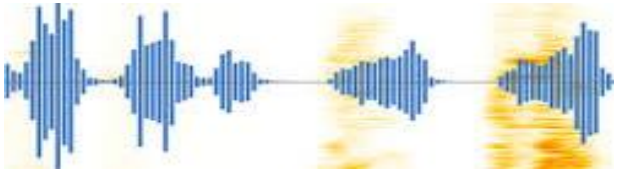
Speech Signal

음성신호

- 음성은 인간의 의사 소통을 위한 소리 형태의 수단



- 각 음성 단어는 모음과 자음이라는 음성이라는 소리 단위의 제한된 세트의 조합

자음 consonants	ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ	ㅈ	ㅎ	
	ㅋ		ㅌ			ㅍ			ㅊ		
	ㄲ		ㄸ			ㅃ	ㅆ		ㅉ		
모음 vowels	ㅏ	ㅓ	ㅗ	ㅜ	ㅡ	ㅣ	ㅐ	ㅒ	ㅚ	ㅟ	
	ㅑ	ㅕ	ㅛ	ㅠ		ㅢ	ㅖ	ㅙ			
			ㅘ	ㅝ			ㅞ	ㅙ			

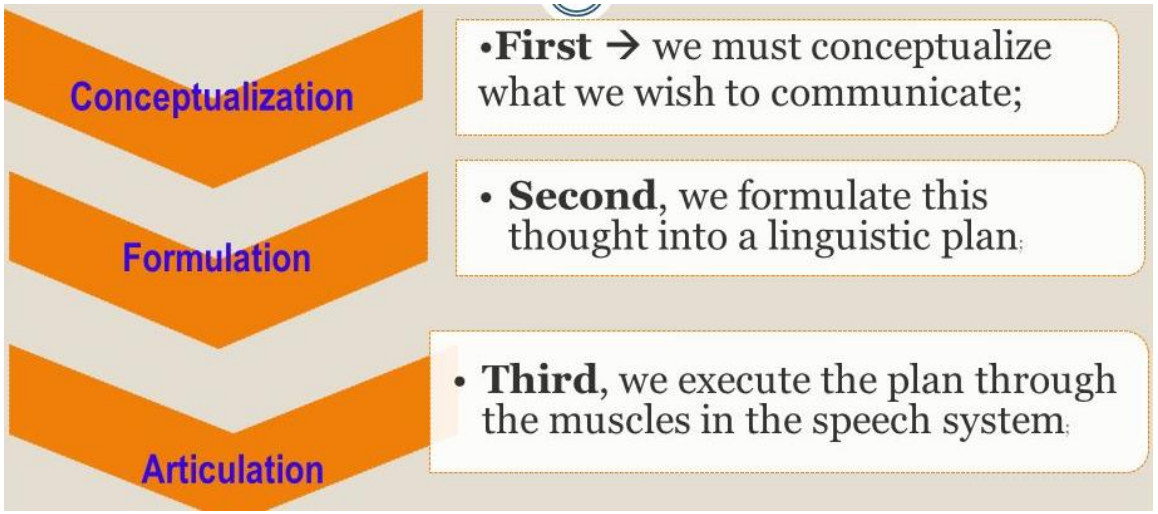
- 한글기반의 한국어는 초성, 중성, 종상으로 구성

음성 생성

- 말하고자 하는 단어들이 선택되고 발성을 위한 준비 단계로 음소들의 조합이 형성된 후, 발성기관을 통해 조음이 만들어지는 전 과정을 의미함.

- 음성의 생성의 3가지 주요한 과정을 포함함.

- Conceptualization
- Formulation
- Articulation

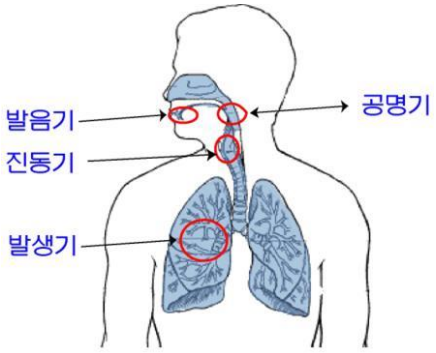


발성 및 조음 (1)

➢ 발성기관

1. 폐는 공기를 모아 놓고 발성 시에 소리의 에너지를 공급

2. 후두와 성대
   - 후두가 성대를 포함하고 있음.
   - 일반적인 호흡 시에는 후두의 연골이 적당히 간격을 두고 벌어져 있다가 발성 시에는 가운데 쪽으로 모임.
   - 무성자음을 발성 시, 성대는 ``s'', ``sh'', and ``f ' 발음 시 완전히 열여 있거나, ``h' ' 발음 시 부분적으로 열려 있음.
   - 유성음은 성대의 떨림에 의해 만들어짐.
   - 성대 떨림의 속도는 주로 그것의 질량이나 긴장도 또는 장력에 의해 결정되는데 통과하는 공기의 압력과 속도도 일부 영향을 줌.

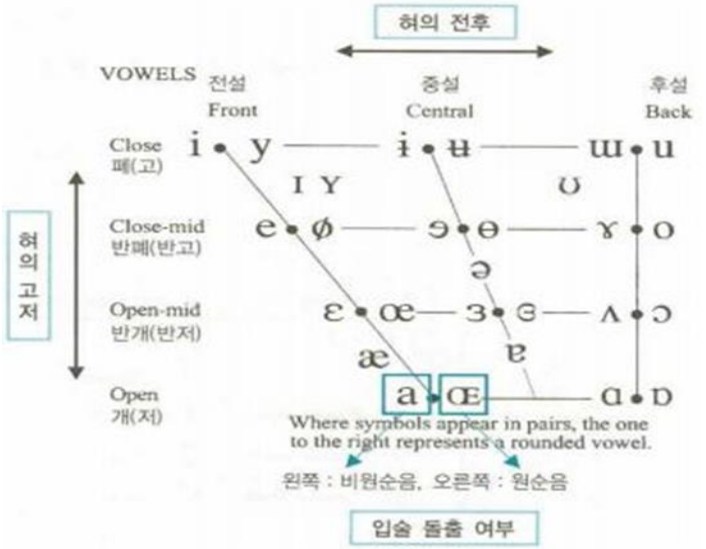


3. 성도: Vocal Tract
   - 성도는 성대부터 입술까지 연결된 하나의 튜브에 코쪽의 비강이 곁가지로 연결된 것으로 여겨짐.
   - 일반적인 성도의 길이는 약17 centimeters로 알려져 있음.
   - 구강은 성도의 끝 쪽에 있지만 가장 중요한 부분으로 입천장, 혀, 입술 및 치아의 상대적 위치의 조정에 따라 그 크기나 모양이 다양해질 수 있음.

발성 및 조음 (2)

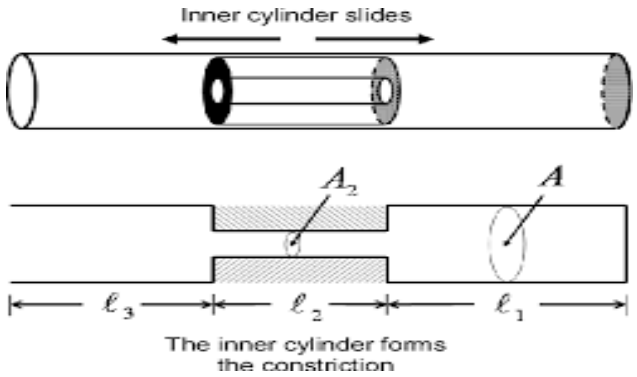
- 음성 조음
  - 음성 소리의 가장 작은 단위는 phoneme(음소) 임. 하나 이상의 phoneme이 결합되어 syllable(음절)을 형성하고 하나 이상의 음절이 결합되어 word(단어)를 형성함.
  - Phoneme 은 두 가지로 구성된다: 모음과 자음(vowels and consonants).
  - 모음은 모두 유성음임. 영어에 사용되는 약 12~21개의 모음이 있음.
  - 자음은 소리의 신속하고 때로는 미묘한 변화를 보임. 자음은 발음의 방법에 따라 파열음 (p, b, t 등), 마찰음 (f, s, sh 등), 비음 (m, n, ng), 유음(r, l) 및 반모음 (w, y)으로 분류됨.

