

How to Speak to Computers

David Andrs

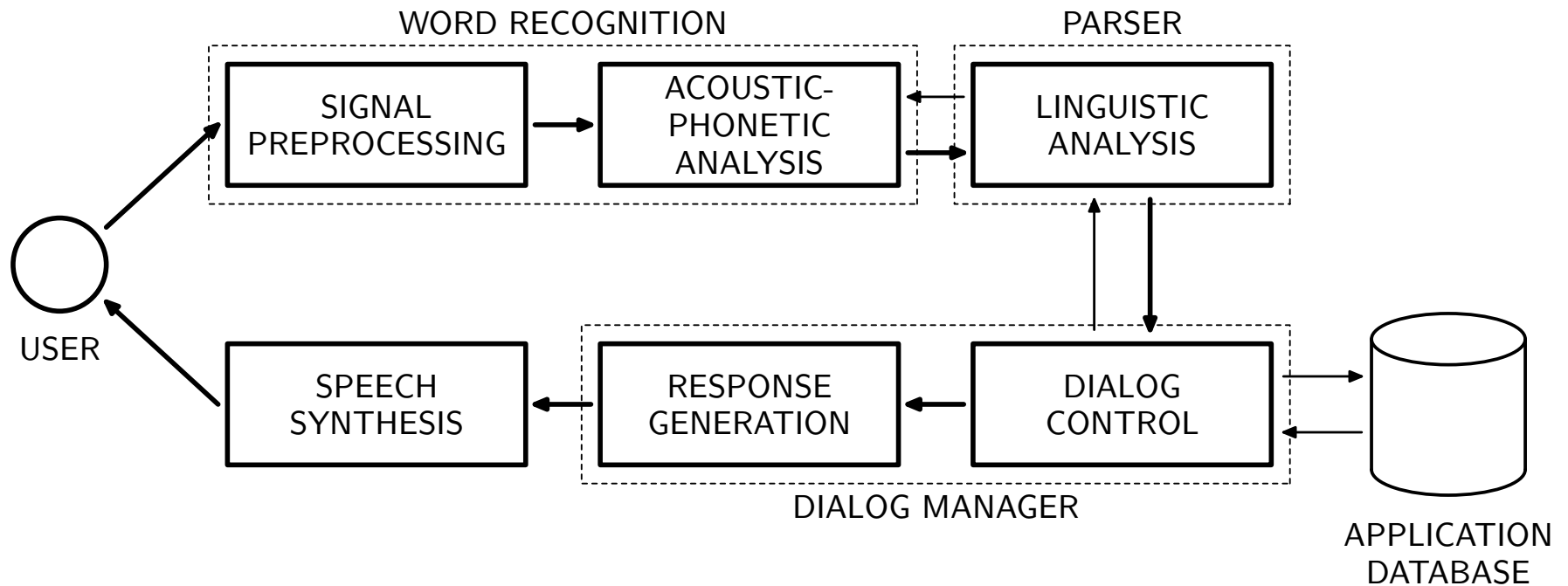
Department of Mathematical Sciences
University of Texas at El Paso

February, 2007

Natural Language Processing

- Text processing
 - analyzing documents, indexing, clustering, searching, ...
 - www.google.com
- **Speech processing**
 - speech analysis
 - * ASR
Speech signal → written form
 - * ASU (= ASR + Linguistic Analysis):
Speech signal → structure reflecting relations
 - speech synthesis
 - dialogue (analysis + synthesis)
 - isolated words
 - continuous speech
 - spontaneous speech

Dialogue Systems



- Speech synthesis
- Speech analysis
- Linguistic analysis
- Dialogue

Speech Synthesis

Speech Synthesis

Goal:

- To generate speech signal in a way that it sounds naturally

Problems:

- Different length of phonemes
(different length of the same phoneme in different words)
- Accent
- Melody
- Coarticulation

Speech Synthesis – Approaches

- Parametrical
 - Frequency analysis → coefficients
 - Generating a wave based on the coefficients
 - Sounds very unnaturally (no prosody, no word/sentence accent, no melody)
- **Concatenation of words/phrases**
 - Building up the sentence from the set of isolated words/phrases
 - Easy to implement
 - How to connect words together to sound naturally (“gaps”)
 - Extending the vocabulary might be hard
(the original speaker is not available, ...)
- Modeling vocal tract
 - Gives the best results
 - No real-time implementation yet, computationally difficult
 - Adjustable voice

