Signal Processing in Speech

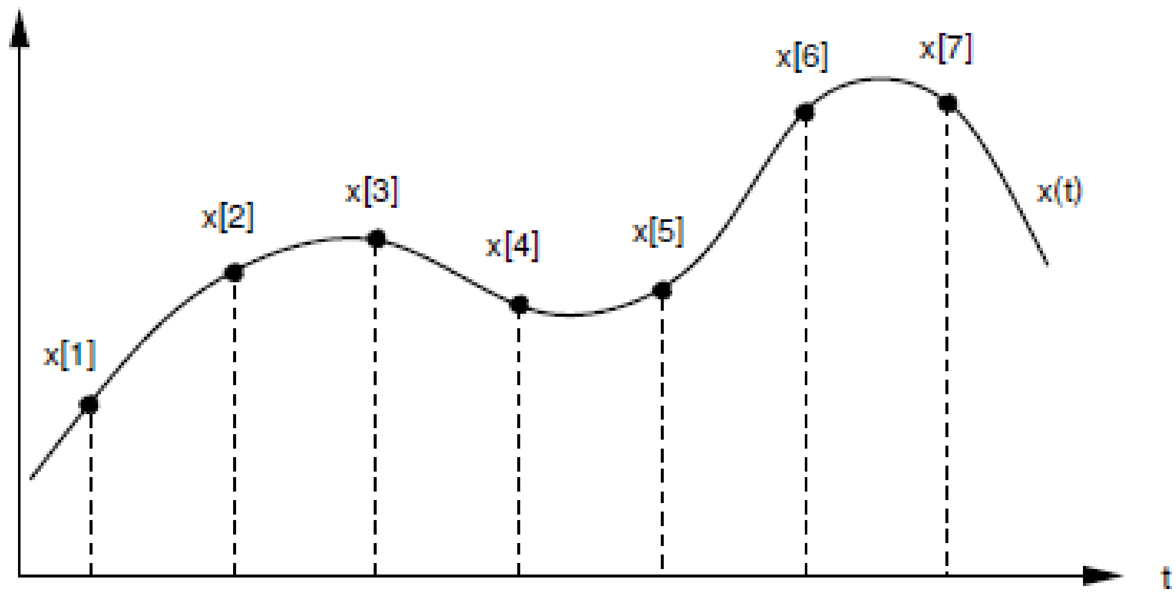
Instructor: Tom Ko

Objectives

- Learn signal processing techniques

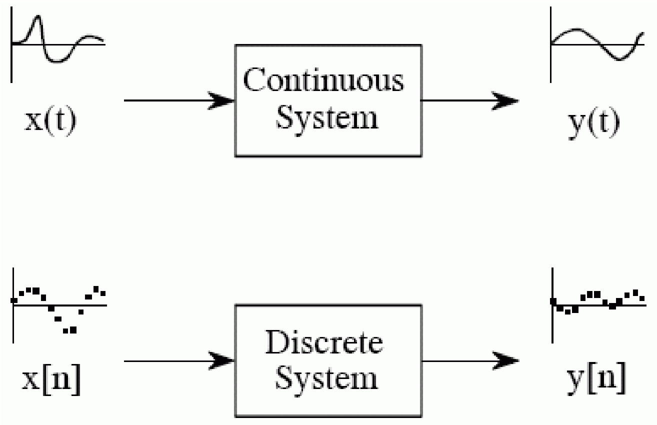
Signals

- A **signal** is a description of how one parameter varies with another parameter.



Systems

- A **system** is any process that produces an *output signal* in response to an *input signal*.



Signals and systems

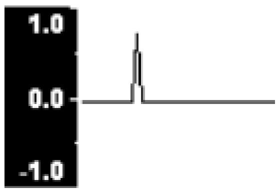
- Any channel that transmits a signal can be regarded as a system. The signals are transformed.
  - The signals transmitted from your vocal fold to your mouth. The vocal tract is considered as a system.
  - The voice signal transmitted from your mouth to my ear. The air in the room is considered as a system.

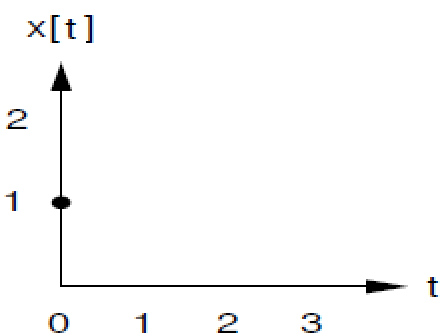
Linear time-invariant (LTI) system

- *Linearity* means that the relationship between the input and the output of the system is a linear map.
  - If x(t) produces y(t), $\sum_k c_k x_k(t)$ produces $\sum_k c_k y_k(t)$
- *Time invariance* means that whether we apply an input to the system now or *T* seconds from now, the output will be identical except for a time delay of *T* seconds.
  - If x(t) produces y(t), x(t−T) produces y(t−T)

An Impulse

$$\text{if } x[t] = \delta[t] = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}$$





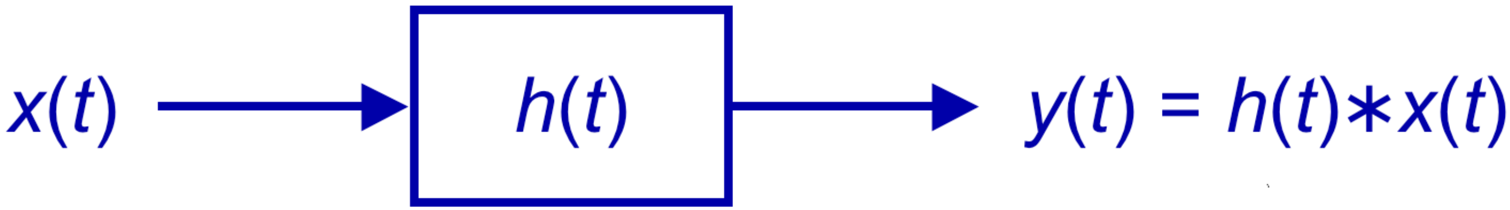
Impulse response

- Is there any way to measure the response of a system?
- In signal processing, the **impulse response**, of a dynamic system is its output when presented with a brief input signal, called an impulse.
- A signal is considered as a sequence of impulses.

$$x(t) \longrightarrow \boxed{h(t)} \longrightarrow y(t) = h(t) \ast x(t)$$

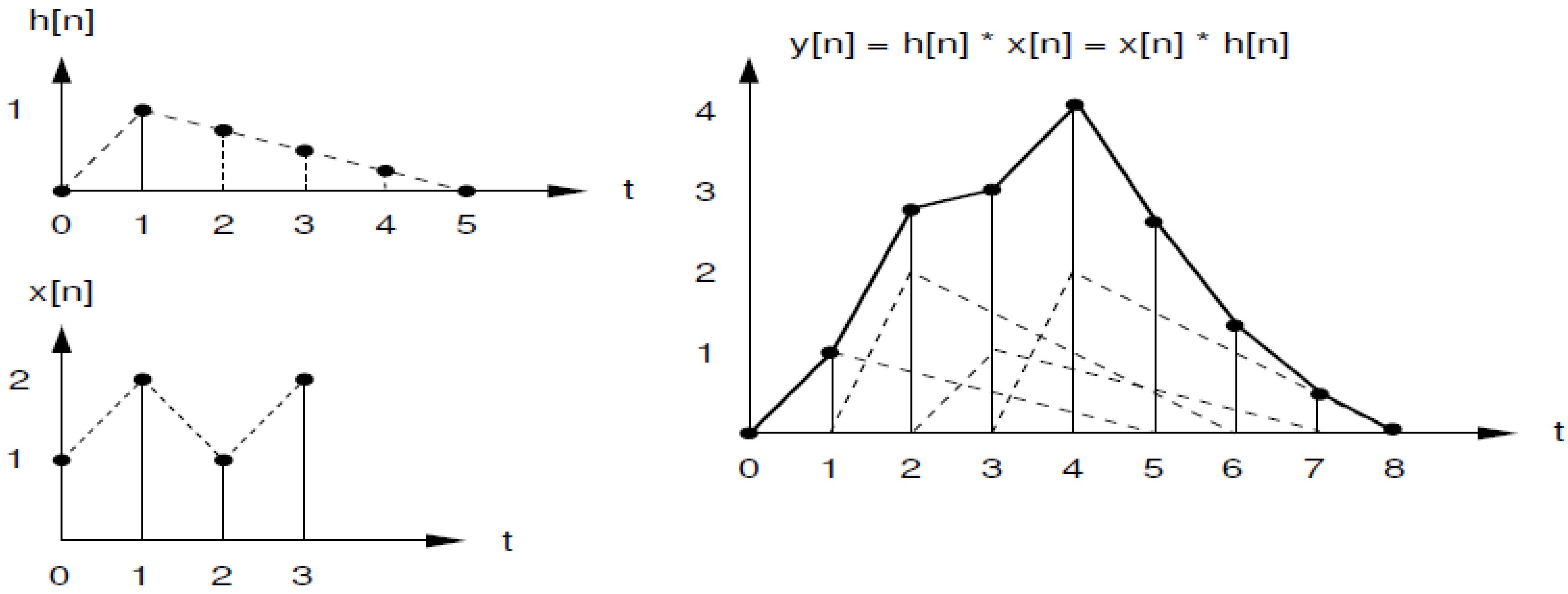
Impulse response

- Giving the system an impulse, the output is its impulse response



- When x(t) contains only an impulse, y(t) = h(t)

