Chapter 3

Fundamentals of Human Speech Production

3.1 **(a)** The regions of voiced speech, unvoiced speech and silence (background signal) are shown in Figure P3.1.1.

(b) The pitch periods (in msec) are indicated by a bracket (in Figure P3.1.1) under the segment of speech to which the estimate corresponds.

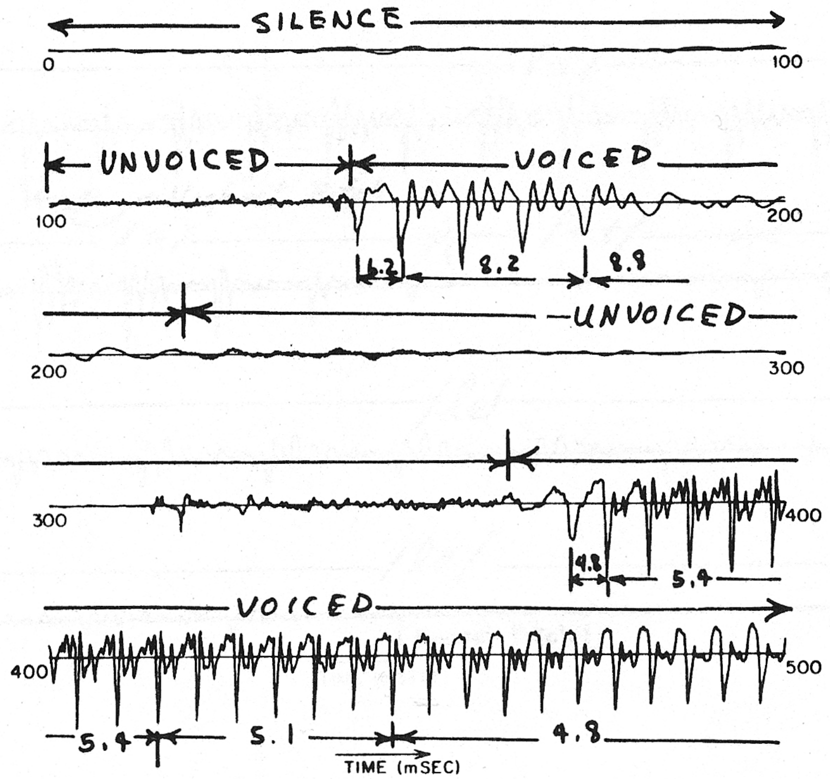


Figure P3.1.1: Locations of regions of voiced, unvoiced and silence.

****

3.2 **(a)** The approximate boundaries between the phonemes are marked in Figure P3.2.1.

(b) The points of lowest and highest pitch are also marked in Figure P3.2.1.

(c) The lowest pitch has a period of about 21.5 msec corresponding to a frequency of 46 Hz. This very low pitch is strongly indicative that the speaker is probably male.

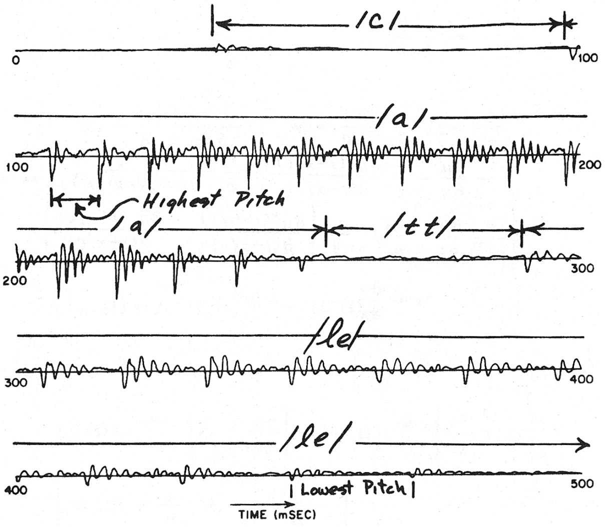


Figure P3.2.1: Time waveform of speech utterance “cattle” with phoneme boundaries and points of lowest and highest pitch marked on the waveform.

