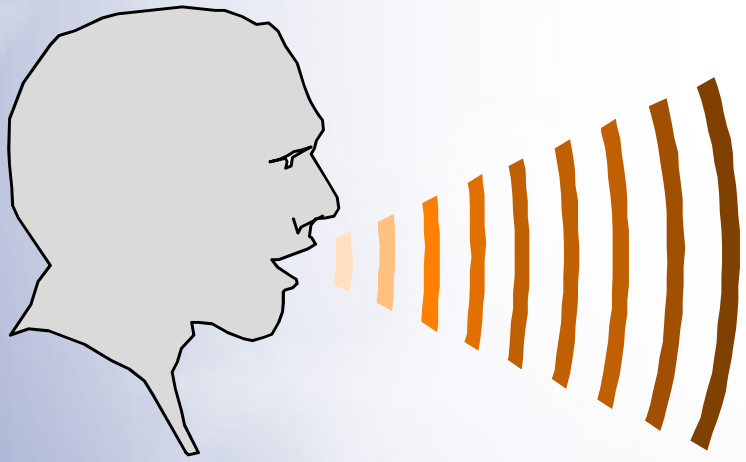
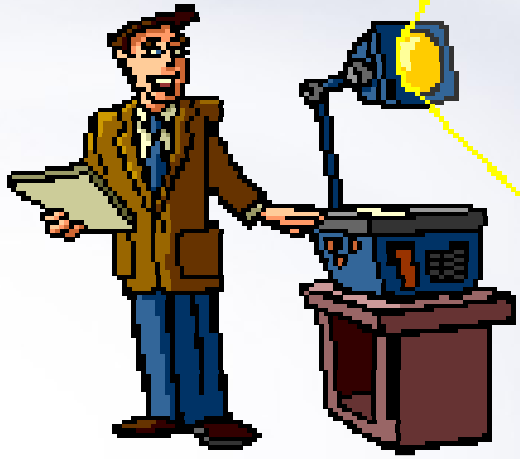
Biometric Authentication



Lecture 11

Behavioral Biometric: Voice



Outline

- Introduction
- Speech & Speaker Identification
- Speaker Verification

Voice Identification

Voice

Voice

❑ Current State

- Utilizes the distinctive aspects of the voice to verify the identity of an individual. The least invasive of the biometric recognition technologies and the most natural to use in speech system.
- Have the most potential for growth, because it requires no new hardware — most PCs already contain a microphone.
- Just say a phrase, about a second long - any language or dialect - chosen by the user. A typical case is AT&T Smart Card.

❑ Feature Set

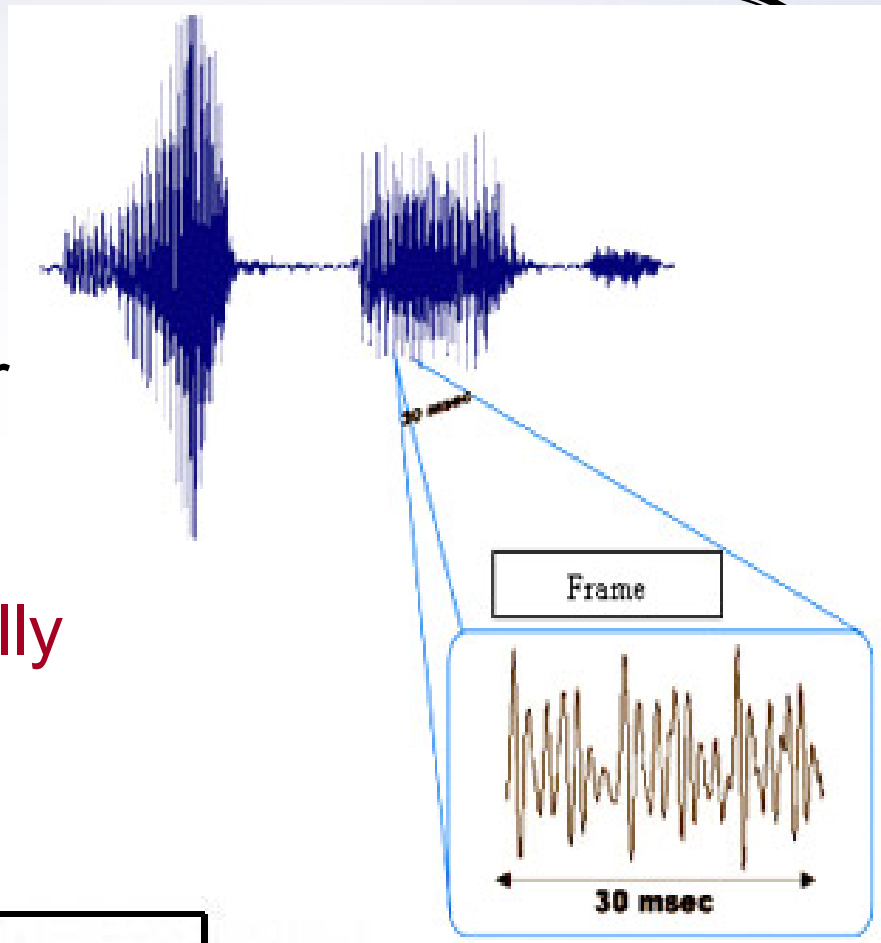
Cadence, frequency, pitch & tone of an individual's voice.

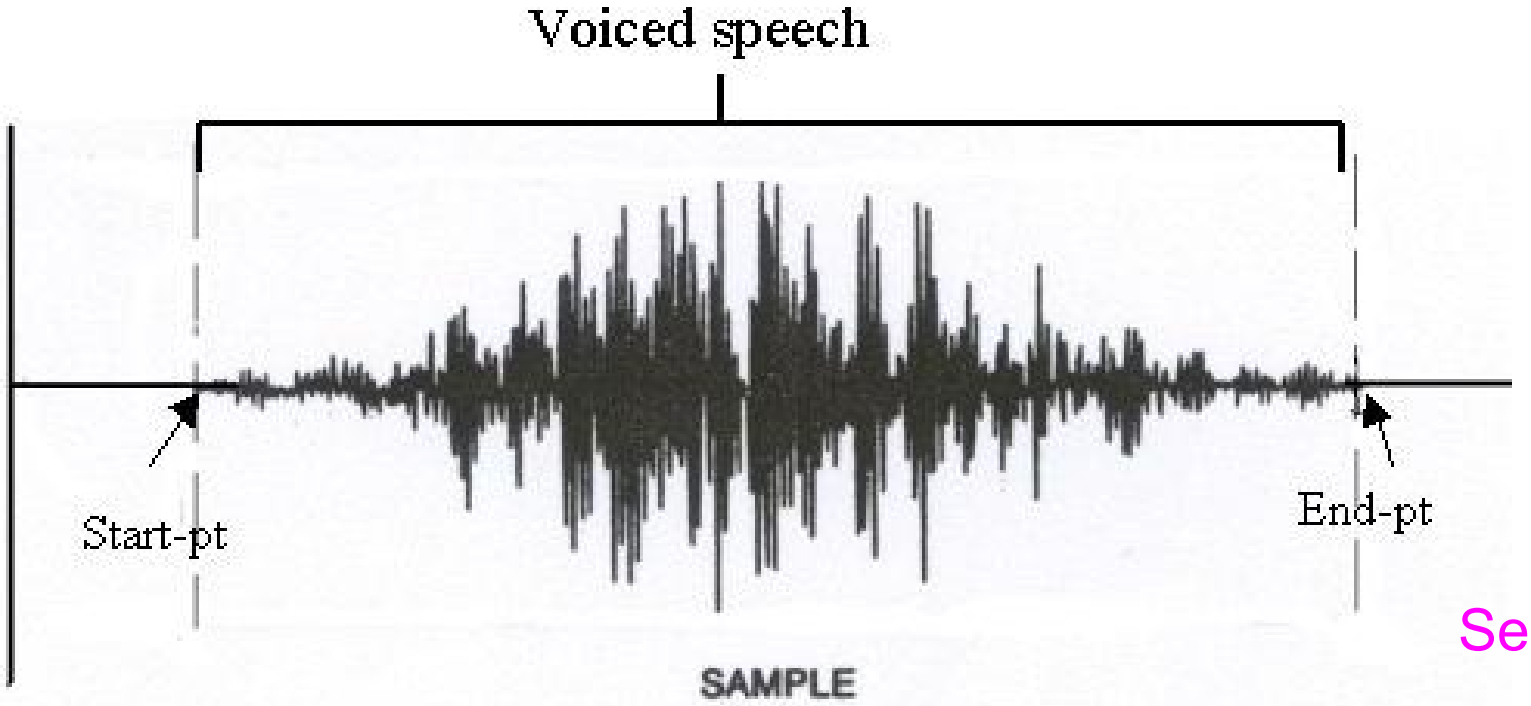
Voice Biometrics

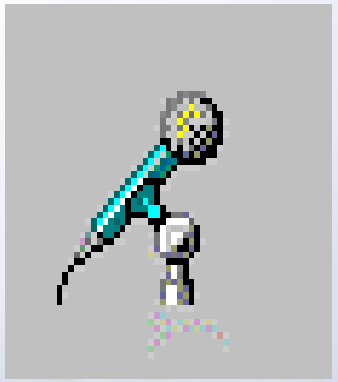
❑ Concept:

--- Authenticating (confirming or denying) a person's claimed identity using people's voice

