11-823: Conlanging

Building a Talking Clock
Festival Speech Synthesis System

http://festvox.org/festival
General system for multi-lingual TTS
C/C++ code with Scheme scripting language
General replaceable modules
  lexicons, LTS, duration, intonation, phrasing,
  POS tagging tokenizing, diphone/unit selection
General Tools
  intonation analysis (F0, Tilt), signal processing
  CART building, n-grams, SCFG, WFST, OLS
No fixed theories
New languages without new C++ code
Multiplatform (Unix, Windows, OSX)
Full sources in distribution
Free Software
http://festvox.org
“I want it to speak like me!”
- Festival is an engine, how do you make voices
- Building Synthetic Voices
  - Tools, scripts, documentation
  - Discussion and examples for building voices
  - Example voice databases
  - Step by Step walkthroughs of processes
- Support for English and other languages
- Support for different waveform techniques:
  - diphone, unit selection, limit domain, HMM
- Other support: lexicon, prosody, text analysers
The CMU Flite project

http://cmuflite.org

“But I want it to run on my phone!”

- FLITE a fast, small, portable run-time synthesizer
- C based (no loaded files)
- Basic FestVox voices compiled into C/data
- Thread safe
- Suitable for embedded devices
  - Ipaq, Linux, WinCE, PalmOS, Symbian
- Scalable:
  - quality/size/speed trade offs
  - frequency based lexicon pruning
- Sizes:
  - 2.4Meg footprint (code+data+runtime RAM)
  - < 0.025 secs “time-to-speak”
Corpus-based Speech Synthesis

- Given natural speech recordings
  - Label the phones/words
- Reconcantenate the units to form new words

Unit Selection Synthesis
  Find “segments” and select appropriate ones

Statistical Parametric Synthesis
  Average multiple examples and generate

Neural Network
  Use neural networks
  Learn mapping from text/phones to audio
Overview

- Design your prompts
  - Test them
- Define your word pronunciations
- Define your phone set
- Setup the voice
- Record the prompts
- Build unit selection voice
  - Find phone alignments
  - Extra parameters
  - Build clusters
- Test it
Designing your prompts

- What will it say:
  - “The time is now, about five past one in the morning”
- Generate 12 or 24 utterances from a basic template
- Carrier sentences are good
  - Makes speaker speak better
  - Makes listener adapt before key information
Designing your Prompts

- Design your carrier phrase
- Plug in each of your actual values
- Don't try to minimize the recordings
  - Better to have word examples multiple times
- Should have word coverage
  - Basic techniques won't allow synthesis of new conjugations
The language Eth

- **Endonym:** eǒ
- **Spoken in the frozen north of Europe 5000 years ago, around the North Sea**
- **By coincidence it's completely understandable by modern Japanese speakers.**
Prompts

- Taidaima, ichi ji go pun gurai, go zen desu.
- Now 1 hour 5 min about, m before copula
- Have initial start (always the same)
- Give time in 5 minute intervals
- Identify before and after noon
Pronunciations

- **Nana** (seven) noun ((N A) 0) ((N A) 0)
- **Hachi** (eight) noun ((H A) 0) ((CH I) 0)
- **Go** (five) noun ((G O) 0))
- **Go** (meridian) noun ((G O) 0))
- ...

...
Phone Defs

- Name clst vc vlng vheight vfront vrnd
  ctype cplace cvox asp nuk
- (A - + l 3 3 – 0 0 0 - -)
- (K - - 0 0 0 – s v - - -)
- (G - - 0 0 0 – s v + - -)
- ...

RAW TEXT START

Phone Defs

- Name clst vc vlng vheight vfront vrnd
  ctype cplace cvox asp nuk
- (A - + l 3 3 – 0 0 0 - -)
- (K - - 0 0 0 – s v - - -)
- (G - - 0 0 0 – s v + - -)
- ...

