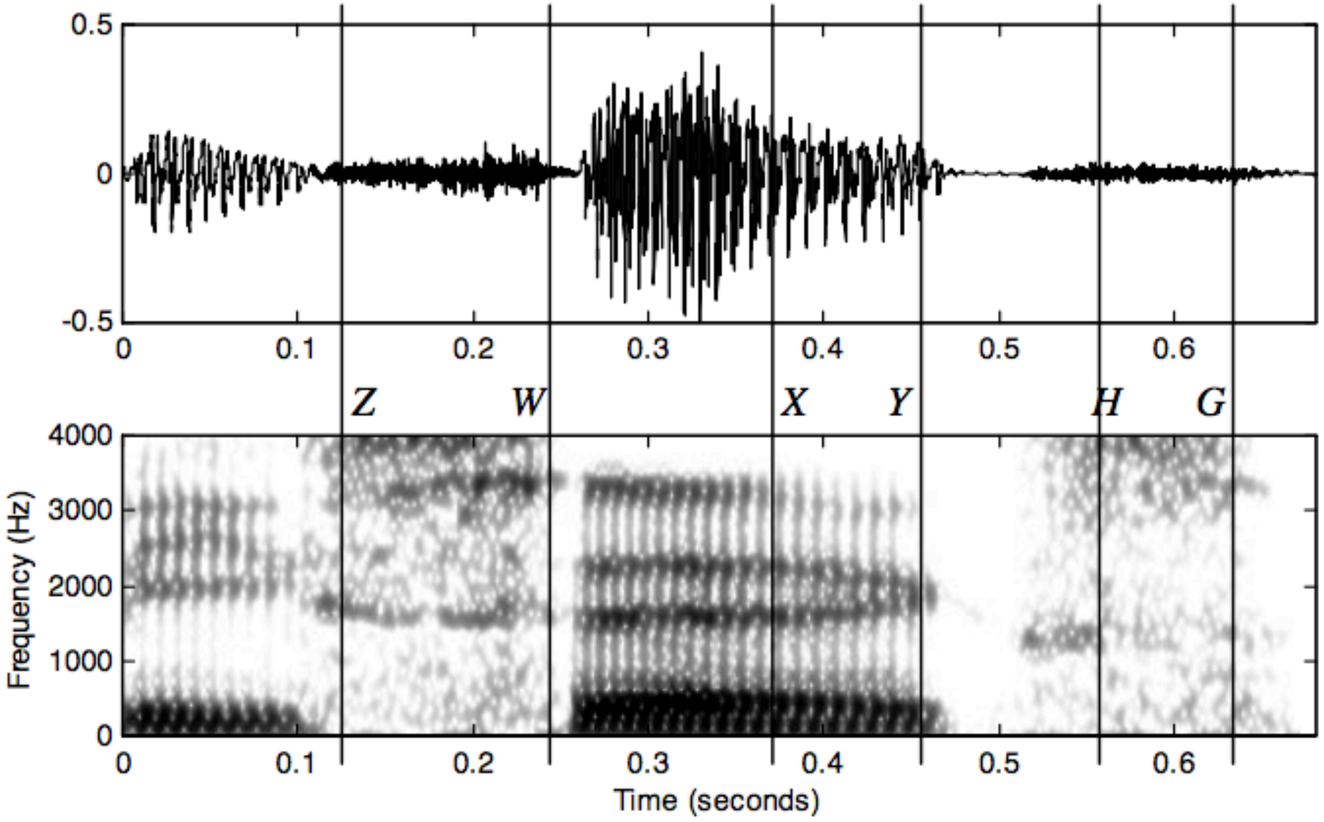


Figure 6.2 (a) Waveform with (b) its corresponding wideband spectrogram. Darker areas mean higher energy for that time and frequency. Note the vertical lines spaced by pitch peri-

Short-Time Fourier Analysis

- Wide-band :
  - 비교적 짧은 윈도우 (< 10ms) -> 시간 해상력이 좋음
  - Combines harmonics
  - Voiced speech vocal fold pulses show as vertical lines

- Narrow-band :
  - 비교적 긴 윈도우 (> 200ms) -> harmonics 볼수 없음
  - Individual harmonics
  - displays formants horizontally
  - No vocal pulses shown

