

## Session 1B Abstracts

# A Corpus Phonetic Study of Contemporary Persian Vowels in Casual Speech

Taylor Jones  
University of Pennsylvania

Modern Persian has six phonemic vowels, however in the last decades there has been disagreement in the literature as to their precise classification, both regarding their appropriate phonological classification and the relevance of historical distinctions of length. With regards to phonological classification, the low-back vowel is particularly controversial, with Lazard [1992], Miller [2013], and Toosarvandani [2004] *inter alia* claiming the low-back vowel is /ɒ/, and Ansarin [2004] and Aronow et al. [2017] arguing for /ɔ/, based on limited acoustic measurements. With regards to length, various linguists argue historical pairings of three short and three long vowels (/i:/ ~ /e/, /u:/ ~ /o/, and /æ/ ~ /ɒ:/) are still phonologically relevant, either outright claiming historical length distinctions still obtain [Lazard, 1992], or that such distinctions inform phonological processes like vowel assimilation [Rahbar, 2009, Toosarvandani, 2004]. The present study aims to settle both debates, through the use of a new forced-aligner, and automatic alignment and vowel extraction of over 40 hours of casual, telephone speech.

This study makes use of the CALLFRIEND FARSI corpus [Canavan and Zipperlen, 1996], a telephone corpus of casual speech among native speakers of Modern Iranian Persian, comprising 100 recordings. The corpus was used to train an HTK-based forced aligner [Young et al., 2002], using the McGill Prosody Lab wrapper [Gorman et al., 2011]. While previous studies have extremely small sample sizes when they use empirical phonetic data at all (e.g., Aronow et al. [2017] uses 90 total vowels from 2 speakers, one male, one female), the present study evaluates 70,711 vowels from 104 speakers.

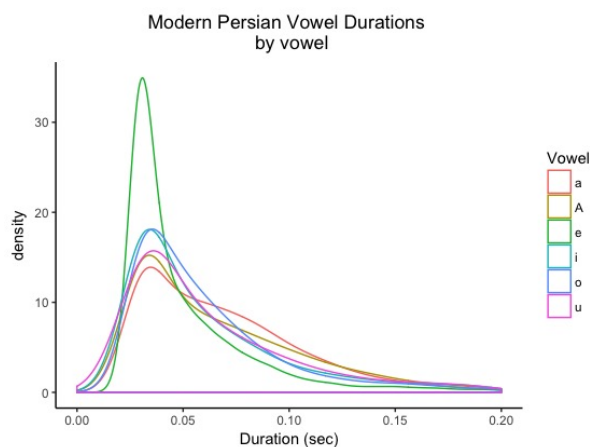


Figure 1: Vowel Durations

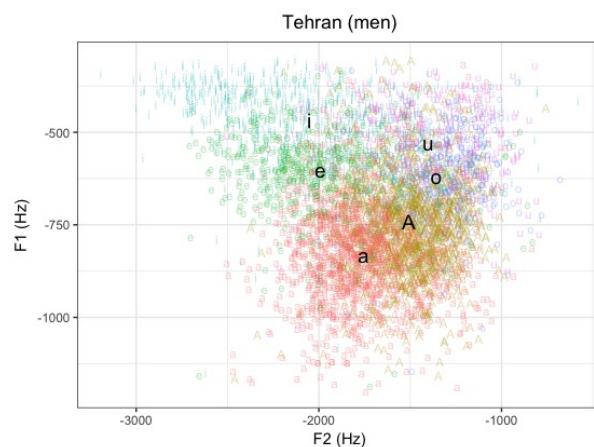
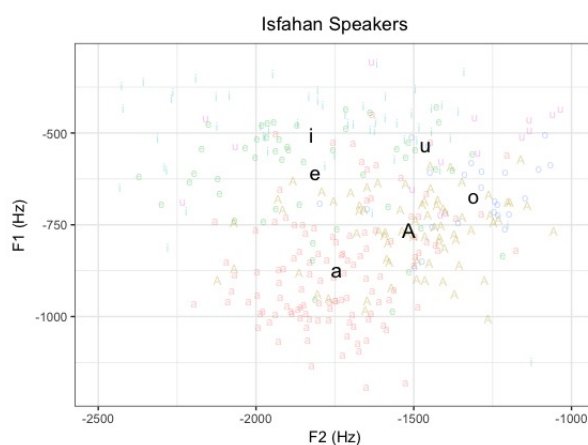


