Predicting and Analyzing Segmental Durations in Sengwato Word-List

Shuo Zhang

December 21, 2013

Part I
Setswana data set and methodology

1. Introduction

The data set in the current project is drawn from the Setswana Project (funded by National Science Foundation, PI: Elizabetn Zsiga, 2011-13). Speech audio data on several dialects of Setswana (spoken in Botswana) are collected over three summers in Botswana by E.Zsiga and O.Tlale (2010-12). In this project we will analyze diclect data from the Sengwato region (with pilot study on Pitsane), and focus on a word list read by speakers from all dialects. An overview of the speakers and dialects is in Table 1.

Table 1: Dialects and speaker overview

dialect	number of speakers used to build model	previous data processing
Sengwato	8	8 speakers segmented and annotated by hand
Pitsane	7	7 speakers annotated by Autoseg 3.0 and corrected by hand

The data of the two dialects are analyzed in this project. In the initial stage, pilot analysis is done on Pitsane dialect. However, since Pitsane was first annotated by Autoseg and then corrected by hand, there is reason to choose Sengwato as a more reliable data set for building the model (for instance, the word duration in Pitsane data is not accurately segmented). Therefore, to limit the scope of this term project, I choose Sengwato as my main data set throughout this analysis.

2. Research questions: predicting segmental durations

The motivating goal of the current project is to build a linear regression model to *predict* the durations of targeted preceding vowel/N and consonant segments, as will be tested by a speech segmentation praat script implemented by me: Autoseg 3.0[1]. This goal is purely oriented towards obtaining a better computational tool that has better accuracy in automatic segmentation of this data set.

The second goal is to *analyze* the effect of preceding segments and consonant type on the duration of stop consonant closure. This goal is oriented towards analysis, and it originates from the analysis of Setswana

[1]AUTOSEG 3.0, by Zhang, see http://zangsir.weebly.com/autoseg-30.html for a complete description and documentation.

Table 2: list of linguistic predictor environments for target stops

stops	b, p, t, d, k
preceding	{null}, ore, xo, N(asal)
following	a, i, u

post-nasal devoicing in Setswana Project. Setswana post-nasal devoicing, illustrated by alternations such as [bat'a] look for, [m-p'at'a] look for me, has received the most attention in the literature, because it goes exactly against a purportedly universal tendency to avoid voiceless obstruents after nasals (Hayes 1999). However, the few phonetic studies that have examined this process in Setswana (Coetzee, Lin, & Pretorius 2007; Tlale 2005; Zsiga, Gouskova, & Tlale 2006) have found a great deal of variation in the phonetic realization of underlying voiced consonants in different positions, suggesting that devoicing is not the best analysis. Gouskove, Zsiga and Tlale (2011) found that consonant closure duration is significantly longer in post-nasal positions, drawing correlation between the duration and the voicing of stops in different environments.

To analyze this pattern, the current word list is designed to cover five types of stop segments /b/, /p/, /d/, /t/, /k/, in a variety of preceding and following environments. The environments can be divided into three types: in isolation (stop is word-initial), intervocalic (2 types of previous segments xo- and ore-), and post-nasal. The choice of following vowel reflects the range of possible combinations of stops and the following vowels in the Setswana inventory. Table 2 provides a summary of the linguistic environments of interest in this study.

Due to the limited space of this term project, I pursue only the first goal here in the current analysis.

3. Methodology

3.1 Overview

In this project, I look at predicting and analyzing segmental durations from a purely statistical perspective. In doing this, only two sets of independent variables are taken into account as predictors: duration data of surrounding segments, and the type of segments in the current word. This approach does not take into account the distribution of various lower level acoustic features that can be extracted from speech audio signal as a cue of segmentation. Nor does it take into account the interaction between duration and other types of high-level measurements in our data, such as percentage of voicing for the consonant closure. In doing so, I limit myself in pursuing a computationally less-expensive method to predict and analyze the segmental durations. In the mean time, the accuracy of analysis and prediction may suffer as a consequence of not knowing what happens underlyingly exactly. In this section, 3.2 addresses methodological issues in *duration prediction.*

