Speech recognition and synthesis

Introduction

Speech recognition and synthesis are most useful if combined into a full Human-Machine dialog system

- Human conversations are extremely efficient and effective interactions
- Spoken dialogs are not like a command-line Question-Answer query session
- Conversations include "control" signals at *low* (pre-verbal) and *high* levels
- Humans speak in *turns*
- In simple automated systems, interactions must be restricted and well structured

Many pictures (and their copyrights) are from [Jurafsky and Martin(2000)]

Introduction

In conversations, timing is everything

- Human dialogs are composed of game-like *moves*
- *Turn* distribution is crucial for effective Human-Machine interactions
  - *who* speaks next
  - *when* should the next speaker start
- Central to human conversations is *projection*
- *Projection* is the ability to predict the
  - *timing* of turns
  - *type* of upcoming moves

Turns

What defines a turn?

- A *single* move in the conversation "game"
- Ends with the *end* of the last utterance
- Utterance *completes* a move
- Does *not* end in a level tone
- Does not end in a *filled* pause (eg, "uuhh")
- Can be followed by a *silent pause*

The end of a turn is a *TRP*, a *Transition Relevance Place*.

Turns: TRPs

Turns and Turn taking. At each TRP of each turn:

- If during this turn the current speaker has selected *A* as the next speaker then *A* must speak next
- If the current speaker does not select the next speaker, *any* other speaker may take the next turn
- If no one else takes the next turn, the *current* speaker may take the next turn

Speech acts

Conversational *moves* are build from *speech acts*

Basic speech acts

- **Assertives:** committing Sp. to something's being the case
  suggesting, putting forward, swearing, boasting, concluding
- **Directives:** attempts by Sp. to get addressee to do something
  asking, ordering, requesting, inviting, advising, begging
- **Commissives:** committing Sp. to some future course of action
  promising, planning, vowing, betting, opposing
- **Expressives:** expressing psychological state of Sp. about state of affairs
  thanking, apologizing, welcoming, deploring
- **Declarations:** changing the world by speech
  E.g. "I resign", "You're fired"

Speech acts

Basic control tasks, handle conversation flow

- **Attention** *someone is listening*
  - Visually, by looking
  - By using *minimal responses* whenever possible
- **Acknowledgment** *move is received*
- **Grounding** *move is integrated, or not*
  - *Okay*, etc.
  - By minimal responses
  - By (partially) repeating previous move
  - By a relevant next move
- **Assessing** *move is judged*
- **Relevant move** *just start a relevant turn*
- *New turn* can subsume *Assessing* can subsume *Grounding* can subsume *Acknowledgment* can subsume *Attention*

Speech acts

Timing of responses

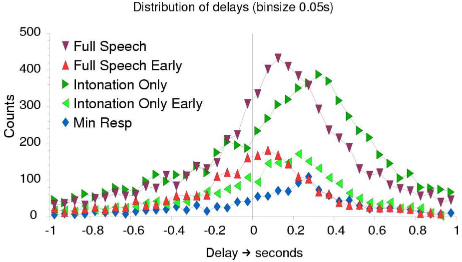
- Respond immediately
- If a *complex* response cannot be given in time, switch to a *simpler*, *faster* response type
- If all else fails, start with an *Uhhhh* placeholder
- Signal problems with a *delayed* response
- Eg, an immediate repeat signals *acknowledgment*, a delayed repeat asks for *confirmation*
- If refusal or repair is dispreferred insert *significant silence*

Minimal responses

Also: Backchannels, continuer, acknowledgment tokens

- *Uh*, *Uhm*, *HmmHmm*, *Yes*, *Sure*, etc.
- Perform the basic control tasks
- Do *not* take a turn
- Do *not* interrupt the speaker
- Are semantically, or even lexically, *empty*
- Keep the conversation going smoothly
- Without visual "feedback", eg, on the phone, a lack of audible minimal responses interrupts the conversation

