

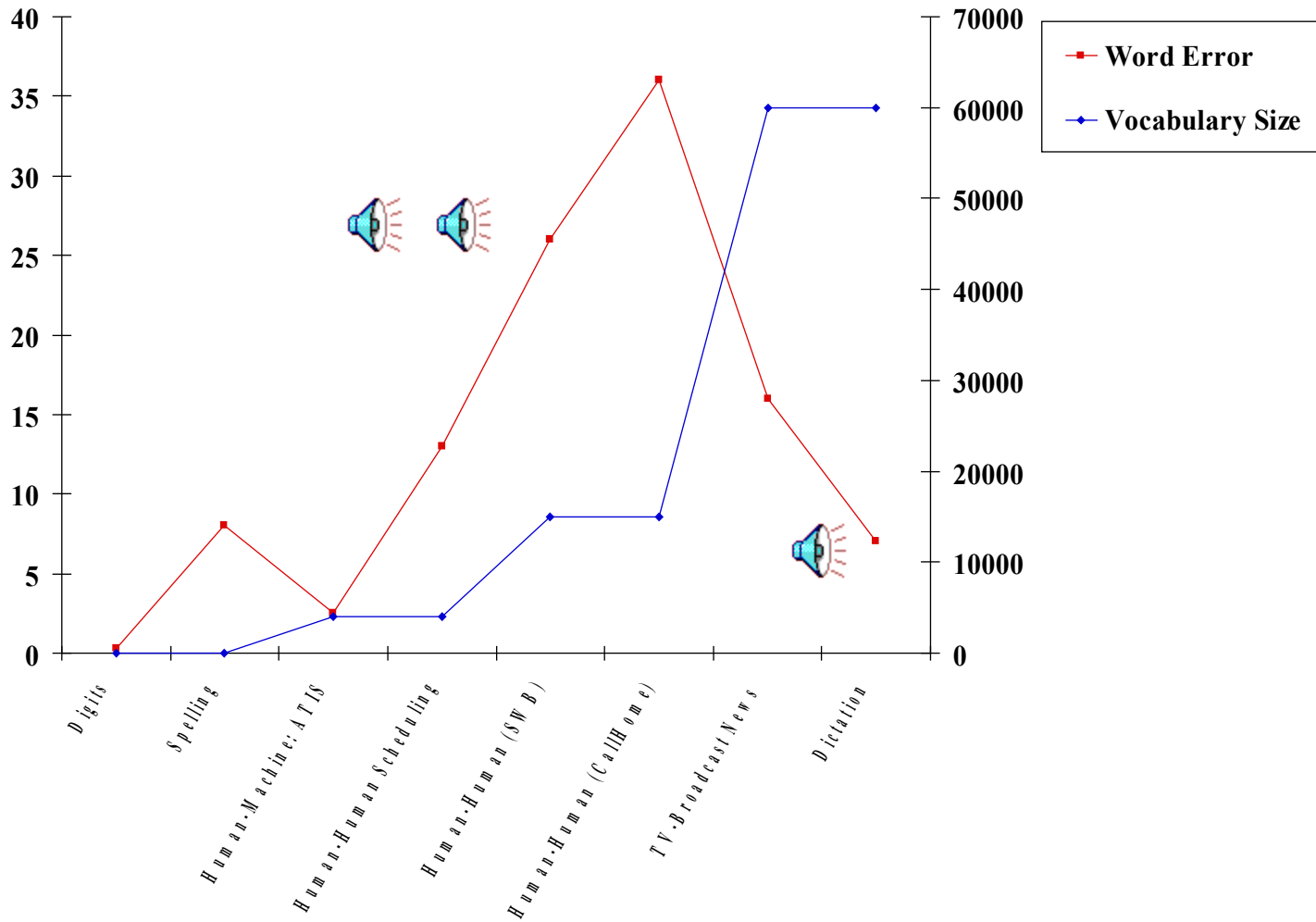
# Speech Recognition

**Alex Waibel**

# Dimensions of Difficulty

- Noise – Environmental, Channel, Reverberation
- Speaker – Male, Female, Children, Elderly
- Acoustic Similarity – Letters, Digits,...
- Vocabulary Size – 10 → 100,000 words
- Speaking Style – Isolated, Continuous Read Speech, Spontaneous, Conversational Speech