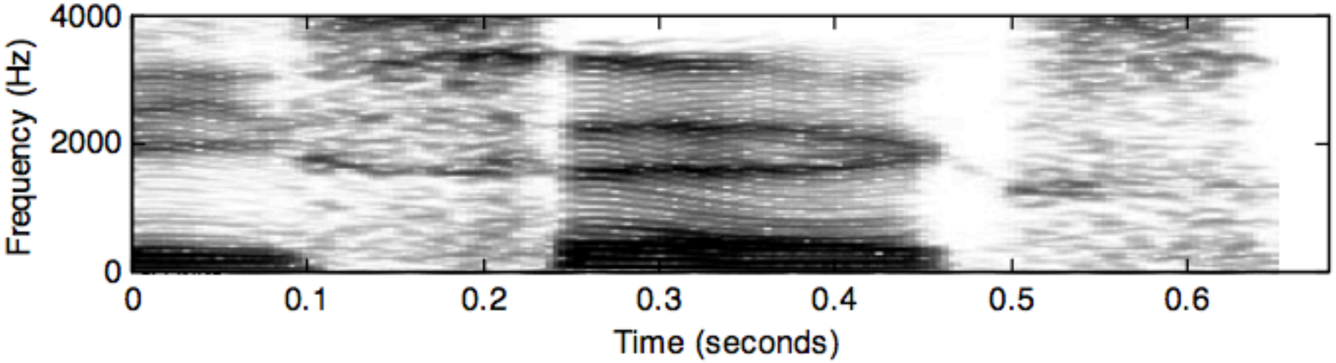
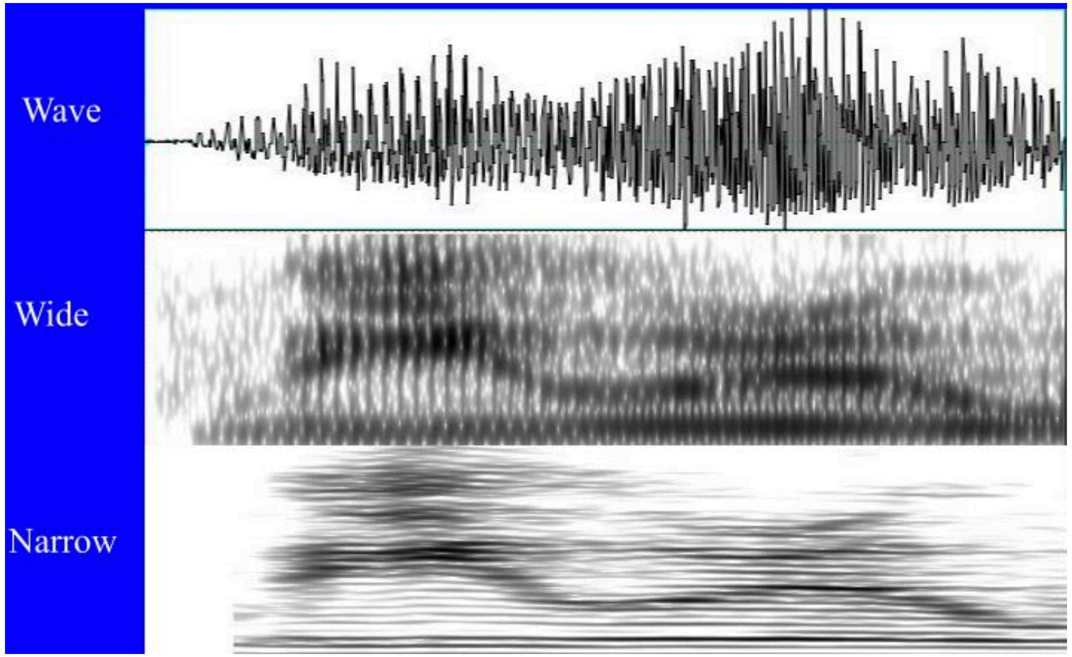


Figure 6.7 Waveform (a) with its corresponding narrowband spectrogram (b). Darker areas mean higher energy for that time and frequency. The harmonics can be seen as horizontal lines spaced by fundamental frequency. The corresponding wideband spectrogram can be seen in Figure 6.2.