Speech Synthesis – TTS

TTS: Text-to-Speech

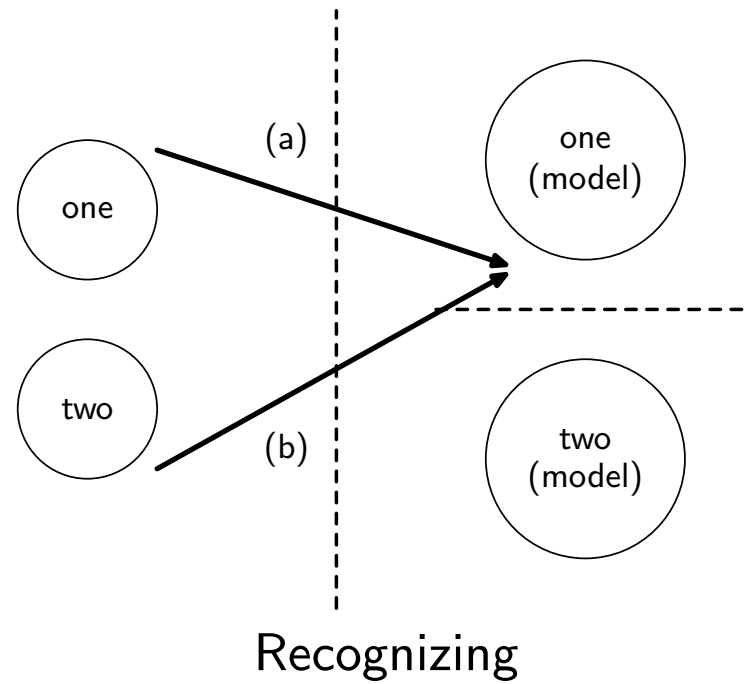
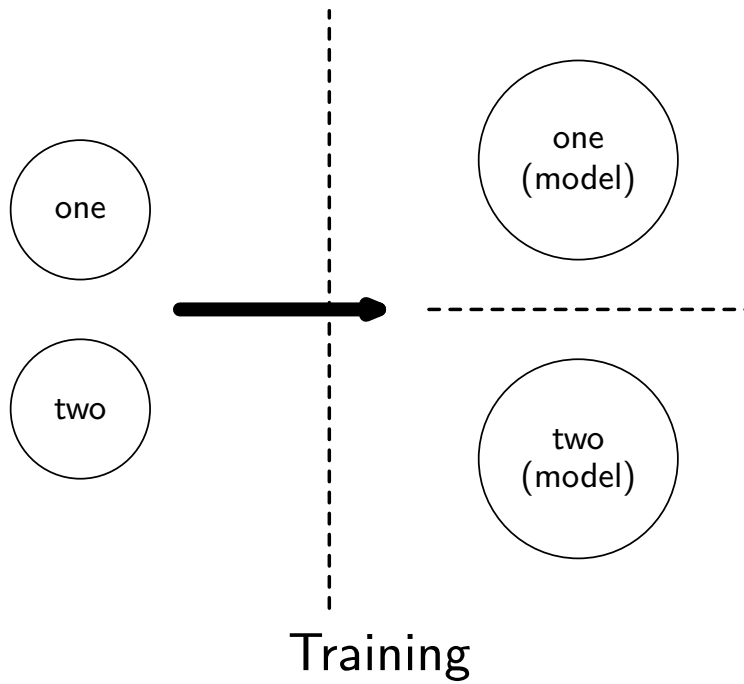
Input: Today is 22th May, 2006

Output: [t'deɪ ɪz ð twentɪ seknd of meɪ, twentɪ oʊ sɪks.]

- Transcription from text to phonetic form
- Marks to modify the signal
 - duration of phoneme
 - changing the melody
- Problems with coarticulation