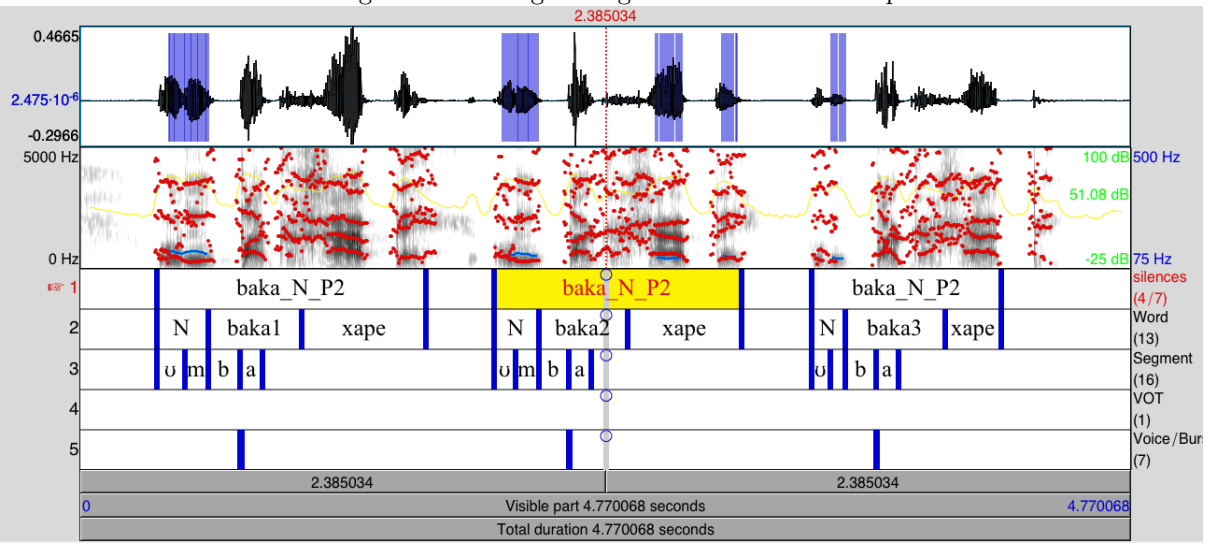
3.2 Predicting durations

The first task is to understand what is the original capacity of Autoseg 3.0, and what information exactly do we need to predict in order to build an accurate segmentation script.

The basic functions of Autoseg 3.0 consist of segmenting and labeling sounds file in Setswana project. Users can load a batch of sound files into Praat Object window and run the script to derive a textgrid for each sound file, containing the segments and labels in three tiers: Token, Word, and Segments (there are two additional tiers not labeled). This is shown in figure 1.

For the purpose of the Setswana Project, we are interested in predicting the durations on the segment tier, even though all three tiers need to be segmented automatically. In the current version of the Autoseg, only one level of reliable automatic segmentation is carried out using the *acoustic features* in the speech signal, that is, the token level (tier 1) segmentation according to a threshold of intensity (as well as a threshold of silence duration). Due to the particular features of speech sound in this data, this token level segmentation is more accurate on the onset of the token (where the onset is no a fricative) and always a bit less accurate

Figure 1: Autoseg 3.0 segmentation: N-baka-xape



on the end point of the token (where there is a decline in intensity before the actual ending of the word). Since we're not interested in the exact location of the end of the word (for the task of Setswana Project), this level of segmentation is good enough and does not need additional correction by hand (or to be improved by modeling). On all other tiers, segmentation is not done with any information from acoustic signal, and duration-based boundaries are inserted in the textgrid purely based on prediction of statistical analysis from the duration of the hand annotated data (in the past, this is the mean of the targeted segment in the hand annotated data).

The true targets of our prediction are the duration of three segments on the `Segment Tier`: The consonant closure (/b/ in Figure 1), the previous segment(/m/ in Figure 1), and the following vowel(/a/ in Figure 1). On the `Word Tier,` the duration of the previous syllable is also relevant (N in Figure 1), whereas the word duration of 'baka' and 'xape' are not as crucial (and are in general assigned a mean duration in each environment). Therefore, the target of the prediction is a set of four durations: N, m, b, a (in Figure 1)[2]. To predict a set of four durations from purely statistical knowledge of durations is a difficult challenge, and low accuracy is expected.

