

# Hearing and Speech Processing

---

SAMARTH GUPTA

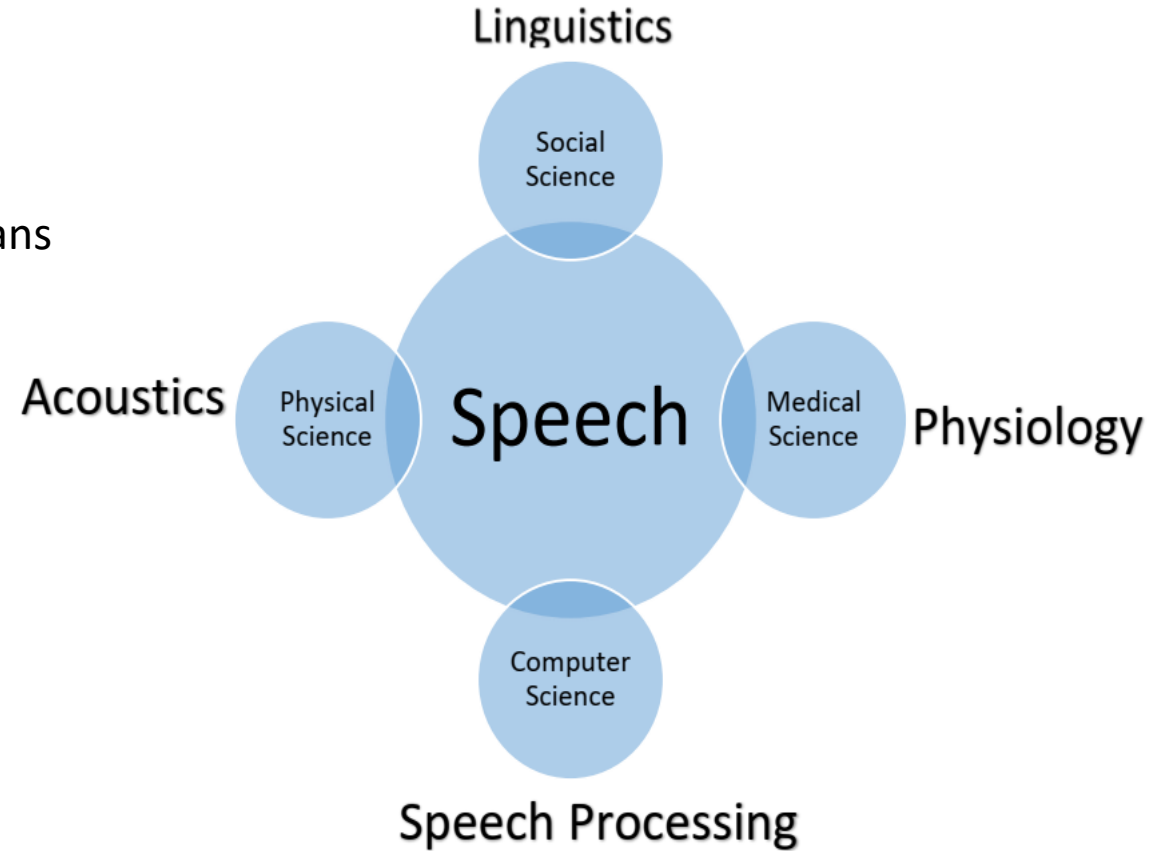
B.TECH. CSE IIND YEAR

ROLL NO.: 1910110338

# Speech

---

- Speech is the most natural form of human-human communications.
- Speech is one of the most intriguing signals that humans work with every day.



# Speech Processing Applications

---

- Automatic translation
- Vehicle navigation systems
- Human computer Interaction
- Content-based spoken audio search
- Home automation
- Pronunciation evaluation
- Robotics
- Video games
- Transcription of speech into mobile text messages
- People with disabilities
- Speech synthesis (Text-to speech)
- Speaker recognition (recognizing who is speaking)
- Speech understanding and vocal dialog
- Speech coding (data rate deduction)
- Speech enhancement (Noise reduction)
- Speech transmission (noise free communication)
- Voice conversion
- Speech recognition (recognizing lexical content)

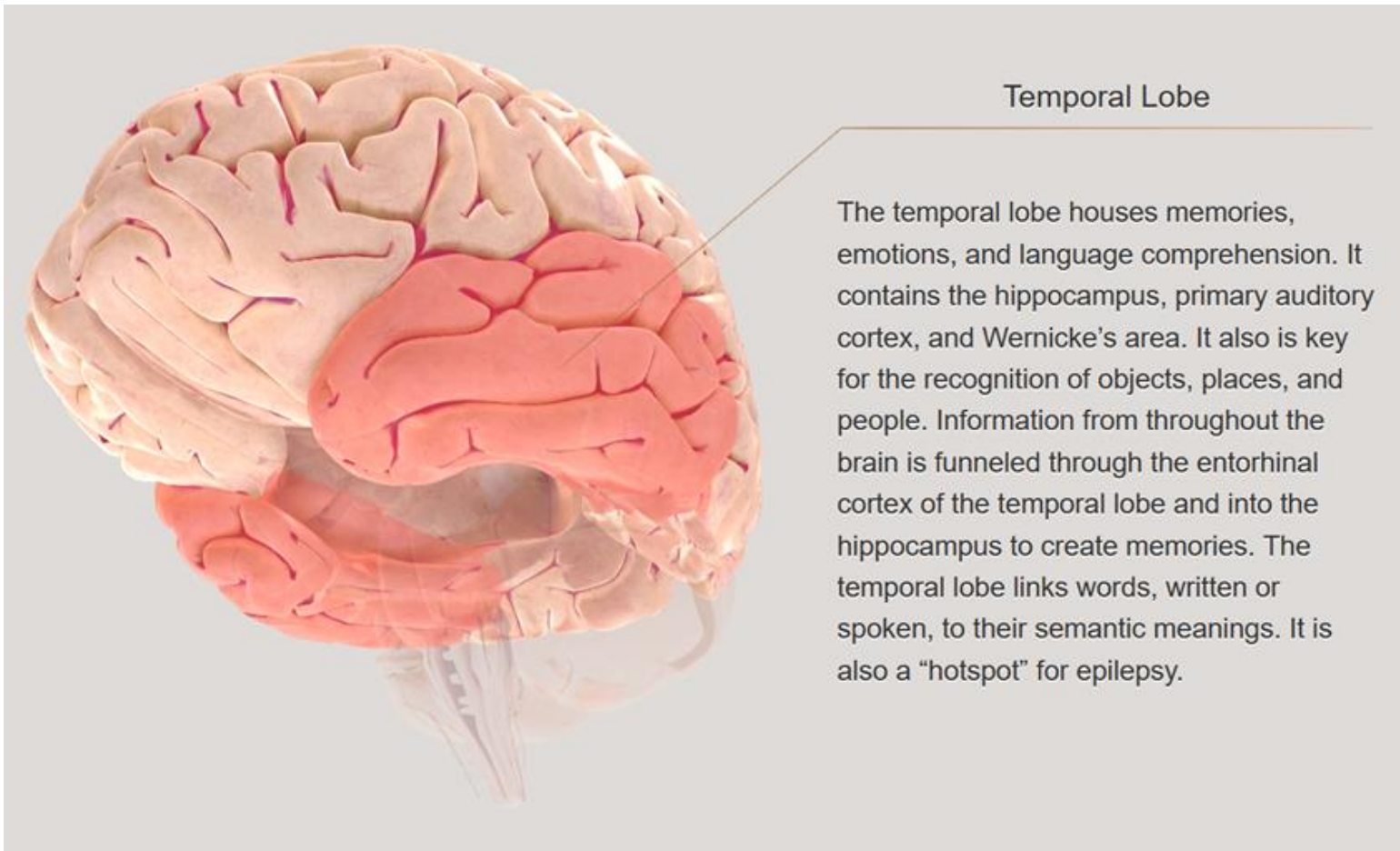
# Physiological perspective: Hearing (Ear)

---

In humans and other vertebrates, **hearing is performed** primarily by the auditory system: mechanical waves, known as vibrations, are detected by the **ear** and transduced into nerve impulses that are perceived by the brain (primarily in the temporal lobe).