조음방법 \ 조음위치			입술소리	혀끝소리	센입천장소리	여린입천장소리	목청소리
			양순음	설단음, 치조음	경구개음	연구개음	후음
			두입술	윗잇몸	센입천장	여린입천장	목청사이
안울림소리	파열음	예사소리	ㅂ	ㄷ		ㄱ	
		된소리	ㅃ	ㄸ		ㄲ	
		거센소리	ㅍ	ㅌ		ㅋ	
	파찰음	예사소리			ㅈ		
		된소리			ㅉ		
		거센소리			ㅊ		
	마찰음	예사소리		ㅅ			
		된소리		ㅆ			ㅎ
울림소리	비음		ㅁ	ㄴ		ㅇ	
	유음			ㄹ			

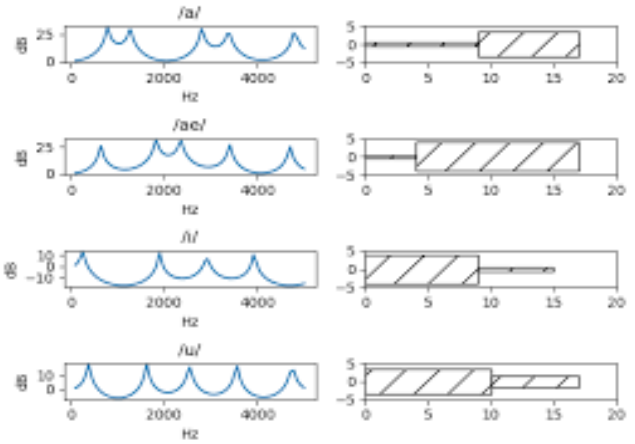


발성 및 조음 (3)

- Vocal Tract Resonances: Formants
  - Phoneme 들은 성도의 공명에 의해 서로 구분될 수 있음.
  - 모음들의 주파수 스펙트럼의 피크들이 발생하는데 이는 피치와는 독립적인 관계이며 이 피크들을 formant라고 함.
  - 저주파수 영역부터 차례로 F1, F2, F3, .. 등으로 명명함.

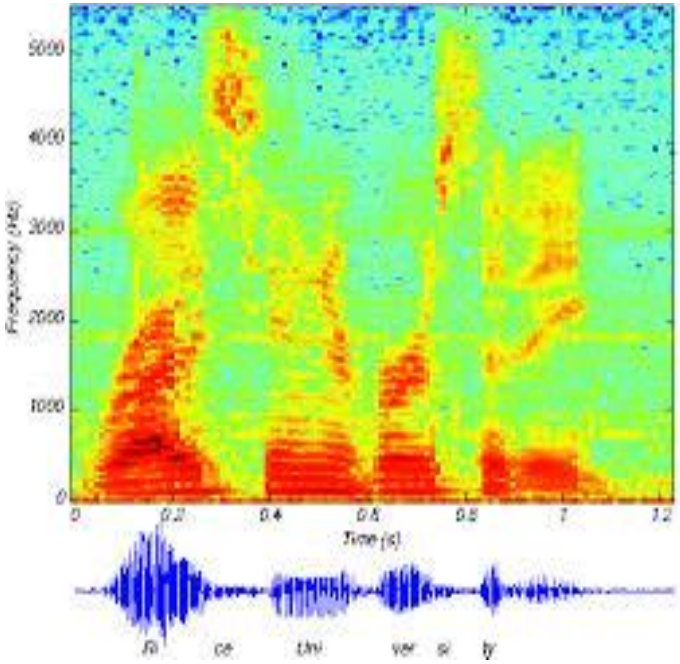


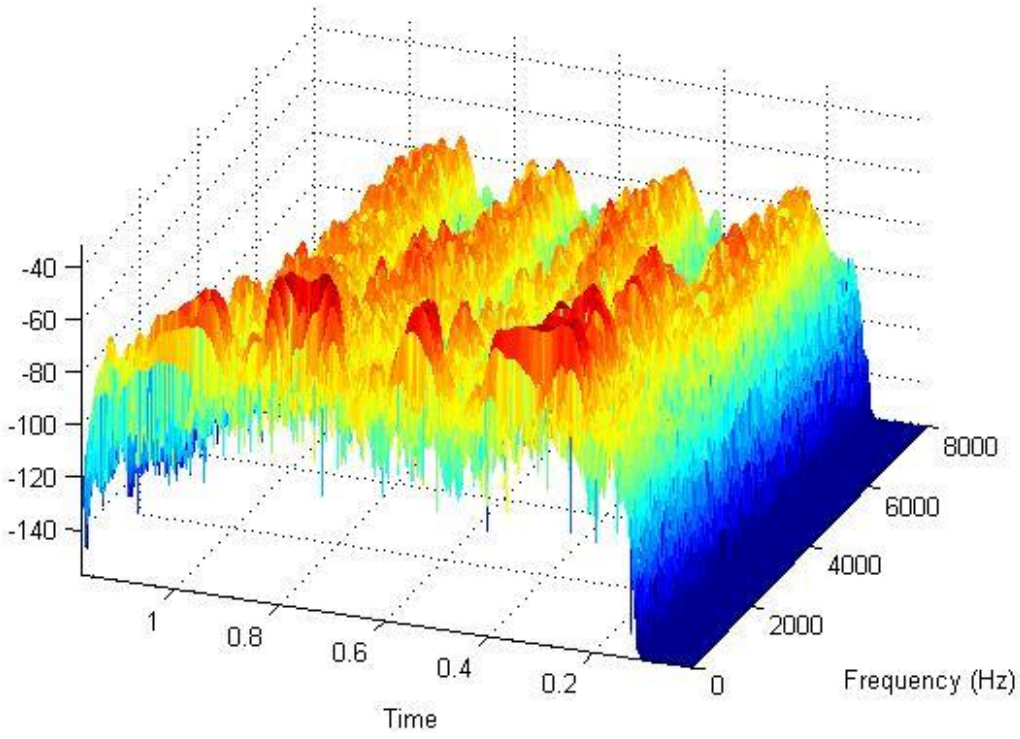
- Vocal Tract Models
  - vocal tract의 모양은 실제로 매우 복잡하지만, 과거에 많은 연구자들에 의해 단순화된 모양으로 모델링 됨.
  - 17cm closed-open cylinder 에서 발생하는 공명은 대략 500, 1500, and 2500 Hz 주파수를 나타내고 이는 모음의 주파수들과 흡사함.



Spectrogram (1)

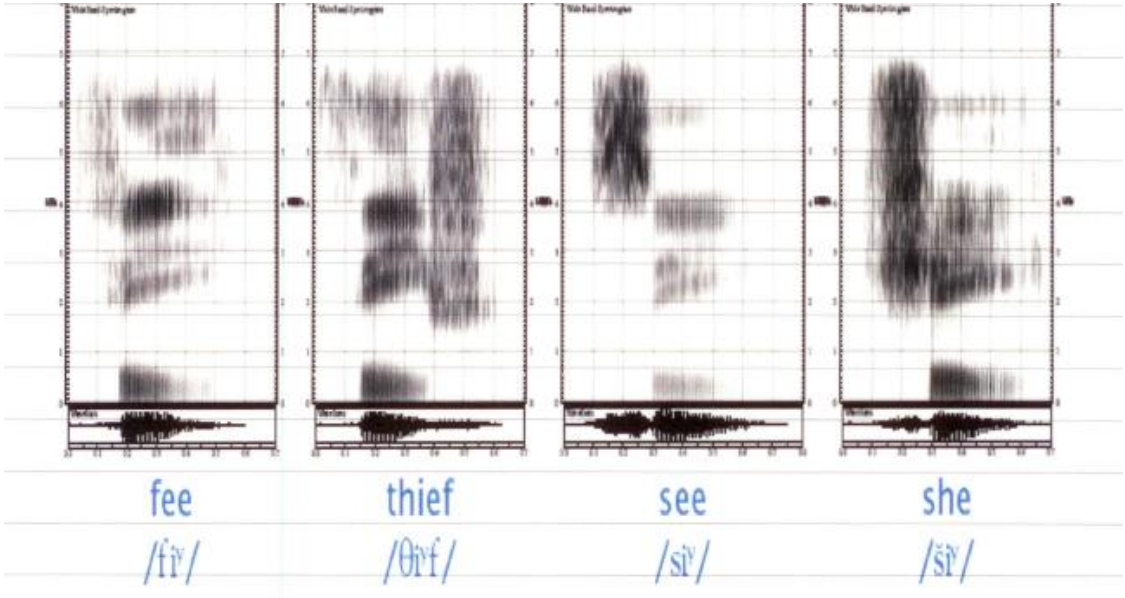
- 스펙트로그램은 소리의 시간은 변화에 따른 주파수 스펙트럼의 정보를 시각화한 것.
- X축은 시간, y축은 주파수, z축은 주파수 스펙트럼은 크기