--- Something done very naturally by people



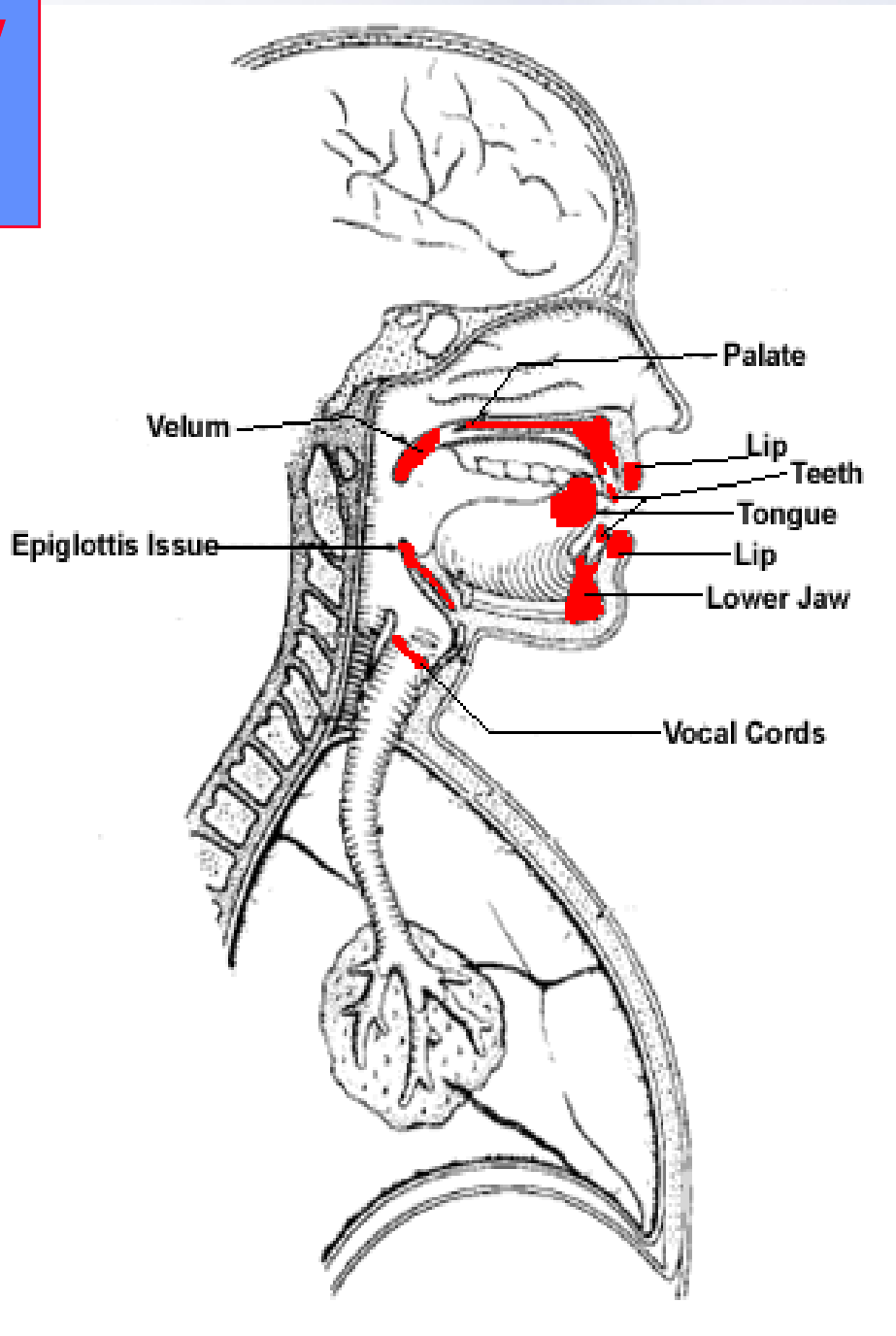




Sensor: microphone

Introduction – How Voice is Generated?

- Voice is generated by glottal pulses through vocal tracts, including epiglottis, lower jaw, tongue, velum, palate, teeth and lips
- Behavioral characteristics (the way people speak), and physiological characteristics (glottis size, lip, tongue shapes) vary



Desirable Properties

Uniqueness

- Everybody has different shape of lip, jar, tongue glottis, and speaking habits

Universality

- Everybody can speak

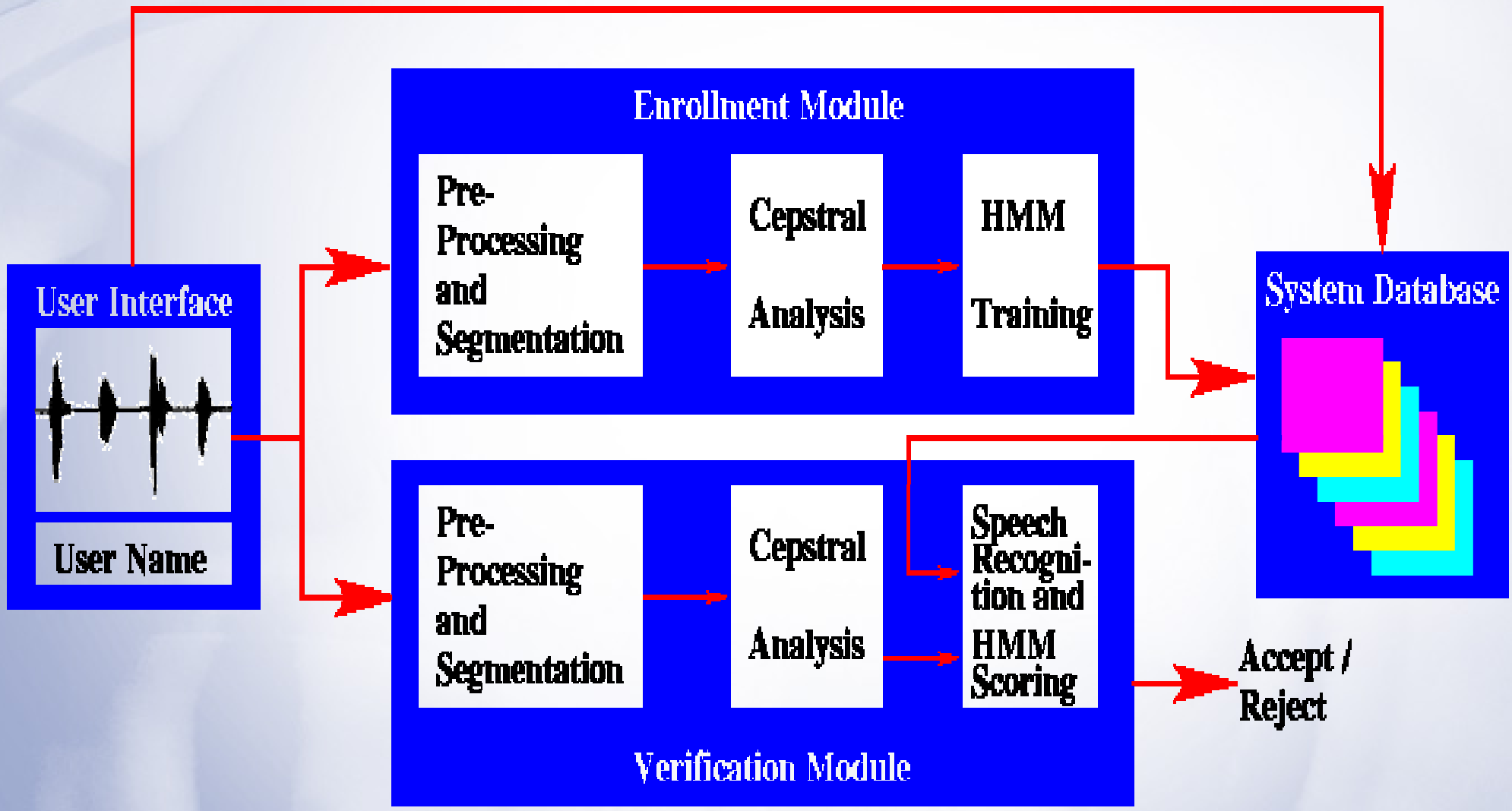
Measurability and Collectability

- Voice can be collected by simple equipment
- Voice signals (spectrum) can be measured

User Friendliness

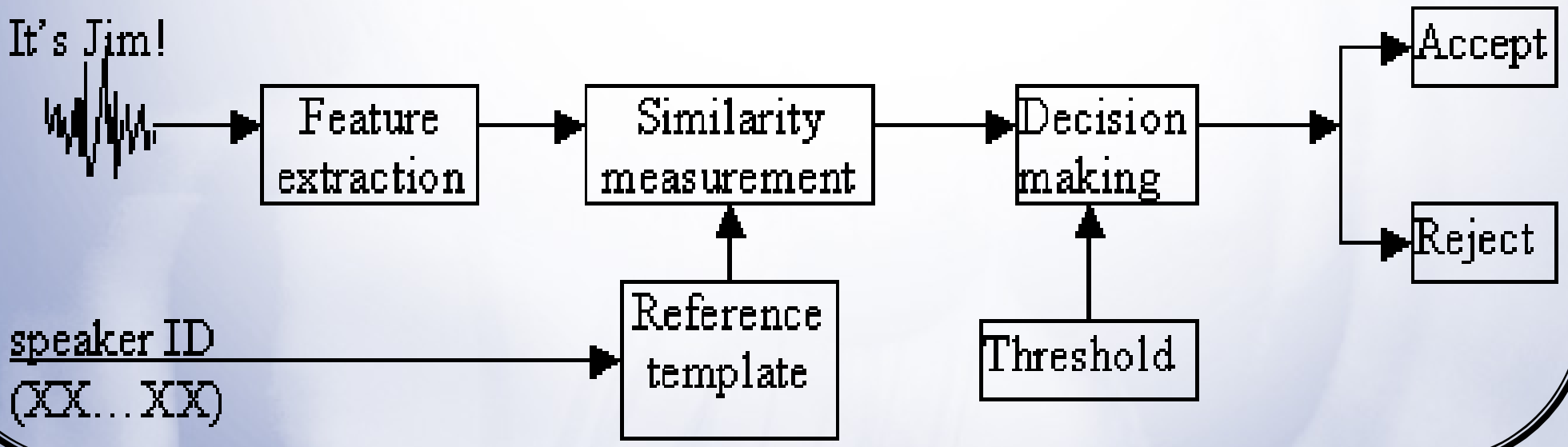
- Voice should be easy and comfortable to be collected and measured

