PHONETIC CHARACTERISTICS OF DEVOICED VOWELS IN UYGHUR

Michael Fiddler

University of California, Santa Barbara
mfiddler@ucsb.edu

ABSTRACT

This paper examines the phonetic characteristics of devoiced vowels in Uyghur. Audio recordings were collected of six female speakers reading 43 words and 15 sentences, which had been selected to explore a variety of characteristics of Uyghur vowel devoicing. Evidence is presented confirming the reliability of three conditioning factors—high vowels, adjacent voiceless consonants, and unstressed syllables. Analysis of a subset of words with the target vowel /i/ shows that the devoiced vowels are realized entirely as sibilant frication when adjacent to sibilant consonants /ʃ/ and /t͡ʃ/ (and thus indistinguishable from the consonant or "deleted") but distinctly present as a segment of aspiration or light frication when between plosives. Duration of devoiced /i/ is found to be shorter than that of voiced /i/. Finally, articulatory considerations relevant to vowel devoicing are discussed.

Keywords: vowel devoicing; duration; Uyghur; Turkic

1. INTRODUCTION

Vowel devoicing in Uyghur is conditioned by vowel height, surrounding consonants, and stress. Hahn's summary is the most thorough and accurate description to date: "High vowels are devoiced between two voiceless consonants (including a glottal stop), in words with more than one syllable usually only in unstressed position" ([4]: 45).

To my knowledge, only one empirical study on Uyghur vowel devoicing has been conducted to date. Tursun & Hemdulla [16] addressed the acoustic properties of devoiced high vowels in Uyghur as compared to regular voiced high vowels. Their results indicated that voiceless vowels are shorter in duration, higher in formant frequency, and lower in intensity than their voiced counterparts. Unfortunately, a number of methodological issues make their findings hard to generalize. Their method of determining voicing was not stated clearly; no mention was made of stress, which would affect interpretation of the acoustic results; no statistical tests were conducted; and the total number of tokens for each cell was not reported. Vowel devoicing thus remains a phonetic topic of interest in Uyghur.

Additionally, vowel devoicing can be realized phonetically in a number of ways. Tursun & Hemdulla suggest that Uyghur voiceless vowels take on some characteristics of voiceless fricatives, but they measured formants as if the vowel quality were still intact [16]. Hahn describes true voiceless vowels, syllabic fricatives, and combinations of both [4], but without acoustic evidence. Authors writing on vowel devoicing in other languages have described full deletion of vowels leading to consonant clusters (see [9, 11, 12]). This paper aims to clarify the precise phonetic nature of the voiceless vowels in Uyghur.

2. METHODS

Data for the present study came from a larger set of recordings designed to explore vowel devoicing, in which six female speakers read 43 one- or two-word tokens and 15 sentences (6 speakers x 5 repetitions = 30 tokens per item). Two minutes of naturalistic data from a TV interview were included as a supplement to the experimental data. The word and sentence tokens were randomized and counterbalanced, with each iteration of the list beginning and ending with a standard decoy word or sentence that was not included in the study. The six participants were native Uyghur speakers who grew up in Xinjiang and were first-year university students at the time of the study. All of them also speak Mandarin fluently and English to varying levels.

The data were segmented in Praat and acoustic measurements were extracted using scripts. Duration was measured from oral release of the previous consonant to the onset of the following consonant, or to point where F2 dispersed for word-final vowels. Praat's 'Fraction of Locally Unvoiced Frames' (under Voice Report) was used as a quantitative measure of devoicing in order to see if the devoicing were categorical or gradient. For this measurement, a separate tier was added in the TextGrid in which aspiration following voiceless stops was segmented separately from the vowels. In this way clearly voiced vowels following aspirated consonants would still measure as fully voiced.

The data was coded for the three conditioning factors—vowel height, consonant environment, and stress. Stress is typically word-final in Uyghur, but there is often variation in speaker production and perception [18]. For this study, word-final stress was expected but confirmed post hoc by comparing the duration of the word-final vowels to the other vowels in the word, as duration is the only reliable acoustic cue for lexical stress [18].

