音声研究　第 22 巻第 1 号
2018（平成 30）年 4 月
13–20 頁

Journal of the Phonetic Society
of Japan, Vol. 22 No. 1
April 2018, pp. 13–20

研究ノート

The Phonetic Structure of Dzongkha: A Preliminary Study[1)]

Seunghun Lee*, ** and Shigeto Kawahara ***

ゾンカ語の音声特徴に関する一考察

SUMMARY: Dzongkha is the national language of Bhutan, but its phonetic nature has not been studied instrumentally in depth. This research note thus explores the phonetic structure of this language, focusing on its three aspects: (i) the vowel quality, (ii) the tonal contrast, and (iii) the four-way laryngeal contrast. The results show (i) that the first three formants are necessary to distinguish the eight vowels of this language, (ii) that the tonal contrast most clearly manifests itself at the onset of syllables, and (iii) that the laryngeal contrast is acoustically differentiated in terms of both VOT and F0 of the following vowels. Although the current analysis is limited in that it is based on the data from a single speaker, it is hoped that it provides a stepping stone toward further analyses of Dzongkha, and comparative phonetic studies of other related languages.

Key words: Dzongkha, vowel, tone, phonation, VOT, F0, consonant-tone interaction

1. Introduction

Dzongkha (a.k.a. Bhutanese) is a Tibeto-Burman language, and is the national language of the Kingdom of Bhutan. According to Ethnologue, it is spoken by about 226,000 speakers[2)]. Despite becoming designated as the national language, however, Bhutan's political situation is such that parents tend to encourage their children to learn English instead of Dzongkha for socio-economic reasons (Nishida 2004). Because of this socio-political situation, Dzongkha is being endangered, and it is important that we document its linguistic properties as soon as possible.

While there is an impressionistic description by Tshering and van Driem (2015), itself a revised version of van Driem and Tshering (1998), in addition to a brief phonetic analysis by Watters (2002), Dzongkha's phonetic structure has not been studied in depth using recent instrumental technologies. From the previous studies (Tshering and van Driem 2015, Watters 2002), we know that Dzongkha has eight contrastive vowels, a two-way tonal contrast (H(igh) and L(ow)), and a four way laryngeal contrast, each category being referred to as "aspirated", "voiceless", "voiced," and "devoiced" (cf. Nishida 2016). Our aim in this research note is to systematically explore the acoustic realizations of these three types of phonological contrasts.

The tone is contrastive in vowel-initial syllables and syllables with sonorant onsets. However, there are tone-consonant restrictions in such a way that obstruents can be followed by only particular types of tones. Concretely, syllables with aspirated and voiceless onsets predictably bear H-tones, and "devoiced" and voiced onsets predictably bear L-tones (Tshering and van Driem 2015, pp. 39–40)[3)].

One general caveat is in order. The phonetic analysis presented in this paper is limited in that it is based on the data from a single speaker, whose speech is made available in van Driem and Tshering (1998); therefore, the findings of this paper should be interpreted with caution, and should be replicated with a larger number of speakers in future research. Meanwhile, it is hoped that we can situate this study as a stepping stone for future research of this language, as well as comparative phonetic studies of other related Tibeto-Burman languages, such as Tamang and Dränjongke.

2. Method

2.1 Vowel Quality and the Tonal Contrast

All the recordings came from van Driem and Tshering (1998). The speaker is Tshering himself, who is a

* International Christian University（国際基督教大学）
** University of Venda（ヴェンダ大学）
*** Keio University（慶應義塾大学）

