

# 발성과 듣기, 음향모델 (Articulation, Listening, Acoustic Modeling)

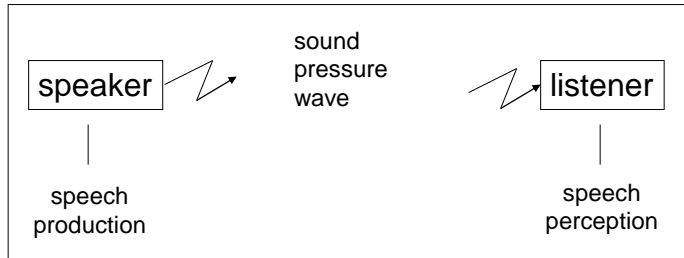
서울대 전기정보공학부  
성원용

Some slides are from CS224S Spoken Language Processing  
Dan Jurafsky Stanford University Spring 2014

## 소리

- 공기압력을 통한 신호전달
- Speech
- Voice
- Sound

## Speech Communication



### Speech Chain

Production: an idea of thought → neurological processes → muscular movements → acoustic sound pressure wave

Perception: acoustic sound pressure wave → human auditory system → neurological signals → the idea of thought

EE493Q: Digital Speech Processing

## 말하는 과정

- 어떤 소리를 내고 싶다 (뇌에서 생각)
- 입과 성대를 움직여서 해당 소리를 낸다. 이 때 귀로 들으면서 피드백을 받는다.
- 마치 악보를 보고 피아노를 치며, 피아노에서 소리가 나오는 것과 비교할 수 있다.

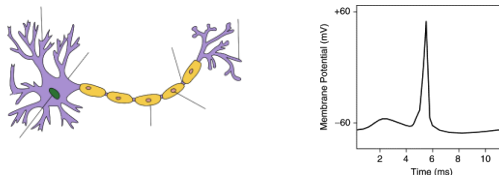
## 음성의 주파수 대역

- 인간은 20Hz ~ 20,000 Hz 사이를 듣는데 보통 음성에서는 200 ~ 4000 Hz 사이를 사용
- 실제로 전달되는 정보량은 1초에 10 byte 도 안된다. 그렇지만 음성이 반사되던가 일부 훼손이 되어도 잘 전달이 된다. 또 음색은 사람을 나타낸다.
- 유성음과 무성음이 있다. 유성음은 모음 등인데, 주기적 신호(pitch)가 있다.



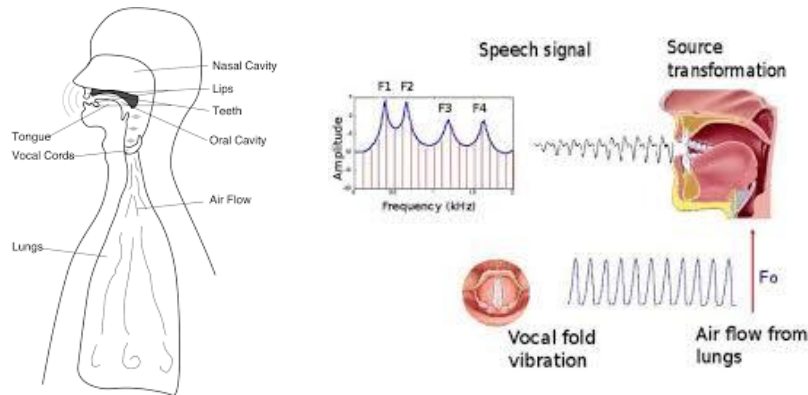
## 신경세포의 동작속도

- 시냅스를 통해서 다른 신경세포 또는 센서로 부터 전기신호를 받는다.
- 신경세포에서 모은 후에 이 것이 threshold 전압을 넘으면 땡 전기신호를 내보낸다.
- 이런 동작을 하는 시간 단위는? 약 10msec
- 이 까닭으로 수kHz의 신호를 발성, 듣기 위해서는 간접적인 방법(도구)을 사용. 발성기간, 귀



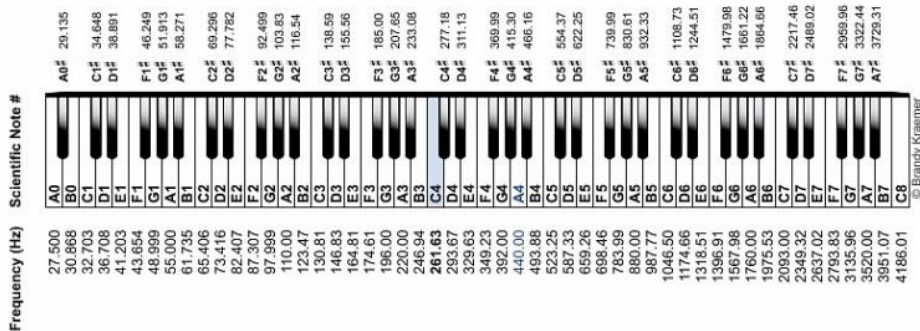
## 발성기관 (Human Vocal System)

- Vocal cord: 유성음(pitch), 무성음
- Vocal tract, mouth: Formant (F1, F2, F3, F4) 결정

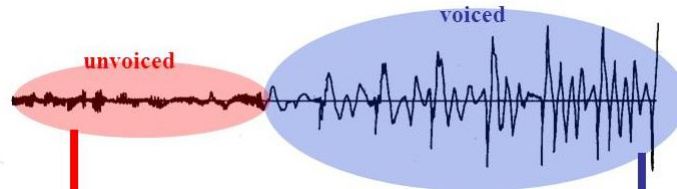


## Piano

- 한 옥타브 - 주파수가 두배 됨.
- 인간은 한번에 4개 이상의 건반을 누름.
- 건반 -> 소리 (주파수)



## 유성음과 무성음

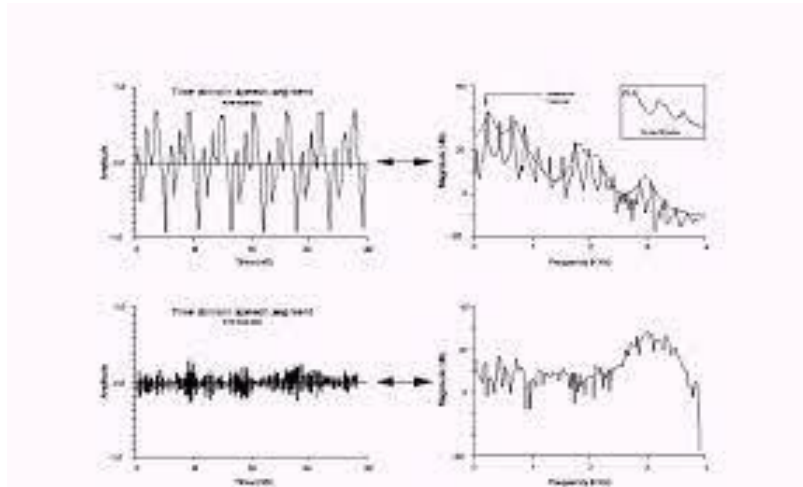


- 유성음
  - pitch가 있다 (규칙적으로 뚝뚝)
  - 노래부를 때 도레미파.. 는 pitch가 결정. Pitch는 음의 뜻에는 큰 관계가 없다.
  - 모든 모음과 일부 자음 (d, z, b, g, v,
- 무성음
  - Pitch가 없다. 보기에 잡음 같다.
  - 일부 자음 (t, s, p, k, f, h, ..

