Durations of words in high and low predictability contexts in Korean non-native English speech

Abraham Demoz Prize Submission

Date Submitted: April 17, 2008

Kelsey Mok
Page Piccinini

Acknowledgements

Ann Bradlow *for her brilliant expertise, for the idea, and being a wonderful mentor*

The Wildcat Corpus *for the recorded materials*

Kristin Van Engen, Rachel Baker, and **Melissa Baese** *for helping with the development of our protocols and advising with critical statistical tasks*

Kristin Van Engen *for graciously copy-editing*

Janet Pierrehumbert *for her guidance*

Rajka Smiljanic *for her wonderful advice*

Introduction

Past studies have demonstrated the presence of hyperarticulation of words produced in low predictability contexts compared to words in high predictability contexts (Lieberman 1963; Scarborough, 2006; Bradlow and Alexander 2007). Specifically, when speakers and listeners can predict an utterance final word they will not take as much care in its articulation as when the word is unpredictable. This hyperarticulation of low predictability words results in greater durations for low predictability (LP) than high predictability (HP) words. Such findings can be understood within the context of the Smooth Signal Redundancy Hypothesis (Aylett and Turk 2004), which states that low levels of redundancy in a word are made-up for with more careful articulation. If one thinks about redundancy in terms of the predictability of a word appearing in a given context, then this theory predicts that the lower likelihood of a word appearing would result in more precise articulation.

In the project documented by Bradlow and Alexander (2007), durations of words in high and low predictability context were measured. It was found that the final word durations of words produced by native speakers in low predictability contexts had a median of 461ms, which was greater than the final word durations in high predictability contexts which had a median of 450ms. This present study will address the question of whether or not non-natives perform the same way as natives for final word durations. The purpose of this experiment is to preliminarily identify the possibility of Korean non-native English speech showing the same pattern of final word durations in both high and low contexts that Bradlow and Alexander (2007) found for the native English speakers.

Where this duration difference lies is also of interest. One component of this change in duration is the process of reduction. Jurafsky et al. posited the “probabilistic reduction

hypothesis", stating that the more predictable a word is the more it is reduced (Jurafsky et al. 2001). In a follow-up experiment, Bell et al. looked at the specific phonemes of function words that were being reduced. They found that changes in vowel duration had the greatest affect for the decrease in duration across the entire word (Bell et al. 2003). This phenomenon has been confirmed by others as well, for as Rebecca Scarborough pointed out, "English speakers hyperarticulate vowels in low frequency words and words with high neighborhood densities relative to the same vowel in high frequency and low density words" (Scarborough 2006). Van Son and Van Santen conducted a study looking at several factors that could affect differences in word duration, one of these was the affect of frequency of the phoneme itself (Van Son and Van Santen 2005). They hypothesized that the more frequent a phoneme was, the more likely it was to be reduced. This hypothesis ended up being confirmed, particularly when the word containing the phoneme was in utterance final position. In the present study, the stimuli used were all of CVC words. Measurements were taken of the full word (CVC), beginning of word to end of medial vowel (CV), and the vowel in isolation (V). This was done in order to figure out where the duration differences between low and high predictability contexts lie compared to the study mentioned above.

While the bulk of work on duration difference of words and phonemes is confined to English, studies have been conducted in other languages as well. This phenomenon will be studied further in this study by taking word final durations in a set of Korean sentences that were modeled after the set of sentences in the Bradlow and Alexander study. These measurements will be compared to the pattern documented in English productions of words in HP and LP sentences.

Methods

Stimuli

Sixteen Korean non-native English speakers were chosen, half male and half female, with each half having an average accent rating of 6.4. These accent ratings were based on an experiment from the Wildcat Corpus. In this experiment, fifty subjects drawn from the Northwestern University 200-level linguistic classes listened to 60 readings by native and non-native speakers of the Stella Passage (The Speech Accent Archive, 2008). The recordings were drawn from the Northwestern University International Summer Institute (ISI) students who participated in a two-part experiment. The first part of the experiment involved completing a task with a partner, thus producing spontaneous speech. In the second part, recordings were taken of: 1) specific words, 2) sentences, and 3) the Stella Passage. The Accent Rating experiment involved the subjects listening to all 60 passages (from all 60 ISI

Accent Ratings by Speaker

Male	Accent Rating	Female	Accent Rating
KO_12	6.89	KOF_05	7.11
KO_08	6.70	KOF_08	7.11
KO_06	6.61	KOF_04	6.89
KO_17	6.52	KOF_13	6.84
KO_01	6.34	KOF_09	6.77
KO_10	6.27	KOF_06	6.07
KO_04	6.20	KOF_10	5.27
KO_02	5.39	KOF_16	5.18
mean	*6.37*	*mean*	*6.41*

participants) and then rating those passages on a scale of 1 to 10 for accentedness, 1 being native and 10 being a very strong non-native accent. Based on these ratings, we were able to compile a selection of Korean non-native English speakers with the males having an accent rating range of 6.89 to 5.37, with a mean of 6.37, and females with a range of 7.11 to 5.18, with a mean of 6.41.

