Long Distance Recordingship: Assessing the Use of Remote Recordings in Acoustic Research on Serbian EFL Students' Acquisition of English VOT

Nina Đukić, University of Belgrade, Serbia

The Paris Conference on Education 2023
Official Conference Proceedings

Abstract
Though research in acoustic phonetics entails laboratory conditions, the rapid technological development accelerated by the COVID-19 pandemic bids the question of remote recording, the success of which could provide phoneticians with more research opportunities. This paper explores the feasibility of remote sample collection in the context of examining the degree of aspiration in initial voiceless stops in Serbian and English with Serbian EFL students. Since the role of positive Voice Onset Time (VOT) in English and Serbian differs significantly, a clear contrast between English long-lag and Serbian short-lag stops might prove challenging for Serbian EFL students. To examine the degree of VOT acquisition, audio recordings are made for 5 advanced and 5 proficient speakers. The participants are firstly recorded in laboratory conditions. Next, the participants are asked to read the same material in a quiet space within their homes, record their speech using mobile phones, and deliver the recordings via email. For each token, both laboratory and remote recordings are examined using the speech analysis software *Praat* (Boersma & Weeninik, 2023). Results indicate that remote recording via smart phones yielded reliable samples with measurable VOT for the voiceless plosives /p t k/ in both English and Serbian. Proficient speakers displayed VOT values that approached native-like patterns. In contrast, the less proficient group exhibited shorter VOT durations, but a significantly clearer distinction between the use of aspiration in Serbian and English. Additionally, a surprising finding shows that proficient speakers assign longer VOT to Serbian plosives too, likely owing to L2 transfer.

Keywords: Remote Recording, Acoustic Phonetics, EFL Acquisition, Serbian Speakers of English, *Praat*

Introduction

The COVID-19 pandemic has given rise to alternative approaches in experimental research, including the field of acoustic phonetics. With a significantly increased risk of contagion, conducting acoustic experiments was rendered unfeasible. Consequently, researchers have turned to alternative means of sample collection, such as remote recording. The ideal conditions for recording have typically included working in a phonetic laboratory – a controlled environment with ample opportunity for monitoring participants, providing additional instruction or simply rerecording. The advancement of technology and participants' general technological proficiency has allowed researchers to test the method of remote recording. This approach generally entails participants using various devices (laptop, mobile) and necessary software to record their own speech in an acoustically appropriate setting and deliver the samples to the researcher thereafter.

Previous research has approached this topic from various standpoints. They range from assessing and comparing the reliability of different devices to comparing how efficiently different types of phonetic data can be recorded and analyzed. Ge et al. compare common acoustic measurements using recordings made by seven different devices in a sound-attenuated lab and a quiet conference room (2021, p. 3985). The focus of their experiment is on the devices' ability to capture linguistically meaningful information in the acoustic signal, as it includes a more comprehensive range of phonetic phenomena. The study reveals varied results for measurements such as fundamental frequency, fricatives, vowels etc. Their results show that F0 (fundamental frequency) in vowels is relatively reliable among devices. (Ge et al., 2021, p. 3987). Vowel formant values, especially for F2, vary significantly among devices, and /i/ is more influenced than /u/. Their findings imply that the choice of recording device is critical in the study of vowels, as some smartphones and ZOOM recordings on laptops can lead to misleading results. Additionally, sampling frequency plays a crucial role in spectral moments, with devices below 40,000 Hz failing to capture energy concentration at high frequencies, important for some fricatives. Recordings made on the cloud using ZOOM have notable duration differences but perform inconsistently in other acoustic measurements. The study suggests that recording locally using the speaker's computer is a better solution. The authors note that various other methods of remote speech data collection are possible, but that researchers should exercise caution and document the steps clearly (Ge et al., 2021, p. 3987).