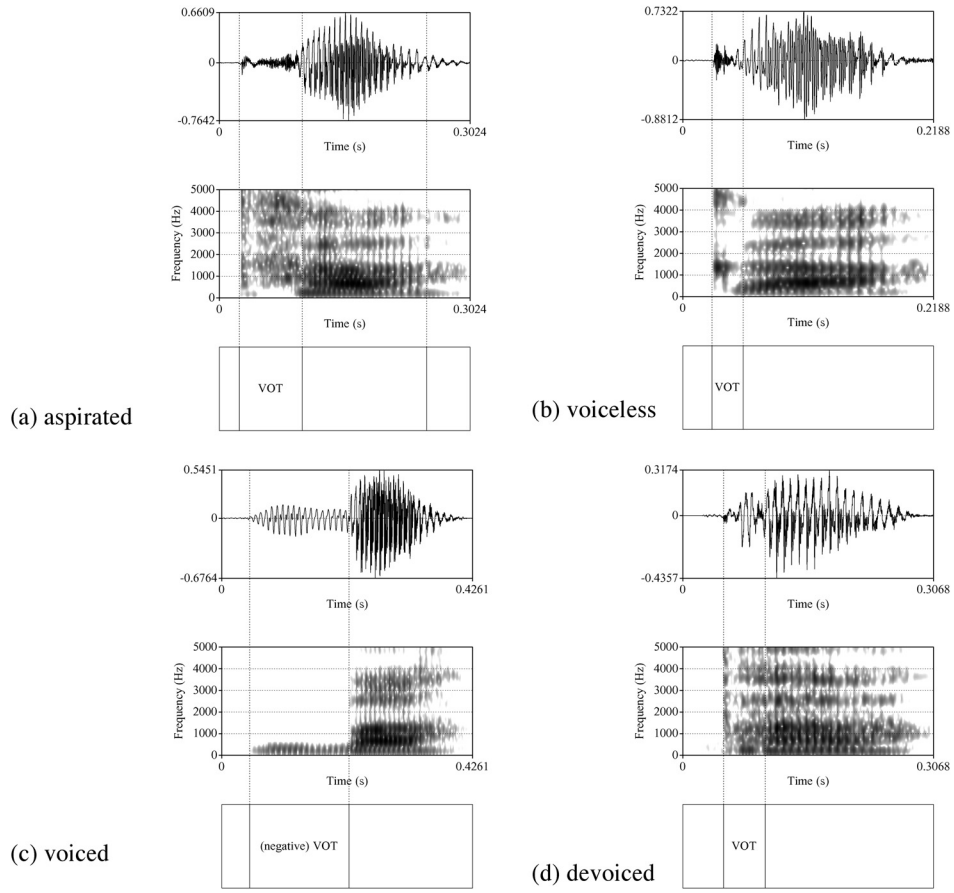

Figure 1 The four laryngeal categories in Dzongkha. Top left (a) = aspirated (= [t^h]); top right (b) = voiceless (= [ka]); bottom left (c) = voiced (= [ba]); bottom right (d) = "devoiced" (= [b̤a]).

male, native speaker of Dzongkha. He was about 30 years old at the time of recording. He worked as the broadcaster of the national TV program in Bhutan. He was born in Thimphu, and raised in Gaselô, in Wangdi Phodrang district. His dialect of Dzongkha should thus be characterized as "Gasebi-kha." All the sound samples were digitized at 44.1 kHz sampling rate.

Dzongkha has eight vowels, transcribed by Tshering and van Driem (2015) as "a", "ä", "e", "i", "o", "ö", "u", and "ü". Among those, five of them have a short-long length contrast (Tshering and van Driem 2015, p. 45; Nishida 2004, p. 20). The vowels with an umlaut sign ("ä", "ö" and "ü") are always long, and these vowels are indeed noticeably longer (ca. 300 ms) than the other vowels (ca. 150 ms) in the recording. Each vowel was read with H-tone and L-tone.

Both F0 and spectral properties of these vowels were analyzed using Praat (Boersma 2001). The first three formant values, averaged across the entire vowel intervals, were extracted. In addition to these vowels read in isolation, the recording included syllabary readings, which included 34 H-tone tokens and 33 L-tone tokens. The F0 patterns of these syllables were analyzed. We also addressed one consonant-tone interaction in Dzongkha by examining 18 syllables with a voiced onset and 16 syllables with what has been referred to as "devoiced" consonants (Tshering and van Driem 2015). The motivation of this analysis came from their impression that "devoiced" consonants are distinguished from other categories in terms of the F0 of the following vowel.

2.2 The Laryngeal Contrast

The syllabary readings of the recording in van Driem and Tshering (1998) included the obstruents of the four laryngeal types, all followed by a vowel [a]. To ana-

lyze the acoustic differences between the four classes of obstruents, the lag between the release of the consonant and the onset of the following vowel is annotated using Praat (Boersma 2001). These intervals are taken to represent VOT of different types of obstruents (see Figure 1). The onset of the vowels was aligned with the point where the vocalic formants started (especially those higher than F1), together with clear periodic energies in the waveform display[4]. Voiced consonants showed clear voicing during closure, despite being word-initial (Figure 1(c)). The closure voicing interval was taken as a negative VOT value. Based on Praat annotations such as those illustrated in Figure 1, the durations of these intervals were automatically extracted using a script.