****

3.3

1. /the/ can be pronounced as /DH/ /UH/ or /DH/ /IY/; /of/ is pronounced as /AA/ /V/; /and/ is pronounced as /AE/ /N/ /D/; /to/ is pronounced as /T/ /UW/; /a/ is pronounced either as /EY/ or /UH/; /in/ is pronounced as /IH/ /N/; /that/ is pronounced as /DH/ /AE/ /T/; /is/ is pronounced as /IH/ /Z/ (S optional); /was/ is pronounced as /W/ /AA/ /Z/ (S optional); /he/ is pronounced as /HH/ /IY/.
2. /data/ can be pronounced as either /D/ /EY/ /T/ /AX/ or /D/ /AE/ /T/ /AX/; /lives/ can be pronounced as either /L/ /AY/ /V/ /Z/ (optional S) or /L/ /IH/ /V/ /Z/ (optional S); /record/ can be pronounced as either /R/ /IH/ /K/ /AO/ /R/ /D/ or /R/ /EH/ /K/ /ER/ /D/.
3. /company/ is pronounced /K/ /UH/ /M/ /P/ /AX/ /N/ /IY/; /happiness/ is pronounced /HH/ /AE/ /P/ /IY/ /N/ /EH/ /S/; /willingness/ is pronounced as /W/ /IH/ /L/ /IH/ /NX/ /N/ /EH/ /S/.
4. The sentence /I enjoy the simple life/ is pronounced as /AY/-/EH/ /N/ /JH/ /OY/-/DH/ /UH/- /S/ /IH/ /M/ /P/ /AX/ /L/-/L/ /AY/ /F/; the sentence /Good friends are hard to find/ is pronounced as /G/ /UH/ /D/-/F/ /R/ /EH/ /N/ /D/ /Z/ (optional S)-/AH/ /R/-/HH/ /AA/ /R/ /D/-/T/ /UW/- /F/ /AY/ /N/ /D/.

****

3.4 The sounds in the word /and/ are /AE/ /N/ /D/ and the (extremely) approximate locations of the sounds are:

- /AE/ samples 400-3800

- /N/ samples 3800-5400
- /D/ samples 5400-6100

The sounds in the word /that/ are /DH/ /AE/ /T/ and the (extremely) approximate locations of the sounds are:

- /TH/ samples 600-1600
- /AE/ samples 1600-3600
- /T/ samples 3600-4800

The sounds in the word /was/ are /W/ /AA/ /Z(S)/ and the (extremely) approximate locations of the sounds are:

- /W/ samples 400-1500
- /AA/ samples 1500-4800
- /Z(S)/ samples 4800-6600

The sounds in the word /by/ are /B/ /AY/ and the (extremely) approximate locations of the sounds are:

- /B/ samples 400-1700
- /AY/ samples 1700-6100

The sounds in the word /enjoy/ are /EH/ /N/ /JH/ /OY/ and the (extremely) approximate locations of the sounds are:

- /EH/ samples 600-1440
- /N/ samples 1440-2860
- /JH/ samples 2860-3500
- /OY/ samples 3500-8000

The sounds in the word /company/ are /K/ /UH/ /M/ /P/ /AX/ /N/ /IY/ and the (extremely) approximate locations of the sounds are:

- /K/ samples 500-1140
- /UH/ samples 1140-2000
- /M/ samples 2000-2440
- /P/ samples 2440-3600
- /AX/ samples 3600-4440
- /N/ samples 4440-5190
- /IY/ samples 5190-6222

The sounds in the word /simple/ are /S/ /IH/ /M/ /P/ /(AX L—EL)/ and the (extremely) approximate locations of the sounds are:

- /S/ samples 0-2590
- /IH/ samples 2590-3290
- /M/ samples 3290-3920
- /P/ samples 3920-5040

- /(AX/ samples 5040-5610
- /L—EL)/ samples 5610-6200

****

3.5 Using the spectrogram (as shown in Figure P3.5.1) to locate the approximate centers of the three vowel regions as:

/enjoy/ - sample 5700 (0.57 sec) - center of /OY/ sound - pitch 80 samples or 125 Hz /simple/ - sample 10850 (1.085 sec) - center of /AX—EL/ sound - pitch 71 samples or 141 Hz /life/ - sample 16580 (1.658 sec) - center of /AY/ sound - pitch 88 samples or 114 Hz

