DATA COLLECTION & PREPARATION FOR SPEECH SYSTEMS

Chevy Levitan
Mentor: Erica Cooper
Director: Dr.Julia Hirschberg

OBJECTIVE

Gather and process data for global speech technologies.

PROJECTS

I. **ENGLISH -> TTS**

II. **LOW-RESOURCE LANGUAGES -> KEYWORD SEARCHING**

- Background
- Methods
- Status
- Future work

TTS >> BACKGROUND

- About

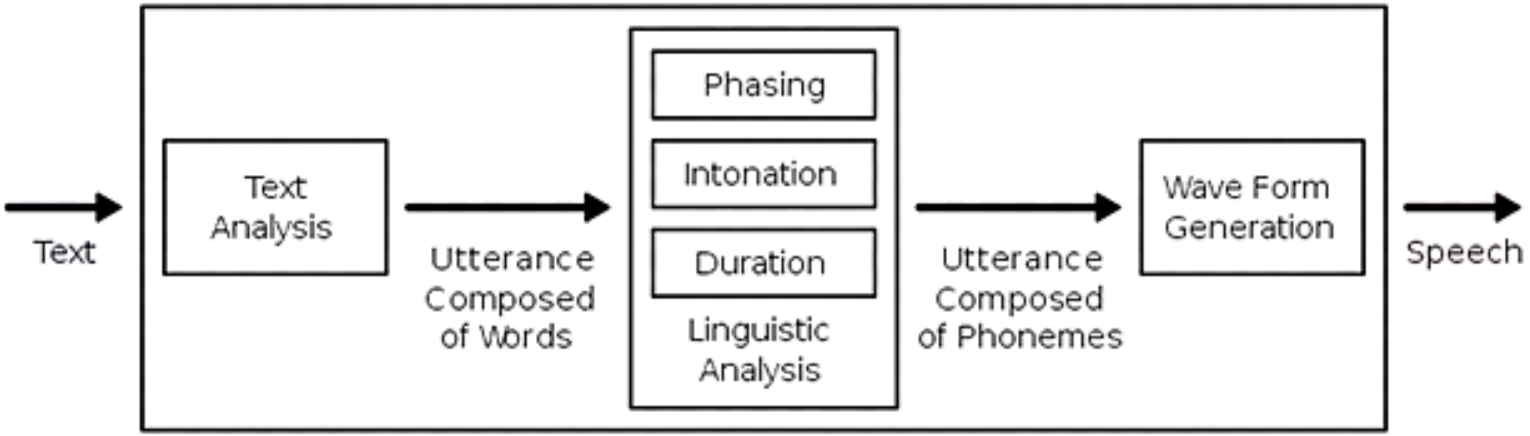
Method	Description	Pros	Cons
Concatenative	form words by stringing together small units of speech	natural sounding, easy to implement	expensive, rigid, large databases
HMM-based	generate waveforms from HMM's	context-dependent, flexible, smaller databases, robust	sounds synthetic

TTS >> BACKGROUND

- Applications
  - assistive technology
    - blind
    - speech impaired
  - phones
    - caller id
    - driving settings

TTS >> BACKGROUND

- Process



TTS >> BACKGROUND

Boston Radio Corpus:

- Designed for TTS
- 7 speakers
- 7+ hours of clean audio
- Transcriptions

TTS >> METHODS

Paragraph -> Sentence:

- Each training segment should be smaller
- Split text and audio
- Each sentence is identified by its speaker and a number (ex: f1a_0001.txt)

TTS >> METHODS

Paragraph -> Sentence:

- **Text**
  a. find (‘.’) in paragraph
  b. list of rules for abbreviations
  c. send each sentence to its own .txt file

- **Audio**
  a. find (‘.’) in .txt file
  b. look up timing in .wrd file for the following word
  c. trim the audio (sox)
     (ex: sox src dest start dur)

TTS >> METHODS

HTS-Speaker Adaptive Demo:

- Install demo
- Configure with default parameters
- Configure with our data

TTS >> STATUS

HTS-Speaker Adaptive Demo:

✓ Install demo

✓ Configure with default parameters

→ Configure with our data

KS >> BACKGROUND

Low-resource Languages:

- Languages that have limited tools at their disposal
- English is high-resource; TTS, ASR…
- Need data to build resources

KS >> BACKGROUND

- Where can we find lots of audio and text data for low-resource languages??

- Internet
  → Free
  → Accessible
  → Global

KS >> BACKGROUND

PROBLEM:

photos, logos, animations, advertisements...

KS >> BACKGROUND

SOLUTION:

BEAUTIFUL SOUP.

KS >> METHODS

- ❑ Select language
- ❑ Find useful websites
- ❑ Scrape

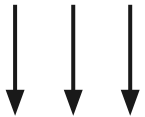
KS >> METHODS

- ✓ Language Telugu
- ✓ Blogs
  1. http://mahojas.blogspot.com/
  2. http://yaramana.blogspot.com/
  3. http://ishtapadi.blogspot.com/
- ✓ Scrape

KS >> METHODS

EXAMPLE :

http://mahojas.blogspot.com/



text sample:

తెలంగాణతో మన భావిసంబంధాలు ఎలా ఉండాలనేది నిర్ణయించుకునే ముందు కొన్ని నిర్దిష్టమైన క్రైటీరియా దృష్టిలో ఉంచుకోవాలి.
౧. వాళ్లు ప్రస్తుతం వ్యవహరిస్తూన్న విధానం, లేదా దానికి మనమిప్పుడు ఫీలవుతున్న ప్రతిస్పందన - ఇవి మన వ్యవహరణ పాలసీకి ప్రాతిపదికలు కాకూడదు. అలా చేస్తే అది మన ఎమోషన్సుని భావితరాలకి బదిలీ చేయడమవుతుంది. అలాంటిది ఎక్కడా ఏ దేశానికీ మేలు చేయలేదు. ఎమోషన్ల బదిలీ మూలాన - మిగలడానికి అవకాశమున్న కొన్ని వంతెనలు కూడా తగలబెట్టబడతాయి.

KS >> STATUS

- Languages: Telugu, Lithuanian
- Scraped ~500 web pages
- Word count: > 100,000

FUTURE WORK

- Data selection
- Audio scraping
- Scrape other languages
  - → Tok pisin
  - → Cebuano
  - → Kurmanji kurdish
  - → Kazakh
- Build synthesizer for low-resource languages

THANK YOU!