Speech Analysis

Basics of Recognition



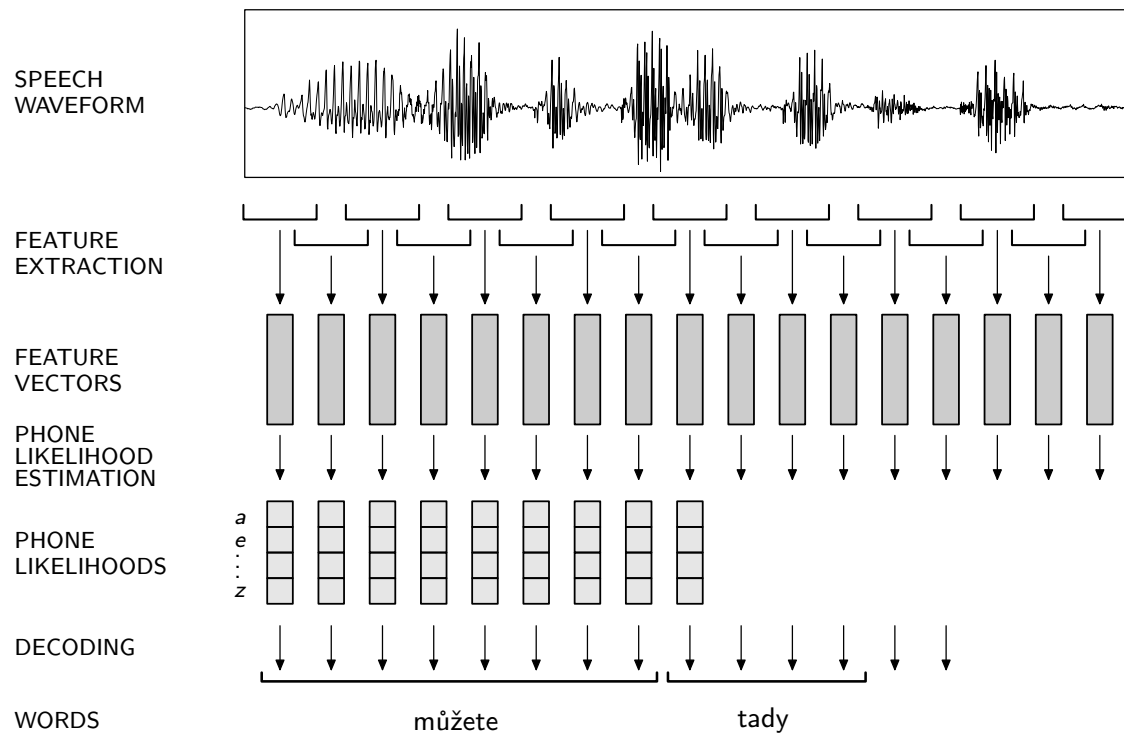
Isolated Words – Analysis

- Command-and-control applications
- Small vocabulary – up to 100 words
- Speaker dependent → Speaker independent
- Practically solved
 - Voice dialing in cellular phones
 - Suitable for eyes-busy and hands-busy applications
- **Principle**
 - A model for each word
 - Calculating the “distance” between spoken word and each model
 - The model with the smallest “distance” is the hypothesis what was said

Continuous Speech – Analysis

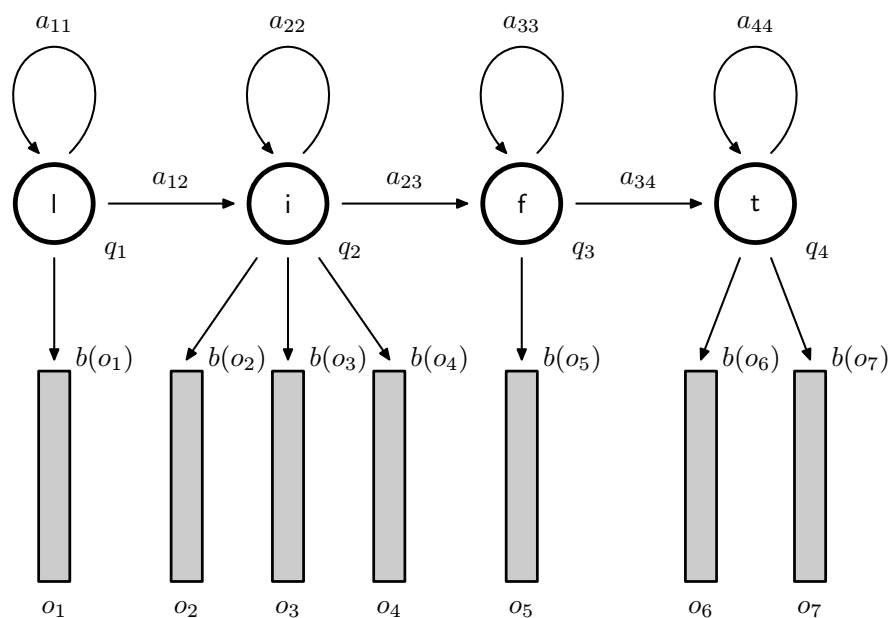
LVCSR: Large Vocabulary Continuous Speech Recognition

- Large vocabulary $\approx 300,000$ words



Continuous Speech – Analysis

HMM: Hidden Markov Models – Example



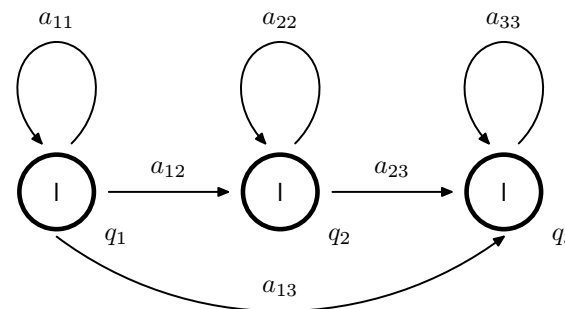
Input: Sequence of feature vectors (observations)