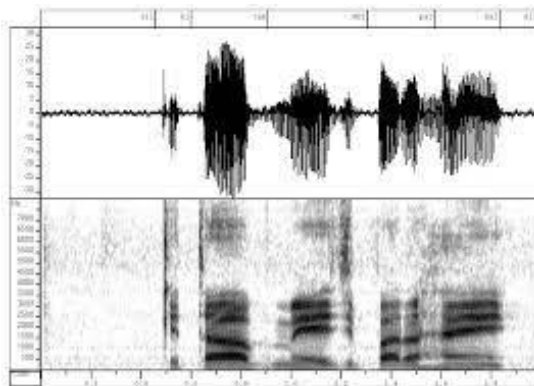
## 음성의 주파수 분석

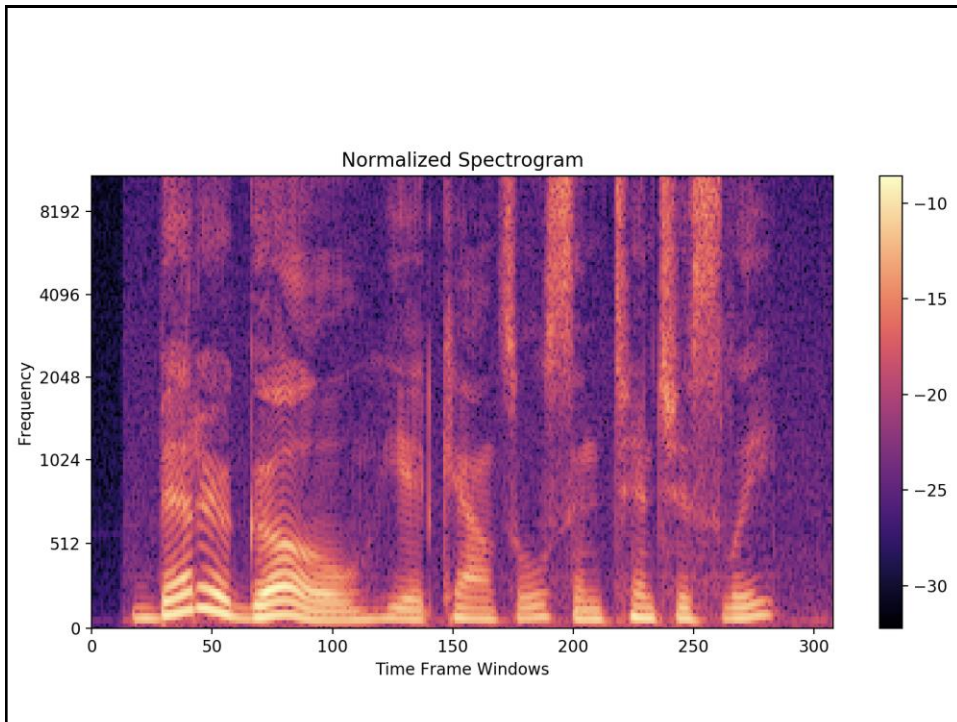
- 음성을 frame 단위로 나눈다 (보통 10msec ~ 20msec). 이 때 서로 좌우 frame 이 약간씩 겹친다. 그래서 sample 의 수는 200 ~ 500개
- 1초의 음성이라면 100~50 frame 이 얻어진다.
- Discrete Fourier Transform 을 각 frame 에 대해서 적용한다.
- 유성음 frame 과 무성음 frame 의 spectrogram 이 매우 다르다. 유성음은 고조파가 있다 (pitch 가 만드는 고조파)

## 유성음과 무성의의 spectrogram (한 프레임에 대해 본 것)



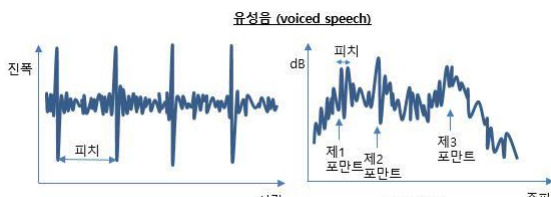
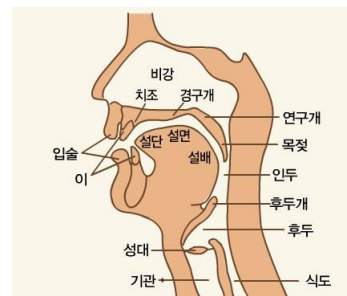
## Spectrogram – 연속된 frame의 spectrum을 연결한 것

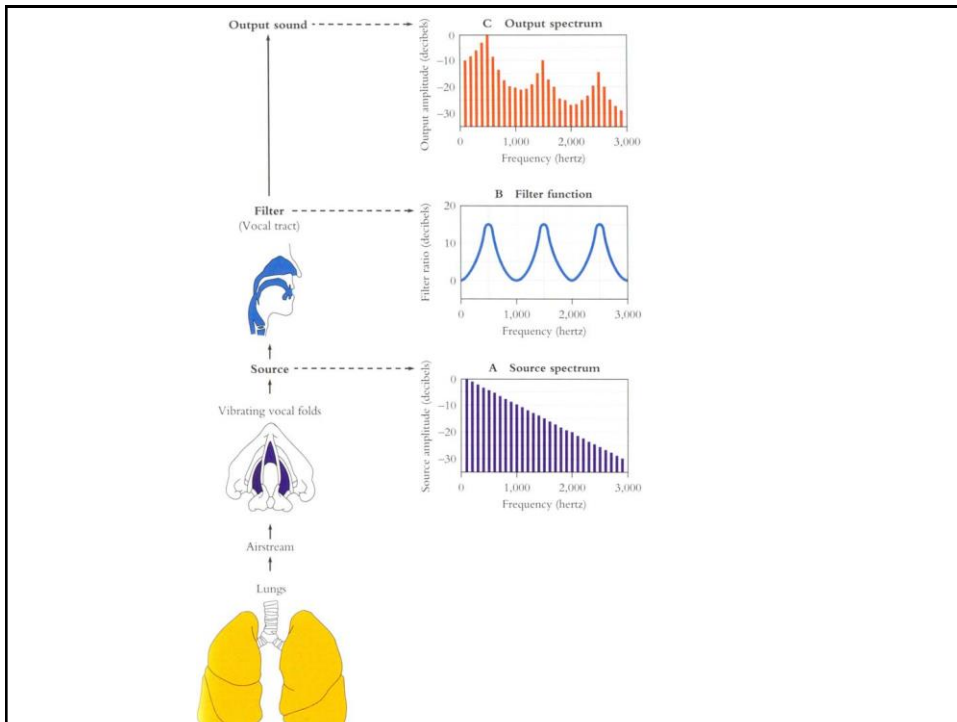




## 조음기관 (소리를 만드는 법)

- 성대 (vocal cord): 소리에 진동을 만들지 안 만들지를 결정 (소리에 진동, 유성음 - 모음과 일부자음)
- 성도 (vocal tract), 입과 혀, 입술, 코: 나오는 음의 특정 주파수를 세게 하거나 또는 작게 한다.
  - Formant를 결정: F1, F2, F3, F4
  - 이 네개 정도의 formant가 무슨 소리 인지(아, 야, ..)를 결정, 이 formant의 위치를 알아야 소리를 안다.