Minimal responses: Timing



Natural and *elicited* minimal responses

- Responses start directly after the TRP, even for the unintelligible signals ($\approx 200ms$).
- Preparations (the *early responses*) start *before* the utterance ends

Early responses are laryngial preparation signals. *Intonation Only* responses are unintelligible *uh* sounds [Wesseling and van Son(2005)][Wesseling and Van Son(2005)]

Conversations: Implicatures

Conversations contain rules of inference

Conversational Maxims of Grice

- **Quantity**: Be *exactly* as informative as required
  - Not *less* informative
  - Not *more* informative
- **Quality**: Speak the *truth*
  - Do not say what you believe is *false*
  - Do not say that for which you lack *evidence*
- **Relevance**: Be relevant
- **Manner**: Be *perspicuous*
  - Avoid *obscurity*
  - Avoid *ambiguity*
  - Be *brief*
  - Be *orderly*

Conversations: Practical dialogs

General conversations are much too complex. Limit *Automatic Dialog Systems* to practical dialogues

Dialogues that are focused on a concrete task, eg,

- Task-oriented
- Information seeking
- Advice and tutoring
- Command and control

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Conversations: Adjacency pairs

Practical dialogues contain many controlled turn switches, called Adjacency pairs

- Question ⇒ Answer
- Proposal ⇒ Acceptance/Rejection
- Apology ⇒ Acceptance/Rejection
- Summons ⇒ Answer

Conversations: Example dialogue

C_1: . . . I need to travel in May.
A_1: And, what day in May did you want to travel?
C_2: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_2: And you're flying into what city?
C_3: Seattle.
A_3: And what time would you like to leave Pittsburgh?
C_4: Uh hmm I don't think there's many options for non-stop.
A_4: Right. There's three non-stops today.
C_5: What are they?
A_5: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C_6: OK I'll take the 5ish flight on the night before on the 11th.
A_6: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_7: OK.

Figure 19.4 Part of a conversation between a travel agent (A) and client (C).

- No real minimal responses
- *Uh Hmm* as an *Acknowledgment*
- *OK*, *Right*, and repeating dates as *Grounding*
- A lot of *Question-Answering* pairs
- A lot of *Implicatures* (licensed inferences)

Automatic Dialog System basics

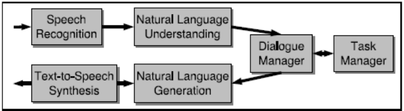


Figure 19.5 Simplified architecture of the components of a conversational agent.

Three part system

- Speech recognition and understanding
  - ASR front end with adapted language model
  - NLP back end for task related semantic parsing
- Language generation and speech synthesis
  - TTS output, can be simple phrase concatenation
  - Frame based or simple grammar sentence generator
- Dialog management
  - Task related manager
  - Task Database back-end

Recognizer

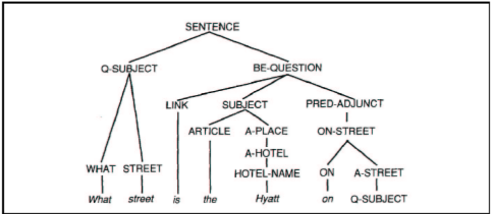


Figure 19.6 A parse of a sentence in the TINA semantic grammar (from (Seneff, 1995)). PLACEHOLDER FIGURE.

Recognizer must deliver semantic message

- Semantic context-free grammar (SCFG) for TINA
- Mixes words and concepts
- Hand written rules

[Jurafsky and Martin(2000)]

Recognizer

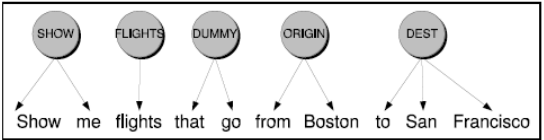


Figure 19.7 The Pieraccini et al. (1991) HMM model of semantics for filling slots in frame-based dialogue systems.