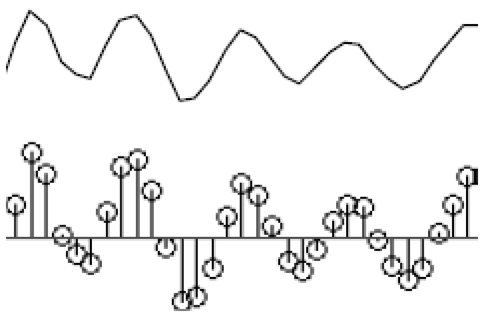
Convolution



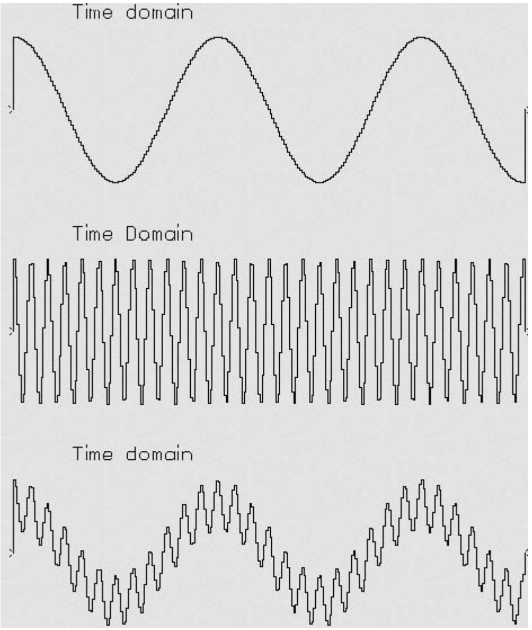
$$y[n] \;=\; x[n] * h[n] = \sum_{m=-\infty}^{\infty} x[m]\, h[n-m]$$

Superposition of signals

- A signal can be considered as a train of impulses.



- It can also be regarded as a superposition of individual signals (components) with different frequencies.



Sinusoidal signals

- One of the most important signals is the sine wave or *sinusoid*

$$x_0[n] = A_0 \cos(\omega_0 n + \phi_0)$$

  - where A_0 is the sinusoid's amplitude, ω_0 the angular frequency (in radians) and $\varnothing_0$ the phase.

Complex number representation

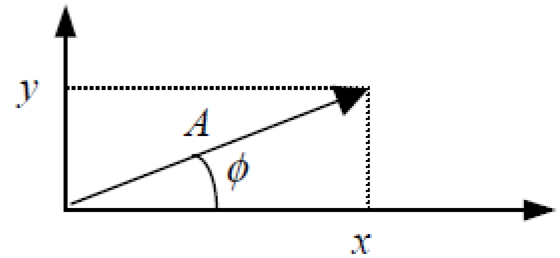


Figure 5.4 Complex number representation in Cartesian form $z = x + jy$ and polar form $z = Ae^{j\phi}$. Thus $x = A\cos\phi$ and $y = A\sin\phi$.

- **Euler's formula** establishes the fundamental relationship between the trigonometric functions and the complex exponential function. It states that

$$e^{j\phi} = \cos\phi + j\sin\phi$$

- The sinusoid signal can be expressed as

$$x_0[n] = A_0 \cos(\omega_0 n + \phi_0) = \mathrm{Re}\{A_0 e^{j(\omega_0 n + \phi_0)}\}$$

Sum of two sinusoid signals

- Sum of two sinusoid signals with same frequency but different amplitude and phase

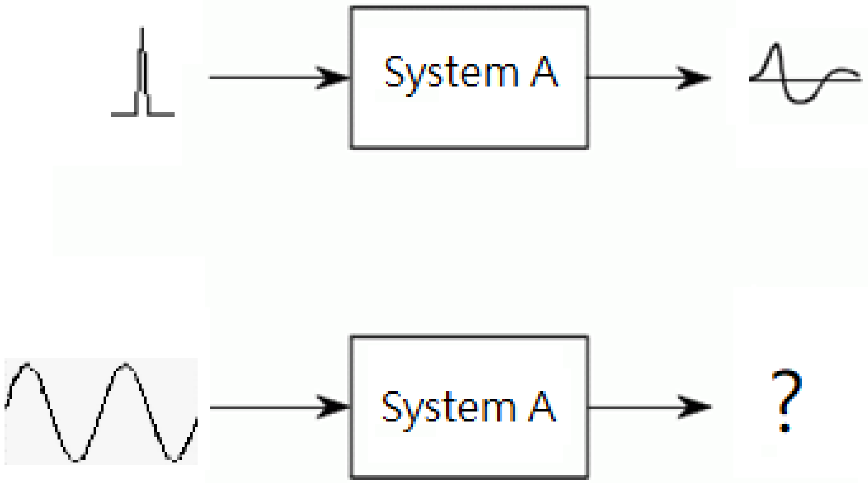
$$A_0 e^{j(\omega_0 n+\phi_0)} + A_1 e^{j(\omega_0 n+\phi_1)} = e^{j\omega_0 n}\left(A_0 e^{j\phi_0} + A_1 e^{j\phi_1}\right) = e^{j\omega_0 n} A e^{j\phi} = A e^{j(\omega_0 n+\phi)}$$

- Taking the real part on both side:

$$A_0 \cos(\omega_0 n + \phi_0) + A_2 \cos(\omega_0 n + \phi_1) = A\cos(\omega_0 n + \phi)$$

- Thus, **the sum of two sinusoids of the same frequency is another sinusoid of the same frequency** with a new amplitude and a new phase.
  - For the computation of the new amplitude and phase, please refer to the book *Spoken Language Processing: A Guide to Theory, Algorithm and System Development, ch 5.1.1*

Sinusoidal signals to LTI system



Sinusoidal signals to LTI system

- Now we want to see the output of a LTI system with impulse response *h*[*n*] when the input is a complex exponential.
- Substituting $x[n] = e^{j\omega_0 n}$ in the convolution formula:

$$y[n] = \sum_{k=-\infty}^{\infty} h[k] e^{j\omega_0 (n-k)} = e^{j\omega_0 n} \sum_{k=-\infty}^{\infty} h[k] e^{-j\omega_0 k} = e^{j\omega_0 n} H(e^{j\omega_0})$$

  - which is another complex exponential of the same frequency where the amplitude is multiplied by the complex quantity $H(e^{j\omega_0})$ given by

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h[n] e^{-j\omega n}$$

Eigensignals

- Since the output of a LTI system to a complex exponential is another complex exponential, it is said that complex exponentials are *eigensignals* of LTI systems, with the complex quantity $H(e^{j\omega_0})$ being their *eigenvalue*.

Discrete-time Fourier transform

- The quantity $H(e^{j\omega})$ is defined as the *discrete-time Fourier transform* of the impulse response h[n].
- It is a function of ω, which is the frequency of the input sinusoid signal.
  - Input signals with different frequency will result in different response.
  - The Fourier transform transforms the function from time domain to frequency domain.
- The Fourier transform $H(e^{j\omega})$ of *h*[*n*] is called the system's *frequency response* or *transfer function*.