To reduce our level of difficulty, I performed an additional set of experiments to use intensity change in speech signal to automatically segment the consonant closure from the sourrounding sonorant segments. Using 0.05s as the minimum silent interval duration and -25dB as threshold for silence, I successfully segmented the previous syllable (N) from the consonant closure (b) (see Figure2). This reduces our task by half: of the set of four unknown durations, now we only need to predict two (Previous segment duration (/m/) and following vowel duration (/a/)). An additional advantage of this is that we also have more continuous predictors that are known: namely, the durations of the PreN, the consonant closure, and the whole word duration.

3.3 Data Preprocessing

In this step, several praat scripts are developed in order to format and extract information from the textgrid into a csv format for analysis in R. The principle components of this set of tools include:

1. `add-tier-word.praat`: This script format the original hand-labeled data of Sengwato to include a 'token' tier, in order to conform to the format of Autoseg segmented Pitsane data and the requirements of data structure of the `extract-info.praat` that produces the csv files.

[2]Alternatively, u, m, b, a. But among the duraiton of N, u, and m, we only need to measure two of them and we can derive the third.

Figure 2: Automatic segmentation by intensity of N-baka-xape: segmental level

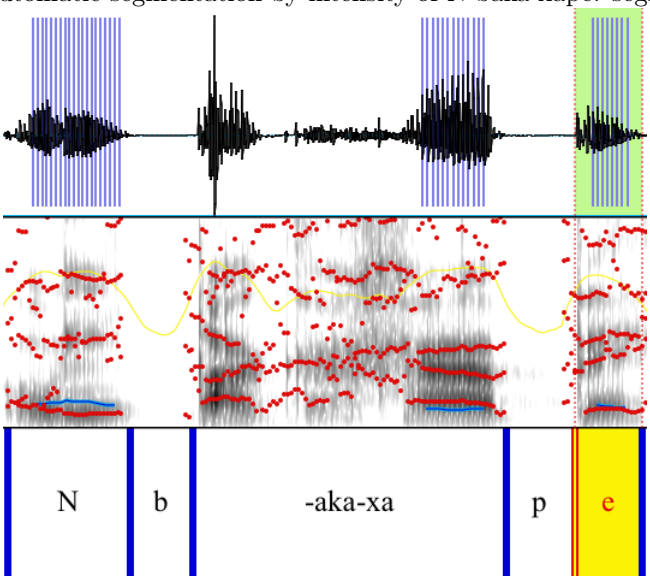


Table 3: List of variables in the csv file

variable names	PreSeg	PreSegDur	Pre1	Pre1Dur
example in Figure 1	N(tier 2)	N duration	/m/(tier 3)	/m/ duration

variable names	C(consonant type)	Cdur	WordDur	FolV
example in Figure 1	/b/(tier 3)	/b/ duration	duration of this token	/a/ (tier 3, following /b/)

variable names	FolVDur	Speaker	PreEnv
example in Figure 1	/a/ duration	P2 (P=Pitsane, S=Sengwato)	N(=nasal, or x=xo, o=ore, iso=isolation,)

2. `extract-info.praat`: This is a main script to extract segment types and segment durations for each token of the repetition (there are typically three or four tokens/repetitions in a file for the same word iterm). The script iterates through all files for all speakers specified under a directory, and produce a csv file for each speaker. Below is an explanation of the column names in the csv files (Table 3).

3. `clean-data.praat`: After a csv file is produced, this script is used to clean up the csv file to eliminate any irregular notations in the coding of the segments (these are due to mistakes and inconsistencies made by hand annotation by multiple annotators). This ensures that all csv files have a uniform format with consistent notations and symbols.

3.4 Statistical Analysis

The csv files of multiple speakers for each dialect are then combined into a data frame using `rbind()` and statistical analysis henthforce is carried out in R.

Part II

Results and discussion

4.Exploratory data analysis on Sengwato

In this section I show selected plots to explore patterns in the data for the prediction of previous segment, following vowel duration, and for analysis of consonant duration. For the complete set of plots and codes, please view the attached R file.

4.1 Prediction of Pre1Dur (Previous segment duration) and FolVDur (following vowel duration)

4.1.1 Linguistic factors

4.1.1.1 Continuous (duration) attributes as predictors

Pilot analysis on Pitsane shows that linear models with continuous attributes as predictors have more predicting power than categorical attributes. Here I explore the distribution of Pre1Dur and FolVDur (dependent variables, to be predicted) as a function of other known duration attributes: WordDur, PreSegDur, and Cdur.