This paper will focus only on the use of mobile devices in recording VOT (Voice Onset Time) in English and Serbian voiceless plosives /p t k/, produced by Serbian speakers of English. Thus, both the aim and analysis of this paper are twofold: the assessment of remote recordings as opposed to laboratory samples which will be used as reference points and, secondly, close examination of VOT acquisition in two groups of Serbian speakers of English who differ in language proficiency.

The connection between aspiration and English stop sounds lies in their manner of articulation. English stops, namely /b d g p t k/, are produced through complex movements within the vocal tract, with a key focus on a high degree of closure or stricture where the articulators come into firm contact (Čubrović, 2009, p. 35). This closure can occur at different places in the vocal tract, leading to the categorization of stops based on their place of articulation. Regardless of their position, all stops go through three stages of articulation: approach, hold or compression, and release (Čubrović, 2009, p. 36). The release stage involves a sudden release of air with an explosive sound, and the energy levels during this release differ based on their voicing.

In general, voiced stops /b d g/ are produced with vibrations of the vocal folds, while voiceless stops /p t k/ are not. Halle et al. also introduce an alternative classification using the terms "tense" and "lax" stops (1957, p. 107), which are also known as "fortis" and "lenis" stops (Čubrović, 2009, p. 37). These labels are considered more precise because they reflect the distinct articulation of these two groups. Tense stops like /p t k/ involve higher pressure build-up against the closure, resulting in a more forceful release (Halle et al., 1957, p. 107). On the other hand, lax stops like /b d g/ also have bursts during the release, but their key difference lies in the presence of more pronounced /h/-like sound, also known as aspiration, especially with tense stops.

Aspiration can be heard in speech, but accurately recording and perceiving this feature on a waveform or audio recording remains a challenge. The distinction between aspirated and unaspirated consonants, however, can be observed through VOT or voice onset time, which refers to the time between the release of the consonant and the onset of voicing for the following sound (Zsiga, 2020, p. 131). By measuring VOT on waveforms, linguists have explored the significance of aspiration across languages and speakers (Halle et al., 1957; Cho & Ladefoged 1999; Kim, 2011; Shimizu, 2011).

Cho and Ladefoged conducted a study on 18 endangered languages to identify VOT universals across different languages. These universals could predict VOT values based on the place of articulation of the consonant. They discovered various patterns in languages from separate language families. One significant universal is that velar stops always have a longer VOT, but this doesn't apply to languages with uvular stops. Additionally, VOT is shortest before bilabial stops and intermediate before alveolar stops for both aspirated and unaspirated stops, with exceptions in Tamil, Cantonese, and Eastern Armenian (Cho & Ladefoged, 1999, p. 208). The authors created four classes of stops based on VOT duration: unaspirated (30ms), slightly aspirated (50ms), aspirated (90ms), and highly aspirated stops (over 90ms) (Cho & Ladefoged, 1999, p. 223). While there are exceptions, this generalization aids in predicting VOT values, speech cues for perception, and the success rate of producing aspiration when learning a foreign language.

The research by Halle et al. focuses on acoustic analysis of plosives and describes the phonetic features of English stops, including their spectral features with sonograms. The study involves both articulation and perception experiments of plosives, where participants identify English plosives in isolation and in syllable form with alternating positions (initial vs. final). Here, aspiration proves to be a significant factor for perception among native speakers (Hale et al., 1957, p. 108).

For non-native speakers, the importance of aspiration depends on their native language. Kim (2011) conducted a contrastive study comparing English and Korean, where Korean speakers of English were examined to determine if accurate use of aspiration is related to language proficiency. The results indicated that more proficient speakers shortened their VOT in English, while less proficient speakers prolonged it. Interestingly, there was also evidence of cross-language phonetic influence, as Korean speakers shortened the VOT in their native language as well.

Shimizu, on the other hand, conducted a comprehensive analysis of Korean, Thai, and Mandarin Chinese in comparison to English with regard to aspiration. The study included minimal pairs and triplets in each language presented within carrier sentences. The acoustic analysis revealed significant L1 transfer in the production of English plosives by non-native speakers, with a considerable delay of voicing and strong aspiration. Additionally, velar stops across all four languages displayed more prominent aspiration compared to bilabial and alveolar stops.