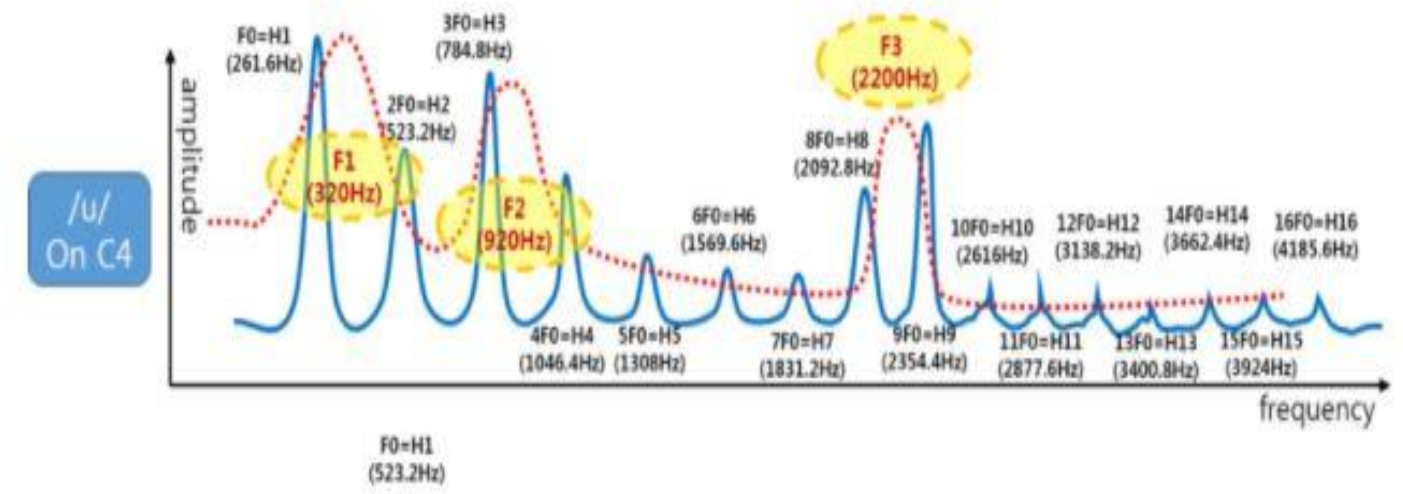
Spectrogram (2)

- 주파수 인 y축 및 진폭 z축은 그래프에 따라 선형 또는 로그 스케일로 표현될 수 있음.
- 오디오는 일반적으로 로그를 이용한 척도를 많이 사용하는 편이고, 데시벨 또는 dB 도 로그로 표현된 소리의 크기를 나타내는 수치임.
- 주파수는 하모닉스 관계를 강조하기 위해서는 선형을 선호하고, 음악적인 톤의 관계를 강조하기 위해서는 로그가 선호됨.



Formant (1)

- Formants는 Gunnar Fant에 의해 ' 음성의 the spectral peaks of the sound spectrum |P(f)| ' 로 정의되었음.
- 음성과학/음성학에서 Formant는 성도의 음향 공명을 의미하는 데도 사용됨.
- 스펙트럼 분석기를 사용하여 얻어진 소리의 주파수 스펙트럼에서 진폭 피크로 측정됨.



Formant (2)

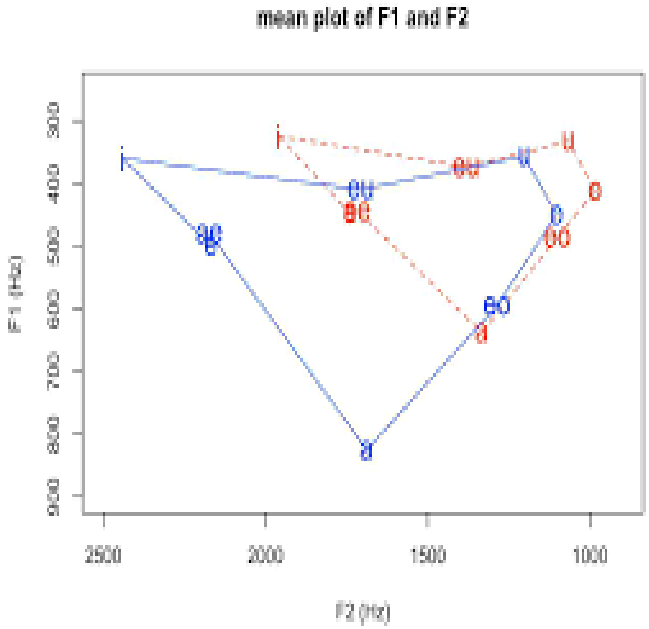
- 포먼트는 인간의 음성으로 구별되거나 의미 있는 주파수 구성 요소임.
- 특히 모음을 구별하는 데 필요한 정보는 모음 소리의 주파수 조합에 의해 정량적으로 표현될 수 있음.
- 즉 모든 모음은 자신만의 포먼트 조합을 가지고 있고 이를 통해 청취자가 그 모음을 인지할 수 있음.
- 포먼트는 저주파 대역부터 순서대로 F1, F2, F3, and so on. 대부분의 경우 두 개, 첫 번째 포먼트인 F1과 두 번째 포먼트인 F2만으로도 모음을 특정하기에 충분함.

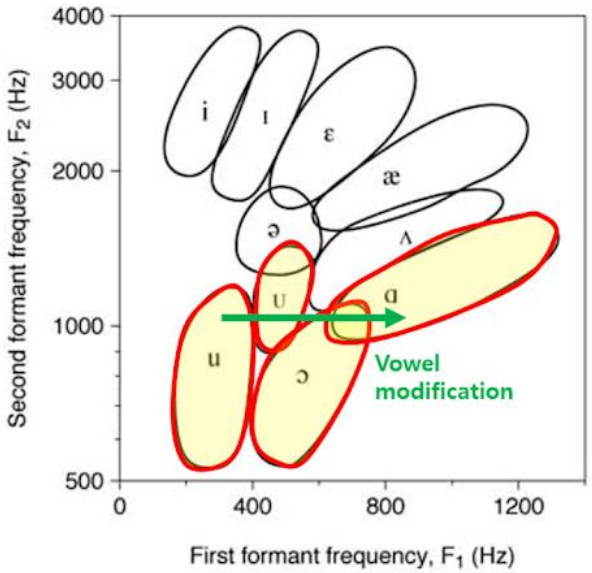
Formant (3)

- 처음 두 포먼트는 모음 품질을 결정하는 데 가장 중요하며, 이것은 종종 모음 품질의 일부 측면을 캡처하기에 충분하지 않지만 두 번째 포먼트에 대한 첫 번째 포먼트의 그래프로 표시하여 분석하는데 사용하기도 함.

