UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Neural basis of individual differences in tonal effects on perceived duration

Permalink
https://escholarship.org/uc/item/9156c5st

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Tung, Tzu-Yun
Yu, Alan

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Neural basis of individual differences in tonal effects on perceived duration

Tzu-Yun Tung (tytung@uchicago.edu)
Department of Linguistics, The University of Chicago

Alan C. L. Yu (aclyu@uchicago.edu)
Department of Linguistics, The University of Chicago

Abstract

Studies in speech perception have consistently found that the perceived duration of a syllable is significantly influenced by the dynamics of the contour of its fundamental frequency (f0). Syllables with a dynamic f0 contour are perceived as longer than those with a flat f0, even though their acoustic duration is identical; high f0 syllables are perceived as longer than low f0 syllables of the same acoustic duration. Yet, while some listeners exhibit the expected perceptual normalization patterns, others show no f0-induced perceptual adjustments.

This study investigates the neural foundation for this individual variability by examining listeners' scalp-recorded frequency-following response (FFR), a measure of phase-locked auditory encoding in humans that has been used to study subcortical processing in the auditory system. Our findings reveal that the FFR predicts listeners' duration estimation performance in different f0 contexts. Additionally, the FFR predicts the magnitude of the f0 influence on perceived duration, which highlights the complex interaction between sensory processing and speech perception.

Keywords: Pitch perception, duration perception, individual differences, FFR, Visual Analog Scale (VAS)

Introduction

Studies in speech perception have repeatedly found that syllables with a dynamic (e.g., rising or falling) fundamental frequency (f0) contour are perceived as longer than those with a flat f0, even though their acoustic duration is identical (Wang, Lehiste, Chuang, & Darnovsky, 1976; Lehiste, 1970). Conversely, high f0 syllables are perceived as longer than low f0 syllables of equal acoustic duration (Yu, 2010; Gussenhoven & Zhou, 2013). An intriguing aspect of this phenomenon is the significant individual variation in perceptual responses to the interaction between f0 dynamics and syllable duration. While some listeners exhibit the expected perceptual normalization patterns, others show no f0-induced perceptual adjustments (Yu, Lee, & Lee, 2014). Such individual variability casts doubts on the robustness of tone-duration interaction and raises questions about the mechanisms underlying the variable interaction between f0 dynamics and perceived syllable duration, or tone-duration interaction for short.

The variable patterns in the effects of f0 dynamics on duration production and perception have been attributed to both articulatory and perceptual factors. Some researchers suggest that the duration differences under different f0 contexts arise from production biases, and the perceptual responses reflect compensatory listening (Gussenhoven & Zhou, 2013). Studies in speech production, for example, have consistently shown that syllables tend to be longer when they have a dynamic f0 contour compared to their flat, stable f0 counterparts, mirroring the perceptual observations (Xu & Sun, 2002; Gordon, 2001; Zhang, 2001). However, syllables with low f0 are produced longer than those with high f0, which is the reverse of the perceptual pattern (Gandour, 1977). Others posit that the perceptual response is primary, and the production difference reflects an articulatory compensation for the psycho-acoustic effects of f0 information on perceived duration (Yu et al., 2014).

This study explores the neural foundation for the above-mentioned observed individual variability. We hypothesize that individual differences in tone-duration interaction in perception stem from variability in the neural encoding of the speech signal. Specifically, the present study investigates how listeners' scalp-recorded FFR, a measure of phase-locked auditory encoding in humans and has been used to study subcortical processing in the auditory system, predicts variation in listeners' duration estimation responses in different f0 contexts. Understanding the neural mechanisms behind this individual variability may provide insights into the origins of the tone-duration interaction. Specifically, we consider two sources of individual variability in neural encoding of speech:

- General neural encoding hypothesis: Individual variation in the tonal influence on duration perception may be governed by general individual variability in auditory encoding robustness, not specific to the neural encoding of pitch dynamics. Thus, we expect individuals with noisier encoding of the speech signal might exhibit different tone-duration interaction patterns from individuals with more robust neural encoding.

- Pitch-specific encoding hypothesis: Given that the perception of the duration of a syllable is partially contingent on the perception of the duration of the sonorous portion of the speech signal, and the sonorous portion of the speech signal generally carries pitch information, we expect a listener's robustness in pitch encoding might modulate how pitch dynamics influence perceived duration.