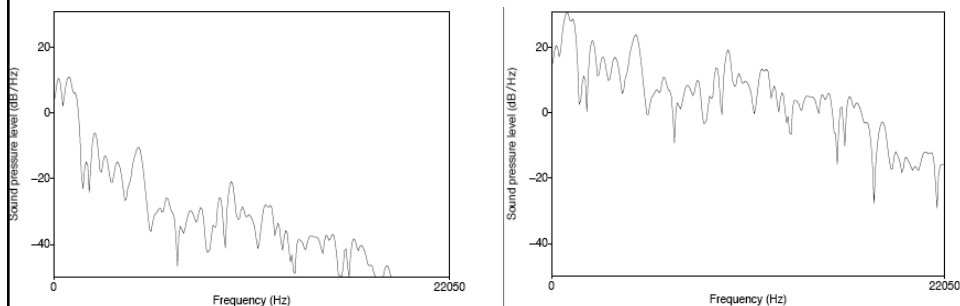
## Pre-emphasis

- Pre-emphasis: boosting the energy in the high frequencies (고주파 영역 강조)
- Q: Why do this?
- A: The spectrum for voiced segments has more energy at lower frequencies than higher frequencies.
  - This is called spectral tilt
  - Spectral tilt is caused by the nature of the glottal pulse
- Boosting high-frequency energy gives more info to the Acoustic Model
  - Improves phone recognition performance

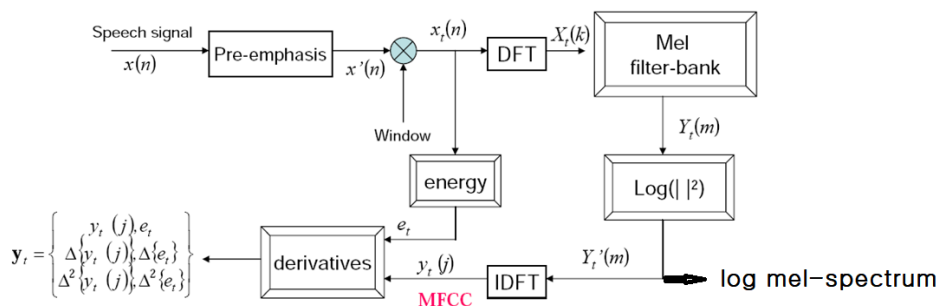


## Example of pre-emphasis

Spectral slice from the vowel [aa] before and after pre-emphasis



## Feature extraction



## Windowing

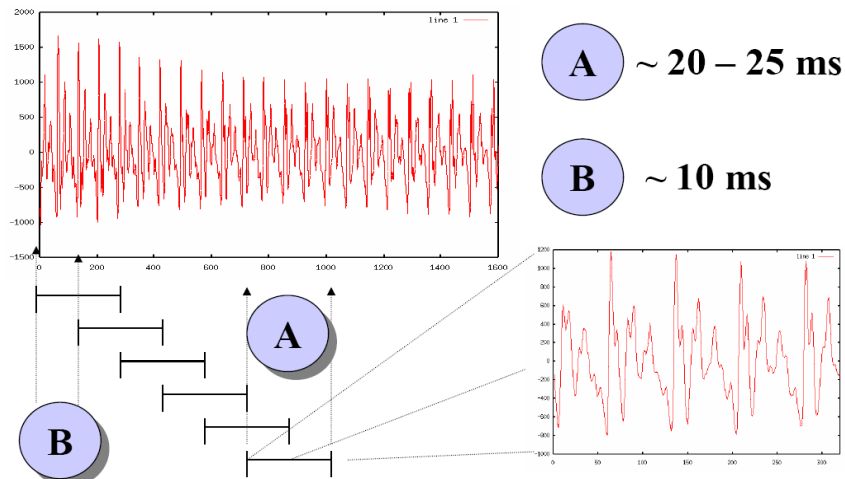


Image from Bryan Pellon

## Windowing

- Why divide speech signal into successive overlapping frames?
  - Speech is not a stationary signal; we want information about a small enough region that the spectral information is a useful cue.
- Frames
  - Frame size: typically, 10-25ms
  - Frame shift: the length of time between successive frames, typically, 5-10ms

## Common window shapes

- Rectangular window:

$$w[n] = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases} \quad \checkmark$$

- Hamming window

$$w[n] = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$



## Discrete Fourier Transform

- Input:

- Windowed signal  $x[n] \dots x[m]$

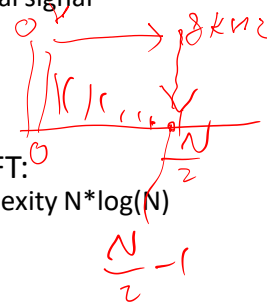


- Output:

- For each of N discrete frequency bands
- A complex number  $X[k]$  representing magnitude and phase of that frequency component in the original signal

- Discrete Fourier Transform (DFT)

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\pi}{N}kn}$$

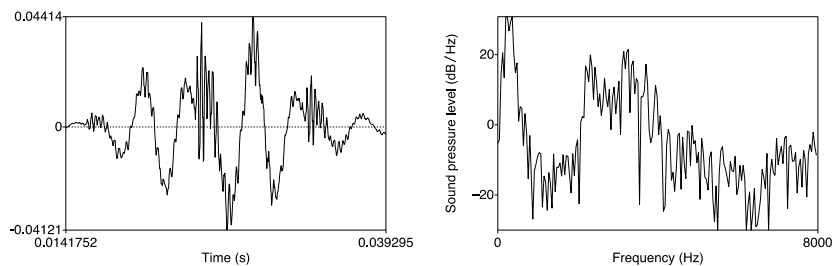


- Standard algorithm for computing DFT:

- Fast Fourier Transform (FFT) with complexity  $N \log(N)$
- In general, choose  $N=512$  or  $1024$