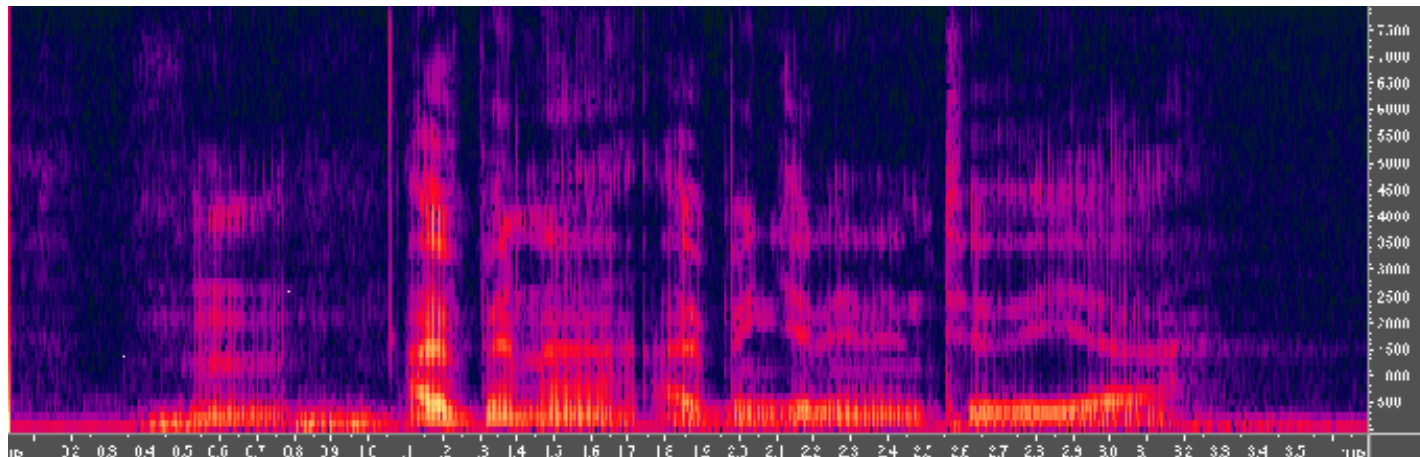
# Speech: State-of-the-Art



# Sloppy Speech

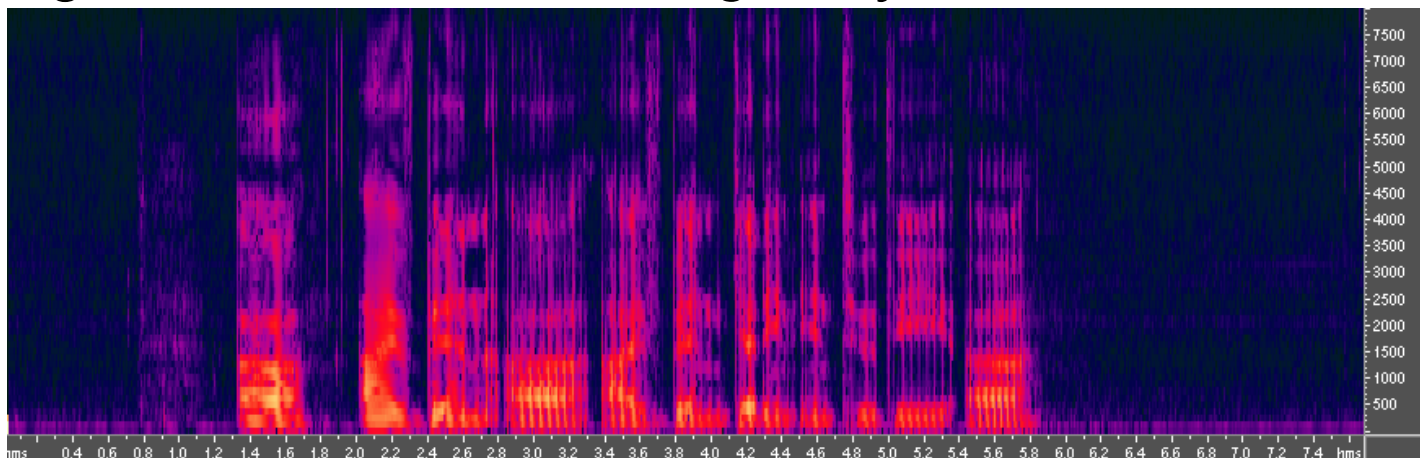
Actual Input: *"I have been I have been getting into"*

Conver-  
Sational  
Speech



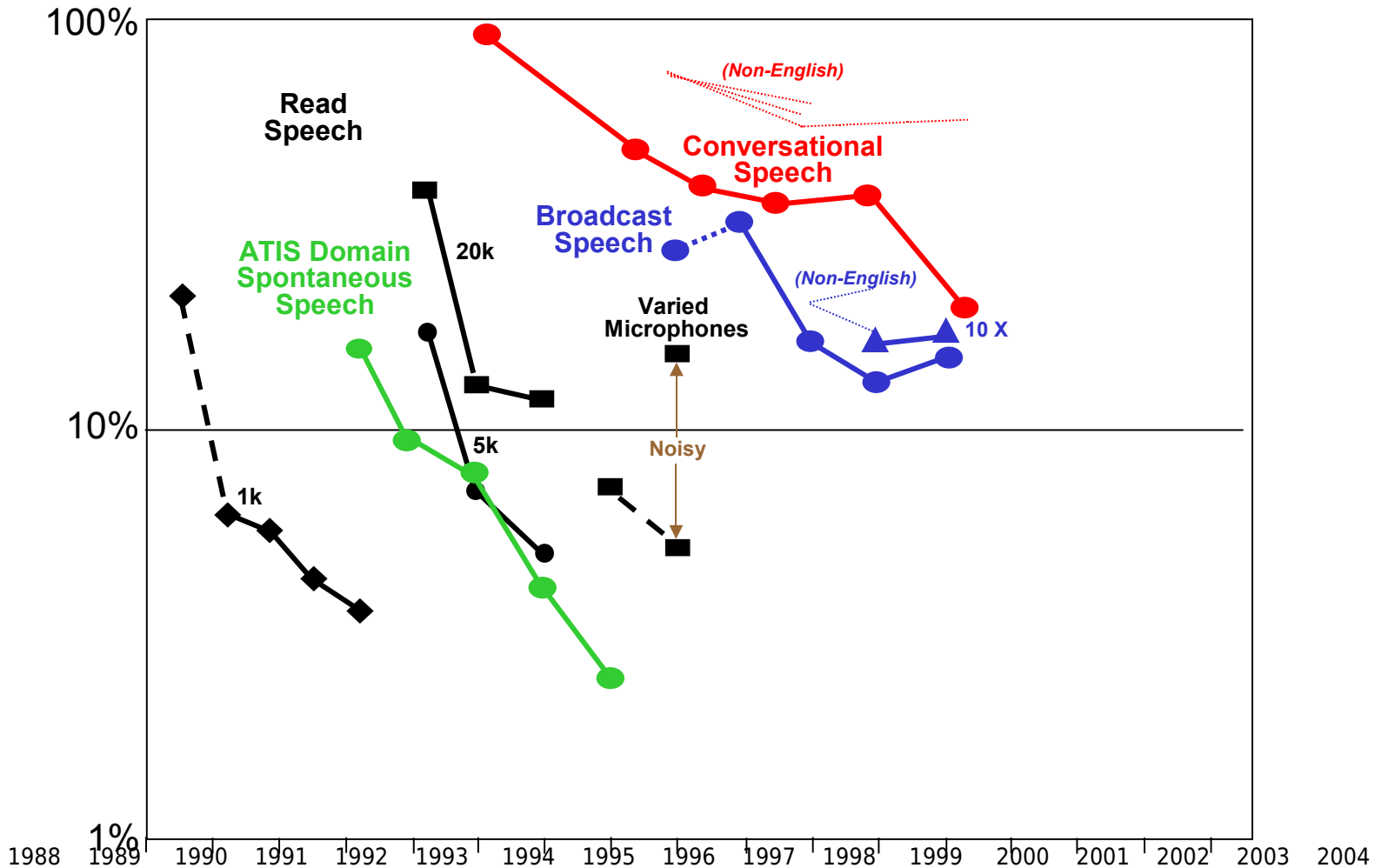
Recognition: *"and I am I being too yeah"*