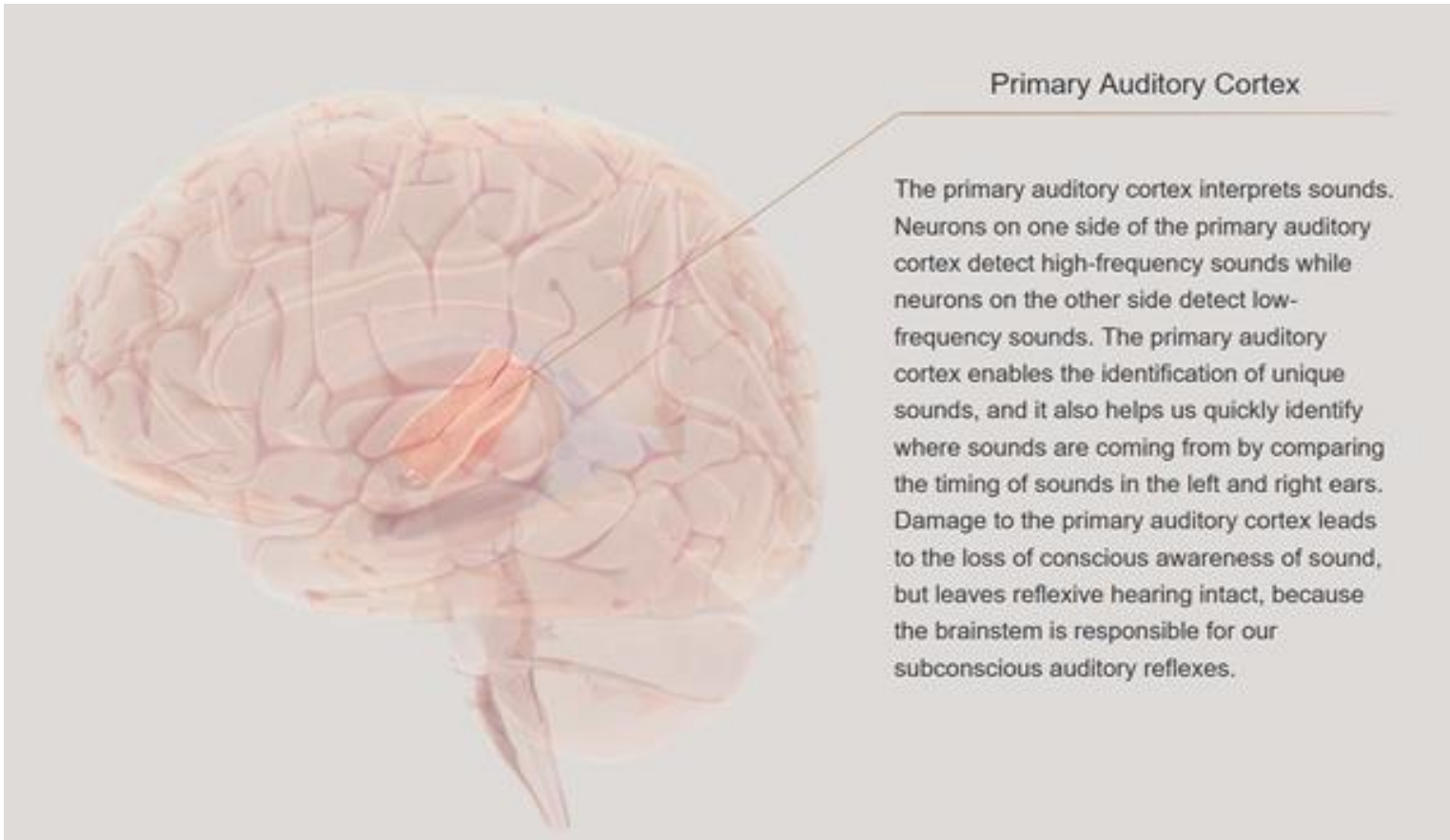
# Temporal Lobe

---



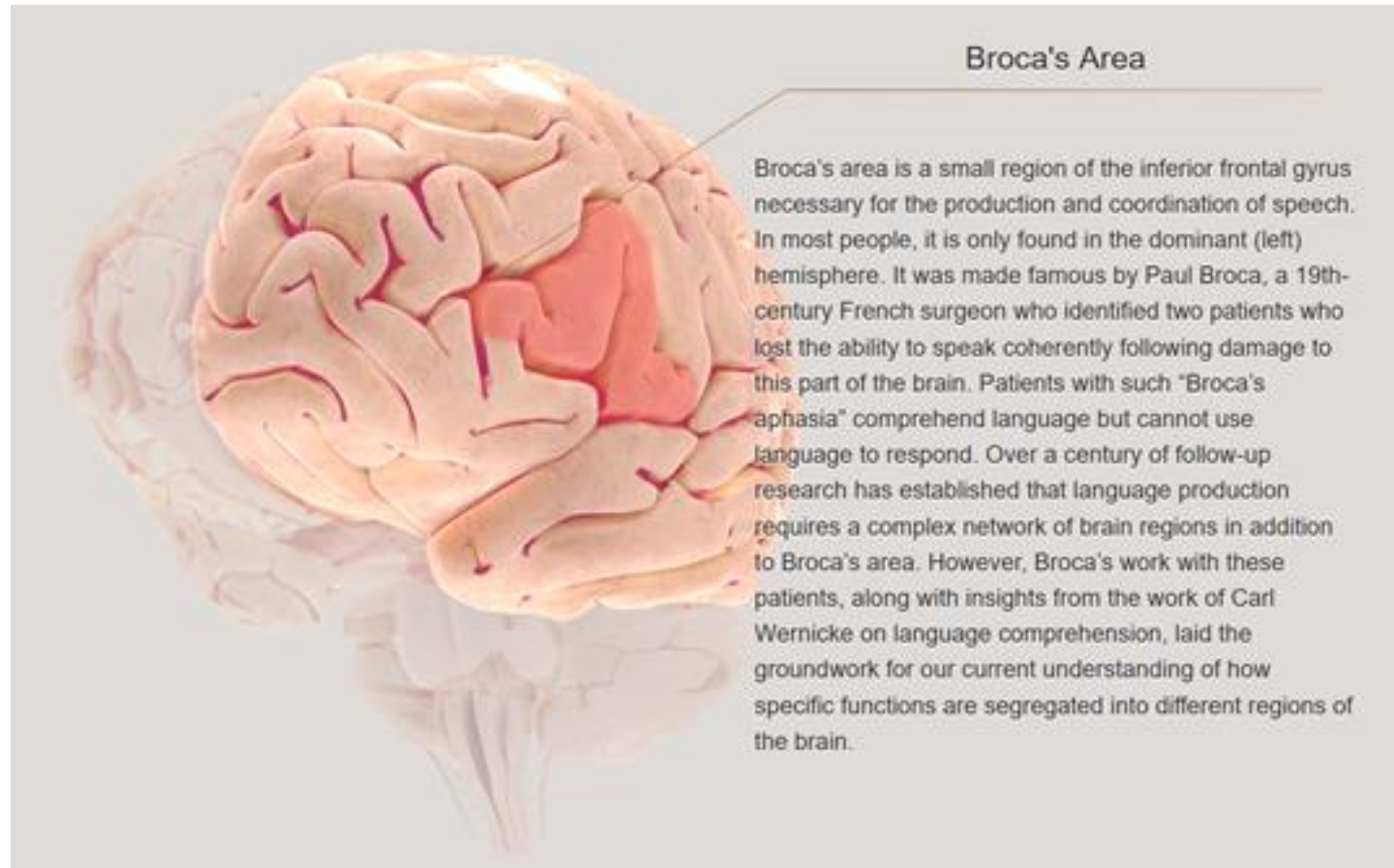
# Primary Auditory Cortex

---



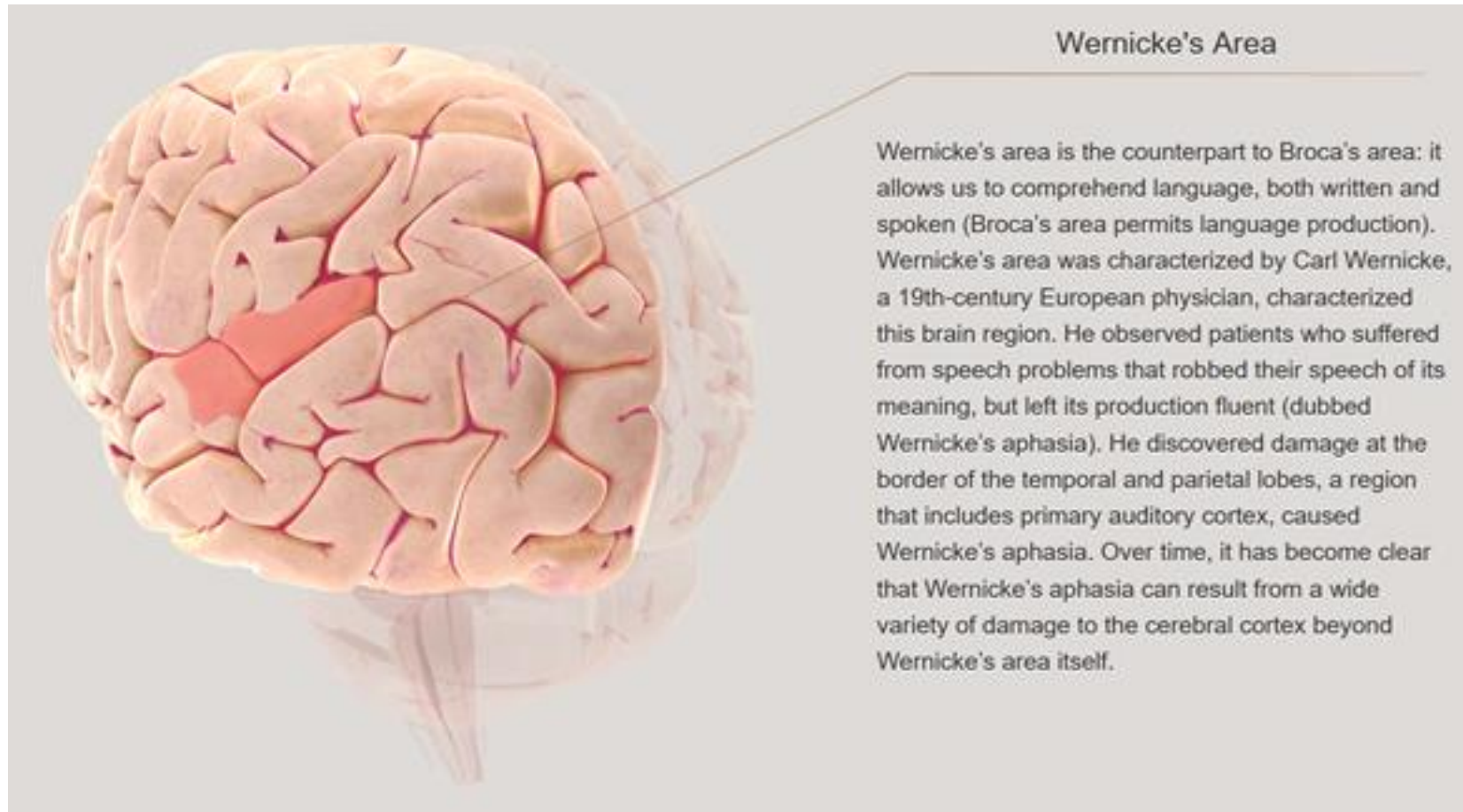
# Broca's Area

---





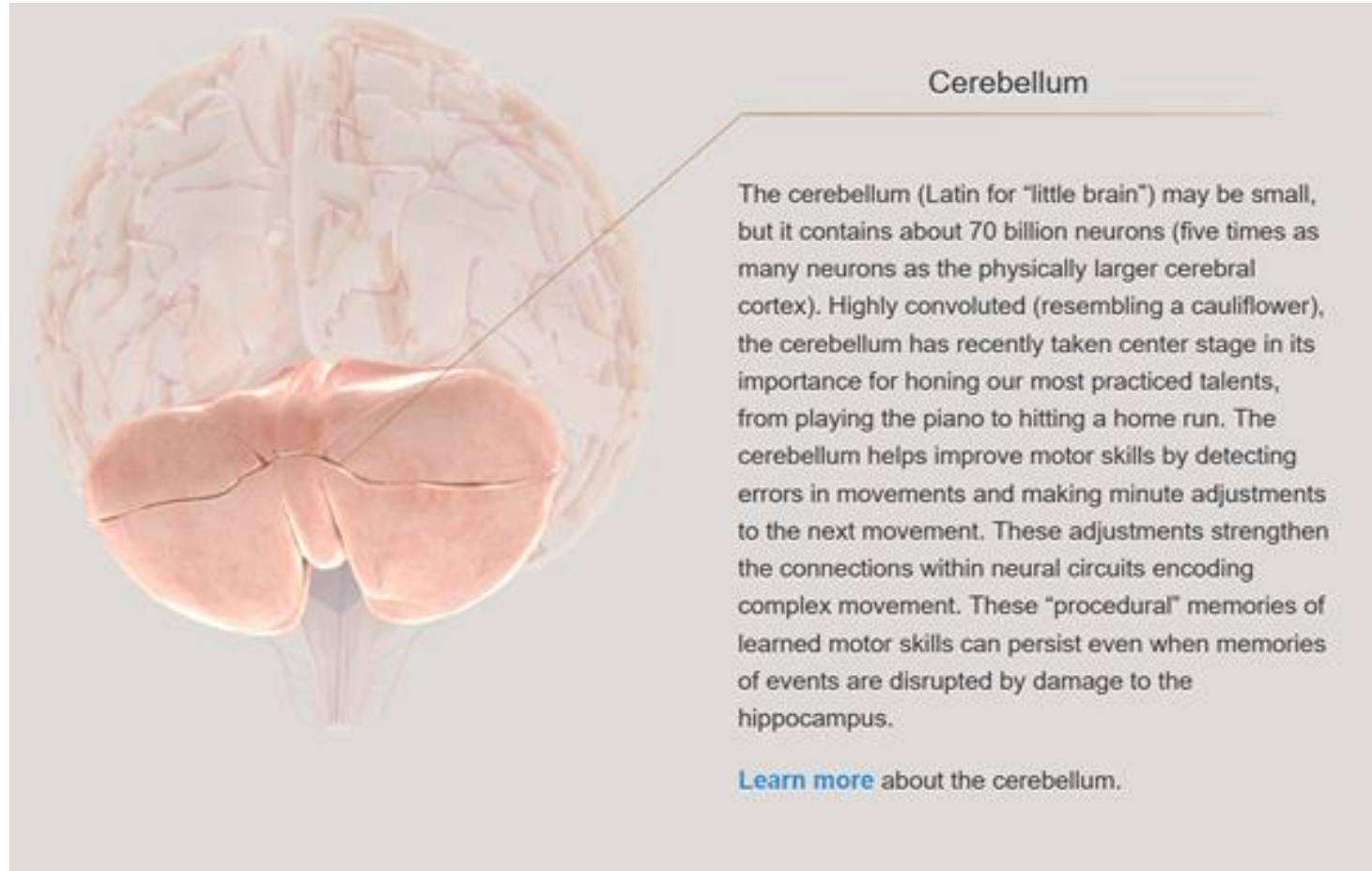
# Wernicke's area



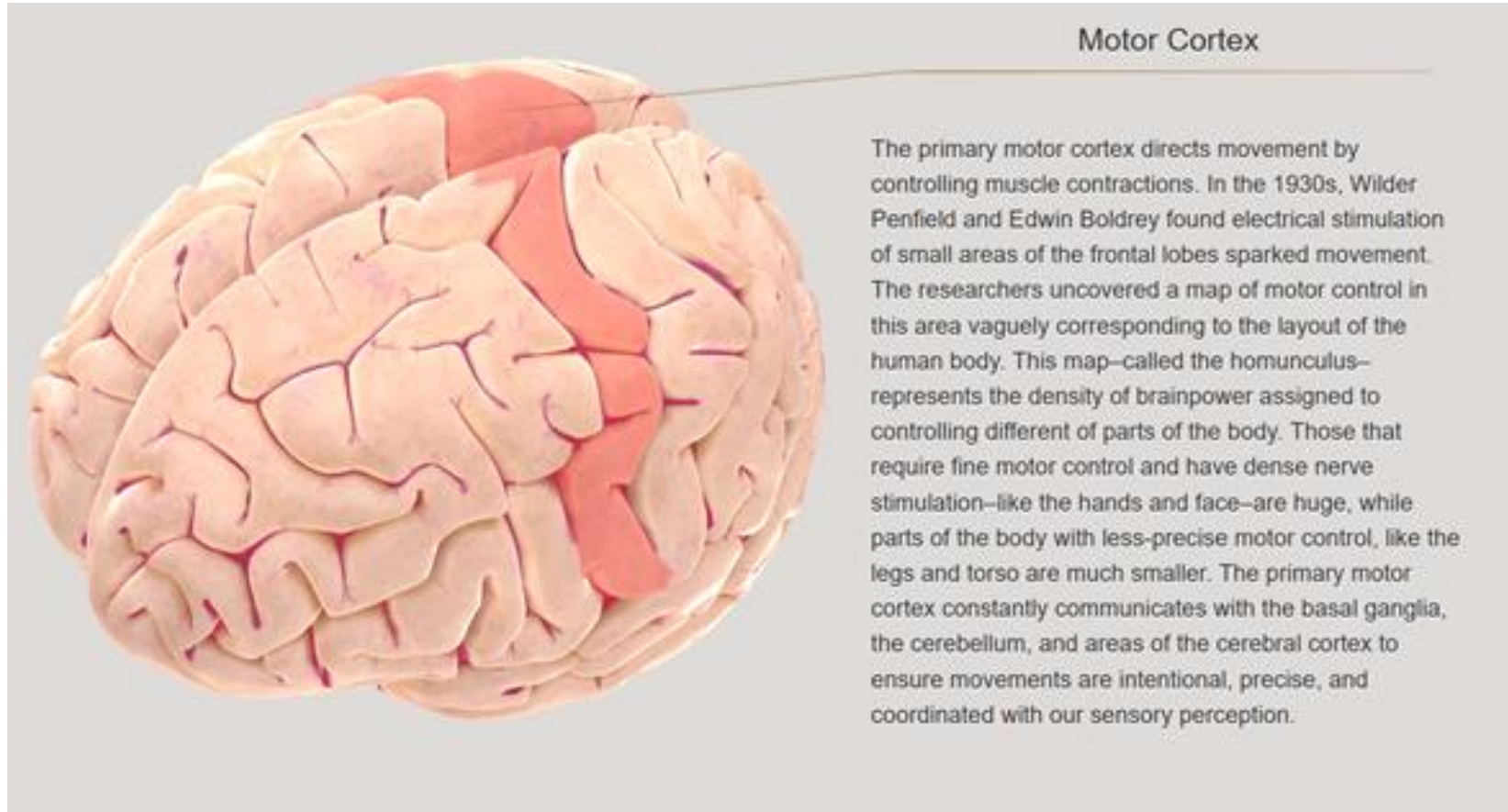


# Cerebellum

---



# Motor cortex



# Neurolinguistics Models

---

**Wernicke-Lichtheim-Geschwind model:** Based on research conducted on brain-damaged individuals, with language related disorders.

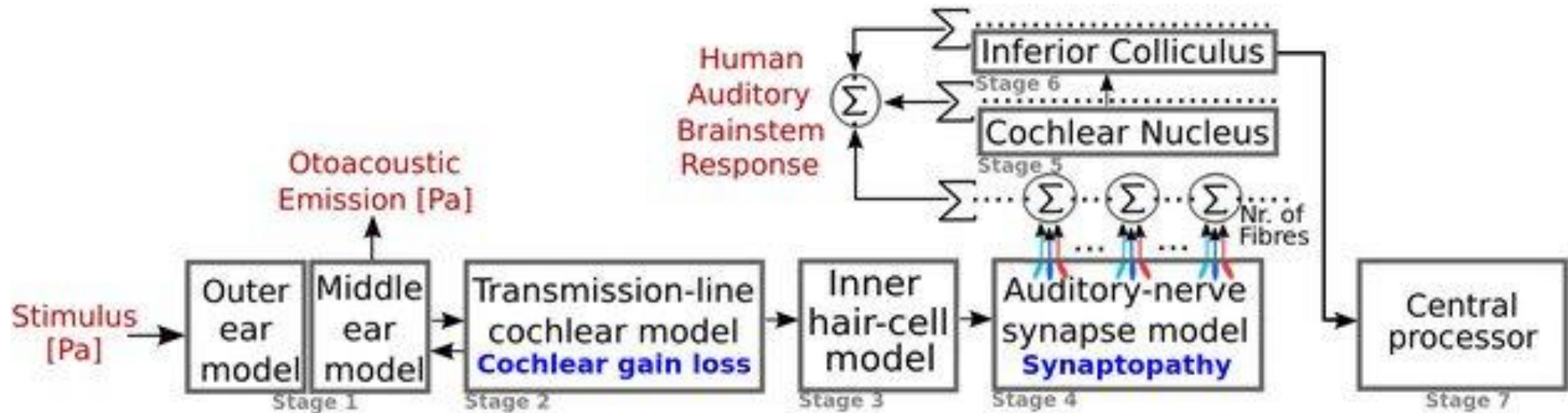
- ❖ Words are perceived via a specialized word reception center (**Wernicke's area**) that is located in the left **temporoparietal junction**.