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h[n]e^{-j\omega n}$$

Frequency response and filters

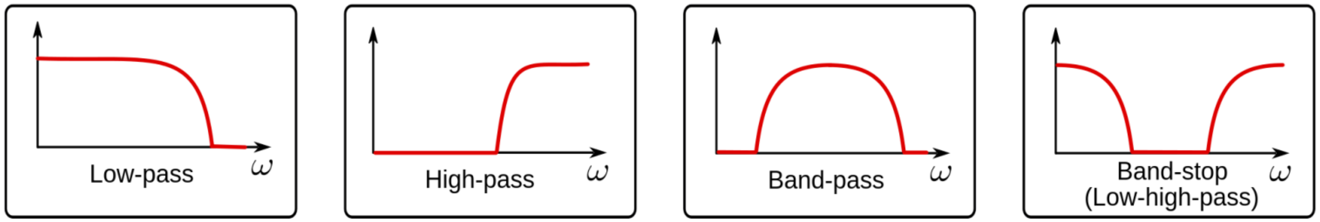
- The Fourier transform results in a complex function , it could be expressed in terms of its polar form (magnitude and phase):

$$H(e^{j\omega}) = \left|H(e^{j\omega})\right| e^{j\arg[H(e^{j\omega})]}$$

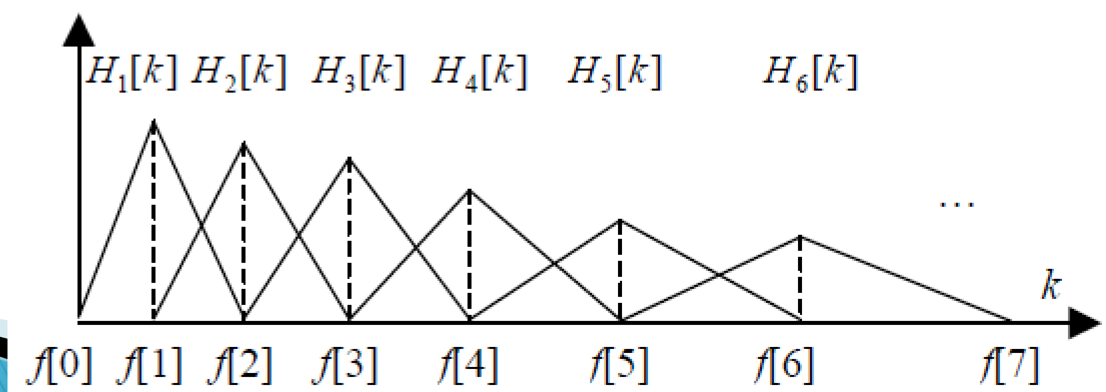
  - The magnitude and phase are also functions of w.
- If $\left|H(e^{j\omega_0})\right| > 1$, the LTI system will amplify that frequency.
- If $\left|H(e^{j\omega_0})\right| < 1$, the system will filter that frequency.
- That is one reason why these systems are also called filters.

Filters and filter bank

- A **filter** is a device or process that removes some unwanted frequency components from a signal.



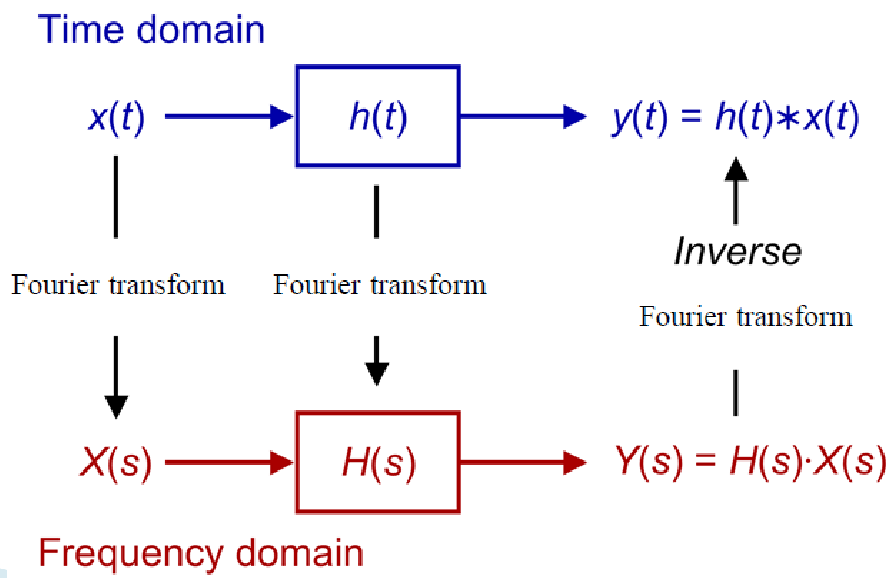
- A **filter bank** is an array of band-pass filters that separates the input signal into multiple components



Property of Fourier transform

- **The convolution theorem:** Let $\mathcal{F}\{\}$ denotes the Fourier transform operator, then
  - $\mathcal{F}\{x[n] * h[n]\} = \mathcal{F}\{x[n]\}\,\mathcal{F}\{h[n]\}$

- The Fourier transform of a convolution of two signals is the product of their individual Fourier transforms.



Fourier transform of signals

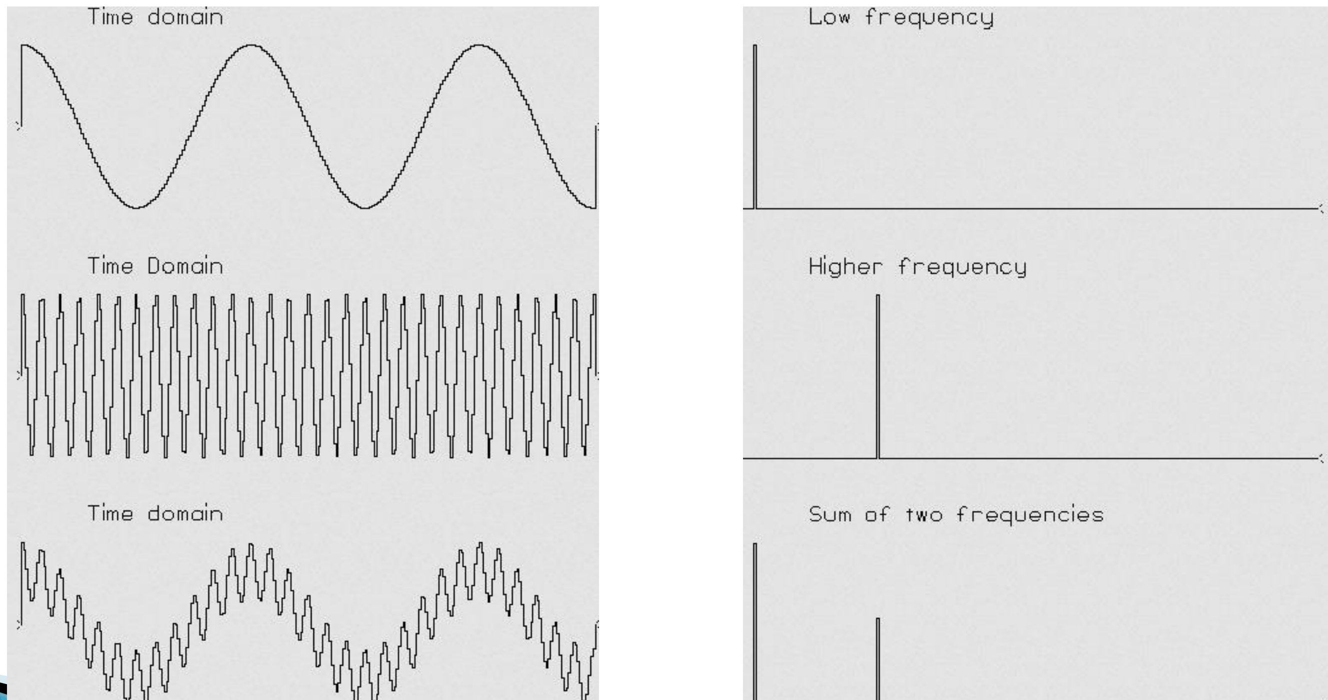
- It is shown that Fourier transform is applied on the impulse response of a system, but it could be applied on any signal.
- When Fourier transform is applied on a signal x(t):

$$x(t) \overset{\mathcal{F}}{\longleftrightarrow} X(j\omega)$$

- $|X(j\omega)|$ determines the relative presence of a sinusoid $e^{j\omega t}$ in x(t)
- $\angle X(j\omega)$ determines how the sinusoids line up relative to one another to form x(t)

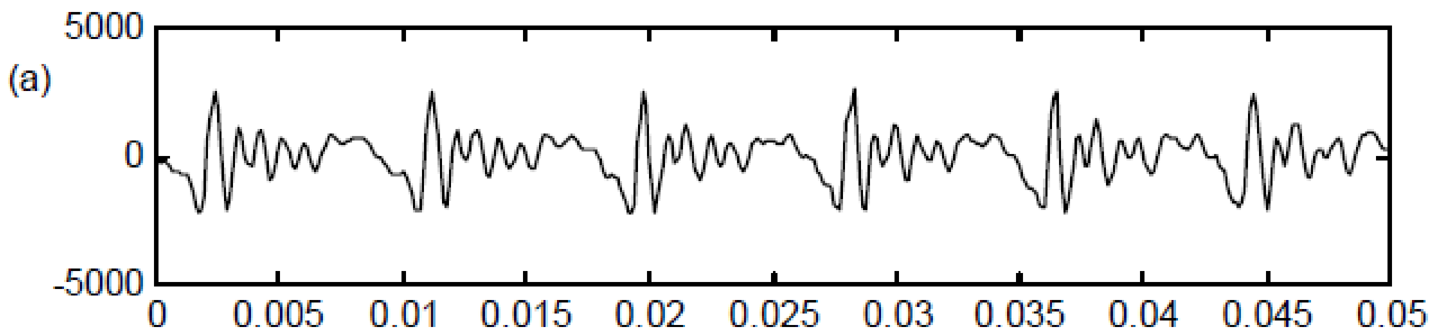
Fourier magnitude

- We can look the strength of each frequency component in the signal.

