Speech resources and tools

Jonas Lindh
http://www.ling.gu.se/~jonas
Outline

- Types of spoken language data
- Why do we need speech resources?
  - Most common types
    - Telephone and other formats...
- What are speech resources used for?
  - Phonetics
  - Behavioral research
- Labeling the speech data
- Speech tools
  - Speech recognition tools
  - TTS tools
  - Speaker Recognition tools
- Speech Corpora
- Questions
What are speech resources

- Recorded speech
- Tools for handling the resources
- Text? Transcriptions?
- Video?
Types of spoken language data

- Read
  - Formal
- Spontaneous
  - Casual
- Monologue or dialogue (or more)
- “Standard” or dialect
- What kind of data do we want?
Why do we need speech resources?

- Speech recognition systems
- Text to speech systems
- Speaker verification systems
- Dialogue systems
- Research...
Most common types

• Read speech
  • usually consists of a word list, a list of sentences or a piece of text ("north wind and the sun").
  • Specific research goal or general database.
  • Tv and/or radio broadcasts (legal etc issues)

• Read speech tends to have the character of a formal monologue performed in a standard dialect
• Speech recognition – dictation (cover variations?)
Most common types

• Dialogue between two or more speakers
  • unscripted speech in almost all cases (acting).
  • ranges between entirely free discussions (less common) and a various types of task-oriented dialogue
    – Calendar task
    – Map task
  • Domain specific or free dialogue...
  • Dialogue systems... (strategies)
Telephone and other formats...

- The same can be recorded over telephone...
  - What happens to recorded speech over landline/gsm etc?
  - Skype?
  - Flight cockpit?
  - Bugging...

- What are we going to use it for?
What are speech resources used for?

- Discourse analysis
  - The grammar/syntax of unscripted spoken language is maybe different from the written?
- Disfluencies
  - “Disturbances” in speech... how, why, can we use it or how can we ignore it (speech recognition/dialogue)?
  - Pausing
- Turn taking
  - An important aspect of the interaction between speakers is turn-taking, i.e. signaling that one has spoken and that it is someone else’s turn to speak
What are speech resources used for?

- Discourse analysis
- Openers (ways to start a conversation).
- Conflict management (how to manage “conflicts” in conversations).
- Indicators of agreement or disagreement.
- etc...
Phonetics

- Segmental level

  - Properties of single speech sounds (monophones)
    - Some ASR system use monophone models (most triphone)

- Phonological vs phonetic realization
  - How much pronunciation can we predict?
Phonetics

- Prosodic level
  - Stress on word and sentence level
    - Which syllable in the word is stressed and how?
    - Which word in the sentence is focused and how?
  - Intonation
    - Lexical tones
    - Suprasegmental “carrier”, statement vs question?
Phonetics

- Duration (quantity)
  - Inherent duration on segmental level /i/-/a/
  - Syllable duration vs articulation rate
Behavioral research

• Multi-modal databases
  • Audio-video for speech and gestures together with other tagging

• Psychological research on emotion

• Psychological research on effect of drug and/or alcohol use

• Sign language?
Labeling the speech data

• Type is dependent of main objective
  
  • Phonological/phonemic transcription
  
  • Phonetic transcription
    – Broad / narrow
  
  • Orthographic transcription
Labeling the speech data

• Phonetic labeling
  • IPA [http://web.uvic.ca/ling/resources/ipa/charts/IPAlab/IPAlab.htm](http://web.uvic.ca/ling/resources/ipa/charts/IPAlab/IPAlab.htm)
  • Segmental (some prosodic info)
    - Narrow
  • aims to maximize the faithfulness to the phonetic signal...
• However, subjective judgements from the labeler..
Labeling the speech data

• Phonetic labeling
  • IPA  http://web.uvic.ca/ling/resources/ipa/charts/IPAlab/IPAlab.htm
  • includes characters that have not been accessible in computer systems... (ASCII standard)
  • The Unicode standard is aimed at expanding the limits of 8-bit encoding...
  • Unicode defines a set of characters aimed to encompass all the worlds writing systems – at present it has a repertoire of almost 110,000 characters...
Labeling the speech data

- Historically databases have usually been made using 7-bit ASCII... still problems...
- Representing IPA symbols in 7-bit ASCII.
Labeling the speech data

- The subjective nature of phonetic labeling along with the number of symbol choices available makes it very difficult to perform searches in phonetically narrow labelled data...