Vowel formant centers

Vowel (IPA)	Formant f_1	Formant f_2
u	**320 Hz**	800 Hz
o	**500 Hz**	1000 Hz
ɑ	**700 Hz**	1150 Hz
a	**1000 Hz**	1400 Hz
ø	500 Hz	**1500 Hz**
y	320 Hz	**1650 Hz**
ɛ	700 Hz	**1800 Hz**
e	500 Hz	**2300 Hz**
i	320 Hz	**2500 Hz**

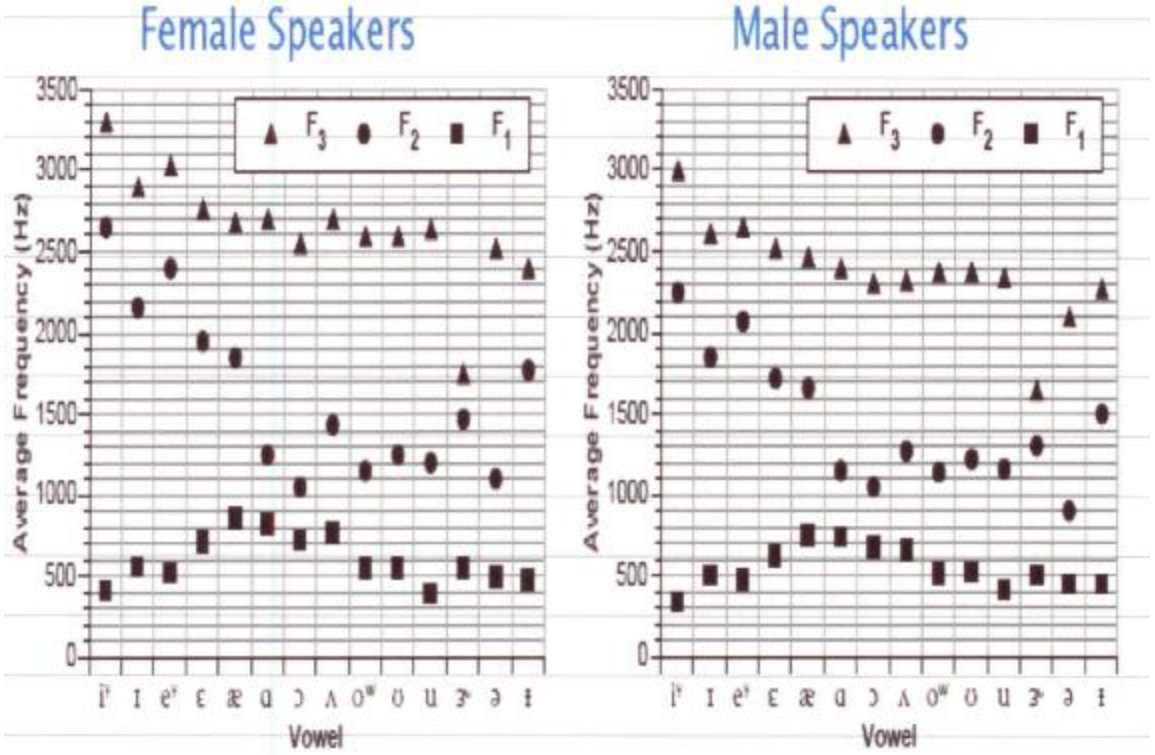




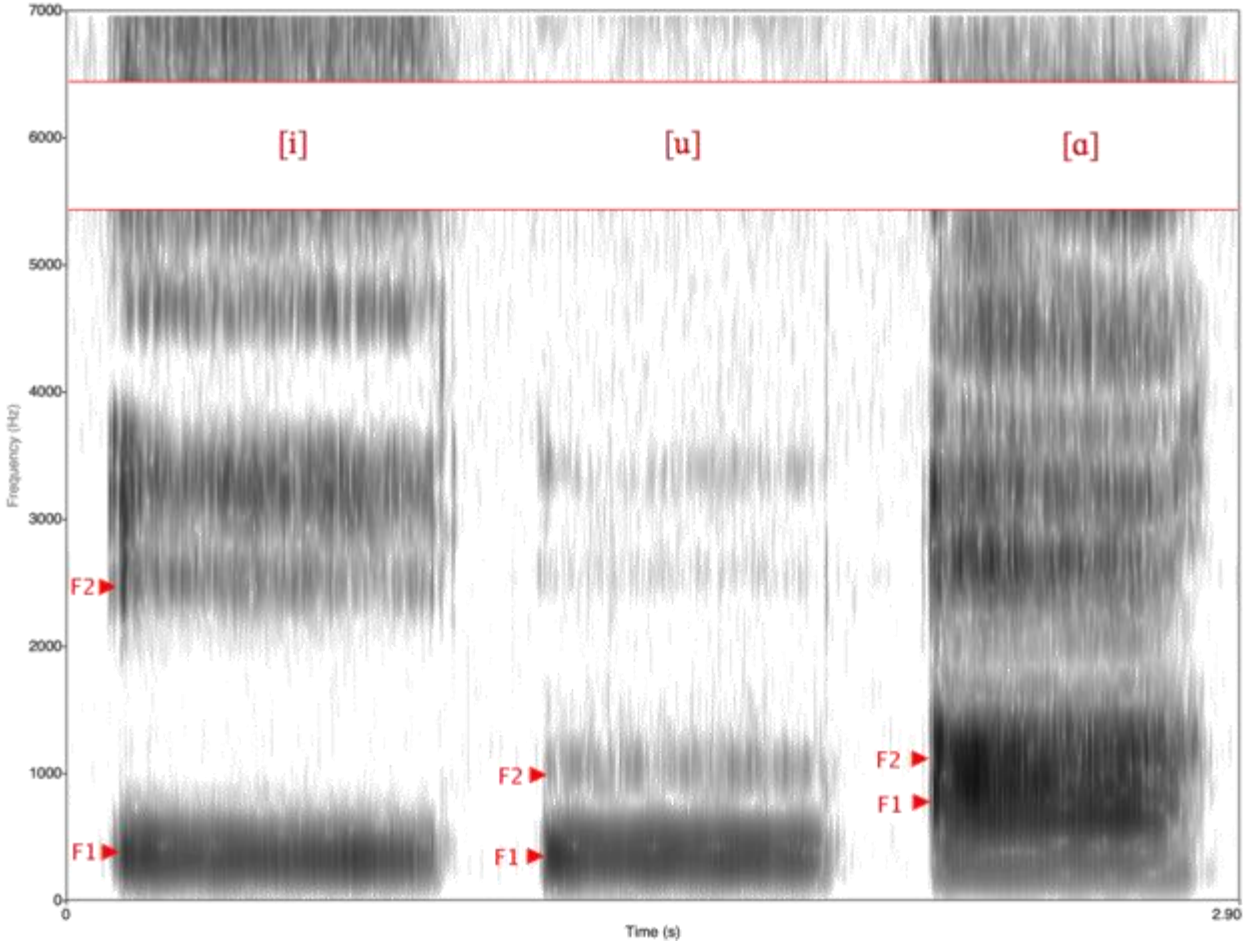
Formant (4)

- Vowel Formant Averages

- 각 모음마다 다르지만, 성별마다 평균적 포먼트의 조합이 다르고, 또한 모든 사람의 목소리가 다른 것 처럼 사람마다 동일한 모음이라도 조금씩 다른 포먼트를 나타냄.