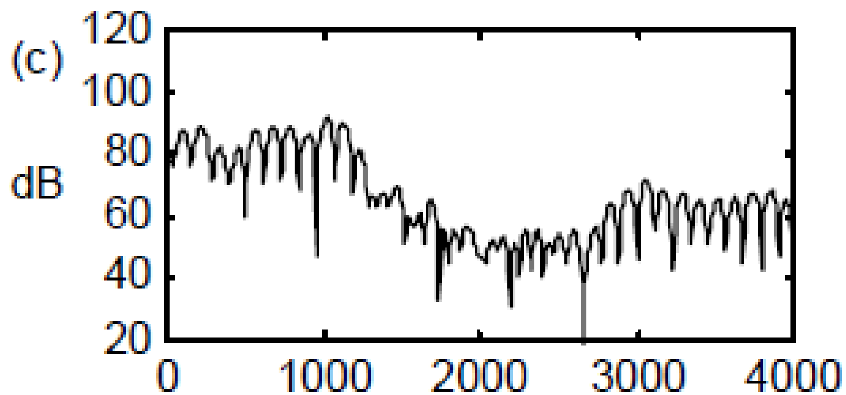


Frequency spectrum

- A signal in time domain:



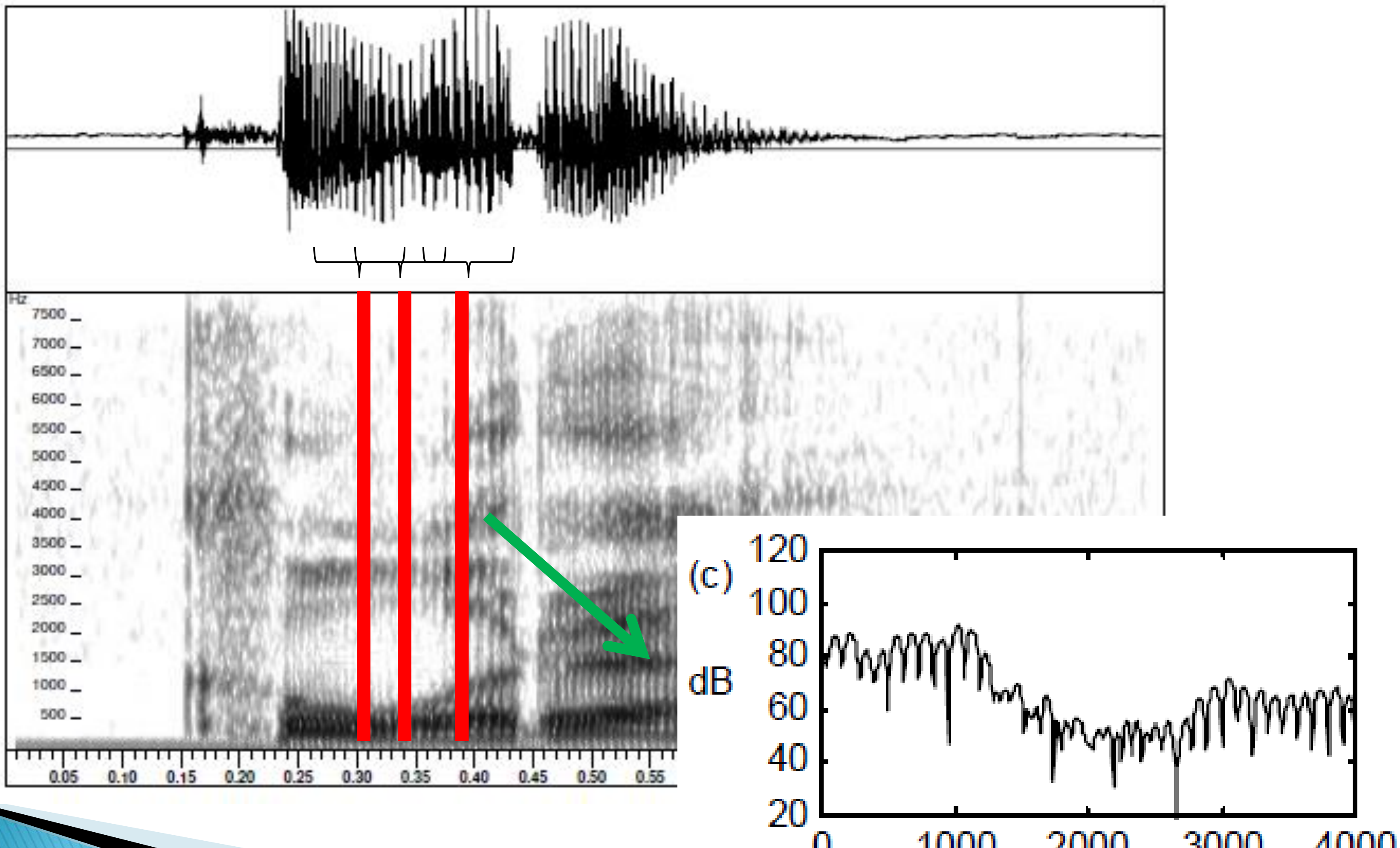
- Its form in frequency domain:



Short-time Fourier Analysis

- We have dealt with periodic signals in our formulation, however, the signal is no longer periodic when longer segments are analyzed.
- **Short-time analysis**: a speech signal is decomposed into a series of short segments.
- In each segment, the signal is assumed to be *stationary.*
  - The region has to be short enough

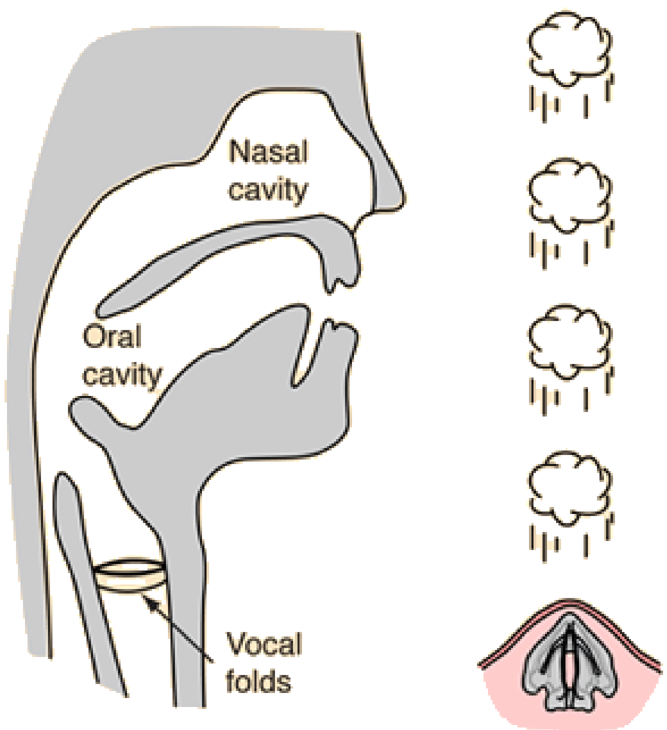
Short-time Fourier Analysis



Spectrogram

- 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.
- A gray scale is typically used to indicate the energy at each point (t, f) with white representing low energy and black high energy.

Speech Production System

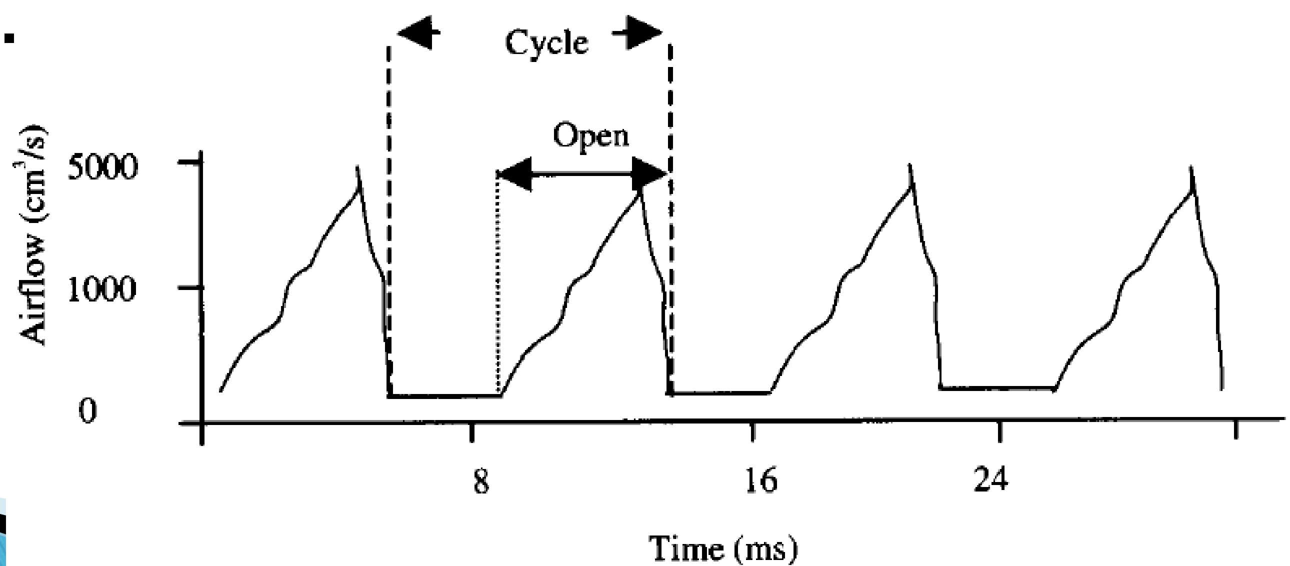


Schematic View of Vocal Tract

- The vocal folds generate periodic impulses.
- The vocal tract acts like a filter of which the impulse response convolutes with the impulses to form the sound.
- The impulse response changes with the shape of the tract.
- Production-based features encode the shape of vocal tract from the signal.

