4.9 Segmentation and Theory behind It

The segmentation was done in *Praat*, using TextGrids, *Praat*'s annotations. Segments are the annotated parts of sound data, superimposed over it, used as references for measurements or for further data extraction. TextGrids allowed very precise placement and manipulation of segments, and, overall, they were very important for consistent work and measurement.

Two methods were used to make annotations in the grids, which enabled the extraction of several types of values: measuring between two points (for measuring time) and extracting values from single points in data (for measuring formants, intensity, and pitch).

Before the actual segmentation of the recorded 16 signal files,[1] each containing 32 sentences, the signal files were individually checked in *Audacity* for errors or strange peaks. Afterwards, they were opened in *Praat*, inspected and segmented[2]. We created 16 TextGrid files containing segmented parts of our recorded sentences.

The segmentation was done on three levels (figure 11): three *Praat* "tiles" (grid elements) were created per TextGrid, for two different types of calculations. One tile (number 3) was named "word" and it included the words within the sentences. The second ("diph", number 1) marked the beginning and the end of the diphthongs within the words. They were both used to extract the length of the referring elements. The words were labelled with corpus word names ("bourse", "Joyce"), while diphthongs followed more detailed procedure. Each diphthong was marked with ASCII[3] letters representing both the first and the second targets, followed by an underscore and length mark ("s" for short and "l" for long). For example, /ɪə/ in "fierce" was labelled "ia_s".

The third tile contained not intervals, but the single points that referred to the two targets within the diphthong. We placed the point marks in the positions that we believed approximately displayed the best articulatory values of the target vowels within a diphthong. The points were labelled by a diphthong name, followed by an underscore, a length mark, again an underscore, and the target number (i.e "babe" had the word label "babe", the diphthong interval label "ey", and the targets "ey_l_1" and "ey_l_2")[4].

[1] 15 recorded by our Serbian speakers, and one by the referent speaker.

[2] The sound data annotation was very time-consuming. It was done in several steps.

[3] The system consisting of mostly English alphabet letters. It is widely supported in software, unlike the more modern Unicode system.

[4] In the chapter about the results the notation is explained again.