Basic Structures of a Speaker Verification System



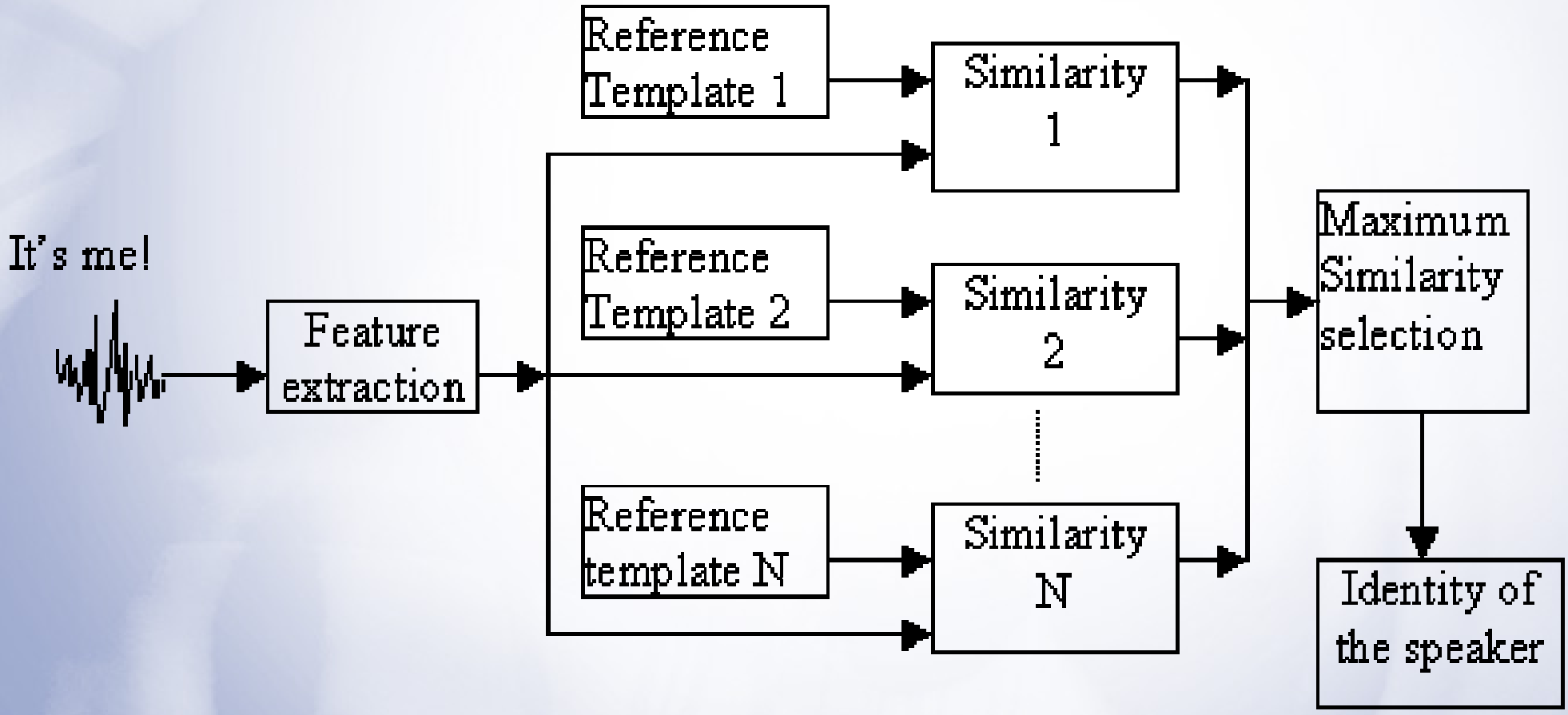
Speaker Verification System

- Aims to answer question: **"Is the speaker who he/she claims to be?"**
- 1-to-1 matching;
- Usually deployed in 2 stages:
  - Enrollment Session: teach and train the system with the speaker's voice samples (voiceprint)
  - Test Session: suing extraction and decision algorithms to verify the speaker's claimed identity

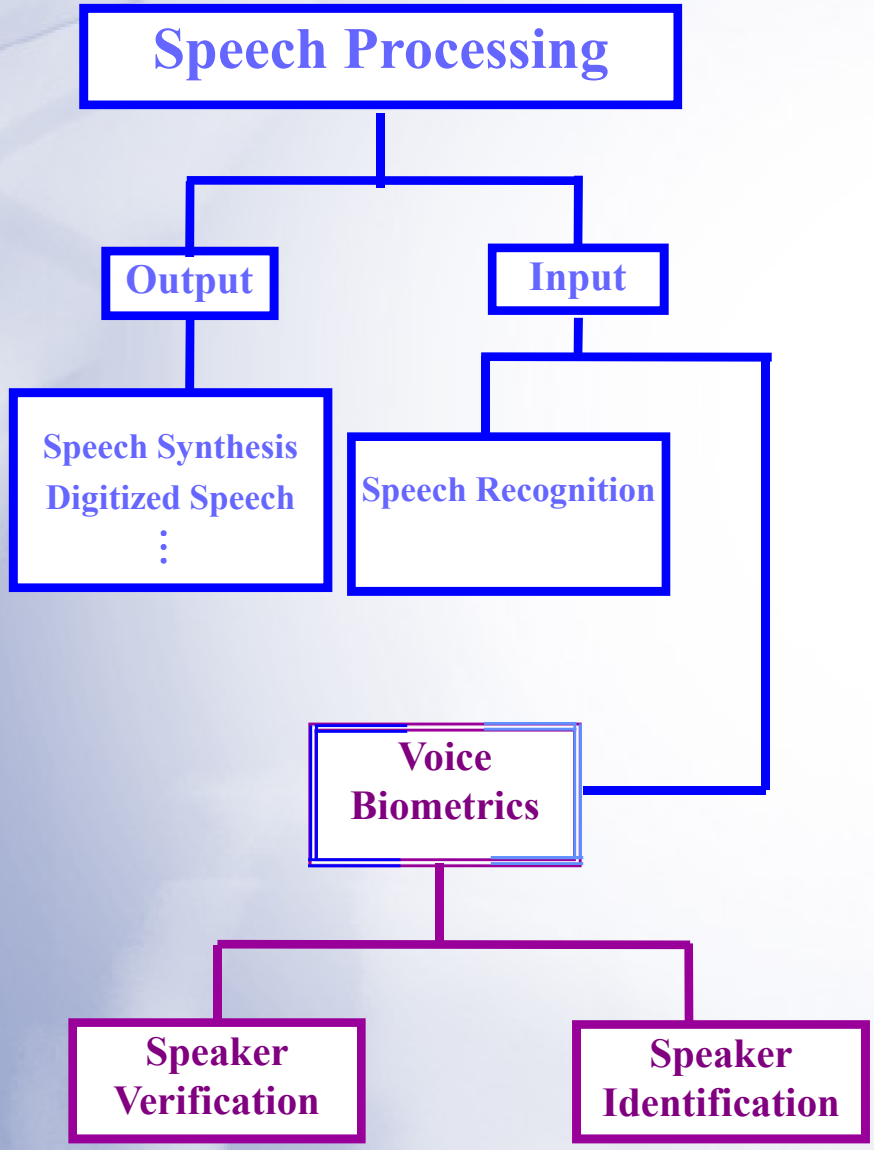


Speaker Identification System

- Similar to Verification System, but the user ID is not know before.



Family Tree: Voice Biometrics



Speaker Verification

- Extract information from the stream of speech.
- Verifies that a person is who she/he claims to be.
- One-to-one comparison.

Speaker Identification

- Extract information from the stream of speech.
- Find out an identity of an unknown person from the voice
- One-to-many comparison.

Speech Recognition

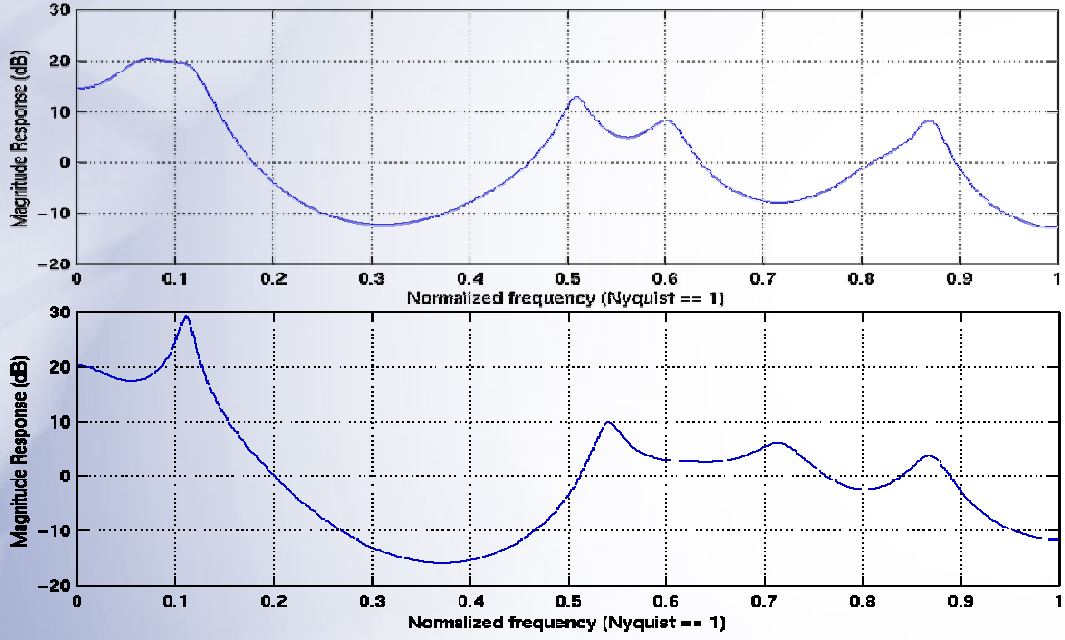
- Extracts information from the stream of speech.
- Figures out what a person is saying.