Figure 2: Tehrani Vowel Space

It was found that there is no empirical phonetic support for the claim that vowel length is distinctive in Modern Persian, with all vowels showing similar distributions, centered

around approximately 30 milliseconds in duration. Furthermore, the empirical evidence suggests that the low back vowel in Modern Persian is not the traditionally claimed /ɒ/, but may be better characterized as /ɔ/. There is some evidence for fronting of /u/ for speakers from Tehran, but not to the extent claimed by Aronow et al. [2017]. There is also strong evidence of regional variation, with distinct vowel spaces for speakers from, e.g., Isfahan, suggesting the focus common in the literature on speakers from Tehran may be limiting our understanding. There is also evidence of socially conditioned variation, with age, education, and gender affecting both low-back vowel raising, and high back vowel fronting.

The results of this study have implications for both our understanding of sociolinguistic variation in Modern Persian and for phonological analysis of Modern Persian, especially insofar as traditional phonological analyses are predicated on the assumption that Persian has two low vowels.



## References

- Ali Akbar Ansarin. An Acoustic Analysis of Modern Persian Vowels. In *9th Conference Speech and Computer*, 2004.
- Robin Aronow, Brian D McHugh, and Tessa Molnar. A pilot acoustic study of modern persian vowels in colloquial speech. *Proceedings of the Linguistic Society of America*, 2:17–1, 2017.
- Alexandra Canavan and George Zipperlen. *CALLFRIEND Farsi LDC96S50*. Linguistic Data Consortium, 1996.
- Kyle Gorman, Jonathan Howell, and Michael Wagner. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193, 2011.
- Gilbert Lazard. *A Grammar of Contemporary Persian*. Mazda Publishers, 1992.
- Corey Miller. Variation in persian vowel systems. *Orientalia Suecana*, 61:156–169, 2013.
- Elham Rohany Rahbar. On Contrasts in the Persian Vowel System. *Toronto Working Papers in Linguistics*, 31, 2009.
- Maziar Doustdar Toosarvandani. Vowel Length in Modern Farsi. *Journal of the Royal Asiatic Society*, 14(03):241–251, 2004.
- Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3:175, 2002.

## **Modified two-component Shepard tones and their application to Sine Wave Speech**

Jon Nissenbaum, Brooklyn College and the Graduate Center, CUNY

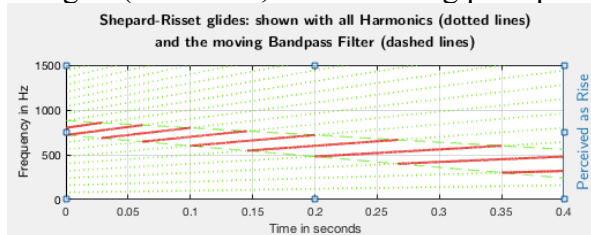
Sine wave speech (SWS), which consists only of several frequency- and amplitude-modulated sinusoids representing vocal tract formants, can elicit perception of words and sentences despite its sparse acoustic structure [1]. For this reason SWS has proven extremely useful as a tool for investigating the perceptual primitives of the segmental content and other aspects of speech. However, SWS contains no information relevant for pitch perception, making it unsuitable for investigating prosody [2, 3] or tone languages [4-6].