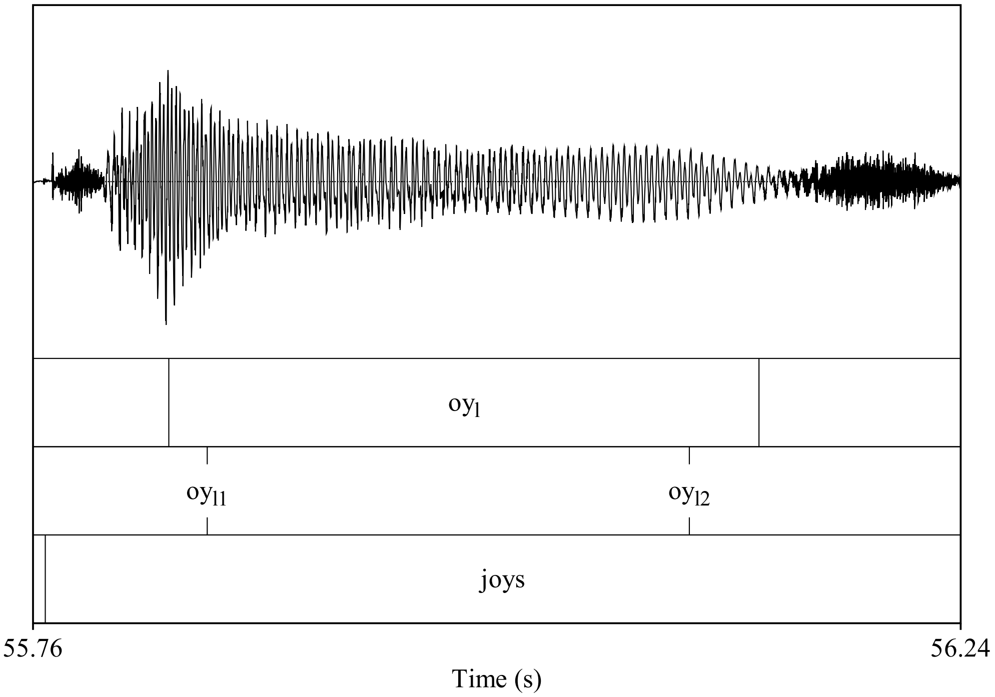

Figure . A TextGrid example, drawn here below a waveform

Table 12 shows the total number of annotations that needed to be placed manually during work in *Praat* after visual inspection and correct settings. The completed 16 TextGrid files with 2046 labels underwent detailed automated check[5] to ensure all segmented elements were properly formatted, placed within correct subordinate elements, and in full number.

Table

The overview of the TextGrid elements in the corpora, per file and the final count

Sound signal	"words" (interval)	"diph" (interval)	"points" (point)	Total
1	32	32	64	128
	For 16 sound signals, total marks:			2048

The segmentation itself was performed after referring to several sources. One of the methodological issues in the research was where to place boundaries that limit the data, later used for measurements. Harrington and Cassidy gave an overview for determining vowel targets (*Techniques* 59) by different authors. In some studies the targets were defined as section or "the formant values at a single time point" (59), while in others authors propose an entire section. In the second instance, it was suggested that 25% of the total steady duration of a vowel be taken as the reference section. However, this can be applied to monophthongs; applying it to diphthongs would bring into focus some other problems, such as defining the limits and lengths of the two targets. Also, a problem related to the steady state in vowels is that such "steadiness" is not substantiated by firm evidence (59).

Another proposed measurement approach was to focus on F1 movement. The authors suggested calculation from the point where "the first format reaches its maximum value" (60). The rationale was that F1 movement is related to the jaw movement, achieving the target

[5] Again, a custom code was written to load the TextGrids and run several checks (diphthong count, matching, etc). The code uses a part of *NLTK* library. More details in the Appendix.

values within a syllable. This approach should be applicable in certain vowel classes, “at least for most phonetically open and mid vowels”, where “F1 in general should be in the shape of inverted parabola whose maximum occurs at the vowel target” (60).

In the chapter about characterising vowels and measurements, Ladefoged writes that “in short monosyllables ... that do not have diphthongs, an interval near the middle of the vowel is usually appropriate” (*Data* 104). However, in diphthongal sounds there should be two points for formant measurement, “one near beginning, but not so close as to be part of the consonant transition”, and the second should be “near the end, but again sufficiently far from any consonantal effects” (150). Kent and Read cite 50 ms as “one fairly reliable temporal constant of stop articulation” during which a transition takes place from a stop to a vowel and from a vowel to a stop (*Acoustic* 116). In the segmentation of tokens for analysis we had taken 50 ms to be an estimate for the transition Ladefoged refers to.