Background

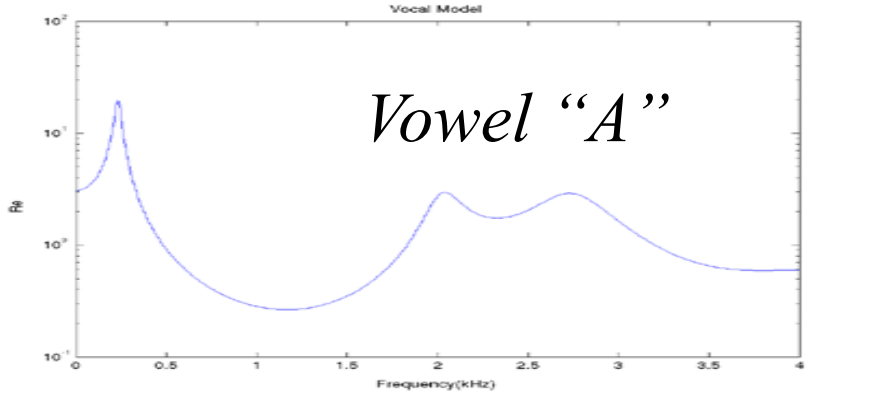
- **Text-dependent speaker ID**
  - Provide utterance of key words or sentences that are the same for training and recognition.
  - Example: Precept(Six Words) of Monk- “唵、嘛、呢、叭、咪、吽
    唵（**an**）嘛（**ma**）呢（**ni**）叭（**ba**）咪（**mei**）吽（**hong**）
- **Text-independent speaker ID**
  - Verifies the identity of the individual who is speaking.
- The performance (of verification) can vary according to:
  - The quality of the audio signal
  - Ambient noise
  - The variation between enrollment and verification devices
- So, the acquisition process usually uses the same device where the verification will take place.

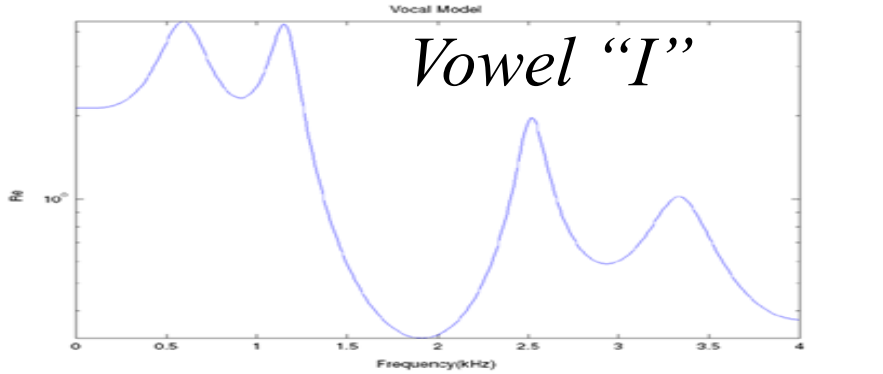
Voice Analysis

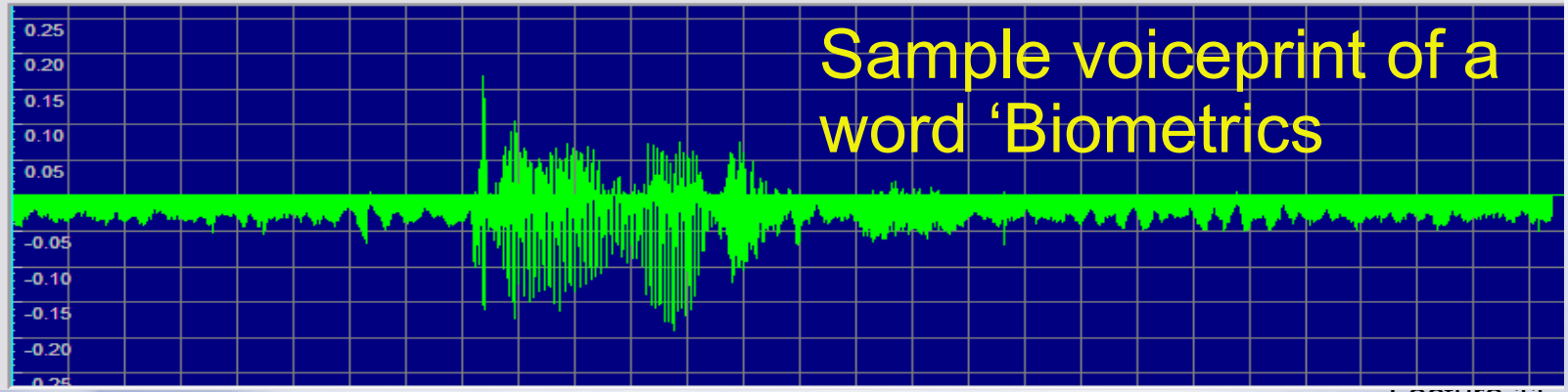
- Differences in the models for two speakers saying the same vowel “A”



- Differences vowel could have different waveforms.







Speaker Verification

System processes:

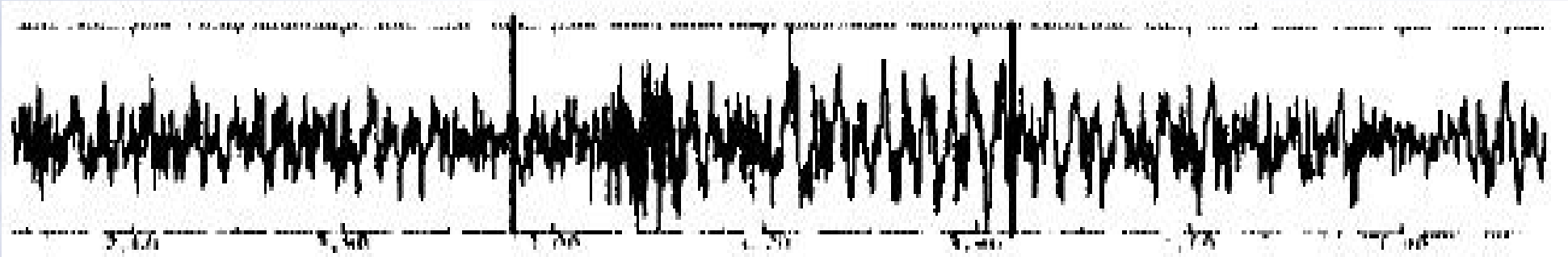
- Waveform acquisition
- Feature extraction
- Signal processing
- Feature evaluation

Stage 1: Waveform Acquisition

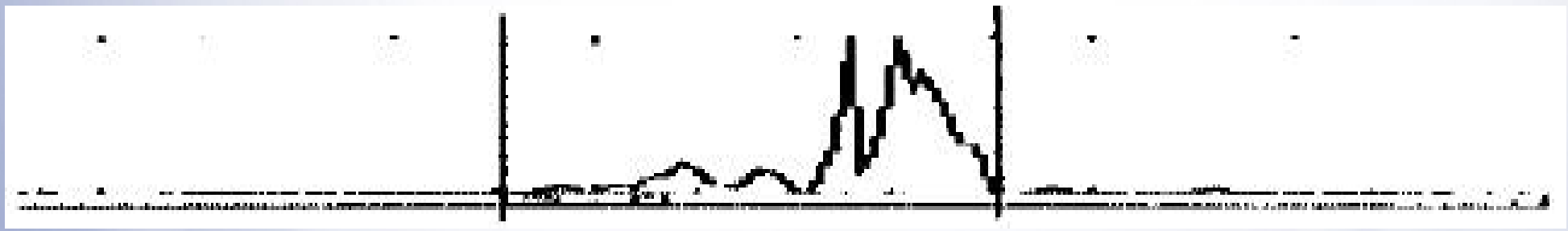
- Speech is captured by the device like microphone.
- The device convert users acoustic wave to analog signal.
- The analog signal is converted to digital signal by an analog-to-digital (A/D) converter.

Stage 2: Signal Processing

- Handle the noise by noise suppression.



- Figure: Voice data with noise



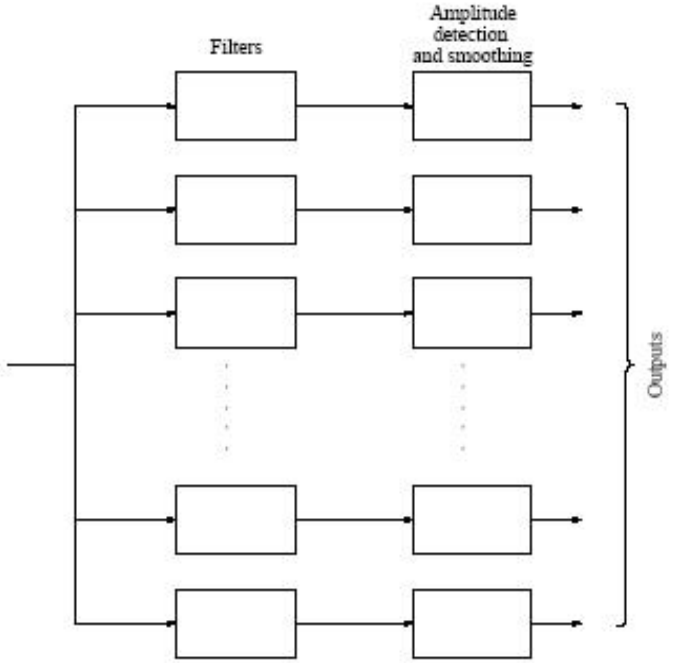
- Figure: noise is suppressed and produce voice data with a clear boundaries