Read  
Speech

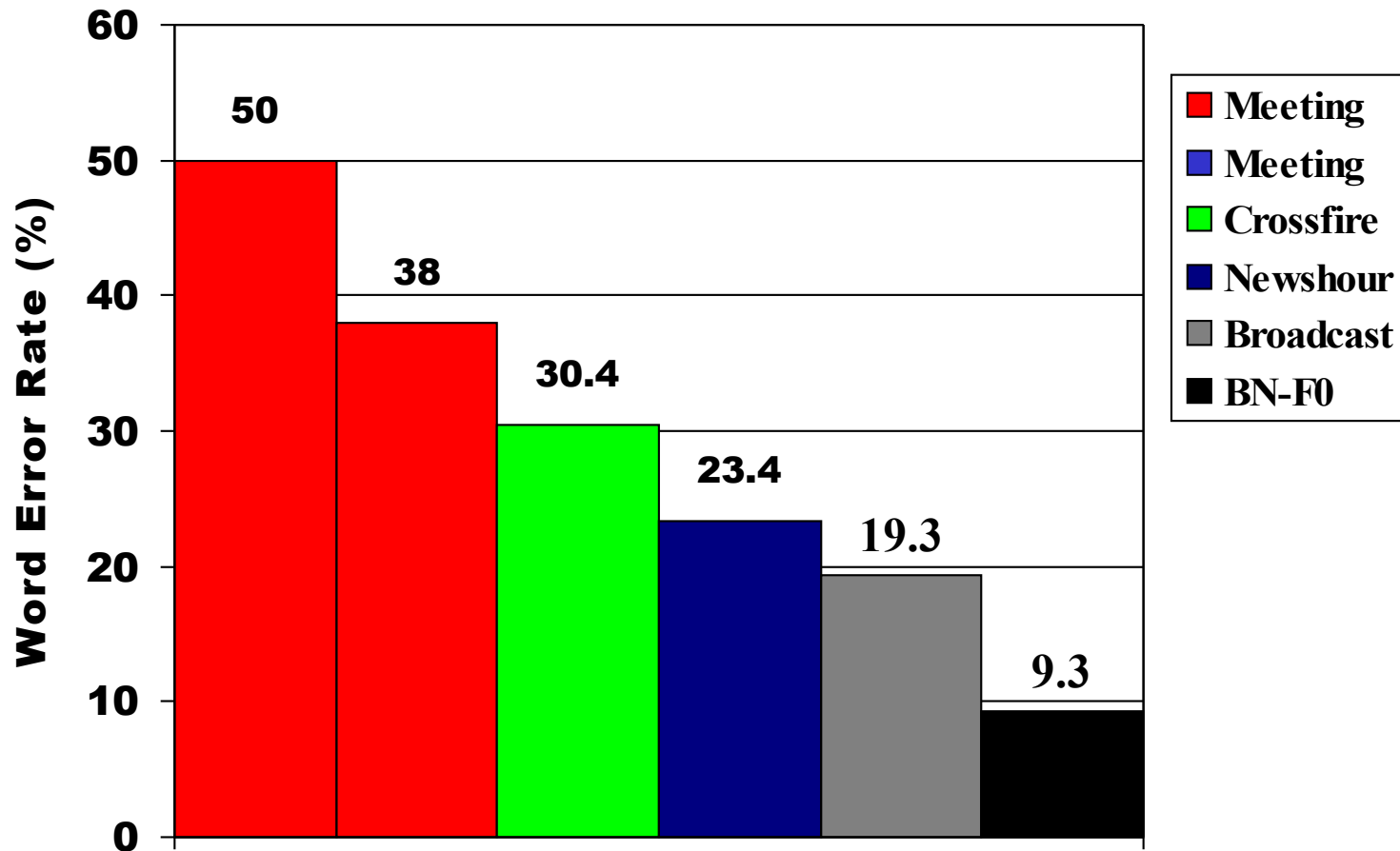


Recognition: *"I have been ties than getting into the"*

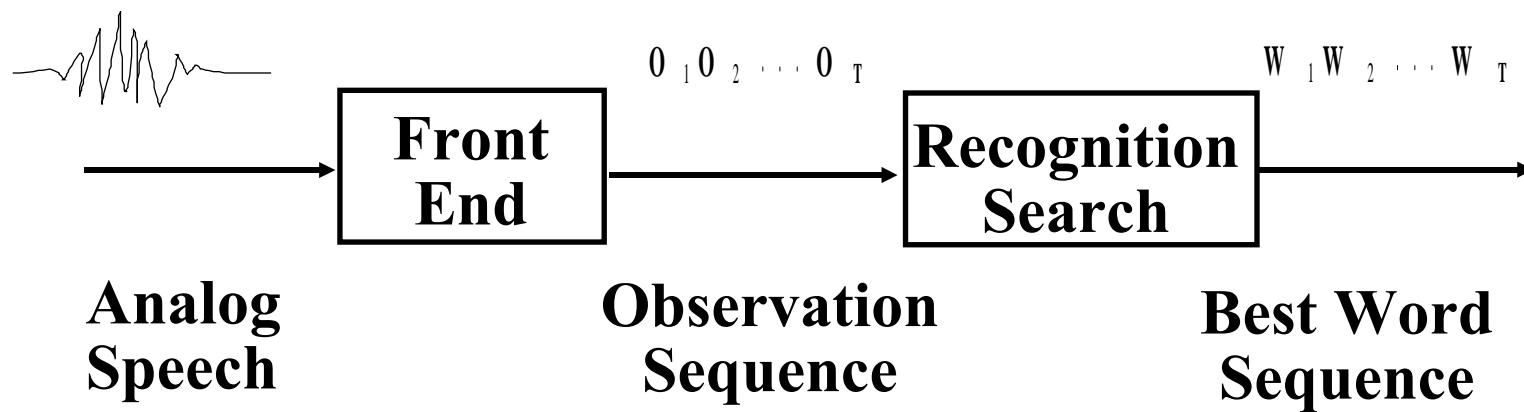
# DARPA Speech Programs: Development of the State-of-the-Art