Concerning the directionality of the effects of neural encoding on the tonal influence on perceived duration, to the extent that accurate duration estimation is informed by one's ability to track f0/pitch, we expect (i) listeners with more accurate pitch tracking to exhibit more veridical duration estimates. That is, listeners with more faithful pitch tracking might engage less higher-order perceptual adjustments across syllables with different pitch dynamics. Conversely, (ii) listeners with less accurate pitch tracking might exhibit less veridical duration estimates, which might lead to stronger perceptual adjustments in duration estimation.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Methods

Participants

Twenty-eight native speakers of American English (15 females, 10 males, 3 non-binary; aged 18-31, mean age = 22 years old) participated in this experiment. They all grew up in households where all members spoke primarily English until they reached the age of 12. Other than English, some participants were also exposed to Spanish, German, French, Maltese, and Mandarin. None of them reported speech or hearing difficulties, or history of neurological deficits. Participants received US$30 in compensation for their time.

Stimuli

For the VAS task, a recording of the syllable [biːtʃ] *beach* was manipulated in Praat (Boersma & Weenink, 2017) to create stimuli varying the vocalic duration in 50 ms increments, yielding four durations (100 ms, 150 ms, 200 ms, 250 ms).

For the FFR acquisition session, a recording of the syllable [biːtʃ] *beach* with an invariant vocalic duration of 200 ms and a total duration of 340 ms (measured from the stop release to the end of the frication noise of the affricate) was used. The f0 of the target stimuli for both the VAS task and FFR acquisition session was then manipulated to create four f0 contours (*tone* henceforth) using the parameters given in Table 1.

Table 1: The f0 (Hz) values of the four tone types for the stimuli.

Tone	f0 Onset	f0 Offset	Description
T11	100	100	Low
T55	180	180	High
T15	100	180	Rising
T51	180	100	Falling

Experimental design

Participants were seated comfortably in a soundproof booth and completed a VAS task to provide a behavioral measure of the effect of f0 on syllable duration perception. Afterwards, brainstem electroencephalography (EEG) recordings were collected in the same booth. The whole experimental session took about two hours, with breaks after the VAS task, as well as between blocks during the EEG session. Each participant also completed a background questionnaire prior to the in-person experiment session.

Behavioral VAS task

The VAS task (Munson, Schellinger, & Carlson, 2012) was administered through Qualtrics. Participants heard each target syllable once, and used the cursor to indicate their perceived durations on a visual analog scale ranging from "short" on the left to "long" on the right (see Figure 1). For each trial, the cursor was always reset to the midpoint of the scale. The response time had no limit, and the next trial began immediately after a cursor response was detected. In addition, four practice trials preceded the test session to familiarize the participants with the VAS task. The practice trials were two randomized repetitions of the T11 stimuli with a vocalic duration of 100 ms and 250 ms. Since the VAS task itself tested duration estimation specifically, the practice trials contained two duration levels without varying the tonal factor due to time constraints.

Figure 1: The visual analog scale for the VAS task.

EEG/FFR data acquisition and preprocessing

In the EEG/FFR data acquisition session, participants heard the target syllables through inserted earphones (ER-1, Etymotic Research, Elk Grove Village, IL) played in alternating polarities (1000 per polarity) with an inter-stimulus interval ranging among 390, 400, and 410 ms. The four tokens were presented in separate blocks with block order randomized across participants.

The onset of each speech token was recorded through a Stim-Trak box (Brain Products GmbH, Gilching, Germany) connected to an auxiliary channel of the EEG amplifier (actiCHamp amplifier, Brain Products GmbH). This timing information enables time-locking EEG signals accurately with the auditory stimuli onset for subsequent analyses.

Following Bidelman, Moreno, and Alain (2013), the FFRs were continuously recorded from an Ag-AgCl electrode placed at vertex (Cz, active), with a ground electrode at mid-forehead, and two reference electrodes at mastoids (M1, M2). We used the BrainVision Recorder software (Brain Products GmbH) with a sampling rate of 25 kHz. Electrode impedances were kept below 5 kΩ. All recordings were conducted in a soundproof booth. Brainstem electroencephalography recordings were collected while participants listened passively to the auditory stimuli seated comfortably in the booth. To minimize myogenic artifacts, participants were instructed to relax and refrain from extraneous body movement, and to ignore the stimuli as they watched a silent movie throughout the recording session.