This talk describes a new method for creating SWS, modified to add a minimal but powerful cue for pitch, thereby expanding the range of perceptual phenomena that SWS is capable of probing so as to include tone and prosody. In order to achieve this, we discarded the lowest sinusoid of the traditional SWS replica (representing the first vocal tract formant, F1), and replaced it with a “modified Shepard-Risset tone”: a two-component tone glide formed from a bandpass whose center frequency tracks F1. The bandpass was wide enough at any timepoint for exactly two harmonics of an independently specified fundamental frequency ( $f_0$ ) contour. In its most general form, a Shepard-Risset tone can be characterized as a complex pitch glide whose individual harmonic components pass through a particular frequency region. Shepard's original experiment [7] used sequences of discrete tones composed of harmonics spaced at octave intervals within a wide frequency range; at each successive tone, as the harmonics moved up (or down), they passed in and out of a bell-shaped spectral envelope, creating the effect of an endlessly rising or falling scale. Risset [8] demonstrated that the same effect could be achieved by scrolling the harmonics through the frequency region, forming a glide instead of a discrete scale.

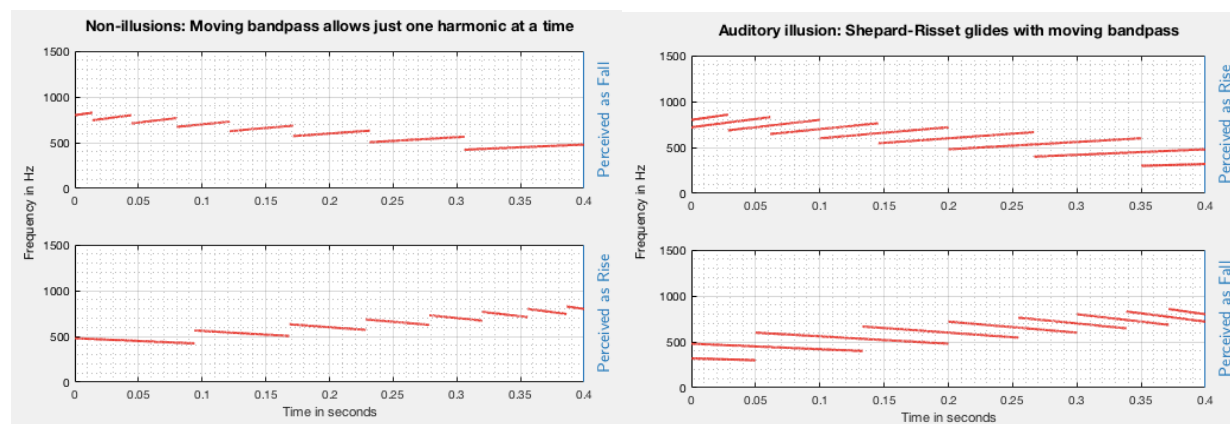
Our method for creating pitch-enhanced SWS representations takes advantage of this perceptual effect but makes two important modifications to the Shepard-Risset paradigm: (1) Instead of using an invariant amplitude envelope to filter the harmonics, we allowed our bandpass filter's center frequency to vary over time, tracking the frequency of F1 (and, moreover, our bandpass had a rectangular rather than a bell-shaped envelope so that there was no attenuation of the harmonics as they passed in and out of its range). (2) Instead of using octave spacing of harmonics (which, in the Shepard-Risset illusion, creates ambiguity as to the height of the fundamental frequency), we used consecutive harmonics of a well-defined and unambiguous fundamental frequency contour.

Crucially, the  $f_0$  was never included as one of the two frequency components of the complex tones that we synthesized. Consequently the perception of pitch relied on the “missing fundamental” effect [9] in a rather extreme way: listeners recovered the  $f_0$  solely on the basis of an adjacent pair of higher harmonics (ranging between the 2nd and 10th). It has been shown previously [10, 11] that tones consisting of 2–3 harmonics (and even, under special circumstances, just one higher harmonic [12]) can elicit perception of pitch despite the absence of any  $f_0$  component. The prior studies using fewer than three harmonics found that masking the complex tones with noise was required for pitch perception. The present study used no masking; the use of pitch glides in conjunction with a moving bandpass filter turned out to be sufficient to induce simultaneous impression of harmonic direction and formant direction.

An initial experiment demonstrated that a bandpass whose center frequency fell from 800-400 Hz over 400ms, when intersected with harmonics of a rising  $f_0$  (80-160Hz, elicited strong perception of a rising contour, and vice versa (*see fig. right*). Strikingly, the pitch at the end of this complex “rise” was perceived as the peak despite the fact that the actual frequency components were lower than at any previous point in the stimulus.