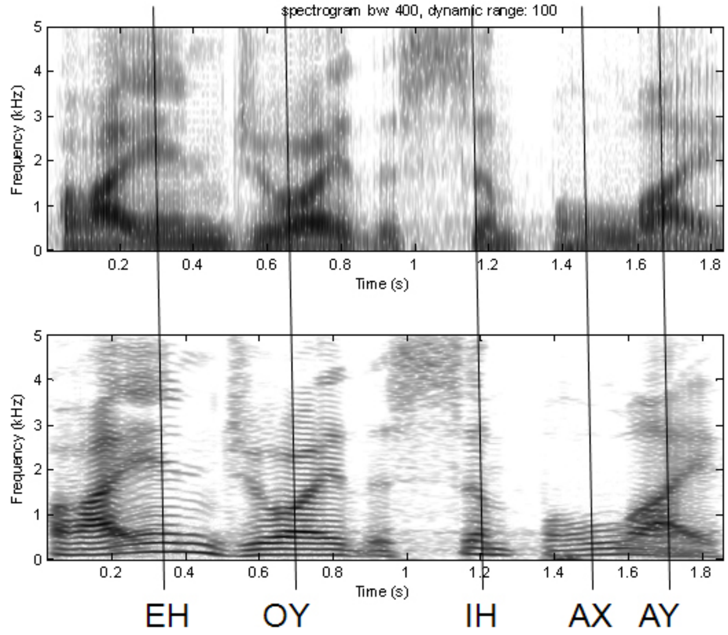


Figure P3.5.1: Locations of center of vowel sounds in utterance.

****

3.6 S samples 1-445 U samples 445-607 V samples 607-2363 U samples 2363-3828 V samples 3828-7032 U samples 7032-7450 V samples 7450-8466 U samples 8466-9144 V samples 9144-11435 U samples 11435-12599 V samples 12599-13202 U samples 13202-14222 V samples 14222-18631 U samples 18631-20357

The regions of voiced, unvoiced and silence are marked on the waveform plot of Figure P3.6.1.

****

3.7 **(a)** The region for the merged /D/ phonemes is approximately samples 4300-4700.

(b) The region for the /IY/ sound in the word "each" is approximately samples 8800-9600.

(c) The region for the /CH/ sound in the word "each" is approximately samples 10,000-10,500.

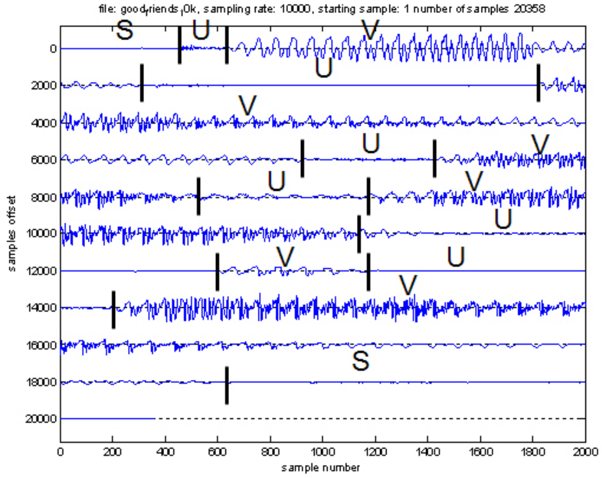


Figure P3.6.1: Waveform with marked regions for voiced, unvoiced, and silence (background) for the sentence "Good friends are hard to find".

(d) We estimate the fundamental frequency for the voiced segment on the first line of the waveform plot by counting the number of pitch cycles over the duration of the voiced region. We see that there are approximately 19 periods over the 1900 sample voiced region, giving a period of 1200/19=63.16 samples. At the sampling rate of 8000 Hz, we covert the 63.16 samples to a period of 7.89 msec. Finally we convert the period to the pitch frequency giving an average fundamental frequency of 126.7 Hz.

(e) There are about 21 phones in the text of the spoken utterance and they take about 17,000 samples or 2.125 seconds at the sampling rate of 8000 samples/second. Assuming independent phonemes with 6 bits per phoneme, we estimate the bit rate of the utterance as the product of the number of phonemes/second (21/2.125) with the number of bits/phoneme (6) giving a total bit rate of about 60 bits/second.

****

3.8 The narrowband spectrogram is the one at the top. The narrowband spectrogram is characterized by a wide time duration with narrow frequency bandwidth; hence it is able to resolve individual pitch harmonics in frequency, leading to a series of pseudo-horizontal striations in the plot.

