The difference between shrieks and shrugs: Spectral envelope correlates with changes in pitch and loudness

K. Jake Patten[1†] and Michael K. McBeath[2]
[1] College of Health Solutions, Arizona State University, Tempe, Arizona, USA
[2] Department of Psychology, Arizona State University, Tempe, Arizona, USA
[†] Corresponding author: kjp@asu.edu

Introduction

Stimulus intensity can impact the perception of pitch. Stevens (1935) and, later, Gulick (1971) demonstrated that, for non-dynamic tones, pitch and loudness are positively correlated for frequencies above 2 kHz but become negatively correlated below 250 Hz. More ecologically valid, dynamic sounds, however, function differently. Pitch and loudness are strongly correlated throughout the range of audible frequencies. Participants experience an illusory increase in pitch as auditory objects draw near and a similar illusory decrease in pitch as objects recede (Neuhoff & McBeath, 1996). The experienced change can be as much as 8 semitones when the physical change amounts to only 2; a full half-octave greater than the actual change (McBeath & Neuhoff, 2002). This bias may have arisen because natural sounds (up to roughly 1 kHz), musical instruments, and human and animal vocalizations naturally exhibit simultaneous increases and decreases in f_0 and intensity (Johnston, 2009; McBeath, 2014; Wang, Astfalck, & Lai, 2002). In particular, when moving from whispering to normal speech to shouting over noise to yelling in anger, frequency consistently rose along with intensity (Scharine & McBeath, 2018). Past research has also shown that shouting is often correlated with a consistent shallowing of the spectral envelope and a reduction in harmonicity (Raitio et al., 2013; Wallmark & Allen, 2020), as well as being identified as one of two primary organizational non-dynamic dimensions of timbre (Patten, McBeath, & Baxter, 2018). The current study investigates the existence of a reliable correlation and natural regularity between spectral envelope, fundamental frequency, and intensity.

To test this, the current study uses North American vowel phonemes, which – when controlled for fundamental frequency and intensity – are changes in timbre that are well-known to all participants. One well-known phenomenon pertaining to vowel sounds is intrinsic fundamental frequency (If_0); the finding that some North American vowels are often voiced higher than others (Crandell, 1925). This production bias may be partly explained by anatomical constraints, as some researchers have found evidence of this bias persisting across cultures (Whalen & Levitt, 1994). Furthermore, high vowels are far more susceptible to f_0 changes when a speaker's larynx is raised than low vowels (Sundberg & Nordstrom, 1976). However, other researchers find evidence of If_0 in some registers of tonal languages and not others or of If_0 not existing at all (Connell, 2002; Zee, 1980). It is possible that If_0 is not solely determined by biophysical constraints, but by the perception of the pitch of vowels and other natural sounds.

Method

Participants in all experiments were undergraduate students enrolled in introductory psychology and speech and hearing science classes at Arizona State University. The average age in all three experiments ranged from 19.4 to 22.6 years.

Experiment 1A. Participants listened to 10 North American monophthongs as well as the steady [a] and [o] portions of [aI] and [ou] spoken by a professional voice actor and recorded in a B_T environment. All vowels were digitally altered to hold f_0 and intensity constant. Participants were first asked to rate the similarity (1 – 10, latter being most similar) of all paired vowels. These ratings were used to derive the multidimensional scaling (MDS) solution in Figure 1. Next, participants were asked to order the vowels from highest to lowest pitch, allowing for a correlation to be computed between the MDS solution and tone height.

Experiment 1B. The MDS solution from 1A informed the choice of phoneme in this and subsequent experiments. Participants were asked to move a mouse on their screen to indicate perceived changes in pitch and loudness (all participants indicated both in two separate, counterbalanced blocks) for vowels that changed in timbre ([i] to [ʌ]), f_0, and/or intensity.

Experiment 2. This experiment consisted of an analysis of different song types to understand how the perceptual and production bias of the pitch of [i] and [ʌ] is manifest in different song types. Scat, a vocalization style where the mouth is used to replicate instruments, was hypothesized to exhibit the largest production bias as the singer would use all techniques to increase their vocal range. Conversational interviews from podcasts were hypothesized to exhibit a small, though significant difference. Finally, lyrical songs – with constraints at the word, phrase, or musical score level – were hypothesized to show no difference. The f_0 of the first 30 instances of both [i] and [ʌ] were observed for six instances of each song type.

Experiment 3. To further demonstrate the bias in using [i] to produce high f_0 sounds, participants were tasked with recreating sounds, both high and low, outside their vocal range (60 Hz and 8 kHz sine waves). Participants' were unconstrained in the vocalizations they could make, though all used vowels.