Complex, two-component tones constructed in this manner contrasted sharply with single-component glides formed by passing the same set of harmonics through a narrower bandpass. Glides consisting of just one component at any given time were strongly perceived as having a pitch direction determined by the *bandpass*, not by the direction of the harmonics (*see fig. 2 below*).



A second experiment replaced F1 in a set of SWS replicas with these two-component tones and elicited robust perception of contrasting pitch contours. Two pairs of stimuli were used: modified and unmodified SWS replicas of the sentences “I ate it raw” and “I ate it now.” The final words in each of these sentences were synthesized using F1 transitions identical to those shown in the tone glides (*fig. 2, right side*): the replica of the word “Raw” includes a 400 ms span during which F1 rises from 400 to 800 Hz, while “Now” includes a 400 ms span in which F1 falls from 800 to 400 Hz. With unmodified SWS, listeners tend to perceive (unnatural) pitch-like contours that are determined by the frequency contour of the first formant [2, 3]. Thus “Raw” is perceived as rising in pitch while “Now” is perceived as falling. In our modified SWS replicas, by contrast, we induced perception of pitch on each of these words going in the opposite direction, with a fall on “Raw” and a rise on “Now.”

After demonstrating the pitch illusion and describing the method of creating the complex tones, we will conclude by discussing future directions and broader significance of pitch-enhanced SWS.

**References:** [1] R Remez et al. (1981), *Science* 212.4497: 947-950. [2] R Remez & P Rubin (1984), *Perception Psychophys* 35.5: 429-440. [3] R Remez & P Rubin (1993), *JASA* 94. [4] S Rosen & SNC Hui (2015), *JASA* 138.6: 3698. [5] YM Feng et al. (2012), *JASA* 131.2: EL133. [6] Y Han & F Chen (2017), *JASA* 141.6: EL495. [7] RN Shepard (1964), *JASA* 36.12: 2346. [8] JC Risset (1971), 7th Intl Congress on Acoustics, Budapest. [9] JF Schouten (1938), *Proc. Koninklijke Nederlandse Akademie van Wetenschappen* 41: 1086–93. [10] RJ Ritsma (1962), *JASA* 34.9: 1224. [11] J Hall & R Peters (1981), *JASA* 69.2: 509. [12] A Houtsma & J Goldstein (1972), *JASA* 51.2: 520. [13] GF Smoorenburg (1970), *JASA* 48.4: 924. [14] T Houtgast (1976), *JASA* 60.2: 405.

An investigation of the articulatory correlates of vowel anteriority  
in Kazakh, Kyrgyz, and Turkish using ultrasound tongue imaging  
Jonathan North Washington  
Swarthmore College

This study uses ultrasound tongue imaging (Stone, 2005)—or UTI—to examine the articulatory correlates of the vowel anteriority contrast in three Turkic languages: Kazakh, Kyrgyz, and Turkish. It is demonstrated that each of these languages exhibits a single anteriority contrast, but that it is implemented differently in Turkish (primarily through the position of the tongue body) than in Kazakh and Kyrgyz (through a correlated position of tongue body and tongue root), despite greater acoustic similarity between the vowel systems of Turkish and Kyrgyz.

It has long been understood that the anterior and posterior distinction present in the vowel systems of many languages is implemented during articulation by the front-back position of the tongue body (Ladefoged & Maddieson, 1996). A second anteriority distinction has also been documented for a number of languages of Africa—namely, that of tongue root position (Lindau, 1978; Stewart, 1967). In languages employing this latter distinction, any given vowel is both tongue-body front or back *and* tongue-root advanced or retracted—i.e., both anteriority contrasts are used independently of one another.