A 20 ms analysis window was created at the onset of the following vowel, and the average F0 within that analysis window was calculated for each type of consonant. The analysis is based on a small number of tokens produced by a single native speaker (aspirated = 5; voiceless = 7; devoiced= 11; voiced = 13). We thus did not attempt to apply statistical analyses to the results of the syllabary reading.

While the syllabary reading tokens may offer "clear" information about the phonetic structure of Dzongkha, as syllabary readings are arguably free from lexical factors that may influence speech production (e.g. Baese-Berk and Goldrick 2009, Gahl 2008, Munson and Solomon 2004, Scarborough 2012, 2013, Wright 2004), they may be "artificial" in the sense that they are not produced as meaningful units in Dzongkha. To overcome this limitation, since the recording in van Driem and Tshering (1998) also included pronunciation of basic vocabulary in Dzongkha, we analyzed the VOT and the F0 in the following vowel using these real words, in the same way that we used for the syllabary readings. The *N*s analyzed using these real words are: aspirated = 26; voiceless = 49; "devoiced" = 9; voiced = 57. We conducted statistical analyses based on these tokens.

3. Results

3.1 The Formant Characteristics of the Vowels

We first started by exploring the acoustic nature of each vowel in Dzongkha ("a", "ä", "e", "i", "o", "u", "ö", "ü"). This first analysis is based on vowel-only readings, each vowel produced with H-tone and L-tone. Figure 2 plots the standard F1 and F2 chart of these eight vowels, which shows that for those vowels without umlaut signs ("a", "e", "i", "o", "u"), their F1 and F2 distribute in the expected F1–F2 regions (except that

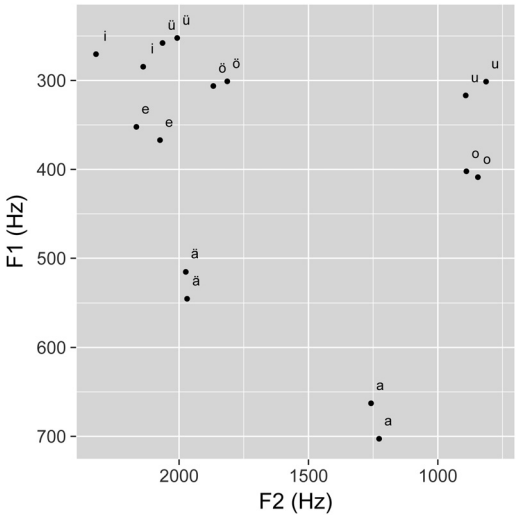

Figure 2 The F1–F2 vowel chart of Dzongka vowels. The analysis is based on the vowel-only reading tokens. Each vowel is produced twice, once with H-tone and once with L-tone.

perhaps, the F1 values of mid vowels distribute closer to those of high vowels than halfway between high and low vowels). We also observe that umlauted versions have lower F1—and, more clearly, higher F2—compared to non-umlauted versions, which suggests that they are likely to be fronted versions of the corresponding non-umlauted vowels (i.e. umlaut represents frontness, as in German[5]); i.e. "ä" = /æ/, "ö" = /ø/, "ü" = /y/ (Johnson 2003, Reetz and Jongman 2008, Stevens 1998). The lowering of F1 in umlauted vowels can potentially be understood as a consequence of an additional palatal gesture associated with the fronting of vowels[6]. Finally, we observe that in Figure 2, four types of vowels are clustered in the left-top region ("i", "e", "ö", "ü").

In order to explore how these four vowels ("i", "e", "ö", "ü") are distinguished acoustically, we examined their F3, which is known to distinguish unrounded front vowels from rounded front vowels (Reetz and Jongman 2008, p. 184). The results appear in Figure 3, which plots F3 values on the y-axis and F2 values on the x-axis. As expected, F3 distinguishes unrounded front vowels ("e", "i") and rounded front vowels ("ö", "ü"), in that the latter group has much lower F3.

By way of summary, Table 1 shows the first three formants of the eight vowels in Dzongkha (the values are based on the H-toned tokens).