- A speech database with phonetic labeling requires a parallel phonemic or orthographic representation to enhance search-ability...
Labeling the speech data

- **Praat** – TextGrid becoming standard
  
  [http://www.fon.hum.uva.nl/praat/](http://www.fon.hum.uva.nl/praat/)

- **Wavesurfer** [http://www.speech.kth.se/wavesurfer/](http://www.speech.kth.se/wavesurfer/)

- **Transcriber**

- **...and there are many others free and commercial...**
Speech recognition tools

- Language Model – Text $P(W_i)$

- Word, commands, keywords, dictation, dialogue or spontaneous speech

- Speaker dependent vs independent?

- Training and test database
Speech recognition tools

![Word Error Rate vs. Level Of Difficulty](image)
Speech recognition tools

- CMU Sphinx  http://cmusphinx.sourceforge.net/
  - JAVA  http://cmusphinx.sourceforge.net/sphinx4/

- HTK Toolkit  http://htk.eng.cam.ac.uk/


- CSLU toolkit  http://www.cslu.ogi.edu/toolkit/
Speech recognition tools

• Acoustic Modeling

• SpeechDat [http://www.speechdat.org/]
  • ex. Swedish 6000 speaker (5000 landline, 1000 mobile)
  • Dialect, age, gender tagging
  • Documentation [http://www.docstoc.com/docs/78265921/LRE-63314-SPEECHDAT]
  • More resources in the end...
TTS tools

- G2P (Grapheme to Phoneme converter)
- Is it going to say a few words, limited domain or anything...
  - Articulatory synthesis [http://www.haskins.yale.edu/facilities/asy-demo.html](http://www.haskins.yale.edu/facilities/asy-demo.html)
  - Formant synthesizer [http://www.speech.kth.se/wavesurfer/formant/](http://www.speech.kth.se/wavesurfer/formant/)
    - $P(f) = S(f) \times T(f) \times R(f)$ (source, filter and radiation)
- Concatenation
  - Simple
  - Unit selection
TTS tools

● Speaker dependent!

● Database dependent on type, most common unit selection using hours of speech...

● Test texts/Evaluation
  - Mean Opinion Score (MOS http://en.wikipedia.org/wiki/Mean_opinion_score )
  - Preference Test (system vs system or system vs human)

● Diphone most common back-off...
TTS tools


- **Festival** [http://www.cstr.ed.ac.uk/projects/festival/](http://www.cstr.ed.ac.uk/projects/festival/)

- **Festvox** [http://www.festvox.org/](http://www.festvox.org/)

- **BOSS** [http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/boss](http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/boss)
Speaker Recognition tools

• Much in common with speech recognition.

• Voices for training (sometimes on specific phrases – using speaker dependent verification).

• Several sessions.

• Different recording equipment.
Speaker Recognition tools

- MASV (Based on HTK extension)  http://www.bas.uni-muenchen.de/Bas/SV/

- Alize (former Mistral)  http://alize.univ-avignon.fr/index_en.html

- Data for training and testing...
More Speech resources

- Linguistic Data Consortium
  http://www.ldc.upenn.edu/Catalog/byType.jsp

- Most downloaded
  http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

- NIST
  http://nist.gov/itl/iad/mig/

- Swedia dialect database
  http://swedia.ling.gu.se/
More Speech resources

- Speaker recognition
- Training and test data NIST
  http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2011S05
Resources?

- Search for resources and tools for your native language...

  - http://www.ldc.upenn.edu/Catalog/catalogSearch.jsp
Questions

• Build groups (smallest 2, biggest 4)

• Are there any resources in your native language?
• Come up with a major challenge for speech technology research and development regarding resources and tools.
• Can you think of any questions you have or something you have not understood?