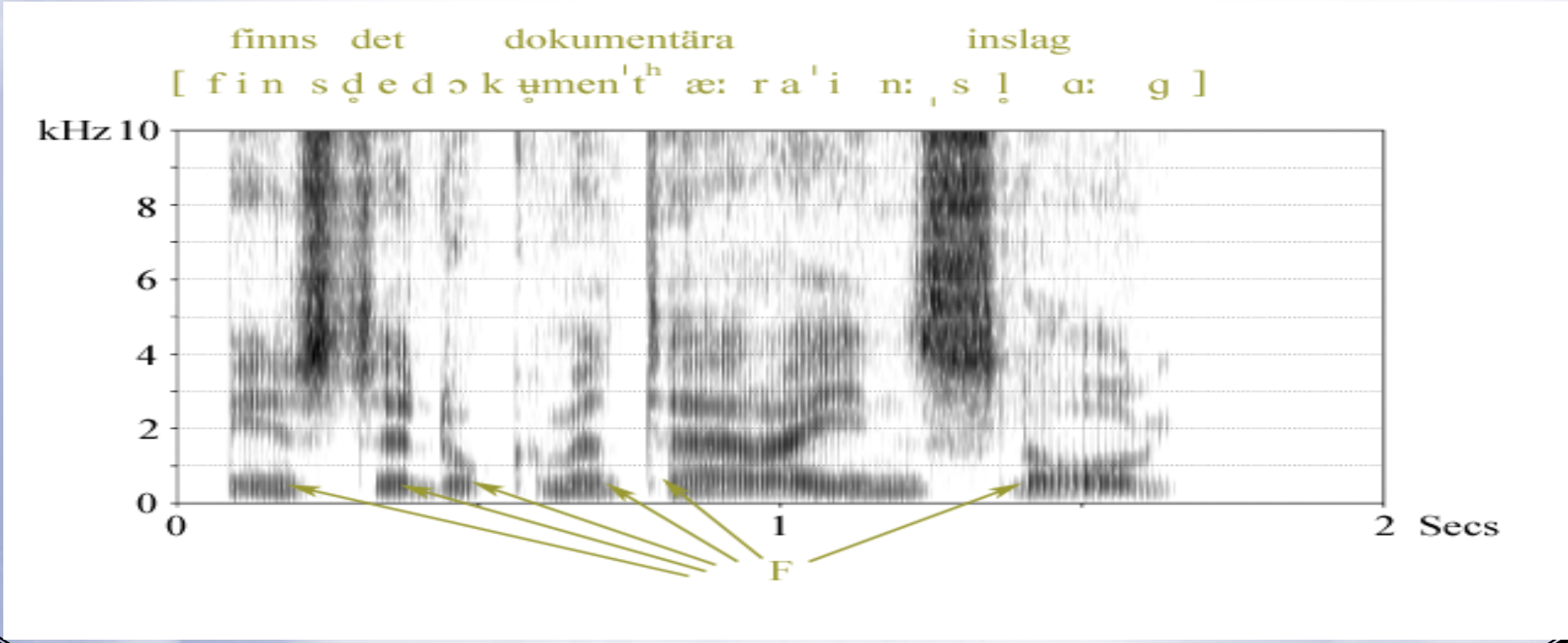
Stage 3: Feature Extraction

- Feature extraction is the estimation of variables, called a feature vector.
- The following will give a brief summary of some features. These features may be used in combination in the real system.
  - Frequency-Ban Analysis
  - Identification from Spectrograms
  - Use of Coarticulation
  - Formant Frequencies
  - Pitch Contours
  - Features derived from Linear Prediction

Frequency-Ban Analysis

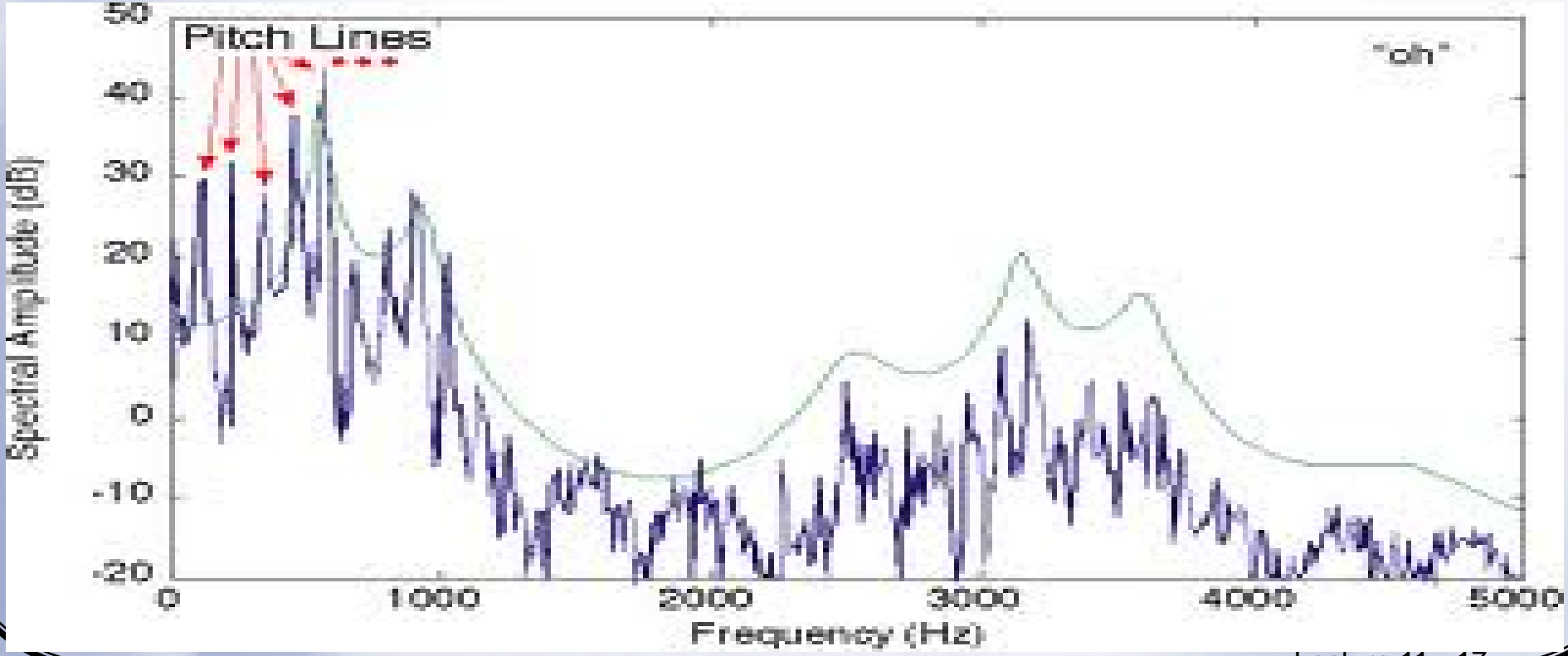
- A filter bank system is used in frequency-ban analysis.
- The outputs are sampled and produce spectrum information for comparison.