In the offline analysis, the evoked responses were downsampled to 10 kHz, bandpass-filtered from 80 Hz to 2,500 Hz, with a notch filter of 60 Hz. Those responses were then epoched from 50 ms pre-stimulus to 320 ms post-stimulus, and baseline-corrected to the mean voltage of the pre-stimulus region. Trials with excessive noise were rejected (T11: mean rejection rate=16%, range=5-24%; T55: mean rejection rate=16%, range=1-24%; T15: mean rejection rate=16%, range=2-24%; T51: mean rejection rate=15%, range=4-25%). Afterwards, all trials were averaged for each stimulus condition for each participant.

EEG data analysis

Following Liu, Maggu, Lau, and Wong (2015) and Krizman and Kraus (2019), we derived six neural measures to index (i) pitch-tracking accuracy and (ii) general neural encoding robustness. Each measure was obtained for each stimulus condition for each participant, to be compared subsequently with the behavioral VAS data. While the test stimuli are 340-ms in length, we only analyzed the first 200-ms of the FFR epochs because the f0 information is most salient and can be most reliably detected during the initial 200 ms, i.e. the vocalic portion of the stimuli. To derive those pitch-tracking measures, we first used the autocorrelation technique to assess the f0 periodicity and phase-locking of the FFR response and the corresponding acoustic stimulus (Wong, Skoe, Russo, Dees, & Kraus, 2007). To better examine the f0 fluctuations in response to the different tonal contours, we employed 40-ms sliding time windows with 39-ms overlap for the autocorrelation analyses. The signal was divided into 40-ms time bins, successively shifted in 1-ms steps from, and correlated with a delayed version of itself. The lag time at which the maximum correlation occurred was recorded, which served to identify the period of the fundamental frequency for each condition for each participant. For each 40-ms time bin, Pearson's r was obtained for each 1-ms interval, and the maximum (peak) autocorrelation value was documented.

The following time, f0/pitch, and amplitude measurements were extracted to determine how well a listener's brainstem encodes the timing, periodicity, and the spectral envelop of the evoking stimuli:

Pitch Strength (PS), a measure of periodicity and phase locking of the response, was calculated by averaging of all autocorrelation peaks (r values from -1 to 1) from all time bins for FFRs per condition per participant. Higher pitch strength indicates higher periodicity.

Pitch Error (PE) (in Hz), which provides a direct estimate of the pitch encoding accuracy of the FFR over the duration of the stimulus, was calculated by averaging the Euclidian distance between the stimulus f0 and response f0 at all time points. The stimulus f0 is shifted in time to match the response f0 according to individual neural lag values identified per condition per participant; neural lag (in ms) (Liu et al., 2015) refers to the time shift to gain the maximum positive correlation between stimulus and response waveforms, and can be determined by the cross-correlation technique. It approximates the FFR latency caused by the neural conduction time of the auditory system.

Mean Squared Difference (MSD) is computed by averaging squared differences between the stimulus f0 and the response f0 at all time points. This is related to Pitch Error as they both measure pitch-tracking accuracy. Notably, in the case of MSD, a fixed neural lag of eight ms was used, following Krizman and Kraus (2019). MSD also weighs the contribution of the larger errors more heavily than PE. Lower PE or MSD suggests smaller differences between the stimulus and response, thus better pitch-tracking.

Stimulus-to-Response Correlation (SRC) measures the strength and direction of the linear relation between stimulus f0 and response f0, taking individualized neural lag into account. This measure ranges from -1 to 1, with higher values representing higher positive correlation; the closer the measure to zero, the weaker the Stimulus-to-Response Correlation.

Root Mean Square (RMS) amplitudes (in μV) of the response waveform is a global measure of the strength of neural encoding for each condition for each participant, with individual neural lag taken into account. The RMS index averages all sample points in the waveform and reflects the magnitude of neural activation over the entire FFR period.

Signal-to-Noise Ratio (SNR) provides the ratio of the RMS amplitude of the response over the RMS of the pre-stimulus region (50 ms).

While Pitch Strength (PS), Pitch Error (PE), Mean Squared Difference (MSD), and Stimulus-to-Response Correlation (SRC) relate to pitch-tracking accuracy, Root Mean Square (RMS) and Signal-to-Noise Ratio (SNR) reflect more general neural encoding robustness.