Results

Experiment 1A. Participants' ratings were used to construct the MDS solution in Figure 1. The *y*-dimension of the solution correlates with rated pitch (though all phonemes were presented at a constant f_0 and intensity), r (10) = .90, $p < .001$. This is greater than the correlation of the *y*-dimension – ostensibly, tone height or experienced pitch – and the second formant, r (10) = .76, $p < .05$. The *x*-dimension of the solution correlates with harmonicity, r (10) = .60, $p < .05$.

Experiment 1B. Psychophysical functions derived from the results reveal that moving from [i] to [ʌ] is equivalent to a .38 semitone decrease in f_0 and a .75 dB decrease in intensity.

Experiment 2. A repeated measures ANOVA reveals an overall mean difference of fundamental frequency between [ʌ] (M = 209.95 Hz), [I] (M = 250.90 Hz), and [i] (M = 261.45 Hz) phonemes [main effect of phoneme, F (2, 438) = 27.02, $p < .001$, $\eta^2 = .04$] and between interviews (M = 147.53 Hz), songs (M = 251.48 Hz), and scat (M = 323.29 Hz) [main effect of type of recording, F (2, 438) = 90.95, $p < .001$, $\eta^2 = .42$]. Importantly, all three hypotheses regarding the extent of the production bias across song type were borne out by a significant interaction between phoneme and recording type, F (4, 438) = 17.15, $p < .001$, $\eta^2 = .05$. This can be seen in Figure 2.

Experiment 3. The [i] phoneme is chosen much more often to produce a high-pitched sound than other phonemes and the [ʌ] phoneme is selected more often for replicating low-pitched sounds, χ^2 (5, 68) = 64.57, $p < .001$, Cramer's V = .44. This is shown in Figure 3.

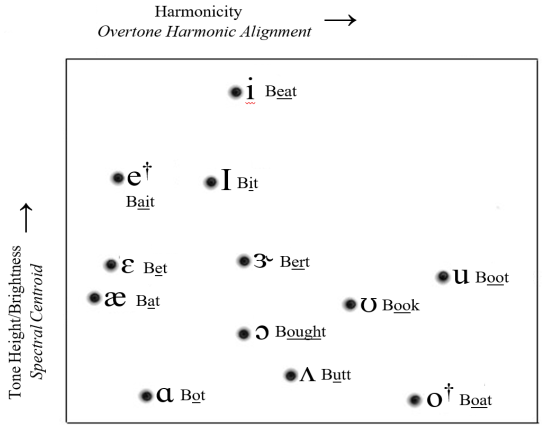

Figure 1. The MDS solution from Experiment 1A. Axes are labeled by the timbre dimension they represent and how that dimension was calculated.

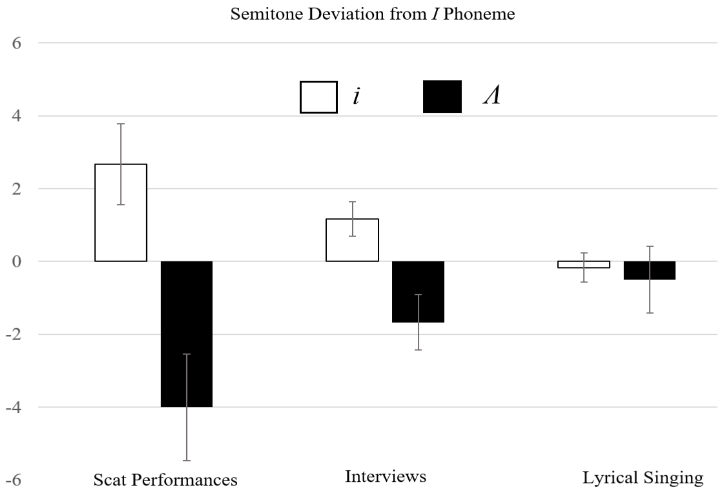

Figure 2: Semitone deviations from [I] for [i] and [ʌ]for scat songs, interviews, and lyrical singing. There was a significant difference for scat songs and natural conversation, though the latter exhibited the bias to a lesser degree. There was no difference for lyrical singing.