The specific set of sentences from the Wildcat Corpus were developed by Bradlow and Alexander (2007). From the list of 60 sentences (30 high predictability contexts and 30 low predictability contexts) from the Bradlow and Alexander (2007) experiment, a total of 10 sentences (5 high predictability contexts and 5 low predictability contexts) were chosen with final words "sheets", "feet", "sleeves", "bird", and "days" (please refer to appendix for the contexts in which these words were spoken).

These words were chosen because of particular properties and landmarks that were easy to identify for measurement purposes, for example, fricatives and stops in the onset and offset of the final words. A further specification for these words was that the word preceding the target word (final word) did not end with a fricative which would otherwise result in a double fricative which would make the onset of the target word difficult to identify. There were instances where the plain speech articulation caused two fricatives to line up in a row, for example, in the low predictability context sentence for "sleeves": *He looked at the sleeves*, where the schwa before "sleeves" was dropped. In all other cases, however, the target word was preceded by a word ending in a vowel. Between all of the speakers (n=16) and tokens (n=10) for each speaker, 160 word durations were taken.

Duration Measurement Protocols

Keeping in mind that non-native English speakers had the potential to use and articulate words differently, our protocols had to be flexible in the application of this experiment. Using Praat, three kinds of duration measurements (full word (CVC), beginning of word to end of medial vowel (CV), and vowel (V)) were taken using representational articulatory cues from the

waveform, spectrogram, and audio in that order. Boundary judgments and decisions were supported by at least two out of the three cues from the waveform, spectrogram, or audio.

Choosing the boundary for the onset of the target words was fairly straightforward. If the target word had a fricated onset, the signaling cue was a large amount of energy in the waveform and spectrogram; and in more detail, high frequency activity in the spectrogram and the beginning of irregular waves on the vowel regularity lead-in to the target word. Similarly, if the target word began with a plosive, the identifying cue was a burst of energy in both the waveform and spectrogram. The decisions here were supported by the corresponding audio.

Dealing with offsets was much more difficult because of two factors: our target words were word-final and our subjects were non-native English speakers. Both factors played an ample role in creating voicing and articulatory elements and cues which made it difficult to judge where exactly the end of the offset phoneme was located. In some instances, for example, the offset phoneme was followed by voicing or glottalization and it was difficult to say whether or not the source was the speaker or an external element such as a paper turn. If a word ended with frication (such as an /s/), the offset was marked where there was a reduction of high frequency noise; the end of a glottalized segment (as in the end of “feet”) was marked at the end of visible voicing in the waveform and voice bar in the spectrogram; and the end of a stop was marked where there was a visible qualitative loss of high frequency energy in the spectrogram and waveform after the release.