Based on similar acoustic properties and accompanying phonetic and phonological patterns, sources like Ard (1983), Rialland & Djamouri (1984), and Svantesson (1985) began to demonstrate that the vowel systems of various Tungusic and Mongolic languages of northern Asia are very similar to the “dual-anteriority” systems documented in languages of Africa. Vajda (1994) has further argued that the single anteriority contrast in the vowel system of Kazakh is one of tongue root position and not tongue body position. This is the only known claim of a tongue-root-only vowel anteriority system, and the only claim based on phonetic data for a tongue-root contrast in a Turkic language. Vaux (2009) hypothesises on phonological grounds that most Turkic languages have a phonologised redundancy between tongue-root and tongue-body posteriority.

The present study is the first to use articulatory data (in the form of UTI) to investigate which region(s) of the tongue is/are involved in the anteriority contrast in Turkic languages. Kazakh is examined with Vajda’s (1994) claims in mind; in addition, two other Turkic languages are examined: Kyrgyz, a close relative of Kazakh with a notably different vowel system, and Turkish, a more distant relative of the two which has received considerably more attention in the linguistics literature.

Native speakers of each language were recorded (12 total participants to date) reading similarly structured carrier phrases containing words with a balance of consonant contexts for each vowel and a range of syllable structures and number of syllables. Short vowels in open initial syllables of multi-syllabic words were measured for all speakers to avoid effects of vowel harmony, prosodic position, syllable structure, and phonemic vowel length contrasts. The imaged tongue surface at the midpoint of each monophthong was hand-traced, and the first and second formants at the same time index were measured.

To impressionistically understand the data, averaged traces (with standard deviation bands) for each vowel type were plotted. Additionally, a measure of tongue region differentiation was developed to understand the role that the position of different areas of the tongue plays in the anteriority contrast of each language. This measure is a calculation of the ratio of the number of degrees separating the region of most positive and most negative difference in tongue position during the articulation of anterior and posterior vowels from the point at which the tongue positions for these two categories of vowels overlaps. It has the potential to be speaker-agnostic, allowing for generalisations at the level of the linguistic variety. Both types of plot are shown in figure 1 for a speaker of Turkish and a speaker of Kyrgyz.

These ratios are found to be consistently around 1.0 for Turkish and 2.0 for Kazakh and Kyrgyz. Together with observations about the averaged traces, this leads to the conclusion that in

Turkish, the anterior and posterior vowels are contrasted using primarily the tongue body—much like other languages with a single anteriority contrast—while Kazakh and Kyrgyz contrast anterior and posterior vowels using the positions of the tongue body and the tongue root combined. In other words, Kazakh and Kyrgyz exhibit an anteriority contrast where tongue body position and tongue root position are coupled, as predicted by Vaux (2009), while Turkish does not. These findings suggest that the tongue root may not always simply be “along for the ride” in vowel systems where the tongue body and tongue root positions are not involved in separate contrasts.