HMM concept grammar

- $\underset{C}{argmax}\ P(C|W) = \underset{C}{argmax}\ P(W|C) \cdot P(C)$
- $P(W|C) = \prod_{i=2,N} P(w_i|w_{i-N+1}, \ldots, w_{i-1}, c_i)$
- $P(C) = \prod_{i=2,M} P(c_i|c_{i-M+1}, \ldots, w_{i-1})$
- Trained on a concept-labeled corpus

[Jurafsky and Martin(2000)]

Recognizer

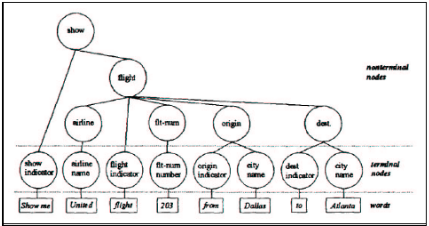


Figure 19.8 Computation of P(C) via the structure of a sentence in the Hidden Understanding Model. (from (Miller et al., 1994)) PLACEHOLDER FIGURE.

Data fragmentation problem

- Identical names can be different concepts
- Eg, cities as *origin* and *destination*
- Use a modified SCFG for $P(C)$
- Add SCFG rules for concepts, i.e. non-terminals

[Jurafsky and Martin(2000)]

Recognizer

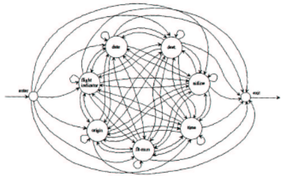


Figure 19.9 The computation of $P(C)$ from the Probabilistic RTN corresponding to the Flight concept, from (Miller et al., 1994). PLACEHOLDER FIGURE.

$P(C)$: Probabilistic finite state concept network

- Enter and Exit states
- Each arrow has a probability
- Circles indicate origin, destination, flight indicator, airline, etc.

[Jurafsky and Martin(2000)]

Speech Generator

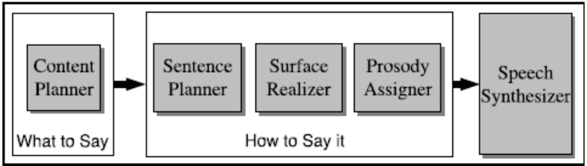


Figure 19.10 Architecture of a generator for a dialogue system, after Walker and Rambow (2002).

Concept to speech

- The database manager generates an abstract message
- Modelled into a sentence structure
- Surface form, i.e. the words, are generated
- Prosody generated from words and content,
- Fed into a TTS system

[Jurafsky and Martin(2000)]

Dialog management

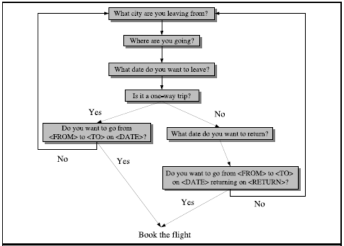


Figure 19.11 A simple finite-state automaton architecture for a dialogue manager.

Finite state automata

- Simple dialog states
- Good for form filling dialogues (frames)
- Can handle frame switching (stochastically)

[Jurafsky and Martin(2000)]

Dialog management

	Prompt Type	
Grammar	**Open**	**Directive**
Restrictive	*Doesn't make sense*	System Initiative
Non-Restrictive	User Initiative	Mixed Initiative

Figure 19.12 Operational definition of initiative, following Singh et al. (2002).

Who takes the initiative

- Machine prompts all user actions ⇒ Finite state script
- User asks questions ⇒ Single frame
- Machine allows some user initiatives ⇒ Frame switching
- Negotiation ⇒ Plan based models

[Jurafsky and Martin(2000)][Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Assignment: Week 9 Automatic evaluation of Mandarin tone pronunciation I

Compare student's pronunciation to synthetic tones generated by eSpeak. If the differences are too large, the utterance is rejected.