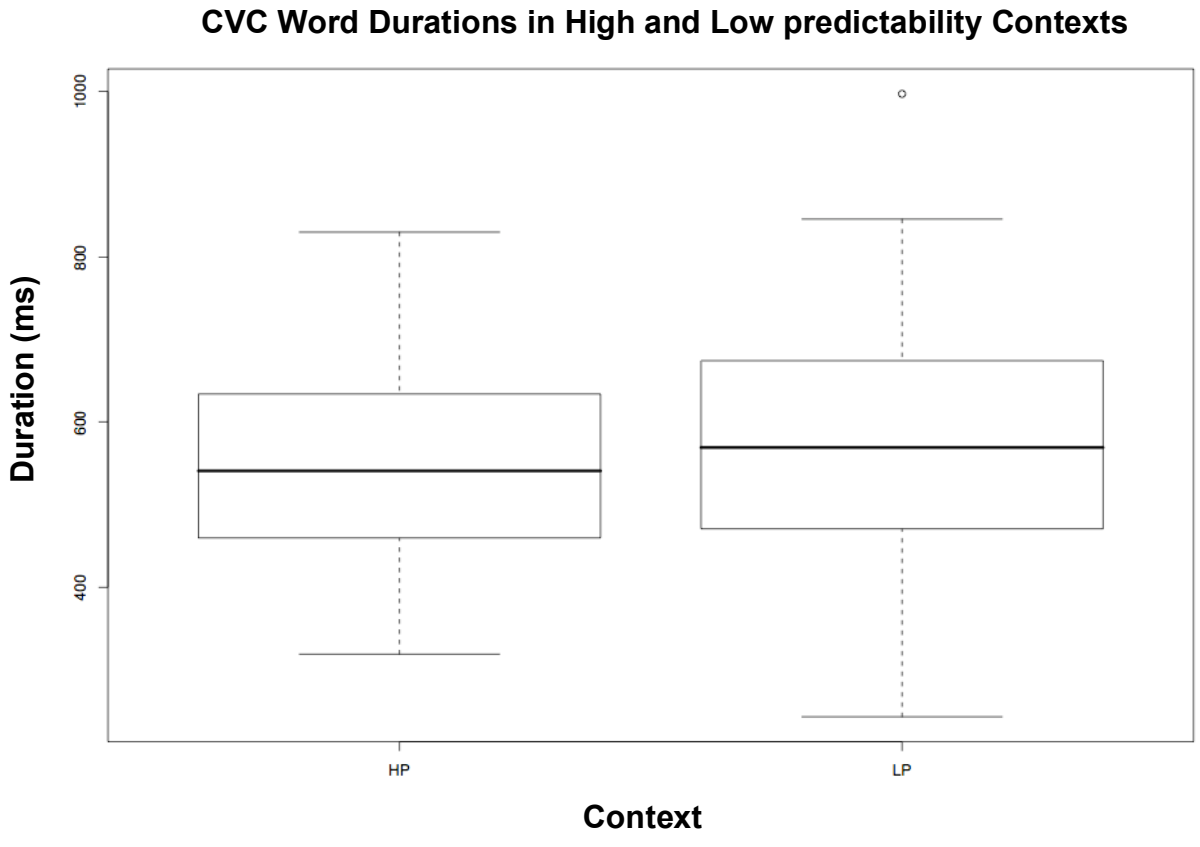
The phoneme /r/ as in “bird” was included as part of the vowel because most of the time the landmarks marking the end of the vowel and start of the approximant were vague and unclear. Because the target words were controlled in the word-to-word comparisons of this experiment, including /r/ within the vowel did not affect the particular results we hoped to find.

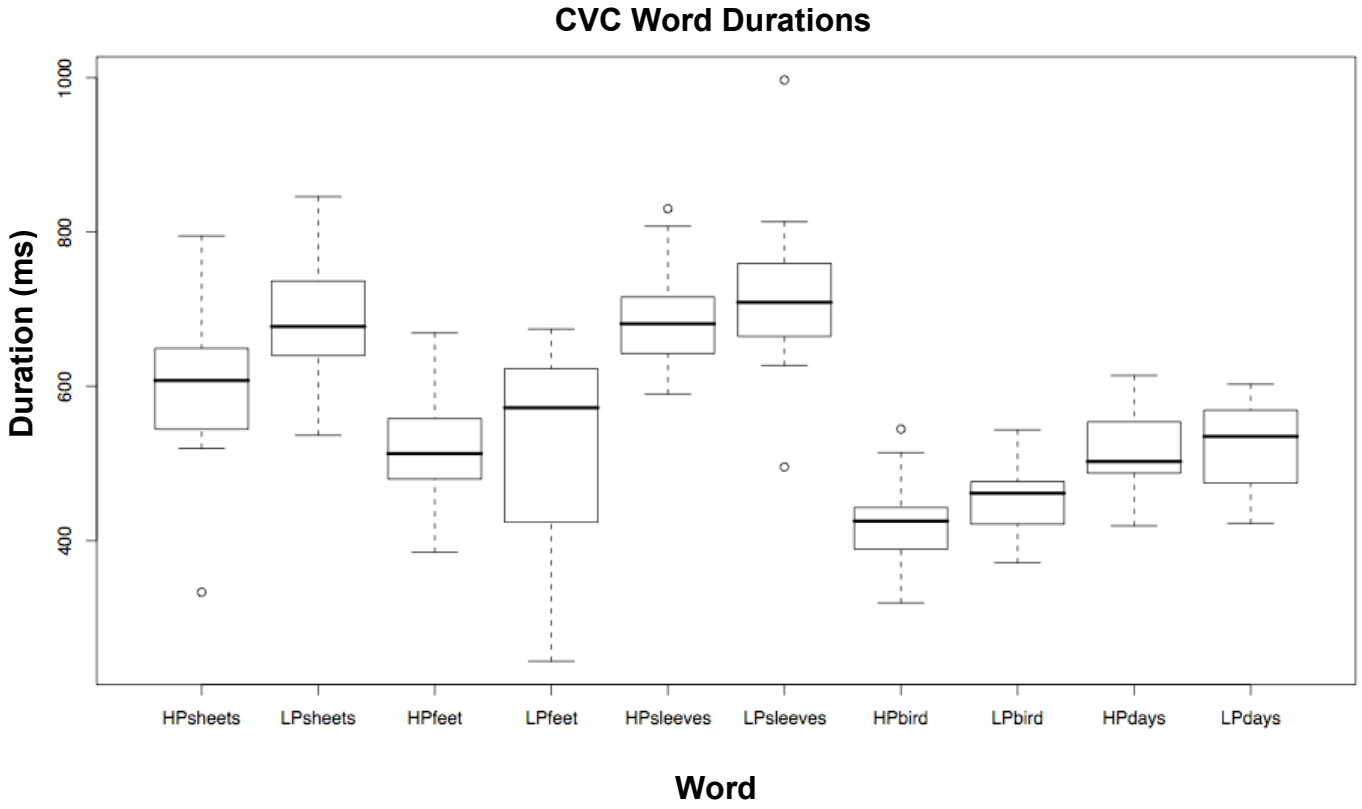
The protocols to measure to the end of the vowel were very similar to measuring the target word's onset: measure to the end of regularity of the vowel and do not include fricatives or plosive in the target vowel's offset with support in argumentation from the waveform, spectrogram, and audio. In measuring the vowel duration of "sleeves" the /l/ was included as part of the vowel just as /y/ and /r/ were, since that produced clearer boundaries and more consistent measurements.