- ❖ This region then projects to a word production center (**Broca's area**) that is located in the left **inferior frontal gyrus**.
- ❖ It became extremely difficult to identify the basic properties of each region, because:
  - All language input was thought to funnel via Wernicke's area
  - All language output to funnel via Broca's area

# Wernicke-Lichtheim-Geschwind model

---

- ❖ Lack of clear definition for the contribution of Wernicke's and Broca's regions to human language.
  - Became extremely difficult to identify their homologues in other primates.
- ❖ Advent of the fMRI, its application for lesion mappings
  - Showed that this model is based on incorrect correlations between symptoms and lesions.
- ❖ Door opened to new models of language processing in the brain.

# Wernicke-Lichtheim-Geschwind model



# Auditory Ventral Stream

---

- ❖ The auditory ventral stream (AVS)
  - Connects **auditory cortex** with the **middle temporal gyrus** and **temporal pole**
  - Which in turn connects with the **inferior frontal gyrus**.
  - Responsible for sound recognition
  - Known as the auditory 'what' pathway.

# Auditory Ventral Stream

---

## ❖ Sound recognition:

- AVS is involved in recognizing auditory objects.
- At the level of the primary auditory cortex, recordings from monkeys showed higher percentage of neurons selective for learned melodic sequences in area R than area A1
- Study in humans demonstrated more selectivity for heard syllables in the anterior Heschl's gyrus (area hR) than posterior Heschl's gyrus (area hA1).