For the present study, a small subset of five disyllabic words containing an unstressed /i/ between voiceless consonants as the target vowel was selected for analysis—*piˈkir* 'idea,' *ʔikˈki* 'two,' *kiˈʃi* 'people,' *t͡ʃiˈqim* 'payment,' and *t͡ʃiˈqiʃ* 'come up.' Several other words were selected for relevant contrasts—*qaˈt͡ʃan* 'when,' *t͡ʃaˈtaq* 'trouble,' *qaˈpaq* 'eyelid,' *myˈʃyk* 'cat,' *pɛˈqɛt* 'only,' *toˈqat͡ʃ* 'small round thing,' *nimiʃˈqa* 'why,' and *sypɛt* 'adjective.' Due to the small size of the subset, only descriptive statistics are reported here. It is hoped that future studies using larger, more rigorously balanced experimental data sets and spontaneous corpus data as in [11] on Japanese can develop this line of research further.

3. RESULTS

3.1. Reliability of conditioning factors

To examine the reliability of the three conditioning factors described by Hahn [4], Table 1 reports the mean Fraction of Locally Unvoiced Frames (FLUF) for the entire data set in a 2x2x2 matrix. A clear distinction emerges between the top left cell, representing the ideal devoicing environment of high vowels in unstressed syllables between voiceless consonants, and all the other cells. On average 74% of the frames were unvoiced for the vowels in the ideal environment, while no notable devoicing occurred in non-high vowels, stressed vowels, or vowels adjacent to voiced consonants.

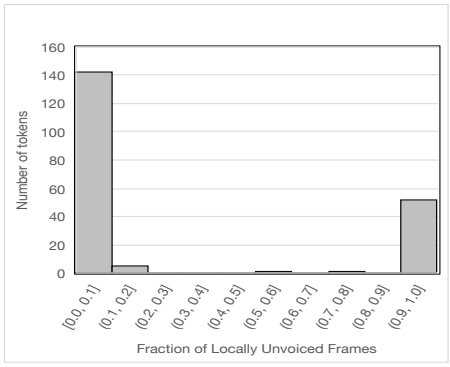
Table 1: Mean FLUF for all vowels in original data set

		Both adjacent consonants voiceless	One/both adjacent consonants voiced
High	Unstressed	0.74 n=1101, sd=0.38	0.08 n=882, sd=0.25
	Stressed	0.09 n=215, sd=0.26	0.01 n=682, sd=0.03
Non-high	Unstressed	0.03 n=497, sd=0.08	0.06 n=491, sd=0.17
	Stressed	0.01 n=618, sd=0.04	0.01 n=971, sd=0.04

This preliminary analysis treated all the data together to get a first impression of the three factors, with no attempt to separate words by number of syllables, target vowels, open vs. closed syllables, or data type (words, sentences, and naturalistic speech). Future analysis could make more fine-grained sub-groups and employ linear modeling to estimate the predictive power of the three factors. The present analysis, though, suggests that the three factors are reliable—only when all three factors align do speakers consistently devoice the vowels.

The FLUF results for all the /i/ vowels in the subset are presented in the histogram in Figure 1. (Histograms for individual speakers looked the same.) The polar distribution indicates that rather than being a gradient phenomenon as in many languages with vowel devoicing [3], Uyghur vowel devoicing is a categorical phenomenon.

