

Origins of Computer Science

# Automatic Speech Recognition

*Ksenia Shalnova*

*07 December 2007*

# Speech Recognition



Easy for humans and difficult for computers



- Words consist of a sequence of sounds called *Phones*
- *Phone* is the name of the sound while a *phoneme* is the name of the underlying concept in the speaker's brain
- The set of *phonemes* for a particular language is the smallest number of different sounds needed to distinguish all its words

# Speech Recognition

Easy for humans and difficult for computers

Phoneme æ	Phones <i>(or acoustic realisations of a phoneme æ)</i>
<i>Dad</i> [dæd]	
<i>Man</i> [mæn]	

The **GOAL** of Speech Recognition is to identify a **sequence of phonemes** based on the acoustic realisations or phones

# Speech Recognition

Easy for humans and difficult for computers

<b>Human ear</b> Detects frequencies in the range <b>20Hz – 20 kHz</b>	<b>Computer</b> Can also detect frequencies <b>above 20 kHz</b>
<b>Human's brain</b> Perceives speech in a language and life context, i.e. <b>understands speech</b>	<b>Computer</b> <b>Can computer really understand speech?</b>

*The more I try to understand what it means for a machine to understand, the less I understand...*

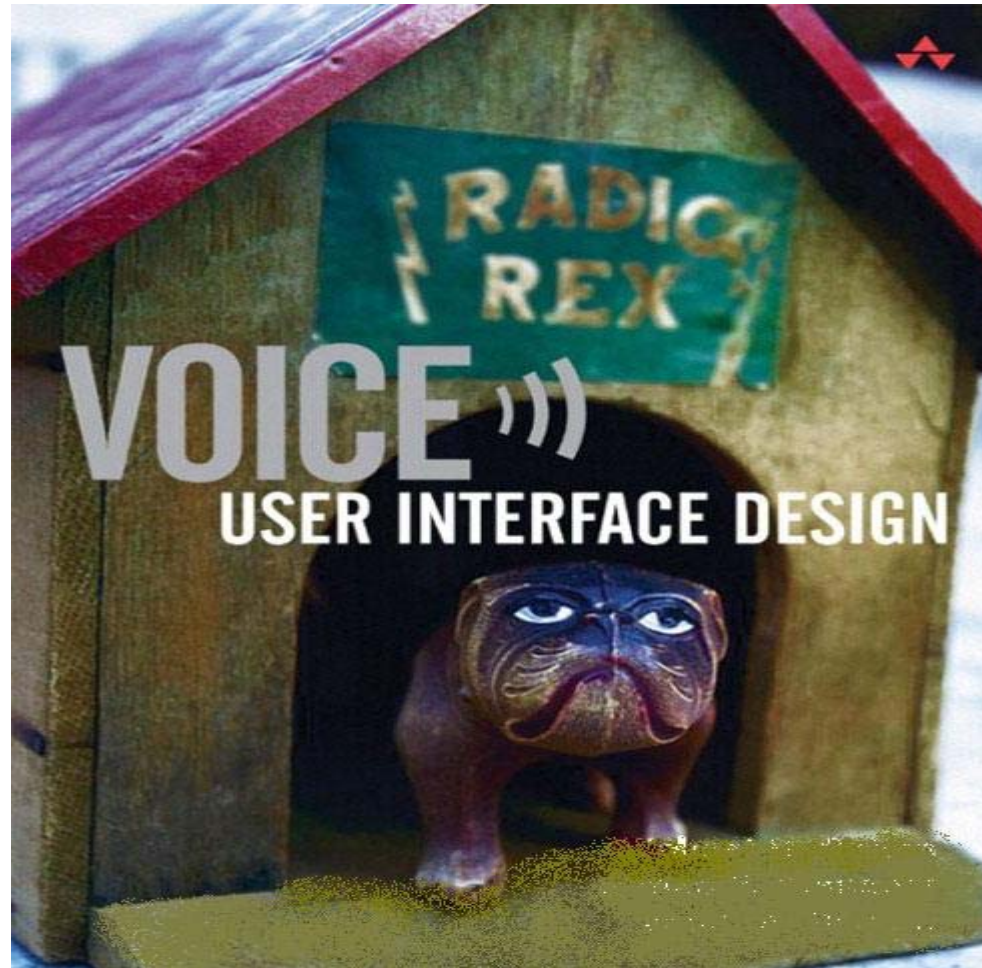
(2001:A space Odyssey, Arthur Clarke).