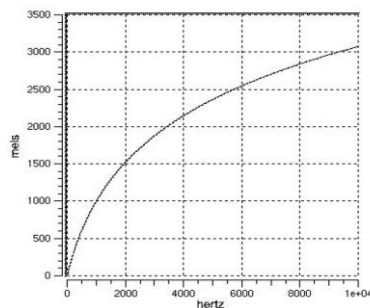
## Discrete Fourier Transform and Spectrum

- A 25 ms Hamming-windowed signal from [iy]
  - And its spectrum as computed by DFT (plus other smoothing)



## Mel-scale

- Human hearing is not equally sensitive to all frequency bands
- Less sensitive at higher frequencies, roughly  $> 1000$  Hz
- I.e. human perception of frequency is non-linear:



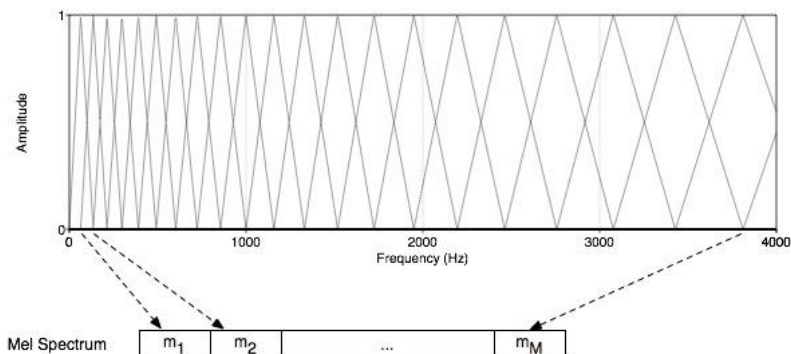
## Mel-scale

- A mel is a unit of pitch
  - Pairs of sounds
    - perceptually equidistant in pitch
    - are separated by an equal number of mels
- Mel-scale is approximately linear below 1 kHz and logarithmic above 1 kHz

$$Mel(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right)$$

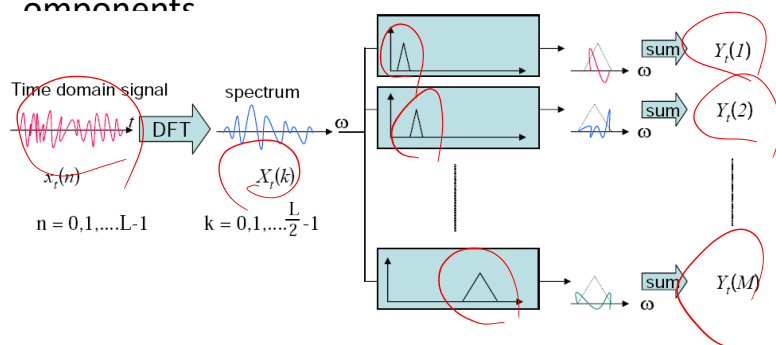
## Mel Filter Bank Processing

- Mel Filter bank
  - Roughly uniformly spaced before 1 kHz
  - logarithmic scale after 1 kHz

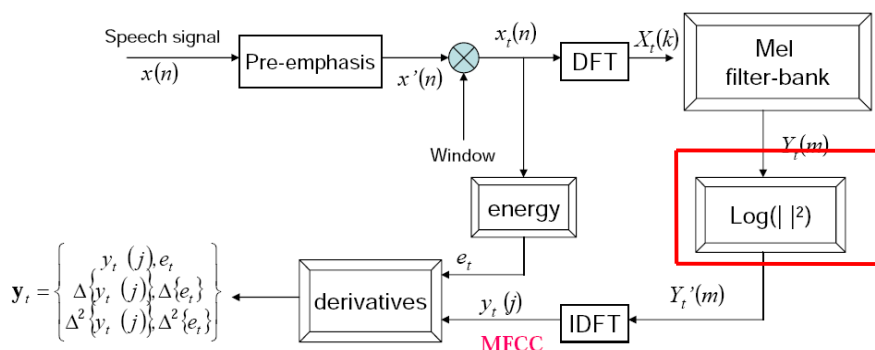


## Mel-filter Bank Processing

- Apply the bank of Mel-scaled filters to the spectrum
- Each filter output is the sum of its filtered spectral components

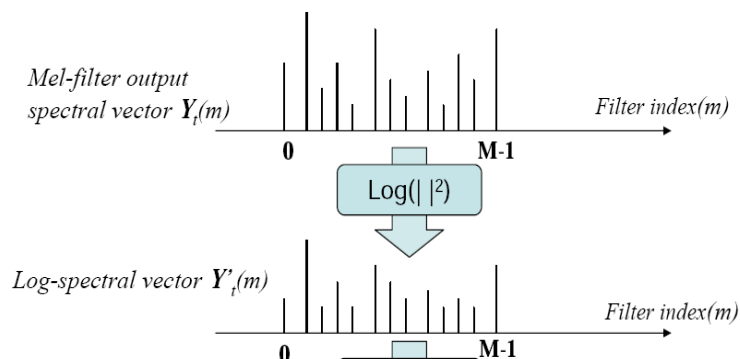


## MFCC



## Log energy computation

- Compute the logarithm of the square magnitude of the output of Mel-filter bank



## Log energy computation

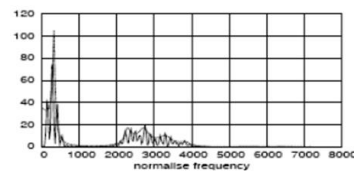
- Why log energy?
  - Logarithm compresses dynamic range of values
    - Human response to signal level is logarithmic
    - humans less sensitive to slight differences in amplitude at high amplitudes than low amplitudes
  - Makes frequency estimates less sensitive to slight variations in input (power variation due to speaker's mouth moving closer to mike)
  - Phase information not helpful in speech

# The Cepstrum

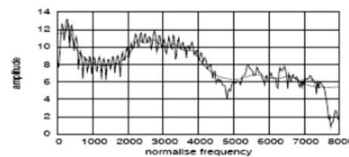
- One way to think about this
  - Separating the source and filter
  - Speech waveform is created by
    - A glottal source waveform
    - Passes through a vocal tract which because of its shape has a p  
articulate filtering characteristic
- Articulatory facts :
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of oral cavity, some harmonics are amplified more than others

# The Cepstrum

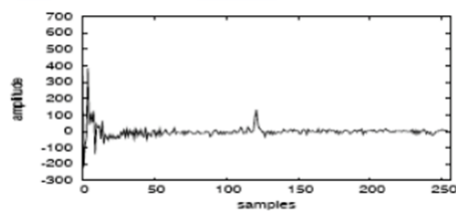
- The spectrum of the log of the spectrum



