Voice Onset Time in English Voiceless Initial Stops in Long Read and Spontaneous Monologue Speech of Thai Students with English as a Second Language

Chanakan Wittayasakpan
Chulalongkorn University

Abstract. Proximity of L1 and L2 VOT (Voice Onset Time) values illustrates that L1 stop categories are used to acquire L2 ones. While a number of studies have found that VOT is very sensitive to many factors, how speech time and VOT correlate is still understudied. This study tests whether L1 transfer in terms of VOT would enhance as speech time elapses. Six university students with Thai as L1 and English as L2 were selected to produce long spontaneous monologue speech and long read monologue speech produced. VOT values in English initial voiceless stops were then segmented and analysed using the mix-effects model. The result reveals that raw VOT in spontaneous speech is significantly shorter than in read speech and no correlation between adjusted VOT and elapsed speech time is found. does not correlate with elapsed speech time. Implications of the result are discussed in terms of stylistic variation and second language acquisition.

Keywords: voice onset time; second language acquisition; L1 transfer; Thai; English

---

1 Introduction

Voice onset time, hence shortened as VOT, is the temporal lag of vocal fold vibration following a stop consonant release, or, in other words, the interval between the burst of a stop consonant and the onset of voicing. VOT serves as a phonetic cue to distinguish voicing categories and is generally classified into three ranges: voice lead, short lag, and long lag. Across languages, VOT values could differ even in the same category, and this is the case for English and Thai. In both languages, stops are produced in three places of articulation: bilabial, alveolar, and velar. However, English has two types of stops: voiced and voiceless, whereas Thai has three types: voiced, voiceless unaspirated, and voiceless aspirated. English voiceless stops and Thai voiceless aspirated stops fall into the long lag VOT category (Kessinger & Blumstein, 1997) but have different ranges of VOT values. VOT values in monolingual English voiceless stops range from minimally greater than 30 milliseconds to approximately 90 milliseconds (Lisker & Abramson, 1964). On the other hand, those in monolingual Thai voiceless aspirated stops range from approximately 40 milliseconds to 120 milliseconds (Lisker & Abramson, 1964; Shimizu, 1996; Shimizu, 2011). In summary, despite the overlap, Thai voiceless aspirated stops have been found to have higher range as well as higher mean VOT values than English voiceless stops.

VOT in stops in English produced by Thai ESL speakers was examined by Shimizu (2011). In the study, Thai ESL participants were asked to produce stops in their native tongue, Thai, and their L2, English. As a result, English voiceless stops produced by the participants have VOT values close to Thai voiceless aspirated stops. This illustrates that Thai speakers use their native stop categories to acquire English ones.

Whether this L1 transfer will enhance with time is thus worth exploring for it could lead to better understandings of L1 transfer and bilingual phonetic realization.

Besides L1, plenty of studies (e.g., Lisker & Abramson, 1964; Yao, 2009; Smith et al., 2015) have illustrated that VOT is sensitive to many factors, such as speech rate, place of articulation, follow vowel height, phonetic context, utterance position, and speaker styles. One of the understudied factors that could affect VOT is speech time. The reason time is an interesting factor is because speech variability tends to intensify during a period of long speech, resulting in reduced accuracy of speech recognition (Frankish et al., 1992). Thus, a question arises if VOT duration values will increase or decrease as a form of this variability.

So far studies regarding the correlation of VOT and time have mostly dealt with code switching. Balukas and Koops (2015) focused on English VOT and Spanish VOT of Mexican bilinguals in spontaneous code-switching and found that English VOT values rose in the first few seconds after code-switching and then stabilised. However, the same effect was not found in Spanish stops which were the first language of the participants. A similar result was found in the study by Piccinini and Arvaniti (2015). As time progressed from a code-switch point, Spanish VOT values remained steady whereas English VOT values became higher. In such studies, the focus was on dyadic speech and time elapsed from code-switching apparently lasted up to 30 seconds. Though there are studies into VOT in longer speech without code-switching, they tend to focus on analysing mean VOT values (e.g., Hillman & Gilbert, 1977; Grosjean & Miller, 1994) and not its correlation with time.