# Applications of Automatic Speech Recognition

- Domestic applications
- Data entry and retrieval
- Telecommunications
- Assistive technologies
- Command and control
- Education

# First Automatic Speech Recognition

## 1911



# History of Automatic Speech Recognition

*Unlike Speech Synthesis, “proper” ASR became possible only with modern computers.*

- 1952 – started at AT&T Bell Labs. Isolated digit recognition in a noise-free environment
- 1964 – Neural Networks
- From 1970-s – applying Dynamic Time Warping
- From 1980-s – applying Hidden Markov Models; incorporating linguistic knowledge, noise immunity. ASR reaches the commercial market

# History of Automatic Speech Recognition

## **Audio-related areas in 1950-ies which made possible creation of ASR:**

- Phoneme recognizer
- Formants discovery
- Using Spectrum information

## **Non-audio areas related to research in ASR:**

- Psychological and biological aspects on hearing
- Pattern recognition
- Decision theory

# History of Automatic Speech Recognition (Digit recognition in Early Days)

## Training

*Parameter extraction based on spectrogram analysis*

Template "zero"

Template "one"

...

Template "nine"

*Spoken word*

Parameter extraction

Unknown template

Template match

*Output a digit if good match, otherwise no output*

## Recognition (or Testing)

# History of Automatic Speech Recognition (Early Days)

## **Conditions for successful recognition**

- Totally noise-free environment (no doors closing, no dogs barking, no engines working etc.);
- A very small number of templates for training, preferably completely different in pronunciation (“one” and “nine” could be mixed due to their similarity);
- The same speaker for training and recognition

# History of Automatic Speech Recognition (Dynamic Time Warping)

Used for comparison of acoustic patterns of different length and with mismatches.

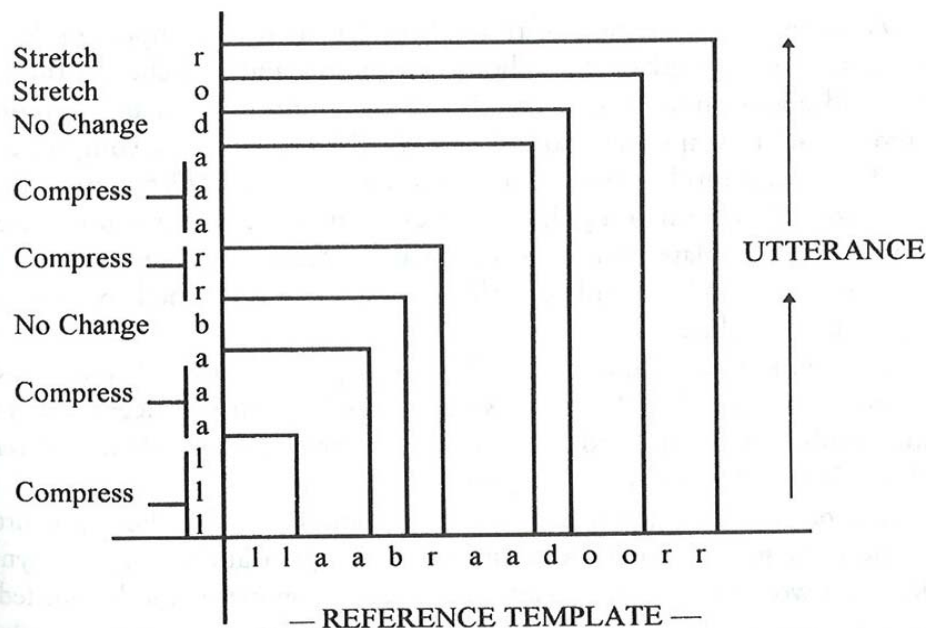


Fig.1. DTW using letters to represent windows of speech.