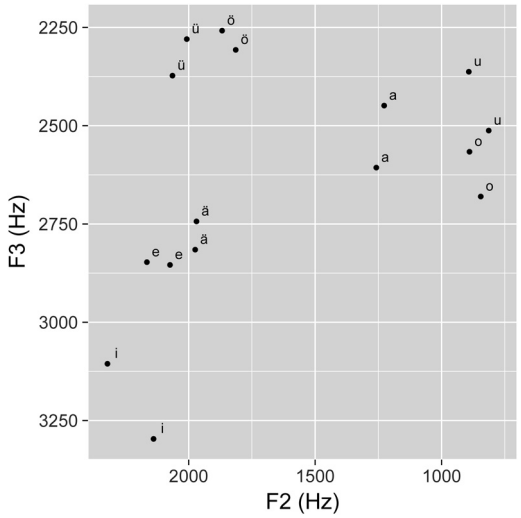

Figure 3 The F2–F3 vowel chart of the Dzongka vowels.

Table 1 The first three formant values of the eight vowels (averaged over the entire vocalic intervals) in Dzongkha (Hz). The values are based on H-toned tokens.

Vowel	F1	F2	F3
“a”	703	1227	2449
“e”	367	2074	2854
“o”	409	845	2680
“i”	285	2139	3297
“u”	317	892	2363
“ä”	545	1969	2743
“ö”	306	1868	2258
“ü”	252	2007	2280

3.2 Tonal Realizations

Figure 4 shows F0 curves of H-toned and L-toned syllables, based on the vowel-only readings, the same dataset that was used in Figures 2 and 3. The F0 contours were obtained by dividing the vocalic intervals into five equally-timed windows, and taking the average F0 values within each window[7]. It shows that H-toned and L-toned syllables are separated clearly at the onset of syllables, and the differences are neutralized toward the end for some vowels (see also Watters 2002 for a similar finding). The tonal difference seems to persist throughout the syllables for “e”, “i”, and “u”[8].

Since the tonal patterns are comparable—if not identical—across different vowel qualities, Figure 5 shows the average F0 plots of H-toned and L-toned syllables, based on all syllabary reading tokens, most of which had an onset consonant. Figure 5 is based on an analysis that is the same as that of Figure 4, although it

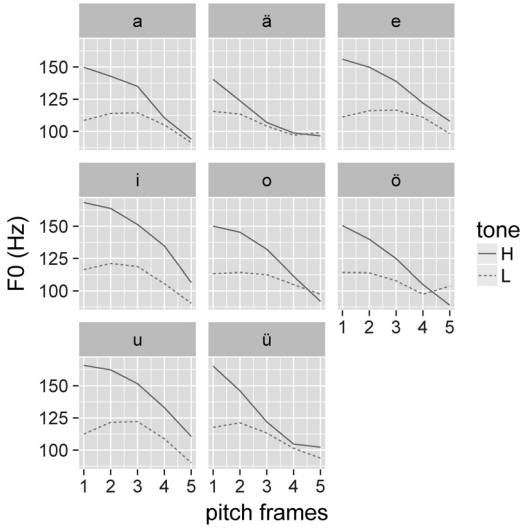

Figure 4 F0 movement of two types of tones, separated by vowel. The analysis is based on the vowel-only reading. H-tones = solid lines; L-tones = dotted lines.

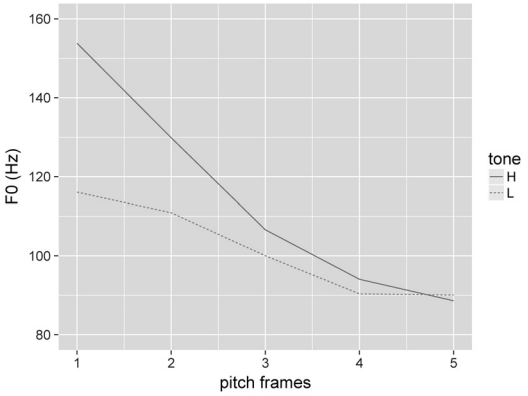

Figure 5 F0 differences of all syllables. The analysis is based on the syllabary readings.

targets only vocalic intervals. On average, at the onset of the syllables, H-toned and L-toned syllables differ by 30–40 Hz; the differences in F0 get smaller toward the end of the syllables, and are not observed in the fifth, final frame. What is emerging through our analysis is that tonal differences in Dzongkha most clearly manifest themselves at the onset of vowels.