When examining the correlation between VOT and time, it is worth including different speech styles. Previous studies have found that VOT values in isolated words tend to be higher than those in words read in sentences (e.g., Baran et al., 1977; Chodroff & Wilson, 2017). This is in line with the study by Nakamura et al. (2008), which found that, comparing with read speech, the spectral distribution was significantly reduced, and phonemes varied more in spontaneous speech. Considering that, a greater degree of variability in spontaneous speech could result in more substantial change than or even a different trend from that of read speech.

The present study has two aims. The first aim is to test if L1 transfer in terms of voice onset time varies with time in long monologue speech. L1 transfer here refers to VOT values in English voiceless stops produced by Thais, which are close to those of Thai voiceless aspirated stops. The second aim is to test if the presence or absence of such variation is the same in different running speech styles, that is, in spontaneous speech and in read speech. Thus, I have two hypotheses. First, VOT values in should significantly increase as speech time elapses because speech production should become more accented, i.e., more similar to Thai, resulting in higher VOT values. Second, VOT values in spontaneous speech should be lower and vary more greatly than those in read speech, resulting in a steeper slope.

2 Methods

2.1 Participants

Six university students, aging from 18 years old to 23 years old, participated in this study. All participants reported that they had Thai as their first language and English as their second language and used only Thai at home. All had studied in an international school or English programme during their primary education and none of them had lived outside Thailand for more than six consecutive months. To ensure the ability

to fulfil the tasks, all of the selected participants also had been trained to debate and were capable of making a 7-minute speech without interruption.

2.2 Tasks and Recording Procedures

For spontaneous speech, a debate in Asian parliamentary format was hosted on Mixidea.org, a website for online debating. Each team consisted of three speakers and each speaker was assigned to give a 7-minute speech. The motion was released 30 minutes before the debate, so each team had half an hour to prepare its case. This was to ensure that speech was spontaneous as the given preparation time would not suffice to write an entire script and would compel all speakers to improvise. The motion on the floor was 'This House would punish natural or legal persons who are accused of cultural appropriation'. The motion and the wordings were chosen with the aim to ensure a sufficient number and skewed distribution of the stops throughout each speaker's speech since the discussion must revolve around 'culture' and 'punish'. Interruption, point of information, and clapping were not permitted during speech so as to ensure continuity of long speech as well as to minimise noise. Recordings were conducted in two means. The first means was the recording function coming with mixidea.com. Each participant was also required to co-record using their own phone. The sound to be analysed was chosen based on minimum noise and minimum missing signals. Eventually, two files from mixidea.com and four files from participants' phones were used.

For read speech, each participant was asked to read the article 'Kept Women' by Marina Benjamin (2013). The article was excerpted and rearranged so as to evenly distribute the stops and make the speech last approximately seven minutes, which was the expected time of spontaneous speech. Practice before recording was allowed and participants were asked to finish the entire speech in one recording so as to ensure the continuity of speech. All read speech files were recorded using participants' phones.

2.3 Analysis

Spontaneous speech was manually transcribed. All speech was then auto-segmented with WebMAUS Basic service (Kisler et al., 2017) provided by Bavarian Archive for Speech Signals (BAS). VOT in initial voiceless stops was then manually segmented in Praat (Boersma & Weenink, 2017) and measured in milliseconds. Each VOT value was coded along with the time at which the stop was produced, its place of articulation, its following vowel duration, its following vowel height, the participant who produced the token, and the word containing the stop. The speech time here is defined as the time from the onset of the first word. Based on the transcription, approximately 1401 were expected, approximately 903 tokens from spontaneous speech and 498 tokens from read speech. Mean overall time is 7.21 minutes for spontaneous speech and 7.36 minutes for read speech.

2.4 Exclusion Criteria

The analysis excluded tokens which underwent problems. The first is stops without clear burst signals. This was most likely caused by technical problems, such as movement of participants' headphones/microphones or background noise in participants' settings. The second is stops which underwent deletion. As a consequence of these problems, identifying the point at which a release occurred became inaccurate, if not impossible, since no clear burst signal could be detected. The third and the most common is stops which

underwent affrication or frication. Participants, especially P1 and P3, tended to produce /t/ as fricative and /p/ as affricate. Apparently, this was due to regional dialects and/or free variants of the participants. Though VOT values could be measured, they were excluded since they were not from stops and could create noise for the analysis. The fourth is stops whose following vowels underwent devoicing, as it prevented accurate identification of the onset of vowels and resulted in drastically higher VOT values. And the last is stops which underwent voicing, as they would yield negative VOT and fall out of the scope of this study. After the step, the final number of tokens is 890 tokens, 523 tokens from spontaneous speech and 367 tokens from read speech.

