11-737 Multilingual NLP

Speech TTS
Speech Synthesis

Carnegie Mellon University
Language Technologies Institute
Text To Speech

- Three levels
- From characters to words
  - Tokenization, dealing with non-standard words
- From Words to Pronunciations
  - Lexical look up
  - Prosody generation
- From Pronunciations to Waveforms
  - Making the actually noises play
This is a pen

My cat who lives dangerously has nine lives.

He stole $100 from the bank.

He stole 1996 cattle on 25 Nov 1996.

He stole $100 million from the bank.

It's 13 St. Andrew St. near the bank.

Its a PIII 1.5Ghz, 512MB RAM, 160Gb SATA, (no IDE) 24x cdrom and 19" LCD.

My home pgae is http://www.geocities.com/awb/.
Non-standard words

- Words without simple pronunciations
- Numbers
  - May have different pronunciations depending on context
- Dates, times etc
  - Special (common) conventions
- Abbreviations, Letter Sequences, Unknown Words
  - Dept, CIA, CASA, CDROM
- Language ID (in code mixed settings)
Pronunciations

- Need a lexicon
  - There are ways to bootstrap lexicons in under resourced languages
- Can use graphemes
  - Works well (depending on the language and amount of data you have)
  - But how to correct wrong words?
  - How do you deal with borrowed (English) words.
Prosody

● Phrasing
  ● Chunking in “breath” groups
  ● Often function/content word distinctions help
  ● \( p(B \mid \text{pos context}) = p(B) \times p(\text{pos context} \mid B) \)

● Intonation
  ● Underlying tune, may have accents on words
  ● Phrasal tune can affect people’s interpretation

● Duration
  ● From data (but hard to get really natural)
  ● Speed of delivery is both number of breaks and length of phones
Expressiveness

• Generating style
  • Not just the choice of words
  • Prosody will make interpretation different
    • Voice often sounds “professional” so it can be used in lots of situations
    • Maybe want polite, commanding, certainty
    • Probably don’t need sad, happy and angry voices

• Dialect (or register)
  • Want to voice to be appropriate for the task
  • Often the “capital”’s accent (cf. BBC English)
Waveform Generation

- Formant synthesis
- Random word/phrase concatenation
- Phone concatenation
- Diphone concatenation
- Sub-word unit selection
- Cluster based unit selection
- Statistical Parametric Synthesis
- Wavenet Neural Synthesis
Corpus Based Synthesis

- Unit selection (Hunt and Black, ICASSP 1996)
  - Select appropriate sub-word units of speech from large databases of natural speech
- Statistical Parametric Speech Synthesis (Zen et al CSL 2009)
  - (HMM-synthesis)
  - Generative models. Generate from “averaged” speech units
- Wavenet (van den Oord et al 2016)
  - Direct prediction of waveform, no explicit internal models

- All methods require “good” databases to train from
  - The right type of data (domain, style, phonetic/prosodic coverage)
Voice building tools

• Festvox Voice Building Tools (festvox.org)
  • Databases, and database design tools
  • make_nice_prompts give nice text to record
    • High-frequency words, 5-15 words, automatically from text (e.g. wikipedia/news)
  • Alignment tools for speech/text for selection of “best” examples
  • Pronunciation (UniTran) and Pronunciation LTS training scripts
  • Evaluation tools
  • Deployment through Festival and Flite (android)

• Festvox has been used for over 750 languages
  • But it's not always plug and play
Finding Data

- Audio Books/Religious Readings
  - Well recorded, single speaker, read speech
  - Requires segmentation (Interslice in Festvox)
  - But can be hard to find for lower resource languages

- Broadcast news
  - With subtitles (but typically less accurate that audio books)
  - Good speakers, good accent, read/performed speech

- Audio records from youtube
  - Multi-speaker, not well recorded, maybe not much actual speech
  - Even with subtitles (e.g. movies) there will be music in background

- ASR datasets
  - Maybe too varied, but you can sub-select for most similar speakers
  - Multi-speaker TTS databases can work, but depends on recording quality
Cross Lingual Models

• Synthesize with “nearby” language
  • Can work, especially if target language listeners are used such an accent
• Can be done naively
  • Just put the text in for the target language and use the pronunciation models
  • General Indic Synthesis sometimes do this (it sort of works)
• Improving cross-lingual models
  • Change the pronunciation models (LTS)
  • Still use the borrowed language phonemes, but have in-language pronunciation
Evaluation

- No good reliable objective measure
  - Can use a Cepstral Disortion measure against natural speech
- Human evaluation is still necessary
  - At least two dimensions
  - Preference
  - Understandability
- Blizzard Challenge (2005-)
  - From standardized databases build a voice
  - Compare the results on unseen sentences through listening tests
  - MOS, Semantically Unpredictable Sentences
- SUS are “det adj noun verb det adj noun”
  - “The sorrowful premieres sang the ostentatious gymnast”
  - Ask listeners to transcribe what they hear
  - Typically a high WER so we can distinctions quality
TTS Summary

- TTS components
  - Text analysis
  - Pronunciation and Prosody
  - Waveform generation
- With good data, building support is “quite” well defined
- Finding data and tidied it up is hard
- Evaluation requires some human listening
- Cross-lingual synthesis can work with care
TTS Discussion Point

- Zero-shot TTS: Choose a language to make a TTS engine for, and a second language to use to synthesis it. (E.g. Dutch from German)
- How much do each of the TTS components affect the target language output: text analysis, pronunciation, prosody, acoustics.
- What social/political consequences are there for a potentially “foreign” accented result.