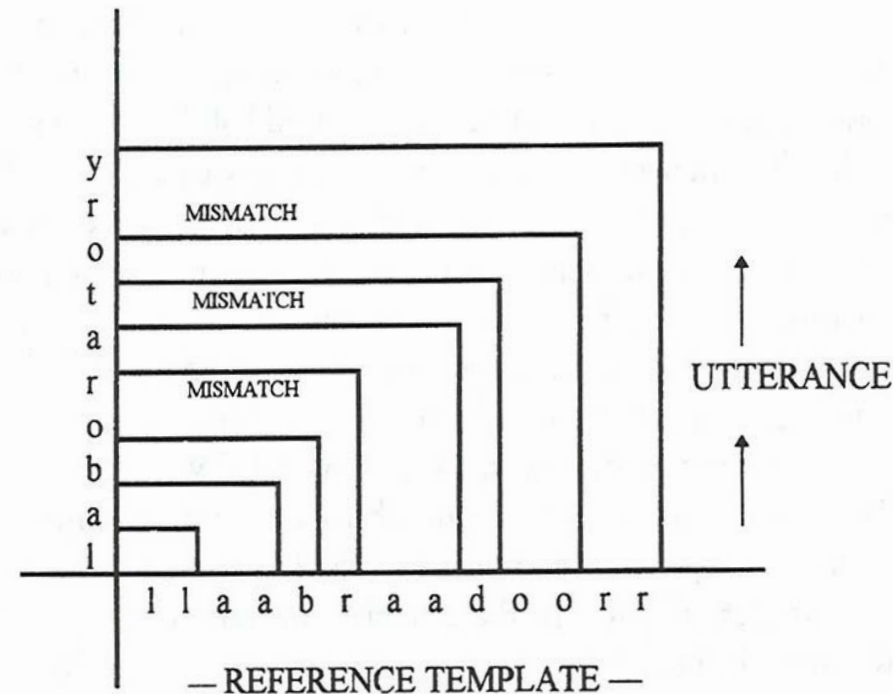
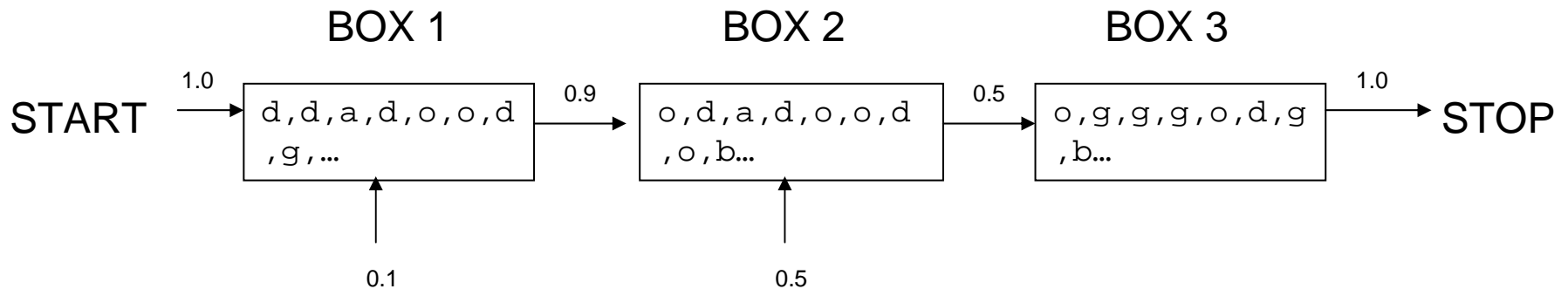


Fig.2. DTW using letters to represent windows of speech with mismatches.

# Present of Automatic Speech Recognition

## Hidden Markov Models

Recognition of a word *dog, doog, ddog ...*



**Probabilistic process in two stages:**

1. Tile observation (letters b,d,a,...)
2. Box selection (or state selection)

# Present of Automatic Speech Recognition

## Hidden Markov Models

### Why Markov?

Name of state transitions where the next state (or box) is determined only by the current state (or box).