1. Generate reference utterance:
   `espeak -v zh "shuo1 hao3 zhong1 wen2" -w reference.wav`
2. Generate test utterance: Record it or use espeak (with errors!) and read with Praat
3. Calculate the Pitch of both test and reference utterances
4. Normalize the reference utterance to obtain the same mean and standard deviation (Hz or Semitones) as the test utterance. (reject if the standard deviation is too small). Either:
   - Use Modify→`Formula...` `(self - MeanRef)*(SDtest/SDref) + MeanTest`
   - Resynthesize the reference with the new Pitch and Standard deviation
5. Select test and the normalized reference pitches → To DTW... (fix start and end, no restrictions)
6. Query for the final distance. Reject if too large

Assignment: Week 9 Automatic evaluation of Mandarin tone pronunciation II

Make "`To DTW...`"visible for Pitch objects. Change to shown:

Praat→Preferences→Buttons...→Actions N-Z→Pitch(2): To DTW...
The values for the above procedure should be compared to the same values obtained by generating incorrect test utterances with eSpeak, eg, "shuo1 hao4 zhong1 wen2" or "shuo3 hao4 zhong1 wen4" using different speeds and pitch and compare them to the reference utterance.

Try to find out what kind of errors can be found this way using several four syllabic phrases. What are good boundaries for "bad" pronunciation? Why?

Example sentences and a translator can be found at the MDBG Chinese English dictionary
http://us.mdbg.net/chindict/chindict.php
(note that this dictionary uses a 5 to indicate the neutral tone)

Chinese examples:

shuo1 hao3 zhong1 wen2	Speak Good Chinese
bei3 jing1 da4 xue2	Beijing University
xue2 sheng1 hen3 mang2	Students are busy
chi1 he1 wan2 le4	Eat drink and be merry
qin1 peng2 hao3 you3	Friends and family
zi4 xing2 che1 sai4	Bicycle race

Assignment: Week 9 Automatic evaluation of Mandarin tone pronunciation III

Use Praat scripting to automate the above procedure. That is, from an input list of 4 syllabic Chinese (pinyin) phrases:

1. Select a phrase
2. Call eSpeak and generate the reference phrase
3. Read it and play it to the subject
4. Record, generate, or read the test phrase (last for evaluation of script)
5. Normalize the reference phrase
6. Use DTW to determine the distance
7. Give feedback
8. Clean up
9. Pause and next phrase

(see: Praat help or http://www.fon.hum.uva.nl/david/ba_spc/2008/scripting.pdf)

Further Reading I

James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent.
Toward conversational human-computer interaction.
AI Magazine, Winter, 2001.
URL http://www.cs.rochester.edu/research/cisd/pubs/2001/allen-et-al-aimag2001.pdf.

P. Boersma.
Praat, a system for doing phonetics by computer.
Glot International, 5:341–345, 2001.
URL http://www.Praat.org/.

P. Boersma and D. Weenink.
Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

FSF.
GNU General Public License.
Web, June 1991.
URL http://www.gnu.org/licenses/gpl.html.

Daniel Jurafsky and James H. Martin.
Speech and Language Processing.
Prentice-Hall, 2000.
ISBN 0-13-095069-6.
URL http://www.cs.colorado.edu/~martin/slp.html.
Updates at http://www.cs.colorado.edu/

Further Reading II

Helmer Strik, Albert Russel, Henk van den Heuvel, Catia Cucchiarini, and Lou Boves.
A spoken dialog system for the dutch public transport information service.
Int. Journal of Speech Technology, 2:121–131, 1997.
URL http://lands.let.ru.nl/literature/strik.1996.4.ps.
Link is to an older version.

W. Wesseling and R. J. J. H. van Son.
Timing of experimentally elicited minimal responses as quantitative evidence for the use of intonation in projecting TRPs.
In *Proceedings of Interspeech2005*, Lisbon, 2005.

Wieneke Wesseling and R.J.J.H. Van Son.
Early Preparation of Experimentally Elicited Minimal Responses.
In *Proceedings of SIGdial 2005*, September 2005.
URL http://www.fon.hum.uva.nl/rob/Publications/ArtikelSIGdial2005.pdf.

Appendix A

Copyright License