In addition to the analysis of these F0 differences due to lexical H-tone vs. L-tone contrast, we also analyzed one type of consonant-tone interaction. Specifically, we examined 18 syllables with a voiced onset consonant and 16 syllables with what Tshering and van

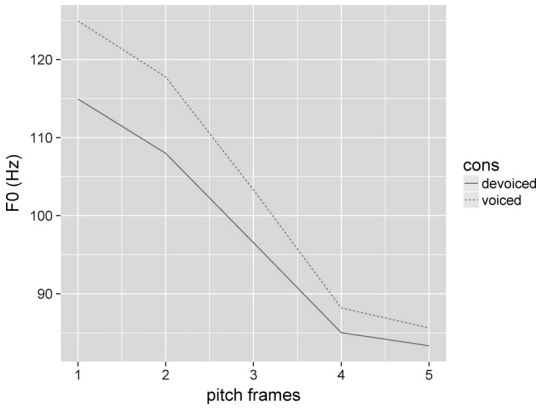

Figure 6 Effects of "devoiced" consonants on F0. The y-axis scale is identical to that of Figure 5.

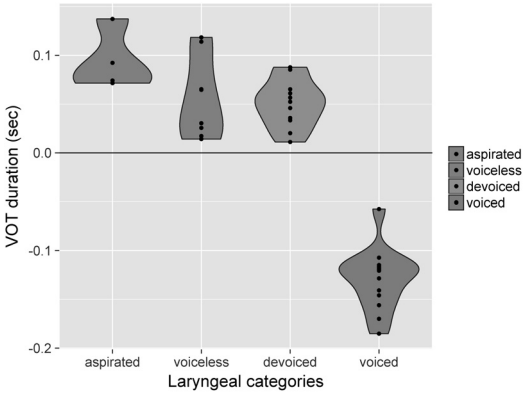

Figure 7 VOT of the four laryngeal categories (based on the syllabary reading).

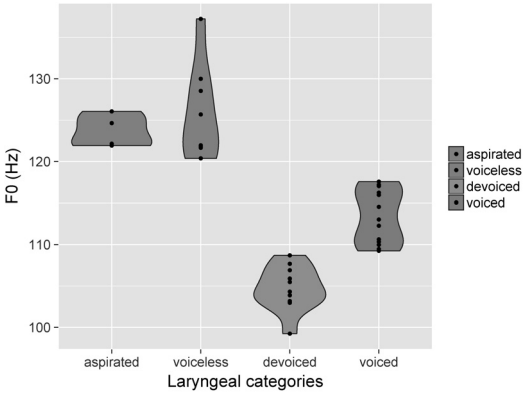

Figure 8 F0 of the four laryngeal categories (based on syllabary reading). The first two types of categories are generally H-toned, and the last two types are generally L-toned.

Driem (2015) and Watters (2002) referred to as a "devoiced" onset consonant. Recall that the lexical tone of these syllables was generally limited to L-tones (Tshering and van Driem 2015). The result, which appears in Figure 6, shows that the F0 is higher after voiced consonants than after "devoiced" consonants, the pattern that is opposite from what is expected if "devoiced" consonants were voiceless, as voiceless consonants usually raise F0 of the surrounding vowels (e.g. Hombert et al. 1979, Kingston and Diehl 1994, Lee 2008). Our conjecture at this point is that these consonants are actually breathy consonants, which are known to lower F0 of the surrounding vowels cross-linguistically (e.g. Baumbach 1987, Lee 2008; cf. Halle and Stevens 1971). This conjecture is supported by the impressionistic description offered by Tshering and van Driem (2015, p. 56) "In articulatory terms, devoiced consonants are unvoiced, but, in contrast to the voiceless consonants, they are followed by a murmured or 'breathy voiced' vowel."

3.3 Laryngeal Contrast

Figure 7 is a violin plot which shows the VOT values of the four laryngeal categories, based on all the syllabary readings. We observe that voiced consonants are separated from the remaining of the three categories in that they all have negative VOT values (i.e. closure voicing); their closure voicing is usually longer than 100 ms. Among the other three categories, aspirated consonants show the largest VOT values, which are close to or slightly shorter than 100 ms. Voiceless and "devoiced" consonants show intermediate values (around 50 ms). One important question that arises is thus how these two categories are phonetically distinguished, which we turn to next.