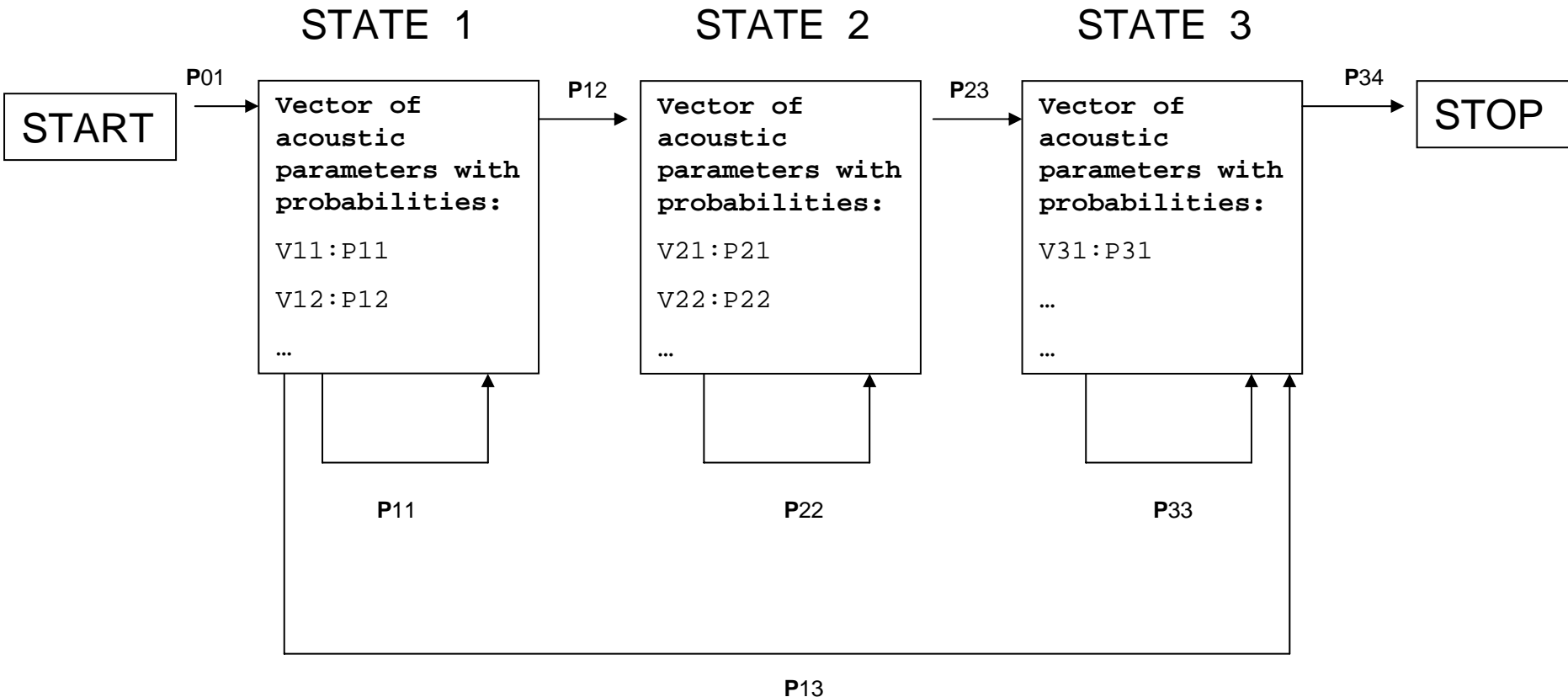
### Why Hidden?

Actual state sequences are hidden from us. **D-o-o-g** may come from state sequence Box1-Box1-Box2-Box3 or from state sequence Box1-Box2-Box2-Box3.

# Present of Automatic Speech Recognition

## Hidden Markov Model for sound recognition

HMM model for recognition of 3 successive sounds



# Present of Automatic Speech Recognition

How to improve the performance given there is a probability of errors in recognition of sounds?