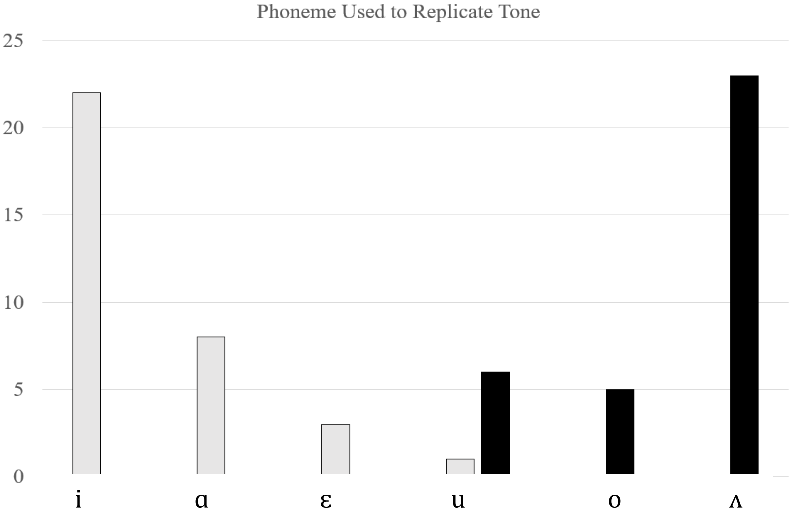

Figure 3. Phonemes used to replicate high (8 kHz) and low (60 Hz) tones in Experiment 3.

Discussion

Experiment 1A illustrated the way vowel phonemes are organized in cognitive space when f_0 and intensity are held constant. This experiment also demonstrated that the two most salient dimensions of vowel phoneme organization are tone height and harmonicity. The extreme vowels of [i] and [ʌ] informed the construction of Experiment 1B, which quantified the extent to which timbre impacts perceptions of frequency and intensity. This finding also suggests the existence of a natural regularity between spectral envelope, f_0, and intensity. Experiment 2 tested the prevalence of this natural regularity in a condition where singers would want to exploit it (scat singing), in a natural setting (interviews), and a condition where the regularity would be suppressed (lyrical singing). Finally, Experiment 3 pushes the findings of Experiment

2 further by tasking non-singers with replicating sounds outside their range. On average, participants used the [i] phoneme to replicate a very high tone and the [ʌ] phoneme to replicate a very low sound. Overall, these three experiments demonstrate a high degree of correlation between spectral envelope, f_0, and intensity. This correlation can be used to enhance the function of synthetic speech, voice recognition, hearing aids, and communication compression algorithms. These findings also support the use of the natural regularities framework to investigate timbre through illusions and discover new truths about perception and hearing.

References

Connell, B. A. (2002). Tone languages and the universality of intrinsic f_0: Evidence from Africa. *Journal of Phonetics*, *30*, 101-129.

Crandall, I. B. (1925). The sounds of speech. *The Bell System Technical Journal, 4*(4), 586-639.

Gulick, W. L. (1971). *Hearing: Physiology and Psychophysics*. New York: Wiley.

Johnston, I. (2009). *Measured Tones. Boca Raton*, FL: CRC Press.

McBeath, M. K. (2014). The Fundamental Illusion. *Paper presented at the 55th annual meeting of the Psychonomic Society*, Long Beach, California.

McBeath, M. K., & Neuhoff, J. G. (2002). The Doppler effect is not what you think it is: Dramatic pitch change due to dynamic intensithy change. *Psychonomic Bulletin and Review, 9*(2), 306-313.

Neuhoff, J. G., & McBeath, M. K. (1996). The Doppler illusion: The influence of dynamic intensity change on perceived pitch. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 970-985.

Patten, K. J., McBeath, M. K., Baxter, L. C. (2018). Harmonicity: Behavioral and neural evidence for functionality in auditory scene analysis. *Auditory Perception and Cognition, 1*(3-4), 150-172.

Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., & Alku, P. (2013). Analysis and synthesis of shouted speech. *INTERSPEECH*, 1544-1548.

Scharine, A. A. & McBeath, M. K. (2018). Natural regularity of correlated acoustic frequency and intensity in music and speech: Auditory scene analysis mechanisms account for integrality of pitch and loudness. *Auditory Perception & Cognition*, *1*(3-4), 205-228.

Stevens, S. S. (1935). The relation of pitch to intensity. *Journal of the Acoustical Society of America, 6*(3), 150-154.

Sundberg, J. & Nordström, P-E. (1976). Raised and lowered larynx – The effect on vowel formant frequencies. *STL-QPSR, 17*(2-3), 035-039.

Wallmark, Z. & Allen, S. E. (2020). Preschoolers‘crossmodal mappings of timbre. *Attention, Perception, and Psychophysics, 82*, 2230-2236.

Wang, C., Astfalck, A., & Lai, J. C. S. (2002). Sound power radiated from an inverter-driven induction motor: Experimental investigation. *IEE Practical Electrical Power Applications, 149*(1), 46-52.

Whalen, D. H. & Levitt, A. G. (1994). The universality of intrinsic f_0 of vowels. *Haskins Laboratories Status Report on Speech Reseach SR-117/118*, 1-14.

Zee, E. (1980). Tone and vowel quality. *Journal of Phonetics, 8*(3), 247-258.