Based on existing research on remote recording, as well as VOT acquisition in English as a foreign language, it can be expected that remote recording using smart phones can provide reliable samples so long as participants receive clear and detailed instructions beforehand. Additionally, the VOT results are expected to demonstrate a direct correlation between the speakers' level of English and their use of aspiration in speech: the higher the level, the more native-like the VOT values will be.

Methodology

A total of ten participants took part in this experiment. Five participants are in their first year of bachelor studies, studying for a degree that is not related to English (medicine and electrical engineering). Nonetheless, all five participants studied English during their primary and secondary education. The remaining five are all first-year students studying languages, including English as a foreign language. All participants are female and their age ranges between 20 and 26. For the purpose of distinguishing between these two groups of participants, the first five participants who do not study English at university level will hereinafter be referred to as *advanced speakers* whereas the participants who study English at university level will be termed *proficient speakers*.

The recording materials consist of target words with word-initial plosives in both English and Serbian, all of which were embedded into carrier sentences. For English, the words used were *pick, tick,* and *kit*, with the Serbian counterparts *pik*, *tik*, and *kit*, which include the same distribution of sounds.

Recordings were made in two different locations for the purpose of quality comparison. The first round of recording took place in the Belgrade Phonetics Lab of the Faculty of Philology, University of Belgrade. All recordings were made directly in the latest version of the *Praat* software (Boersma & Weenink, p. 2022), at the input frequency of 44100 Hz. The participants read sentences shown on PowerPoint slides. Each participant read the sentences three times, which amounted to a total of 180 lab-made tokens for both languages, i.e. 90 in Serbian and 90 in English. Next, participants were instructed to find a quiet space in their homes and read the same slides, while recording their voice using voice-recording applications on their phones – all participants possessed new generation smart phones with the necessary application. The home-made voice recordings were then delivered for analysis via email, with the exact same number of tokens – 180. Together with lab-made tokens, the total number of tokens analyzed for this research was 360.

After the recording was completed, each target word was cut from the original recording and used as a token for a twofold analysis. Firstly, each individual home-made token was compared to its lab counterpart to determine whether the sample was clear enough to analyze the VOT values. The criterion for marking a remote sample as *acceptable* were clearly visible borders between plosive production stages, which would consequently allow the measurement of VOT. Only the values of acceptable remote samples were then entered into a table in Excel, which was used to calculate mean values and draw conclusions.

Discussion: Remote Recording

To assess the reliability of the remote recordings, the VOT values obtained from the remote samples were compared with those recorded in a controlled laboratory environment. The laboratory samples, which served as reference points, were collected while working in a traditional phonetic laboratory setting. What follows is an example of side-by-side comparison between two pairs of recordings by an advanced (Figure 1.1. and 1.2.) and proficient speaker (Figures 2.1. and 2.2.).