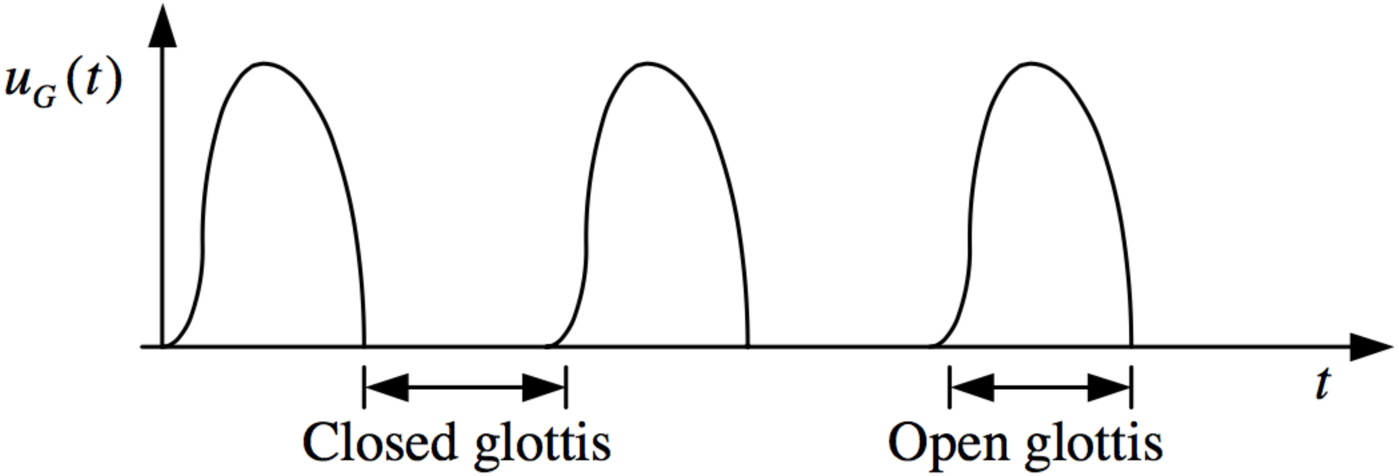


Acoustical Model of Speech Production

- Acoustic theory analyzes the laws of physics that govern the propagation of sound in the vocal tract.
- Such a theory should consider 3D wave propagation, the variation of the vocal tract shape with time, losses due to heat conduction and viscous friction at the vocal tract walls, softness of the tract walls, radiation of sound at the lips, nasal coupling and excitation of sound.

Acoustical Model of Speech Production

1) Glottal Excitation



- the vocal cords constrict the path from the lungs to the vocal tract.
- As lung pressure is increased, air flows out of the lungs and through the opening between the vocal cords (glottis).
- At one point the vocal cords are together, thereby blocking the airflow, which builds up pressure behind them. Eventually the pressure reaches a level sufficient to force the vocal cords to open and thus allow air to flow through the glottis.
- This condition of sustained oscillation occurs for voiced sounds.