3 Results

Firstly, let us summarize the distribution of raw VOT values. Figure 1 illustrates central tendencies and the variability of VOT values from both spontaneous and read speech. The symbol '+' signifies the mean value. The VOT means and standard deviations for both speech styles are given in Table 1.

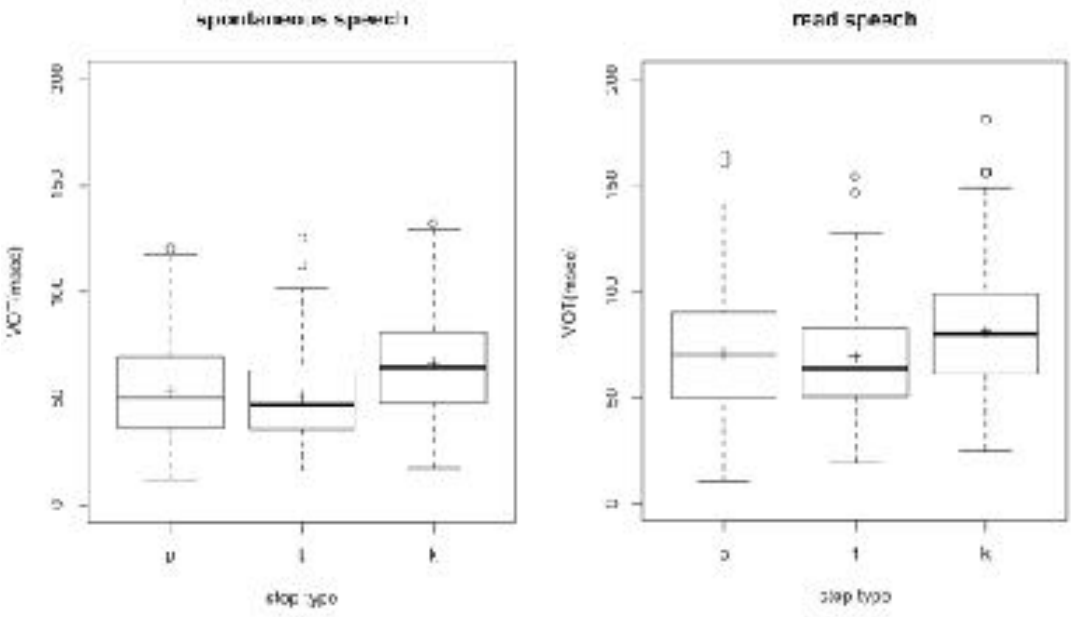

Figure 1: *Spontaneous and read speech distributions of raw VOT values.*

Table 1: *VOT means (ms), standard deviations (ms), amount of tokens.*

	Spontaneous speech			Read speech		
	mean	SD		mean	SD	
/p/	53.705	24.301	n = 116	70.601	31.952	n = 119
/t/	50.469	21.379	n = 113	68.888	26.590	n = 85
/k/	66.541	23.754	n = 294	81.080	28.243	n = 163
grand mean	56.905	23.144	n = 523	73.523	28.928	n = 367