Results

CVC words

A statistical analysis of the mean and median data for the full word (CVC) measurements showed that Korean non-native English speech did not follow the same patterns as native English speech in terms of the final word in high predictable contexts having a shorter duration than in low predictable contexts.





Statistical landmarks in final word CVC durations in HP and LP contexts (milliseconds)

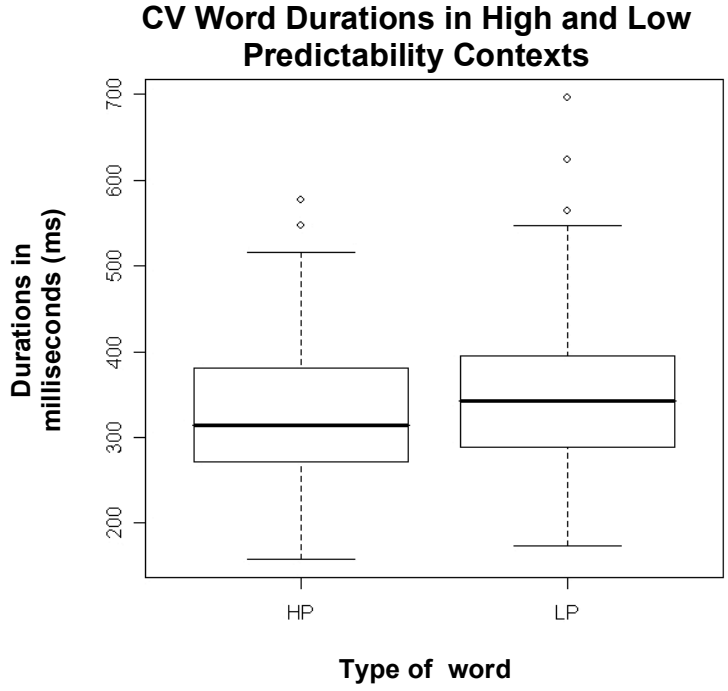
	HP sheets	LP sheets	HP feet	LP feet	HP sleeves	LP sleeves	HP bird	LP bird	HP days	LP days
mean	594.1	676.9	517.1	518.0	688.3	714.0	423.1	451.8	512.7	524.3
median	607.5	677.5	512.8	572.2	681.0	708.8	425.4	461.4	502.5	535.1
std. dev.	96.3	80.3	77.7	131.7	71.3	107.6	54.5	47.4	54.3	57.0
min	333.2	536.6	385.0	243.7	590.0	495.5	319.3	371.3	419.3	422.4
max	794.8	845.8	669.2	674.1	830.1	996.9	544.6	543.5	614.1	602.8