Excitation from Vocal Cord

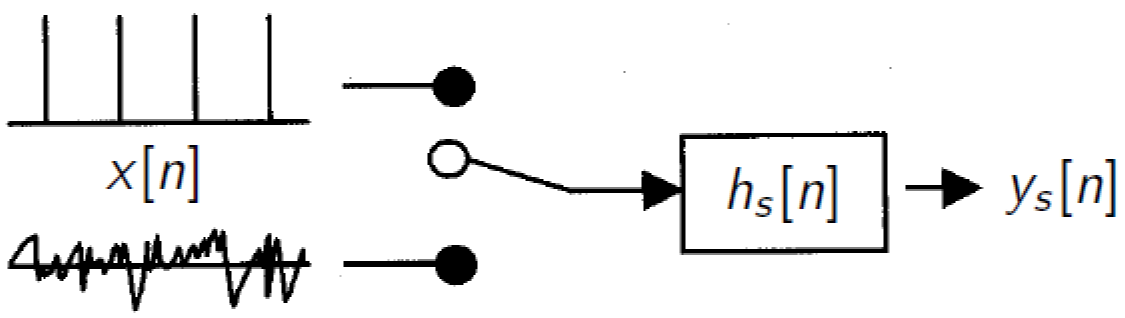
- To produce a voiced sound, the vocal cords open and close and produce a train of impulses.
- Its frequency is called the fundamental frequency F0 which gives the perception of pitch.



Source-filter model

- Human speech production system is modeled by the source-filter model
  - Voiced speech is excited by a periodic train of impulses
  - unvoiced speech is excited by random white noises
- Each distinct speech sound, s, has its own distinct filter (shaped by the vocal tract) represented by its impulse response h_s [n]
- The output speech y_s [n] of a sound s is the result of convolution between its excitation x[n] and its impulse response h_s [n].

$$y_s[n] = x[n] * h_s[n]$$



Fundamental frequency (F0)

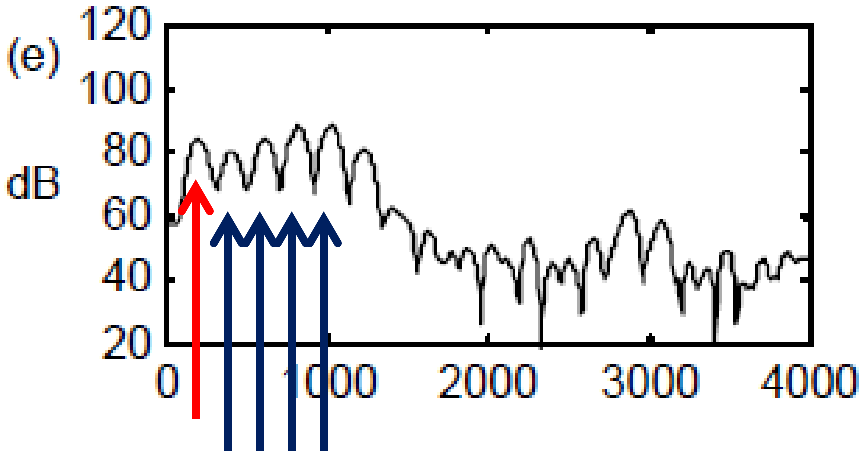
- It is defined as the lowest frequency of a periodic waveform.
- It is related to the pitch, but not exactly.
  - Pitch properly refers to a percept rather than a parameter of speech production.
- It can be defined as rate of the vocal fold vibration.
- Fundamental frequency range:
  - A typical adult male: 85 to 180 Hz,
  - A typical adult female: 165 to 255 Hz
  - Children: 200 to350 Hz

Harmonic

- The human voice is not a pure tone (as produced by a tuning fork).
- Human voice is composed of a fundamental tone and its upper harmonics.
    - Upper harmonics are multiple of the fundamental frequency.
- As long as the harmonics are precise multiples of the fundamental, the voice will sound clear and pleasant.
- If non-harmonic components are added, hoarseness will be perceived in relation to the intensity of the noise components in the frequency spectrum.

Harmonic

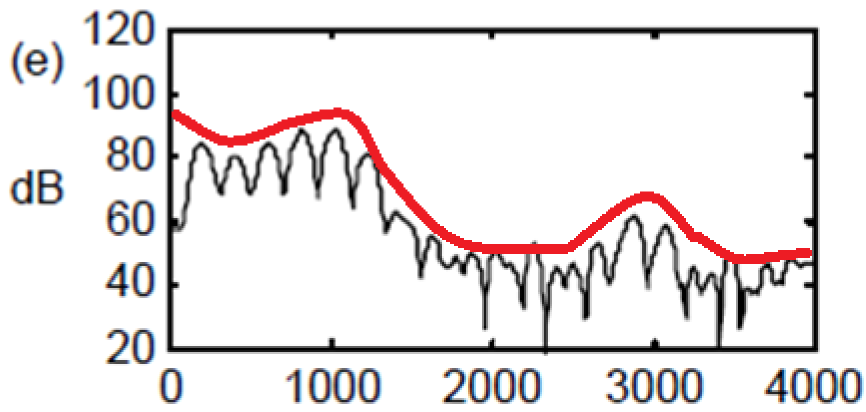
- Short-time spectrum of female voiced speech (vowel /aa/ with F0 of 200Hz):



- Now recall the frequency response of a filter, some frequency components will be amplified, some will be reduced.

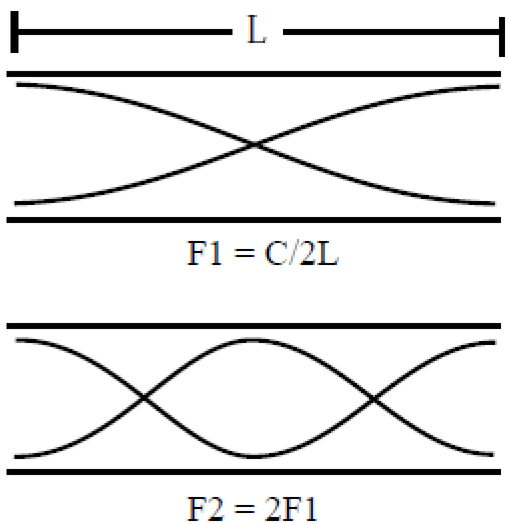
Formants

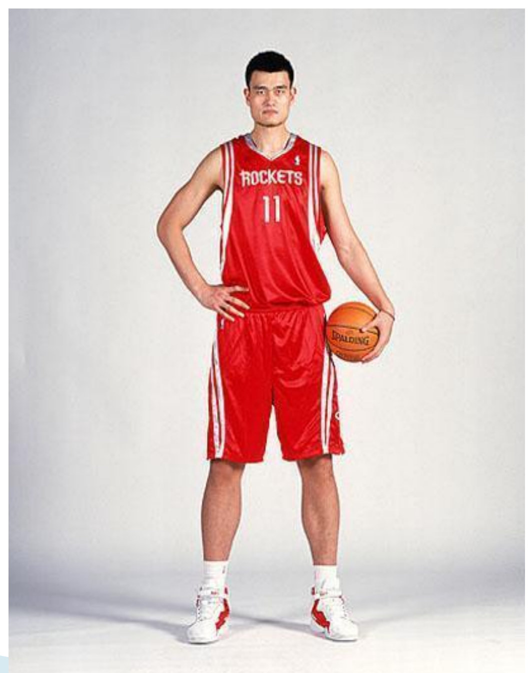
- When the wave travels in the vocal tract, resonance occurs.
- The shape of these tubes determines the resonance frequencies, called formants: F1 is the 1st formant, F2 is the 2nd formant, etc.
- A formant can be defined as a peak, or local maximum, in the spectrum.
- They determines the vowels that you hear.