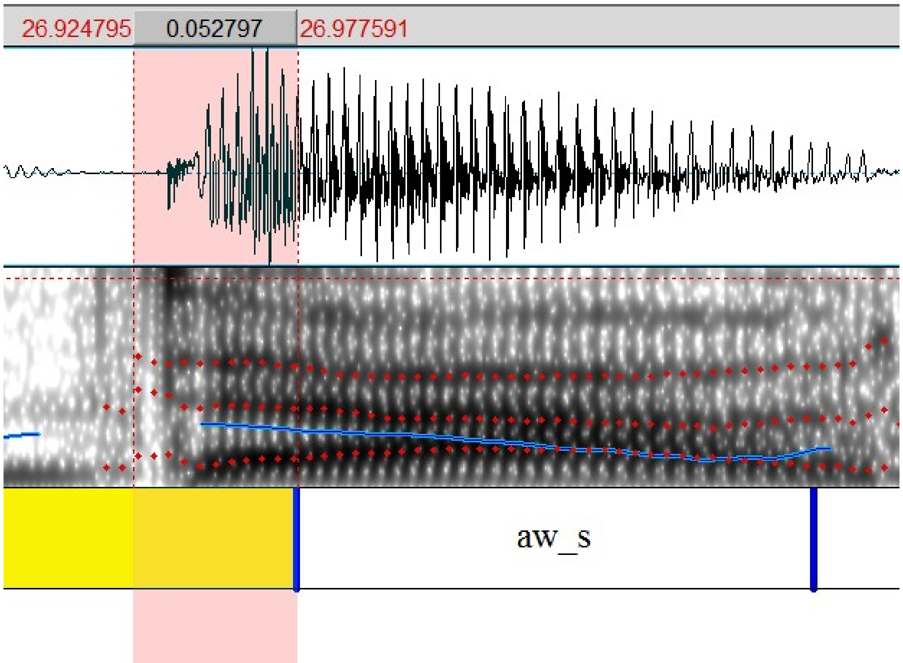

Figure . An approximate 50 ms after the plosive release, the region after which the beginning of a segment is placed (the first line in aw_s part). The waveform and the spectrogram show the word “doubt”. The formant and pitch lines were visible since they help in segmentation.

During the segmentation phase, we used both approaches, each for the different use: F1 movement was usually a good reference for reaching the target value, so no measurement occurred before F1 reached the expected maximum. The measurement points of the two diphthong targets were placed approximately in the middle of the targets (the points were used to measure F1, F2, F3, intensity and pitch).

However, these were not sufficient techniques for the measurements larger than a single point: in the measuring of a whole-diphthong (and whole-word) duration, the segmentation included placing one boundary at the start of the TextGrid segment, and the other at the end. The temporal segmentation was no easy task. It was done at the first step, before placing points for the single-value measurements; it exacted several manual “passes” through the corpus and detailed inspection of waveforms, spectrograms, formants, and pitch (they were all very important signals to determine where a word or diphthong approximately began, and where they ended).

This was a complex issue due to coarticulation, the variety of speaking styles our speakers employed, and the fact that our speakers were females. For example, Speaker 8 had no voicing in vowel that preceded the voiceless plosive. If we were relaying only on voicing as a criterion for a vowel, then this speaker would, in some examples, have had an extremely short vowel duration (which would not have been true, compared with other data and the place of the plosive release). In some words in data recorded by Speaker 14, a sudden pitch drop was evident just before the plosive, as a mark of the diphthong end, because the speaker was speaking so fast that other cues were hard to discern.

The fact that all analysed data originated form female speakers meant taking care about higher frequencies, spectral resolutions and formant cues. Even though we were aware of the differences between male and female voice, as well as of methodological issues,[6] our only concern (since we had no mixed corpus) was to take good measurements and have consistent procedures.[7]

The result of the above considerations was a set of rules for temporal segmentation of diphthongs, to some extent influenced by the theoretical assumptions in this chapter, and to some extent by our own in-practice observations:

1. After a plosive, a boundary was placed 0.04 to 0.05 seconds after the release or voicing. This was considered sufficient to reduce the influence of the consonant. With some speakers this could have been even more, with some less, but a precise boundary would have been extremely difficult to determine: it would have lacked consistency and would have been influenced by subjective factors.
2. The second boundary was placed after whatever a speaker pronounced that was supposed to be a diphthong. For example, if /ðɛətu/ was pronounced with “rhotic r” instead of /ə/ as the second target, the boundary was placer after /r/. This included, but was not limited to, the lack of voicing and pitch drop. This rule was crucial in determining the temporal domain of diphthongs (errors in words were on much lower scale, because words last longer).
3. When placing the second boundary before the fricative /s/, on average three cycles in waveform were indicative of the boundary limit. This seemed to be an interesting consistency throughout the corpus.
4. When preceded by a voiceless plosive, the second boundary was placed after the voicing in the segment ends, where applicable.