Spectrum



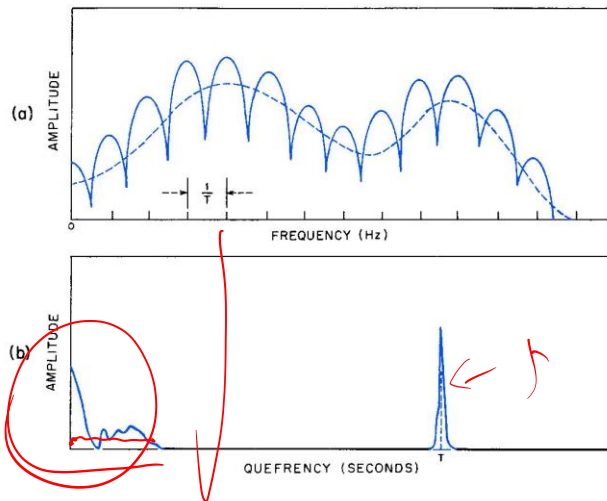
Log spectrum



Spectrum of log spectrum



## Thinking about the Cepstrum



## Mel Frequency cepstrum

- The cepstrum requires Fourier analysis
- But we're going from frequency space back to time
- So we actually apply inverse DFT

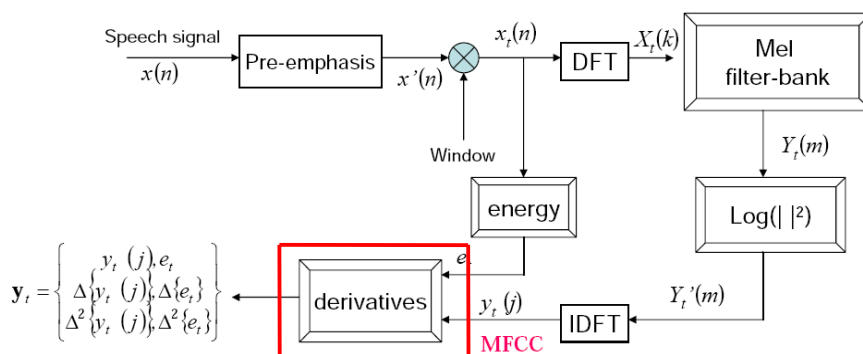
$$y_t[k] = \sum_{m=1}^M \log(|Y_t(m)|) \cos\left(k(m - 0.5)\frac{\pi}{M}\right), \quad k=0, \dots, J$$

- Details for signal processing gurus: Since the log power spectrum is real and symmetric, inverse DFT reduces to a Discrete Cosine Transform (DCT)

## Another advantage of the Cepstrum

- DCT produces highly uncorrelated features
- If we use only the diagonal covariance matrix for our Gaussian mixture models, we can only handle uncorrelated features.
- In general we'll just use the first 12 cepstral coefficients (we don't want the later ones which have e.g. the F0 spike)

## MFCC

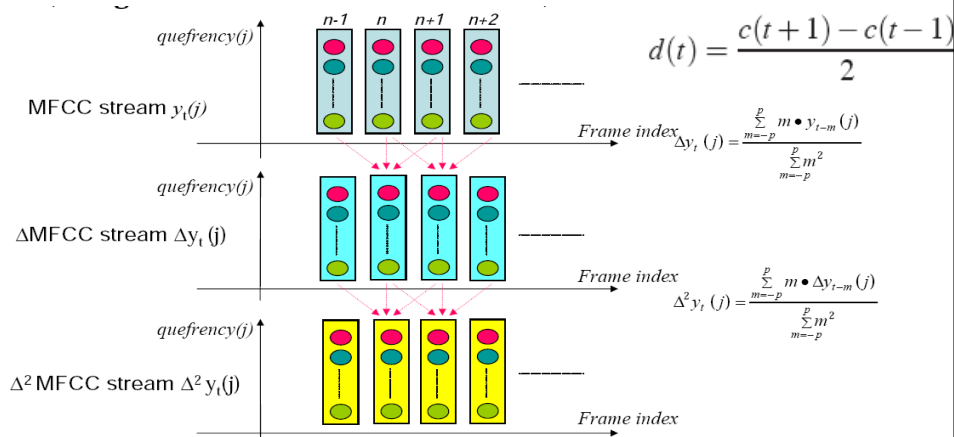


## “Delta” features

- Speech signal is not constant
  - slope of formants,
  - change from stop burst to release
- So in addition to the cepstral features
- Need to model changes in the cepstral features over time.
  - “delta features”
  - “double delta” (acceleration) features

## Delta and double-delta

- Derivative: in order to obtain temporal information



## Typical MFCC features

- Window size: 25ms
- Window shift: 10ms
- Pre-emphasis coefficient: 0.97
- MFCC:
  - 12 MFCC (mel frequency cepstral coefficients)
  - 1 energy feature
  - 12 delta MFCC features
  - 12 double-delta MFCC features
  - 1 delta energy feature
  - 1 double-delta energy feature
- Total 39-dimensional features

## Why is MFCC so popular?

- Efficient to compute
- Incorporates a perceptual Mel frequency scale
- Separates the source and filter
- IDFT(DCT) decorrelates the features
  - Necessary for diagonal assumption in HMM modeling
- There are alternatives like PLP

## Feature extraction for DNNs

### Mel-scaled log energy

- For DNN (neural net) acoustic models instead of Gaussians
- We don't need the features to be decorrelated
- So we use mel-scaled log-energy spectral features instead of MFCCs
- Just run the same feature extraction but skip the discrete cosine transform.

## 청각기관의 구조

## 듣기과정

말하기 과정:

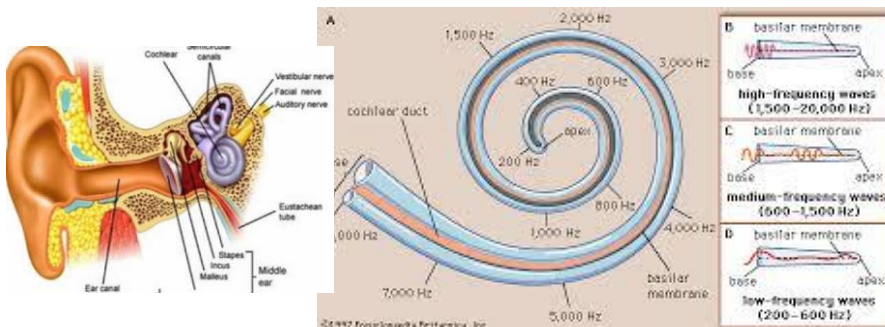
- 어떤 소리를 내고 싶다 (뇌에서 생각)
- 입과 성대를 움직여서 해당 소리를 낸다. 이 때 귀로 들으면서 피드백을 받는다.
- 마치 악보를 보고 피아노를 치며, 피아노에서 소리가 나오는 것과 비교할 수 있다.