Figure 8 is a violin plot which shows the results of the F0 analyses; recall that what has been measured are the average F0 values of the 20 ms analysis windows, placed at the onset of the following vowels. Figure 8 shows that voiceless and "devoiced" consonants, which showed comparable VOT profiles, are separated out in terms of this measure. In addition, voiced consonants show lower F0 than voiceless consonants, an observation that is compatible with cross-linguistic observations (e.g. Hombert et al. 1979, Kingston and Diehl 1994, Lee 2008).

Since the number of syllabary reading tokens was limited, to the degree that we were not able to apply statistical analyses, we analyzed the real words recorded in van Driem and Tshering (1998). Figure 9 shows the results of the VOT analysis based on these real word to-

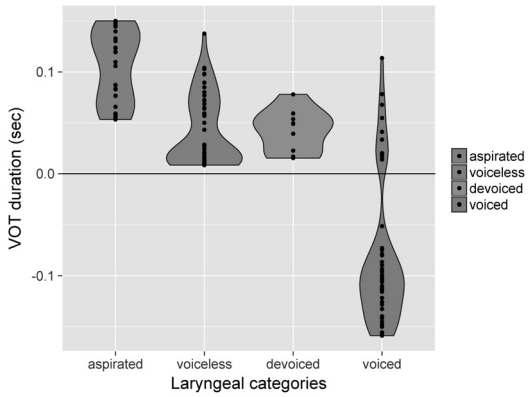

Figure 9 VOT of the four laryngeal categories (based on real words).

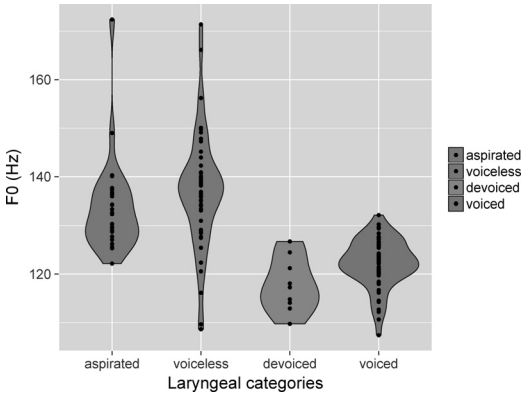

Figure 10 F0 of the four laryngeal categories (real words).

kens. It turned out that the results look very similar to what we observed in syllabary readings (Figure 7), except that we observe several tokens of positive VOT values for the voiced category. This may be related to the fact that all of these consonants were pronounced word-initially in isolation, and hence initiating closure voicing was particularly challenging (Hayes 1999, Kingston and Diehl 1994, Westbury and Keating 1986)[9]. To statistically assess the differences between the four laryngeal categories, a one-way ANOVA with VOT duration as the dependent variable and the four laryngeal categories as the independent variable, was run, whose effect was significant ($F(3, 137) = 101.4, p < .001$). Multiple post-hoc comparisons using Tukey Honest Significant Difference (HSD) tests show that all the differences but the difference between voiceless and "devoiced" are significant, all at the $p < .001$ level.

The patterns of F0 in Figure 10 more or less follow the same pattern that we observed in syllabary reading (Figure 8): aspirated and voiceless consonants show high F0 in the following vowels; "devoiced" consonants show the lowest F0 and voiced consonants show slightly higher F0.

Statistically, one way ANOVA shows that there is a significant effect of the laryngeal categories on F0 ($F(3, 137) = 31.12, p < .001$); Tukey HSD tests show that there are no significant differences between aspirated and voiceless categories; both aspirated and voiceless consonants show higher F0 than devoiced and devoiced consonants, all at the $p < .001$ level. No statistical differences were observed between voiced and "devoiced" consonants.

Table 2 summarizes how the four laryngeal categories are distinguished in Dzongkha.

Table 2 How the laryngeal contrast is differentiated in Dzongkha.

	VOT	F0
aspirated	long	high
voiceless	short	high
"devoiced"	short	low
voiced	negative	low

4. Conclusion

The aim of this research note was to explore the basic phonetic structure of Dzongkha. Our preliminary analyses have revealed that (i) Dzongkha's eight vowels are distinguished in terms of F1, F2, and F3, (ii) the lexical H-tone vs. L-tone contrast manifests itself at the onset of syllables, and (iii) the four-way laryngeal contrasts are distinguished in terms of both VOT and F0 of the following vowels. As declared at the outset of this paper, these conclusions have limitations in the sense that they are based on speech produced by a single speaker. We are working to seek other speakers of Dzongkha to examine the generality of the current findings. We also aim to compare the current findings with related Tibeto-Burman languages, including Dränjongke and Tamang.