RAW TEXT END
Preliminaries

- `export ESTDIR=/home/awb/speech_tools`
- `export FESTVOXDIR=/home/awb/festvox/`

- `mkdir eth_clock`
- `cd eth_clock`
- `$ESTDIR/src/unitsel/setup_clunits cmu eth awb`
Language dependencies

- Copy your prompt list to etc/txt.done.data
  ( time_0001 “Taidaima, ….” )
  ( time_0002 “Taidaima, ….” )
- Add your lexical entries to
  festvox/cmu_eth_awb_lexicon.scm
- Add your phoneset definitions to
  festvox/cmu_eth_awb_phoneset.scm
- Map your phoneset to English in
  festvox/cmu_eth_awb_lexicon.scm
- Add your phoneset to festival/clunits/all.desc
My language isn't English, this can't be done!

– Yes it can!

We do this to allow automatic phone labeling

A (bad) rendering of English phones will match your actual phone list (really it will)

Vowels more like Vowels, than Consonants

Consonants more like Consonants than Vowels

KH A P L A Q

k aa p l aa pau
Dynamic Time Warping

- We have synthesized prompts
  - With phone labels
- We have recorded prompts
  - Without phone labels
- We can align the two prompts
  - Then map synth labels to recorded labels
Dynamic Time Warping

Template

Sample Speech
DTW algorithm

For each square

\[ \text{Dist}(\text{template}[i], \text{sample}[j]) + \]

\text{smallest_of (Dist}(\text{template}[i-1], \text{sample}[j]) \]

\[ \text{Dist}(\text{template}[i], \text{sample}[j-1]) \]

\[ \text{Dist}(\text{template}[i-1], \text{sample}[j-1]) \]

Remember which choice your took (count path)
Build Cross Lingual Prompts

- ./bin/do_build build_prompts_waves
  - Synthesized into prompt-wav/*.wav
  - Labels in prompt-lab/*.lab

- Play these waveforms to check them
- Look at the prompt-lab/*.lab files
Record the Prompts

- ./bin/promp_them etc/txt.done.data
  - Displays text, plays prompt
  - Records for the right amount of time
  - But it won’t work for you

- Use audacity
  - Record each prompt
  - Export them as 16KHz mono riff
  - Put them in recording/*.wav
  - ./bin/get_wavs recording/*.wav

- Take care to get them right
  - Minimize silence at beginning and end
Align with DTW

- ./bin/make_labs prompt-wav/*\.wav
  - Produces lab/*\.lab
  - Check them (by hand)
  - Use wavesurfer to view them

- ./bin/do_build build_utts
  - Build the utterance structure
  - Words/Syls/Segments/Duration etc
Automatic Labeling
Parameterization and Build

- ./bin/do_build do_pm
  - Find pitch periods (glottal closure)

- ./bin/do_build do_mcep
  - Find spectral properties
  - At each pitch period

- ./bin/do_build build_clunits
  - Build unit selection synthesizer
  - Find clusters of similar phones
Running the Voice

- festival festvox/cmu_eth_awb_clunits.scm

... 

festival> (voice_cmu_eth_awb_clunits)

...

festival> (SayText “Tadaima, ...”)

...

festival> (set! utt1 (SayText “Tadaima, ...”))

...

festival> (utt.save.wave utt1 “eth_11:30.wav”)
Issues

- Recordings aren't right
  - Too much silence
  - Wrong format
- Alignment doesn't work
  - English mapping to confusable
- Something else
  - You are building a new language
  - Maybe there is a new challenge
- Ask if you get stuck
  - Package up the whole voice directory
- See class website for long details of build
Homework for Part 1

Submitted by email by noon to awb@cs.cmu.edu and lsl@cs.cmu.edu, with 11-823 in the subject

- Name of your language
- Short background about your language
- List of prompts you will record
- List of phonemes you will use
- List of word pronunciations
- Write up with gloss of prompt(s) and explanation of other decisions you have made
Homework for Part 2

Submitted by email by noon to awb@cs.cmu.edu and lsl@cs.cmu.edu, with 11-823 in the subject

- Name of your language
- Short update about your language
- Final list of prompts you record
- Tar/zip version of whole voice directory
- At least 2 synthesize novel examples
- If possible something that didn't work
Function to map 24hr clock to your textual description

- 03:14 → “the time is now almost quarter past three in the morning”
- This can be done in Festival (or any other programming language and have it call Festival to generate the waveform file)