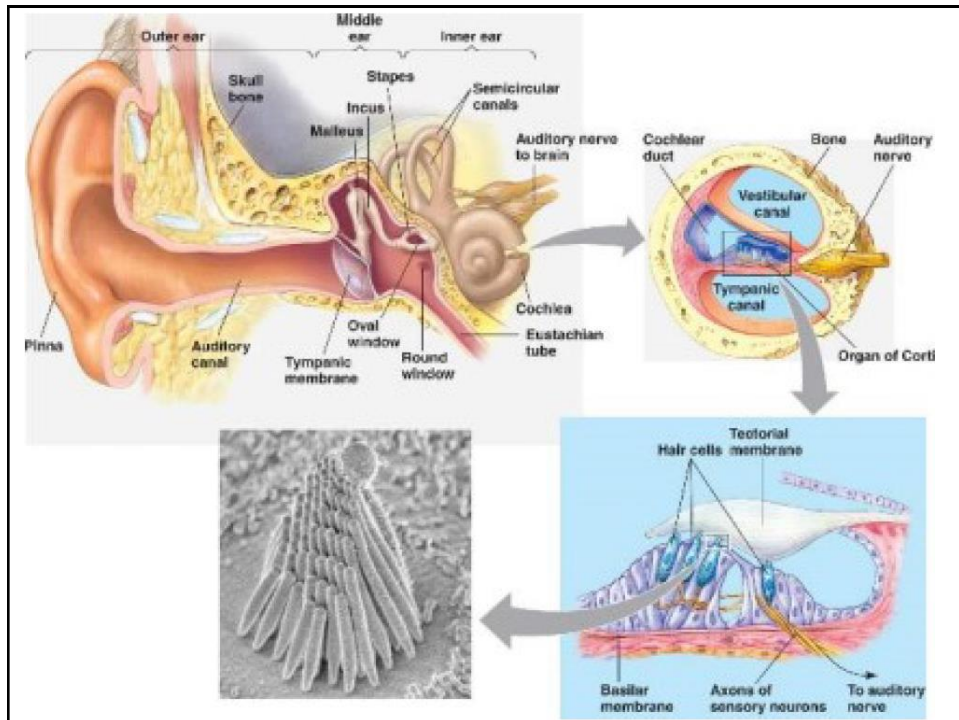
듣기과정:

- 소리가 들어온다.
- 이 소리를 주파수 영역으로 바꾼다. (소리를 듣고 피아노의 건반위치를 안다) - 귀의 달팽이관
- 뇌에다 각 주파수 영역에 어떤 세기의 신호가 왔는지를 알려준다.
- 뇌가 어떤 소리인지를 판다.

## 듣는 기관 (귀)

- 소리가 들어가면 주파수에 따라 달팽이관의 특정위치에 진동을 만듦.
- 달팽이관의 중심부분이 낮은 주파수, 바깥이 높은 주파수임.
- 그 곳에 연결된 hair cell을 통해 신경을 자극함



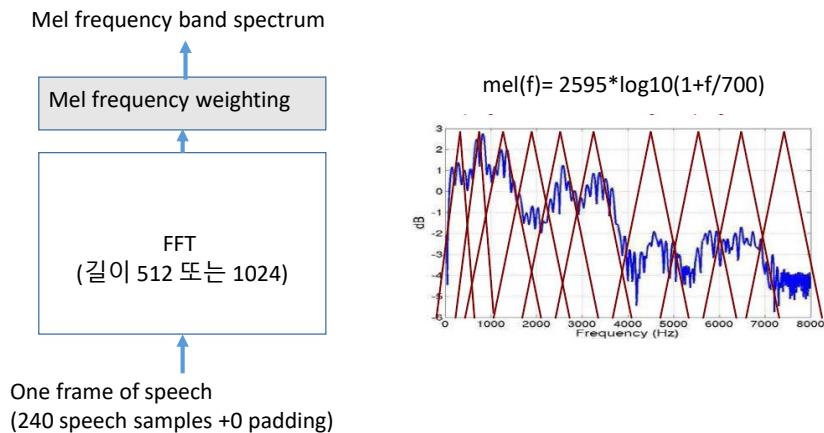


- <https://m.blog.naver.com/PostView.nhn?blogId=ling1134&logNo=70165398365&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F>