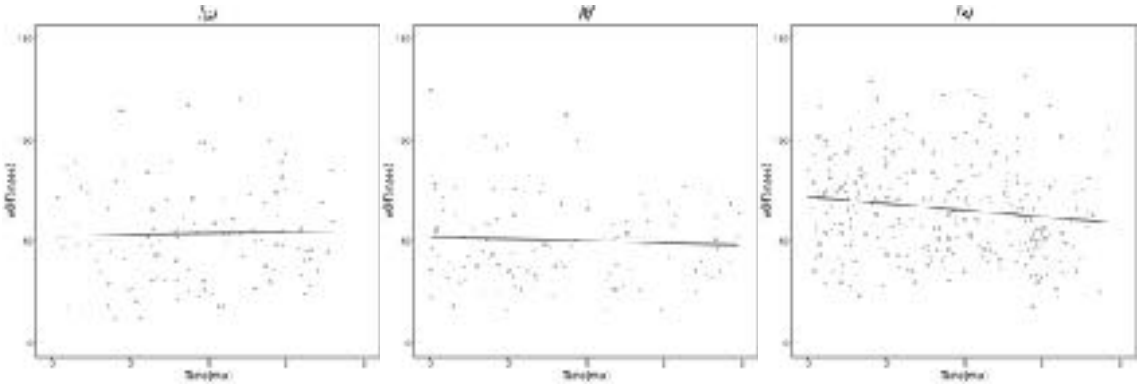

Figure 2: *Spontaneous speech VOT values by time.*

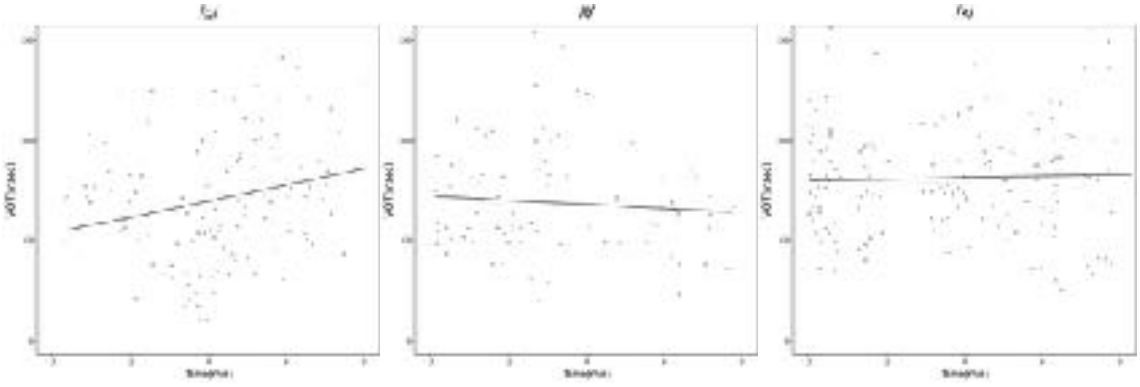

Figure 3: *Read speech VOT values by time.*

From Figure 1 and Table 1, in all stop types VOT values in read speech are on average 16.6 milliseconds higher than those in spontaneous speech. The standard deviations in read speech are also higher, indicating that the range of VOT values in the style is both higher and wider. In both styles, /p/ has slightly higher

VOT values than /t/, while /k/ has the highest VOT values. A major overlap between /p/ and /t/ could be seen in both speech styles. The mean VOT values and standard deviations in read speech are close to those of English voiceless stops produced by Thai speakers in Shimizu (2011).

Next, let us turn to the relationship between raw VOT and values speech time. As mentioned, if English voiceless stops became more similar to Thai voiceless aspirated stops, VOT values would be expected to increase. In the scatterplots in Figure 2 and Figure 3, linear regression lines are imposed to illustrate trends. In spontaneous speech, the /p/ line shows a slightly upward trend whereas the /t/ line shows a slightly downward trend. However, the slopes are not steep, indicating that neither /p/ nor /t/ undergoes any significant change. Only one stop type, /k/, shows a visible decrease in VOT values, signifying that the temporal gap between the burst of /k/ and the onset of voicing becomes shorter as time progresses in spontaneous speech.

The trends are distinct in read speech. While /k/ is the only stop that undergoes a clear change in spontaneous speech, it is the most stagnant stop in read speech. In contrast, /p/ here has an upward trend with a steeper slope than that in spontaneous speech, while /t/ still has the same slightly downward trend. It should be reasonable to conclude here that no systemic correlation between raw VOT values and speech time is found in either speech style.

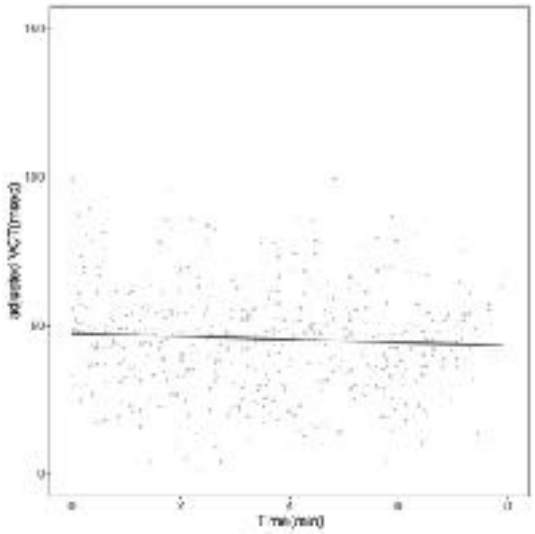

Figure 4: *Regression line showing slight decrease in spontaneous speech (slope=-0.52).*