The wideband spectrogram is characterized by a narrow time duration with wide frequency bandwidth; hence it resolves pitch periods in time, leading to a series of vertical striations in the plot.

****

3.9 **(a)** The bottom spectrogram is the wideband spectrogram.

(b) The fundamental frequency at $t = 0.18$ seconds is estimated from the narrowband spectrogram where we see that at $t = 0.18$ seconds, there are about 19 harmonics in a frequency band from 0 to 2500 Hz, giving an estimate of the fundamental frequency of about 131 Hz.

(c) From the narrowband spectrogram we see that the fundamental frequency is decreasing in the region from $t = 1.6$ to $t = 1.8$ seconds.

(d) From the wideband spectrogram we can estimate the values for the first three formant frequencies as $F_1 = 700$ Hz, $F_2 = 1700$ Hz and $F_3 = 2400$ Hz.

(e) The location of the merged /D/ phonemes is the region around $t = 0.6$ seconds (approximately).

****

3.10 By identifying the features of the four spectrograms and comparing them to the presumed features for the given words, it can readily be deduced that the four spectrograms correspond to the following words:

- The top left spectrogram corresponds to the word “was”
- The top right spectrogram corresponds to the word “enjoy”
- The bottom left spectrogram corresponds to the word “company”
- The bottom right spectrogram corresponds to the word “enjoy”

****

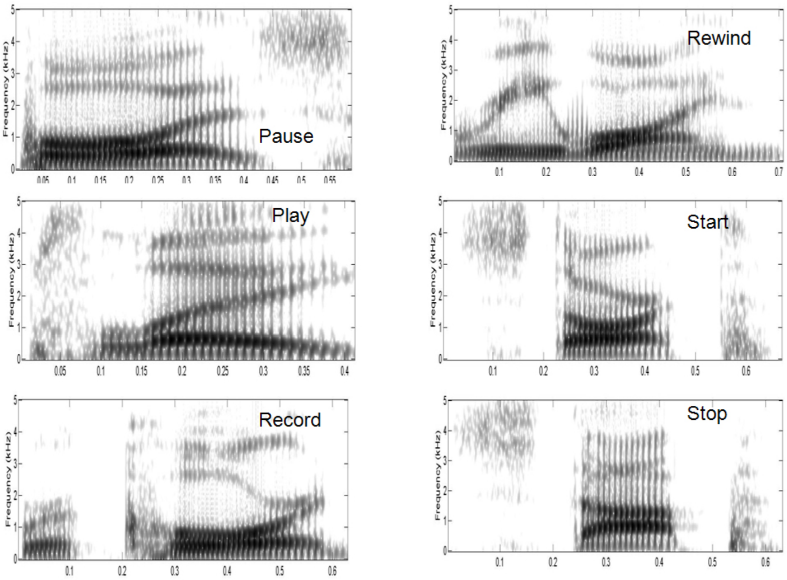


Figure P3.11.1: Spectrograms with word labels of one version of each of the control words for a voice controlled cassette tape system.

3.11 By examining the features of the six control words it is relatively easy to match the individual spectrograms to the spoken command words, as shown marked in Figure P3.11.1.

Thus the control word “rewind” is an all-voiced utterance with a voiced stop, /D/, at the end; hence it’s spectrogram is in the top row, second column. (Note that the positions of the spectrograms are different from that given in the problem statement.)

There are two words that start with strong fricatives, namely “start” and “stop”. The feature in the spectrogram that distinguishes these two control words is the falling third formant for the /R/ sound in “start”. Hence we can readily assign the spectrogram in the third row, first column to “start”, and the spectrogram in the second row, second column to “stop”.

There is only one spectrogram that ends with a strong fricative, namely “pause” and hence we can assign the spectrogram in the second row, first column to “pause”.

Of the remaining two spectrograms, corresponding to the words “record” and “play”, we can easily see that the spectrogram in the third row, second column has an initial stop consonant (/P/) and is followed by an all-voiced region; hence it represents the word “play”.

Finally by the process of elimination, the remaining spectrogram in the first row, second column corresponds to the word “record”. We see the stop gap of the /K/ in “record” which serves to verify the analysis for this word.