[생리학] 기계식 수용(Mechanoreception) -2

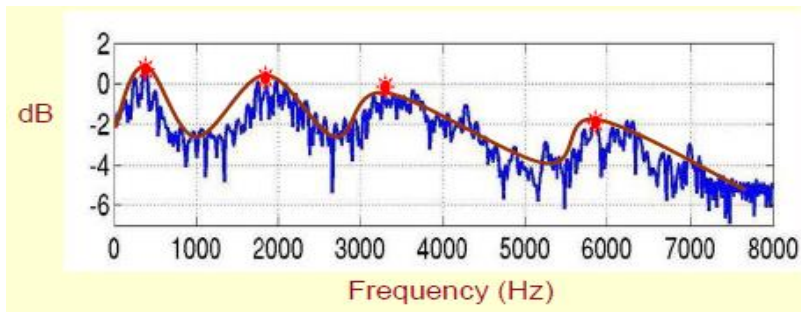
## Mel Scaled-Filter Bank

- Human auditory system based

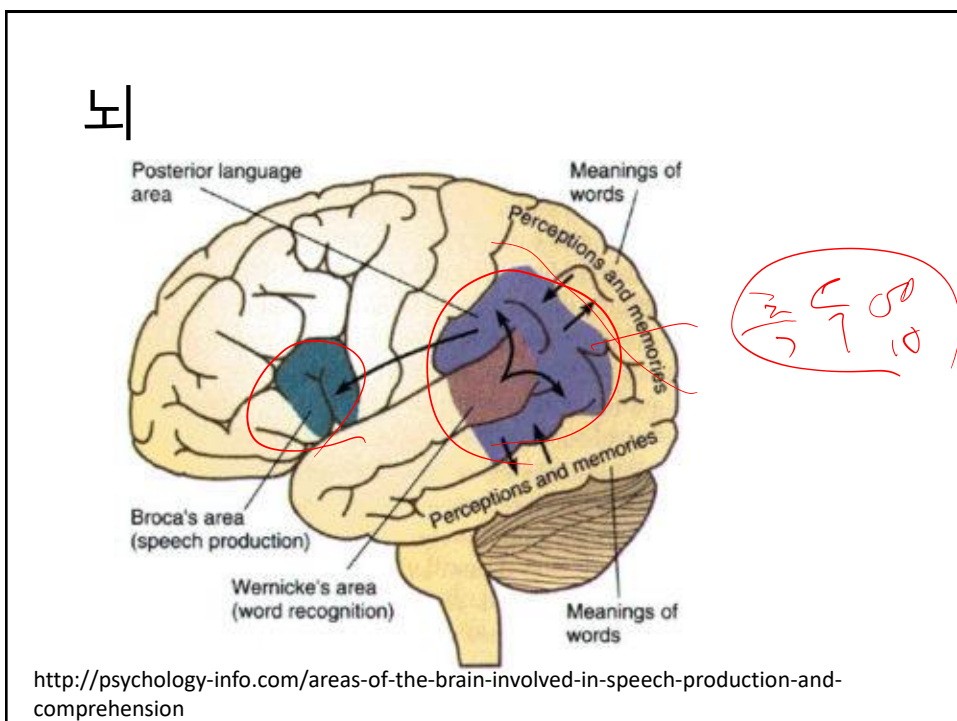
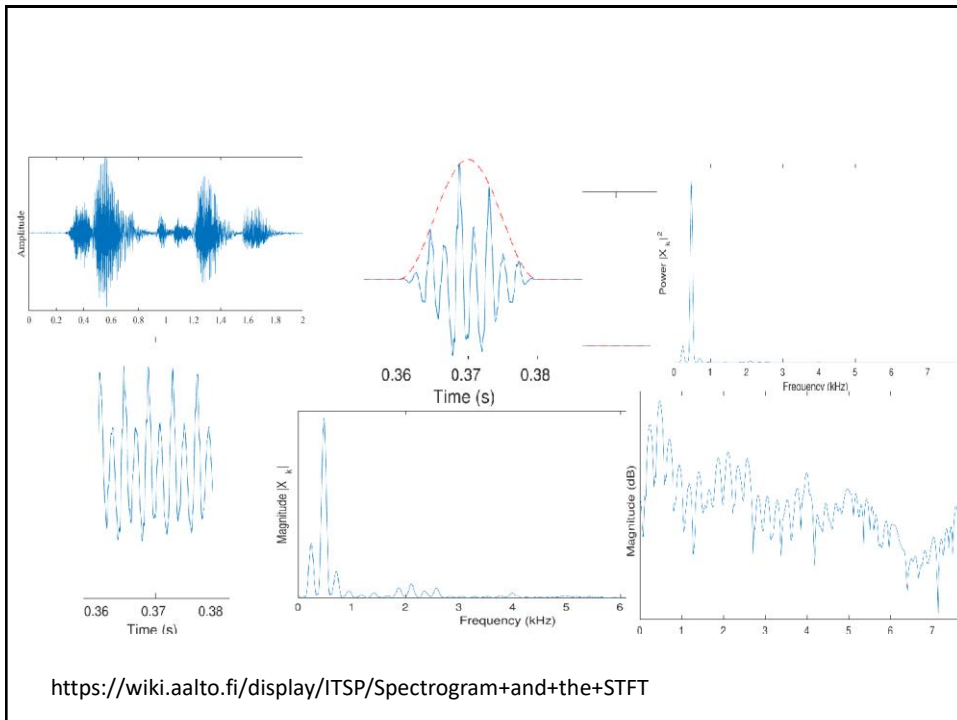


## Speech features

- Feature – 개수가 너무 많으면 계산 많고 복잡하다. 반면 너무 적으면 인식을 떨어진다.
- 인간귀의 특징을 고려 (낮은 주파수의 해상도 크다).
- Spectral envelope이 중요. Pitch에 의해 생기는 골의 영향이 적어야함.
- 벽의 반사 등 주변 환경의 영향을 제거할 필요가 있다.

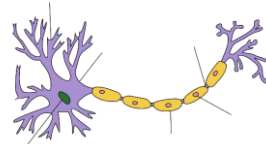
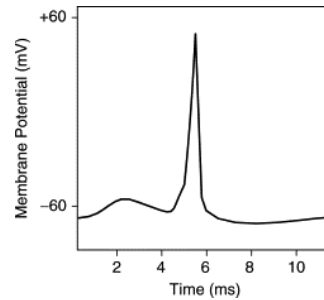






## 음성발성 및 듣기 과정의 이해

- 인간은 소리의 주파수 변화로 듣는다 (time domain이 아니라 frequency domain이 중요, phase 중요치 않음)
- 인간의 뇌활동에 의한 control은 대개 10 ms (10 밀리세컨드) 전후의 속도로 동작한다.
- 인간이 내는 소리의 스펙트럼(vocal tract가 결정)은 약 10msec 동안은 안정적이다.



## 음성인식과 feature extraction

- 귀의 달팽이관에서 일어나는 일을 수행
- 소리 -> 주파수 영역으로 바꾼다
- 이때 낮은 주파수는 좀 미세하게, 높은 주파수는 좀 성기게 채집한다. (Mel frequency filter bank)

# 음성(speech)의 구조와 모델링

## 음성의 모델링

- 음성은 매우 짧은 시간 (약 20msec)에서는 스펙트럼이 변하지 않는다. (phoneme state)
- Phoneme – 우리가 언어학적으로 정의하는 어떤 발음 (발음기호), 그런데 이 발음은 아주 정해진 것이 아니라 옆의 발음에 따라 변한다.
- 단어 – 앞의 phoneme 의 sequence
- 문장 – 단어의 sequence

## Acoustic modeling

- Phoneme: 언어학자가 정의한 기본 발음단위.
- 영어의 경우 40~50개의 phoneme

(There are total of 78 phonemes used in TIMIT database, out of which 46 phonemes are of English language (American), 1 phoneme for silence(sil), 1phoneme for short pause(sp) and the remaining 30 are stressed phonemes.)

- 하나의 phoneme 도 그 음을 처음 발음하는 frame, 중간 frame, 끝날 때 frame에서의 발음이 다른데, 이 경우 이를 통째로 모델하는 것이 mono-phone, 세개의 state로 나누어서 model 하는 것이 tri-phone 이다.

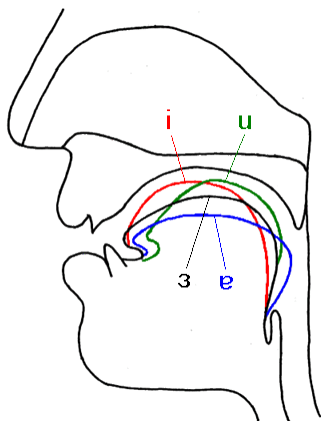
	Phone Label	Example		Phone Label	Example		Phone Label	Example
1	iy	beet	22	ch	choke	43	en	button
2	ih	bit	23	b	bee	44	eng	Washington
3	eh	bet	24	d	day	45	l	lay
4	ey	bait	25	g	gay	46	r	ray
5	ae	bat	26	p	pea	47	w	way
6	aa	bob	27	t	tea	48	y	yacht
7	aw	bout	28	k	key	49	hh	hay
8	ay	bite	29	dx	muddy	50	hv	ahead
9	ah	but	30	s	sea	51	el	bottle
10	ao	bought	31	sh	she	52	bcl	b closure
11	oy	boy	32	z	zone	53	dcl	d closure
12	ow	boat	33	zh	azure	54	gcl	g closure
13	uh	book	34	f	fin	55	pcl	p closure
14	uw	boot	35	th	thin	56	tcl	t closure
15	ux	toot	36	v	van	57	kcl	k closure
16	er	bird	37	dh	then	58	q	glotal stop
17	ax	about	38	m	mom	59	pau	pause
18	ix	debit	39	n	noon	60	epi	epenthetic
19	axr	butter	40	ng	sing			silence
20	ax-h	suspect	41	em	bottom	61	h#	begin/end
21	jh	joke	42	nx	winner			marker

Table 2. 61 TIMIT original phone set.

aa, ao	aa
ah, ax, ax-h	ah
er, axr	er
hh, hv	hh
ih, ix	ih
l, el	l
m, em	m
n, en, nx	n
ng, eng	ng
sh, zh	sh
uw, ux	uw
pcl, tcl, kcl, bcl, dcl, gcl, h#, pau, epi	sil
q	-

Table 3. Mapping from 61 classes to 39 classes, as proposed by Lee and Hon, (Lee & Hon, 1989). The phones in the left column are folded into the labels of the right column. The remaining phones are left intact. The phone 'q' is discarded.

