Temporal aspects of talker variability in lexical tones

Kristine M. Yu, University of Massachusetts Amherst

Patterns of phonetic variation—whether they transcend individual phonetic/phonological units or not—are contingent on the phonetic parameters they are defined over. To elucidate how such patterns come into play in speech perception, it is thus critical to consider what parameters to use to characterize these patterns. Talker variability in the acoustic realization of lexical tones offers an informative case study in this respect. Most work on talker variability takes the methodological abstraction of defining patterns over one or two phonetic parameters measured at a single time point or interval, e.g. VOT, F1 measured at steady state. But investigating talker variability for tones forces us to confront the fact that patterns of variation occur over a signal unfolding in time: even phonemic level representations of tones, e.g. ˩˧, acknowledge this. Here, we use mathematical analysis of a multi-speaker tonal production corpus of Bole, Mandarin, Cantonese, and Hmong that we collected (Yu, 2011) to show that classic z-score and related models of talker variability are not supported. Instead, perceptual data suggests that talker variability may have effects on fine-grained detail in the shape of f0 contours, over a local temporal window extending over multiple syllables.

Rose (1987)'s classic work on the normalization of tonal f0 defined normalization as aiming "for a maximum reduction in between-speaker variance without sacrificing the desideratum of making perceptual sense". Because the percept of pitch is a function of more acoustic parameters than just f0, and since perceptual normalization was not well studied at the time, Rose (1987) considered only the acoustic criterion of reducing between-speaker variance in comparing z-score transforms versus transforms based on fraction of f0 range. Since then, both types of normalizations have remained ubiquitous in the tone literature in linguistic phonetics and automatic speech recognition, e.g. Shen (1994); Zhu (2004); Zhang et al. (2004); Levow (2006); Zhang and Liu (2011); Rose (2014); Li and Chen (2016). However, we show here that both mathematical analysis of f0 contours and recent evidence on temporal aspects of tone perception[1] do not support z-score and fraction of f0 range transforms as models of between speaker variation.

Following Liberman (2010), we consider possible mathematical models of between speaker variation in tonal f0 based on their algebraic properties, rather than comparing some measure of between-speaker variance as a result of applying different normalizations. We observe that both z-score transforms, e.g. (1), and fraction-of-range transforms, e.g. (2), are special cases of the general linear equations in (3), where $y_i(spkr)$ is the transformed f0 value at timepoint i for a speaker $spkr$, and a_{spkr} and b_{spkr} are speaker-specific constants.

$$y_i(spkr) = \frac{\mathrm{f0}_{i,spkr} - \overline{\mathrm{f0}}_{spkr}}{\sigma_{spkr}} \tag{1}$$

$$y_i(spkr) = \frac{\log(\mathrm{f0}_{i,spkr}) - \log(\mathrm{f0}_{min,spkr})}{\log(\mathrm{f0}_{max,spkr}) - \log(\mathrm{f0}_{min,spkr})} \times 4 + 1 \tag{2}$$

$$\begin{aligned} y_i(spkr) &= a_{spkr} \times \mathrm{f0}_{i,spkr} + b_{spkr} \\ y_i(spkr) &= a_{spkr} \times \log(\mathrm{f0}_{i,spkr}) + b_{spkr} \end{aligned} \tag{3}$$

If disparate tonal realizations across speakers are indeed related by a mathematical model of the form (3), then we expect a particular geometric signature in plots of scaling relations between speakers (Liberman, 2010): a linear relation for $\langle \mathrm{f0}_{i,spkr_m}, \mathrm{f0}_{i,spkr_n} \rangle$, a linear relation for $\langle \mathrm{f0}_{i,spkr_m} - \mathrm{f0}_{i,spkr_n}, \mathrm{f0}_{i,spkr_n} \rangle$, and an inverse $(1/x)$ relation for $\langle \mathrm{f0}_{i,spkr_m}/\mathrm{f0}_{i,spkr_n}, \mathrm{f0}_{i,spkr_n} \rangle$, where $spkr_m$ and $spkr_n$ are different speakers, and $\mathrm{f0}_{i,spkr}$ may be substituted with $\log \mathrm{f0}_{i,spkr}$. However, we show that this is not the case in our tonal production corpus.

[1] Recent work has also better elucidated the role of non-f0 acoustic parameters in tonal pitch perception (e.g. Garellek et al. (2013); Kuang (2013a,b); Yu and Lam (2014)), but we abstract away from those here.

Moreover, the speaker-specific parameters for z-score and fraction-of-range transforms are typically calculated over all data from a speaker in a corpus, while $f0_{i,spkr}$ is often extracted only from the current/target syllable or rime. But Wong and Diehl (2003); Huang and Holt (2009); Lee et al. (2009); Yu and Lam (2014) suggest that a more local preceding temporal window is enough. We also show in a perceptual experiment manipulating whether native Cantonese listeners heard only the target syllable, or the preceding and/or following syllable as well, that $f0_{i,spkr}$ in the following syllable is crucial for tonal identification. Acoustic analysis shows that peak delay pushes critical information about the slope of contrastive rising tones into the following syllable.