Effect of Vocal Tract Length

- It has been well-established that longer vocal tracts are associated with lower formant frequencies.
- This affect the pitch you perceive.





Vowels and consonants

- Vowels and consonants are two principal classes of speech sounds.
  - Vowels are voiced. You can change the pitch when you pronounce it.
  - Consonants are unvoiced. They are produced when there is significant constriction or obstruction in the vocal tract.
- Vowel examples: aa ih uw ae ao
- Consonant examples: b, p, g, d, t, f, s, sh

Words	far	ill	mood	gas	all
Phonetic transcription	f aa r	ih l	m uw d	g ae s	ao l

Vowels and consonants

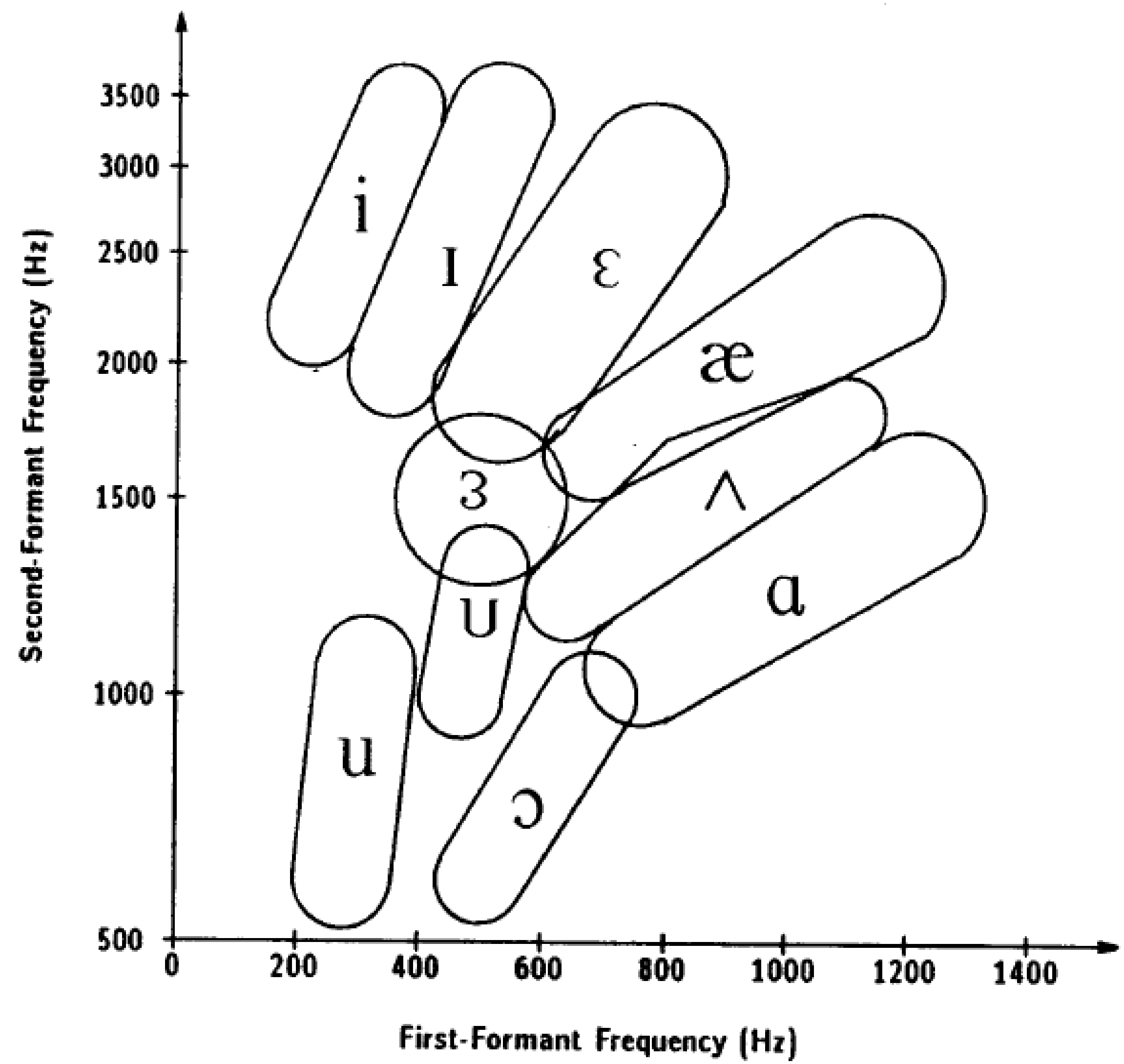
Vowel	
basic vowel:	No obstruction in the vocal tract; resonance iy, ih, ae, aa, ah, ao, ax, eh, er, ow, uh, uw
diphthongs:	Moving from one basic vowel to another. ay, ey, aw, oy

Consonant	
plosives:	Complete blockage of airflow. Start with a short silence. b, p, d, t, g, k (stops)
nasals:	Like a stop but air is channelled to the nasal passage. m, n, ng
fricatives:	Air is forced through a narrow opening so that an aperiodic hissing noise is created. f, v, s, z, th, dh, sh, zh
afficatives:	Begin as a stop but end in a fricative. ch, jh
liquids:	Consonants with little frication. l, r (semi–vowels)
glides:	Basically /y/ = /iy/ and /w/ = /uw/ but shorter in duration and unstressed y, w (semi–vowels)

Formants of some vowels

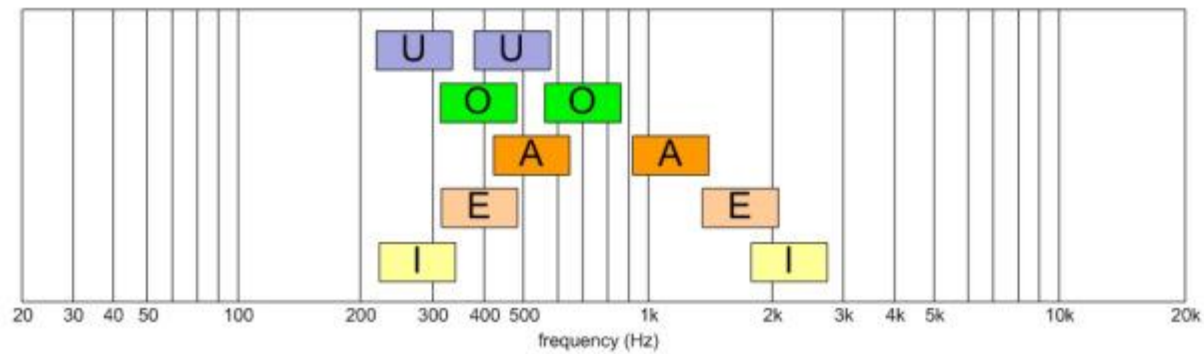
Vowel	Example	Average F1 (Hz)	Average F2 (Hz)
iy	feel	300	2300
ih	fill	360	2100
ae	gas	750	1750
aa	father	680	1100
ah	cut	720	1240
ao	dog	600	900
ax	comply	720	1240
eh	pet	570	1970
er	turn	580	1380
ow	tone	600	900
uh	good	380	950
uw	tool	300	940