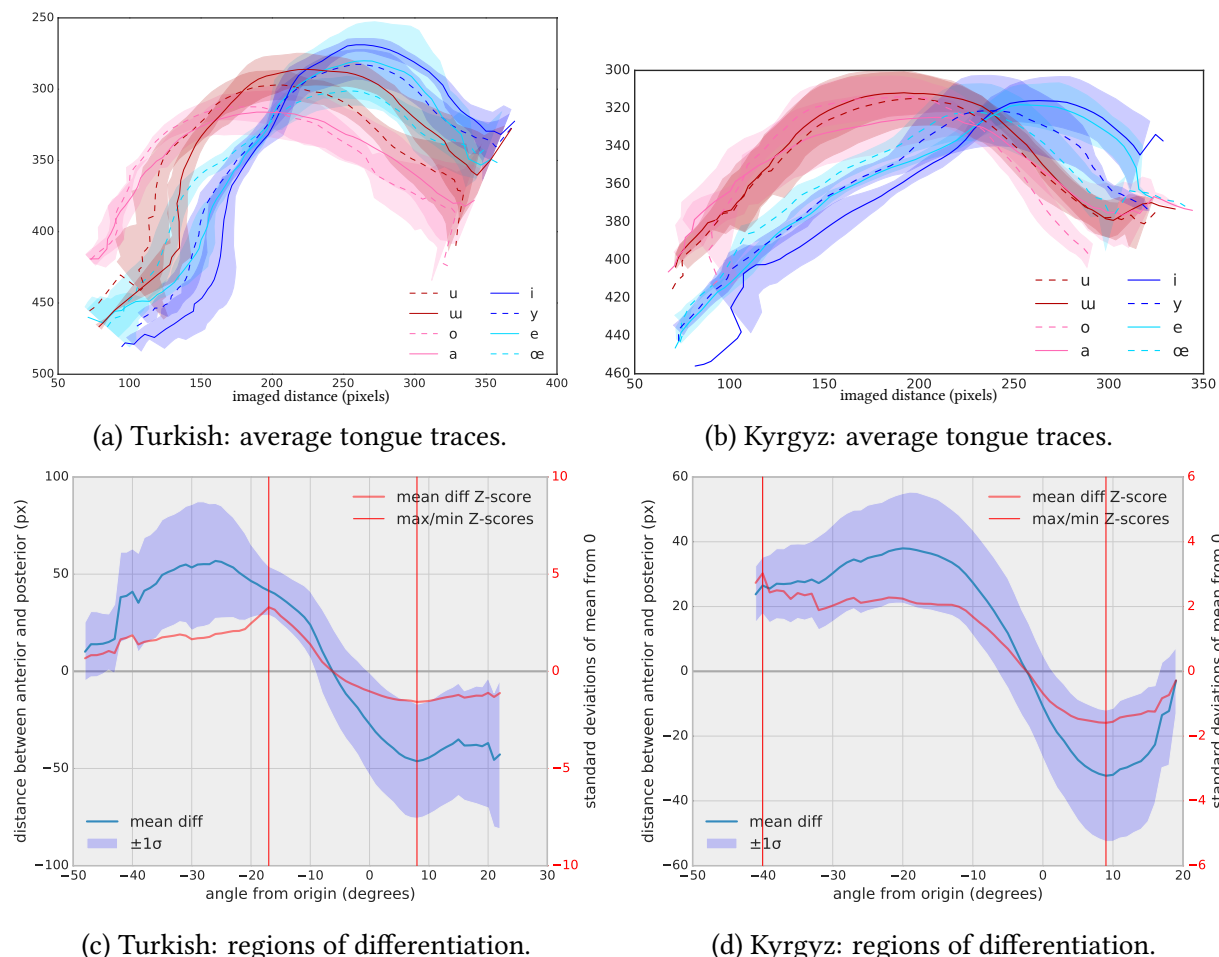


Figure 1: Plots of tongue traces (with standard deviation bands) for vowel categories of Turkish and Kyrgyz (one speaker each), and plots presenting the regions of maximum differentiation in tongue shape between anterior and posterior vowels. Anteriority increases from left to right.

- Ard, Josh (1983). “A sketch of vowel harmony in the Tungus languages”. In: *Paper in Linguistics* 16.3-4, pp. 23–43. doi: [10.1080/08351818309370594](https://doi.org/10.1080/08351818309370594).  
 Ladefoged, Peter & Ian Maddieson (1996). *The Sounds of the World's Languages*. Oxford: Blackwell.  
 Lindau, Mona (Sept. 1978). “Vowel Features”. In: *Language* 54.3, pp. 541–563.  
 Rialland, Annie & Redouane Djamouri (1984). “Harmonie vocalique, consonantique et structures de dépendance dans le mot en mongol khalkha”. In: *Bulletin de la Société de Linguistique de Paris* 79, pp. 333–383.  
 Stewart, John M. (1967). “Tongue Root Position in Akan Vowel Harmony”. In: *Phonetica* 16.4, pp. 185–204.  
 Stone, Maureen (2005). “A guide to analysing tongue motion from ultrasound images”. In: *Clinical Linguistics & Phonetics* 19.6-7, pp. 455–501. doi: [10.1080/02699200500113558](https://doi.org/10.1080/02699200500113558).  
 Svantesson, Jan-Olof (1985). “Vowel Harmony Shift in Mongolian”. In: *Lingua* 67, pp. 283–327. doi: [10.1016/0024-3841\(85\)90002-6](https://doi.org/10.1016/0024-3841(85)90002-6).  
 Vajda, Edward J. (1994). “Kazakh Phonology”. In: *Studies on East Asia*. Ed. by Edward H. Kaplan & Donald W. Whisenhunt. Vol. 19: “Opuscula Altaica: Essays Presented in Honor of Henry Schwarz”. Western Washington University, pp. 603–650.  
 Vaux, Bert (2009). “[atr] and [back] Harmony in the Altaic languages”. In: *Investigations into Formal Altaic Linguistics: Proceedings of WAFL3*. Ed. by Sergei Tatevosov, pp. 50–67.