Acoustical Model of Speech Production

2) Lossless Tube Concatenation

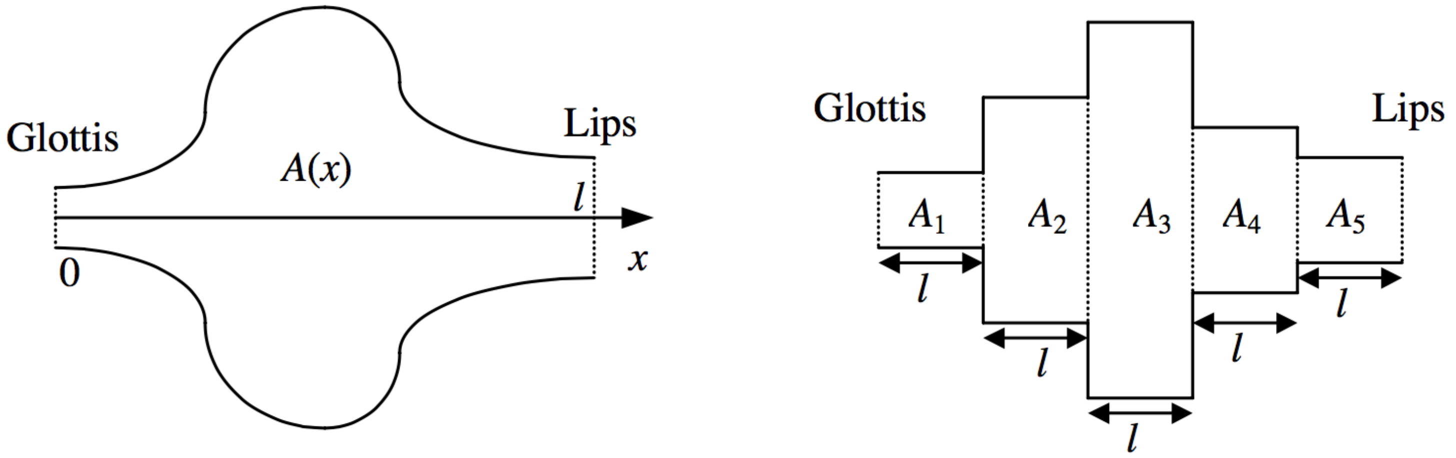


Figure 6.9 Approximation of a tube with continuously varying area $A(x)$ as a concatenation of 5 lossless acoustic tubes.

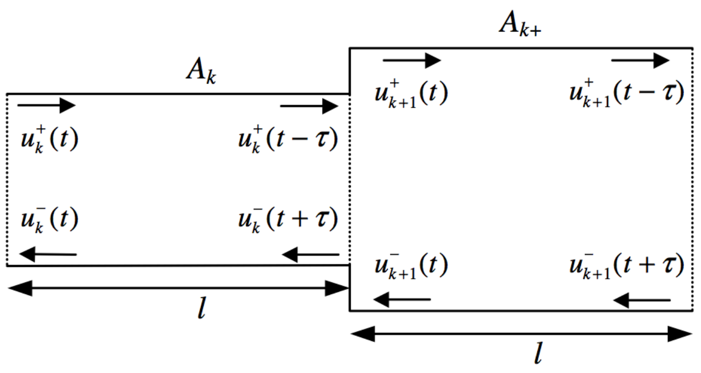


Figure 6.10 Junction between two lossless tubes.

- A widely used model for speech production is based on the assumption that the vocal tract can be represented as a concatenation of lossless tubes.
- The constant cross-sectional areas {Ak } of the tubes approximate the area function A(x) of the vocal tract.

