Lecture 1: Speech Data, Transcriptions, Transcription Tools, and Dictionaries

Lecturer: Mark Hasegawa-Johnson (jhasegaw@uiuc.edu)
TA: Sarah Borys (sborys@uiuc.edu)

May 16, 2005

1 Speech Corpora

Speech corpora used in this course will include some of the following.

Name	Style	BW	Hrs	Spkrs	Transcriptions
TIMIT	Read	7kHz	14	640	TXT, WRD, PHN
NTIMIT	Read	3.5kHz	14	640	TXT, WRD, PHN
WS97	Conversation	3.5kHz	3.5	2283	TXT, WRD, PHN, Stress, TOBI
Switchboard 1	Conversation	3.5kHz	349	4188	TXT, WRD
Broadcast News	TV	7kHz	Big	Big	TXT
Radio Speech	Read	7kHz	3.5	7	TXT, WRD, TON, BRK, LBL=PHN
AVICAR	Read	7kHz	50	100	TXT

1. TIMIT (/workspace/timit) [10]. Transcriptions: WRD (manual word boundary transcription), PHN (manual ARPABET coarse-grained allophonic transcription). SX sentences are "phonetically balanced," i.e., designed so that they cover all of the possible triphone combinations of English. each word). This corpus fits onto 1 CD, is better transcribed than most corpora, is universally cited, and is only $100 from LDC (http://www.ldc.upenn.edu), making it one of the best available bargains in the world of transcribed speech corpora.

2. NTIMIT (/workspace/fluffy1/ntimit-train, /workspace/fluffy1/ntimit-test) [5, 7] is the TIMIT corpus passed through real-world telephone channels: low bandwidth, signaling tones, echo, acoustic noise, electrical noise. Sampling rate of the corpus distributed by LDC is still 16kHz, but bandwidth is 3.5kHz (telephone-band), therefore we usually downsample to $F_s = 8$kHz.

3. WS97 (/workspace/fluffy1/train-ws97, train-ws96, misc-ws97): Phonetic transcriptions are more detailed than TIMIT. Each phone is transcribed using ARPABET (like TIMIT), then annotated using the most perceptually salient modifier, where modifiers may include manner change (e.g. fricated stops), nasalization, or creak. Prosodic prominence is transcribed using a highly non-standard perceptually based notation [4, 3]. About 200 utterances have also been TOBI-transcribed at UIUC [9]. Because it's conversational speech, coverage of phoneme combinations is MUCH worse than TIMIT. For the same reason, it is quite common to find really dramatic allophonic reduction and assimilation phenomena (e.g., "I don't know" produced as a single nasalized vowel). Transcriptions are available from http://www.icsi.berkeley.edu/real/stp/. Speech data are available by request, from steveng@cogsci.berkeley.edu, for the cost of shipping. Although these speech data are excerpted from the Switchboard 1 corpus, the 1998 Switchboard re-segmentation project [1] eliminated the possibility of time aligning the two corpora.

4. Switchboard 1 (/workspace/fluffy1/switchboard-1) [2]. Speech data is distributed in a difficult format (SPHERE two channel format, samples alternating, each sample encoded in 8-bit mu-law). Microsoft WAV files are available on IFP network in wav_chopped. Word boundary time transcriptions are available for free from Mississippi State (http://www.isip.msstate.edu) [1]. Speech data is available

from LDC for about $2000. More recent, larger telephone speech corpora available from LDC include Switchboard 2 (released in 3 sub-corpora, 1999-2001) and the Fisher corpus (about 3000 hours; first subcorpus release was 2005); utterance-level transcriptions are available for these corpora, but not word boundary times.

5. Broadcast News (/workspace/mickey1/BN97) includes announcer speech (read), on-air conversations, on-the-street interviews, and telephone speech. Data includes many hours for a few speakers (announcers), and small segments from a large number of speakers (interviewees, etc.). Transcriptions do not include time alignments of the beginning and ending of every word; alignment times are only marked once per "utterance unit."

6. Radio Speech Corpus (/workspace/Radio_Speech_Corpus) [8]. Possibly the largest ToBI-transcribed corpus in English; transcriptions include TON (ToBI tone labels) and BRK (ToBI break indices). WRD files contain time-aligned word transcriptions, and POS files contain time-aligned part of speech transcriptions. Also includes automatic phoneme transcriptions (ALA, LBA) and some files have manually aligned phoneme transcriptions (ALN, LBL).

7. AVICAR [6] (/workspace/mickey1/AVICAR_DIST) includes 7-channel audio and 4-channel video recordings of 100 talkers reading digits, telephone numbers, letters, and TIMIT sentences in a moving car.