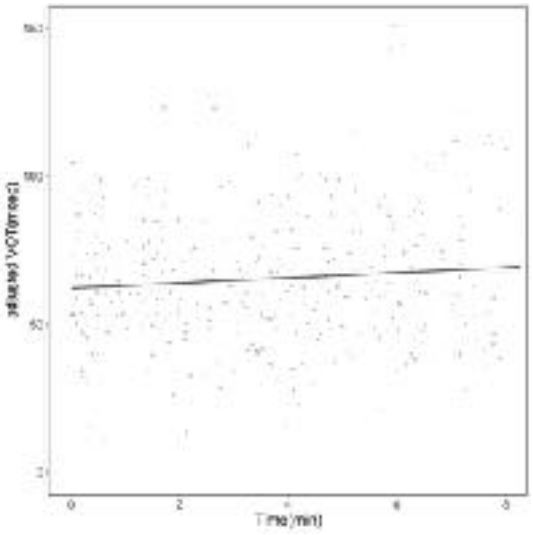

Figure 5: *Regression line showing slight increase in read speech (slope=0.85).*

As stops in running speech are greatly influenced by their environment (Yao, 2009; Smith et al., 2015), a linear mixed-effect model using random intercepts was constructed using lme4 package (Bates et al., 2015) in the statistical software R. This is to minimise and control variables which may affect VOT values. The model was adapted from Balukas and Koops (2015) by using the same coded fixed and random effects but without a logistic transformation. This is due to the fact that their study focused on code switching and the phonetic convergence was present only in the earlier part of speech, so the relation between VOT durations and the time from a code-switch point was non-linear. On the other hand, the present study aims to establish a linear relationship between VOT durations and the time from the onset of the first word since the effect should persist throughout speech. On that account, a logistic transformation is not included. Here, coded fixed effects are a) speech time and b) following vowel duration. The following vowel duration is used as an indirect measurement of speech rate as well as stress. Coded random effects are a) place of articulation, b) vowel height, c) word containing the stop, and d) participant who produced the token.

Figure 4 and Figure 5 show the trends of VOT values after mixed effects are calculated. All the grey dots in the scatterplots have been intercept-adjusted. In spontaneous speech, VOT values slightly drop as time elapses. In read speech, VOT increases slightly but with a greater slope than that from spontaneous speech. Here, VOT values in both speech styles do vary but very much slightly. Each minute that goes by, VOT durations become shorter by 0.52 milliseconds in spontaneous speech and longer by 0.85 milliseconds in read speech. So, the change in neither speech style tends to be greater than seven milliseconds throughout seven minutes of speech time. Recognising that there is much room for VOT durations to extend up to 120 milliseconds, which is the upper bound of Thai voiceless aspirated stops, the resulted change is barely significant. An F-test via Kenward-Roger approximation also affirms that elapsed time is not a statistically significant predictor of VOT in either speech style. (p-value = 0.24 in read speech, 0.20 in spontaneous speech). In conclusion, there is no clear relationship between speech time and adjusted VOT values.

4 Discussion

Both hypotheses fail. VOT values do not significantly increase as speech time elapses. Though lower, VOT values in spontaneous speech do not vary more significantly that those in read speech. Thus, with the present study, it should be reasonable to conclude that Thai ESL speakers tend to produce English initial voiceless stops without the VOT duration values becoming closer to those of Thai voiceless aspirated stops. In other words, they tend to adhere to a narrow range of VOT values throughout long speech in both spontaneous and read speech. Thus, L1 transfer does not vary with time.

Let us begin with discussing L1 transfer. English voiceless stops in read speech in the present study have the mean VOT values and standard deviations rather close to English voiceless stops produced by Thai ESL speakers in Shimizu's (2011) study. The same could not be said for spontaneous speech for Shimizu's study analysed VOT in read isolated words. Though more information is needed for comparison, it would not be inconsistent with the previous study to conclude that Thai ESL speakers produce English voiceless stops in read speech, whether in citation forms or running speech, with VOT values close to those in Thai.