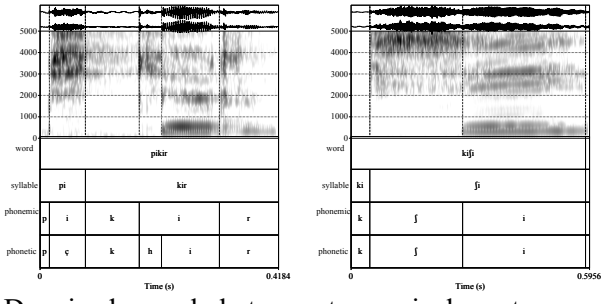
Figure 1: Fraction of Locally Unvoiced Frames (FLUF), subset with /i/ target vowels



2.2. Acoustic description of subset of voiceless vowels

In the present Uyghur data set, devoicing of vowels adjacent to voiceless sibilant consonants /s/, /ʃ/, and /t͡ʃ/ was realized as total deletion or blending with the adjacent consonants for, so that it was impossible to mark a distinct segment for the vowel on the TextGrid. In the subset of vowels selected for detailed analysis in this paper, the /i/ vowels in the first syllables of *kiˈʃi*, *t͡ʃiˈqim*, and *t͡ʃiˈqiʃ* were inseparable from the [ʃ] (see *kiʃi* in Figure 2).

Figure 2: Devoiced vowels in *piˈkir* and *kiˈʃi*

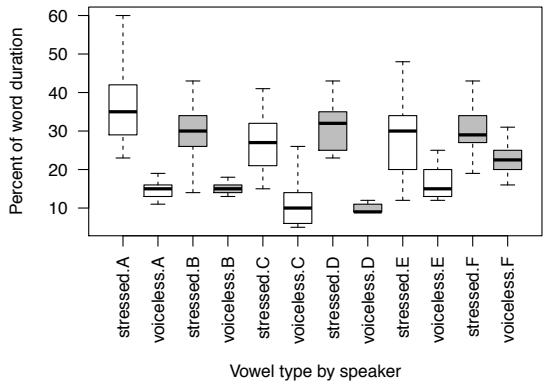


Devoiced vowels between two voiceless stops were clearly present but totally voiceless. The /i/ vowels in the first syllables of *piˈkir* 'idea' and *ʔikˈki* 'two' were distinct from the adjacent stop consonants. On the spectrogram they appeared as aperiodic high-frequency noise, similar to aspiration or voiceless palatal frication. In some cases there were darker bands of energy between 2000-4000 Hz, but it was never clear enough to call formant structure (see *piˈkir* in Figure 2). It was not difficult, however, to hear the difference between voiceless /i/ in these words and voiceless /u/ and /y/ in other words in the data set, even for vowels that were blended or

coarticulated with adjacent sibilants. Analysis of the spectral peak or center of gravity of the aperiodic noise would be an interesting topic for future study (cf. [14] on Japanese).

To compare duration of voiced and voiceless /i/, the ideal environment would probably be two disyllabic words with /i/ in the unstressed first syllable flanked by voiceless consonants in one word but a voiced consonant blocking the devoicing in the other. Unfortunately the wordlist for this study did not contain such a word with a voiced consonant. The next best, then is to compare the duration (normalized as percent of word duration) of the voiceless unstressed /i/ in the first syllables of *piˈkir* and *ʔikˈki* with the voiced stressed /i/ in the second syllables of *piˈkir*, *kiˈʃi*, *ʔikˈki*, *t͡ʃiˈqim*, and *t͡ʃiˈqiʃ*, with results shown per speaker. Six tokens of the first /i/ in *ʔikˈki* were not devoiced and were therefore removed from the voiceless subset. Grouped together, all speakers' voiceless /i/ tokens (mean 15.7, sd=5.6, n=54) were much shorter than the voiced stressed tokens (mean 30.4, sd=8.5, n=147). Individual results for speakers A-F (Figure 3) showed the same trend, with slightly less difference between voiceless and stressed /i/ for speakers E and F. The mean FLUF measurements were 0.97 (sd=0.2, n=54) for voiceless /i/ and 0.02 (sd=0.1, n=147) for stressed /i/. Note that for this subset the voicing threshold (under Pitch…Advanced pitch settings) was set to 0.70, as this setting produced more reasonable results when checked against the amount of voicing visible on the waveform.