What might be an alternative to linear relations of the form (3) as models for talker variability in tonal f0? We suggest examining talker variation in terms of coefficient weights of basis functions for f0 curves, over running local windows including at least the preceding and following syllable. The rationale for this is that tonal contrasts can hinge on fine-grained differences in timing of falls in f0 (Remijsen, 2013; Remijsen and Ayoker, 2014), and perceptual and neurolinguistic work show that listeners are sensitive to the curvature of tonal f0 contours (Chandrasekaran et al., 2007; Krishnan et al., 2009; Barnes et al., 2012). By parameterizing f0 curves in terms of a set of basis functions–empirically (e.g. functional PCA) or pre-determined (e.g. orthogonal polynomials)–we can decompose talker variability in terms of its effects on the temporally detailed shape of f0 contours.

References

Barnes, Jonathan, Nanette Veilleux, Alejna Brugos, and Stefanie Shattuck-Hufnagel. 2012. Tonal center of gravity: a global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology* 3:337–383.

Chandrasekaran, Bharath, Ananthanarayan Krishnan, and Jackson T. Gandour. 2007. Experience-dependent neural plasticity is sensitive to shape of pitch contours. *NeuroReport* 18:1963–1967. URL `http://journals.lww.com/neuroreport/Abstract/2007/12030/Experience_dependent_neural_plasticity_is.17.aspx`.

Garellek, Marc, Patricia Keating, Christina M. Esposito, and Jody Kreiman. 2013. Voice quality and tone identification in white hmong. *The Journal of the Acoustical Society of America* 133:1078–1089.

Huang, Jingyuan, and Lori L. Holt. 2009. General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America* 125:3983–3994. URL `http://link.aip.org/link/?JAS/125/3983/1`.

Krishnan, Ananthanarayan, Jackson T. Gandour, Gavin M. Bidelman, and Jayaganesh Swaminathan. 2009. Experience-dependent neural representation of dynamic pitch in the brainstem. *NeuroReport* 20:408–413.

Kuang, Jianjing. 2013a. Phonation in tonal contrasts. Doctoral Dissertation, University of California Los Angeles, Los Angeles, CA.

Kuang, Jianjing. 2013b. The tonal space of contrastive five level tones. *Phonetica* 70:1–23.

Lee, Chao-Yang, Liang Tao, and Z.S. Bond. 2009. Speaker variability and context in the identification of fragmented mandarin tones by native and non-native listeners. *Journal of Phonetics* 37:1–15. URL `http://www.sciencedirect.com/science/article/B6WKT-4T5CGNX-1/2/a57c2905985d2a7909e1a9a8a6a4f006`.

Levow, Gina-Anne. 2006. Unsupervised and semi-supervised learning of tone and pitch accent. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, 224–231.

Li, Qian, and Yiya Chen. 2016. An acoustic study of contextual tonal variation in Tianjin Mandarin. *Journal of Phonetics* 54:123 – 150.

Liberman, Mark. 2010. COGS 501 – FDA homework # 2. URL `http://www.ling.upenn.edu/courses/Fall_2010/cogs501/FDA_HW2.html`.

Remijsen, Bert. 2013. Tonal alignment is contrastive in falling contours in Dinka. *Language* 89:297–327.

Remijsen, Bert, and Otto Gwado Ayoker. 2014. Contrastive tonal alignment in falling contours in shilluk. *Phonology* 31:435–462.

Rose, Phil. 1987. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication* 6:343–352.

Rose, Phil. 2014. Transcribing tone – a likelihood-based quantitative evaluation of Chao's tone letters. In *INTERSPEECH-2014*, 101–105.

Shen, Zhongwei. 1994. The tones in the Wujiang dialect. *Journal of Chinese Linguistics* 22:278–315.

Wong, Patrick C. M., and Randy L. Diehl. 2003. Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language & Hearing Research* 46:413–421. URL `http://jslhr.asha.org/cgi/content/abstract/46/2/413`.

Yu, Kristine M. 2011. The learnability of lexical tones from the speech signal. Doctoral Dissertation, University of California Los Angeles, Los Angeles, CA.

Yu, Kristine M., and Hiu Wai Lam. 2014. The role of creaky voice in Cantonese tonal perception. *Journal of the Acoustical Society of America* 136:1320–1333.

Zhang, Jie, and Jiang Liu. 2011. Tone sandhi and tonal coarticulation in Tianjin Chinese. *Phonetica* 68:161–191.

Zhang, Jin-Song, Satoshi Nakamura, and Keikichi Hirose. 2004. Tonal contextual f0 variations and anchoring based discrimination. In *SP-2004*, 525–528.

Zhu, Xiaonong. 2004. F0 normalization: How to deal with between-speaker tonal variations? *Linguistic Sciences* 2:3–19.