Notes

1) This paper is based on the talks given at the 31st annual meeting of the Phonetic Society of Japan and Seoul International Conference on Speech Science 2017, whose proceedings papers appeared as Lee et al. (2017a) and Lee et al. (2017b), respectively. This research is supported by the Strategic Japanese-Swiss Science and Technology Programme of JSPS and SNSF. We would like to thank George van Driem, Hyun Kyung Hwang, Hanna Kaji, Fuminobu

Nishida, Tomoko Monou, Jeremy Perkins, Haruka Tada, and Karma Tshering for their help on this and related projects. Two anonymous reviewers provided useful comments which improved presentation and the analyses of this paper. All remaining errors are ours.

2) https://www.ethnologue.com/language/dzo, last access Feb., 2018.

3) Watters (2002, p. 17) provides a near-minimal pair /ʃii/ 'field' vs. /ʃi/ 'die', in which the first word has L-tone and the second word has H-tone. We have consulted Karma Tshering, a native speaker of Dzongkha, and also the speaker who provided the data for the current study. He informed us that he does not know how to pronounce the first word (i.e. /ʃii/ with L-tone). There are two possibilities for this discrepancy. One is that Watters (2002) is dealing with a different dialect of Dzongkha. The second is that this is simply a typo.

4) An anonymous reviewer raised a potential concern to the effect that the onset of the vowels should have been aligned with the onset of voicing, rather than the onset of higher formant structure, the latter of which comes slightly later. The rationale is that we should follow the original definition of VOT by Lisker and Abramson (1964). This difference between the onset of voicing and the onset of higher formant structure often occurs in natural languages because supralaryngeal gestures and laryngeal gestures are not perfectly synchronized, although they are undoubtedly coordinated (i.e. articulatory binding: Kingston 1985, 1990, Silverman 1995, Shaw and Kawahara 2018). Thus, while we appreciate this reviewer's comment, we would like to point to the fact that it is not uncommon to identify the onset of vowels using formant structures higher than F1. For example, Davidson (2010) defines vocalic intervals as "a period of voicing...with formant structure containing *a visible second formant* that ended with abrupt lowering of intensity at the onset of [the following consonant]" (p. 276: emphasis added). Kawahara (2006) likewise states "[t]he onset of [the preceding vowel] was set where *F3* becomes visible" (p. 552: emphasis added). In this sense, our estimates of VOT may be longer than what we would have obtained if we set the onset of the vowels to the onset of vocalic voicing (Lisker and Abramson 1964). Ultimately, however, we believe that what is more important is consistency within the analysis, rather than arguing how we should decide where the vowels start. It is most likely that thanks to articulatory binding, the onset of voicing and the onset of higher formant structure should be highly correlated after all.

5) Tshering and van Driem (2015) state that "[t]he Dzongkha vowel ö has no English counterpart. The Dzongkha vowel ö is like the vowel [œ] in French *oeuf*, German *plötzlich* or Dutch *lus* (p. 52)." This impressionistic statement is compatible with the result of the acoustic analysis.

6) Alternatively, umlauted vowels may have lower F1 because of their longer duration. It is known that in Japanese, long vowels are more dispersed from each other than short vowels are (Hirata and Tsukada 2009).

7) This process was automated using a scripting function in Praat.

8) In all syllables, tones fall toward the end. This may be due to the fact that they were read in isolation, and declarative sentences in Dzongkha have sentence-final low tones (Nishida 2004). Future research should use a frame sentence to address this issue of whether the observed fall in pitch is due to phrasal/sentential tones.

9) Some important remaining questions include whether these voiced consonants with positive, rather than negative, VOT would be appropriately perceived as voiced, and if so, how. A perception experiment is necessary to address these questions.

References

Baese-Berk, M. and M. Goldrick (2009) "Mechanisms of interaction in speech production." *Language and Cognitive Processes* 24, 527–554.

Baumbach, E. (1987) *Analytical Tsonga grammar*. Pretoria, South Africa: University of South Africa.

Boersma, P. (2001) "Praat, a system for doing phonetics by computer." *Glot International* 5, 341–345.