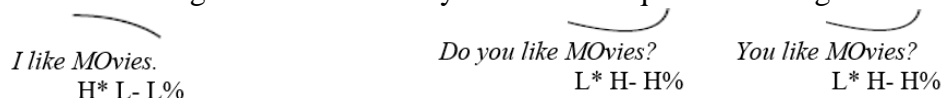
# The influence of pitch contour on Mandarin speakers' perception of English stress

Yaobin Liu

Stony Brook University

**BACKGROUND:** Previous studies on L2 stress perception have mostly focused on words in isolation or in invariable intonational contexts (see Archibald 1993, 1997, Wang 2008, Peperkamp et al 2010, etc.). This paper reports on a study exploring the influence of different intonation contours, i.e. falling (declarative) and rising (yes/no question), on nonnative speakers' stress perception. The acoustic correlates of English lexical stress include duration, intensity, pitch (F0) and vowel quality, all of which serve as active cues in perceiving stress, despite the dynamics of their relative weighting (Fry 1955, 1958, 1965, Lieberman 1960, Lehiste 1970). As a result, pitch and stress may not interact in an unambiguous way. When intonation contour is imposed on stress contour, the stressed syllable may not necessarily have the highest pitch in a sentence as in citation forms. In citation forms and declarative sentences (intonation: H\* L-L%), the nuclear pitch accent bears a high tone. However, at the end of a yes/no question (YNQ) (intonation: L\* H-H%), the nuclear pitch accent bears a low tone with high pitch on the following phrase accent and boundary tone (Ladefoged & Johnson 2011), as exemplified in (1).

(1) a. declarative: falling contour                      b. yes/no & echo question: rising contour

  
I like MOVies.                      Do you like MOVies?                      You like MOVies?  
H\* L-L%                      L\* H-H%                      L\* H-H%

Mandarin, as a tone language, utilizes pitch differently where it signals lexical contrasts. Empirical studies on both production and perception have found F0 to be the primary acoustic correlate of Mandarin tone (Howie 1970, 1976, Chuang et al 1972, Massaro et al 1985, Jongman et al 2006). Presumably, the phonemic nature of tone in Mandarin would render its speakers more sensitivity to pitch than to other cues when they are exposed to English lexical stress. This conjecture was borne out in Wang's (2008) study where only F0 was found to have "a decisive effect on the stress judgments by Chinese learners of English", unlike native speakers. Archibald's (1997) longitudinal study also suggested Chinese learners paid attention only to pitch, which was carried over as part of the lexical entry from L1 to L2. Drawing on these studies, we should predict variability of their stress judgments in different intonation contexts, as intonation is primarily a function of pitch contour in English.

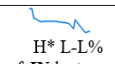

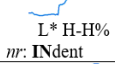

**RESEARCH QUESTION:** Accordingly, it was hypothesized that Mandarin speakers would be more subject to the influence of pitch contour in stress perception than English speakers; specifically, they would misperceive stress when the stressed syllable bears a low tone, as under the rising pitch contour, while native speakers are less likely to do so for they rely on multiple cues. The hypothesis could be tested by answering the questions: i) Do Mandarin speakers misperceive stress in a rising pitch contour significantly more than in a falling pitch contour? Do English speakers display a similar pattern? ii) Do Mandarin speakers perceive stress in a rising pitch contour significantly differently than English speakers? Does the falling pitch contour display a similar pattern?

**EXPERIMENT:** As a commonly adopted method (Lieberman 1960, Smith 2016), 12 minimal pairs of noun-verb alternation were used, where the nouns uniformly had initial stress and the verbs final stress, e.g. INdent (n)-inDENT (v). These items were placed in two kinds of intonation contour, namely falling (declarative) and rising (echo question), both of which used the same syntactic template "This word is \_\_\_\_./?". The two stress



patterns, indexed by part of speech (POS) (noun  $\equiv$  initial stress, verb  $\equiv$  final stress), together with the two intonation contexts form a matrix of 4 experimental conditions: noun-in-falling (*nf*), verb-in-falling (*vf*), noun-in-rising (*nr*), verb-in-rising (*vr*). The target materials, interspersed with an equal number of fillers, were randomly and evenly distributed among participants using a Latin square design. 38 Mandarin learners of English and 15 native American English speakers listened to the audio stimuli and identified the POS of the last word of each sentence by making a forced choice between “noun” and “verb” merely based on its stress pattern they heard. The experiment was conducted via Qualtrics.

**RESULTS & DISCUSSION:** 636 tokens were collected for the experimental items and the accuracy rate was calculated for each of the four conditions. The summary of the perceptual results, in contrast with actual stress patterns (illustrated with “indent”), is shown in the table below; the statistical strength of these results was confirmed with a mixed effects logistic regression model.

Condition	Native language	Predominantly perceived stress position
 H* L-L% <i>nf</i> : INdent	English	INdent (78%)
	Mandarin	INdent (86%)
 H* L-L% <i>vf</i> : inDENT	English	inDENT (93%)
	Mandarin	inDENT (84%)
 L* H-H% <i>nr</i> : INdent	English	INdent (58%)
	Mandarin	inDENT (61%)
 L* H-H% <i>vr</i> : inDENT	English	inDENT (93%)
	Mandarin	inDENT (81%)

First, in the falling contour, Mandarin speakers matched with native speakers (85% vs. 86%), setting up the baseline that the Mandarin group was capable of perceiving stress in a nativelike fashion under default intonational circumstances. Second, there was a main effect of pitch contour ( $p < 0.001$ ), suggesting that both

Mandarin speakers and English speakers were influenced by pitch contour in stress perception, with lower accuracy rates (60% vs. 76%) in rising pitch conditions in general. To zoom in, the significance of the effect was found only in initially-stressed words, with the difference of 47% for Mandarin speakers and 20% for English speakers between the falling contour and the rising contour, but not in finally-stressed ones (4% vs. 0%). Third, although the English group was also influenced by pitch contour, which was unexpected, there was a significant interaction between native language and pitch contour for words with initial stress ( $p < 0.05$ ), suggesting that the effect of pitch contour was explicitly stronger for Mandarin learners than native speakers at least when they were perceiving initial stress. This contrast is corroborated by the uniquely problematic status of the *nr* condition against the other three conditions. Under this condition, stressed syllables take on a low pitch while unstressed syllables take on a high pitch, as opposed to other conditions where stressed syllables and unstressed syllables are generally aligned with a high pitch and a low pitch respectively. For native speakers, the effect of this “misalignment” in *nr* can be offset possibly by increased duration, or increased intensity, or simply the pitch change itself, or all of them, but can exert itself in Mandarin learners’ interlanguage system and polarize their stress perception, given their susceptibility to tone in native grammar.

**SELECTED REFERENCES:** Archibald, J. 1997. The acquisition of English stress by speakers of nonaccentual languages: Lexical storage versus computation of stress. *Linguistics*. Fry, D. B. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27, 765-768. Jongman, Allard, Wang, Yue, and Sereno, Joan. 2006. Perception and Production of Mandarin Tone. Ladefoged, P., & Johnson, K. 2011. *A course in phonetics* (6th ed.). Boston, MA: Wadsworth. Wang, Q. 2008. Perception of English stress by Mandarin Chinese learners of English: An acoustic study. Ph.D. dissertation, University of Victoria.