Figure 3: Pre1Dur by (a)WordDur, (b)Cdur, (c)PreSegDur

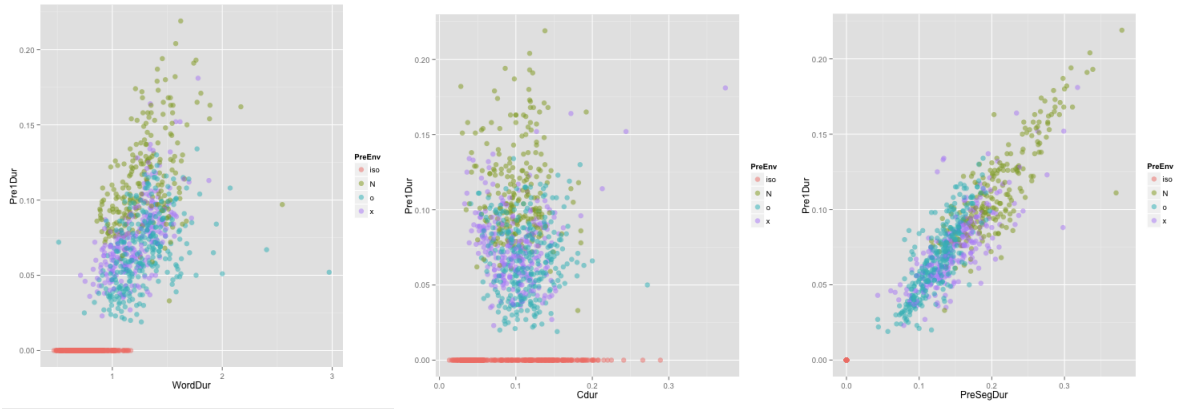


Figure 3 shows a series of plots exploring the three known continuous attributes as predictors of Pre1Dur (grouped in color by previous environment PreEnv). First, it shows that Pre1Dur is somewhat correlated with WordDur, and not surprisingly, quite well correlated with PreSegDur (notice that the red dots represent the isolation case, in which the Pre1Dur and PreSegDur are always 0). Second, of the three meaningful groups here, the intervocalic environments (o=ore, x=xo) seem to pattern together in one cluster, where as the nasal condition has a more distinct pattern. This pattern is overall characterized by the longer duration of Pre1Dur for the nasal /m/ or /n/. In addition, in Figure 3(c) it clearly shows that nasal prefix has both longer syllable duration as well as longer nasal segmental duration. Third, this set of plots show that PreSegDur is an especially good predictor to separate out the post-nasal cluster and intervocalic cluster.

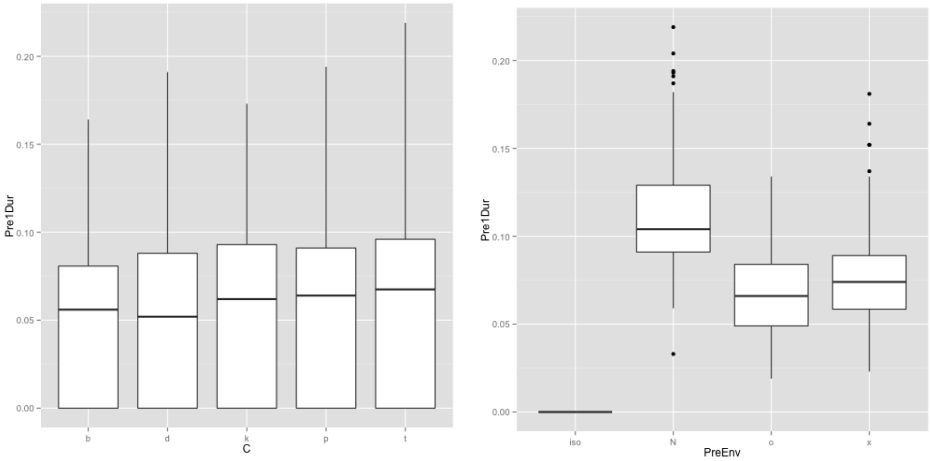
Figure 4 shows a set of plots exploring the three continuous predictors as functions of FolVDur. The pattern in this set is less clear: FolVDur form a dense single cluster, except in isolation case where the vowel duration is predictably longer than other conditions, while the word duration is shorter. Otherwise we do not see any interesting patterns that makes prediction easier.