# Computational Perspective: Speech Processing

---

Areas of interest in Speech Processing:-

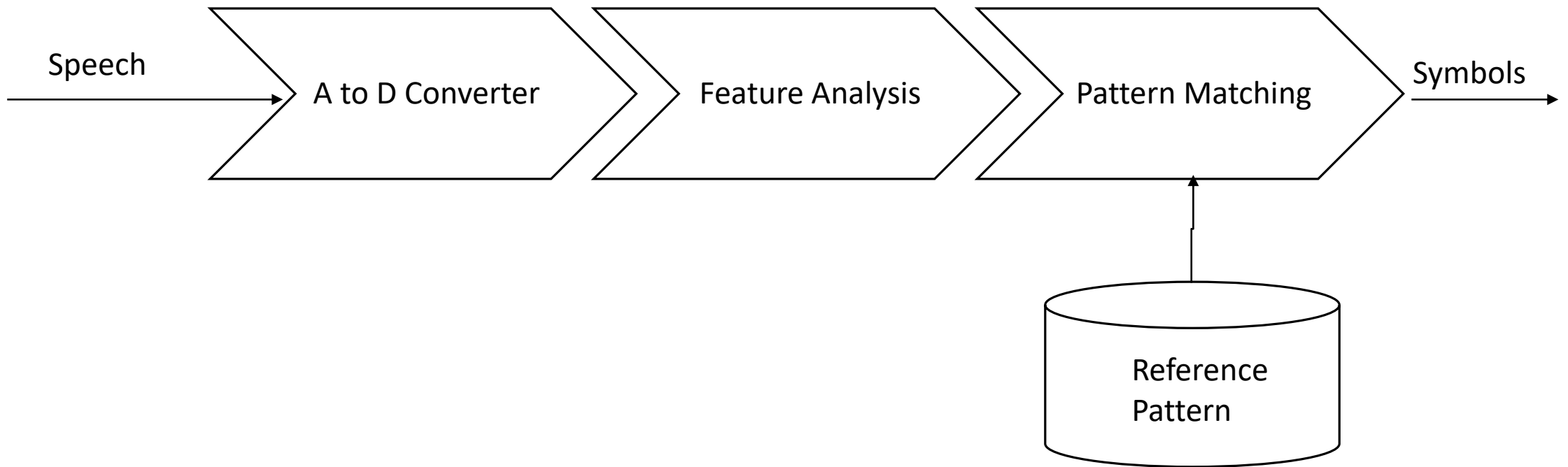
- Speech coding
- Speech Synthesis
- TTS/STT
- Speaker verification
- Speaker Diarization
- Speech Recognition
- Word Spotting
- Automatic indexing of speech recordings

**Speaker Recognition** - *Speaker recognition* is the process of automatically recognizing who is speaking on the basis of individual information included in speech signals. It has the following types:-

- Closed set
- Open set
- Text independent
- Text dependent

# Speaker recognition: Pattern matching

---



# Speech Representation: Models

---

- Temporal features
- Low energy rate
- Zero crossing rate (ZCR)
- 4Hz modulation energy
- Pitch contour
- Spectral features
- Spectral Centroid (sharpness)
- Speech Processing
- Spectral Flux (rate of change)
- Spectral Roll-Off (spectral shape)
- Spectral Flatness (deviation of the spectral form)
- Linear Predictive Coefficients (LPC)
- Cepstral coefficients
- **Mel Frequency Cepstral Coefficients (MFCC): human auditory system**
- Harmonic features: sinusoidal harmonic modelling
- Perceptual features: model of the human hearing process
- First order derivative (DELTA)

# Major Problems in Speech Processing

---

- ❖ **Acoustic variability:** The same phonemes pronounced in different contexts will have different acoustic realization (coarticulation effect)
- ❖ **The signal is different when speech is uttered in various environments:**
  - Noise
  - Reverberation
  - different types of microphones.
  - Speaking variability: when the same speaker speaks normally, shouts, whispers, uses a creaky voice, has a cold, or **is wearing a mask**
  - Speaker variability: Different speakers have different timbers and different speaking habits

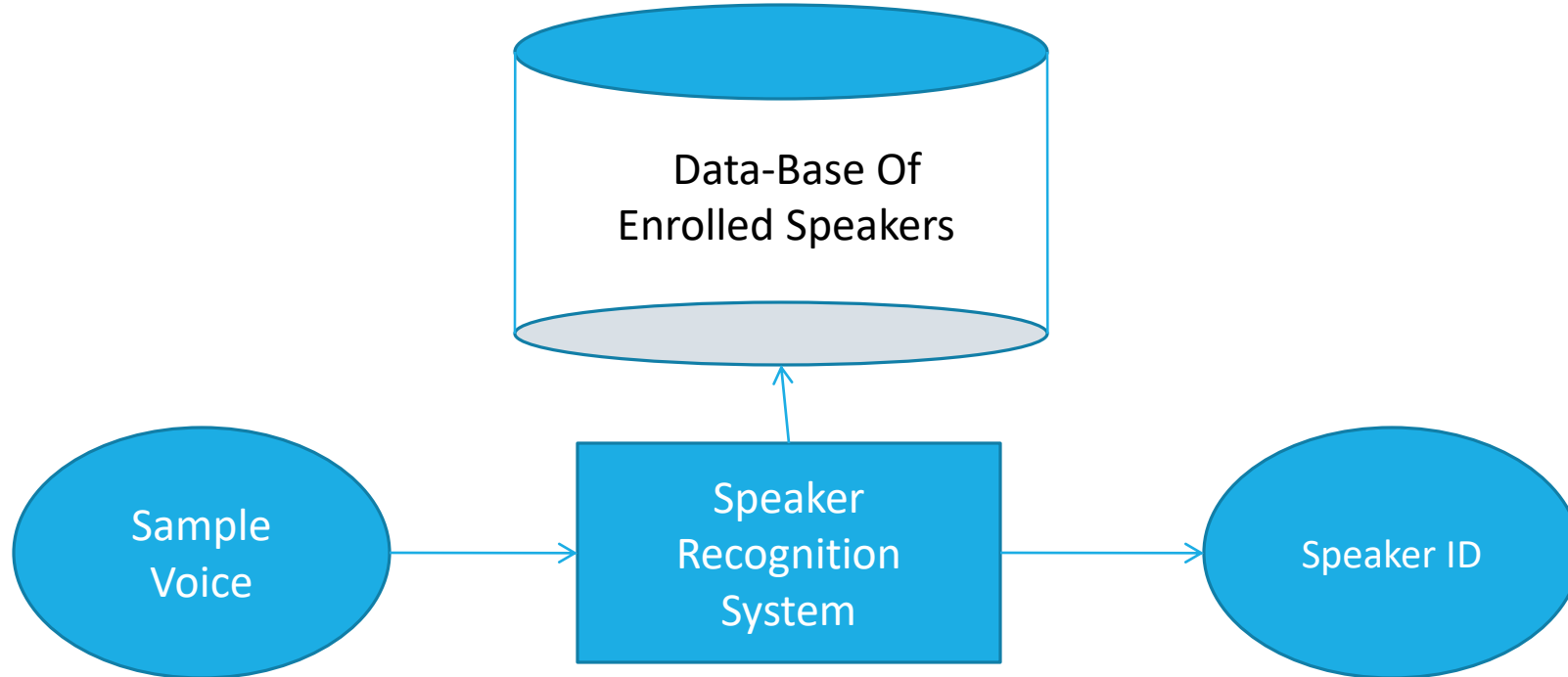
# Major Problems in Speech Processing

---

- ❖ **Linguistic variability:** the same sentence can be pronounced in many different ways, using many different words, synonyms, and many different syntactic structures and prosodic schemes
- ❖ **Phonetic variability:** due to the different possible pronunciations of the same words by speakers having different regional accents
- ❖ **Lombard effect:** Noise modifies the utterance of the words (as people tend to speak louder)
- ❖ **Continuous speech:** Words are connected together (not separated by pauses or silences)
  - It is difficult to find the start and end points of words
  - The production of each phoneme is affected by the production of surrounding phonemes
  - The start and end of words are affected by the preceding and following words
  - The rate of speech (fast speech tends to be harder)

# Speaker Recognition System

Captures speech signal and identifies who the speaker is out of a set of known speakers



# Closed Set of Speakers

---

Test speaker **definitely** belongs to the training samples.