Acoustical Model of Speech Production

3) Source-Filter Models of Speech Production

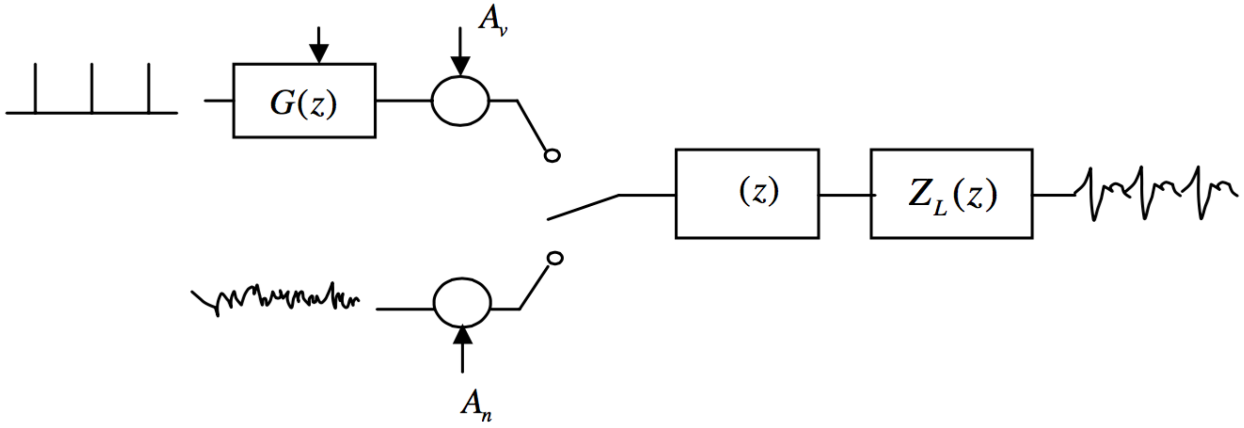
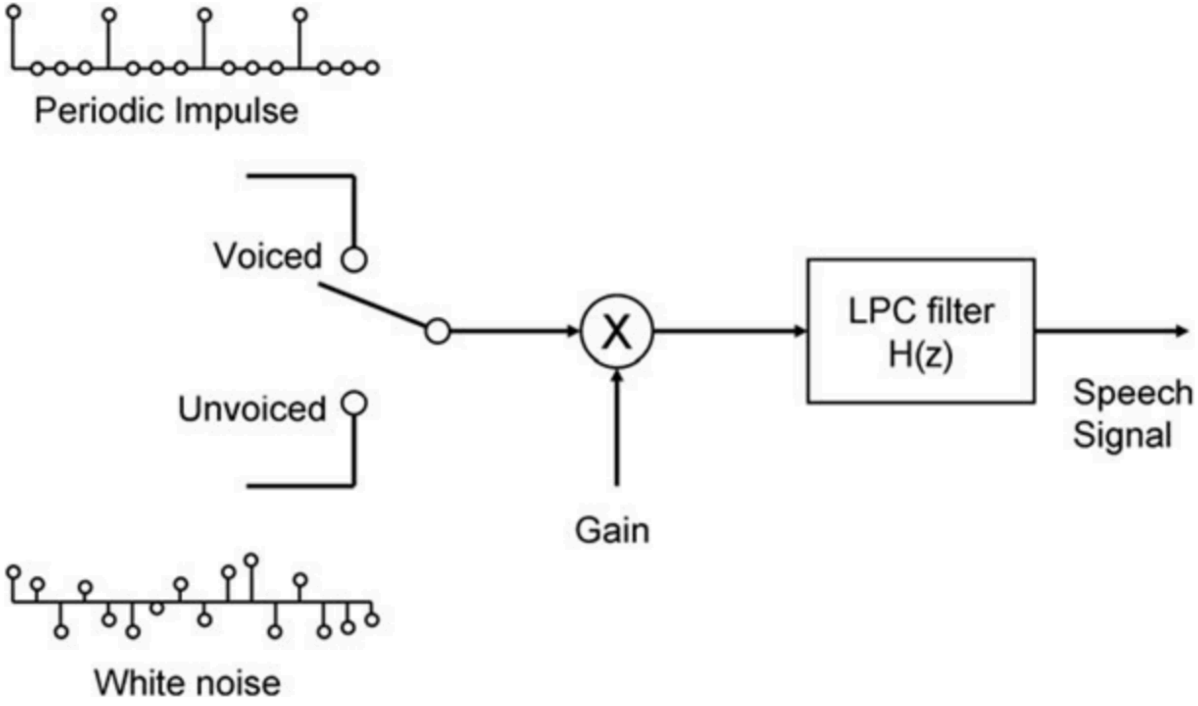


Figure 6.14 General discrete-time model of speech production. The excitation can be either an impulse train with period *T* and amplitude A_v driving a filter *G*(*z*) or random noise with amplitude A_n .

- speech signals are captured by microphones that respond to changes in air pressure. Thus, it is of interest to compute the pressure at the lips PL (z) , which can be obtained as

$$P_L(z) = U_L(z)Z_L(z) = U_G(z)V(z)Z_L(z)$$

Linear Predictive Coding



- 기본 가정 :
  - 현재의 음성 샘플이 이전의 음성 샘플들의 근사적인 선형 결합으로 표현 가능 .
  - 현재 음성 표본값을 과거의 표본값들로부터 예측하고, 그 잔차(차분) 성분 만을 부호화 -> 데이터가 작아지므로 압축도 가능

- LPC 해석: 성도의 주파수 특성을 변화시켜가면서 각각 다른 음성을 발생시키는 과정을 분석하여, 유성음 및 무성음에 따라 입력 신호의 크기 또는 주기 등의 변화에 대한 각종 계수를 구하는 과정
- 특징 :
  - 음성 주파수 스펙트럼 상에서의 특징을 상대적으로 적은 수의 파라미터 만으로 비교적 정확하게 표현 가능
  - 선형 예측 분석에 따른 계산량이 크지 않음