Results

EEG data

We first examined how tonal contours were encoded in subcortical responses. Figure 2 shows the group-averaged FFRs elicited by the four stimulus conditions (Low T11, High T55, Rising T15, Falling T51) on the right panel, as well as the corresponding acoustic waveform of the four acoustic stimuli on the left panel. The initial 200 ms of the acoustic waveform include the stop and the vowel, while the remaining segment shows the frication noise of the affricate.

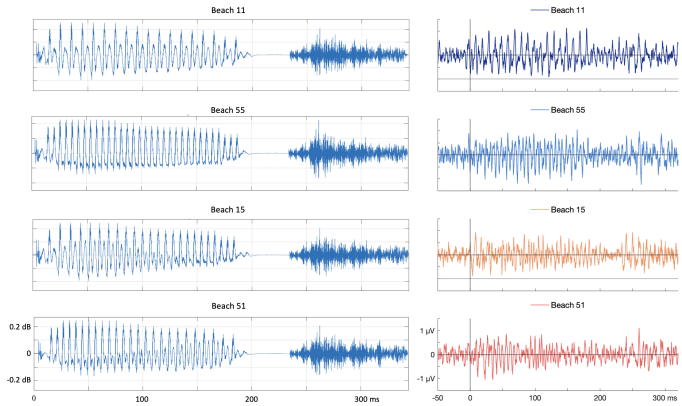

Figure 2: The acoustic waveform of the four acoustic stimuli (left) and the corresponding grand-average FFRs for the four stimulus conditions (right).

In order to examine the accuracy of pitch-tracking at the group level, Figure 3 shows the time-course of periodicity, the f0, extracted using autocorrelation for group-averaged FFR, relative to the f0 of the corresponding stimulus (grey

dots). The figure suggests that FFR tracks the f0 of the low tone stimulus, T11, the best, but not as accurately with respect to the rising (T15) and falling (T51) f0 contours.

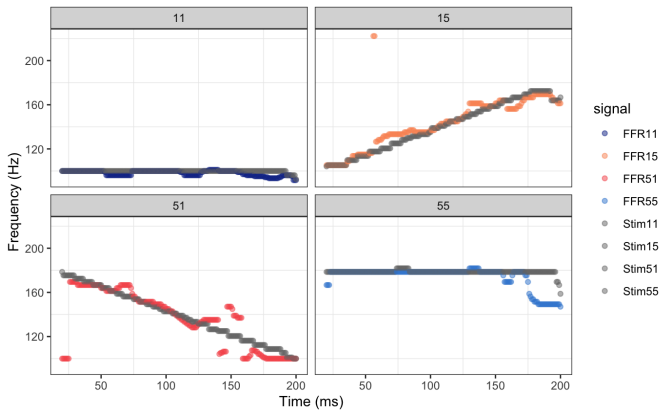

Figure 3: The pitch extraction (f0) for grand-average FFR over time (non-grey colored dots) relative to f0 for acoustic stimulus (grey dots) calculated from autocorrelation respectively for the four stimulus conditions.

Brain-behavior relations

To study how a listener's neural encoding is related to his/her perceptual judgments, this section examines the brain-behavior relationship. The VAS scores were modeled using linear mixed-effects regression fitted in `R`, using the `glmmTMB` package and with the beta distribution family, as it is more appropriate for data that are confined within the interval (0, 1). Before fitting the model, all values of the response variable were normalized within the range (0, 1). The beta distribution requires that all y values meet this constraint, values of 0 were replaced with a small epsilon value (1e-6), and values of 1 were replaced with 1 minus this epsilon value to ensure that all response values are compatible with the beta distribution without significantly altering the data.

Model selection started with a baseline regression model that tested the effects of stimulus DURATION, TONE (T55, T11, T15, and T51), and their interaction. The model also included by-subject random intercepts as well as by-subject random slopes for DURATION and TONE, to allow for subject-specific variation with respect to these variables. A model that included by-subject random slopes for the interaction between DURATION and TONE did not converge. Interactions between the stimulus fixed factors and the neural indices were tested by comparing between models with and without the inclusion of a fixed/random factor and/or interaction. Only interactions that improved model-likelihood significantly were retained. DURATION, treated as a continuous measure, as well as all neural indices were centered and z-scored. TONE was coded as three contrasts: contrast 1 = T55 vs. T11; contrast 2 = T15 vs T51; contrast 3 = Contour Tones [T15, T51] vs. Level Tones [T55, T11]. The final model in `glmmTMB` format is: `VAS ~ DURATION * TONE + TONE * SRC + (1 + DURATION + TONE|PARTICIPANT)`.