Formant (5)


7000
6000
5000
4000
3000
2000
1000
0
Frequency (Hz)
[i]
[u]
[ɑ]
F2
F1
F2
F1
F2
F1
0
2.90
Time (s)

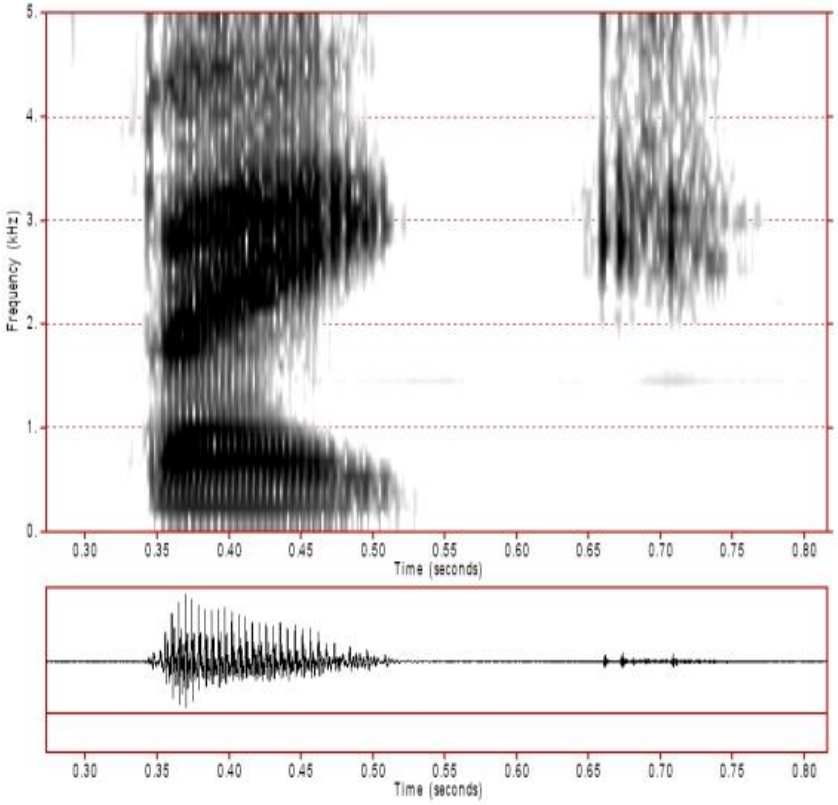
Consonants (1)

- 음성학에서 자음은 성도의 완전 또는 부분 폐쇄로 발생하는 소리임.
- 예를 들어 [p]는 입술이 동반된 발음이고, [t]는 혀의 끝 쪽이 입안의 앞쪽에서 조음되고, [k], 혀의 안쪽과 관련되고, [h] 목으로 내는 소리이고, [f] [s]는 공기 흐름의 통로를 좁혀서 난기류를 생성하여 만들어 내는 소리, [m] [n] 비강을 통한 공기흐름을 통해 조음됨.
- 자음은 주파수 조합들이 시간의 흐름과 함께 다양한 변형을 만들어내는 것인데, 이는 입이나 목구멍 등으로 공기의 흐름을 막거나 일부 막는 방법으로 만들어짐.

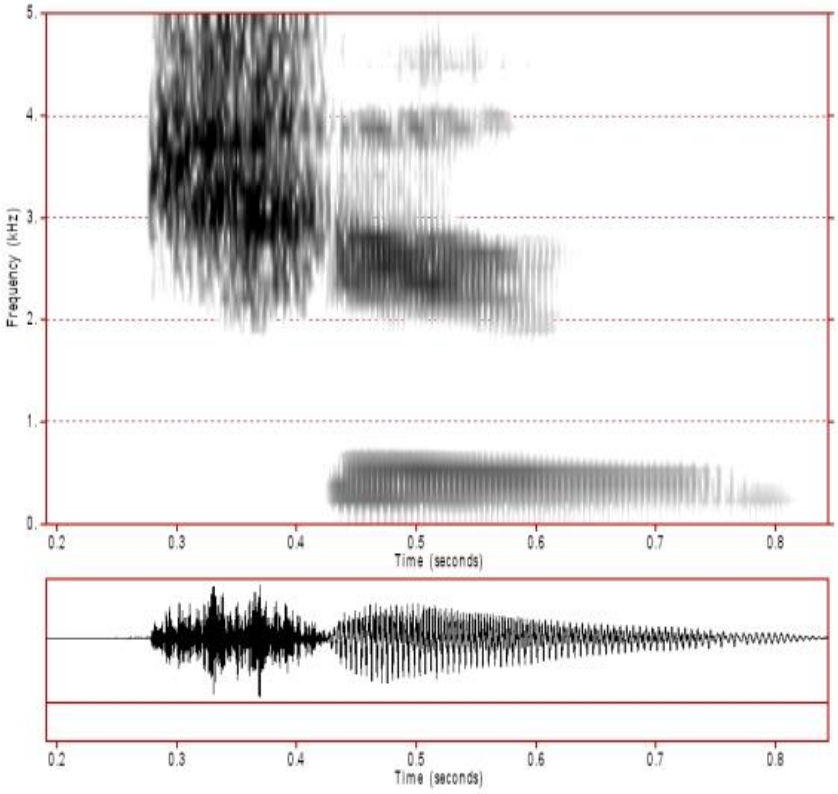
		양순음	설단음	경구개음	연구개음	후두음
파열음	예사소리	ㅂ	ㄷ		ㄱ	
	된소리	ㅃ	ㄸ		ㄲ	
	거센소리	ㅍ	ㅌ		ㅋ	
파찰음	예사소리			ㅈ		
	된소리			ㅉ		
	거센소리			ㅊ		
마찰음	예사소리		ㅅ			ㅎ
	된소리		ㅆ			
비음		ㅁ	ㄴ		ㅇ	
유음			ㄹ			

Consonants (2)

"BAKE"

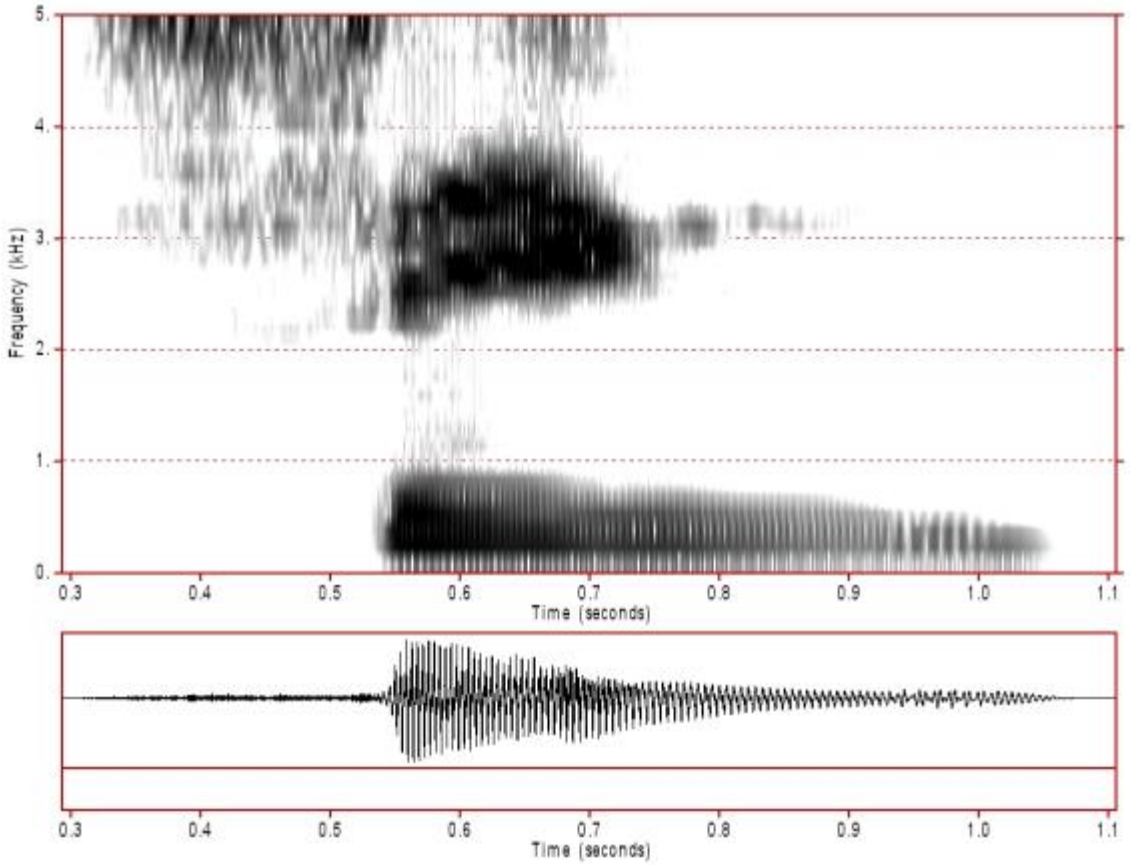


"CHEW"



Consonants (3)

- "SING"



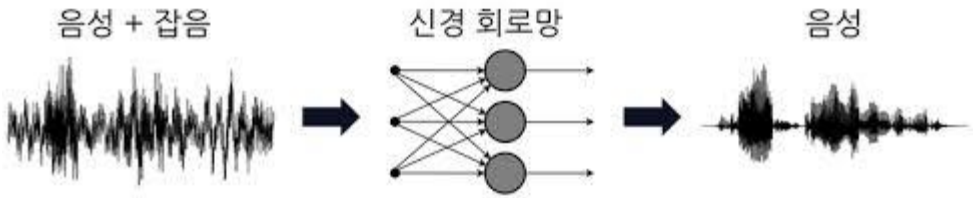
Speech Processing

- 음성 처리는 음성 신호 및 이러한 신호의 처리 방법에 대한 연구임.