Figure 3: Normalized duration of voiceless /i/ and stressed /i/, speakers A-F (grey color for readability)



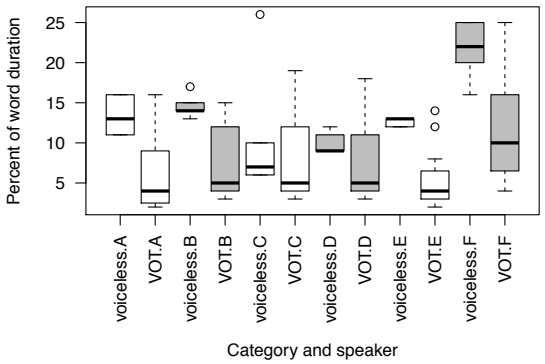
The fact that the voiceless vowels adjacent to sibilants were essentially deleted raises the question of whether the observed duration of devoiced vowels between stops is long enough to constitute a real vowel or whether it was just a brief transition between the articulation of the two stops. This question will be addressed from three angles—by comparing them with the duration of unstressed voiced /i/, by comparing them with the VOT of voiceless stops, and by examining the duration of the sibilant segments adjacent to/coarticulated with the devoiced vowels.

As the wordlist for this study did not contain voiced unstressed /i/ in a disyllabic word, the duration of voiced unstressed /i/ can be estimated by looking at the difference between stressed and unstressed tokens of a different vowel, such as /a/ in the words *qaˈt͡ʃan* 'when,' *t͡ʃaˈtaq* 'trouble,' and *qaˈpaq* 'eyelid'. In these words the ratio between the normalized duration of the stressed and unstressed tokens was 1.60:1. Following that ratio, a normalized duration of about 19 would be expected for voiced unstressed /i/, which is slightly longer than the duration observed for voiceless unstressed /i/ above, but still much closer than the 30.4 observed in stressed /i/.

There was an unstressed voiced /i/ in one of the sentences in the original data set, in the first syllable of *nimiʃˈqa* 'why?'. Its normalized duration, adjusted by a factor of 1.5 to account for a trisyllabic vs. disyllabic word, averaged 14.0 (n=30, sd =2.95). This is actually shorter than the normalized duration of voiceless /i/ reported above. The comparison is not ideal, though, because *nimiʃˈqa* came from a sentence rather than a word, it has three syllables rather than two, and /i/ is between nasals rather than stops.

The second angle for assessing whether devoiced /i/ between stops is actually deleted is looking at the VOT of voiceless stops. Syllable-initial voiceless stops in Uyghur are generally aspirated [4]. It is possible the duration observed between the two stops is just the normal amount of aspiration following the release of the voiceless stops. To test this, the VOT of the /p/ preceding voiced /ɛ/ in *pɛˈqɛt* was compared with the duration of voiceless /i/ in the first syllable of *piˈkir*.

Figure 4: Normalized VOT vs voiceless vowel duration after /p/, speakers A-F (grey color for readability)

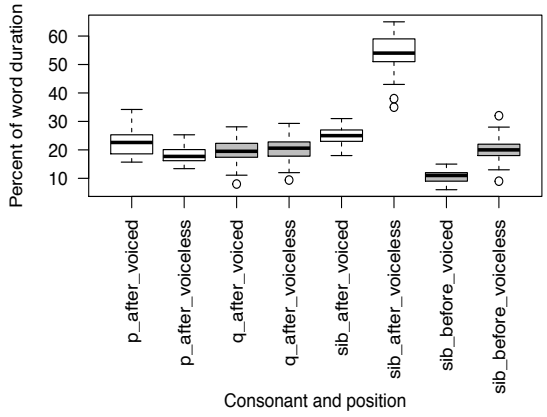


As Figure 4 shows, the median duration of voiceless /i/ was longer than the VOT of /p/ before voiced /ɛ/ for each speaker. For speakers C and D, there was a great deal of overlap in the distributions. For all speakers grouped together, the normalized duration of the voiceless /i/ (mean 13.8, sd=5.3, n=30) was more than twice as long as the VOT of /p/ in *pɛˈqɛt* (mean 5.1, sd=2.9, n=30). However, the 60.5ms mean for voiceless /i/ is still within a normal range for VOT of aspirated consonants as reported by

Cho & Ladefoged [1]. Further study of the VOT of voiceless stops in Uyghur would be needed to strengthen this line of argument.