	Coef	SE	z value
(Intercept)	-0.278	0.133	-2.096 *
DUR	0.951	0.145	6.575 ***
T55 v. T11	-0.237	0.233	-1.018
T15 v. T51	0.840	0.404	2.079 *
CT v. LT	-0.139	0.237	-0.589
SRC	0.137	0.130	1.055
DUR:T55 v. T11	0.178	0.053	3.346 ***
DUR:T15 v. T51	-0.022	0.053	-0.425
DUR:CT v. LT	-0.095	0.037	-2.543 *
T55 v. T11:SRC	-0.413	0.152	-2.716 **
T15 v. T51:SRC	-0.673	0.412	-1.633
CT v. LT:SRC	0.716	0.256	2.797 **

Table 2: Regression Results. CT = Contour Tone; LT = Level Tone; SRC = Stimulus-response Correlation

As expected, the model shows a significant main effect of DURATION (β=0.951, z = 6.575, $p < 0.001$). There is also a main effect of TONE. Specifically, T15 is perceived as longer than T51 (β=0.84, z = 2.079, $p < 0.05$). There are also significant interactions between DURATION and TONE. Figure 4 shows the model predictions for the interaction. Specifically, the DURATION:TONE$_{T55vs.T11}$ interaction (β=0.178, z = 3.346, $p < 0.001$) suggests that the difference in VAS score between stimuli with a T55 tone and those stimuli with a T11 tone is smaller with longer stimuli. The DURATION x TONE$_{Contourvs.Level}$ (β=-0.095, z = -2.543, $p < 0.05$) suggests that the contour tone stimuli (i.e. T15 or T51), on average, are rated as shorter than stimuli with a level tone (T55 or T11) with longer stimuli. An examination of Figure 4 shows that this reduction is likely driven by the increase in perceived duration of T55 with longer stimuli, however.

With respect to the interactions with the neural indices, there is a significant interaction between TONE and Stimulus-Response Correlation (SRC). The model predictions are shown in Figure 5. While the perceived duration of T11 is stable across SRC, there is a gradual decline in perceived duration of T55 with greater SRC. This is confirmed by the significant interaction between TONE$_{T55vs.T11}$ and SRC (β= -0.413, z = -2.716, $p < 0.01$). Figure 5 also suggests that the perceived duration of T51 varies between individuals with different SRC. This is partly reflected in the significant interaction between the TONE$_{Contourvs.Level}$:SRC interaction (β= 0.716, z = 2.797, $p < 0.01$), with individuals with higher SRC (i.e. more robust pitch tracking) experiencing contour tones as longer than level tones, while individuals with lower SRC experience the opposite difference. This effect is largely driven by the fact that the perceived duration of T55 decreases while the perceived duration of T51 increases across SRC. This is confirmed by a series of pairwise post hoc comparisons conducted using the `emmeans` package in `R` to determine significant differences (α = 0.05) in each tonal environment con-

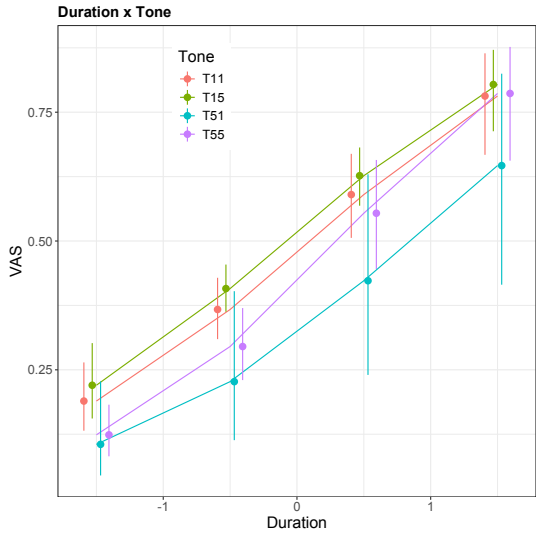

Figure 4: Interaction between Tone and Acoustic Duration (centered and z-scored).

dition across SRC, with p-values corrected using the Tukey HSD method.