- 신호는 일반적으로 디지털 표현으로 처리되므로 음성 처리는 음성 신호에 적용된 디지털 신호 처리로 간주될 수 있음.

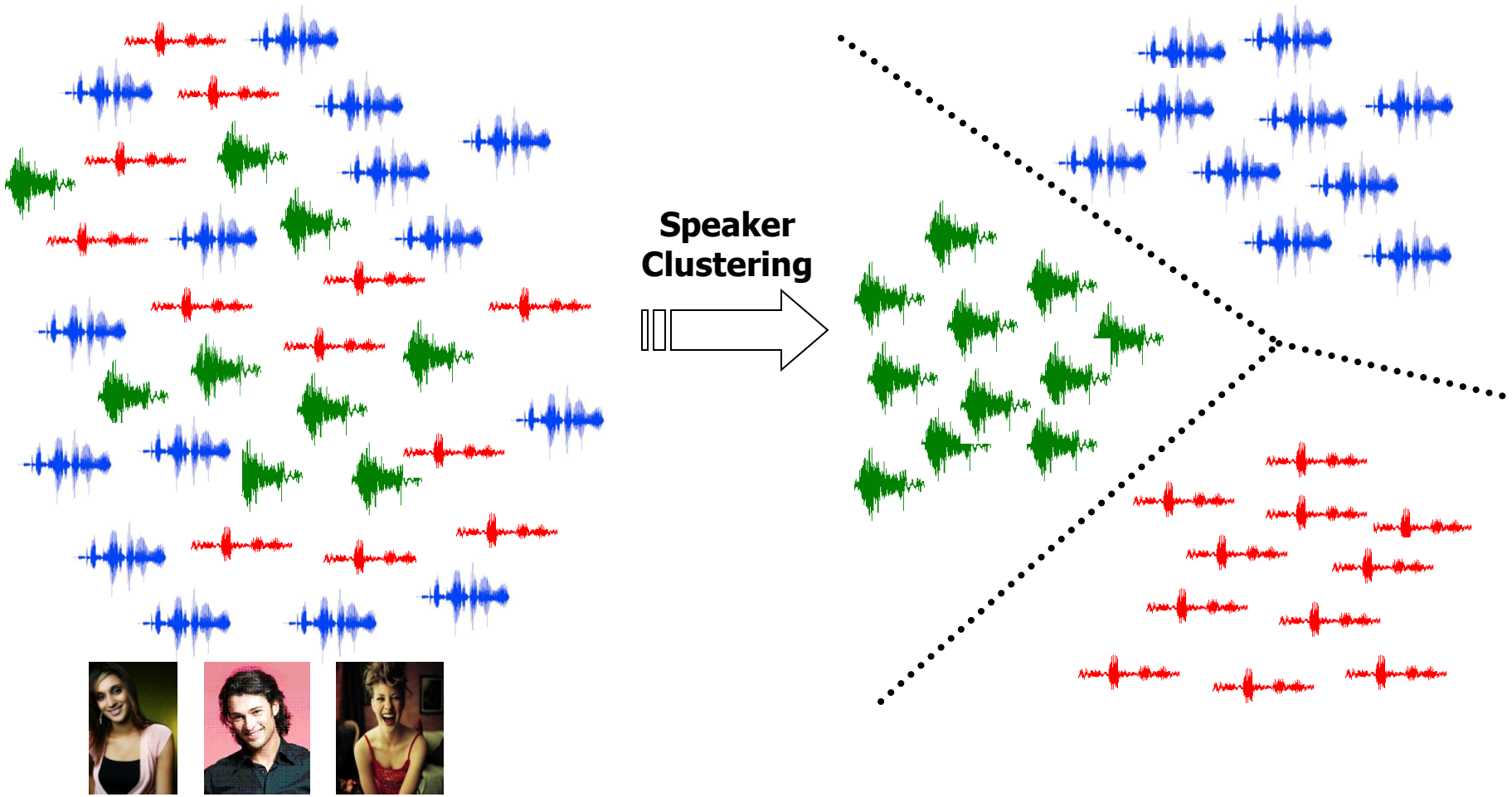
- 음성 처리는 다음 범주로 나눌 수 있음.

- Speech recognition: 음성 신호를 언어적 내용과 관련된 텍스트로 변환함.
- Speaker recognition: 말하는 사람이 누구인지를 인식함.
- Speech synthesis: 텍스트를 특정 목소리의 음성신호로 변환함.
- Speech enhancement: 오디오 관점의 노이즈 감소 등을 통해 음성 신호의 인식성과 신호의 명확성 등 품질을 향상시킴.



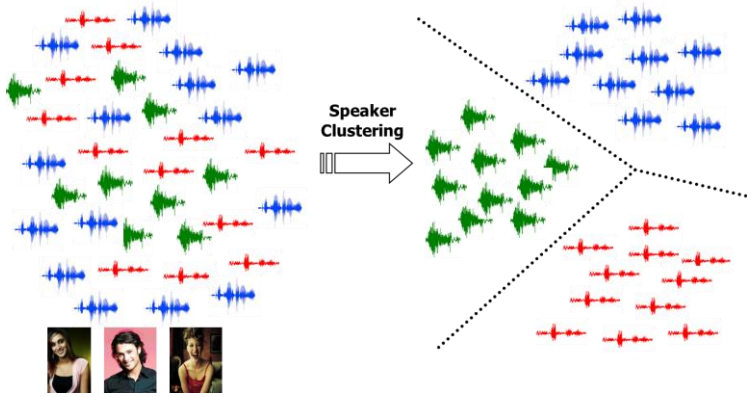
Speaker Recognition

- 화자, 즉 말하는 사람의 음성을 구분하고 분류함.

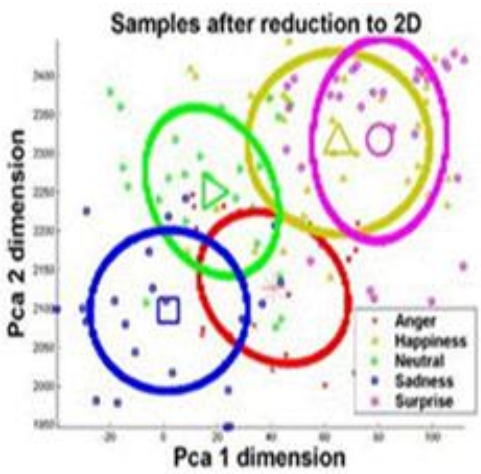


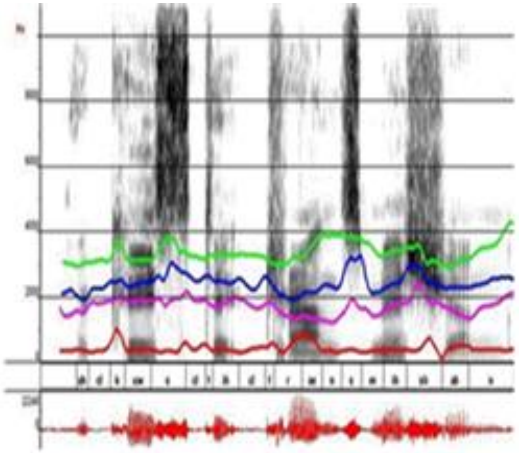
음성정보

- 음성은 인간의 의사 소통을 위해 다양한 정보를 전달하는 수단



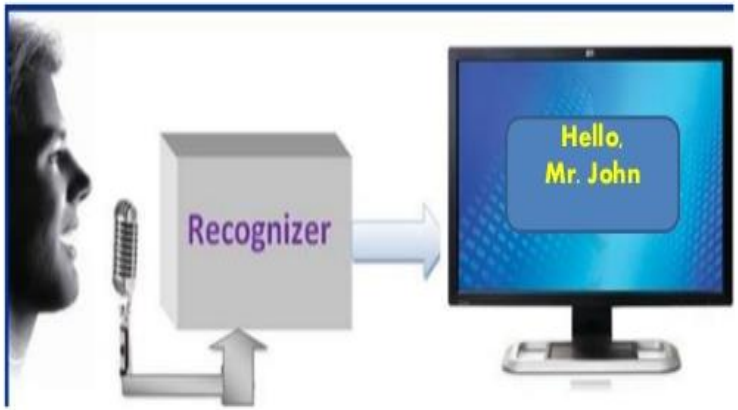
- 음성신호는 말 또는 텍스트 이외에 말하는 사람의 identity, 감정, 상황 등 다양한 정보를 포함