Cepstral Processing

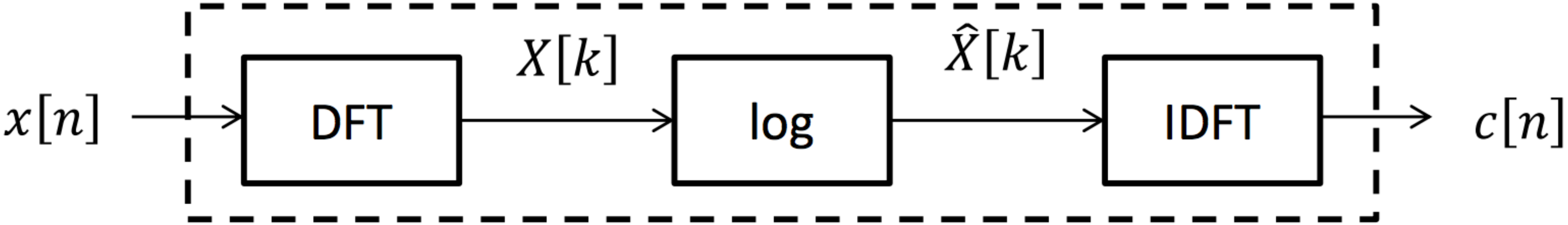
- The cepstrum is defined as the inverse DFT of the log magnitude of the DFT of a signal

$$c[n] = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]\}|\}$$

  - where $\mathcal{F}$ is the DFT and $\mathcal{F}^{-1}$ is the IDFT

- For a windowed frame of speech $y[n]$, the cepstrum is

$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}\right|\right) e^{j\frac{2\pi}{N}kn}$$



Cepstral Processing

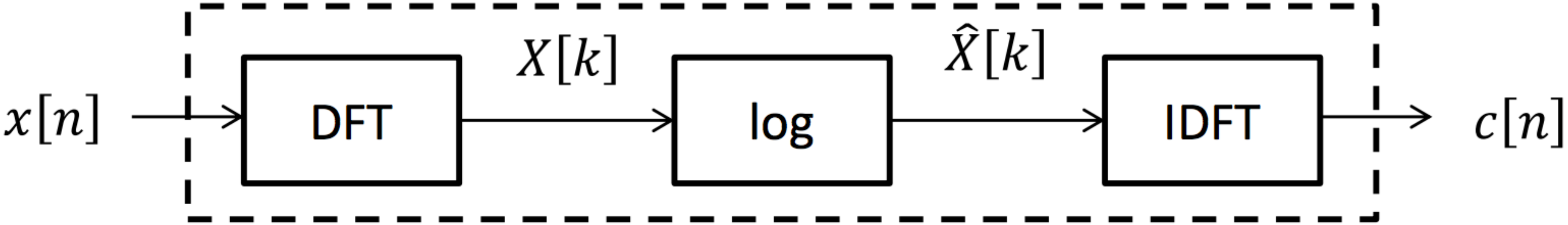
- The cepstrum is defined as the inverse DFT of the log magnitude of the DFT of a signal

$$c[n] = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]\}|\}$$

  - where $\mathcal{F}$ is the DFT and $\mathcal{F}^{-1}$ is the IDFT

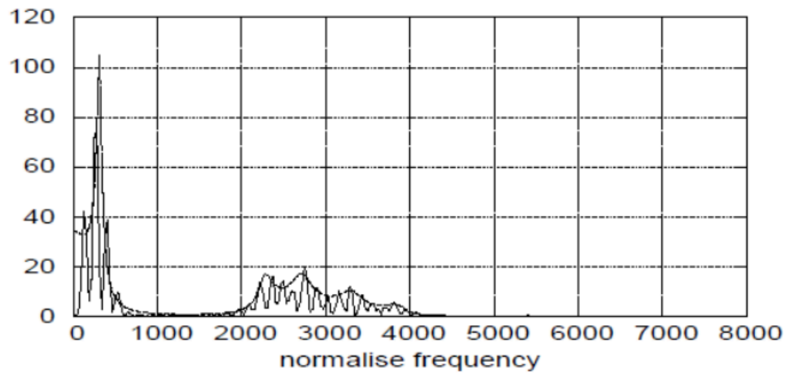
- For a windowed frame of speech $y[n]$, the cepstrum is