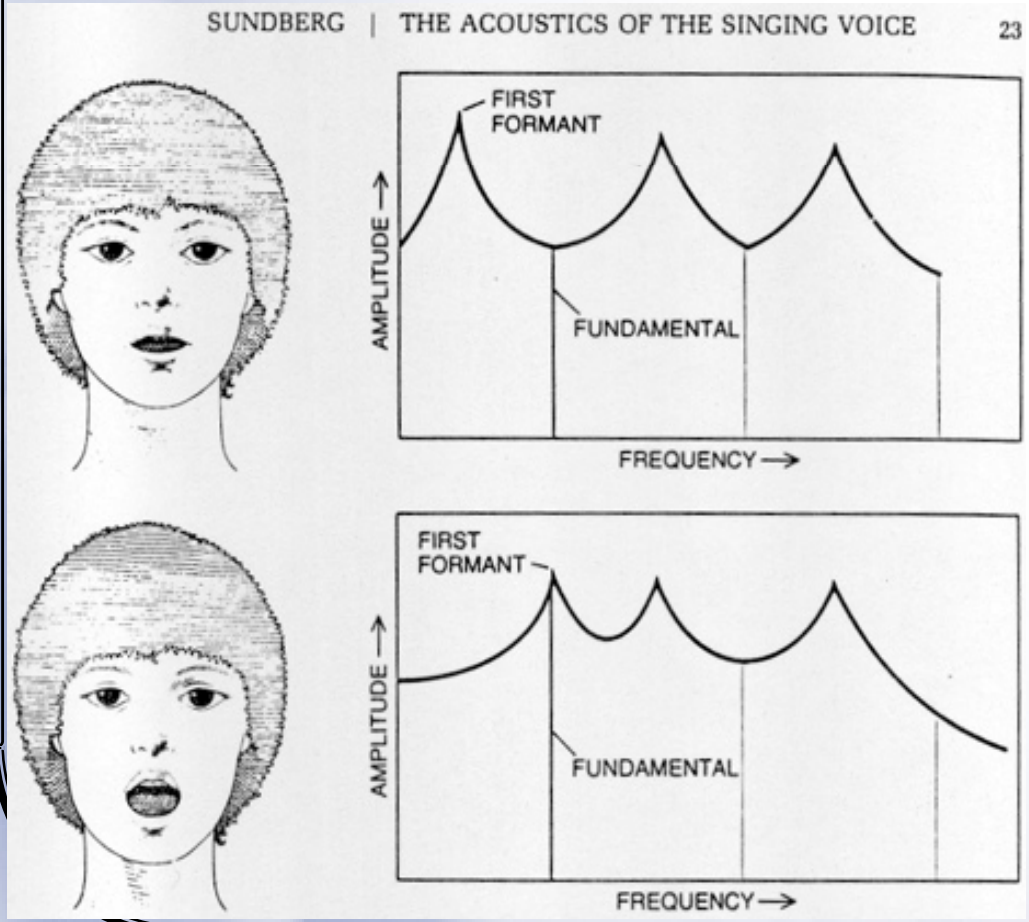
Identification from Spectrogram

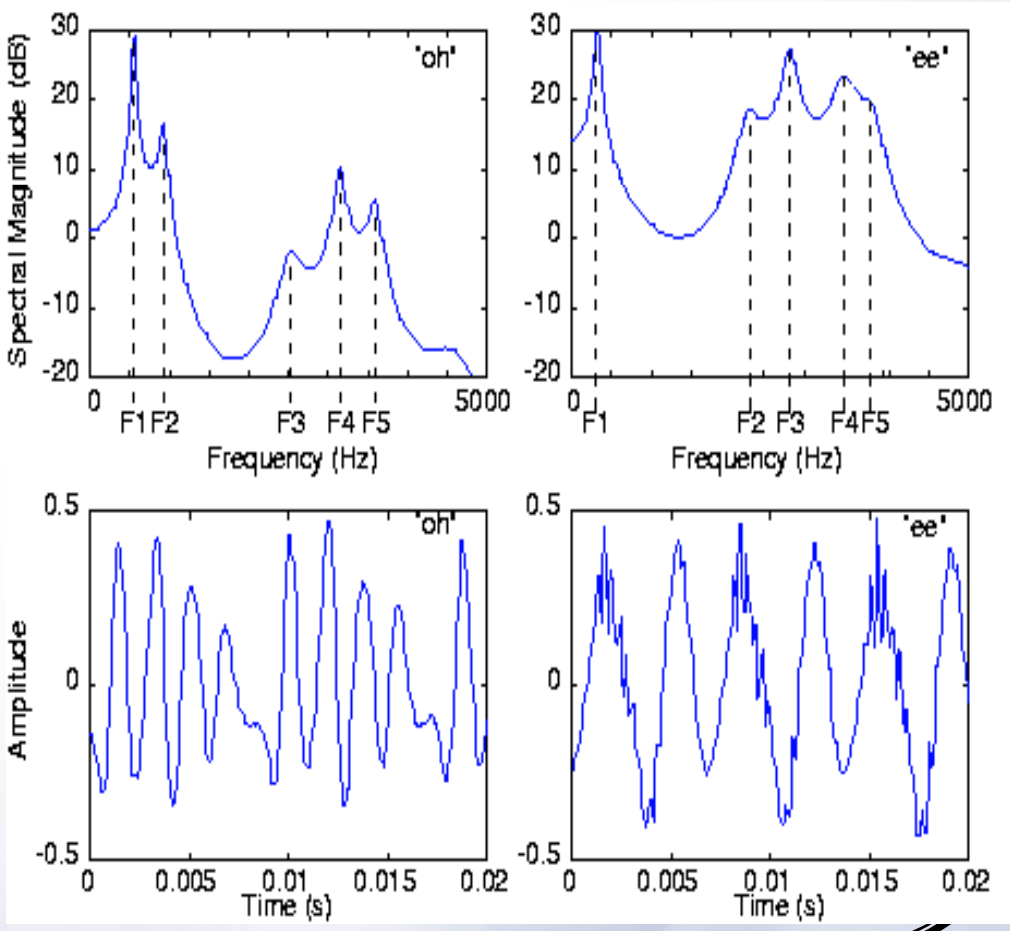
- A spectrogram is a time-varying spectral representation (forming an image) that shows how the spectral density of a signal varies with time.
- Spectrogram
  - Energy distribution of speech signal
  - Visual comparison of spectrogram to recognize the speaker



Formant Frequencies

- The resonances in the spectrum are referred as a formants.
- Different speech organs produce different regions of resonances. The **position** of the resonances can determine the differences between the speakers.





Use of Coarticulation

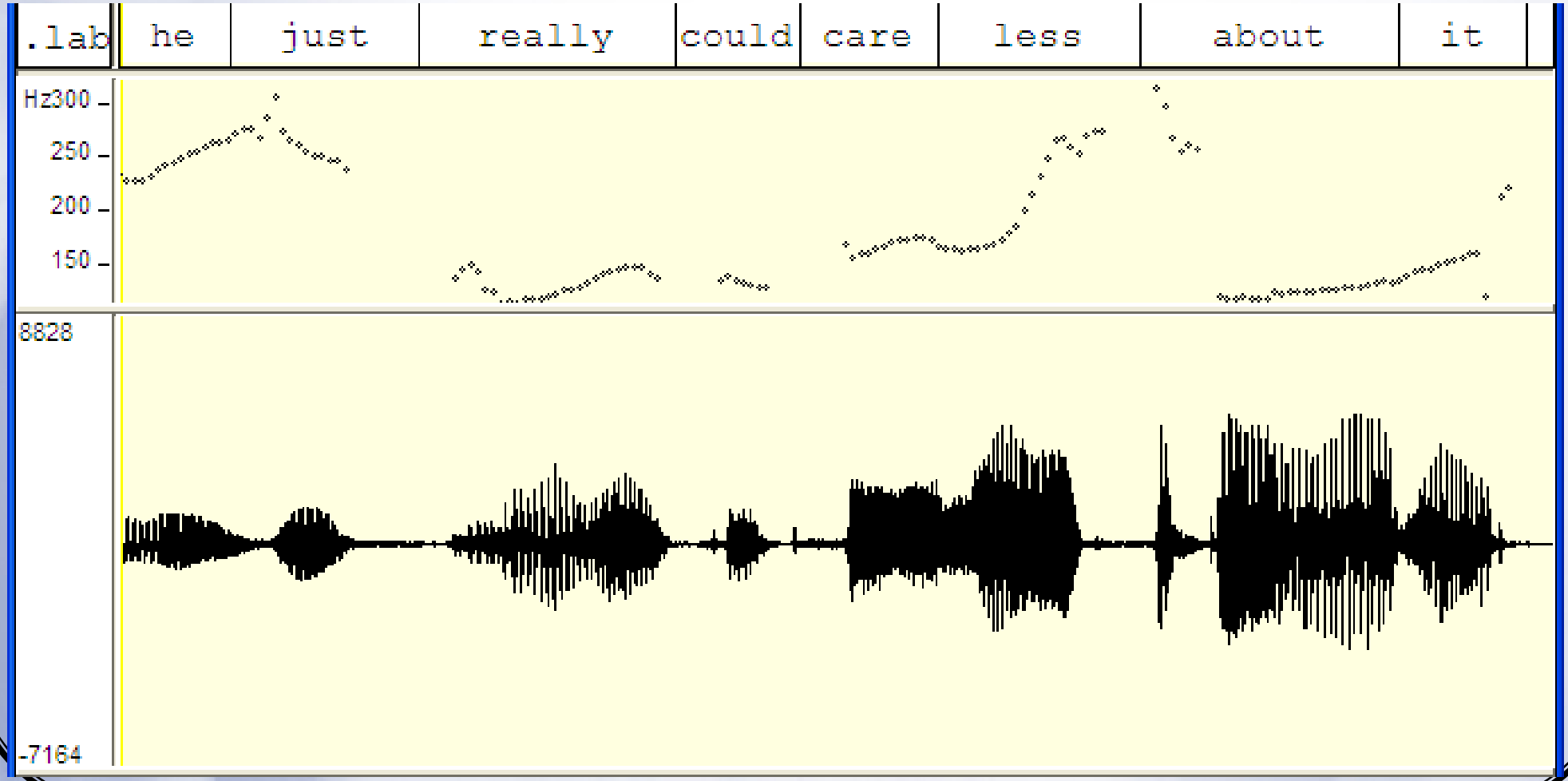
- During the transition from one sound to another, the speech organs prepare to produce new sound while some traces of the old sound will still retain. This is called coarticulation.
- Speaker can be recognized by analyzing the points of spectrograph where coarticulation takes place.

Example: Tongue Twister

顺南边来了个喇嘛，手里提了五斤塌目，顺北边来了个哑巴，腰里别着个喇叭，提了塌目的喇嘛要拿五斤塌目去换北边哑巴腰里别着的喇叭，别着的喇叭的哑巴不愿意拿喇叭去换提了塌目喇嘛他的塌目，提了塌目的喇嘛就急了，拿起了五斤塌目打了别着的喇叭哑巴一塌目，别着的喇叭的哑巴也急了，顺腰里摘下喇叭，打了提了塌目喇嘛一喇叭，也不知道喇嘛的塌目打了别着的喇叭的哑巴一塌目，还是别着的喇叭的哑巴打了提了塌目的喇嘛一喇叭，喇嘛回家炖塌目，哑巴回家吹喇叭。

Pitch Contours

- The variations of the pitch (fundamental frequency) during the period of utterance give a “contour” that can be used as a feature for speaker recognition



Features derived from Linear Prediction

- Linear prediction is derived from two equations:

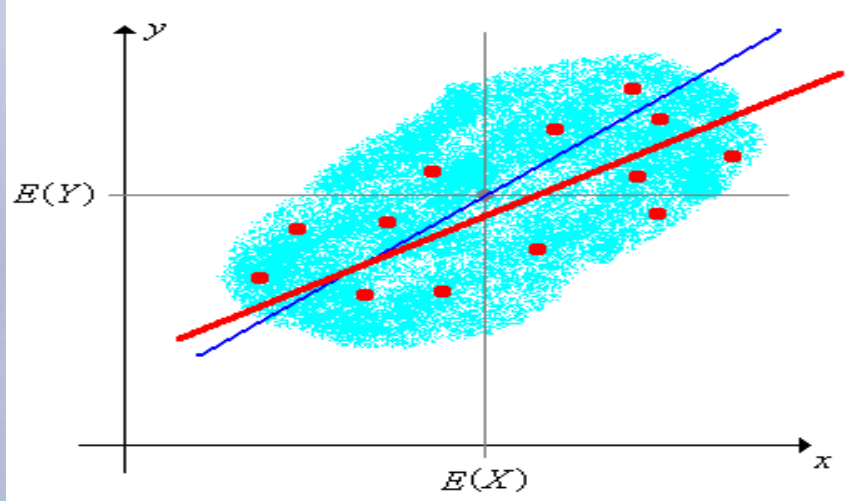
$$\bar{x}_k = \sum_{m=1}^{P} a_m x_{k-m}$$

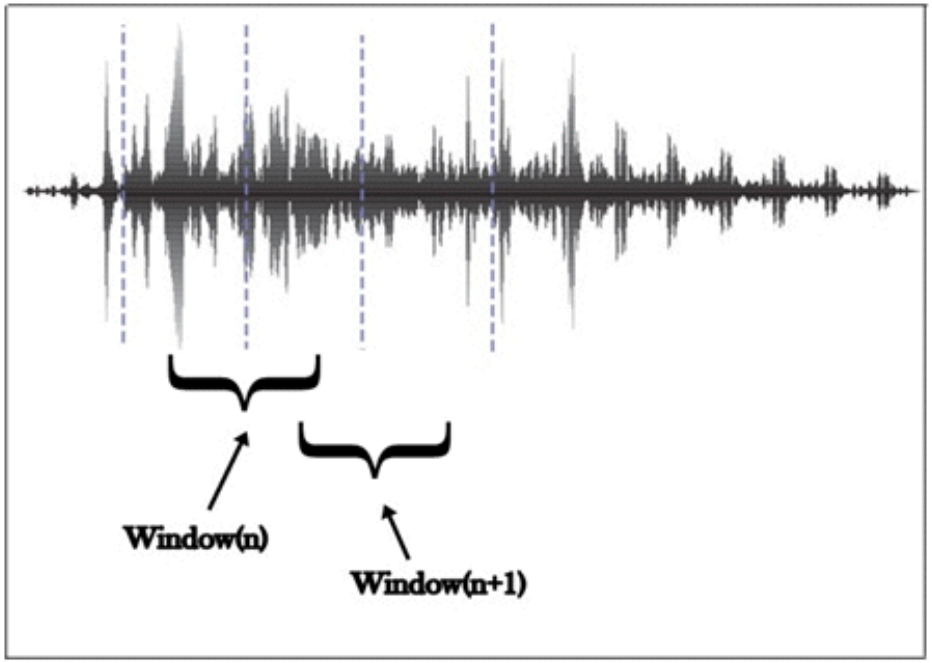
Find the speech sample by linear prediction

$$\varepsilon = \sum_{k=1}^{S} \varepsilon_k = \sum_{k=1}^{S} (x_k - \bar{x}_k)^2$$