The third angle of inquiry is the duration of the sibilant adjacent to the voiceless vowel. If the sibilant is longer than usual, it is possible that sibilant noise is actually a method of realizing the voiceless vowel. (Alternatively, the vowel could be deleted but the sibilant lengthened to compensate.) Figure 5 shows that the normalized duration of [ʃ] after a voiceless vowel in *kiˈʃi* (mean=53.7, sd=7.4, n=30) was longer than the duration of [ʃ] after a voiced vowel in *myˈʃyk* (mean=25.0, sd=3.0, n=30); and likewise the [ʃ] portion of /t͡ʃ/ before a voiceless vowel in *t͡ʃiˈqiʃ* 'come up' (mean=20.2, sd=4.9, n=30) was longer than the corresponding [ʃ] before a voiced vowel in *t͡ʃaˈtaq* 'trouble' (mean=10.9, sd=2.1, n=30). Results by speaker showed no clear differences. This trend was not observed in stop consonants. The normalized duration of /p/ (closure to release) after voiceless /i/ in *syˈpɛt* (mean=18.4, sd=3.0, n=30) was actually shorter than /p/ after a voiced vowel in *qaˈpaq* (mean=22.8, sd=5.1, n=30), and the duration of /q/ after voiceless /i/ in *t͡ʃiˈqiʃ* and *t͡ʃiˈqim* (mean=20.3, sd=3.6, n=57) was about the same as the /q/ after voiced vowels in *toˈqat͡ʃ* and *pɛˈqɛt* (mean=19.4, sd=3.8, n=60).

Figure 5: Normalized duration of stops and sibilants adjacent to voiced and voiceless vowels (grey color for readability)



In sum, the three angles of investigation, though subject to certain limitations, seem to converge on the conclusion that voiceless vowels between stops are indeed present rather than deleted and that voiceless vowels adjacent to sibilants are also not deleted but realized as homorganic sibilant frication.

4. DISCUSSION

The realization of a voiceless vowel as frication of some kind is not altogether surprising from an articulatory perspective. A vowel between two obstruents has two tight constrictions on either side of it. The devoiced vowels are high vowels, for which the oral opening would be rather narrow even if they were voiced. Ohala notes that high vowels, "by virtue of their high close constriction, impede the flow of air and constitute 'almost' obstruents. In conjunction with other factors, they can reduce ΔP_{glot} enough to extinguish voicing" ([13]: 3). When the articulators are only separated with a tiny opening in the transition between the stop closures in words like *piˈkir* or not opened any further than the preceding or following sibilants in words like *kiˈʃi* or *t͡ʃiˈqim*, then frication is the natural consequence. If the duration of the frication makes up for the duration that would be expected in a vowel, then the frication very well may be the way of realizing the voiceless vowel. If this is so, then the phonotactics and syllable structure are preserved, with the frication as the syllable nucleus.

The articulatory explanation of the frication also provides an answer to the question of why it is the high vowels that devoice in Uyghur. The oft-cited explanation for the cross-linguistic tendency of high vowels to devoice more than other vowels is that high vowels are shorter in duration than non-high vowels and thus their gestures are more easily overlapped by consonant gestures (see [3, 8]). This makes sense in languages where the devoicing happens as a gradient (as [8] found for Korean), where the likelihood of devoicing is highly sensitive to rate of speech (cf. [6] on Turkish), or where non-high vowels also devoice in very fast speech (cf. [10] on Japanese, [17] on Spanish). However, in the present data for Uyghur, devoicing happened consistently even in the reading style of experimental conditions, and virtually no devoicing was observed of any non-high vowels (see Table 1). These distributional patterns combined with the observation of frication suggest that the close constriction of high vowels than low vowels may be more relevant than their short duration.