# Recognition of Speech in Meetings

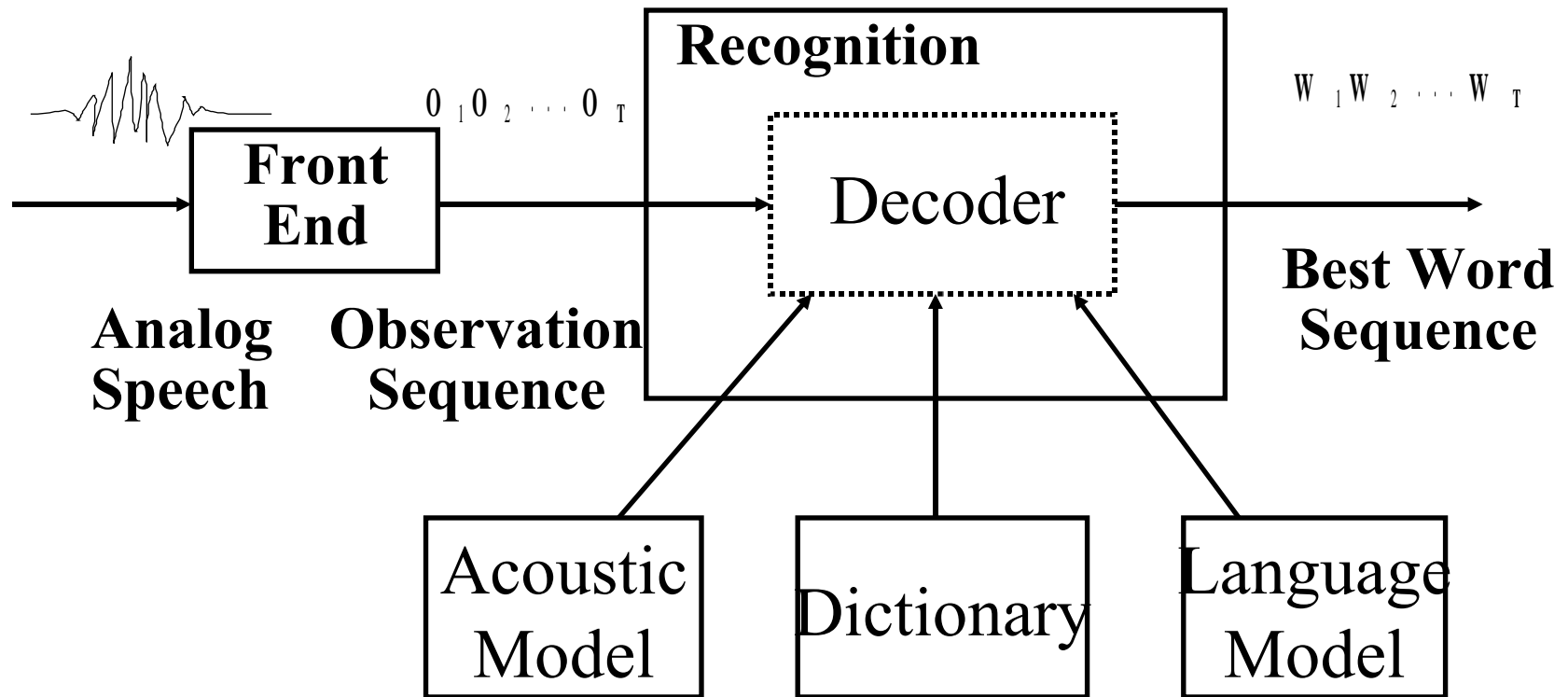


# Speech Recognition (System Overview)