[6] Kent and Read in the chapter “Speaker Variables: Age and Sex” quote Titze: “One wanders, for example, if the source-filter theory of speech production would have taken the same course of development if female voices had been the primary model early on”. (154)

[7] More about this in *Praat* settings below.

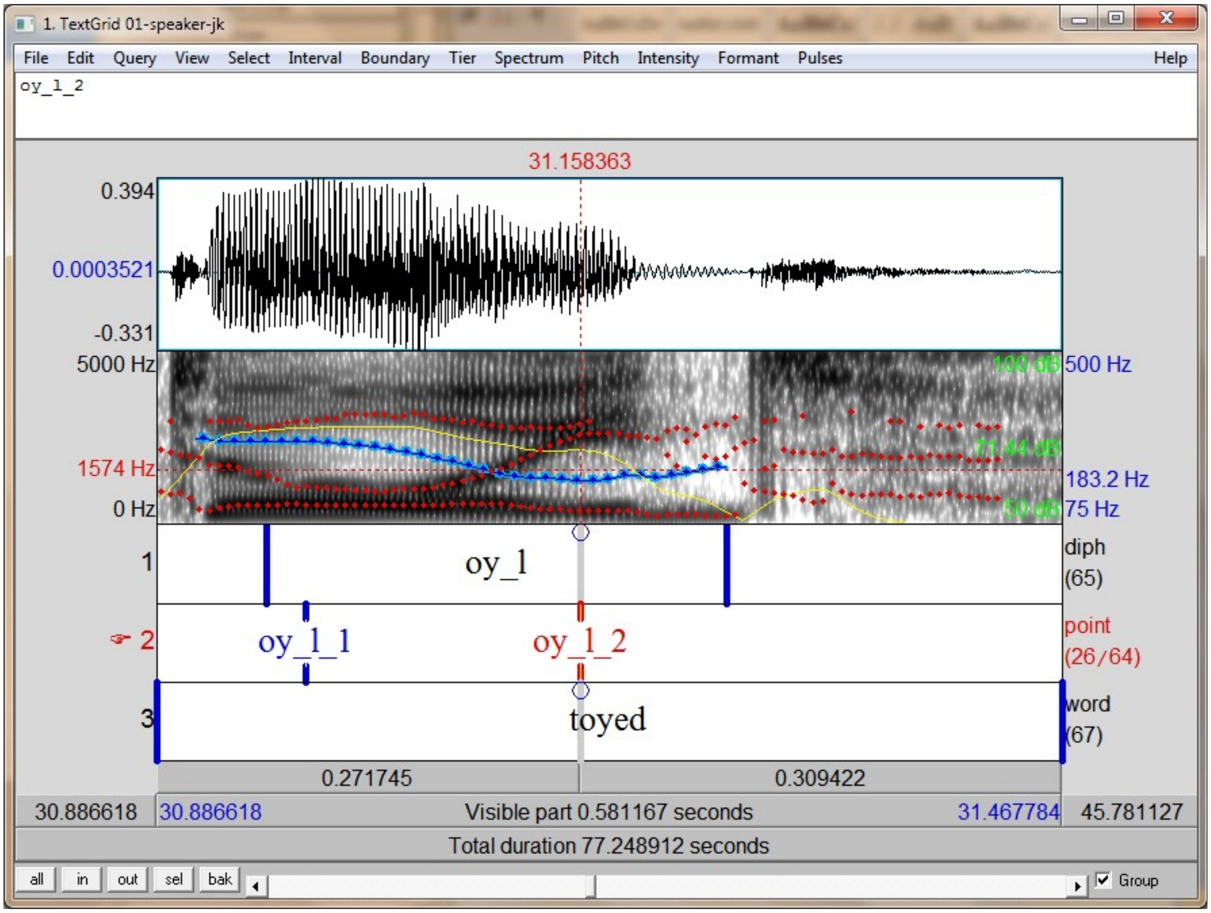

Figure . Selecting the target points for the diphthongs

The primary rules for selecting the points used as the places for measuring formant, pitch and intensity were:

1. The target points must be within the temporal boundaries.
2. The target points must be in the section where F1 reached sustained maximum value.
3. The target point must be selected while optimal upper formant frequency was set.
4. The target points must be within the expected range (formant curves must have expected forms).

