Speech Corpora

Speech corpus – a large collection of audio recordings of spoken language. Most speech corpora also have additional text files containing transcriptions of the words spoken and the time each word occurred in the recording.

When you conduct research on speech you can either (1) record your own data or (2) use a ready-made speech corpus.

Recording your own data:
Linguists usually collect their own data in a phonetics laboratory where there is a sound-attenuated booth and high-quality recording equipment. They ask speakers to read words or phrases that have been chosen specifically for the experiment. Words are read in the same "carrier phrase" in order to control for outside factors.

Say "heed" two times.
Say "hid" two times.
...

Using a speech corpus:
If you decide to use a speech corpus for your research, the Linguistics Department at Stanford has many available. Corpora are located either on:

- the AFS server
- the corpus computer in the Linguistics Department
- CDs, which can be checked out

See the corpora webpage for detailed information about corpora available and gaining access: http://www.stanford.edu/dept/linguistics/corpora/

Speech corpora can be divided into two types:
(1) Read speech

- Excerpts from books
- News broadcasts
- Word lists
- Number sequences

(2) Spontaneous Speech

- Dialogs and meetings – free conversations between 2 or more people
- Narratives – one person telling a story
- Map-tasks – two people are each given a map that other person cannot see. The maps are identical, except that one has a route specified. The person with the route must explain it to the other person.
- Appointment-tasks – two people are given individual schedules and are supposed to find a free time to meet.
- "Wizard of Oz" simulations – modeling a real-life situation, like booking a flight

Examples of English Speech Corpora in the Linguistics Department

Speech Corpus	Type of data	Size	Type of Annotation
TIMIT	Read sentences	630 speakers each reading 10 sentences 8 US dialects	Orthographic Phonetic
Broadcast News	News reports	104 hours of television and radio broadcasts	Orthographic
TIDIDIGITS	Connected digit sequences	326 speakers each reading 77 digit sequences	Orthographic
Switchboard	Phone conversations between strangers on an assigned topic	2400 conversations 543 speakers Many US dialects	Orthographic Some phonetic
CallHome	Phone conversations with family and close friends.	120 conversations Up to 30 min each	Orthographic
ICSI meetings	Weekly meetings of various research groups	72 hours 53 speakers	Orthographic
HCRC Map Task	Map-task	18 hours 62 speakers (mainly Scots English)	Orthographic
ATIS	Flight booking	36 speakers	Orthographic

The vast majority of corpora are in English, but other languages are available as well: Arabic, Bulgarian, Cantonese, Czech, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Portuguese, Russian, Spanish, Tamil, Vietnamese.

Advantages of using a speech corpus:

(1) Time saving – no need to collect and process recordings
(2) Large amounts of data
(3) Searchability
(4) Real language usage

Disadvantages of using a speech corpus:

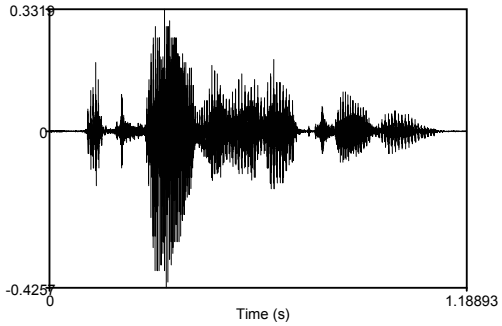
(1) Recording quality often lower than in a phonetic laboratory
(2) Too much information – may need to work on subsets
(3) Messy - not as controlled as speech collected in a phonetic laboratory
(4) Currently only available for mainstream languages

Types of Annotation

In order for speech corpora to be useful for research they need to be labeled in some way. At the minimum the words spoken are transcribed in standard orthography. Sometimes additional linguistic information is provided: syllables, sounds, intonation, disfluencies, filled pauses (um, uh). Phonetic transcription is usually done in ARPABET (see chart below).