2 File Formats

2.1 Speech File Formats

1. SPHERE.

   - Where it is used: LDC distributions, data provided to participants in competitions sponsored by NIST (National Institute of Standards, http://www.nist.gov/speech).
   - File extension: In TIMIT, the files are called WAV, but are actually in SPHERE format; other data in SPHERE format usually has the SPH file extension.
   - Header: A SPHERE header has a length of exactly 1024 bytes, in ASCII (type `man ascii`). In order to read the header of file SX43.SPH, you can just type `more SX43.SPH`.
   - Data: Speech data can be encoded in any number of channels, any number of bytes per sample, any sampling rate, and any encoding scheme: check the file header. TIMIT and NTIMIT are encoded as little-endian short integers, so that they can be easily read using matlab or (on a little-endian architecture, e.g., Intel) using C `fread`. Switchboard is encoded using 8-bit mu-law encoded samples, alternating between channel A and channel B of the conversation.
   - Useful tools: SPHERE 2.6 (http://www.nist.gov/speech) includes C functions for reading and writing SPHERE.

2. WAV/RIFF. Microsoft WAV format is pretty much an internet standard, and can be read directly into matlab or most other tools. File extension is usually `WAV`.

2.2 Transcription File Formats

ALL important transcription formats are formatted plaintext. If you are unsure what type of file you are dealing with (or what sub-format it has, in the case of SPHERE files), use `more` or `less` to look at it.

1. SPHERE/NIST

   - Where it is used: LDC distributions, NIST competitions
   - File Extension: Variable. Utterance-level segmentation is contained in .TXT (TIMIT) or × -trans.text (Switchboard) files. Word-level segmentation is contained in .WRD (TIMIT) or × -word.text (Switchboard). Phone-level segmentation may be in PHN files.

- Header: usually none.
- Data: SPHERE format transcriptions consist of one segment per line. Each line matches the following pattern:

  `[CORPUS\_NAME] START\_TIME END\_TIME LABEL1 [LABEL2 ...] \# [COMMENT]`

  CORPUS_NAME is optional; if present, it must be a string containing no spaces. START_TIME and END_TIME may be written in samples (TIMIT, NTIMIT: at 16kHz sampling rate), or in seconds (Switchboard). There may be many sequential labels per segment, e.g., a TXT file includes the entire sentence in a single line. Hash marks the beginning of a comment.

2. ESPS (Entropic Signal Processing Systems)

   - Where it is used: Radio Speech Corpus.
   - File Extension: Variable. Radio Speech uses .WRD, .TON, .BRK, .LBA, and .LBL.
   - Header: specifies the number of fields (`nfields`). Ends in # on a line by itself.
   - Data: Each line specifies a point in time. Usually, it is assumed that the end time of one segment is equal to the start time of the next segment. Format of each line:

     `START\_TIME COLOR LABEL1 [LABEL2 ...]`

     START_TIME is always in seconds. COLOR is an integer, specifying the colormap entry of the color used to label this particular time alignment. The purpose of COLOR is to allow color-coded display of multiple transcriptions at the same time; to my knowledge, only the original WaveSurfer makes use of this detail.

3. MLF (Master Label File).

   - Where it is used: HTK (the hidden Markov modeling toolkit) usually requires all other transcriptions to be converted into this format before training or testing a recognizer. MLF is convenient because it puts transcriptions for many speech files into a single transcription file.
   - File Extension: MLF.
   - Header: HTK will reject an MLF file not following this format: The file header is one line containing the characters "`#!MLF!#`". The section of the file corresponding to one speech file is initiated by a line naming the file, and is ended by a line containing a period (`.`) on a line by itself. In theory, the filename line is relatively flexible, but in practice, HTK breaks unless it contains exactly the characters

     `"*/FILENAME.lab"`

     where FILENAME is replaced by the root name of the speech file (everything that is not part of the path or the extension).
   - Format of a data line is

     `START\_TIME END\_TIME LABEL1 [LABEL2 ...]`

     START_TIME and END_TIME are optional; if the first non-space character on the line is not an integer, HTK will assume that the start time and end time of this line are not specified. If START_TIME and END_TIME are present, they must be integer values, specified in 100ns units (thus the line "450000 480000 DX" specifies a flap exactly 30ms long, starting at 0.45 seconds, and continuing until 0.48 seconds).

4. TextGrid.

   - Where it is used: this is the file format for Praat transcriptions.
   - File Extension: always TextGrid (e.g., SX43_PHN.TextGrid).

- Header, Data formats: These are too complicated to describe here, but relatively easy to learn if you need to. TextGrid is the only format listed here that can contain many different non-synchronous transcriptions of the same speech waveform in a single transcription file. The data format is designed using envelopes, rather like HTML, but with a different syntax. The outermost envelope is the file; then the transcription tier; then the segment or point. Within the segment envelope are contained fields (separate lines) specifying start time, end time, and label. The point envelope is similar to the segment envelope, but specifies just one time, rather than two times.

5. SGML

   - Where it is used: Broadcast News
   - File Extension: SGML.
   - Header: one line containing the filename, e.g. "m970927.sgml".
   - Data: in standard SGML envelopes, with types like "episode" (wrapping a broadcast program), "section" (wrapping a section with a particular content and a particular channel quality), "turn" (wrapping the dialog turn of one speaker), and "time" (specifying an alignment time at which the speech file may be broken; the "time" envelope is usually not closed by a /time marker).