****

3.12 The sounds of these two sentences are as follows:

1. “She eats some Mexican nuts”: /SH/ /IY/ - /IY/ /T/ /S/- /S/ /UH/ /M/ - /M/ /EH/ /K/ /S/ /IH/ /K/ /IH/ /N/ - /N/ /UH/ /T/ /S/.
2. ”Where roads top providing good driving”: /WH/ /EH/ /R/ - /R/ /OW/ /D/ /S/ - /S/ /T/ /AA/ /P/ - /P/ /R/ /UH/ /V/ /AY/ /D/ /IH/ /NX/ - /G/ /UH/ /D/ - /D/ /R/ /AY/ /V/ /IH/ /NX/.

The sound at the end of each word is virtually identical to the sound at the beginning of the following word – hence there is a high degree of sound co-articulation across words, making it virtually impossible to reliably identify word boundaries in this spoken context.

****

3.13 It is relative easy to decode the text of the sentence without the vowels as “To give you some idea of the amount of work required in this course”. However, without the consonants, the text is virtually undecodable. The actual sentence is “Bear in mind that it accounts for half the grade in this course”.

****

3.14 The most commonly occurring words in this list are mono-syllabic function words which account for the top 56 word, and 93 of the top 105 words in the list. About 89% of the top 105 words from this list are mono-syllabic, so about 11% have more than one syllable.

The poly-syllabic words in this list are:

“about”, “into”, “only”, “any”, “over”, “even”, “after”, “alos”, “many”, “before”, and “because”.

****

3.15 The way to determine the most likely vowels is to measure the first three formant frequencies (in or around the middle of the vowels) and then compute a weighted distance between the formants of the unknown vowel sound, and the formants in the vowel chart of Chapter 3. Using this criterion (and using various MATLAB tools to measure the frequencies accurately, and to measure distances to target vowels accurately) we get the following matches:

1. row 1, column 1 – the measured formants are 703, 1113, and 2402 Hz, and the closest vowel match is /AA/ with a weighted distance of 0.05, and no other vowel being close to this minimum distance. The word that was spoken was “sob” and the vowel was /AA/.
2. row 1, column 2 – the measured formants are 557, 1299 and 1563 Hz, and the closest vowel match is /ER/ with a weighted distance of 0.16, with no other vowel being close to this minimum distance. The word that was spoken was “bird” and the vowel was /er/.

3. row 2, column 1 – the measured formants are 313, 2510 and 2939 Hz, and the closest vowel match is /IY/ with a weighted distance of 0.19, with no other vowel being close to this minimum distance. The word that was spoken was "cease" and the vowel was /iy/.
4. row 2, column 2 – the measured formants are 596, 1943 and 2930 Hz, and the closest vowel match is /EH/ with a weighted distance of 0.23 with no other vowel being close to this minimum distance. The word that was spoken was "set" and the vowel was /EH/.
5. row 3, column 1 – the measured formants are 811, 1855 and 2734 Hz, and the closest vowel match is /AE/ with a weighted distance of 0.28 with no other vowel being close to this minimum distance. The word that was spoken was "sat" and the vowel was /AE/.
6. row 3, column 1 – the measured formants are 283, 830 and 2793 Hz, and the closest vowel match is /UW/ with a weighted distance of 0.26 with no other vowel being close to this minimum distance. The word that was spoken was "boot" and the vowel was /UW/.

****

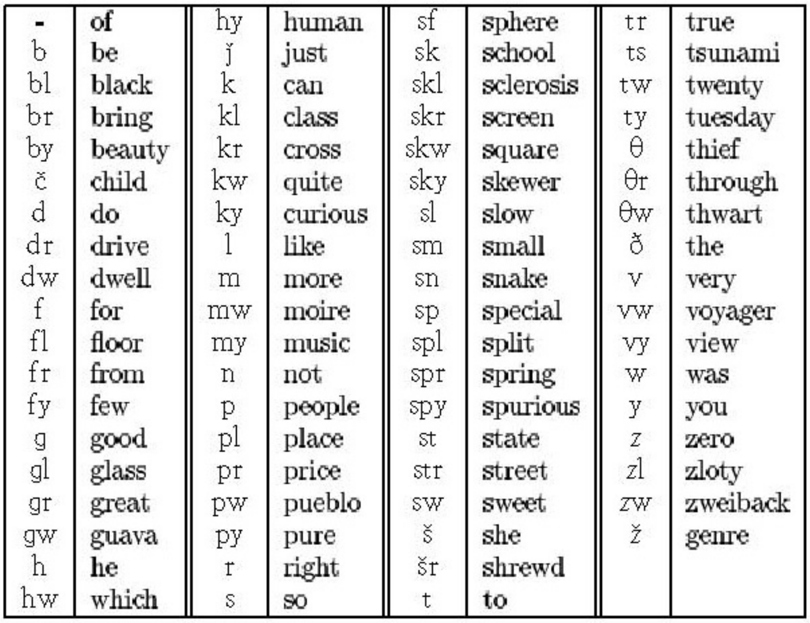
3.16 The transcription and the place and manner of articulation for the consonent sounds of the sentence "I enjoy the simple life" are as follows: /AY/ - /EH/ /N/ /JH/ /OY/ - /DH/ /UH/ - /S/ /IH/ /M/ /P/ /AX/ /L (EL)/ - /L/ /AY/ /F/

