Text-to-speech Synthesis

Outline

- Where TTS is used
- Why do you need to know about TTS
- Types of TTS engines
- Issues in formatting output for TTS

Where is TTS used

- Spoken Dialogue Systems
  - Speaking responses, asking questions
- Reading email
- Reading webpages
- Aids to the handicapped
- TTS is used for much more than speaking isolated utterances!

TTS and SDS

- The voice of the system is the first/last thing people remember
- Choice of TTS engine depends on factors relating to SDS
- Criteria related to, but different from those used to decide on which NLG engine to use

Primary issues

- Vocabulary size
  - Is the domain large/contain many proper names?
  - Are new customers added routinely?
  - Are addresses part of what has to be spoken?
- Types of input
  - Will the input be well-formed?
  - Does it contain free-form text, with abbreviations, symbols, etc?
  - Will system need to read from email/webpages?
  - Will values from databases need to be inserted into speech?

Primary issues

- Vocabulary growth
  - Will new words need to be added to the domain (e.g., new restaurant names)?
  - Will new words conform to standard spelling conventions?
- Expressivity of voice
  - Is it important that the voice be able to convey emotion?
  - Will it be necessary to generate stress on certain words?
- Personality of voice
  - Is there a "persona" associated with the application?
  - Should the application be friendly/professional/humorous?

Types of TTS

- Formant-based
- Concatenative synthesi
- Pre-recorded speech

Formant-based synthesis

- Based on source-filter model of speech synthesis
- Sound propagation in acoustic tube provides model
- Human vocal tract is acoustic tube
- Very popular in 1980's and 1990's, largely replaced now by concatenative synthesis
- Stephen Hawking's voice

Concatenative synthesis

- Pre-recorded segments of speech are glued together (concatenated) to form entire utterance
- Made possible by advances in search/memory
- Basic unit is usually the diphone, a sequence of two adjacent phones
- Some signal manipulation needed to smooth transitions in waveform
- Larger units preferred, if available in corpus
- Approach used in most commercially available systems today


Pre-recorded speech

- What many commercial applications demand
- Most human-like, but most labor-intensive
- A form of concatenative synthesis
  - Largest possible chunks determined a priori
  - Recorded by professional speakers, often under supervision
  - New words/proper names synthesized separately and inserted

Formant based synthesis: pros

- Consistent
- Easy to modify things like pitch, phone duration
- Can make synthesis more expressive along certain dimensions
  - Contrastive stress:

    "Babbo has *very* highly rated food quality."

    "This model is preferred by customers interested in *sound* quality."
  - Question intonation:

    "And your telephone number?"



Formant-based synthesis: cons

- Can sound monotonous
- Less human-like than good concatenative synthesis
- Not personalizable
- Might be necessary if goal is unrestricted generation of text

Concatenative synthesis--pros

- When unit inventory is sufficiently large, can sound very good
- Large corpora now available for storing many units

Concatenative synthesis--cons

- Can be near-unintelligible on unusual sequences
- Difficult to modify pitch
- Significant effort necessary to record speech for new voice/language

Pre-recorded speech: pros

- Most natural sounding of all TTS methods
- Working with voice "talent" gives system developer maximum flexibility for developing unique persona

Pre-recorded speech--cons

- Exhaustive inventory of possible responses must be derived a priori
  - Limited flexibility for changing/adding prompts
  - Chicken-and-egg problem—coming up with good responses typically iterative in initial stages of development
- TTS for insertion of new words should match pre-recorded speech
- Voice talent needs to be constantly available for new prompts
- Very time-consuming

Text normalization/tokenization

- Determining individual words
- Converting abbreviations to full-form words
  - "St." → "Street" (or "Saint")
  - "MA" → "Massachusetts"
- Expanding/reordering phrases with symbols
  - "$100.00" → "one hundred dollars"
  - j.smith@dcs.shef.ac.uk → "j dot smith at d c s dot shef dot ac dot u k"
  - "mon. 8/5/06" → "Monday the eighth of may two thousand six" (or "Monday may eighth two thousand six")
- Expanding acronyms
  - "MI5" → "M I five" (but "NASA" !-> "N A S A")

Part of Speech tagging

- Mark words for basic parts of speech (typically noun, verb, adjective, preposition)
- Enables intelligent phrase boundary marking
  - "Turn left at the corner and proceed two blocks to the next light."
- Enables disambiguation of homonyms
  - "Read the instructions for the particular area in which you live."