Output: Sequence of states

Continuous Speech – Analysis

HMM: Hidden Markov Models – Implementation

- Phonetic alphabet (≈ 30 elements)
- 3-state HMM for every element
- Lexicon (set of words)
- Grammar (how words are connected to build up a sentence)
- or language models (capable to estimate the following word(s))



Continuous Speech – Analysis

Decoding: Viterbi algorithm

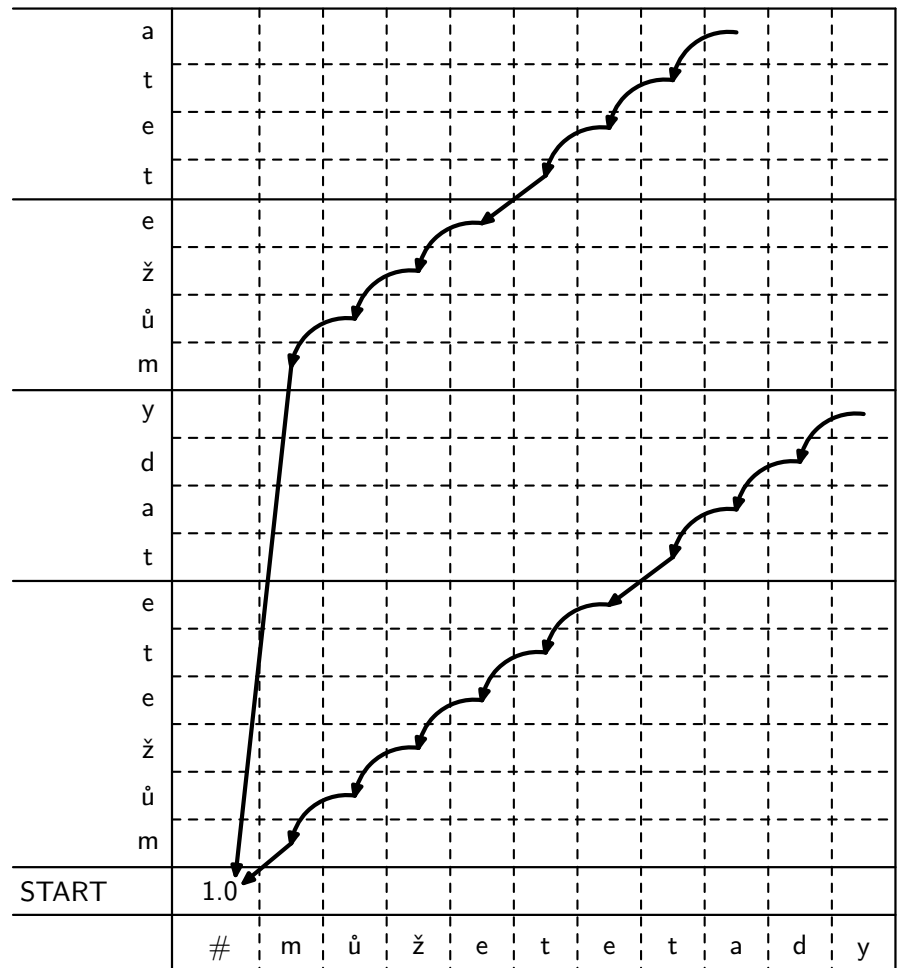
- Word Segmentation
- Forward + Backward phase