- /N/ – alveolar, voiced nasal sound
- /JH/ – alveolar-palatal, voiced fricative
- /DH/ – dental, voiced fricative
- /S/ – alveolar, unvoiced fricative
- /M/ – bilabial, voiced nasal
- /P/ – bilabial, stop
- /L/ – alveolar, voiced glide
- /F/ – labiodental, unvoiced fricative

****

3.17 Figure P3.12.1 shows a list of the word-initial consonants that occur at the beginning of English words. The word-initial consonant pairs consist of the following combinations:

1. /HH/ followed by /W/ and /Y/
2. /B/ followed by /L/, /R/ and /Y/
3. /D/ followed by /R/ and /W/
4. /G/ followed by /L/, /R/ and /W/
5. /P/ followed by /L/, /R/, /W/ and /Y/
6. /T/ followed by /R/, /W/, /Y/ and /S/
7. /K/ followed by /L/, /R/, /W/ and /Y/
8. /M/ followed by /W/ and /Y/
9. /V/ followed by /W/ and /Y/
10. /Z/ followed by /L/ and /W/
11. /F/ followed by /L/, /R/ and /Y/
12. /TH/ followed by /R/ and /W/



-	of	hy	human	sf	sphere	tr	true
b	be	ǰ	just	sk	school	ts	tsunami
bl	black	k	can	skl	sclerosis	tw	twenty
br	bring	kl	class	skr	screen	ty	tuesday
by	beauty	kr	cross	skw	square	θ	thief
č	child	kw	quite	sky	skewer	θr	through
d	do	ky	curious	sl	slow	θw	thwart
dr	drive	l	like	sm	small	ð	the
dw	dwell	m	more	sn	snake	v	very
f	for	mw	moire	sp	special	vw	voyager
fl	floor	my	music	spl	split	vy	view
fr	from	n	not	spr	spring	w	was
fy	few	p	people	spy	spurious	y	you
g	good	pl	place	st	state	z	zero
gl	glass	pr	price	str	street	zl	zloty
gr	great	pw	pueblo	sw	sweet	zw	zweiback
gw	guava	py	pure	š	she	ž	genre
h	he	r	right	šr	shrewd		
hw	which	s	so	t	to		

Figure P3.12.1: List of word-initial consonants that occur at the beginning of English words.

13. /S/ followed by /L/, /W/, /M/, /N/, /P/, /T/, /K/ and /F/
14. /SH/ followed by /R/

The general rule is a consonant followed by a glide, or the fricative /S/ followed by a glide, a nasal, a voiceless stop or the fricative /F/. The place of articulation of the initial consonant and the place of articulation of the following glide sound (/W/, /L/, /R/ or /Y/) is generally unrelated as the articulators glide between initial configuration and that of the glide.

The only word-initial consonant triplets are the combinations: 1. /S/ /K/ followed by /L/, /R/, /W/ and /Y/ 2. /S/ /P/ followed by /L/, /R/ and /Y/ 3. /S/ /T/ followed by /R/ The general rule here is /S/ as the initial consonant, followed by an unvoiced stop (/K/ or /P/ or /T/) followed by a glide (/L/, /R/, /W/ or /Y/). Again, the place of articulation of the initial /S/ consonant (alveolar) is generally unrelated to the place of articulation of the stop consonant that follows (bilabial for /P/, alveolar for /T/ and velar for /K/) or the place of articulation of the glide sound (/W/, /L/, /R/ or /Y/) that follows the word-initial consonant pairs.

****