Generating pronunciation

- Two methods:
  - Look word up in dictionary
  - Apply symbol-to-sound rules
- Original MITalk had 10K word dictionary (considered enormous)
- CMU dictionary: 127K words

Symbol-to-sound mapping

- Not always straightforward
- Main problems come in proper names (people, product names, place names, company names)
  - Slough, Worcester, Leominster
  - Diazepam, cyclobenzaprine
- Can be dependent on language of origin
  - Bertucci, Dvorak, Deng Xiaoping
- When all else fails, change the input:
  - "Samer Al-Naser" → "suhmeer al naasehr"




Prosody assignment

- Assigning intonation/pauses to utterance
- Intonation
  - Pitch (fundamental frequency)
  - Content words assigned pitch accent
  - Function words typically have no pitch accent
  - Words at phrasal boundaries typically have falling intonation
  - Compound words vs. individual words
    - "blackbird" vs. "black bird"

- Pauses
  - Mark phrases in utterance for boundaries

Prosody assignment in pre-recorded speech

- If variation is required, separate waveforms must be recorded
  - Apology #1 different from apology #2?
- Certain words (e.g., numbers) will always be concatenated
  - Each number needs rising/falling pitch contour
    - "two six four three eight two four"

Synthesizing the waveform: the final step

- Formant based:
  - Synthesize sound from underlying model of vocal tract
  - Prosody incorporated into waveform at time of synthesis
  - Computational cost in model itself

Concatenative synthesis

- Unit selection
  - Size of underlying inventory of sounds
  - Search algorithm for finding best match
    - Simple matching of diphone-to-diphone not sufficient
    - Co-articulation effects must be taken into account
- Concatenating units
  - Simply stringing units together does not quite work
  - Phase/spectral mismatch at boundaries problematic
  - Some modification to signal necessary
    - PSOLA (Pitch Synchronous Overlap and Add)
    - MBROLA (Multi-band Resynthesis Overlap and Add)

How do you evaluate TTS?

- Possible metrics:
  - Are people able to understand?
    - How to measure?
  - Do people like it?
- Other considerations:
  - Does it match the "persona" of the system?
    - A movie line might want to be friendly, a bank might want some gravitas

One final question

- After taking this course, which component of a SDS do you think is the most *interesting*? Which would you like to work on?

Prosody:
from words+phones to boundaries, accent, F0, duration

- Prosodic phrasing
  - Need to break utterances into phrases
  - Punctuation is useful, not sufficient
- Accents:
  - Predictions of accents: which syllables should be accented
  - Realization of F0 contour: given accents/tones, generate F0 contour
- Duration:
  - Predicting duration of each phone

Waveform synthesis:
from segments, f0, duration to waveform

- Collecting diphones:
  - need to record diphones in correct contexts
    - l sounds different in onset than coda, t is flapped sometimes, etc.
  - need quiet recording room, maybe EEG, etc.
  - then need to label them very very exactly
- Unit selection: how to pick the right unit? Search
- Joining the units
  - dumb (just stick'em together)
  - PSOLA (Pitch-Synchronous Overlap and Add)
  - MBROLA (Multi-band overlap and add)

TTS

Issues in TTS

- Pre-processing text
  "Drive in at the rear of the building."
  "The drive in outside of town is closed."
- Unit-based stress
  "The White House denied all charges today."
  "It's the white house at the end of the block."
- Number sequences
  $3485.00
  3485 Fairview Terrace

Issues in TTS

- Abbreviations
  - Hants.
  - NASA
  - MI5
- Proper names
  - Bertucci's
  - Deng Xiaoping