Formants of some vowels

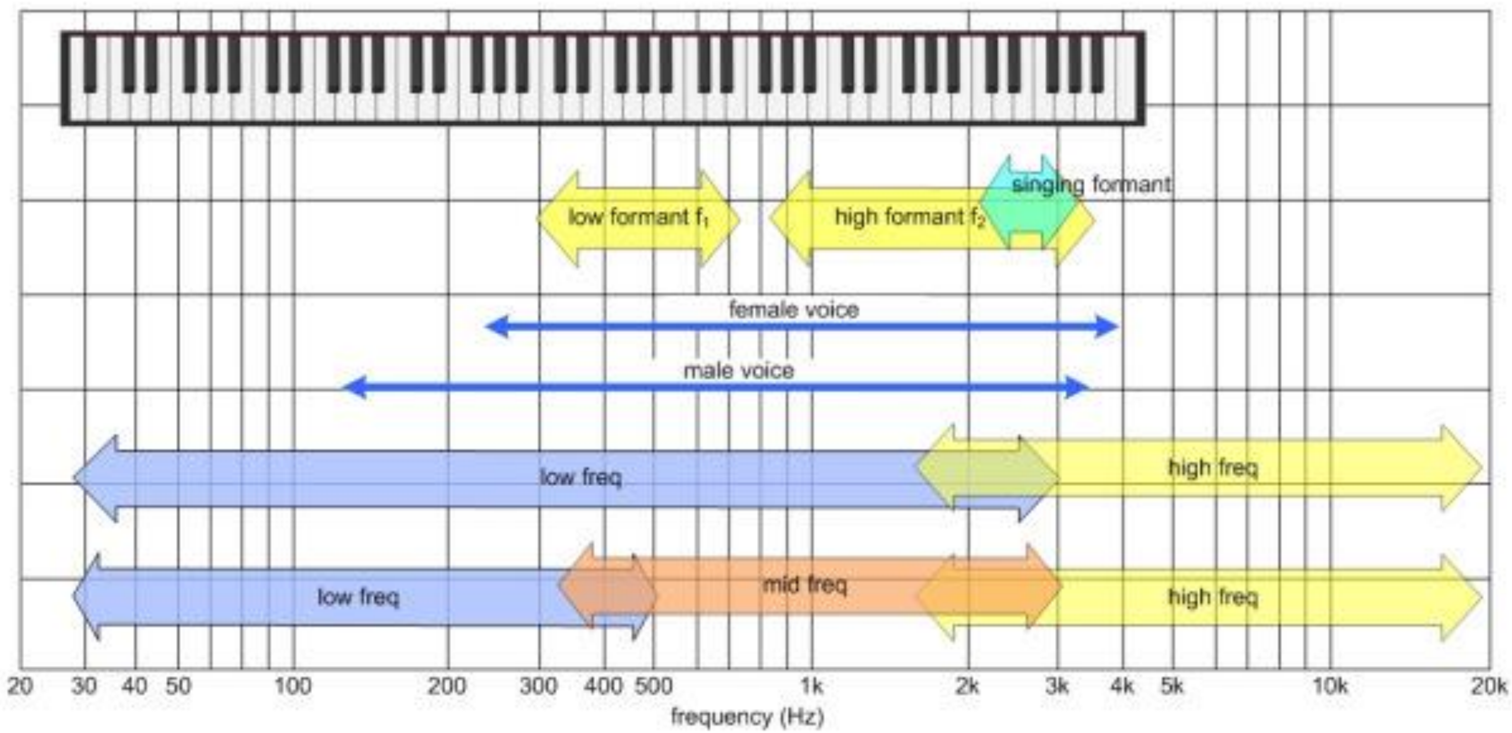


First two formants

- The lower speech formant f1 has a total range of about 300Hz to 750Hz
- and the higher speech formant f2 has a total range of about 900Hz up to over 3000Hz.



Speech frequency range



Phonemes and phones

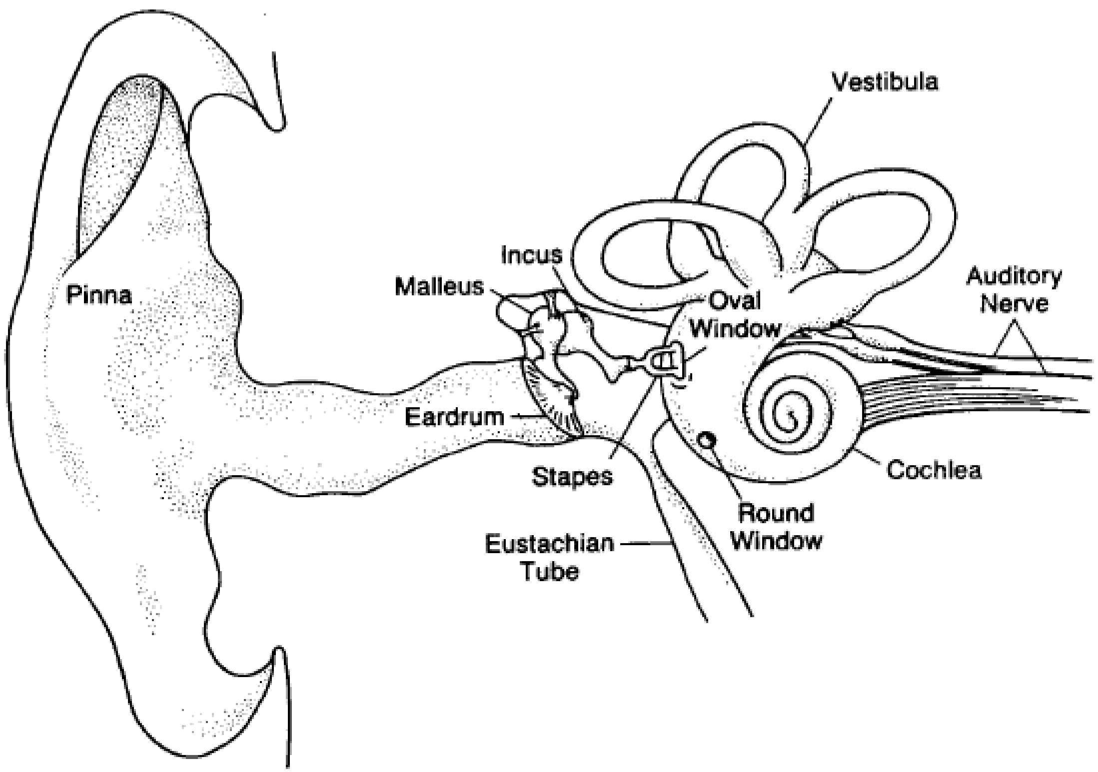
- Phonemes are the minimal speech units in a language that can serve to distinguish one word from another.
  - There are about 40-60 phonemes in English
- Phones are the acoustic realization of phonemes (the sounds).
- Phonemes can be broadly categorized into vowels and consonants.

Co-articulation

- A phoneme produced in isolation is very different from its realization under different acoustic contexts.
- The influence by its neighboring phonemes is called co-articulatory effects.

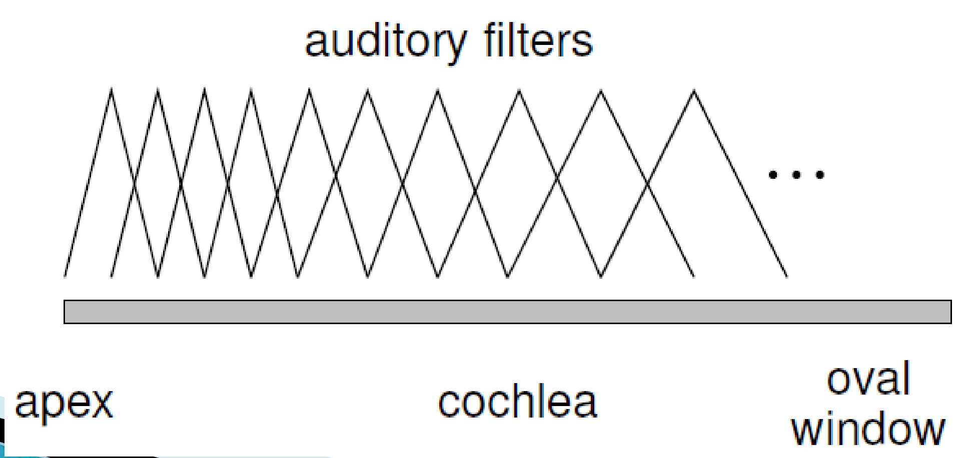
Speech Perception System

- Average lengths of various parts:
  - ear canal = 2.5cm,
  - middle-ear = 1.3cm (vol. = $6cm^3$)
  - cochlea = 3.5cm.



Cochlea as a Filter Bank

- Inside the cochlea is the basiliar membrane along which auditory nerves run.
  - Each location of basiliar membrane is most sensitive to a particular frequency called the characteristic frequency
  - It seems the inner ear is performing frequency analysis like Fourier Transform!

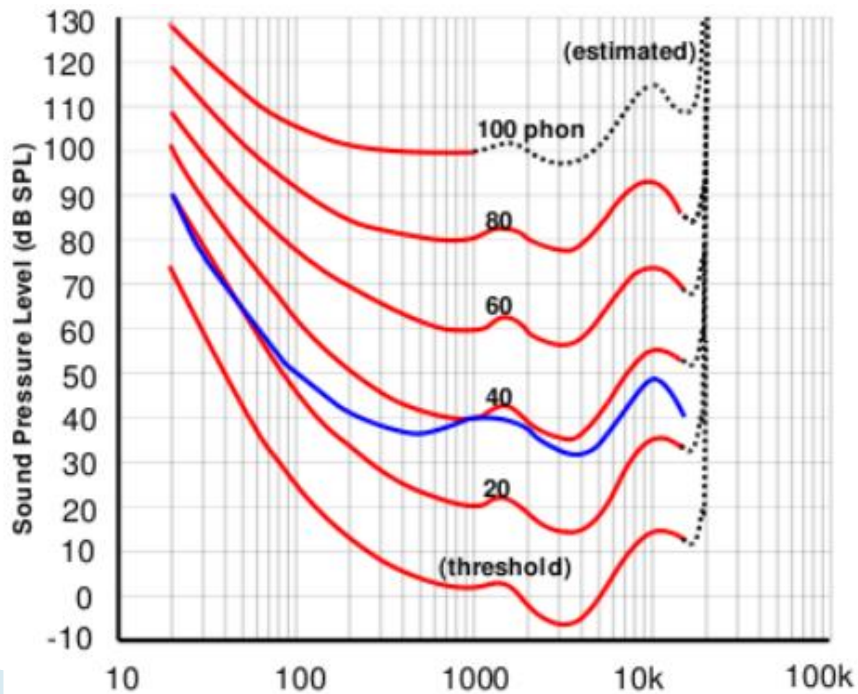


Cochlea as a Filter Bank

- The basilar membrane is roughly regarded as a filter bank — a set of overlapping bandpass filters.
  - Filters closest to the oval window respond to the high frequencies.
  - Filters closest to the apex respond to the low frequencies
- Each filter has an almost constant ratio of center frequency to bandwidth
  - Filters with high center frequency has a wider bandwidth.
- Thus, our inner ear has a higher resolution for low frequencies than for high frequencies in our common frequency scale (in Hz).
  - The common linear frequency scale is different from the perceptual frequency scale

