LING 120:
Language and Computers

Semester: Fall '17

Instructor: Sowmya Vajjala

Iowa State University, USA

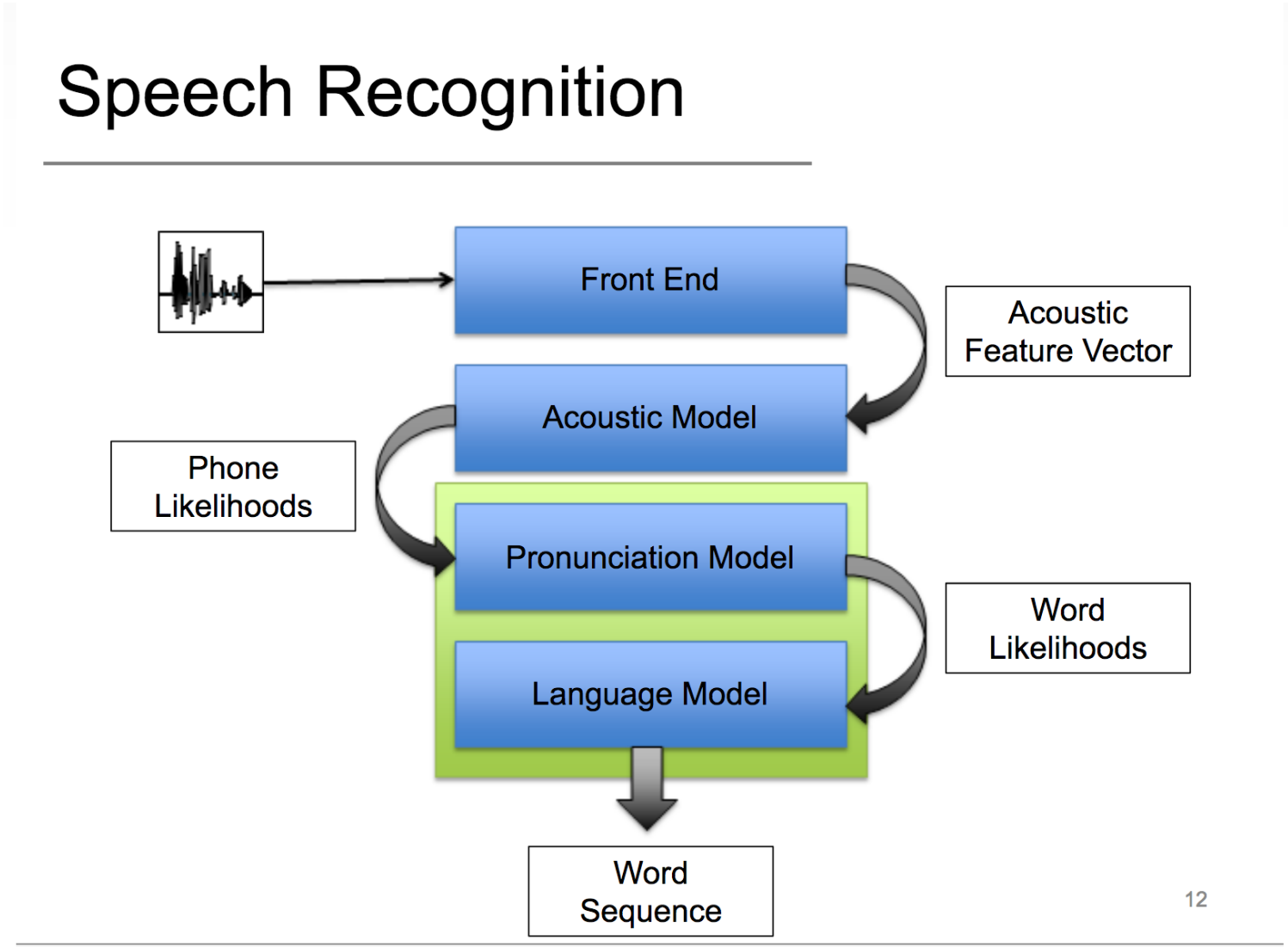
8 November 2017

Outline

- Recap of last class
- Speech Synthesis
- Group exercise on scoring speech for proficiency
- Next class: Conclusion of speech processing

Automatic Speech Recognition - Recap

ASR general process



source: Andrew Rosenburg's Speech processing lecture.