The marked contrast of the raw VOT values between the two styles is in line with previous studies (Baran et al., 1977; Chodroff & Wilson, 2017), which found that VOT in read speech was longer than that in spontaneous speech. It also strongly supports stylistic variation, that is, attention and awareness affect stop articulation. When participants read, they tend to be more aware of their speech, resulting in more articulation rate and thus higher VOT values. In contrast, in spontaneous speech like a debate speech, they tend to be less aware of their production since they have to constantly engage themselves with the content at hand, resulting in less articulation rate and thus lower VOT values. Also, long spontaneous monologue usually contains many linguistic/pragmatic constraints and repetition. Many times, words either are repeated as a means to buy time to think mid-speech or could be predicted based on the syntactic structure and the context. This greater degree of redundancy and less amount of information permit VOT reduction while still maintaining intelligibility of speech (Coker & Umeda, 1975; Baran et al., 1977).

A possible explanation for the stability of VOT values is that even if L1 transfer is present, ESL learners still try to maintain VOT values of English voiceless stops in a particular range so that they will not overlap with those of Thai stops. Basically, it could be considered a way in which a bilingual attempt to keep a set of phonetic properties of stops separated and exclusive to each language. A study into Thai stops in long speech produced by ESL Thai learners is needed to compare and prove the hypothesis.

Though minimal, the difference between the resulting trends, i.e., that VOT values in spontaneous speech tend to decrease while those in read speech tend to increase, is worth discussing. The downward trend in spontaneous speech could stem from vocal fatigue after long continuous speech production. Since the effect of vocal fatigue on stops was minimal (Caraty & Montacié, 2010), the downward trend turned out to be only slight. Regarding the upward trend in read speech, I offer two explanations. First, the upward trend itself could be the result of L1 transfer which was minimised, if not neutralised, by the attempt to separate stop categories for each language. It is apparent in only read speech since the effect was overshadowed by neither cognitive load nor redundancy. Second, the increase could also be a form of compensation. Since participants tended to be more aware of their speech when reading, it could follow that they were aware of fatigue, which should result in less articulation or greater imprecision. Consequently, they tried to compensate it by carefully articulating, thus resulting in higher VOT values than the earlier part of reading in which they did not try to compensate fatigue.

5 Conclusion

The present study examines VOT in long spontaneous speech and long read speech elicited from Thai speakers with English as a second language. The results show that VOT values in spontaneous speech are lower than those in read speech. The marked contrast between VOT values in the two speech styles support stylistic variation and constraints in spontaneous speech. Also, no correlation between VOT values and speech time is found in either speech style. One possible explanation is that a bilingual tries to maintain the phonetic exclusivity for each language. Though insignificant, the downward trend in spontaneous speech could stem from vocal fatigue while the upward trend in read speech could be due to minimised L1 transfer or compensation.

Further studies may use tasks without a topic to obtain spontaneous speech in order to reaffirm the result and may also group participants according to levels of English fluency to test if the result would differ among people from different levels. Since this study includes only English stops, future studies should also examine Thai stops in long speech produced by native Thai speakers to investigate whether the trends would be similar when there is no L1 transfer effect.

6 References

Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, 19(4), 423–443.

Baran, J. A., Laufer, M. Z., & Daniloff, R. (1977). Phonological contrastivity in conversation: A comparative study of voice onset time. *Journal of Phonetics*, 5(4), 339–350.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Benjamin, M. (2013). Kept Women. [Online] Retrieved from *Aeon*: <https://aeon.co/essays/why-do young-rural-women-in-china-become-mistresses>.

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program, version 6.0.33]. Retrieved 13/01/2020, from *Praat*: <http://www.praat.org/>.

Caraty, M. J., & Montacié, C. (2010). Multivariate analysis of vocal fatigue in continuous reading. INTERSPEECH, 2010.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61, 30-47.

Coker, C. H., & Umeda, N. (1975). The importance of spectral detail in initial-final contrasts of voiced stops. *Journal of Phonetics*, 3, 63–8.

Frankish, C., Jones, D., & Hapeshi, K. (1992). Decline in accuracy of automatic speech recognition as a function of time on task: fatigue or voice drift?. *International Journal of Man-Machine Studies*, 36(6), 797–816.