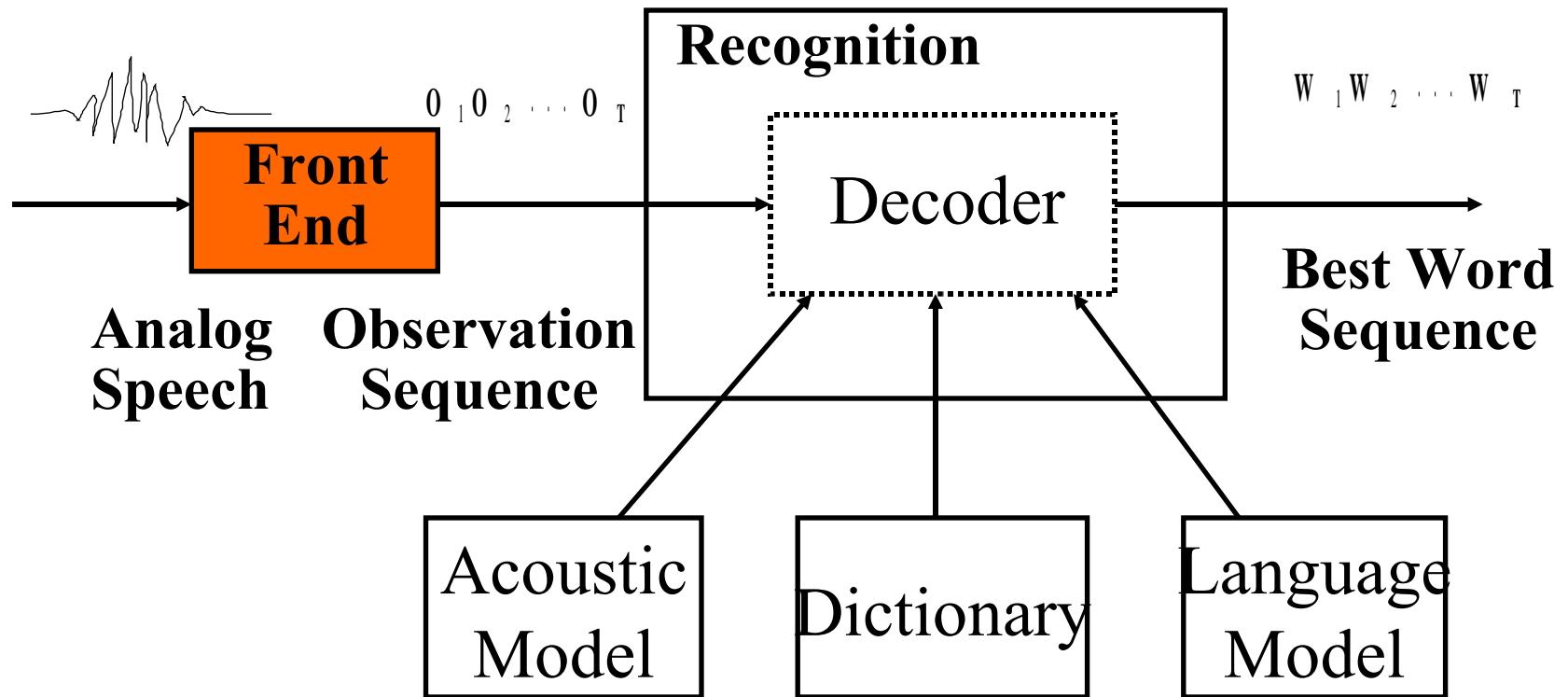
# Speech Recognition (Components)

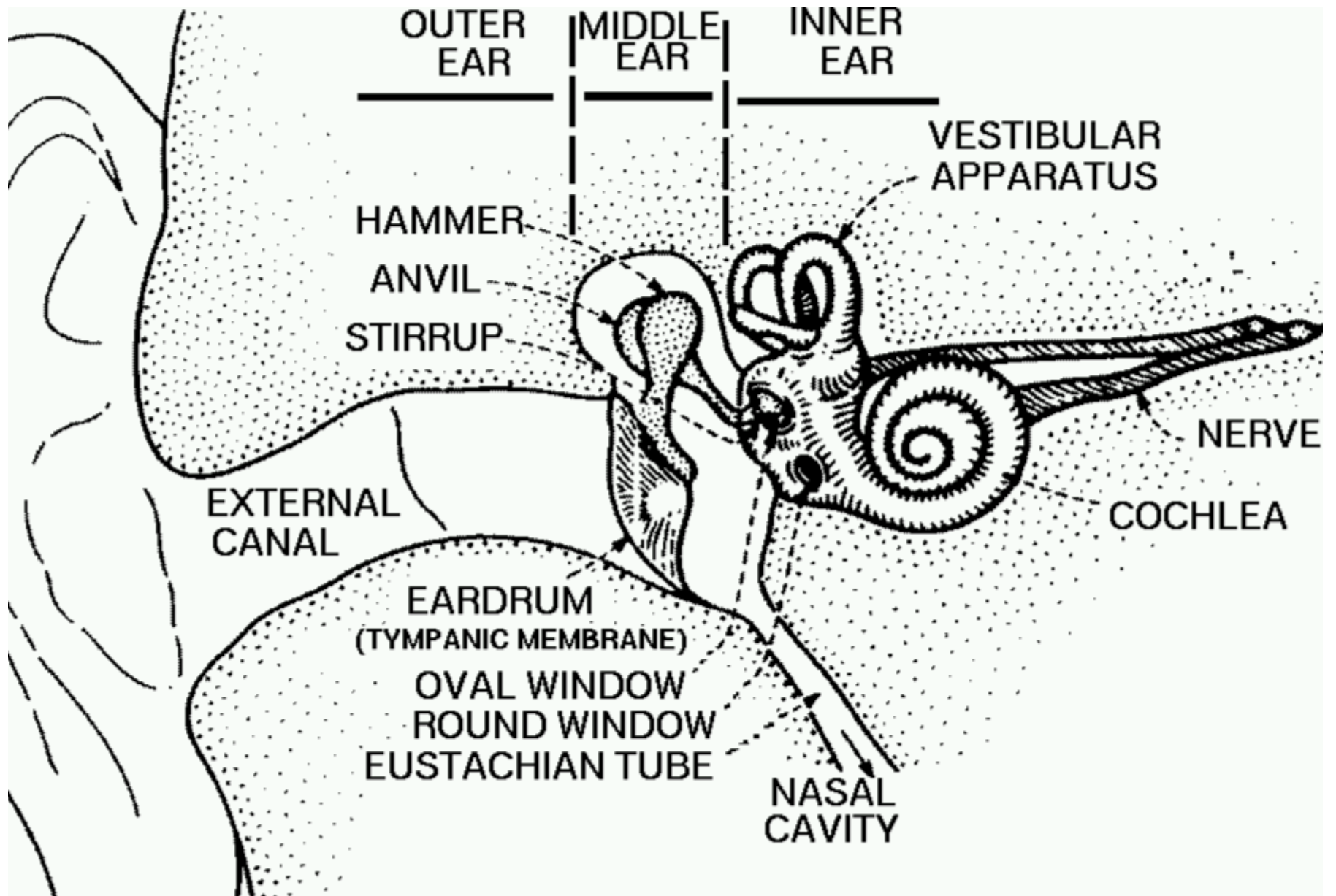
- Recognizer Components:



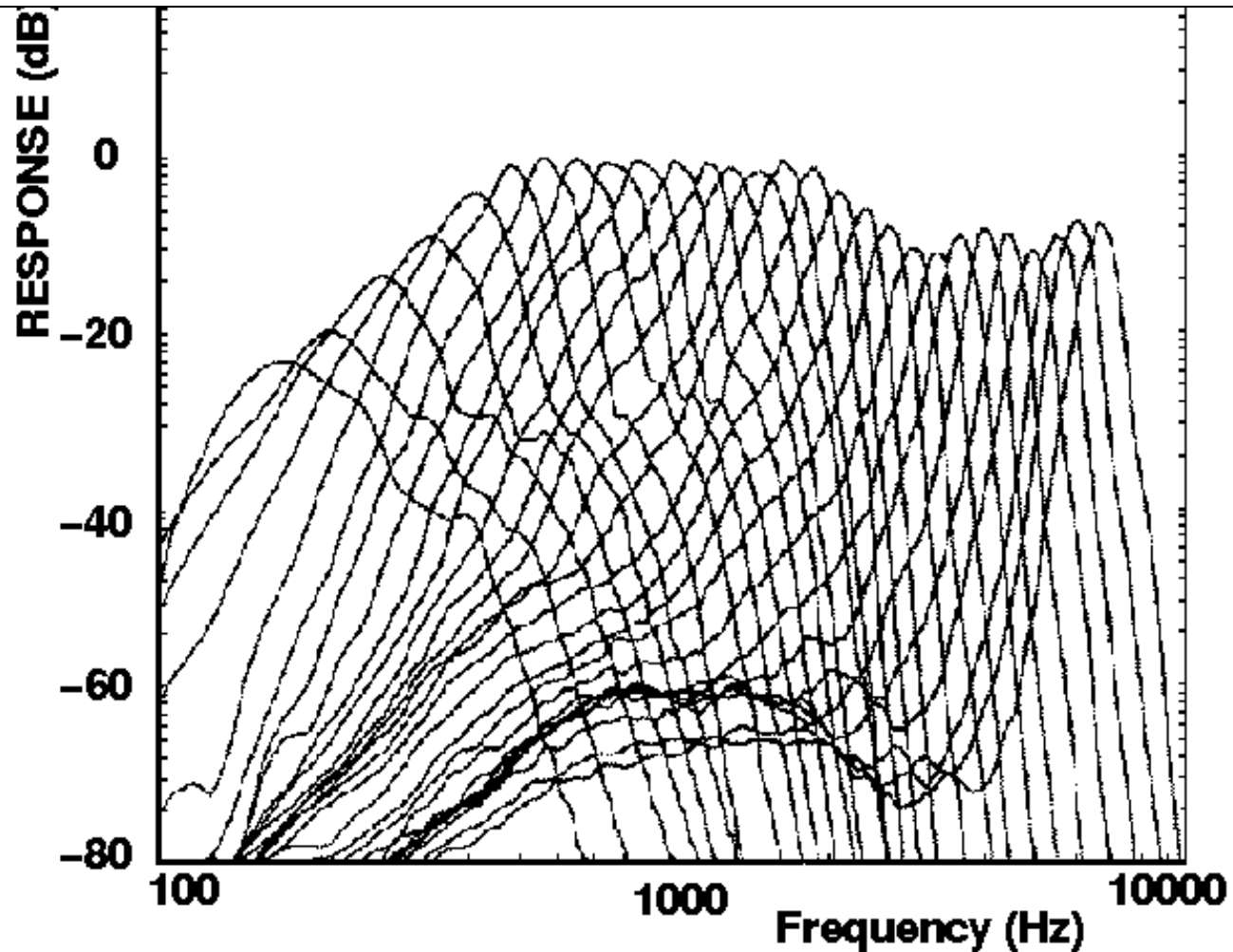
# Speech Recognition (System Components)

- Recognizer Components:



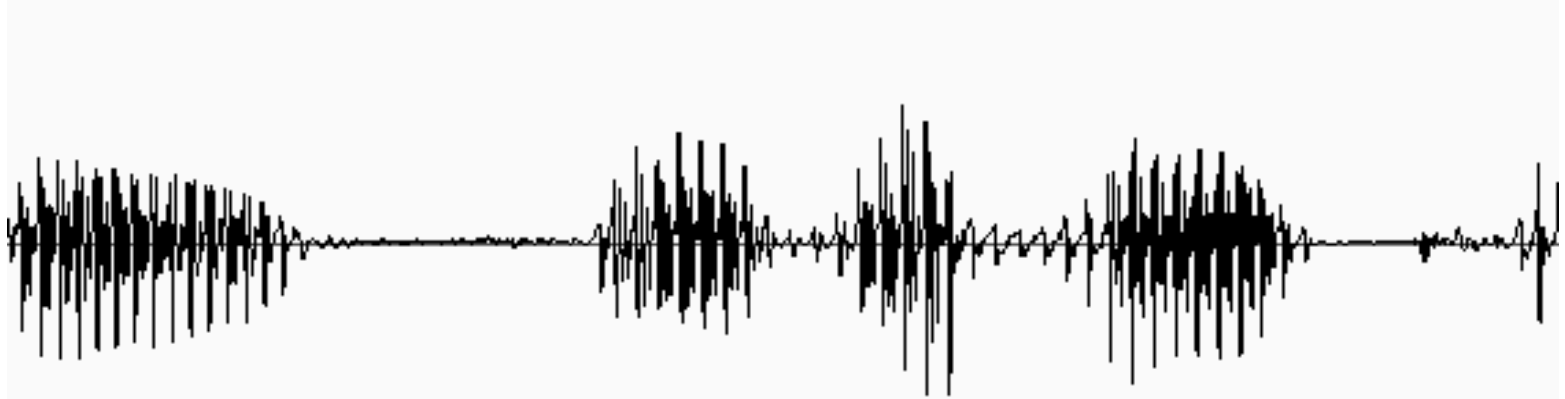


# Frequency Response of the Basilar Membrane



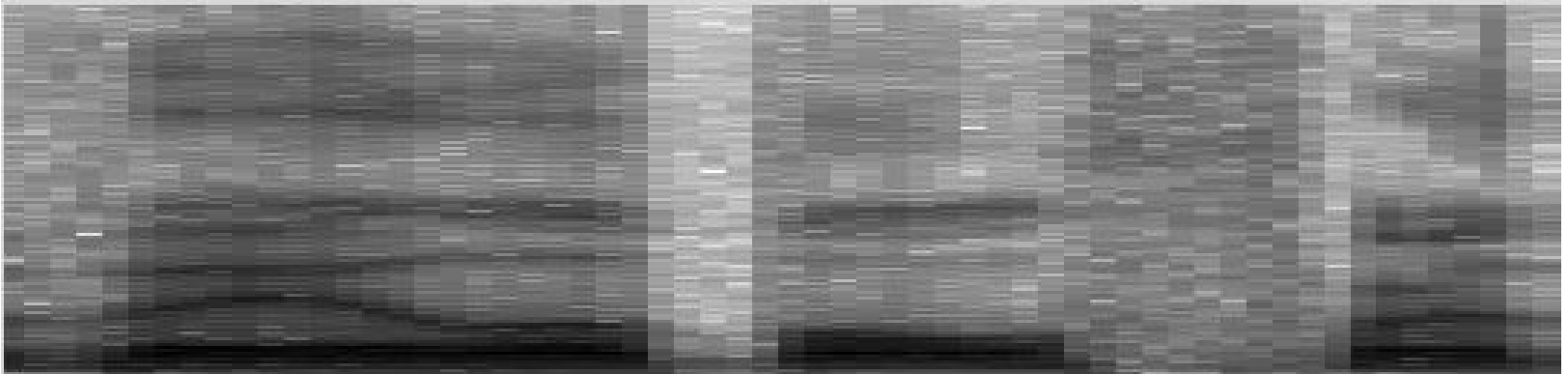
# Analog to Digital

- Sampled wave form
- Sampling Rate, Sample Resolution, Example: 16 kHz, 16 bit



# Front End Processing

- Reduce influence of undesired components --> Accuracy
- Reduce amount of data --> Speed
- Spectrum Contains most Important Information
- Most Popular Candidates: Mel-Scale FilterBank, LPC, Cepstral Coeffs

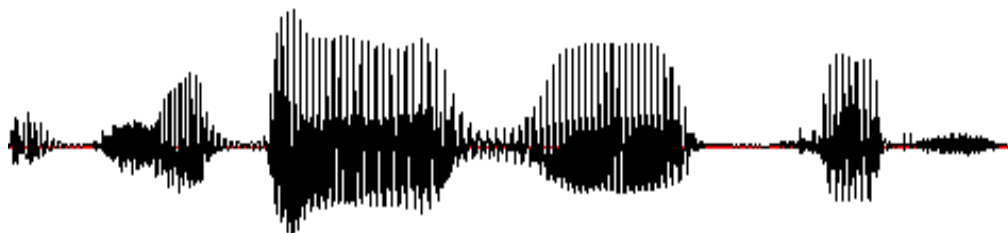


# Typical Steps

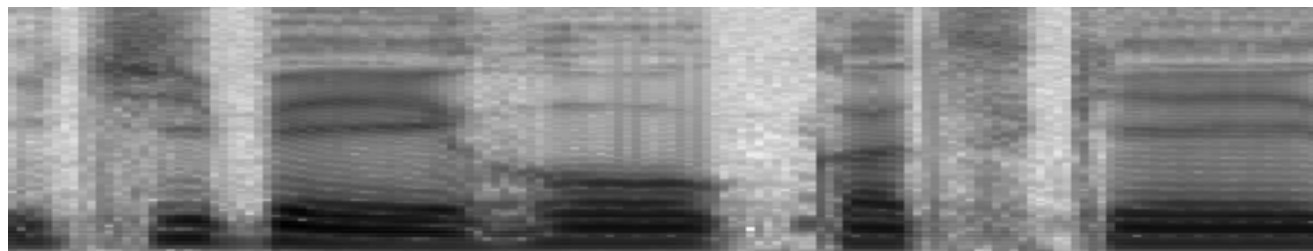
- Anti-Aliasing Filter
- AD Conversion
- Windowing
- FFT
- Compute Power-Spectrum
- Mel-Scale Filter Bank Coefficients
- Or:
  - Compute LPC or Cepstral Coefficients