Equal loudness curve

- The sensitivity of the human ear changes as a function of frequency
- The **range** of **human hearing** is generally considered to be 20 Hz to 20 kHz
- Humans are most sensitive to sounds around 2-4 kHz



Equal-loudness contours (red) (from ISO 226:2003 revision)
Original ISO standard shown (blue) for 40-phons

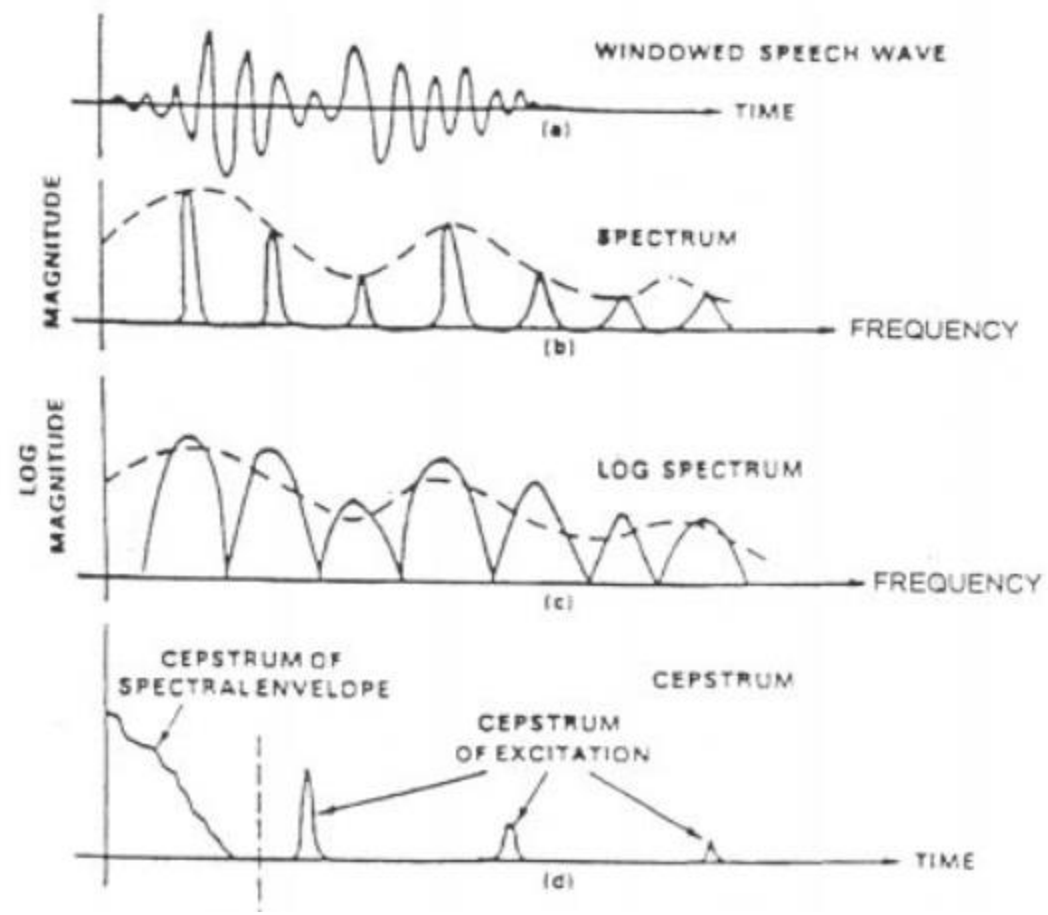
Features for ASR

- What kind of features are most important for ASR?
- Speech recognition is somehow a task of recognizing human voices.
- Human voices can be broken down into phones.
- What features are important for recognizing phones?

Cepstrum

- Recall that formants are important for classifying vowels.
- How to extract the formant information?
  - Extract the envelope information in the spectrum.
- Apply a transform again on the "signal" in the spectrum.
- The frequency domain will be converted into time domain again.
- The word cepstrum comes from reversing the word spectrum: "spec" -> "ceps"

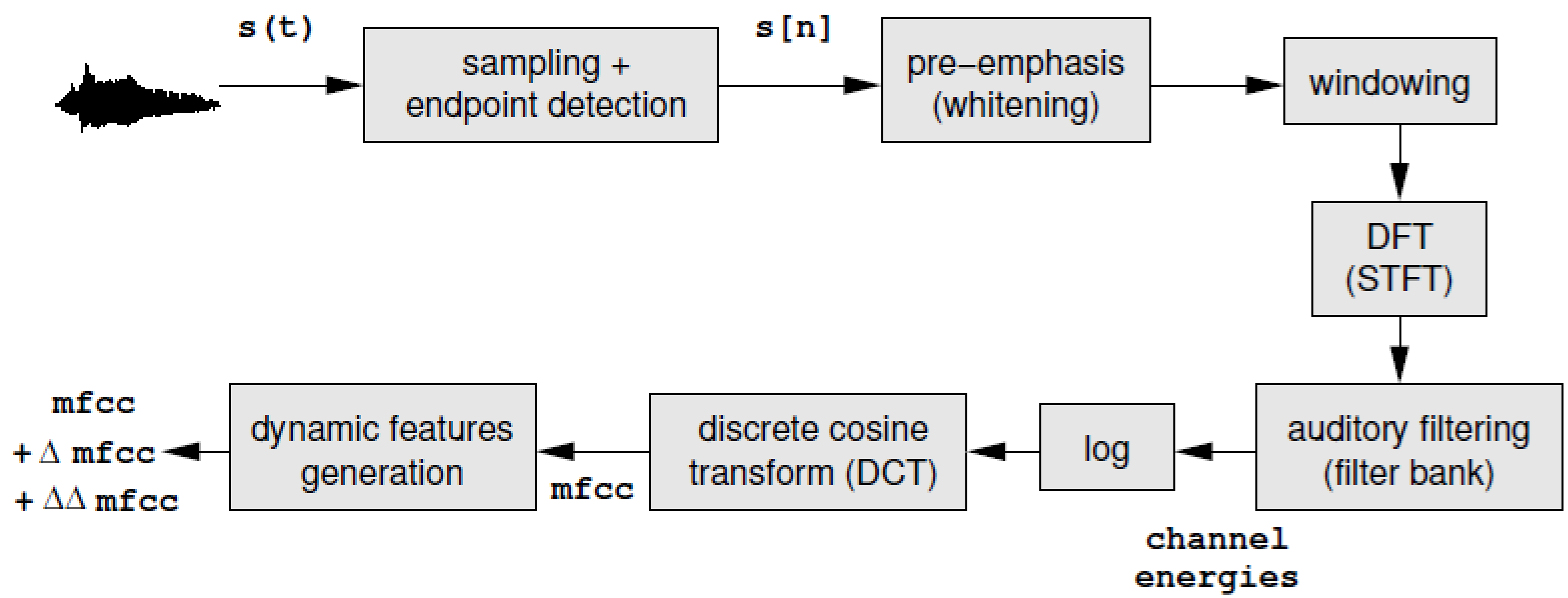
Cepstrum



WINDOWED SPEECH WAVE
TIME
(a)
MAGNITUDE
SPECTRUM
FREQUENCY
(b)
LOG MAGNITUDE
LOG SPECTRUM
FREQUENCY
(c)
CEPSTRUM OF SPECTRAL ENVELOPE
CEPSTRUM
CEPSTRUM OF EXCITATION
TIME
(d)

Mel-Frequency Cepstrum Coefficients

- Mel-Frequency Cepstrum Coefficients (MFCC) is the mostly commonly used feature for ASR.
- Its derivation is based on speech perception model.



Reading list

- Davis, S., Mermelstein, P.: *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. IEEE Trans. Acoust., Speech Signal Process. **28**(4), 357–366 (1980)
- **Spoken Language Processing: A Guide to Theory, Algorithm and System Development**