Grosjean, F. (1980). Temporal variables within and between languages. In H. W. Dechert, & M. Raupach (Eds.), *Towards a cross linguistic assessment of speech production* (pp. 39–53). Frankfurt: Lang.

Grosjean, F., & Miller, J. L. (1994). Going in and out of languages: An example of bilingual flexibility. *Psychological Science*, 5(4), 201–206.

Hillman, R. E., & Gilbert, H. R. (1977). Voice onset time for voiceless stop consonants in the fluent reading of stutterers and nonstutterers. *The Journal of the Acoustical Society of America*, 61(2), 610-611.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, 25(2), 143–168.

Kisler, T., Reichel U. D., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45, 326–347.

Lisker, L., & Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *WORD*, 20(3), 384–422.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, 10, 1–28.

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2), 171–184.

Piccinini, P., & Arvaniti, A. (2015). Voice onset time in Spanish–English spontaneous code switching. *Journal of Phonetics*, 52, 121–137.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. *International Congress of Phonetic Sciences*, 607–610.

Shimizu, K. (1996). *A Cross-Language Study of Voicing Contrasts of Stop Consonants in Asian Languages*. Seibido Publishing Co.

Shimizu, K. (2011). A Study on VOT of Initial Stops in English Produced by Korean, Thai and Chinese Speakers as L2 Learners. *International Congress of Phonetic Sciences*, 1818–1821.

Stuart-Smith, J., Sonderegger, M., Rathcke, T. & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, 6(3-4), 505–549.

Yao, Y. (2009). Understanding VOT Variation in Spontaneous Speech. *UC Berkeley Phonology Lab Annual Report,* 5(5), 29–43.

7 Appendices

7.1 Appendix One

Table A1: *Speech time.*

	Spontaneous speech	Read speech
P1	6.57	7.50
P2	7.24	8.22
P3	6.34	6.44
P4	7.24	7.41
P5	7.56	6.23
P6	7.01	6.50
Mean	7.21	7.36

Table A2: *Fixed effects summary.*

	Spontaneous speech			Read speech		
	estimate	Std. Error	t-value	estimate	Std. Error	t-value
intercept	47.259	7.308	6.467	62.194	7.5672	8.219
time	-0.523	0.403	-1.296	0.8516	0.7147	1.192
vowel duration	179.896	27.804	6.47	111.2686	33.5296	3.319

Table A3: *Random effects summary.*

	Spontaneous speech		Read speech	
Group	Variance	SD	Variance	SD
stop type	36.03	6.003	16.22	4.028
word	78.91	8.883	141.37	11.89
vowel height	22.31	4.723	40.63	6.374
participant	144.03	12.001	109.67	10.473
residual	335.77	18.324	564.31	23.755

Table A4: *Random intercept values for the variable 'stop type'.*

	Spontaneous speech	Read speech
/p/	-2.892	-3.193
/t/	-3.499	0.752
/k/	6.391	2.441

Table A5: *Random intercept values for the variable 'following vowel height'.*

	Spontaneous speech	Read speech
high	-4.700	-4.812
low	2.105	6.083
mid	2.595	-1.271

Table A6: *Random intercept values for the variable 'participant'.*

	Spontaneous speech	Read speech
P1	-2.725	-3.347
P2	5.008	14.368
P3	15.852	2.181
P4	6.489	6.210
P5	-6.531	-4.502
P6	-18.092	-14.910

Acknowledgements

First of all, I am extremely grateful to Assistant Professor Sujinat Jitwiriyanont for inspiring me to pursue phonetics, encouraging me to send my paper to conferences, and advising this paper from a term project of the Acoustic Phonetics subject to even after the course ended. I'm deeply indebted to all my dearest linguistics seniors who have helped me understand linguistics, conferences, and researching inexpressibly deeper: P' Ponds Vatcharit, P'Tuner Pongbodin, and P'Tongla Ponrawee. This paper would not have been possible without P'Beam Asawa, my debate senior who helped me survive R and obtain the very result. Thanks also to P'Sun Suntisuk, P'Pao, P'Est, Book, U-Center friends, and my family for accompanying me

throughout this project, which unfortunately coincided with the pandemic, and making the progress a lot less hard for me.