모음

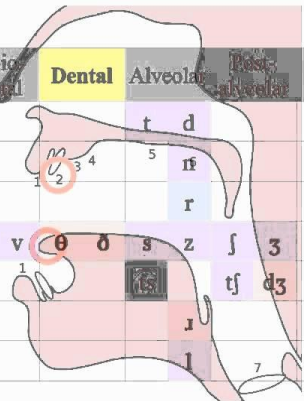


	front	central	back	
close	heed i:		u: shoe	
half-close	hid I	the ə	u put	
half-open	head e	bird ɜ:	ɔ: saw	
open	had æ	cut ʌ	ɑ: ɒ hod	
	unround		hard round	

# 자음

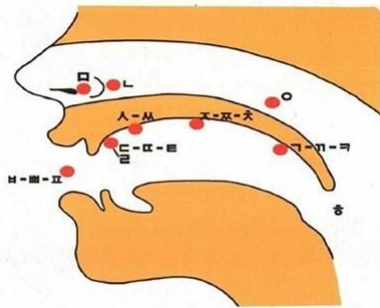
## Dental Consonants

Зубные согласные



	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar	Glottal
Stop	p b			t d			k g ʔ	
Nasal		m		n ñ			ŋ	
Trill				r				
Fricative		f v	θ ð	s z	ʃ ʒ	x		h
Affricative				tʃ dʒ				
Approximant				ɹ		j	w	
Lateral appr.				l				

## 우리말의 자음 발성

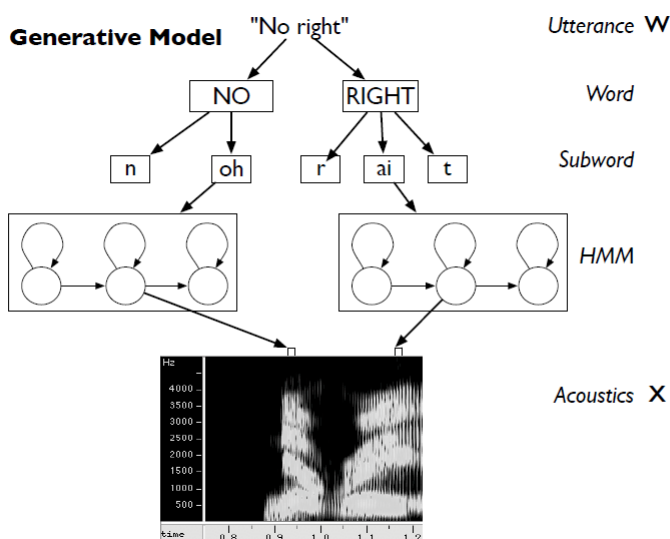


		양순음	치경음	경구개음	연구개음	성문음
폐쇄음 (파열음)	평음	ㅂ	ㄷ		ㄱ	
	격음	ㅃ	ㄸ		ㅋ	
	경음	ㅍ	ㅌ		ㄴ	
마찰음	평음		ㅅ			ㅎ
	경음		ㅆ			
파찰음	평음			ㅈ		
	격음			ㅊ		
	경음			ㅌ		
비음		ㅁ	ㄴ		ㅇ	
설측음			ㄹ			

## Tri-phone

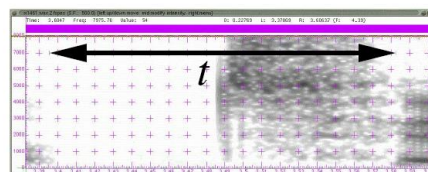
- 어떤 phoneme 을 발음하는 과정을 앞부분, 중간 부분, 뒷부분 이렇게 세부분 (tri)의 state 로 나누어서 모델링
- 이 경우 각 부분은 비교적 일정한 주파수 특성을 보인다.
- 앞부분과 뒷부분은 이어지는 발음에 따라 달라진다 (context dependent). 따라서 triphone 의 개수는 엄청 많아질 수 있는데 (수만) 이를 줄여서 수천개로 만들어 쓴다. 이를 CD-triphone states 라 한다. 이 CD-triphone states 가 보통의 hidden Markov model에 사용된다.

## Hierarchical modeling of speech

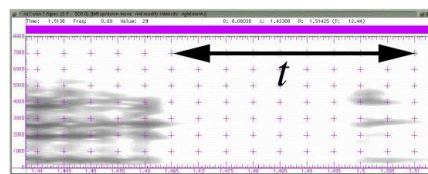


## Phonetic context

- Context – 좌우의 발음에 의해서 가운데 발음이 영향을 받음 (coarticulation).
- /n/ in ten (dental) and tenth (alveolar)

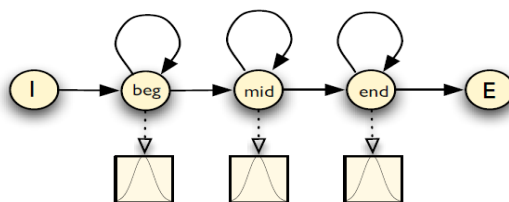


"tube"

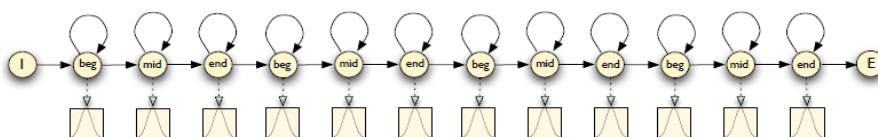


"suit"

## Three state phone models (triphone)



/t/



/s/

/ih/

/k/

/s/

"six"



## 단어의 발음

- Phoneme 또는 알파벳을 연결함 (sequence)
- [/bʊk/]
- book

## Acoustic modeling

- 음성을 듣고서 발음으로 표시함. 이 때 표현하는 발음의 단위에 따라
  - Triphone states (매우 작은 시간 단위로 표현)
    - 매우 안정된 주파수 특징을 보이기 때문에 인식이 쉽다.
    - 이를 꺾어 맞추어서 단어와 문장으로 만들기가 복잡하다. 더 복잡한 hidden Markov model.
    - 전체적으로 정확도 좋다.
  - Monophone (40~70개의 phoneme 으로 표현)
  - Grapheme (알파벳 글자로 표현) – 수십개, 1초에 몇 번 나옴.
  - Wordpiece – 수백 ~ 수만
  - Word – 수십만 단어, 1초 정도의 시간