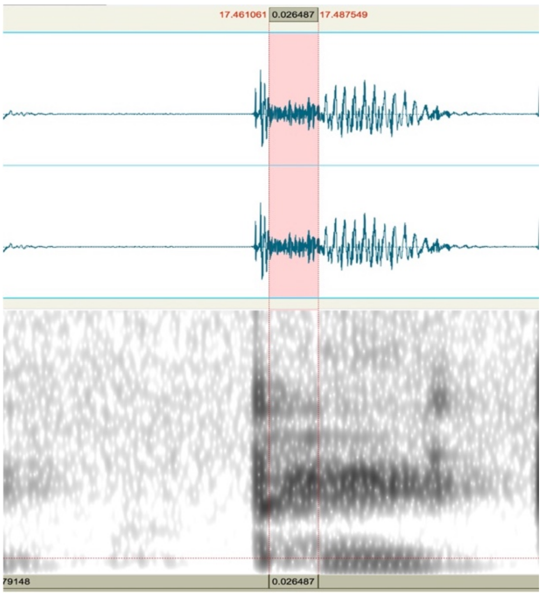


Figure 1.1. /pick/ - Proficient Speaker, LAB

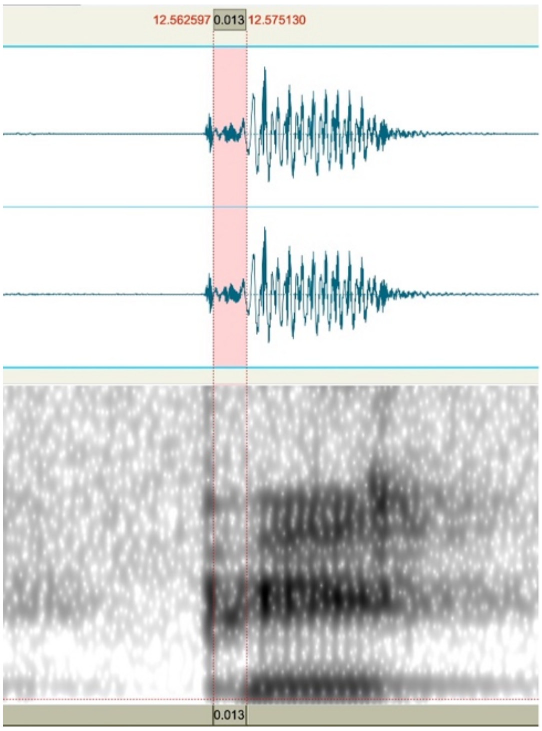


Figure 1.2. /pick/ - Proficient Speaker, HOME

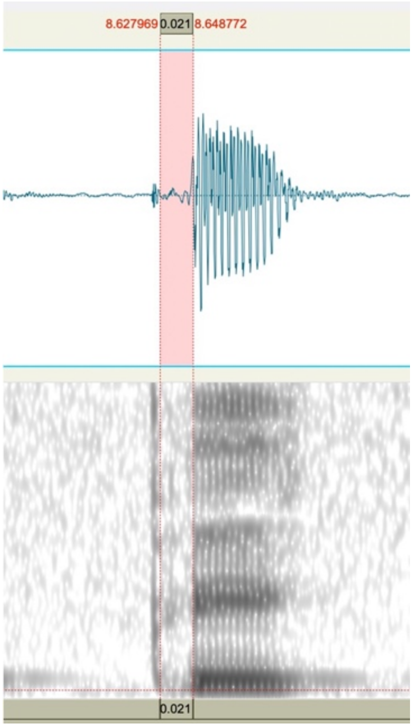


Figure 2.1. /pɪk/ (Eng) - Advanced Speaker, LAB

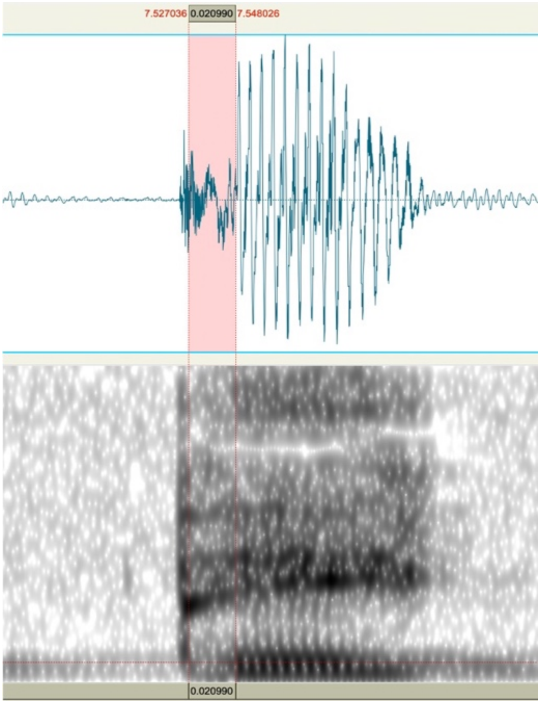


Figure 2.2. /pɪk/ (Eng) - Advanced Speaker, HOME

The analysis of the VOT values from both remote and laboratory recordings showed a high degree of consistency. Apart from a slight increase of background noise in the remote recordings, which can be observed by a darker shade of gray in the spectrograms' background, there were no statistically significant differences between the mean VOT values obtained from the two recording methods for each voiceless plosive in both English and Serbian. This finding suggests that remote recording is a viable alternative to traditional laboratory recordings for collecting acoustic data related to VOT.

Furthermore, the remote recordings exhibited clear patterns of aspiration in the VOT values for the voiceless plosives. The visibility of aspiration in the release stage of the voiceless stops was clear in both the remote and laboratory recordings, proving the successful capture of this phonetic feature through remote means, which is in line with the initial hypothesis of this paper.

Discussion: VOT Acquisition

Before presenting the collected values from this research, it is essential to provide reference values for English stops produced by native English speakers. These reference values, extracted from Kim (2011), will serve as reference points for a more reliable assessment of the study's results.

NATIVE SPEAKERS	p	t	k
AVG	58	70	80

Table 1: Native speakers' mean VOT values for English (Kim 2011: 4)

Table 1 confirms the universals regarding VOT duration, with velar stops exhibiting the longest offset interval, followed by intermediate alveolar VOT, and bilabial /p/ having the shortest VOT (Cho & Ladefoged, 1999, p. 208). What follows are the mean values for Advanced (Table 2) and Proficient (Table 3) Serbian speakers of English respectively.

ADVANCED L2 SPEAKERS	SRB			ENG		
	p	t	k	p	t	k
AVG	12.79	15.53	41.2	56.31	45.82	64.22

Table 2: Advanced speakers' mean VOT values for Serbian & English

In our experiment, VOT measurements were taken for the interval between the release of the plosive and the onset of voicing of the following sound, expressed in milliseconds (ms). Each target word had three repetitions per speaker, and the tables above present the average values for /p t k/ for both groups of speakers.

Comparing the values with the reference (Table 1), it becomes apparent that although advanced speakers demonstrate a more pronounced lag in word-initial stops, their VOT still differs significantly from that of native speakers. As expected, velar stops have the most prominent VOT in both languages, in accordance with the presented mean values (Cho & Ladefoged, 1999). Additionally, the mean VOT values for Serbian align with the initial hypothesis, showing less prominence and significantly shorter VOT durations compared to English.

PROFICIENT L2 SPEAKERS	SRB			ENG		
	p	**t**	**k**	**p**	**t**	**k**
AVG	15,88	19,335	56,5425	62,56	68,53	84,3525

Table 3: Proficient speaker's (mean) VOT values for Serbian & English

The values for proficient speakers are much closer, if not fully aligned, with the VOT values of native speakers. However, there are a few surprising observations: the average VOT value for the velar stop /k/ is the most prominent, slightly longer than that of native speakers. On the other hand, the VOT values for the Serbian /k/ are significantly higher than those of advanced speakers. Spectrograms illustrating the unusually longer VOT in English and Serbian can be seen in Spectrogram 1 and Spectrogram 2, respectively (see Appendix).

Thus, it can be assumed that higher proficiency and increased exposure to a language with a long-lag aspiration feature may affect the non-native speaker's VOT production in their mother tongue. These findings are in line with Kim's results (2011) and demonstrate the "bi-directional influence" (Grosjean, 1989) of languages. In other words, significant exposure to a foreign language (L2) is likely to influence the learner's production of their native language (L1) as well.

Regardless of the reference values, it is evident that proficient speakers have acquired the aspiration feature to a greater extent and have no difficulties in its application during pronunciation. Conversely, while advanced speakers demonstrate a slightly less prominent use of aspiration, they exhibit the ability to make a greater distinction between VOT values in Serbian and English.

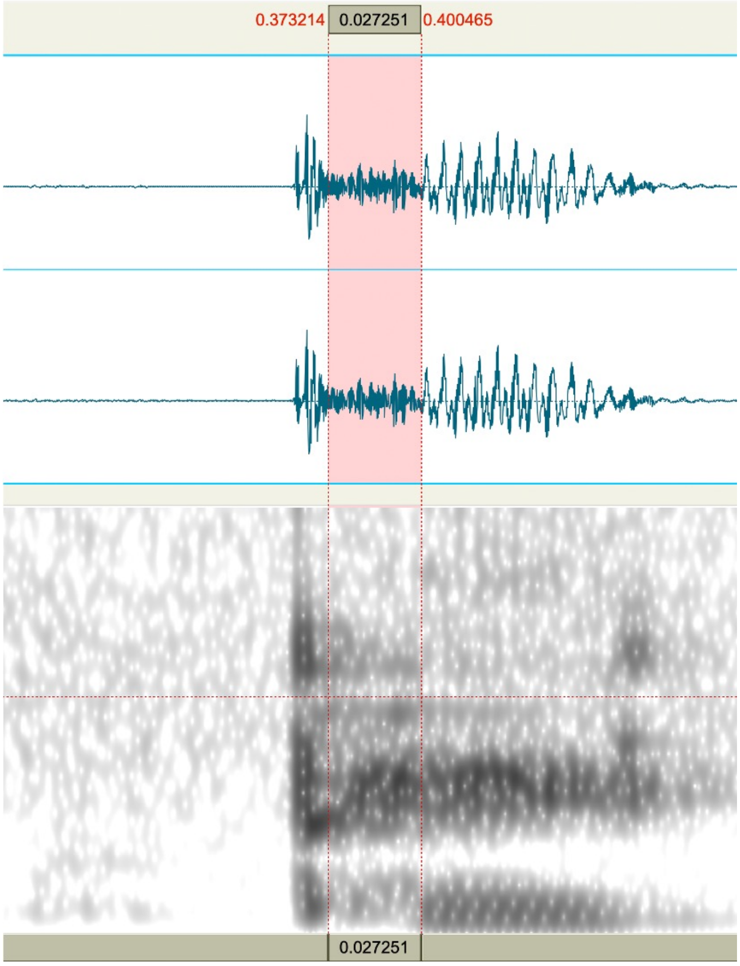
Conclusions

The results of the study indicate that remote recording using smart phones yielded reliable samples with measurable Voice Onset Time (VOT) values for the voiceless plosives /p t k/ in both English and Serbian. Participants in the study were able to follow clear and detailed instructions on how to record their speech, ensuring the successful acquisition of usable data. However, this conclusion is limited only to the use of mobile phones and their reliability in the context of VOT analysis. A larger sample with more variables (devices or phonological phenomena) would have to be included in the future to draw further conclusions on the matter. On the other hand, the data from remote recordings also allowed for the comparison of VOT acquisition between two groups of Serbian speakers of English who differed in their language proficiency. The more proficient group of speakers displayed VOT values that approached native-like patterns. In contrast, the less proficient group exhibited shorter VOT durations, but a significantly clearer distinction between the use of aspiration in Serbian and English. Based on the side-by-side comparison in Tables 2 and 3, it has been concluded that proficient speakers are more inclined to assign longer VOT intervals, but this tendency extends to Serbian plosives too, most likely owing to L2 transfer. It is important to note that this research is limited to voiceless plosives in combination with a single vowel. In order to fully explore the role of VOT in Serbian and English, as well as their correlation, a more comprehensive sample is needed in the future, which would assess VOT in more varied contexts.