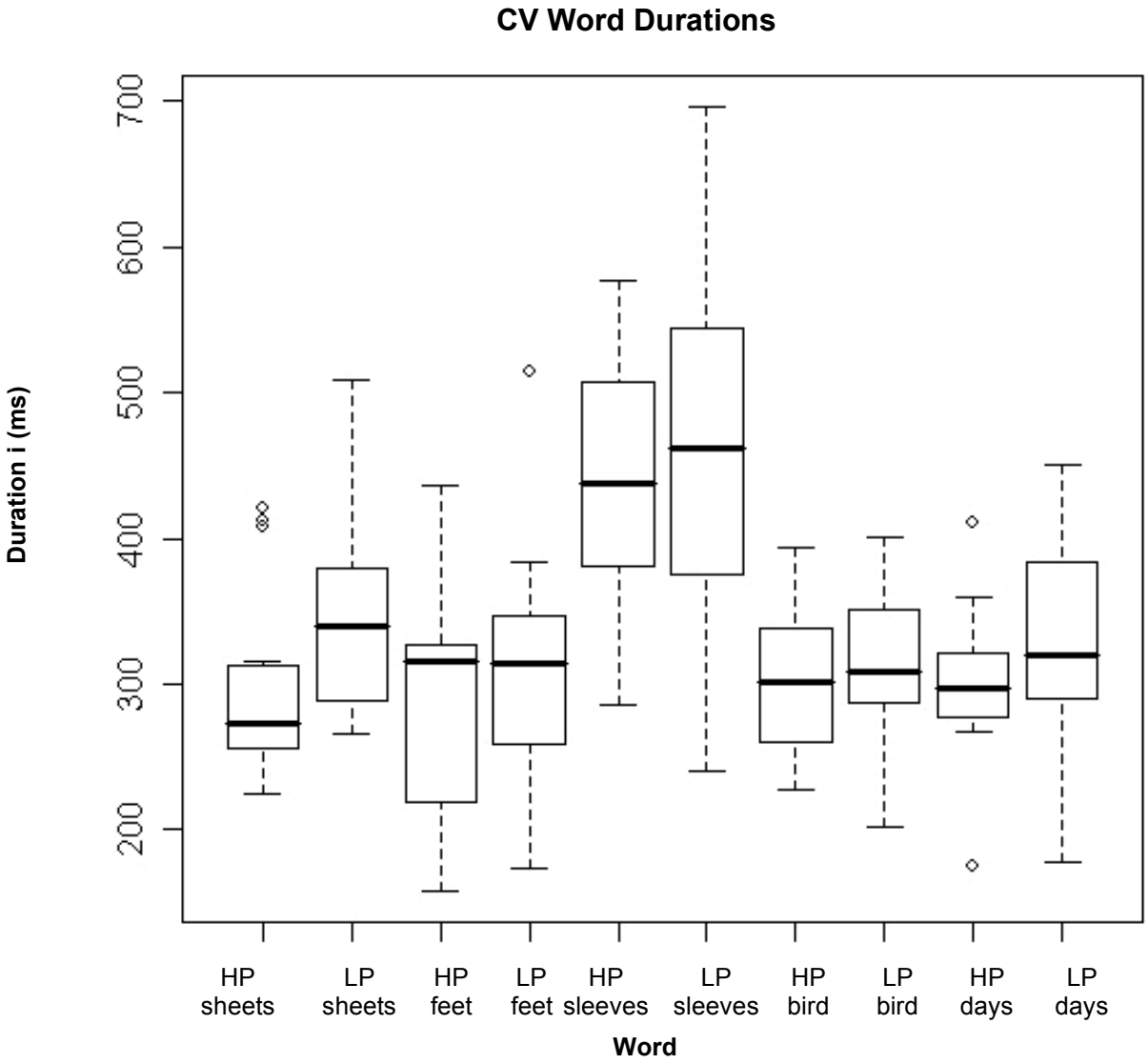
Two paired t-tests were performed on the data. The first tested the significance of the difference between the average duration of target word in high predictability context across all speakers and the average duration of the same target word in low predictability context. The second tested the significance of the difference between the average duration of all of the target words in high predictability context and the average duration of the same target words in low predictability context of all speakers against each other. The first t-test yielded a p-value of 0.1016 ($df = 4$) and the second t-test yielded a p-value of 0.01748 ($df = 15$). In accordance with

the mean data, the first statistical test shows that words in LP context are not longer than words in HP context; however, the second t-test shows that words in LP context are indeed longer than words in HP context when comparing within speakers.

CV words

Because the word durations in LP contexts were not significantly greater than the words in HP contexts, there was motivation to investigate this phenomenon further. In order to do this, a decision was made to choose another way to measure these target words. This second method dictated that the final consonant would be left off the measurement, leaving only the first consonant and vowel (CV) for analysis. The main reason to execute this method was because of the nature of phrase-final words, such as the target words featured in this study. After looking at some of the target words, it was noted that some final sound segments (in this case, the final C) in words in phrase-final position tended to take on characteristics that were reduced and difficult to measure with consistency.




CV Word Durations
Duration i (ms)
200
300
400
500
600
700
HP sheets
LP sheets
HP feet
LP feet
HP sleeves
LP sleeves
HP bird
LP bird
HP days
LP days
Word

Another visual analysis of the two preceding graphs of CV Word Durations shows that, similar to the set of CVC measurements, words in LP contexts are longer in duration that words in HP contexts. Again, the two paired t-tests that were performed in the previous measurement method (CVC), were again executed. The first t-test yielded a p-value of 0.01724 ($df = 4$) and the second t-test yielded a p-value of 0.001432 ($df = 15$). Because the p-value was less than 0.05, it can be said that this method of measurement yielded results that showed that words in LP context were greater than words in HP context.