# Front End Preprocessing

Recording:



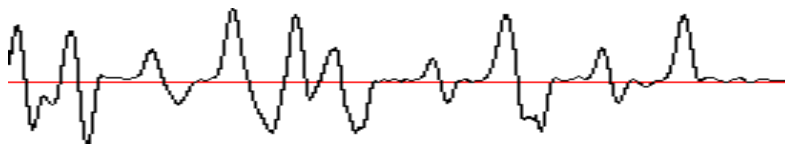
Spectrum:



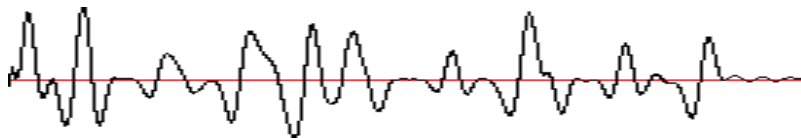
Power:



$\Delta$ -Power:

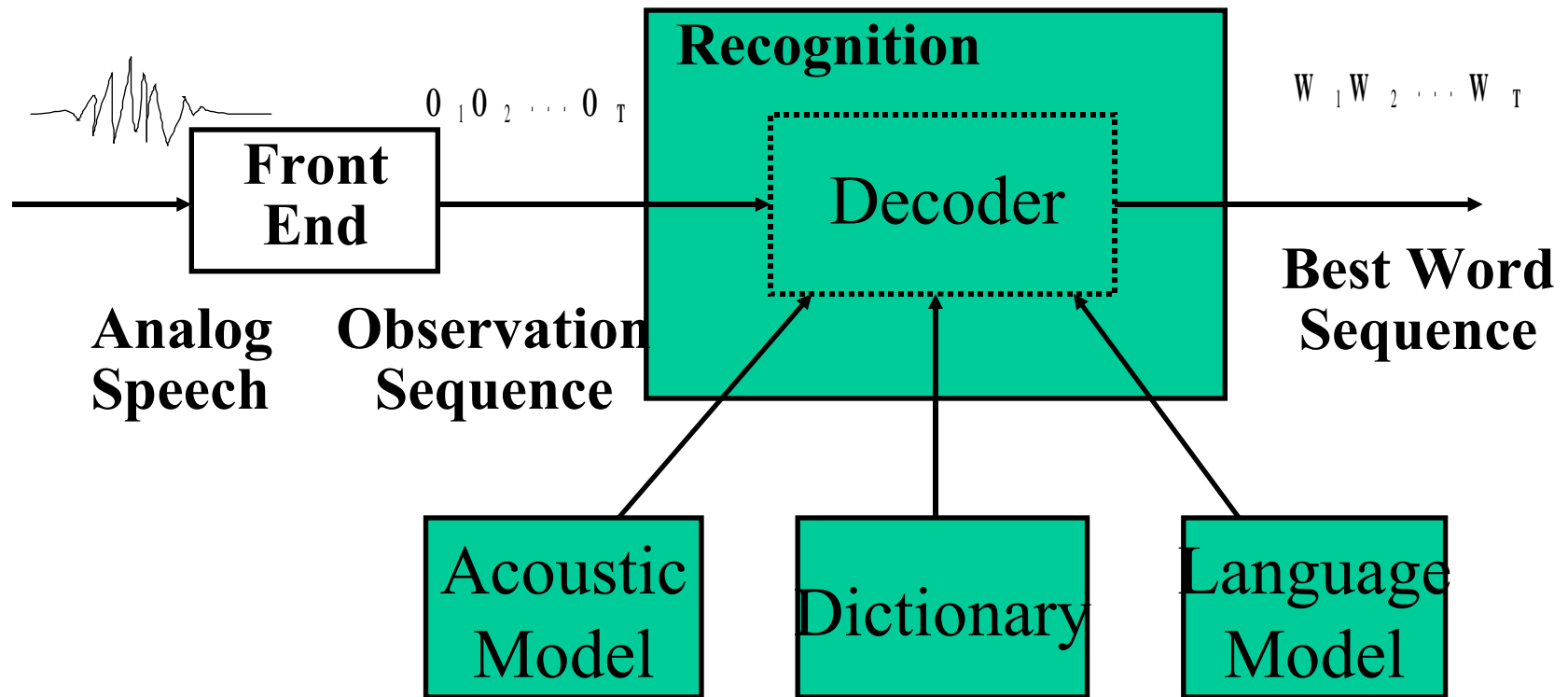


$\Delta\Delta$ -Power:



# Speech Recognition (System Components)

- Recognizer Components:



# Speech Recognition

- Goal:
  - Given acoustic data  $A = a_1, a_2, \dots, a_k$
  - Find word sequence  $W = w_1, w_2, \dots, w_n$
  - Such that  $P(W | A)$  is maximized

## Bayes Rule:

$$P(W | A) = \frac{\overset{\text{acoustic model (HMMs)}}{P(A | W)} \cdot \overset{\text{language model}}{P(W)}}{P(A)}$$

**$P(A)$  is a constant for a complete sentence**

# Speech Recognition (Components)

- Recognizer Components:

