Lecture 1: Phonetic and Prosodic Transcription Codes

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

May 17, 2005

1 Phonetic Transcription Codes

On systems that use unicode, phonemes can be labeled using the International Phonetic Alphabet (http://www2.arts.gla.ac.uk/IPA/fullchart.html). Unfortunately, speech corpora are usually transcribed in plaintext (also known as ASCII, `man ascii`, a subset of the Latin-1 alphabet), therefore phonemes must be encoded using an ASCII code. Two codes are standard: a case-sensitive code that uses one character per phoneme (1-character ARPABET), and a case-insensitive code that uses up to two characters per phoneme (2-character ARPABET).

Tense vowels and Diphthongs:

IPA	ARPABET 1	ARPABET 2	Example
ɑ	a	aa	father
æ	@	ae	bat
ɔ	c	ao	bought
aw	W	aw	bout
aj	Y	ay	bite
e	e	ey	bait
i	i	iy	beat
o	o	ow	boat
oj	O	oy	boy
u	u	uw	school

Lax vowels and Schwa:

IPA	ARPABET 1	ARPABET 2	Example
ə	x	ax	about
	x	ix	attribute
ɛ	E	eh	bed
ɪ	I	ih	bit
ʊ	U	uh	book
ʌ	A	ah	buck

Glides and Liquids (voiced /h/ is more like a glide in English, unvoiced /h/ is more like a fricative. Placement depends on your philosophy).

IPA	ARPABET 1	ARPABET 2	Example
h	h	hh	hi
ɦ	h	hv	ahead
l	l	l	lead
l̞	L	el	bottle
r	r	r	roof
r̞	R	er	bird
r̞	R	axr	butter
w	w	w	wall
j	y	y	yacht

Nasal Consonants:

IPA	ARPABET 1	ARPABET 2	Example
m	m	m	mom
m̩	xm	em	bottom
n	n	n	new
n̩	N	en	button
ŋ	G	ng	sing
ŋ̩	xG	eng	tossing

Fricative Consonants:

IPA	ARPABET 1	ARPABET 2	Example
f	f	f	frank
v	v	v	very
θ	T	th	think
ð	D	dh	that
s	s	s	silly
z	z	z	zoom
ʃ	S	sh	shelf
ʒ	Z	zh	azure

Stops and Affricates:

IPA	ARPABET 1	ARPABET 2	Example
p	p	p	pool
b	b	b	bite
t	t	t	tip
d	d	d	dog
tʃ	C	ch	child
dʒ	J	jh	judge
k	k	k	clip
g	g	g	good
ʕ	q	q	Batman

Various dictionaries and transcriptions denote stop closures and silences in different ways. First, you must understand that a stop consonant is marked by three instants in time: the closure instant (when the mouth closes), the release (when the mouth opens), and the voice onset (when voicing begins). The period of time from closure to release is called the "closure segment." The period of time from release to voice onset is sometimes called the "release segment."

- TIMIT breaks every stop into two segments: the closure segment, and the release segment. TIMIT uses the stop label /p,t,k,b,d,g/ to denote the "release segment." The "closure segment" (the silence) is labeled /pcl,tcl,kcl,bcl,dcl,gcl/.
- SPHINX uses just ONE SEGMENT to denote any given stop, but considers released stops (the d in "adapt") and unreleased stops (the d in "adman") to be DIFFERENT PHONEMES. /p,t,k,b,d,g/ are used to denote released stops (covering the segment from closure to voice onset). Unreleased stops (from closure until the boundary of the next phoneme) are labeled /pd,td,kd,bd,dd,gd/.
- "epi" is used in TIMIT to denote the short "epinthetic" silence inserted between a fricative and nasal, e.g., in "insert." All other transcription systems consider this silence to be part of the nasal.
- TIMIT uses "h#" to denote sentence-initial or final silence, and "pau" to denote pauses in the middle of the sentence. Most other systems use "sil" to denote silence of any kind.
- Radio Speech Corpus uses "brth" to denote breath noise. Other kinds of noise are marked with great detail in Switchboard.