4.9.1 *Praat* Settings Methods

Spectrogram settings in *Praat* were applied having in mind the instructions about the differences in measurements in male and female voices. Thus, we expected to find approximately one formant per 1200 Hz in our research (Ladefoged, *Data* 125), because all of our speakers were females.

Our primary references for the settings, segmentation and measurements came from Ladefoged, notes from the *Praat* help files, Weenin's instructions (*Speech Signal Processing With Praat*), Harrington, and other sources.[8] These were used to assemble a set of methods that we hoped to be suitable for an approximation (as any measurement is an approximation) of the digital data we recorded with our speakers.

The upper limit for spectrogram settings in *Praat* ranged from 3400 to 3800 Hz. The generally suggested value was about 1000 Hz for male speaker and 1100 Hz for female speaker per formant. After examining our data it was obvious that 1100 Hz, as suggested in *Praat*, was a very low upper frequency in formant calculations. When 3300 Hz, which

[8] Most of works about acoustic phonetics in our working bibliography had notes about the analysis.

corresponds to 3 formants in the 1100 Hz range, was set in *Praat*, the program could rarely determine the most probable F3 values. This was a problem, because of the narrow differences some speech sounds have in the F2/F3 range (most notably upper front vowels).

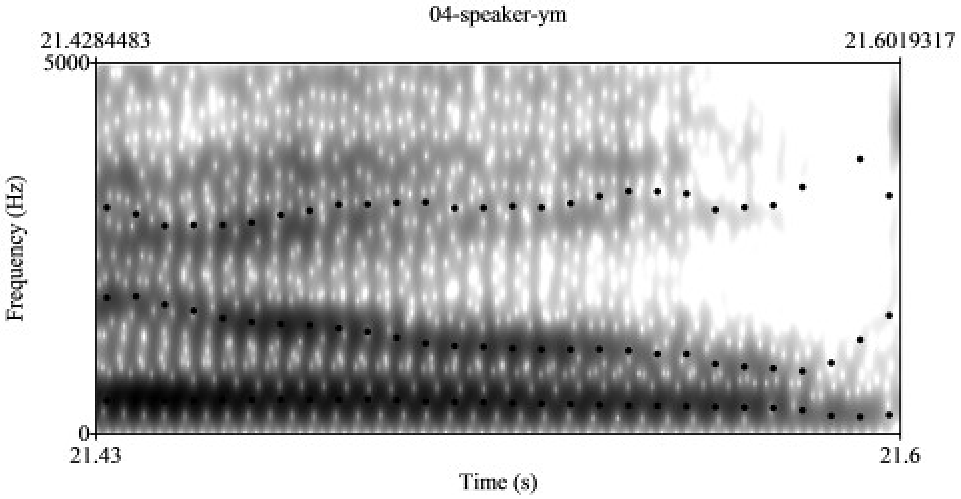

Figure . The diphthong and the formant marks from “joke”, after 3800 Hz was applied as the maximum frequency for the first three formants

However, our knowledge of formant distribution within vowels and vocalic articulatory features enabled us to introduce a consistent method for setting the upper frequency limit in formant calculations. The method consisted of gradual increase of the upper frequency for the calculation of the first three formants, until, in the region of /ɪ/, a significant amount of F3 values (re-calculated by a program) became visible (drawn upon the spectrogram). The vocal /ɪ/ was used in this step because of the near position of F2 and F3. Of course, even this is an approximate selection, but when accompanied with the spectrum observation, it provided a good orientation for measuring formant values within targets of the diphthongs.