Find the prediction error

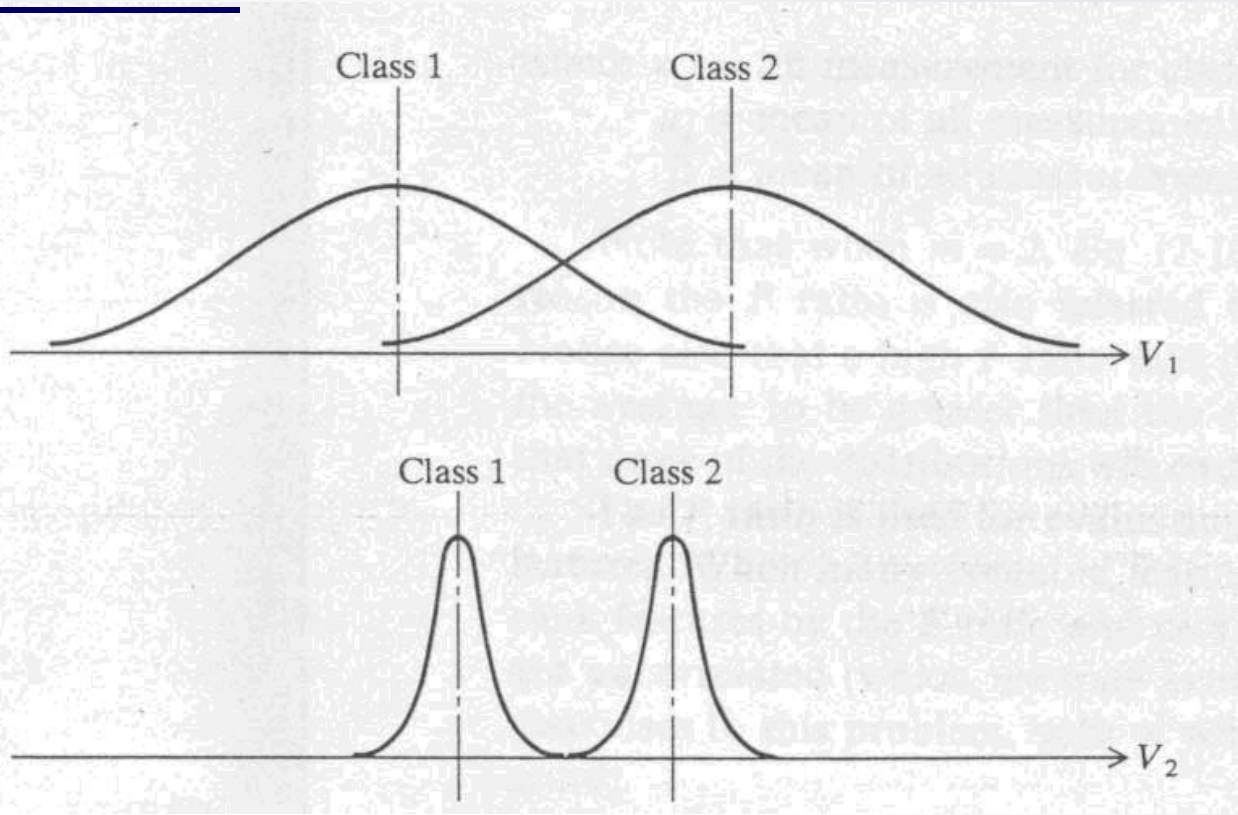
- The two equations generate the linear prediction coefficients for each speech, which can be used to determine the differences between two or more speeches.





Stage 4: Evaluation

- After selection of the features, we need to see how well they separate the different classes
  - Measure the performance of the features



- One of the function that evaluate features is called F-ratio:

$$F = \frac{\text{variance of the means (over all classes)}}{\text{mean of the variances (within classes)}}$$

- For speaker verification, high F-ratios are desirable

Challenges: Comparison & Decision

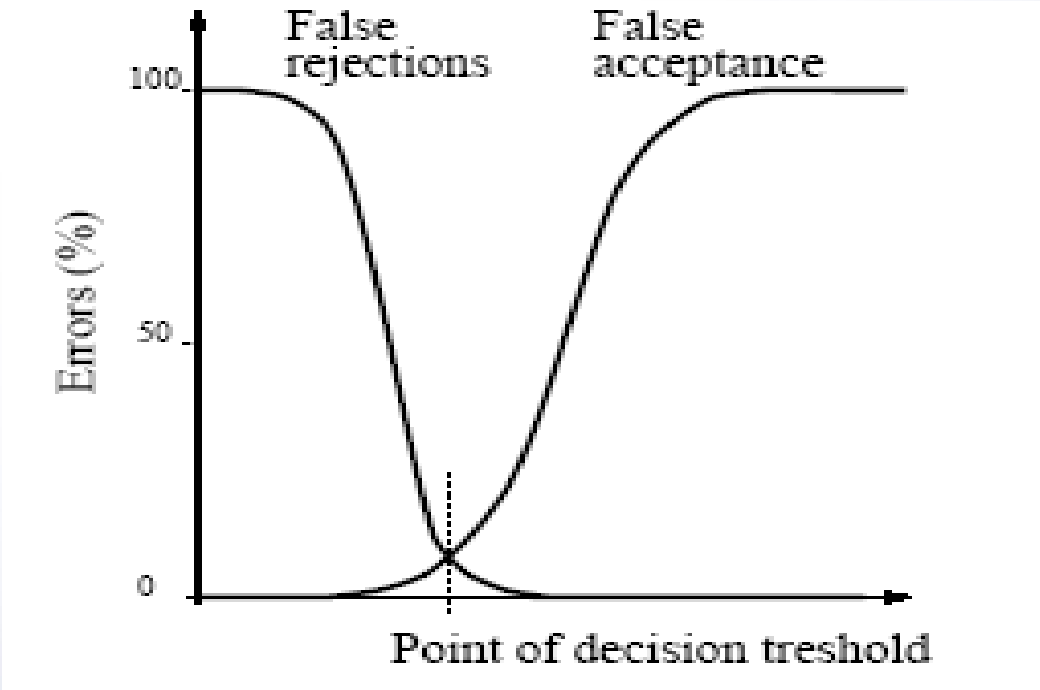
- **Intra-class variability**
  - ***A person may sound differently***

    Methods: Trains more an individual voice, remembers the key characteristics of this person

- **Inter-class similarity**
  - ***Two people may have similar tones, pitch or frequency***

    Method: all features matching others are removed, the remained features are the unique characteristics of a person

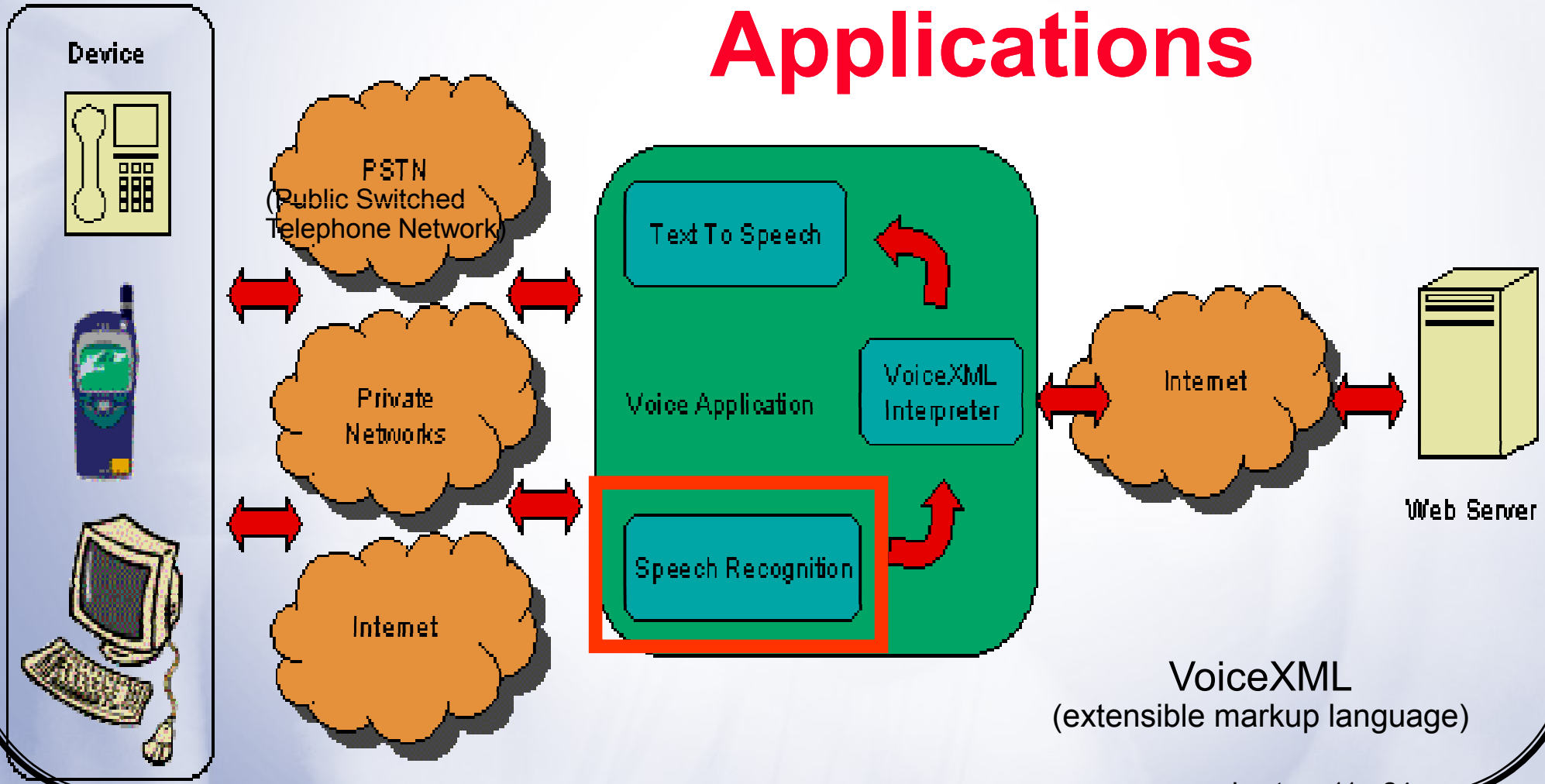
- Matching decision is controlled by a threshold of reliability or acceptance
- Each application should set its specific acceptable and rejection level