2 Prosody

2.1 Lexical Stress

Every content word contains one syllable with primary stress, and possibly other syllables with secondary stress. Some dictionaries (e.g., pronlex) mark syllabification and stress. Annotation codes are described in the distribution notes for each dictionary, but a relatively common system (used by Radio Speech) marks lexical stress using "+digit" after the vowel, e.g.,

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

2.2 Prosodic Phrasing

Words are grouped into rhythmic phrases. The phones in the coda of the syllable preceding a phrase boundary are lengthened [3], and phrase-initial phones may be glottalized [1].

The Tones and Break Indices (ToBI) system of prosodic transcription allows for five different levels of prosodic break between succeeding words. Level 1 is a normal word boundary. Level 0 is used to separate words that are not pronounced as separate words, meaning that co-articulation effects occur across the word boundary; these words are said to be a clitic group. Level 3 denotes an "intermediate phrase boundary" (ip); intermediate phrase boundaries tend to occur at a subset of the syntactic phrase boundaries. Level 4 denotes an "intonational phrase boundary" (IP); IP boundaries tend to occur at most clause boundaries (e.g., sentence boundaries, dependent clause boundaries, extra-positional dialog marker boundaries, parenthetical clauses). Level 2 is used for a break that is marked by one of the two correlates of an intermediate phrase boundary: either the boundary tone, or durational cues (phoneme lengthening and/or pause), but not both. In the Radio Speech Corpus, break indices are marked in the BRK files.

Level 3 and Level 4 breaks are marked by intonational (F0) correlates. After the last pitch accent in an intermediate phrase (ip), all remaining voiced frames are marked with the same low or high phrase tone (L- or H-); a downstepped intermediate phrase is one that ends in a high tone, but at a lower (explicitly downstepped) value relative to previous high tones (denoted !H-). An intonational phrase boundary is marked by an extra low or high movement on its very last syllable. Intonational phrase boundary tones sometimes, but not always, mark the type of the syntactic clause coinciding with the IP boundary: L-L% marks the end of a declaration, H-H% sometimes marks a yes-no question, L-H% marks a clause boundary after which the talker intends to continue speaking. In the Radio Speech Corpus, phrase tones are marked in the TON files.

Break indices and tones are listed here:

Boundary	Break Index	Possible Phrase Tones
Clitic	0	none
Word	1	none
Intermediate	3	L-, H-, !H-
Intonational	4	L-L%, L-H%, H-L%, H-H%, !H-L%

2.3 Phrasal Prominence

Typically at least one word per intermediate phrase is more "prominent" than the others.

In the notation of Greenberg [2], prominent syllables are marked with a number: 1.0 for syllables with extremely high perceptual prominence, 0.5 for syllables with moderate perceptual prominence, 0.0 for syllables that are not prominent.

ToBI notation proposes that there are categorically different types of phrasal prominence, and that the phonological categories correlate with F0 movement over the prominent syllable. For this reason, phrasal prominence in ToBI is called "pitch accent:" some types of "pitch accent" (L*) may involve no F0 movement at all, but the *lack* of F0 movement is considered to be phonologically distinctive (distinguishing L* from H* or !H*, and allowing the talker to thereby convey meaning). In English ToBI notation, there are seven different types of pitch accent, shown in the table below. Each of these is composed of one or two consecutive tones, of type L (low), high (H), or downstepped high (!H; an F0 peak that is explicitly lowered with respect to some previous F0 peak). The tone marked '*' is implemented over the prominent syllable; the other tone, if specified, is implemented over a neighboring syllable (or over the first or last part of the prominent syllable). Some distinctions are difficult (e.g., L+H* vs. H*), therefore, to date, University of Illinois Switchboard transcriptions only distinguish L* vs. H* [4]. The Radio Speech TON files include nine different types of phrasal prominence, as listed here:

Tone on Prominent Syllable	Labels
Low	L*, L*+H
High	L+H*, H*, H*+L
Downstepped High	!H*, H+!H*
Transcriber Unsure	*?
Tone Unmarked	*

References

[1] Laura Dilley, Stefanie Shattuck-Hufnagel, and Mari Ostendorf. Glottalization of word-initial vowels as a function of prosodic structure. *J. of Phonetics*, 24:423–444, 1996.

[2] 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.

[3] Colin Wightman, Stefanie Shattuck-Hufnagel, and Mari Ostendorf andPatti Price. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91(3):1707–1717, March 1992.

[4] 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.