Input: Sequence of states

Output: Word sequence

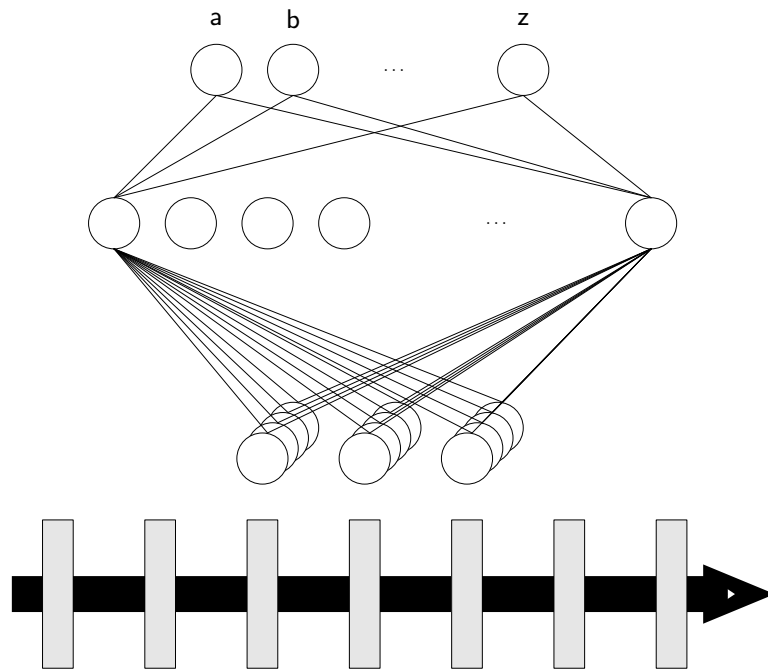
2 segmentations

- m ů ž e t e | t a d y
- m ů ž e | t e t a d y



Continuous Speech – Analysis

ANN: Artificial Neural Network



- Multilayer perceptron
- Hidden layer
- Output layer
(phoneme probabilities)

Input: Feature vectors

Output: Sequence of vectors of phone probabilities

Continuous Speech – Analysis

Language Models

$$W = w_1 w_2 \dots w_N$$

$$P(W) = P(w_1 w_2 \dots w_N) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \dots P(w_N|w_1 w_2 \dots w_{N-1}).$$

$$\hat{P}(w_1, w_2, w_3) = \frac{F(w_1, w_2, w_3)}{F(w_1, w_2)}$$

- Introduced by F. Jelinek
- Basis of state-of-the-art speech recognizers
- N-gram language models (bigrams $N = 2$, trigrams $N = 3$)

Problems

- **Large enough corpus** – Good estimates of $P(W)$
- **Sparse data problem**

Continuous Speech – Sparse Data Problem

**Estimate the probability of unseen N-grams
by redistributing the total probability mass.**

Smoothing techniques

- Add-one

$$\hat{P}(w_1, w_2) = \frac{F(w_1, w_2) + 1}{F(w_1) + |V|}$$

- Good-Turing Discounting
- Back-off

$$\tilde{P}(w_i|w_{i-2}w_{i-1}) = \begin{cases} P(w_i|w_{i-2}w_{i-1}), & \text{if } F(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha P(w_i|w_{i-1}), & \text{if } F(w_{i-2}w_{i-1}w_i) = 0 \end{cases}$$

- Deleted Interpolation

$$\hat{P}(w_n|w_{n-1}w_{n-2}) = \lambda_1 P(w_n|w_{n-1}w_{n-2}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n) \quad \sum_i \lambda_i = 1$$

Linguistic Analysis

Linguistic Analysis

- Syntax
- Semantics
- Pragmatics

Example: Traffic Lights

- Syntax
 - **Red**, Orange, Green
 - Red, **Orange**, Green
 - Red, Orange, **Green**
- Semantics
 - **Red** — Stop
 - **Green** — Go
- Pragmatics
 - Pedestrian signals are for pedestrians, not for cars and vice versa

Linguistic Analysis – Syntax

- Mainly when processing written text (documents)
- Spontaneous speech is agrammatical
 - ⇒ useless to waste computer time with syntax analysis

I want to go from Texas to Florida

```
(S (NP I)
  (VP want
    (S (VP to
      (VP go
        (PP from
          (NP texas))
        (PP to
          (NP florida)))))))))
```

Linguistic Analysis – Semantics

- No general formalism for describing semantics is available.
 - How to describe *yellow, square, wind, love?*
- User specific
 - User assigns the meaning (*good vs. evil*)

Applications:

- Always domain specific, very limited

To somehow extract the meaning of utterance

- Semantic interpretation: translating a natural language to a formal language

Techniques

- Keyword spotting
- Shallow parsing
- Unified Field Objects

Linguistic Analysis – Keyword Spotting

I want to **go from** *El Paso* **to** *Houston*

- **go** → *action*
- **from** *El Paso* → *source(El Paso)*
- **to** *Houston* → *destination(Houston)*

Linguistic Analysis – Shallow Parsing

- Not strictly defined
- Set of techniques to extract the meaning
- Mostly uses keyword spotting
- It is not necessary to do complete parsing