Statistical landmarks in final word CV durations in HP and LP contexts (milliseconds)

	HP sheets	LP sheets	HP feet	LP feet	HP sleeves	LP sleeves	HP bird	LP bird	HP days	LP days
mean	299.46	343.97	295.43	307.61	439.5	463.29	301.95	312.86	300.04	330.06
median	273	339.06	315.4	313.67	438.37	461.73	300.98	308.01	297.31	320.33
std. dev.	62.23	66.49	78.65	79.05	80.3	115.69	48.26	59.03	49.96	68.84
min	224.61	265.34	158.38	173.52	285.44	240.74	226.97	201.52	175.1	178.26
max	420.27	508.95	436.41	514.88	577.5	695.84	394.12	401.29	410.81	450.92

This phenomenon was also shown to be true within the individual words that were measured with an exception for “feet” which in high predictability context, the median for the target word was greater than the median duration of the target word in low predictability context by 2ms. Possible reasons for this deviation include several incidents that were encountered with measuring the final vowel, for sometimes the speakers would elongate or shorten the vowel depending on how the speaker would produce the final /t/. For example, speaker KO_01 glottalized the final /t/ in “feet” in the high predictability context in which the vowel was significantly longer (257ms) and 1.93 times the vowel duration of the same word in low predictability context (133ms). In addition to this glottalization effect, there were times when no /t/ was pronounced at all, making it harder to decided where the end of the vowel was, since the word was sentence final and thus no additional articulation followed.

Of the other four other words that did show final words in HP context having a shorter duration than the final words in LP context, the following median differences were found in

descending order: “sheets” with a difference of 66ms, “sleeves” with 23ms, “days” with 23 ms, and “bird” with 7 ms.

Final CV word durations (medians)

	sheets	feet	sleeves	bird	days
HP	273	315.4	438.37	300.98	297.31
LP	339.06	313.67	461.73	308.01	320.33
Difference	66.06	-1.74	23.36	7.03	23.02