## **Example:**

Speaker set for training phase:  $S = \{s_1, s_2, s_3, s_4, \dots, s_{20}\}$

Speaker for testing phase: from  $S$



# Text Independent System

---

Characterized by:

The training and the test utterances need not match.

Example:

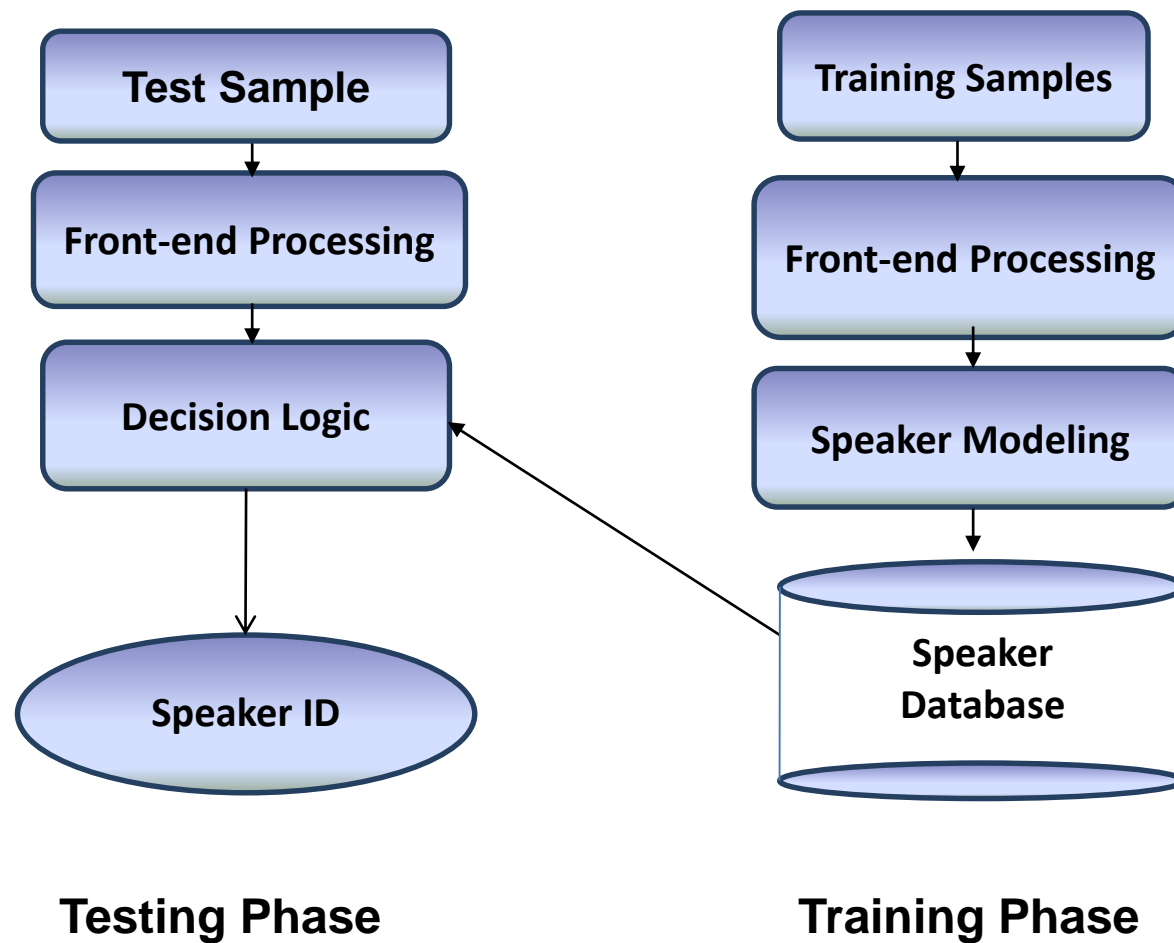
Text during training phase:

"Hello, How are you ?"

Text during testing phase:

"Good Morning, I am fine."

# Block Diagram of Speaker Recognition



# Front-end processing

---

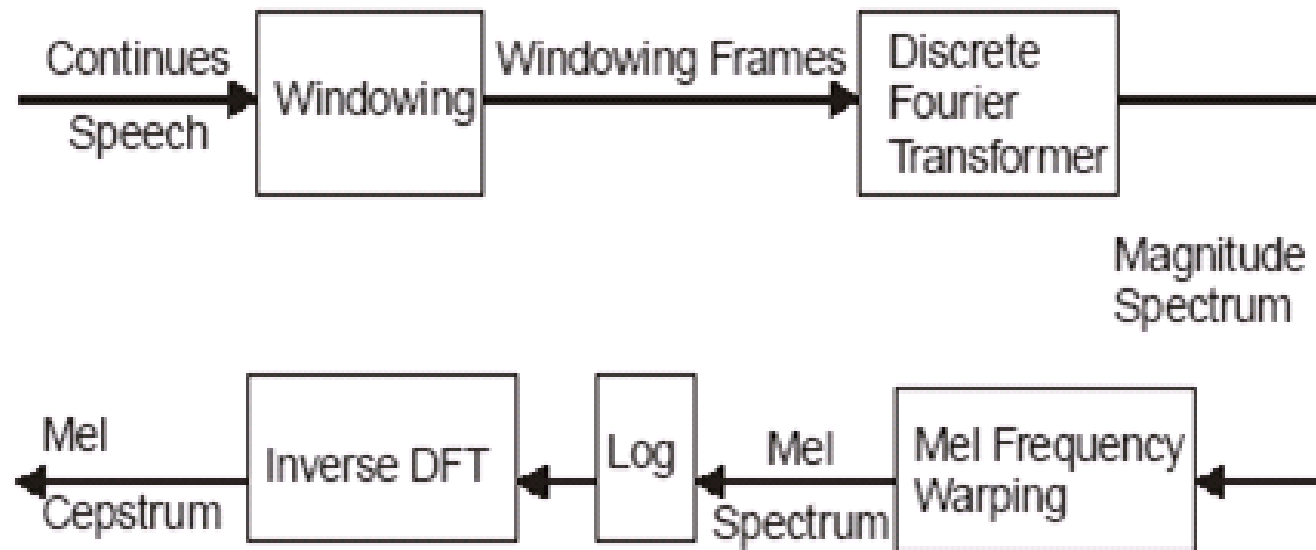
Recording of speech samples

Converts analog speech to digital form by *Sampling (16 KHz)*.