Figure 4: FolVDur by (a)WordDur, (b)Cdur, (c)PreSegDur

4.1.1.2 Categorical (type) attributes as predictors

In this section I explore the categorical predictors: PreEnv and C for Pre1Dur (Figure5), and FolV and C for FolVDur(Figure 6). Here we see that there are some slight differences in the overall durations among different consonant types. Meanwhile, as we have observed, nasal prefix has longer duration than intervocalic prefixes overall. Therefore, PreEnv is a good predictor of the Pre1Dur.

Figure 5: Pre1Dur by PreEnv and C



For predicting FolVDur, we see that there are quite an overlap among consonant types in their following vowel duration. Figure 6 also shows that there exists a considerable difference among vowel types in terms of vowel duration: dur(a)>{dur(i),dur(u)}, with a significant overlap between /u/ and /i/.

4.1.2 Sociolinguistic factors

In the current data set I explore the effect of AgeGroup and Sex as sociolinguistic predictors of durations. As shown in Figure 7, both durations are somewhat positively correlated with age: older AgeGroup has longer durations. In the current data set, we also consider the fact that there are two subjects in the 'old' category (age 51 and 85) and both have a slower speech rate in the reading task (confirmed by a WordDur by AgeGroup plot, not shown here). It is not clear whether their age difference (which raises issues for grouping them together) or age similarity has an effect on duration, or it is random. On the other hand, plots by Sex (not shown here) show that there is not a significant difference between male and female speakers in their durations (even though females have overall slightly longer FolVDur).

Figure 6: FolVDur by C and FolV

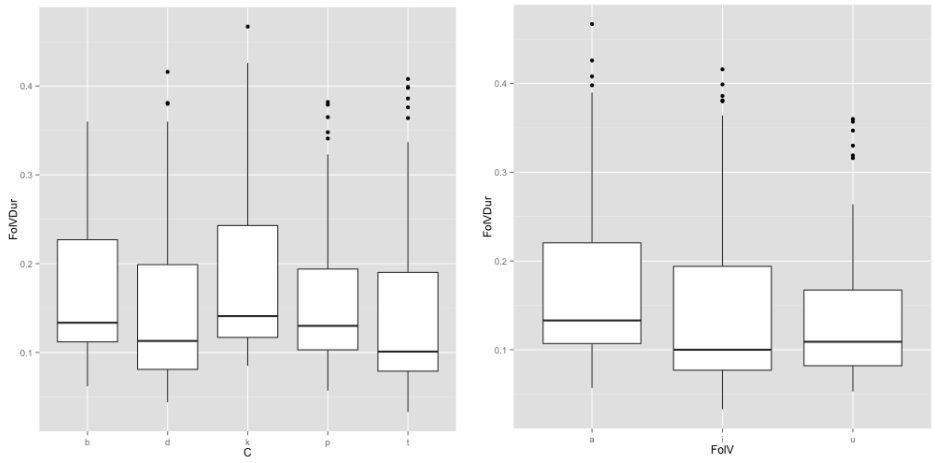


Figure 7: Pre1Dur and FolVDur by AgeGroup

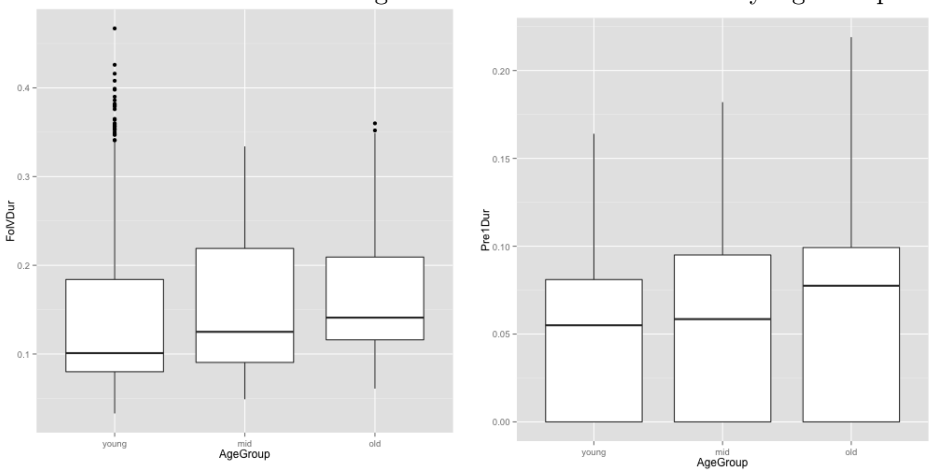


Table 4: linear model: Predict Pre1Dur and FolVDur using continuous attributes predictors