Evaluation of ASR

- Word Error Rate
- Sentence Error Rate
- In the case of dialog systems involving spoken input: concept error rate

Question from last class

- Q: Code-switching refers to people switching between languages while speaking or writing. It is not very common in US, but in multi-lingual societies, it is quite common. What or how do you think ASR systems should be tuned to such scenarios?

Question from last class

- Q: Code-switching refers to people switching between languages while speaking or writing. It is not very common in US, but in multi-lingual societies, it is quite common. What or how do you think ASR systems should be tuned to such scenarios?
- A: the answer I was expecting to see: Record code-mixed conversations and use that for acoustic-pronunciation-language models instead of using one language!

Question from last class

- Q: Code-switching refers to people switching between languages while speaking or writing. It is not very common in US, but in multi-lingual societies, it is quite common. What or how do you think ASR systems should be tuned to such scenarios?
- A: the answer I was expecting to see: Record code-mixed conversations and use that for acoustic-pronunciation-language models instead of using one language!
- Microsoft Research-India's CodeMixing project: `https://pocomixmaadi.wordpress.com/` and the post on Pronunciation modeling for Code-mixing - `https://goo.gl/ijiK9M`

ASR news from Yesterday

Identifying the songs that are playing in the background!

- http://mashable.com/2017/11/07/google-assistant-android-roll-out

Speech Synthesis

source: Speech Synthesis, Chapter 8 in Speech and Language Processing by Jurafsky and Martin, 2nd Edition.

What is Speech Synthesis?

- What is SS?

What is Speech Synthesis?

- What is SS?
- What is the difference between ASR and SS?

What is Speech Synthesis?

- What is SS?
- What is the difference between ASR and SS?
- Where is it useful?

What is Speech Synthesis?

- What is SS?
- What is the difference between ASR and SS?
- Where is it useful?
- Is it easier than ASR?

What is Speech Synthesis?

- What is SS?
- What is the difference between ASR and SS?
- Where is it useful?
- Is it easier than ASR?
- What is particularly challenging about SS?

What is Speech Synthesis?

- What is SS?
- What is the difference between ASR and SS?
- Where is it useful?
- Is it easier than ASR?
- What is particularly challenging about SS?
- What kind of resources and solutions do we need for SS?

What is Speech Synthesis?

- What is SS?
- What is the difference between ASR and SS?
- Where is it useful?
- Is it easier than ASR?
- What is particularly challenging about SS?
- What kind of resources and solutions do we need for SS?
- How would we evaluate SS?