Challenges for Voice Verification / Recognition

- Background Noise
- Quality of Input Device
- Channel Noise
- Extreme Hoarseness

Applications

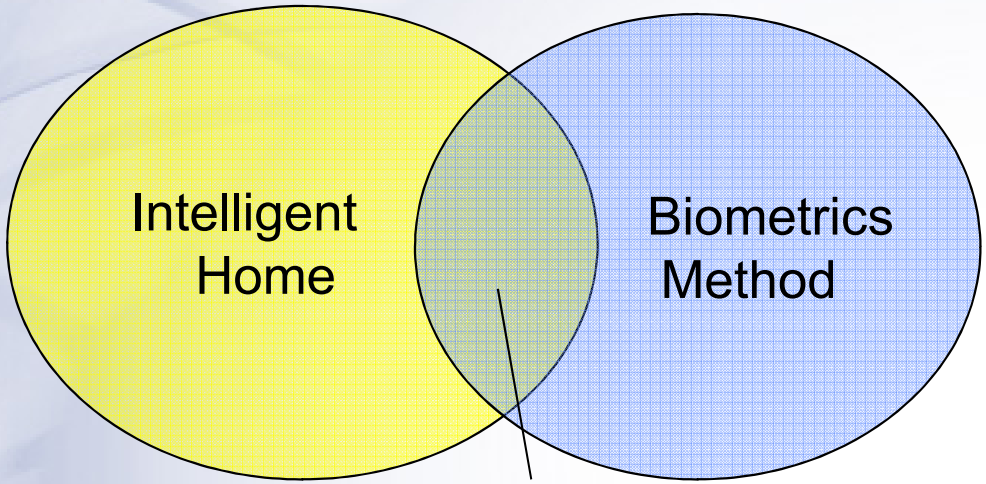


VoiceXML
(extensible markup language)

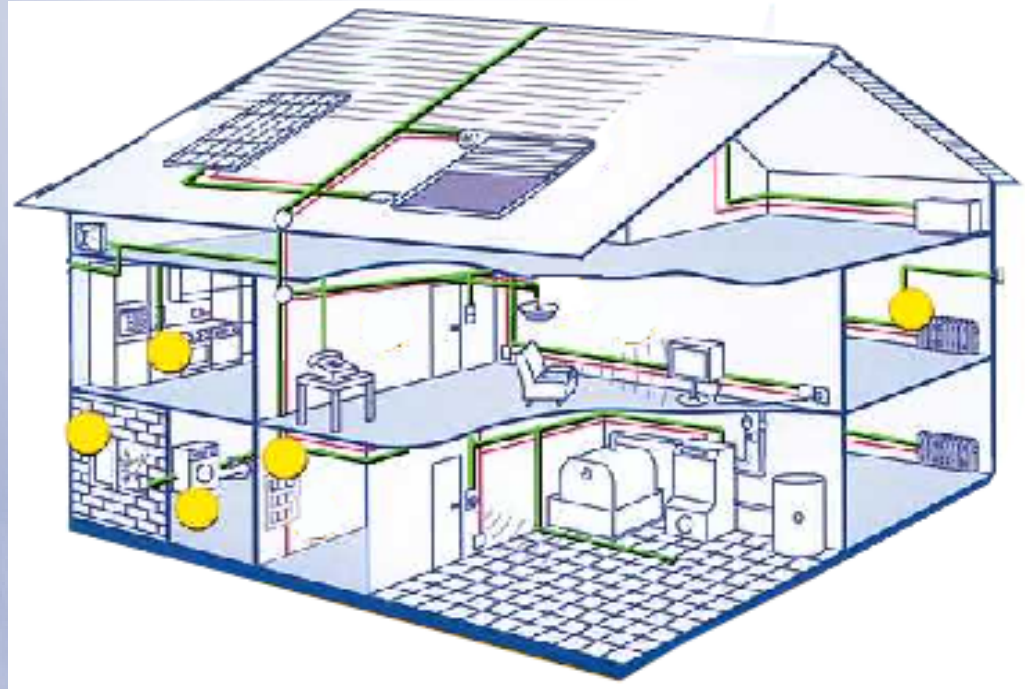
Common Applications

- Call Center Automation
  - *Widely used in all industries (consumer interface)*
    - Airline companies: booking flights, general info, etc.
    - Banking companies: “pay by phone”, account balances, etc.
    - Delivery Services (FedEx): tracking orders, etc.
    - All general customer service systems
- Computer Integration of voice recognition
  - *Personal Computers*
    - Accessibility purposes: voice control of computers
- Integrated into automobiles:
  - Visteon Voice Technology™ used in Infiniti Q45
  - Controls of: Climate; CD player; Navigation system

Some Applications



Remote Intelligent Home



Tomorrow's cell phones will recognize their users' voices.

Voice Summary

Strengths

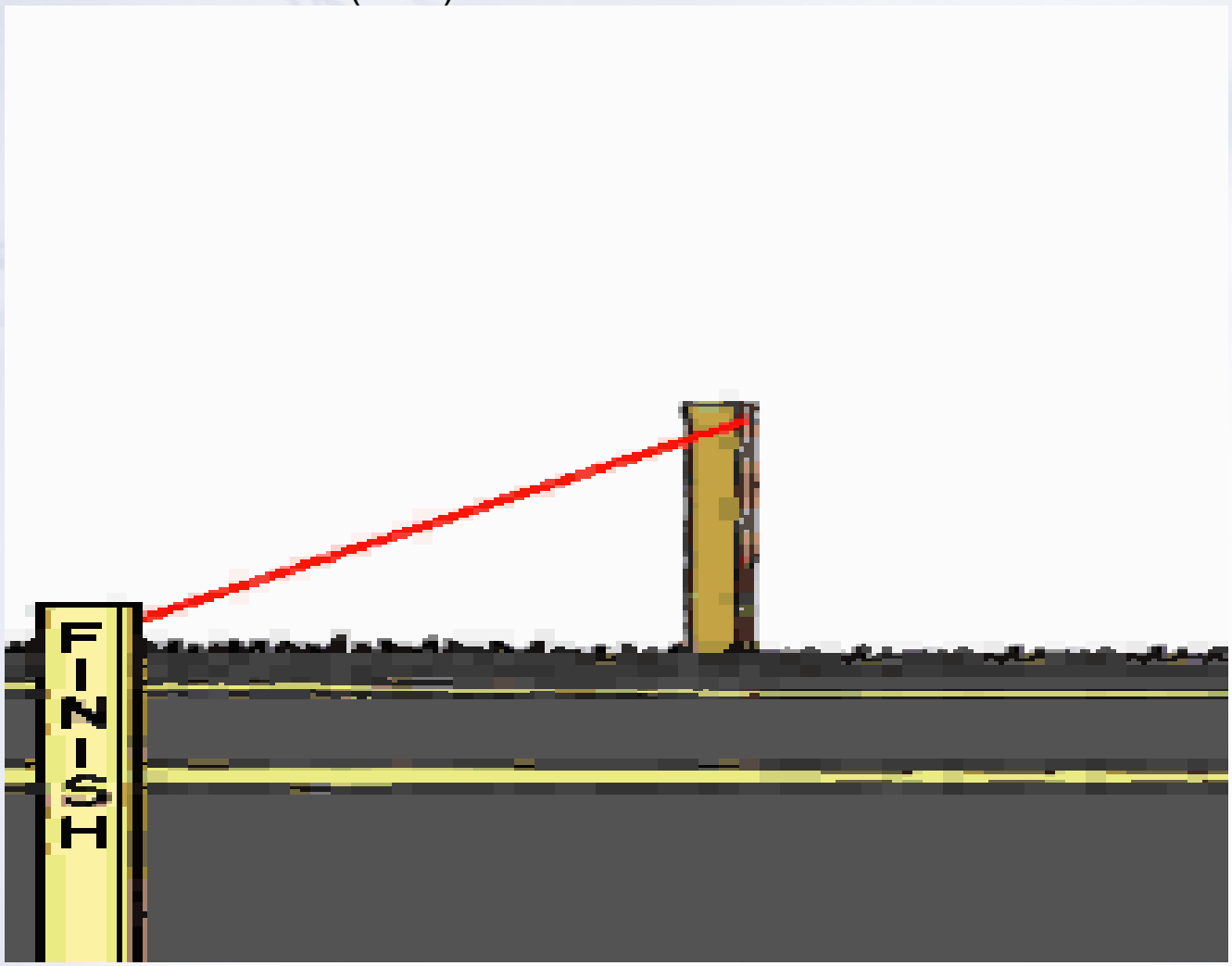
- Unlimited data collection
  - In contrary to fingerprints, face, ear, etc
- Collectable, user friendly (unobtrusive)
  - Well accepted by users
- Economical
  - Cheap equipment

- Widely used
  - Deployed on existing telephony system
- Location dependent

Weaknesses

- Affected by external environment
  - Noisy environment
  - Health condition of users (e.g. heavy colds)

- Degradation of voice quality
  - Through microphone, digitizers, communication channels
- Behavioral nature of voice
  - Affected by stress, fatigue, tempo of the speaker



END