Spoken Dialog Systems: From Theory to Technology

Giuseppe Riccardi*,
Paolo Baggia^,
Pierluigi Roberti*

*AMI^2 Lab, Univ. Of Trento,
^LOQUENDO, Turin
Conversational Machines

1974
University of Padua
(De Biasi, Mian, Tisato)

2007
University of Trento
(Riccardi, Baggia, Roberti)

2037
???

Touch Tones
Synthetic Voice
System Directed
No Error Recovery
Lab Demo

Giuseppe Riccardi,
UNITN
Spoken Dialog Systems

User voice request

Automatic Speech Recognition

Spoken Language Understanding

Dialogue Management

Text-to-Speech Synthesis

Languages

Concepts

Dialogue Management

Giuseppe Riccardi,
UNITN
The Noisy Channel Model

source sentence
If music be the food of love...

NOISY CHANNEL

noisy sentence

DECODER
?Alice was beginning to get...
?Every happy family...
?In a hole in the ground...
?If music be the food of love...
?If music be the foot of dove...
...

guess at original sentence
If music be the food of love...
Automatic Speech Recognition (ASR)

Given the acoustic observation sequence $A=a_1,a_2,...,a_m$ what is the most likely word sequence $W=w_1,w_2,...,w_n$?

$$\hat{W} = \arg \max_{W} P(W | A) = \arg \max_{W} \frac{P(A | W)P(W)}{P(A)} = \arg \max_{W} P(W | A)P(W)$$
Speech Signal Features

- Compression
  - Minimize number of parameters
- Accuracy
  - Discriminate speech sounds
- Robustness
  - Speaker, noise, channel...

Feature Extraction

$s(n)$ → DFT → Mel FB → log → DCT → Feature Extraction

$\hat{W}_i$

"Yes I would like to make a reservation.."
Hidden Markov Models

• Set of states
  - $Q = q_1, q_2...q_N$; the state at time $t$ is $q_t$
• Transition probability matrix $A = \{a_{ij}\}$
  
  \[ a_{ij} = P(q_t = j \mid q_{t-1} = i) \quad 1 \leq i, j \leq N \]
• Output probability matrix $B=\{b_i(k)\}$
  
  \[ b_i(k) = P(X_t = o_k \mid q_t = i) \]
• Initial probability vector $\pi$
  
  \[ \pi_i = P(q_1 = i) \quad 1 \leq i \leq N \]

Constraints:

\[
\sum_{j=1}^{N} a_{ij} = 1; \quad 1 \leq i \leq N \quad \sum_{k=1}^{M} b_i(k) = 1 \quad \sum_{j=1}^{N} \pi_j = 1
\]
HMMs for speech
(beads-on-string model)

• HMM for word “six”
Language Modeling

• Word Sequence Probability (P(W))
  - Vocabulary V ⇒ V^n n-grams
    \[ P(W) = \prod_i P(w_i \mid w_{i-1}^{i-1}) \approx P(w_i \mid h(w_{i-n+1}^{i-1})) \approx P(w_i \mid w_{i-n+1}^{i-1}) \]
    \[ \uparrow \quad \uparrow \quad \uparrow \]
    Chain Rule Markov n-gram

• Unseen Event Estimation
  - Estimating the probability of n-grams with zero counts
    - Back-off
    \[ P_M (w_i \mid w_{i-n+1}^{i-1}) = \tilde{P}(w_i \mid w_{i-n+1}^{i-1}) + \theta(P(w_i \mid w_{i-n+1}^{i-1}))\alpha(\bullet)P_M (w_i \mid w_{i-n+2}^{i-1}) \]
What Counts as Understanding?

some notions

• We understand if we can respond appropriately
  - Question and Answer task
    • How tall is Eiffel Tower?
  - Information retrieval task
    • Can you get me the latest Bush’s speeches

• We understand if we can carry out a task
  - Drive the car from home to work
  - I want to fly from New York to Paris tomorrow business class
Example

- **User:** yes I would like to make a reservation on the afternoon flight from New York to Chicago…

- **ASR:** I’d like to make a reservation on the afternoon **fly** from Newark to Chicago

- **SLU:**
  - Action Determination (e.g. `@action=Request`)
  - Named Entities (e.g. “@city=Newark”)
  - Semantic Model associated with world model (e.g. database)
    - `@departure_city="New Work"
    - `@arrival_city="Chicago"`
Show me morning flights from Boston to SF on Tuesday

SHOW:
FLIGHTS:
ORIGIN:
  CITY: Boston
  DATE: Tuesday
  TIME: morning
DEST:
  CITY: San Francisco
## Standards for Speech Technology

World Wide Web Consortium (W3C)

### Standards Overview

<table>
<thead>
<tr>
<th>Requirements</th>
<th>VoiceXML 2.0</th>
<th>VoiceXML 2.1</th>
<th>SRGS 1.0 (speech grammars)</th>
<th>SISR 1.0 (semantic interpretation)</th>
<th>SSML 1.0 (speech synthesis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3C Recommendation</td>
<td>Proposed Recommendation</td>
<td>Candidate Recommendation</td>
<td>Last Call Working Draft</td>
<td>Working Draft</td>
<td>CCXML 1.0 (telephony, call control)</td>
</tr>
</tbody>
</table>

Giuseppe Riccardi, UNITN
Text-to-Speech Architecture

Text Analysis
- Text Normalization
- Part-of-Speech tagging
- Homonym Disambiguation

Phonetic and Prosodic Analysis
- Grapheme-to-Phoneme (LTS)
- Pitch accent assignment
- Duration computation

Waveform synthesis
- Unit (Diphone) recording and annotation
- Unit Selection and Search
- Spoken unit generation (MBROLA, PSOLA)

Giuseppe Riccardi, UNITN
Spoken Dialog Systems

DM Resources

- Dialog History
- Action Constraints
- Action Ontology
- Domain Concept Ontology

Strategy Planning

Natural Language Generator

Dialog Manager

VXML Page

- TTS
- SpeechGrammar/SLM
- ASR

W3C
Dialog Evaluation

• Measure the effectiveness of conversational Systems
• Subjective metrics
• Objective metrics
  - Word Accuracy (edit distance)
  - Concept Accuracy
    • “What is the prob correct on understanding student’s intent”
  - Task Success Rate
    • “How many times the student successfully registered for the exam”

• UNITN Help-Desk data collection
  - Students (field trial)
  - Task Success Rate 59% (unsupervised)
Research Challenges

• Automatic Speech Recognition
  - Robustness
  - Unlimited vocabulary
  - Human-Human Conversations

• Spoken Language Understanding
  - Model Generalization \(10^N\) concepts
  - Multimodal (speech, visual cues, sensors)

• Dialog Models
  - Stochastic models of human-machine interaction
  - Multimodal Input/Output

• Spoken Language Generation
  - Natural Language Generation & Text-to-Speech

Giuseppe Riccardi, UNITN
Workshop Toni Mian 2007