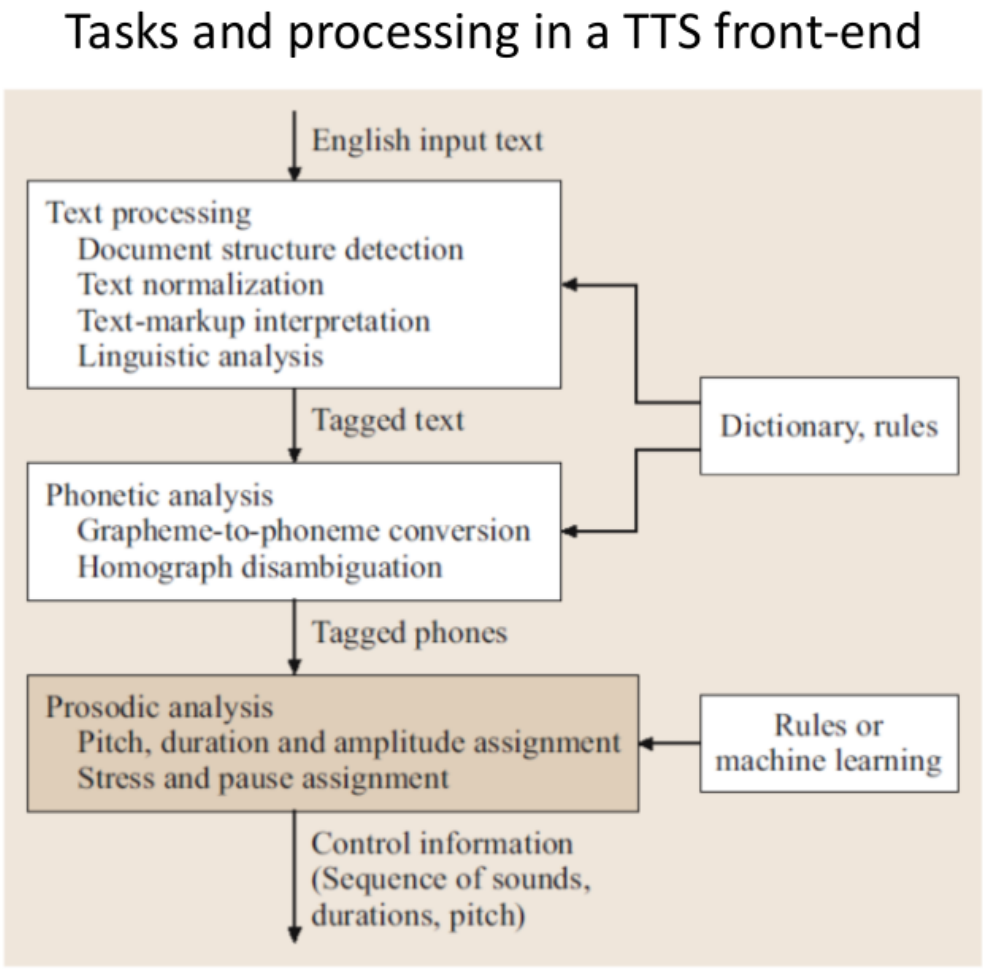
Uses of Speech Synthesis

- Dialog agents design
- Support speech impaired patients with communication aids `https://www.youtube.com/watch?v=UErbwiJH1dI` - Stephen Hawking's speech synthesizer.
- For people who cannot read (and also perhaps blind people?) - as a means to get information
- in Language tutoring (to teach how to pronounce correctly?)

Challenges for Speech Synthesis

- Breaking text down into sounds
- Mapping sound sequences to correct pronunciation (based on context)
- Getting the intonation, prosody etc right.
- Converting to a proper waveform that maps to human language words
- Not sounding too mechanical
- Some languages have straight forward pronunciation. Some languages do not.
- Some languages Tonal i.e., depending on tone, word meaning changes (e.g. Chinese)

Making of a speech synthesizer



[Schroeter, 2008, in Benesty et al., (Eds)]

Some of the important steps

- text processing: identifying sentence and word boundaries, pronunciation, abbreviations, symbols (dollar sign) etc. Converting Mr. to Mister, numbers to word form, knowing what to discard (metadata in emails etc) etc.

Some of the important steps

- text processing: identifying sentence and word boundaries, pronunciation, abbreviations, symbols (dollar sign) etc. Converting Mr. to Mister, numbers to word form, knowing what to discard (metadata in emails etc) etc.
- phonetic analysis: identifying the right pronunciation for the word by analyzing it (converting writing script to phonemes)

Some of the important steps

- text processing: identifying sentence and word boundaries, pronunciation, abbreviations, symbols (dollar sign) etc. Converting Mr. to Mister, numbers to word form, knowing what to discard (metadata in emails etc) etc.
- phonetic analysis: identifying the right pronunciation for the word by analyzing it (converting writing script to phonemes)
- prosodic analysis: taking care of intonation, emotion and rhythm in speech, identifying phrase boundaries - where to pause etc

Some of the important steps

- text processing: identifying sentence and word boundaries, pronunciation, abbreviations, symbols (dollar sign) etc. Converting Mr. to Mister, numbers to word form, knowing what to discard (metadata in emails etc) etc.
- phonetic analysis: identifying the right pronunciation for the word by analyzing it (converting writing script to phonemes)
- prosodic analysis: taking care of intonation, emotion and rhythm in speech, identifying phrase boundaries - where to pause etc
- waveform synthesis based on all the above steps

What resources do we need?