*Feature extraction with MFCC function*

# Complete Pipeline for MFCC

---



# MFCC

---

Spectrum  $\rightarrow$  Mel-Filters  $\rightarrow$  Mel-Spectrum

Say  $\log X[K] = \text{Log}(\text{Mel-Spectrum})$

Now perform Cepstral Analysis on  $\log X[k]$

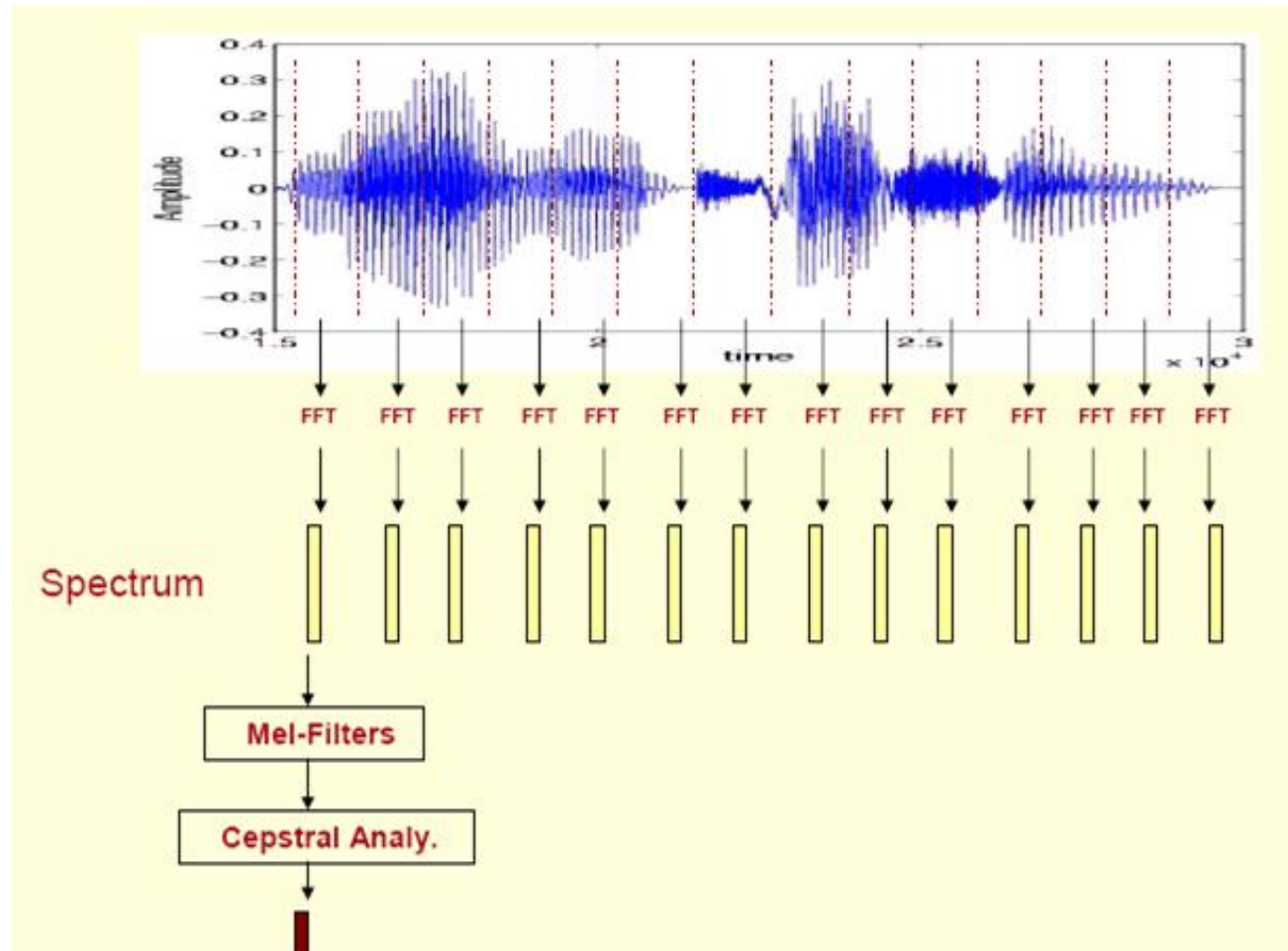
$$\log X[k] = \log H[k] + \log E[k]$$

Taking IFFT

$$x[k] = h[k] + e[k]$$

$h[k]$  represents cepstral coefficient(MFCC)

# Speech Represented as Sequence of Spectral Vectors



# Speaker Modeling

---

*Feature Modeling using following techniques*

(i) Logistic Regression:

(ii) Support Vector Machine (SVM):

(iii) Vector Quantization(VQ):



# Logistic Regression

---

Logistic regression algorithm is a *classification Algorithm*.

Hypothesis

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

is called the *logistic function* or the *sigmoid function*.

# Sigmoid Function

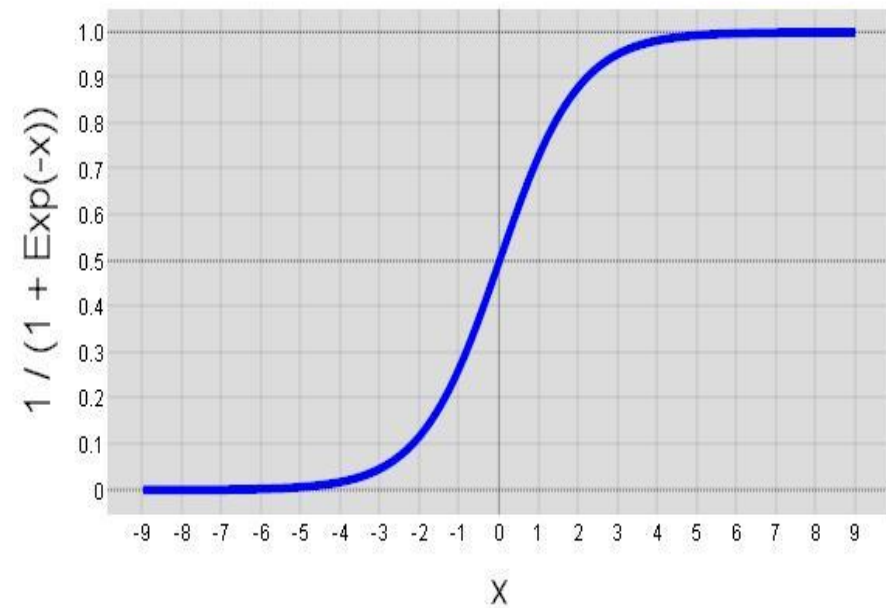
Predict “y=1” if  $\geq 0.5$   $h_{\theta}(x)$

Predict “y=0” if  $< 0.5$   $h_{\theta}(x)$

$g(z) \geq 0.5$  when  $z \geq 0$

$g(z) \geq 0.5$  when  $z \geq 0$

Here  $z = \theta^T(x)$



## Logistic regression cost function

$$\begin{aligned} J(\theta) &= \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)}) \\ &= -\frac{1}{m} \left[ \sum_{i=1}^m y^{(i)} \log h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_{\theta}(x^{(i)})) \right] \end{aligned}$$

To fit parameters  $\theta$ :

$$\min_{\theta} J(\theta)$$

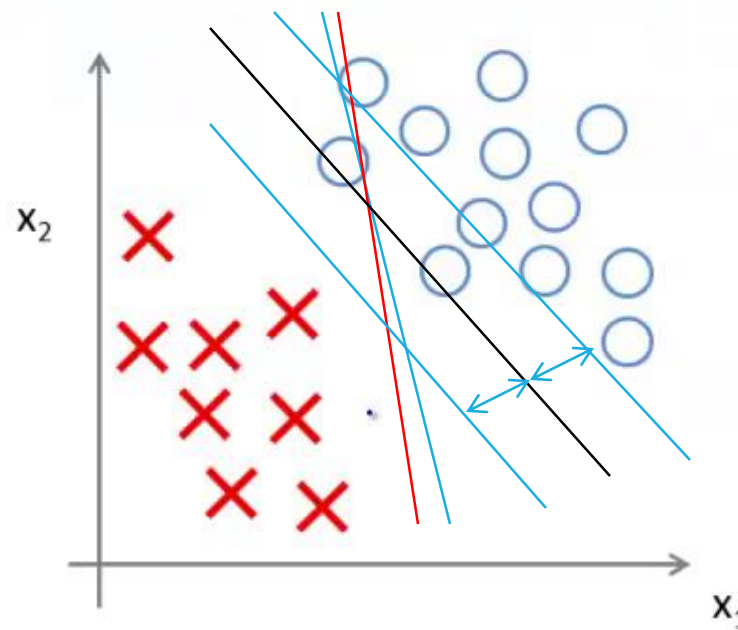
To make a prediction given new  $x$ :

$$\text{Output } h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

# Support Vector Machine

---

## SVM Decision Boundary: Linearly separable case



# Support Vector Machine

---

A support vector machine (SVM) is a supervised learning method and is a powerful way to calculate non-linear functions.