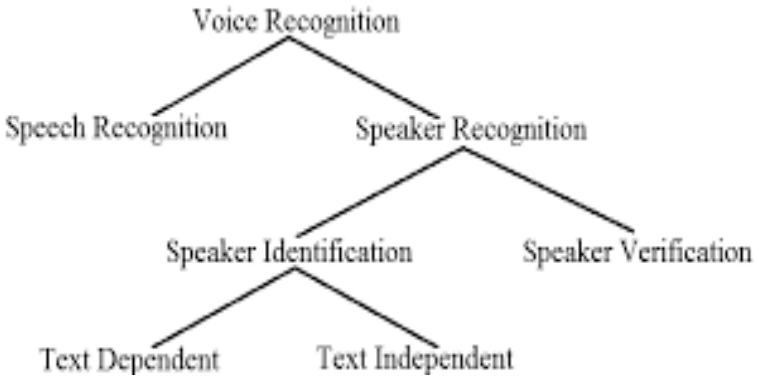




화자인식

- 말하고자 하는 단어들이 중요한 것이 아니라 누가 말하고 있는지를 인식
- Speaker Identification
- Speaker Verification/Authentication
- Text Independent/Dependent





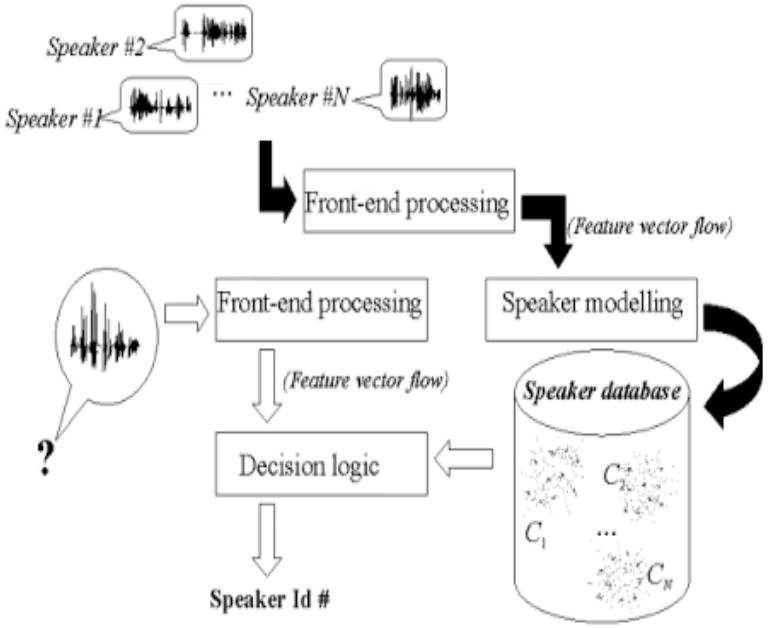
화자인식 과정

➢ 주요 모듈

1. Front-End Processing
   - 정규화
   - 특징추출
   - MFCC

2. Speaker Modelling
   - 추출된 특징 기반으로 확률통계 모델 훈련
   - GMM, HMM, Neural Network 등

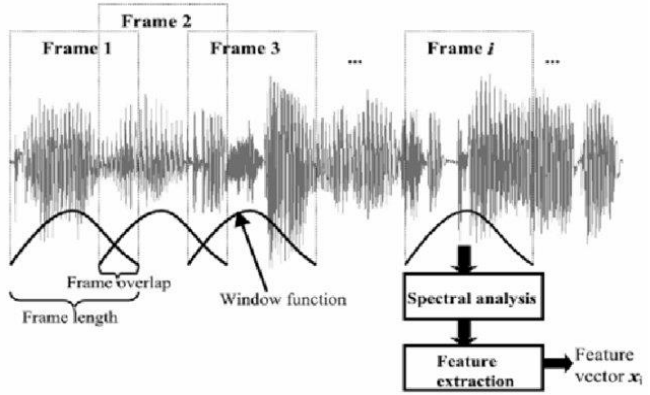
3. Decision
   - 확률적 판단
   - Identification: 가장 높은 확률
   - Verification: 정해진 수치 이상의 확률

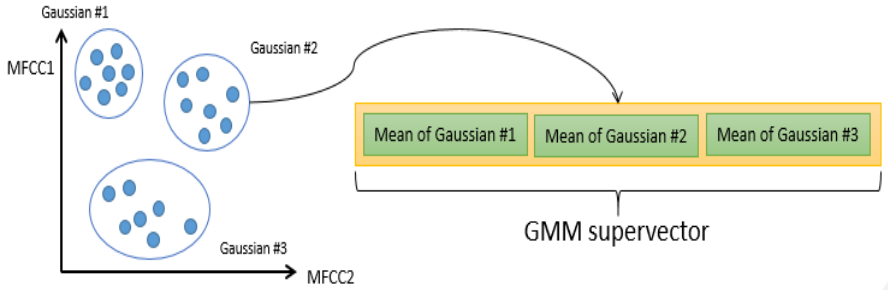


화자인식 관련 특징요소

- 특징벡터
  - MFCC: Mel-Frequency Cepstal Coefficient
    - 사람의 청각시스템을 모방한 Mel 주파수 대역 분할을 통한 특징추출
    - 비교적 simple하여 light한 시스템에 적용하기 좋음.
    - 음성인식이나 화자인식 등에서 가장 많이 사용되어 온 특징

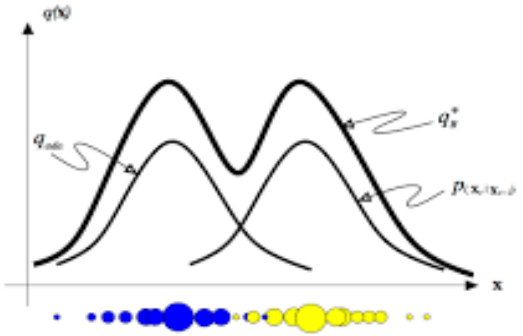
  - Gaussian Supervector
    - GMM의 각 Gaussian mixture의 mean 벡터들을 연결
    - 각 Gaussian mixture들은 서로 다른 음성 특징을 표현하므로, GMM supervector는 특정 화자의 통계적인 발화 패턴을 축약한 벡터임.
    - MFCC는 frame-level로 추출되지만, GMM supervector는 frame 수와는 무관하게 GMM 당 하나씩만 생성하여 고정된 크기를 만들기 용이하여 딥러닝에 사용하기 좋음.