- text processing: lists of symbols, abbreviations etc, their word equivalents, sentence and word segmentation programs etc.
- phonetic analysis: pronunciation dictionaries
- prosodic analysis: pronunciation dictionaries + knowledge about pronunciation in context.
- waveform synthesis: usually needs hours and hours of recordings like in ASR (preferably from one person or two, which is not like ASR!), and some way to store the mapping between words/phrases and sounds.

There "where do I get so much data" question

https://research.googleblog.com/2015/09/crowdsourcing-text-to-speech-voice-for.html

Evaluation of Speech Synthesis

- Typically done by human listeners based on aspects such as intelligibility and quality
- intelligibility: can humans understand the machine utterances correctly?
- quality: does it sound like a natural human voice?

Evaluation of Speech Synthesis

- Typically done by human listeners based on aspects such as intelligibility and quality
- intelligibility: can humans understand the machine utterances correctly?
- quality: does it sound like a natural human voice?
- Intelligibility tests: give two similar sounding, confusable rhyming words and check for the differences in machine pronunciation
- More practical evaluation: make the synthesizer read addresses aloud, give directions etc (whatever is its purpose) and check for intelligibility with this.

Evaluation of Speech Synthesis

- Typically done by human listeners based on aspects such as intelligibility and quality
- intelligibility: can humans understand the machine utterances correctly?
- quality: does it sound like a natural human voice?
- Intelligibility tests: give two similar sounding, confusable rhyming words and check for the differences in machine pronunciation
- More practical evaluation: make the synthesizer read addresses aloud, give directions etc (whatever is its purpose) and check for intelligibility with this.
- Quality: ask several people and get a mean opinion score.

Evaluation of Speech Synthesis

- Typically done by human listeners based on aspects such as intelligibility and quality
- intelligibility: can humans understand the machine utterances correctly?
- quality: does it sound like a natural human voice?
- Intelligibility tests: give two similar sounding, confusable rhyming words and check for the differences in machine pronunciation
- More practical evaluation: make the synthesizer read addresses aloud, give directions etc (whatever is its purpose) and check for intelligibility with this.
- Quality: ask several people and get a mean opinion score.
- Comparing 2 systems: play both and ask humans which one is better

A historical perspective after this background

- Voder, early speech synthesizer from Bell Labs (1939): `https://www.youtube.com/watch?v=0rAyrmm7vv0`
- Ordering a Pizza with a computer voice (1974): `https://www.youtube.com/watch?v=94d_h_t2QAA`
- More such stuff: See the History section in Wikipedia page: `https://en.wikipedia.org/wiki/Speech_synthesis`

Speech Synthesis in Real World

- Google, Microsoft, IBM - all major IT companies have a speech synthesizer software that others can also use in their own programs.
- Google Translate, Maps have one version of this for multiple languages. (Let us take a look at translate's SS)
- Siri's responses to you won't happen without some form of synthesis
- This is not exactly synthesis - but Waze allows you to record your own voice and uses that to give directions later.
- Synthesizing your own voice back to you?: `lyrebird.ai`, `goldenspeaker.las.iastate.edu`

Attendance Exercise

Work in groups of 2–4 people

- We briefly talked about an automated scoring system in text classification class (i.e., classifying English writing into beginners, intermediate, advanced learners) like in exams like GRE, TOEFL etc.
- Scenario: Test taker gets a question, they respond with, say, a 1 minute speech on that, and you get the speech file.
- If we were to do the same kind of classification system with with these files, what do we need?
- What resources do I need for such a classifier? What kind of features should I extract? Once I get all the "features", can I use same classification algorithms and evaluation metrics as for written responses?
- Hint: We already saw speech recognition is possible even with a audio file as input (swiftscribe.ai demo)