$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}\right|\right) e^{j\frac{2\pi}{N}kn}$$

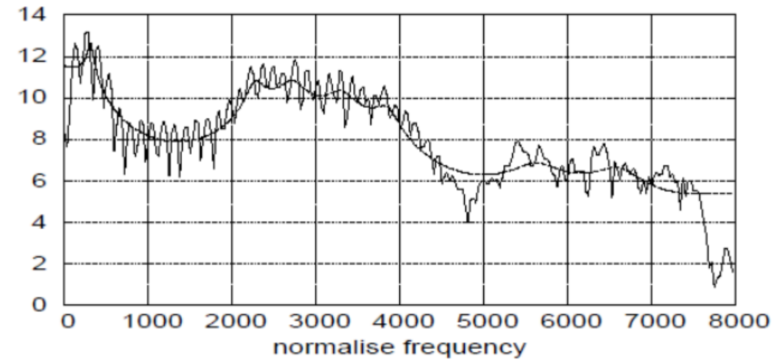


Cepstral Processing

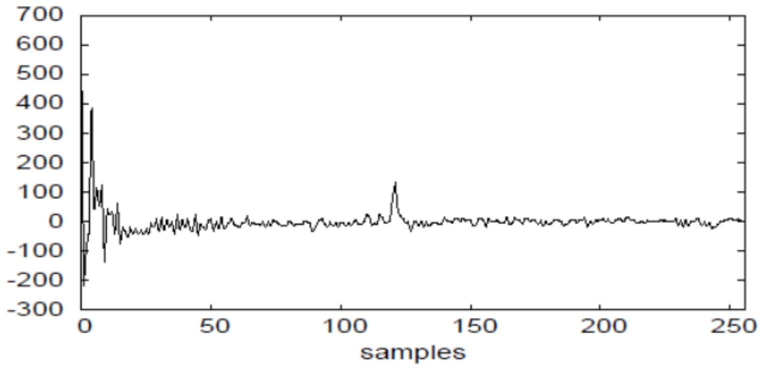
$\mathcal{F}\{x[n]\}$



$\log|\mathcal{F}\{x[n]\}|$



$\mathcal{F}^{-1}\{\log|\mathcal{F}\{x[n]\}|\}$



[Taylor, 2009]

Cepstral Processing

MFCC (Mel-Frequency Cepstrum Coefficients)

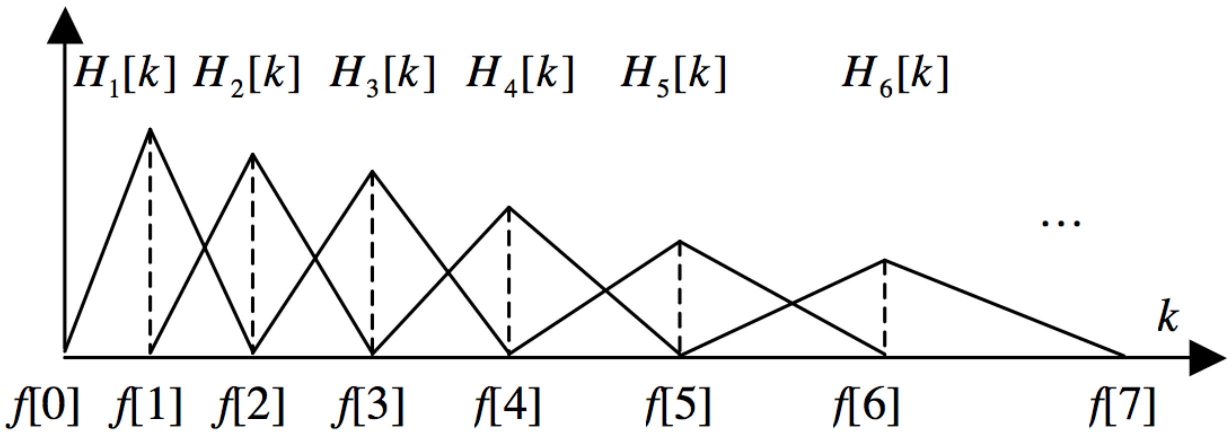


Figure 6.28 Triangular filters used in the computation of the mel-cepstrum using Eq. (6.140).