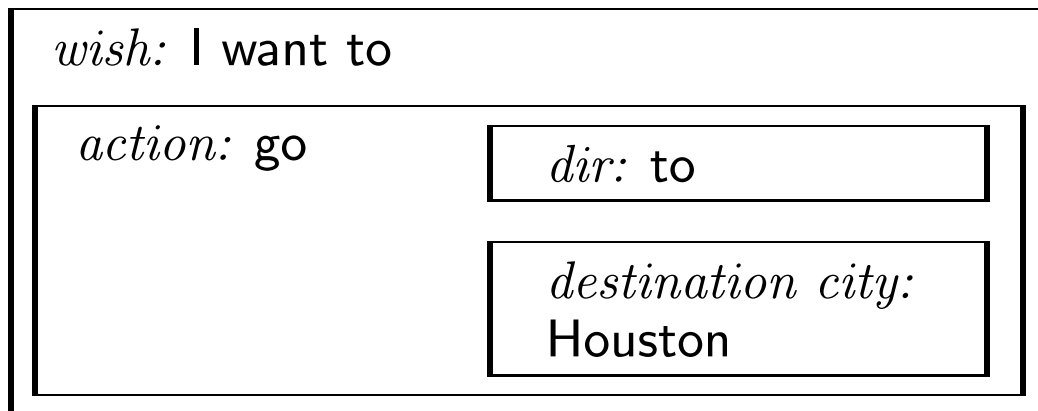
I want to **go from** *El Paso* **to** *Houston*

```
(S (NP I)
  (VP want
    (S (VP to
      (VP go
        (PP from
          (NP el paso))
        (PP to
          (NP houston)))))))))
```

```
(VP go
  (PP from (NP el paso))
  (PP to (NP houston)))
```

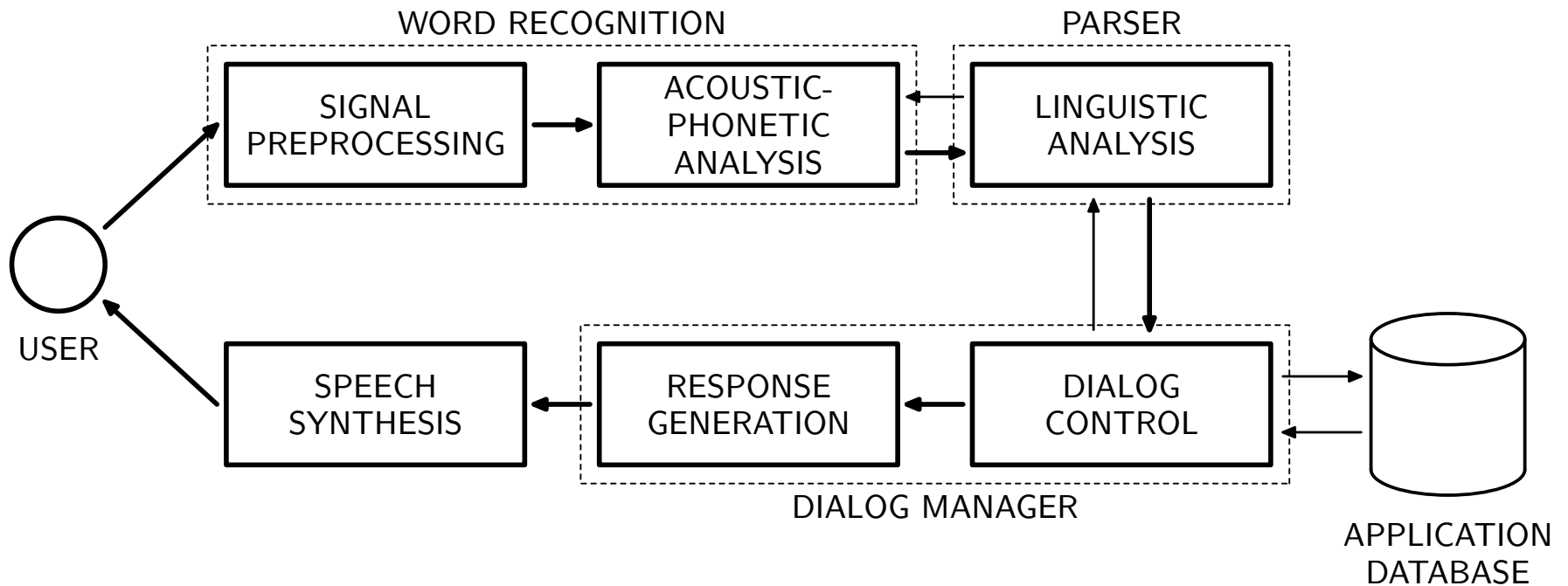
Linguistic Analysis – Unified Field Object

I want to go to Houston.



Dialogue Systems

Dialogue Systems



Dialogue Systems – Types of Dialog Managers

Example: Information retrieval IS

I'd like to travel to Houston.

- Frame based

Source	???
Destination	<i>Houston</i>
When	???