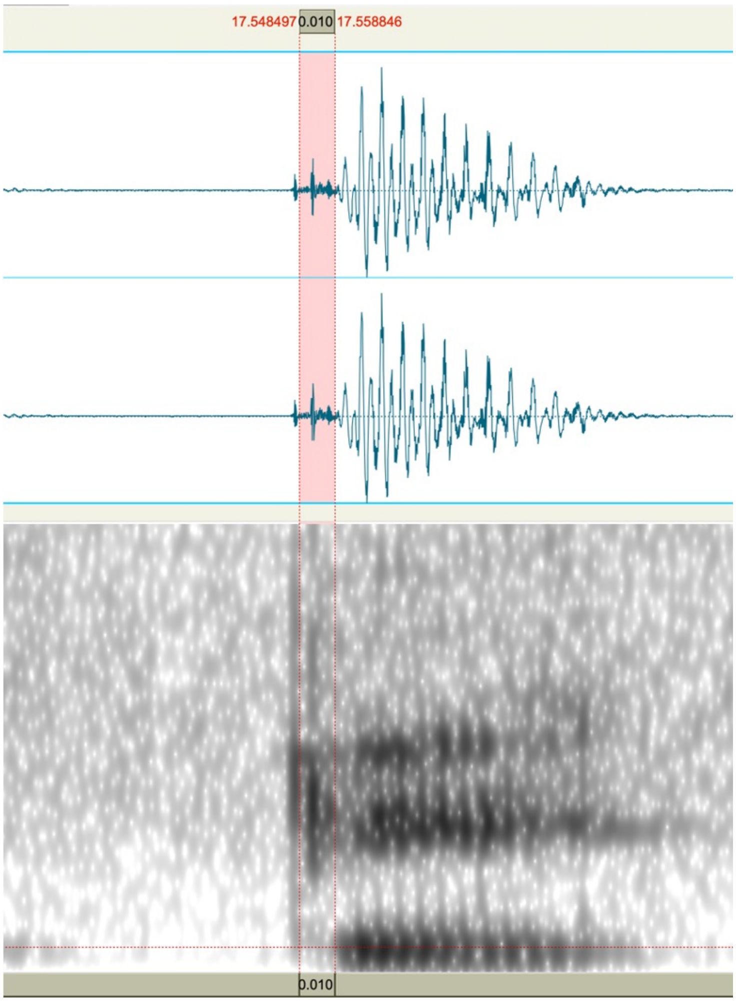
Appendix

The appendix contains 2 spectral representations of recorded speech by speaker P4, who belongs to the proficient group of non-native speakers of English in this study.

Spectrogram 1
Word /pɪk/ (English) by participant P4



Spectrogram 2
Word /pɪk/ (Serbian) by participant P4



References

Boersma, P., Weenink D. (2022). *Praat: doing phonetics by computer* [Computer program]. Version 6.2.20. http://www.praat.org/

Cho, T., Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics, 27*(2), 207–229.

Čubrović, B. (2009). *Profiling English Phonetics*. Belgrade: Philologia.

Flege, J., Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal Of Phonetics, 15*(1), 67–83.

Flege, J. E. (1992). Speech Learning in a Second Language. In C.A. Ferguson, et al. (Eds.), *Phonological Development: Models, Research, Implications* (pp. 565–604). Timonium, MD: York Press.

Flege, J. E. et al. (1995). Effects of age of second-language learning on the production of English consonants, *Speech Communication, 16*, 1–26.

Ge, C., Xiong, Y., & Mok, P. (2021). How Reliable Are Phonetic Data Collected Remotely? Comparison of Recording Devices and Environments on Acoustic Measurements. In *Interspeech* (pp. 3984–3988).

Grosjean, F. (1989). "Neurolinguists, beware! The bilingual is not two monolinguals in one person", *Brain and Language, 36*, 3–15.

Halle, M. et al. (1957). Acoustic properties of stop consonants. *Journal of the Acoustical Society of America, 29*, 107–116.

Kim, M-R. (2011). The relationship between cross language phonetic influences and L2 proficiency in terms of VOT. *Phonetics and Speech Sciences, 3*(3), 3–10.

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

Zsiga, E. (2020). *The Sounds of Language*: *An Introduction to Phonetics and Phonology*. Wiley-Blackwell: New Jersey.

Contact email: djukic.nina96@gmail.com