The articulatory account, however, is stronger as a historical explanation of how high vowels came to be devoiced than as a synchronic explanation of why speakers devoice high vowels now. Contemporary Uyghur vowel devoicing is most likely part of the phonology rather than the phonetic implementation, as it is very consistent and also probably quite old. Devoicing of high vowels (and phenomena like "reduction," "loss," "dropping", or "disappearance") occurs all across the modern Turkic family [7], and similar processes are noted in several Mongolic languages [5, 11] as well as Japanese [10, 12, 14, 15], Korean [8], and Northwest Mandarin [2]). While it is possible that these processes all arose independently or spread across Asia after the Turkic migration, it is also possible that vowel devoicing was historically an areal feature in northeast Asia. If the vowel devoicing processes in e.g. Uyghur in the east and Turkish in the west have the same origin, they must be quite old, as Turkic speakers had spread from present-day Mongolia across Central Asia by 600AD [7].

7. REFERENCES

[1] Cho, T., Ladefoged, P. 1999. Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics* 27, 207-229.

[2] Dwyer, A. M. 1995. From the Northwest China *Sprachbund*: Xúnhuà Chinese dialect data. 從中國西北部的語言區域關係體: 循化話語言材料. *Yuen Ren Society Treasury of Chinese Dialect Data* 元任學會漢語方言資料寶庫 1: 143–182.

[3] Gordon, M. 1998. The phonetics and phonology of non-modal vowels: a cross-linguistic perspective. *Berkeley Linguistics Society* 24, 93-105.

[4] Hahn, R. F. 1991. *Spoken Uyghur*. Seattle: University of Washington Press.

[5] Janhunen, J., ed. 2005. *The Mongolic languages.* London: Routledge.

[6] Jannedy, S. 1995. Gestural phasing as an explanation for vowel devoicing in Turkish. In *Ohio State University Working Papers in Linguistics* 45, 56-84.

[7] Johanson, L., Csató, É. (eds.) 1998. *The Turkic Languages*. London: Routledge.

[8] Jun, S., Beckman, M. E. 1994. Distribution of devoiced high vowels in Korean. *Third International Conference on Spoken Language Processing* 94, 479-482.

[9] Kim, S. S. "Santa." In Janhunen, J, ed. 2005. *The Mongolic languages.* London: Routledge, 346-363.

[10] Maekawa, K., Kikuchi, H. 2005. Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. In: van de Weijer, J., Nanjo, K., Nishihara, T. (eds.) *Voicing in Japanese*. New York: Mouton de Gruyter, 205-228.

[11] McCloy, D., Yurong, Puthuval, S. 2016. Phonetically-conditioned vowel devoicing in Chahar Mongolian. Poster presented at the 90th Annual Meeting of the Linguistic Society of America, Washington, DC.

[12] Ogasawara, N. 2012. Lexical representation of Japanese vowel devoicing. *Language and Speech* 51(1) 5-22.

[13] Ohala, J. J. 1997. Aerodynamics of phonology. *Proceedings of the 4th Seoul International Conference on Linguistics* [SICOL]. Seoul: Linguistic Society of Korea, 92-97.

[14] Tsuchida, A. 1994. Fricative-vowel coarticulation in Japanese: acoustic and perceptual evidence. *Working papers of the Cornell Phonetics Laboratory* 9, pp. 183-222.

[15] Tsuchida, A. 2001. Japanese vowel devoicing: cases of consecutive devoicing environments. *Journal of East Asian Linguistics* 10: 3, 225-245.

[16] Tursun, D., Hemdulla, E. 2010. 维吾尔语中清华元音实验语音学研究. Experimental phonetic study on voiceless vowels in Uyghur. *Journal of Chinese Information Processing*, May 2010, 117-123.

[17] Uber, D. R. 1989. Stress-timing, Spanish rhythm, and particle phonology. *Linguistic Society of America Annual Meeting 1989*.

[18] Yakup, M., Sereno, J. A.. 2016. Acoustic correlates of lexical stress in Uyghur. *Journal of the International Phonetic Association* 46.1, 61-77.