2.3 Acoustic Feature File Formats

This course will always use the HTK file format for periodic vectors (MFCC, spectra, et cetera). HTK file format will be described later.

3 Dictionaries

Essentially all dictionaries have the same format: each line contains one word, followed by its pronunciation. There are subtle differences in the encoding that MUST be considered if you try to merge dictionaries. It is useful to merge dictionaries because English contains more "words" (space-delimited orthographic units) than you thought it did; for example, the 39k Switchboard dictionary contains 10,500 words that are not in the 91k pronlex dictionary (not counting word fragments, of which there are an additional 6000). Some of these are word fragments, some are not.

- Pronlex (the 1997 Callhome English lexicon, LDC lexicon release number LDC97L20; 90988 entries) uses one-letter ARPABET notation, syllable boundaries marked with '.', primary and secondary stresses marked with "' and '+':

  ```
  administration .@dm+In.Istr'eS.In
  ```

- The Mississippi State Switchboard transcriptions include a dictionary with 38910 entries, in dict/switchboard.dict. Phones use 2-letter ARPABET; stress and syllabification are not marked

  ```
  administration ae d m ih n ih s t r ey sh ih n
  ```

- Radio Speech corpus includes a lexicon representing the pronunciation of speaker F1A (Disk1/f1a/labspeech/f1alab.prn). Words end in comma for some reason; phones are labeled in 2-letter ARPABET, syllables are marked by '*', primary and secondary stresses with +1 or +2:

  ```
  administration, ae d * m ih+2 n * ih * s t r ey+1 * sh ax n
  ```

- TIMIT includes DOC/TIMITDIC.TXT (6200 entries). Phones are in 2-letter ARPABET, syllable boundaries unmarked, primary and secondary stress marked with numbers:

  ```
  administration  /ax d m ih2 n ix s t r ey1 sh ix n/
  ```

HTK requires dictionaries to be in Switchboard format, except that HTK does not care what phoneme encoding you use (if you want to distinguish stressed and unstressed vowels, HTK has no problem with that).

If several lines start with the same word, they are interpreted as equally probable alternative interpretations of the same word, e.g.

```
object ax b jh eh k t
object aa b jh eh k t
```

Pronunciation probabilities may be specified as

```
object 0.3 ax b jh eh k t
object 0.7 aa b jh eh k t
```

Of course it would be better to treat these words as linguistically distinct, e.g.

```
object_v ax b jh eh k t
object_n aa b jh eh k t
```

. . . but in that case, you have to make sure that your transcriptions match your dictionary. None of the speech corpora mentioned in Section 1, except Radio Speech, include transcriptions with Part of Speech marked.

4 Transcription and File Manipulation Software

- For reading and writing SPHERE/NIST waveform and transcription files, you can use the SPHERE toolkit (http://www.nist.gov/speech).
- Matlab builds in tools for reading and writing WAV. Tools for reading and writing short-integer SPHERE (TIMIT, but not Switchboard) are available in the VoiceBox toolkit (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html).
- Matlab and perl functions for reading and writing some transcription file formats will be provided on the course web page. These functions should work for the corpora listed in Section 1 (except possibly Broadcast News), but might break for other corpora.
- The most useful tool for viewing and interactively analyzing speech and transcription data is Praat: http://www.fon.hum.uva.nl/praat/. Praat is easy to use, contains signal processing tools capable of performing most standard analysis functions (pitch tracking, formant tracking, spectrograms, LPC, etcetera), includes its own scripting language that can automate these functions, is open-source, and is actively supported by its authors.

References

[1] Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. Resegmentation of switchboard. In *Proc. Internat. Conf. Spoken Language Processing*, 1998.

[2] J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *Proc. ICASSP*, pages 517–520, 1992.

[3] S. Greenberg, H.M. Carvey, and L. Hitchcock. The relation of stress accent to pronunciation variation in spontaneous american english discourse. In *Proc. ISCA Workshop on Prosody and Speech Processing*, 2002.

[4] Steven Greenberg and Leah Hitchcock. Stress-accent and vowel quality in the Switchboard corpus. In *NIST Large Vocabulary Continuous Speech Recognition Workshop*, Linthicum Heights, MD, May 2001.

[5] Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Judith Spitz. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Proc. ICASSP*, 1990.

[6] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang. Avicar: Audio-visual speech corpus in a car environment. In *INTERSPEECH International Conference on Spoken Language Processing*, 2004.

[7] Pedro J. Moreno and Richard M. Stern. Sources of degradation of speech recognition in the telephone network. In *Proc. ICASSP*, volume I, pages 109–112, 1994.

[8] M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. *The Boston University Radio Speech Corpus.* Linguistic Data Consortium, 1995.

[9] Taejin Yoon, Sandra Chavarria, Jennifer Cole, and Mark Hasegawa-Johnson. Intertranscriber reliability of prosodic labeling on telephone conversation using tobi. In *Proc. Internat. Conf. Spoken Language Processing*, 2004.

[10] V.W. Zue, S. Seneff, and J. Glass. Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9:351–356, 1990.