- Task oriented

Task #1: Destination (confirm)

Task #2: When? (fill)

Task #3: Source? (fill)

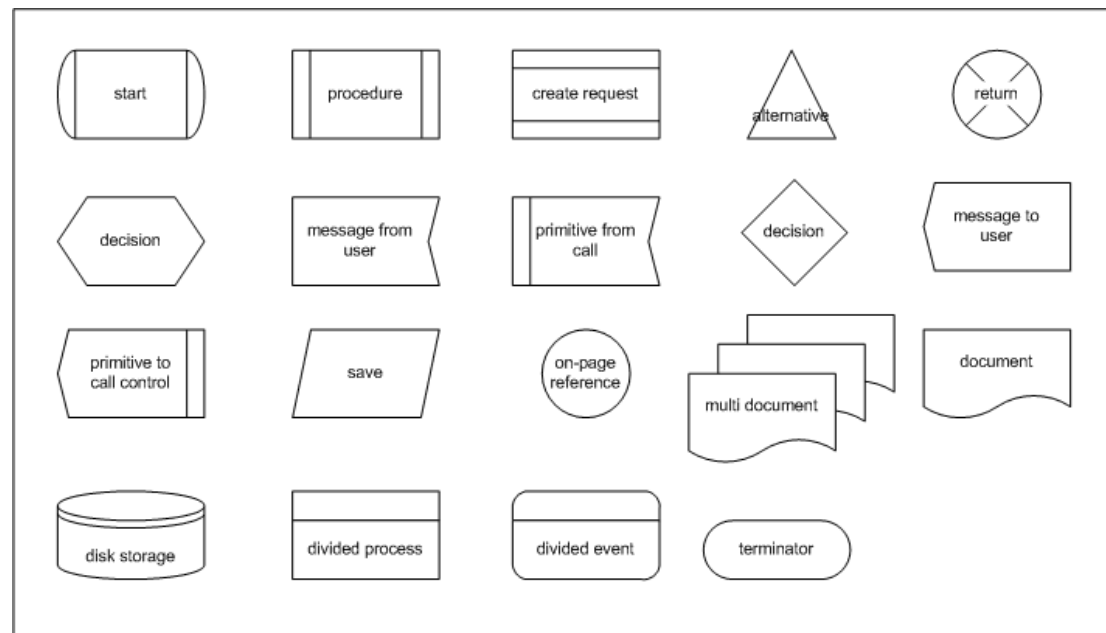
Dialogue Systems – Dialogue Modeling

- Diagrams
- User inputs (DMTF tones, speech, spelling)
 - DMTF tones
 - Speech (grammars: JSGF)
 - Spelling