model	dependent variable	formula	Residual Standard Error	Multiple R-squared	F-statistic	evaluation
1	Pre1Dur	lm(Pre1Dur ~ PreSegDur, p)	0.0131 on 1123 degrees of freedom	0.923	1.35e+04 on 1 and 1123 DF, p-value: <2e-16	n/a
2	Pre1Dur	lm(Pre1Dur ~ PreSegDur + WordDur, p)	0.0131 on 1122 degrees of freedom	0.923	6.76e+03 on 2 and 1122 DF, p-value: <2e-16	not significantly better than model 1
3	Pre1Dur	lm(Pre1Dur ~ PreSegDur + WordDur + Cdur, p)	0.0131 on 1121 degrees of freedom	0.924	4.54e+03 on 3 and 1121 DF, p-value: <2e-16	significantly better than model 1,2($p<0.01$)
4	FolVDur	lm(FolVDur ~ Cdur,p)	0.0785 on 1123 degrees of freedom	0.000306	0.343 on 1 and 1123 DF, p-value: 0.558	n/a
5	FolVDur	lm(FolVDur ~ Cdur + WordDur,p)	0.0668 on 1122 degrees of freedom	0.277	215 on 2 and 1122 DF, p-value: <2e-16	significantly better than model4 ($p<0.001$)
6	FolVDur	lm(FolVDur ~ Cdur + WordDur + PreSegDur, p)	0.0565 on 1121 degrees of freedom	0.483	349 on 3 and 1121 DF, p-value: <2e-16	significantly better than model5 ($p<0.001$)

5. Linear models: continuous and categorical attributes

In this section, I summarize the fixed-effects linear models based on (1)continuous attributes predictors; (2)categorical attributes predictors.

Summary of linear models based on continuous predictors is provided in Table 4. The continuous attributes of known duration values for PreSegDur, Cdur, and WordDur are proved to be superior predictors of Pre1Dur (and to a lesser extent, of FolVDur). Model 1-3 shows that more than 92% of data is explained by the three predictors, with PreSegDur as a main predictor, and Cdur as a significant predictor. The prediction of values is expected to be greatly improved based on this model.

Model 4-6 summarizes prediction of FolVDur. The Multiple R-squared value comes to 0.483, showing PreSegDur and WordDur as the significant factors. This is somewhat unexpected since the consonant is in the most approximity to the vowel, whereas the previous segment precedes the vowel by two segments. Nonetheless, the continuous predictor still explain for nearly 50% of the data.

Table 5 shows the linear models built by using categorical attributes as predictors. Comparing to models with continuous predictors, this set of models has less predictive power, nonetheless it still explain a good portion of the behavior in the data. Model 7 and 8 show that the Previous Environment is a significant predictor of Pre1Dur, whereas the Consonant Type is not. This is somewhat expected given the correlation between part of the prefix and the whole, which is why we do not want to mix these categorical predictors with the continuous ones, as there may be some overlap underlyingly in its explanatory powers of the data.

The FolVDur is far less well predicted with the categorical attributes in model 9 and 10 (multiple r-squared is 0.04). However, the outcome that Consonant Type is a significant predictor of FolVDur is somewhat unexpected. We also notice that this effect is limited, and only applies to /p/ and /b/. Model

Table 5: linear model: Predict Pre1Dur and FolVDur using categorical attributes predictors

model	dependent variable	formula	Residual Standard Error	Multiple R-squared	F-statistic	Summary	Evaluation
7	Pre1Dur	lm(Pre1Dur~ PreEnv,p)	0.022 on 1121 degrees of freedom	0.783	1.35e+03 on 3 and 1121 DF, p-value: <2e-16	PreSegdur(iso) < dur(N) < dur(ore) < dur(xo), p < 0.001 on all levels	n/a
8	Pre1Dur	lm(Pre1Dur ~PreEnv + C,p)	0.0221 on 1117 degrees of freedom	0.783	577 on 7 and 1117 DF, p-value: <2e-16	no significant effects of consonant type	not significantly better than model7
9	FolVDur	lm(FolVDur ~ FolV, p)	0.0773 on 1122 degrees of freedom	0.0325	18.8 on 2 and 1122 DF, p-value: 8.91e-09	FolVdur(a) > dur(i) > dur(u), p < 0.001 on all levels	n/a
10	FolVDur	lm(FolVDur ~ FolV + C,p)	0.0769 on 1118 degrees of freedom	0.0459	8.97 on 6 and 1118 DF, p-value: 1.39e-09	FolVdur (b) > dur (p), p < 0.05, effect of consonant type	significantly better than model9 (p<0.01)
11	FolVDur	lm(FolVDur ~ FolV + C + AgeGroup + Sex,p)	0.07397 on 1115 degrees of freedom	0.1195	16.82 on 9 and 1115 DF, p-value: < 2.2e-16	FolVDur(young, male) < dur (AgeGroup-Mid) < dur (Age-GroupOld) < dur (SexFemale)	significantly better than model 10(p<0.001)

11 significantly improves upon model 10 by including sociolinguistic factors AgeGroup and Sex. Consistent with our exploratory analysis, it shows that mid and old aged females has significantly longer vowel durations than young and/or males.

6. Individiual speaker variation and mixed-effects models

The result of fixed-effects models, especially with continuous predictors, proves sufficient improvement of our predicting powers over the current version of Autoseg (using mean duration). Moreover, plotting shows that there is not a great amount of individual variation in the current data set. Therefore, I will not include a discussion on mixed-effects model here (even though I explored some possibilities in my R file).

7. Model evaluation

The current project has been successful in improving the prediction of our two target durations, Pre1Dur and FolVDur by an multiple R-Squared value of 0.92 and 0.5, respectively. Continuous attributes have been proved to have much more predictive power than categorical attributes in deriving more accurate results using more precise predictor values. This could be in part due to overall variation in speech rate, i.e., correlation between the duration of the whole word and a particular segment. Analytically this may be unattainable or uninteresting in many cases; however, in this particular project, we have made available the precise values of these continuous predictors by using intensity-based automatic segmentation in speech signal. Computationally, I have demonstrated that a statistic model based solely on the distribution of duration, without consideration of acoustic features, can lead to a sufficiently good model in predicting the durations.

The extended model evaluation of the current project is to develop a tool that combines the functionality of Autoseg to extract information from the speech file, then possibly using R to generate duration values using the model built, and to send the values back to Autoseg in order to implement those values on the current speech file. The development of this more complicated tool is beyond the scope of this project, at which point a true test and evaluation will be made with unseen data.

References

[1] Coetzee, Lin, & Pretorius 2007. Post-nasal devoicing in Tswana. In Jürgen Trouvain and William J. Barry, eds. ICPhS XVI. p. 861-864. (pdf) (With Susan Lin and Rigardt Pretorius.)

[2] Hayes, B., 1999. Phonetically driven phonology: the role of Optimality Theory and inductive grounding. In: Darnell, M., Newmeyer, F. J., Noonan, M., Moravcsik, E., Wheatley, K. (Eds.), Functionalism and Formalism in Linguistics, Volume I: General Papers. John Benjamins, Ams- terdam, pp. 243–285.

[3] Hyman 2001.The limits of phonetic determinism in phonology: *NC revisited. In: Hume, E., Johnson, K. (Eds.), The role of speech perception in phonology. Academic Press, San Diego.

[4] Tlale, O., 2005. The phonetics and phonology of Sengwato, a dialect of Setswana. Ph.D. thesis, Georgetown University, Washington, D.C.

[5] Elizabeth Zsiga, Maria Gouskova and One Tlale. 2006. On the status of voiced obstruents in Tswana: Against *ND. In C. Davis, A. Deal, and Y. Zabbal (eds.) The Proceedings of NELS 36. Amherst, MA: GLSA. pp. 721-734.

[6] Gouskova, Maria, Elizabeth Zsiga, and One Tlale. 2011. Grounded constraints and the consonants of Setswana. Lingua 121, pp. 2120-2152.

[7] Lisa Zsiga & One Boyer. 2012 Phonological devoicing and phonetic voicing in Setswana. Paper presented at the 43rd Annual Conference on African Linguistics, March 15-17, 2012 at Tulane University, New Orleans.