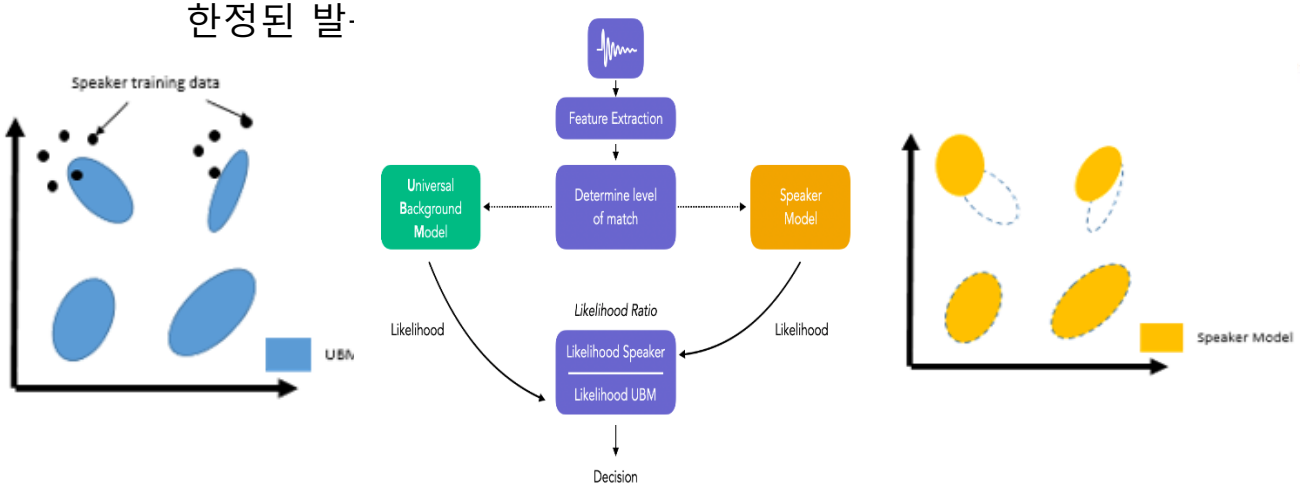


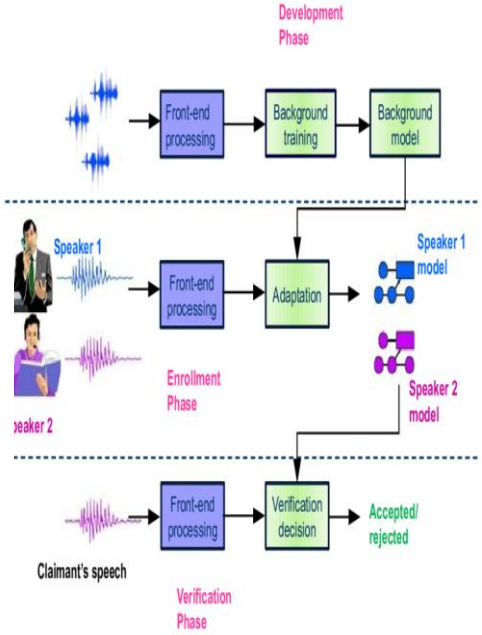
화자 모델

- GMM
  - Gaussian Mixture Model
  - 데이터의 분포를 여러개의 가우시안의 혼합을 통해서 표현
  - 여러개의 Gaussian들을 weighted sum한 것



- UBM
  - Universal background model
  - 최대한 다양한 화자 및 발음의 음성을 수집한 후 생성된 GMM
  - 소수의 화자 발성과 몇 개 안 되는 음성 샘플에 존재하는 한정된 발.





https://maelfabien.github.io/machinelearning/Speech1/#3-voice-activity-detection

DOI: 10.13140/2.1.2933.7283

화자인식 성능측정

➢ For Speaker Identification

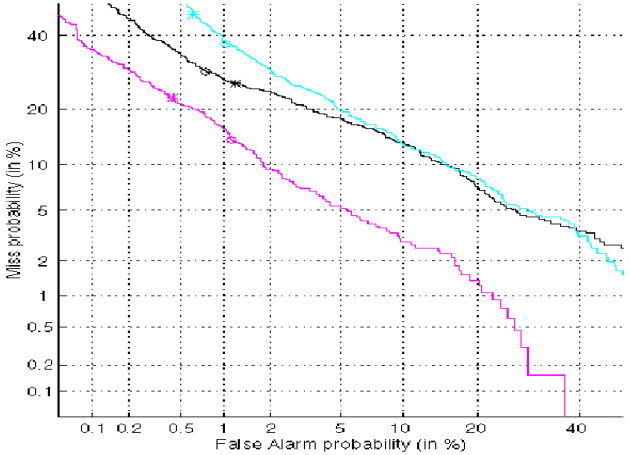
- error rate, E(ID)= n(err)/n(total)

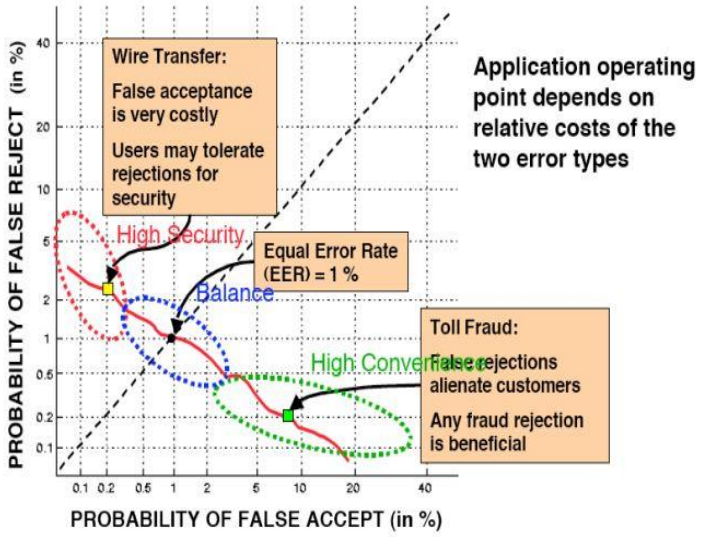
➢ For Speaker Verification

- Type I error: Imposter를 true speaker로 판정함.
- Type II error: True speaker를 imposter로 거절함.

➢ Equal error rate (EER)

- Type I error 과 Type II error 이 같아지는 지점의 오류율
- 최적의 경우에 달성할 수 있는 최소 오류율





화자인식의 실제

- Front-End Processing
  - 음성신호 중간중간의 묵음구간은 모든 사람에게 공통적인 요소로 목소리 구분에 방해됨. 하지만 소량의 묵음은 영향이 미미함. 100msec 이상의 묵음은 제거해 주는 것이 좋음.
  - 음량이 작거나 주변 잡음 대비 음량의 비율이 낮으면 인식에 방해되므로, 녹음 시 최대한 잡음 없이 진행하고 이후 정규화 해주는 것이 좋음.

- Feature Extraction
  - 어떤 특징벡터를 사용하는지, 한 특징벡터에서 몇 차원의 벡터를 사용할지는 실험적, 경험적으로만 알 수 있음.

- Speaker Model
  - 최적의 모델은 없으며, 성능이 좋아지는 방법일 수록 모델 훈련에 필요한 데이터 양이 증가함. 그래서 사용하는 용도 및 데이터 확보 용이성 등에 따라 선택해야 함.

- Decision Making
  - 화자인증에서는 Detection Error Trade-off 를 고려하여 적용하려는 목적이나 기준에 따라 셋팅함.
  - 그 외의 화자인식에서는 특별한 고려사항은 없지만, 추가적인 목적이나 기준이 있다면 그 에 맞는 부가적인 알고리즘을 사용할 수 있음.

응용분야

➢ 보안기술

- 개인고유의 목소리를 통해 본인인증을 하므로, 다른 보안기술들(스마트카드, PIN)과 달리 해킹과 분실의 염려가 없어 안전
- 원격인증이 가능하고 비용경쟁력이 높다는 장점
- 기술이 발전하여 이제는 감기 등의 이유로 인한 목소리 변화, 음성변조, 녹음된 목소리 등으로 인한 문제점들도 해결

기술	화자인식	PIN	스마트카드	지문인식	홍채인식	얼굴인식
비용효과	낮음	낮음	높음	높음	높음	높음
보안 시스템에 의한 접근제한	YES	YES	NO	NO	NO	NO
원격인증	YES	YES	NO	NO	NO	NO
안전성	YES	NO	NO	YES	YES	YES
침입봉쇄	YES	YES	YES	NO	NO	NO
편리성	YES	NO	NO	YES	YES	YES

자료: "3SH Consulting"

➢ 개인화 서비스

- 개인고유의 목소리를 통해 누군지를 인식하여 개인화된 서비스를 제공
- AI 스피커, AI 비서, 자동 회의록 등에 활용