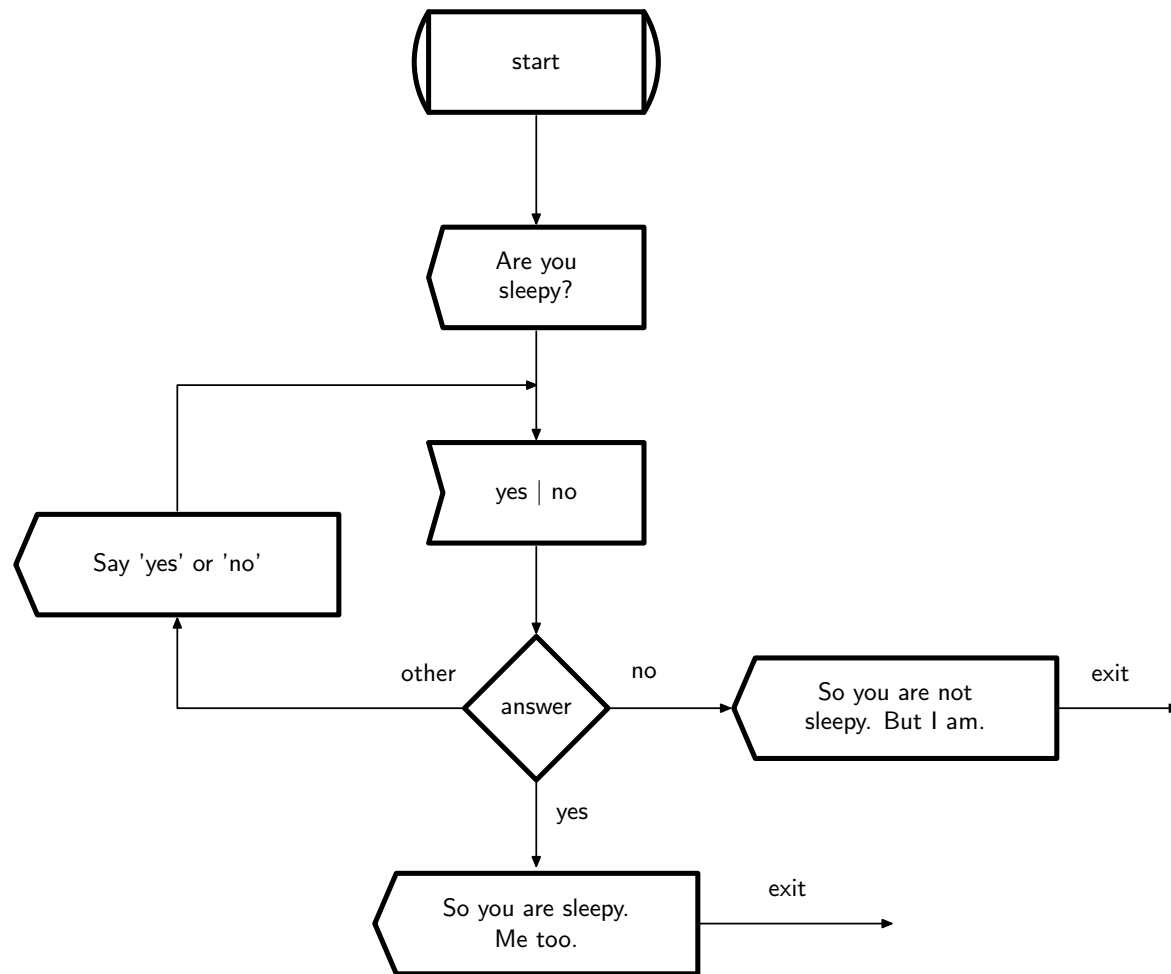
- System responses
- Transitions

Implementation

- VoiceXML
- Interpreted (Elvira)



Dialogue Systems – Dialogue Modeling



Dialogue Systems – Confirmation Strategies

Every user input must be confirmed

- Different ways how to confirm inputs
 - DMTF tones – do not need to be confirmed
 - spelling – playback the recorded signal
 - speech –
- Switching strategies according to the “quality” of user input
- After third negative confirmation → human operator
- System has to control the dialogue
 - The first initiative has to come from the system
- Mixed initiative
 - Listen to the user input – interrupt the synthesis

Dialogue Systems – Confirmation Strategies

A1: *Hello, this is IS, how can I help you?*

B1: I'd like to go from El Paso to Houston.

A2: *Do you want to go from El Paso?*

B2: Yes.

A3: *Do you want to go to Houston?*

B3: Yes.

A4: *When do you want to go from El Paso to Houston?*

B4: Tomorrow.

A5: *Do you want to go from El Paso to Houston tomorrow?*

B5: Yes.

A6: *To go from El Paso to Houston tomorrow, you need to ...*

Dialogue Systems – Confirmation Strategies

A1: *Hello, this is IS, how can I help you?*

B1: I'd like to go from El Paso to Houston.

A2: *Do you want to go from El Paso?*

B2: Yes.

A3: *Do you want to go to **Boston**?*

B3: No.

A4: *What is the destination city?*

B4: Houston.

A5: *Do you want to go to Houston?*

B5: Yes.

A6: *When do you want to go from El Paso to Houston?*

B6: Tomorrow.

A7: *Do you want to go from El Paso to Houston tomorrow?*

B7: Yes.

A8: *To go from El Paso to Houston tomorrow, you need to ...*

Dialogue Systems – Confirmation Strategies

A1: *Hello, this is IS, how can I help you?*

B1: I'd like to go from El Paso to Houston.

A2: *When do you want to go from El Paso to Houston? (confirm + add)*

B2: Tomorrow.

A3: *Do you want to go from El Paso to Houston tomorrow?*

B3: Yes.

A4: *To go from El Paso to Houston tomorrow, you need to ...*

A1: *Hello, this is IS, how can I help you?*

B1: I'd like to go from El Paso to Houston.

A2: *When do you want to go from El Paso to **Boston**? (IS error)*

B2: I want to go to Houston.

A3: *Do you want to go to Houston? (confirmation strategy switched)*

B3: Yes.

A4: *When do you want to go from El Paso to Houston?*

B4: Tomorrow.

Interesting Facts

English

- Type: Isolation (analytic)
- Phrasal verbs
(smaller vocabulary)

- Fixed word order

The dog eat the glue.

The glue eat the dog.

Japanese

- CV type language (Sayonara)
- Hundreds of syllables
- Recognition rate $\approx 97\%$

Czech

- Type: Flektion
- Declination and conjugation
(large number of word forms)

One verb in Czech has ≈ 200 forms
with different meaning

- Free word order

Pes snědl lepidlo.

Lepidlo snědl pes.