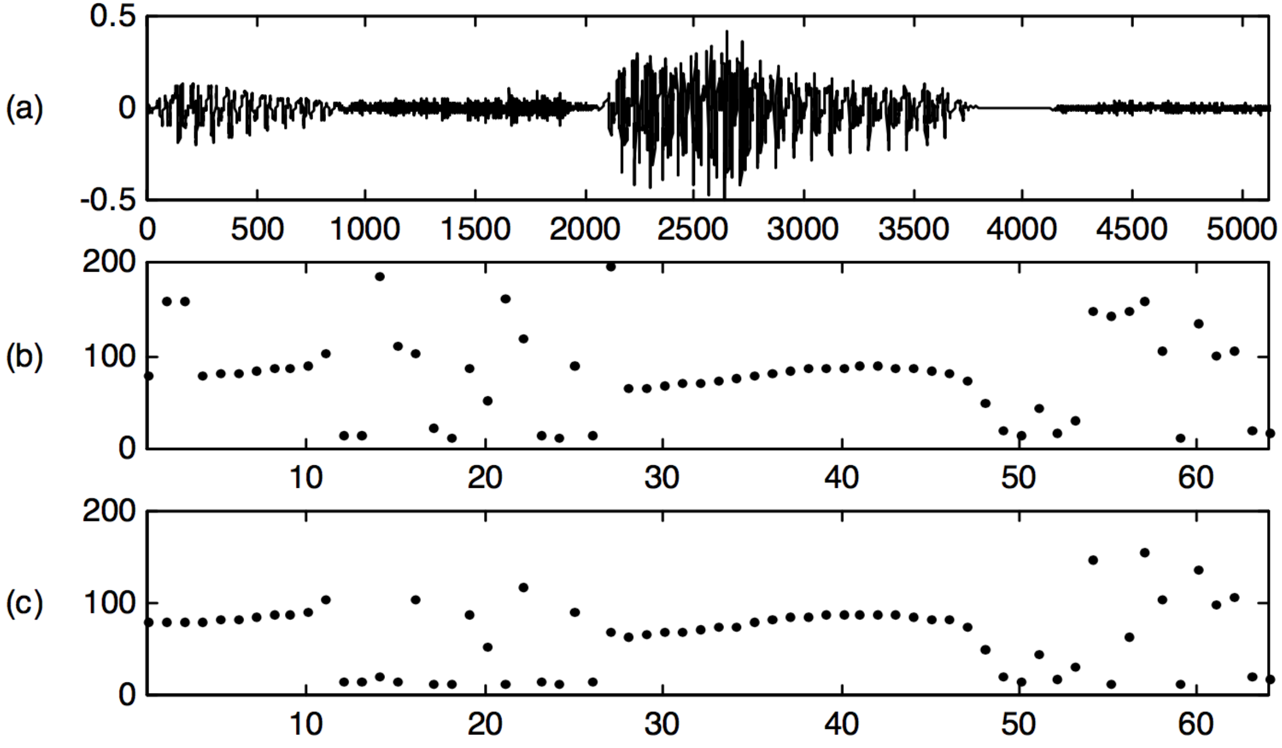
- A representation defined as the real cepstrum of a windowed short-time signal derived from the FFT of that signal.
- The difference from the real cepstrum is that a nonlinear frequency scale is used, which approximates the behavior of the auditory system.
- Davis and Mermelstein showed the MFCC representation to be beneficial for speech recognition.

The Role of Pitch

- Pitch determination is very important for many speech processing algorithms.
- Chinese speech recognition systems use pitch tracking for tone recognition, which is important in disambiguating the myriad of homophones.
- Pitch is also crucial for prosodic variation in text-to-speech systems (see Chapter 15) and spoken language systems (see Chapter 17).

The Role of Pitch

- (a) Waveform and (b) (c) unsmoothed pitch tracks with the normalized cross-correlation method. A frame shift of 10 ms, window length of 10 ms, and sampling rate of 8 kHz were used.
- (b) : standard normalized cross-correlation method
- (c) : having decaying term.



- If we compare it to the autocorrelation method of Figure 6.31, the middle voiced region is correctly identified in both (b) and (c), but two frames at the beginning of (b) that have pitch halving are eliminated with the decaying term.
- the pitch values in the un- voiced regions are essentially random.