Davidson, L. (2010) "Phonetic bases of similarities in cross-language production: Evidence from English and Catalan." *Journal of Phonetics* 38, 272–288.

Gahl, S. (2008) "Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech." *Language* 84, 474–496.

Halle, M. and K. N. Stevens (1971) "A note on laryngeal features." *Research Laboratory in Electronics, Quarterly Progress Report* 101, 198–213.

Hayes, B. (1999) "Phonetically-driven phonology: The role of Optimality Theory and inductive grounding." In M. Darnell, E. Moravscik, M. Noonan, F. Newmeyer and K. Wheatly (eds.) *Functionalism and formalism in linguistics, vol. 1: General papers*, 243–285. Amsterdam: John Benjamins.

Hirata, Y. and K. Tsukada (2009) "Effects of speaking rate and vowel length on formant frequency displacement in Japanese." *Phonetica* 66, 129–149.

Hombert, J.-M., J. Ohala and W. G. Ewan (1979) "Phonetic explanations for the development of tones." *Language* 55, 37–58.

Johnson, K. (2003) *Acoustic and auditory phonetics: 2nd edition*. Malden and Oxford: Blackwell.

Kawahara, S. (2006) "A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese." *Language* 82, 536–574.

Kingston, J. (1985) *Phonetics and phonology of the timing of oral and glottal events*. Doctoral dissertation, University of California, Berkeley.

Kingston, J. (1990) "Articulatory binding." In J. Kingston and

M. Beckman (eds.) *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, 406–434. Cambridge: Cambridge University Press.

Kingston, J. and R. Diehl (1994) "Phonetic knowledge." *Language* 70, 419–454.

Lee, S. (2008) *Consonant-tone interaction in Optimality Theory*. Doctoral dissertation, Rutgers University.

Lee, S., S. Kawahara, H. Kaji and H. Tada (2017a) "The acoustic manifestation of laryngeal contrasts in Dzongkha: A preliminary study." *Proceedings of SICSS 2017*.

Lee, S., S. Kawahara, H. Tada and H. Kaji (2017b) "A preliminary acoustic study of tone in Dzongkha." *Proceedings of the 31st meeting of the Phonetic Society of Japan*, 114–119.

Lisker, L. and A. Abramson (1964) "A cross-language study of voicing in initial stops: Acoustical measurements." *Word* 20, 384–422.

Munson, B. and N. Solomon (2004) "The effect of phonological neighborhood density on vowel articulation." *Journal of Speech, Language, and Hearing Research* 47, 1048–1058.

Nishida, F. (2004) "Zonka-go gasa-hoogen no on'in taikei." *Reitaku University Journal* 78, 13–39.

Nishida, F. (2016) *Zonka-go Kiso 1500-go*. Daigaku Shorin.

Reetz, H. and A. Jongman (2008) *Phonetics*. Oxford: Blackwell-Wiley.

Scarborough, R. (2012) "Lexical similarity and speech production: Neighborhoods for nonwords." *Lingua* 112, 164–176.

Scarborough, R. (2013) "Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation." *Journal of Phonetics* 41, 491–508.

Shaw, J. and S. Kawahara (2018) "The lingual gesture of devoiced [u] in Japanese." *Journal of Phonetics* 66, 100–119.

Silverman, D. (1995) *Phasing and recoverability*. Doctoral dissertation, University of California, Los Angeles.

Stevens, K. (1998) *Acoustic phonetics*. Cambridge, MA: MIT Press.

Tshering, K. and G. van Driem (2015) "The grammar of Dzonkha." ms.

van Driem, G. and K. Tshering (1998) *Dzongkha: Languages of the greater Himalayan region*. Leiden: Research CNWS, School of Asian, African, and Amerindian Studies, Leiden University.

Watters, S. A. (2002) "The sounds and tones of five Tibetan languages of the Himalayan region." *Linguistics of Tibeto-Burman Area* 25, 1–65.

Westbury, J. R. and P. Keating (1986) "On the naturalness of stop consonant voicing." *Journal of Linguistics* 22, 145–166.

Wright, R. (2004) "A review of perceptual cues and cue robustness." In B. Hayes, R. Kirchner and D. Steriade (eds.) *Phonetically based phonology*, 34–57. Cambridge: Cambrdige University Press.

(Received Aug. 30, 2017, Accepted Dec. 21, 2017)