Typically the actual recordings and the annotations are in separate files linked by a common filename. Orthographic and phonetic transcriptions are usually simple text files. You may need to write small scripts to process the transcriptions or at least be able to use simple search commands such as "grep."

Audio Recording:



Orthographic transcription (not time-aligned):

A: What I was doing at, at home, is like I work nights here, so that's another long story that we will talk about. It's funny that I got you though.

Orthographic transcription (time-aligned):

```
A       6.40     0.14  It's
A       6.54     0.20  funny
A       6.74     0.06  that
A       6.80     0.12  I
A       6.92     0.14  got
A       7.06     0.18  you
A       7.24     0.18  though.
```

Phonetic Transcription (IPA: [ɪts fɅi ðʌɾ aɪ ɣatʃu ðoʊ])

```
0.334407  121 h#
0.460000  121 ih t s
0.591176  121 f ah_n
0.650000  121 iy
0.732149  121 dh ah
0.828198  121 dx ay
0.940895  121 g_ap aa
1.140000  121 ch uw
1.339699  121 dh ow
1.464997  121 h#
```

Examples of phonetic research with speech corpora:

- Comparing pronunciations in different dialects
- Comparing pronunciation by males and females
- Flapping across word boundaries in spontaneous speech
- The effect of disfluencies on neighboring words
- Duration of sounds at the end of an utterance
- Pronunciation of unstressed vowels
- The omission of sounds (sound deletion)
- Palatalization across word boundaries – whatcha, gotcha, wouja
- Intonational patterns

In addition to general linguistic research, speech corpora play a crucial role in automatic speech recognition and speech synthesis.

To work with speech, I recommend using Praat. It can be downloaded for free from http://www.praat.org and works on all platforms. (It's a good idea to go through the tutorial first.) Praat lets you measure following things (you will learn about these later in the course):

- Duration
- Vowel formants
- Fundamental frequency (Pitch)
- Intensity (Loudness)

Practice with spontaneous speech

The best part of speech corpora is having physical evidence of how we actually speak on a daily basis. Spontaneous speech is full of surprises! It's fascinating to compare how we think a phrase is pronounced with how someone actually says it in real conversation.

You will hear the following utterances. Transcribe them phonetically using the IPA.

Example 1: It's funny that I got you though.

Example 2: Yeah I guess that about does it.

Example 3: What's what's your most recent one that you've seen.

Example 4: … is you sit down at the table.

Example 5: On Monday I wear the worst looking one.

ARPABET and approximate IPA equivalents

If you work with a phonetically transcribed corpus, most likely the sounds will be transcribed using the ARPABET (developed by the Advanced Research Projects Agency). Since you are learning the IPA in Ling 110, you may find this conversion chart useful for your project.

ARPABET	IPA	ARPABET	IPA
p	p^h, p	l	l
b	b	r	ɹ
t	t^h, t	w	w
d	d	y	j
k	k^h, k	er	ɚ
g	g	iy	i
f	f	ih	ɪ
v	v	ey	eɪ
th	θ	eh	ɛ
dh	ð	ae	æ
s	s	aa	ɑ
z	z	ah	ʌ
sh	ʃ	ax	ə
zh	ʒ	ao	ɔ
hh	h	ow	oʊ
ch	tʃ	uh	ʊ
jh	dʒ	uw	u
m	m	ay	aɪ
n	n	aw	aʊ
ng	ŋ	oy	ɔɪ

Sample Searches

Searching for examples of the word "probably" in the Switchboard Corpus:

```
% cd /afs/ir/data/linguistic-data/Switchboard/Switchboard-Transcripts/swb1/trans
% grep –i "probably" phase*/disc*/*.txt
```

Searching for sequence "what you" in the Switchboard Corpus:

```
% cd /afs/ir/data/linguistic-data/Switchboard/Switchboard-Transcripts/swb1/trans
% grep –i "what you" phase*/disc*/*.txt
```

Many searches however may require a bit of programming to process the data. If this seems daunting you can ask around; someone may already have the program written that you need.