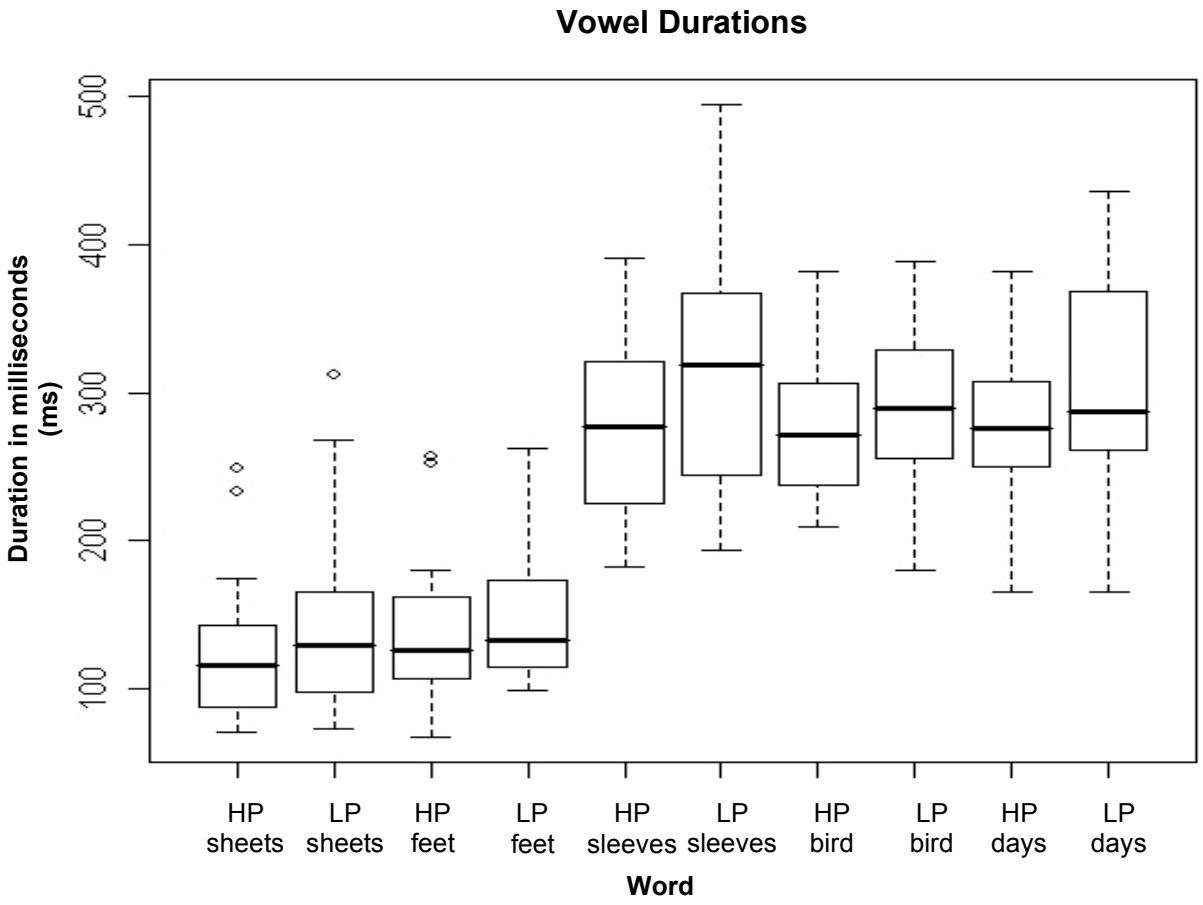
V words

Vowel duration measurements alone showed there were again greater durations for the low predictability contexts, including the word “feet”. This time, however, the duration differences did not follow the same ranking pattern as when /onset+vowel/ were measured: in vowel measurements alone the median differences were: “sleeves” with a difference of 42ms, “bird” with 18ms, “sheets” with 14ms, “days” with 12ms, and “feet” with 7ms.

Vowel lengths in final words (medians)

	sheets	feet	sleeves	bird	days
HP	115.52	125.58	277.43	271.4	275.74
LP	129.7	132.55	319.12	289.7	287.31
Difference	14.18	6.98	41.69	18.31	11.56

The two paired t-tests that were performed in the previous two measurement methods (CVC and CV), were repeated for the V data. The first t-test yielded a p-value of 0.02183 ($df = 4$) and the second t-test yielded a p-value of 0.0007556 ($df = 15$). Because both of the p-values were less than 0.05, the vowels within words that were in LP context were greater than words in HP context both between subjects and between words.



Discussion

In conclusion, all three conditions for which duration measurements were taken, words in LP context were longer than those in HP context; however, this was only statistically significant for CV and V durations. Although this relationship was only significant for one type of measure (namely, CV), this says something about the differences between the productions of native English speakers and non-native English (Korean) speech. While Bradlow and Alexander (2007) showed that the full length of words (CVC) were longer in low predictability contexts than high predictability contexts, the results from this study of non-native English (Korean) speech show that full length words (CVC) are not statistically longer in low predictability contexts than full length words (CVC) in high predictability contexts. The only measurements

(CVC, CV, or V) that did display longer word durations in low predictability contexts than high predictability contexts were those in CV form (without the final C) and V. Because the CV measurements showed significant differences between durations of words (CV) in high and low predictability context and the CVC measurements (including the coda) did not show this such said difference. The only separation between these two measures is that one included the duration of the coda (CVC) in the duration of the full word and the other (CV) did not. The motivating factor for the word having a statistically significant higher duration in the low predictability context than the high predictability context is whether or not the coda is considered as part of the word. Kim and Jongman (1996) made a note on the fact that "coda neutralization" occurs in the Korean language. In the analysis of the present study's final codas, it could be that because coda neutralization occurred so variably in the target words that the inclusion or exclusion of this phenomenon caused the results for CVC durations to be less correlated than if all of the Korean speakers had chosen to utilize coda neutralization or not.

Additionally, looking at the median differences between words, it is interesting to note that the greatest differences were found in these CVC words ending in consonant clusters (in this sampling all clusters ended in /s/) and that the words with the lowest median difference did not. From highest to lowest duration difference, the rankings can be seen as follows: 1) stop+/s/, 2) fricative+/s/, 3) fricative+/nothing/, 4) stop+/nothing/, and 5) stop+/nothing/. It is particularly important to see the difference between the second and third words, "sleeves" and "days". The difference in duration was only 0.34ms, but it appears that the addition of the word final /s/ for both words was enough to produce duration differences. Similarly, the two words with the lowest duration differences, "bird" and "feet" (the latter of which actually had an inverted difference: HP>LP), both ended in stops. This data suggests that there is some kind of hierarchy going on in

onset and vowel articulation of HP and LP words in regards to the articulation of the offset. In this hierarchy consonant clusters result in the greatest duration differences between contexts HP and LP, followed by fricatives alone.