Add higher levels of Language Knowledge

1. Syntax (or word order)

*I am flying to Boston.*

*Flying am I to Boston.*

2. Semantics (meaning of words)

*Is the baby crying?*

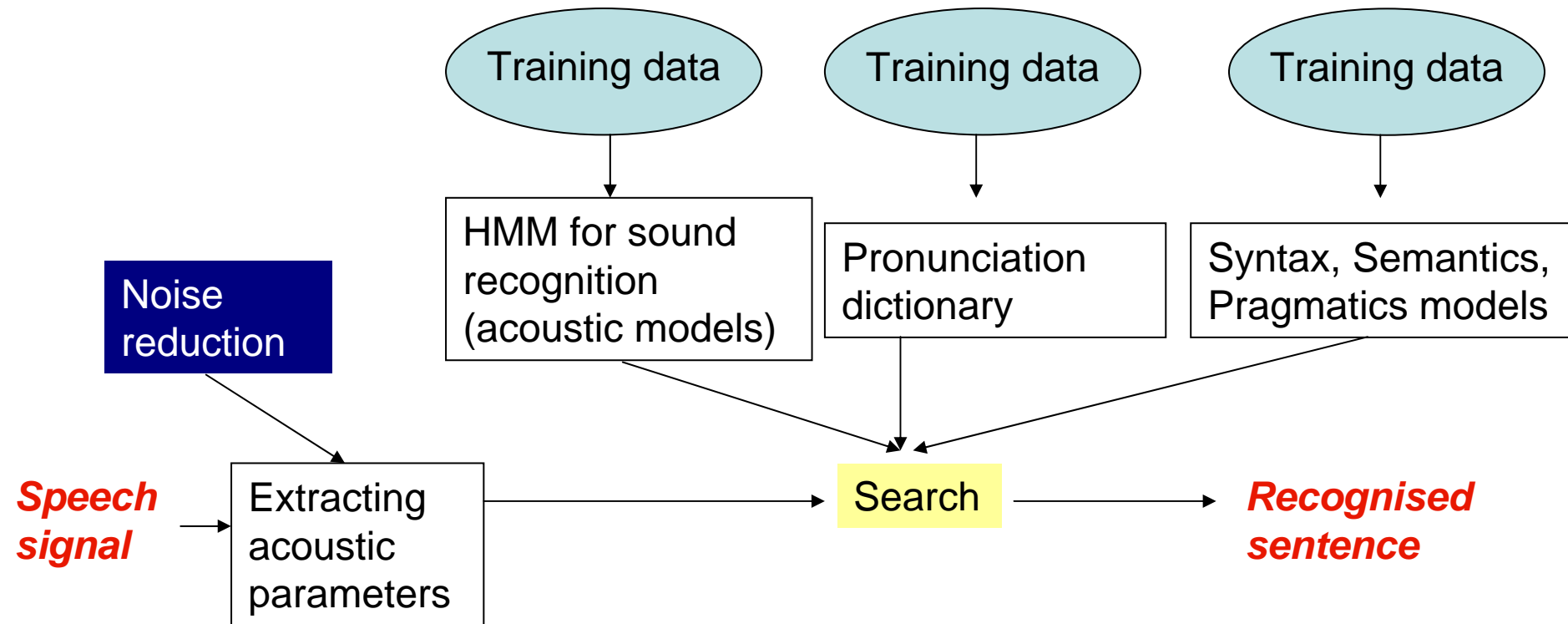
*Is the bay bee crying?*

3. Pragmatics (context of a whole sentence)

*It is difficult to recognize speech.*

*It is difficult to wreck a nice beach.*

# Components in modern ASR



# Types of ASR systems

- Speaker-independent versus Speaker-dependent
- Continuous versus Discrete
- Read speech versus spontaneous speech
- Pronunciation Vocabulary size – small (<20 words) to large (>50,000 words)

# Present and Future of ASR systems

## Technological improvements

- Find an Alternative approach to Hidden Markov Process
- ASR of Spontaneous speech
- ASR in Environmental noise
- ASR of Dialectal speech (or non-native speech)

*Is it a final Goal to recognize all the speech of all the people speaking simultaneously? It is probably beyond human capabilities.*

## Application challenge

- Automatic translation of one continuously spoken language into another in real time
- Speech and Robotics

# Available ASR systems

## **You can try yourself now**

Recognition of isolated words

Automatic railway query service in UK (tel. 08457484950):

Speaker-dependant recognition that requires training for better performance

Windows XP or Windows Vista

## **You have to buy**

Speaker independent

Dictation (IBM, Scansoft)