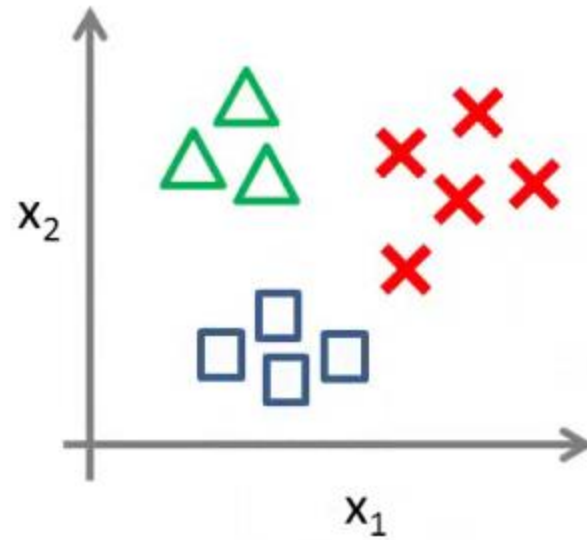
Cost Function



$$\min_{\theta} C \sum_{i=1}^m \left[ y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) \right]$$

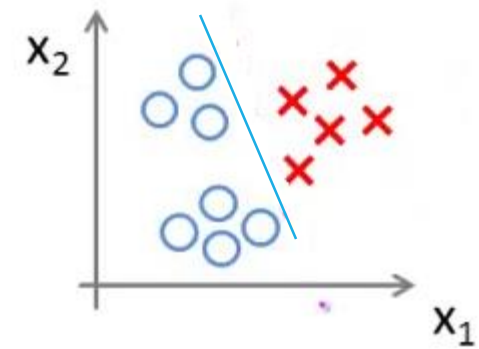
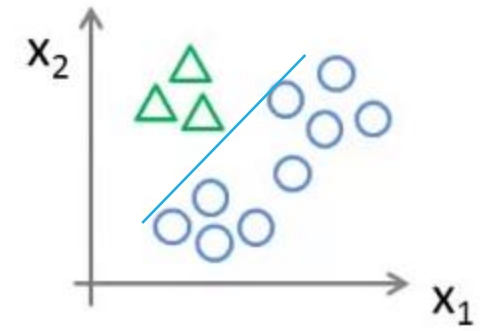
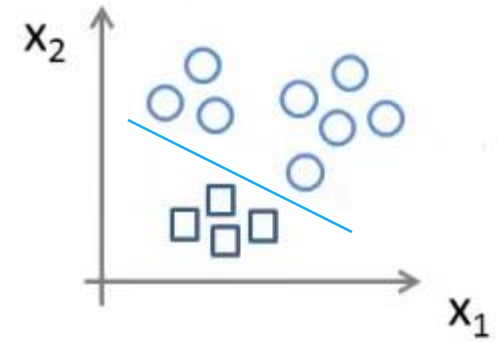
If  $y = 1$ , we want  $\theta^T x \geq 1$  (not just  $\geq 0$ )

If  $y = 0$ , we want  $\theta^T x \leq -1$  (not just  $< 0$ )

## One-vs-all (one-vs-rest):



Class 1:   
Class 2:   
Class 3: 



# Vector Quantization

---

Technique which divides a large set of points into groups having approximately the same number of points closest to them. Each group is represented by its centroid point.

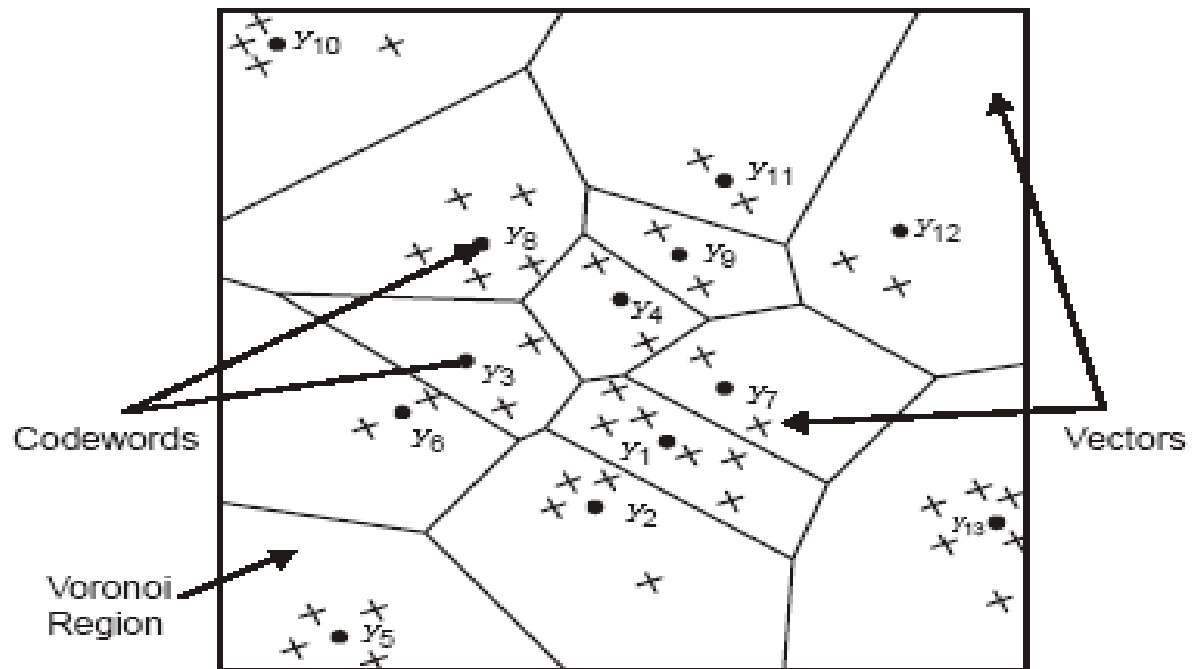
A vector quantizer maps  $k$ -dimensional vectors in the vector space  $R^k$  into a finite set of vectors  $Y = \{y_i; i = 1, 2, \dots, N\}$ .

Each vector  $y_i$  is called a *code vector* or a *codeword*

*Codebook*: set of all the code words



# Codewords in 2-D space



- Input vectors  $\rightarrow x$
- Codewords  $\rightarrow \bullet$
- Voronoi regions  $\rightarrow$  boundary lines

# Result for Closed Set Speaker Recognition Systems using Vector Quantization

---

System Type	Train Data	Test Data	Accuracy
Text Dependent	11 Speakers, speaking “Zero”	Same 11 Speakers, speaking “Zero”	90%
Text Independent	4 Speakers, speaking “No”	Same 4 Speakers, speaking “Yes”	50%
Text Dependent	4 Speakers, speaking “No” and “Yes”	Same 4 Speakers, speaking “No” and “Yes”	90%
Text Dependent	4 Speakers, speaking “Yes” and “No”	Same 4 Speakers, speaking “No” and “Yes”	90%

# Conclusion

---

We have developed a closed-set Speaker recognition system that accepts speech signals as input.

The system shows good accuracy using vector quantization in a Text Dependent environment.

# Future Scope

---

The speaker recognition system may be extended to work with open-set and the use of statistical models like HMMs, GMMs or learning models like neural networks can also be incorporated to make the system much tolerant to variations like accent, speaker wearing masks and extraneous conditions like noise and associated residues and hence make it less error prone.