In addition to codas, vowel duration in particular can be an important cue in whole word duration. It was found that vowels in words that were in low predictability contexts were longer than vowels in words in high predictability contexts; however, with regards to median difference the order of words did not follow that of the /onset+vowel/ measurements. “Feet” still continued to be the word with the lowest difference, but other than that no other words fell in the same ranking order by environment as before. Still, when looking not just at the phonemic environment following the vowel but the entire environment surrounding the vowel a pattern is discovered. (Note: As previously mentioned when measuring “vowel duration” the phonemes /r/ and /l/ were included as “part of the vowel” to provide more consistent measurements. In other words, when “environment surrounding the vowel” is discussed, it is really “environment surrounding /l/+vowel, vowel+/r/, or vowel.) The word with the greatest difference, “sleeves”, had a fricative for both an onset and an offset, the second highest difference, “bird”, had a stop on each end; followed by a word with a fricative onset and stop following the vowel, “sheets”; then stop onset fricative release, “days”; and finally fricative onset and stop offset, “feet”. The pattern present is that similar onset and offset environments around the vowels make it easier for a production difference to be made in vowel duration, particularly when the onset and offset are both fricatives.

After analyzing this data from multiple angles, while it is clear that non-natives follow natives in articulating words in high predictability contexts with a shorter duration than low predictability contexts, it is evident that the phonemic make-up of the word itself also plays a

significant role. Since the amount of data analyzed only qualifies for a pilot experiment, more measurements need to be taken in similar and different phonemic environments to confirm these results. With continued work with the Wildcat Corpus, we are hoping to expand this project to its full potential and acquire a firmer grasp on non-native articulation patterns of English words.

Appendix

Sentences Used in Bradlow, Alexander Experiment

High Predictability	Low Predictability
1. For dessert, he had apple pie.	Mom talked about the pie.
2. The color of a lemon is yellow.	Mom thinks that it is yellow.
3. In spring, the plants are full of green leaves.	She talked about the leaves.
4. She made the bed with clean sheets.	Dad talked about the sheets.
5. The sport shirt has short sleeves.	He looked at the sleeves.
6. He washed his hands with soap and water.	We talked about the water.
7. A wristwatch is used to tell the time.	This is her favorite time.
8. The war plane dropped a bomb.	Dad talked about the bomb.
9. The team was trained by their coach.	We read about the coach.
10. People wear shoes on their feet.	Mom looked at her feet.
11. When sheep graze in a field, they eat grass.	Dad pointed at the grass.
12. Football is a dangerous sport.	This is her favorite sport.
13. Bob wore a watch on his wrist.	He looked at her wrist.
14. Monday is the first day of the week.	This is her favorite week.
15. Rain falls from clouds in the sky.	Dad read about the sky.
16. A quarter is worth twenty-five cents.	He pointed at the cents.
17. People wear scarves around their necks.	She talked about their necks.
18. Birds build their nests in trees.	He read about the trees.
19. My parents, sister and I are a family.	We read about the family.
20. The good boy is helping his mother and father.	Mom pointed at his father.
21. A book tells a story.	We looked at the story.
22. A pigeon is a kind of bird.	We pointed at the bird.
23. At breakfast he drank some orange juice.	Mom looked at the juice.
24. Last night, they had beef for dinner.	He talked about the dinner.
25. A racecar can go very fast.	She thinks that it is fast.
26. Many people like to start the day with a cup of coffee.	Mom pointed at the coffee.
27. I wear my hat on my head.	She pointed at her head.
28. February has twenty-eight days.	There are many days.
29. We heard the ticking of the clock.	She looked at the clock.
30. Elephants are big animals.	He pointed at the animals.

Sentences Used in Our Experiment

High Predictability	Low Predictability
1. She made the bed with clean sheets.	Dad talked about the sheets.
2. The sport shirt had short sleeves.	He looked at the sleeves.
3. People wear shoes on their feet.	Mom looked at her feet.
4. A pigeon is a kind of bird.	We pointed at the bird.
5. February has twenty-eight days.	There are many days.

Bibliography

Aylett, M., and Turk, A. (2004). "The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prominence, and duration in spontaneous speech," Lang Speech 47, 31-56.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," Journal of Acoustic Society of America 113 (2), 1001-1024.

Bradlow, A., and Alexander, J. (2007). "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners," J. Acoust. Society 121, 2339-2349.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). ''Probabilistic relations between words: Evidence from reduction in lexical production,'' Frequency and the Emergence of Linguistic Structure, 229–254.

Kim, H., and Jongman, A. (1996). "Acoustic and perceptual evidence for complete neutralization of manner of articulation in Korean," Journal of Phonetics 24, 295-312.

Lieberman, P. (1963). "Some effects on semantic and grammatical context on the production and perception of speech," Lang Speech 6, 172-187.

Scarborough, R. (2006). "Lexical and contextual predictability: Confluent effects on the production of vowels", presentation at the Tenth Conference on Laboratory Phonology, Paris, France.

The Speech Accent Archive. (2008). George Mason University. http://accent.gmu.edu

Van Son, R.J.J.H., and Van Santen, J. (2005). "Duration and spectral balance of intervocalic consonants: A case for efficient communication," Speech Communication 47, 100-123.