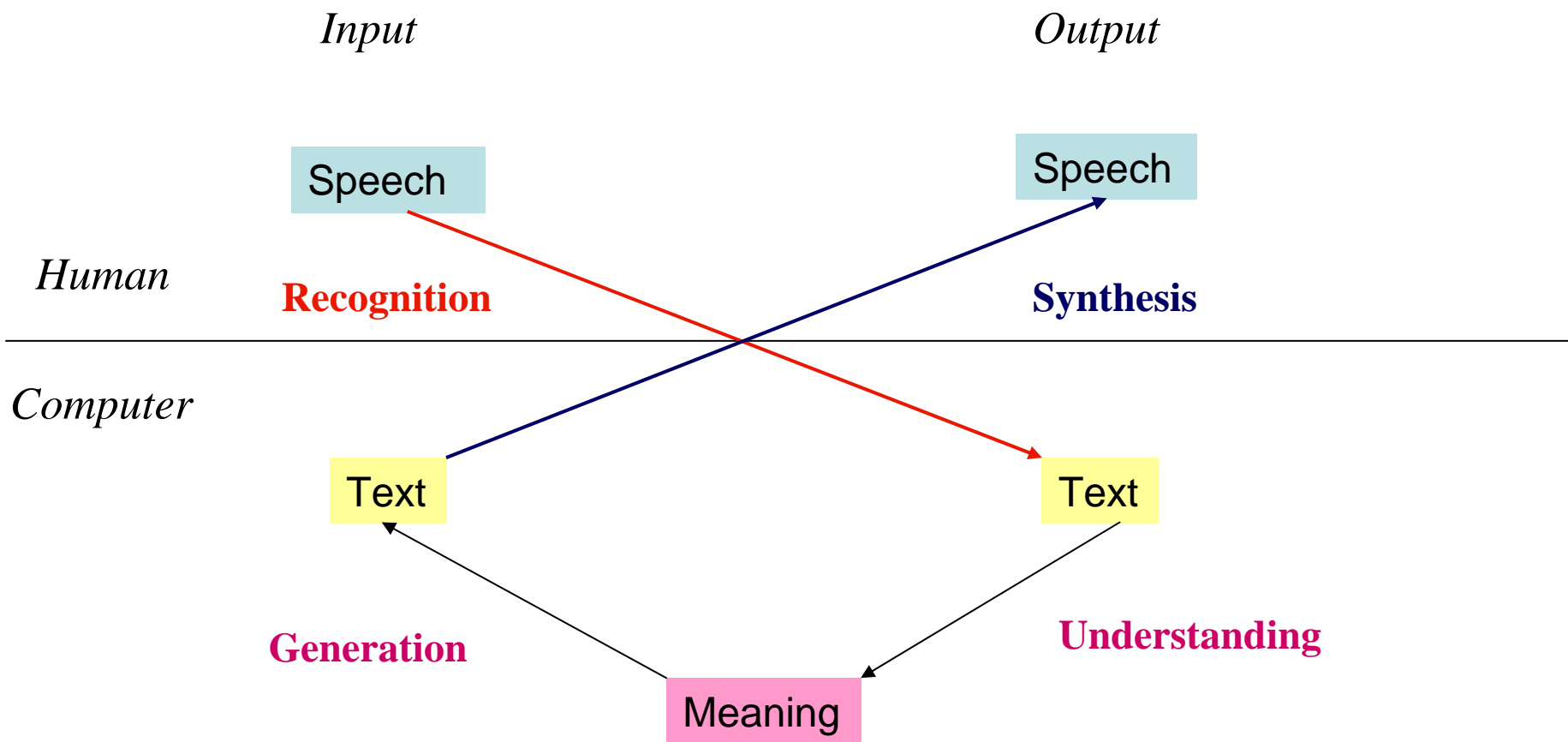
Speaker independent with a restricted vocabulary

Telephone transactions (AT&T, Philips, Speech Works etc.) to be used in automated brokerage system, shipping costs etc.

# Demo of Windows XP Automatic Speech Recognition



# Combining Speech Synthesis and Automatic Speech Recognition



# Speech and Robotics

R. Stiefelhagen, C. Fuegen, P. Giesemann, H. Holzapfel, K. Nickel, A. Waibel. **Natural Human-Robot Interaction using Speech, Gaze and Gestures**, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.