Speech Synthesis

By

Budditha Hettige

Department of Computer Engineering
Speech Synthesis

• What is the task?
  – Generating natural sounding speech on the fly, usually from text

• What are the main difficulties?
  – What to say and how to say it

• How is it approached?
  – Two main approaches, both with pros and cons

• How good is it?
  – Excellent, almost unnoticeable at its best

• How much better could it be?
  – marginally
Architecture of TTS systems

**Text-to-phoneme module**

- Text input
  - Text in orthographic form
    - Grapheme-to-phoneme conversion
      - Orthographic rules
      - Grammar rules
      - Exceptions lexicon
  - Normalization
    - Abbreviation lexicon
  - Phoneme string

**Phoneme-to-speech module**

- Synthetic speech output
  - Phoneme string + prosodic annotation
  - Prosodic modelling
  - Prosodic model
  - Acoustic synthesis
  - Various methods
Input Type

- Concept-to-speech vs text-to-speech
- In CTS, content of message is determined from internal representation, not by reading out text
  - E.g. Database query system
  - No problem of text interpretation
Text-to-speech

• What to say: text-to-phoneme conversion is not straightforward
  – Dr Smith lives on Marine Dr in Chicago IL. He got his PhD from MIT. He earns $70,000 p.a.
  – Have toy read that book? No I’m still reading it. I live in Reading.

• How to say it: not just choice of phonemes, but allophones, coarticulation effects, as well as prosodic features (pitch, loudness, length)
Text normalization

• Any text that has a special pronunciation should be stored in a lexicon
  – Abbreviations (Mr, Dr, Rd, St, Middx)
  – Acronyms (UN but UNESCO)
  – Special symbols (&, %)
  – Particular conventions (£5, $5 million, 12°C)
  – Numbers are especially difficult
    • 1995  2001  1,995  📞236 3017  233 4488
Grapheme-to-phoneme conversion

- English spelling is complex but largely regular, other languages more (or less) so
- Gross exceptions must be in lexicon
- Lexicon or rules?
  - If look-up is quick, may as well store them
  - But you need rules anyway for unknown words
- MANY words have multiple pronunciations
  - Free variation (eg controversy, either)
  - Conditioned variation (eg record, import, weak forms)
  - Genuine homographs
Grapheme-to-phoneme conversion

- Much easier for some languages (Spanish, Italian, Welsh, Czech, Korean)
- Much harder for others (English, French)
- Especially if writing system is only partially alphabetic (Arabic, Urdu)
- Or not alphabetic at all (Chinese, Japanese)
Syntactic (etc.) analysis

• Homograph disambiguation requires syntactic analysis
  – *He makes a record of everything they record.*
  – *I read a lot. What have you read recently?*

• Analysis also essential to determine appropriate prosodic features
Architecture of TTS systems

**Text-to-phoneme module**

- Text input
  - Normalization
    - Text in orthographic form
      - Grapheme-to-phoneme conversion
        - Phoneme string
          - Orthographic rules
            - Exceptions lexicon
              - Abbreviation lexicon

**Phoneme-to-speech module**

- Synthetic speech output
  - Acoustic synthesis
    - Phoneme string + prosodic annotation
      - Prosodic modelling
        - Various methods
        - Prosodic model
        - Grammar rules
Prosody modelling

• Pitch, length, loudness
• Intonation (pitch)
  – essential to avoid monotonous robot-like voice
  – linked to basic syntax (eg statement vs question), but also to thematization (stress)
  – Pitch range is a sensitive issue
• Rhythm (length)
  – Has to do with pace (natural tendency to slow down at end of utterance)
  – Also need to pause at appropriate place
  – Linked (with pitch and loudness) to stress
Acoustic synthesis

- Alternative methods:
  - Articulatory synthesis
  - Formant synthesis
  - Concatenative synthesis
  - Unit selection synthesis
Articulatory synthesis

• Simulation of physical processes of human articulation
• Wolfgang von Kempelen (1734-1804) and others used bellows, reeds and tubes to construct mechanical speaking machines
• Modern versions simulate electronically the effect of articulator positions, vocal tract shape, etc.
• Too much like hard work
Formant synthesis

- Reproduce the relevant characteristics of the acoustic signal
- In particular, amplitude and frequency of formants
- But also other resonances and noise, eg for nasals, laterals, fricatives etc.
- Values of acoustic parameters are derived by rule from phonetic transcription
- Result is intelligible, but too “pure” and sounds synthetic
Formant synthesis

• Demo:
  – In control panel select “Speech” icon
  – Type in your text and Preview voice
  – You may have a choice of voices
Concatenative synthesis

- Concatenate segments of pre-recorded natural human speech
- Requires database of previously recorded human speech covering all the possible segments to be synthesised
- Segment might be phoneme, syllable, word, phrase, or any combination
- Or, something else more clever ...
Diphone synthesis

• Most important for natural sounding speech is to get the transitions right (allophonic variation, coarticulation effects)
• These are found at the boundary between phoneme segments
• “diphones” are fragments of speech signal cutting across phoneme boundaries
• If a language has P phones, then number of diphones is $\sim P^2$ (some combinations impossible) – eg 800 for Spanish, 1200 for French, 2500 for German)
Diphone synthesis

• Most systems use diphones because they are
  – Manageable in number
  – Can be automatically extracted from recordings of human speech
  – Capture most inter-allophonic variants

• But they do not capture all coarticulatory effects, so some systems include triphones, as well as fixed phrases and other larger units (= USS)
Concatenative synthesis

- Input is phonemic representation + prosodic features
- Diphone segments can be digitally manipulated for length, pitch and loudness
- Segment boundaries need to be smoothed to avoid distortion
Unit selection synthesis (USS)

- Same idea as concatenative synthesis, but database contains bigger variety of “units”
- Multiple examples of phonemes (under different prosodic conditions) are recorded
- Selection of appropriate unit therefore becomes more complex, as there are in the database competing candidates for selection
FreeTTS Example

• Way to use existing packages
  – Add jar library
  – Import class

```java
import com.sun.speech.freetts.FreeTTS;
import com.sun.speech.freetts.Voice;
import com.sun.speech.freetts.VoiceManager;
```
private static final String VOICENAME_kevin = "kevin16";
public static void speak(String text) {
    Voice voice;
    VoiceManager voiceManager = VoiceManager.getInstance();
    voice = voiceManager.getVoice(VOICENAME_kevin);
    voice.allocate();
    voice.speak(text);
}

public static void main(String[] args) {
    FreeTTS freeTTS = new FreeTTS();
    speak("Hello How are you");
}
Speech synthesis demo

AT&T Labs, Inc. - Research

Text-To-Speech (TTS) -- Our Demo Speaks Your Text

Home | > Demo | FAQ | Publications | Contact
Wizzard Software | AT&T Natural Voices
TV Commercial

STEP 1  Voice & Language: Audrey .... UK English

STEP 2  Text: [ Selected language only | 300 character limit | Help with UTF-8 or Latin-1 ]
I hope you are enjoying this course unit, L E L A 3 U 4 3 1

STEP 3  Click: SPEAK - or - DOWNLOAD [ restrictions apply~ ]
Questions

Answer the following questions.

1. Draw an architecture of a TTS system.
2. Briefly explain three issues on grapheme-to-phoneme conversion.
3. Compare and contrast articular synthesis and concatenative synthesis.

Identify which of the following is/are true? In each case, justify your answer.

1. ’Diphones’ are fragments of speech signal cutting across phoneme boundaries.
2. High frequency sounds are less informative for ASR.
3. Speaker-independent systems require ’training’ to ’teach’ your individual idiosyncracies
4. to the system.