In practice, this meant that the initial upper frequency setting was 3300 Hz, while the program was set to calculate only 3 formants. Because all of our subjects were young adolescent females, the upper frequency limits were expectedly relatively high. Table 13 shows the upper frequencies used in the measurement settings. They frequencies range from 3400 Hz (two speakers, one of which is our referent speaker, who is in the upper age bracket) to 3900 Hz. The average value is 3620 Hz, with a deviation of 112.5 Hz.

Table

Maximum frequency settings for formant calculations

Speaker	Frequency settings (Hz)
16-speaker-hz	3400
15-speaker-sr	3600
14-speaker-ao	3500
13-speaker-jr	3900
12-speaker-vv	3600
11-speaker-dz	3600
10-speaker-st	3500
09-speaker-tz	3770
08-speaker-ni	3650
07-speaker-lc	3600
06-speaker-ip	3700

05-speaker-gl	3500
04-speaker-ym	3800
03-speaker-im	3800
02-speaker-jj	3400
01-speaker-jk	3600
Average:	3620
Average deviation:	112.5

4.9.2 *Praat* Scripting and *R*

After we created the TextGrids we wrote a *Praat* script to automate the measurement, which was possible because *Praat* has a feature for task automation. The measurements could have been done manually, but this would have been strongly influenced by factors of human error, which could have appeared during reading values of more than 2000 elements in the corpus. Scripting enabled not only the reduction of possible human error, but also consistent and easily manageable measuring.

The script was written to give joint measurements for the signal files. Thus, we had one folder with our corpus files (30 files, out of which 15 sound files and 15 TextGrids), and another for referent files (2 files). The script run by loading all TextGrid file names into *Praat* working space. Once the files were enumerated, the first TextGrid and the corresponding sound file were loaded. The second step was to create analysis objects,[9] while beforehand applying upper limits for each recording individually, as specified in the table above.

The calculations were made on the file level, instead of on the diphthong level which means that the measurements were taken by passing all file content to the program, not just the bits containing words or diphthongs. By this we avoided the influence of “analysis window” (Johnson 32) which reduces important data near the ends of segments.

The next step in the script[10] was to calculate the lengths of diphthongs and words, while checking if they were all present in the TextGrid. Afterwards, pitch, formant, and intensity calculations were read per point (the first and the second diphthong target). The lengths were saved in one file, while formants, pitch and intensity in another.

All data was saved in tab-delimited text files with proper column headings: *file/speaker handle, diphthong, word, time, f1, f1, f3, pitch* and *intensity* labels.

R: A Language and Environment for Statistical Computing

R, or officially *R: A Language and Environment for Statistical Computing*, was our primary tool for statistical analysis, calculations, and drawing graphics. Its value for this paper was immense. The analysis of data in *R* included learning the *R* programming language[11] in order to produce the results and to attest the validity of calculations.

The code in *R* was divided into several sections, the most notable being *fpi* (formants, pitch, intensity), *time* and *graph*. These sections processed the measurement data, and filtered them in the same time. Thus, we had mean values of the first and the second targets for both the corpus and referent recordings. The same sections were used to create calculations from the Marković corpus.

[9] A Praat element with (in our case) calculated formants, intensity and pitch..

[10] One part of the script was more elaborate, and involved a special reading of the data. The time of the diphthong (n) is divided by 10, and for each n1 time span, formant values were read and saved. This file was later to be used in determining the phases of diphthongs and target attainments. However, the analysis proved to be methodologically very challenging.

[11] The author would like to thank people from the official #R channel on Freenode Network for their unselfish help during programming.

The second important aspect of *R* was the plotting function. We created several plotting elements that were used to generate many images in this paper. The basic idea revolved about the F1/F2 graph (Ladefoged, *Data* 131) on which other elements were placed: individual IPA signs or ellipses.[12]

Mlinar, Romeo.
"Pronunciation of English Diphthongs by
Speakers of Serbian: Acoustic Characteristics"
MA Paper. Novi Sad: University of Novi Sad, 2011.

The paper is available at:
<http://www.languagebits.com/files/ma-paper>

[12] Ellipses were calculated using the “car” package for R.