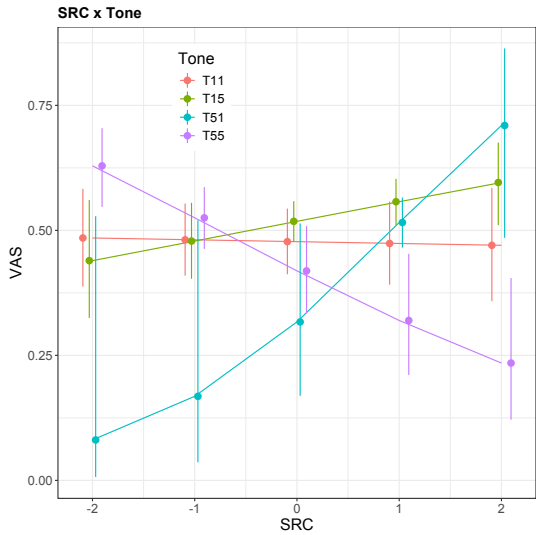

Figure 5: Interaction between Tone and Stimulus-response Correlation (centered and z-scored).

Discussion

In the current study, we examine the neural basis of individual differences in tonal effects on perceived duration. Our participants completed a Visual-Analog-Scale (VAS) task and an FFR recording session. The VAS task examined listeners' perceived duration of syllables with different acoustic duration. Our findings are consistent with previous reports that listeners experience syllables with different pitch dynamics as having different perceived duration, even when the acoustic duration is controlled. Specifically, syllables with a rising pitch contour are heard as longer than syllables with equivalent acoustic duration that have a falling pitch contour. Similar to previous reports, the acoustic duration of the target syllables also modulates the nature of the perceived duration variation. Specifically, while T55 syllables are heard as slightly shorter than T11 syllables, this difference largely disappears with the acoustically longer syllables (e.g. 250 ms). However, unlike previous studies, we did not find a general difference in the perceived duration of T11 and T55. As noted above, the duration estimates of the T55 and T11 syllables are different when the syllable is acoustically short (e.g. 100 ms), with T55 being heard as shorter than T11; this is opposite to what is reported in (Yu et al., 2014). These differences might reflect differences in the sampled populations between studies. The cohort sampled in this study consists mainly of students from a highly selective university, while the participants in (Yu et al., 2014) were recruited on Mechanical Turk. Upon closer examination of the participants' backgrounds, we also found that most of our participants have some musical training (N=20). In order to examine the impact of musical training on tonal effects on perceived duration, we fitted the following model: `VAS ~ Duration * Tone + Tone * Music + (1 + Duration + Tone|Participant)`. The model suggests a potential difference in how participants without music training might have perceived the T11 syllables (β= 0.402, z = 1.874, $p = 0.061$). As shown in Figure 6, while the musicians show the general pattern reported above (i.e. T55 syllables are heard as shorter than the T11 syllables), the non-musicians show an opposite pattern – the T55 syllables are heard as longer than the T11 syllables, reflecting the pattern reported in (Yu et al., 2014). Further examination is needed to ascertain the robustness of this effect of musical training on tone-duration interaction in perception.

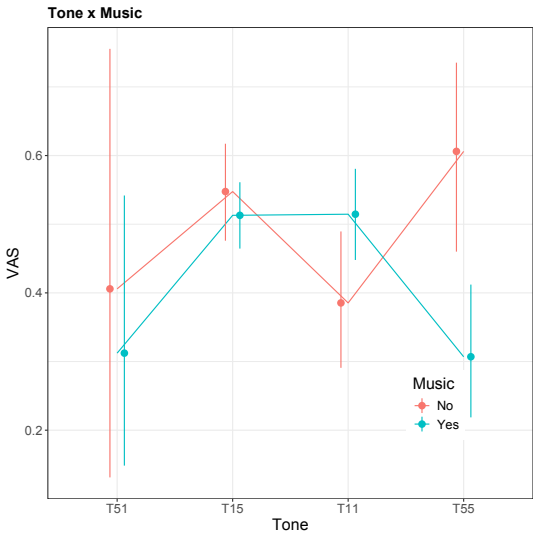

Figure 6: Interaction between Tone and Music.

The FFR investigation also revealed that participants with higher pitch-tracking accuracy, as indexed by the Stimulus-Response Correlation (SRC), show different sensitivities to tonal effects on perceived duration. Participants with lower SRC are more likely to experience syllables with a falling pitch contour as shorter, while syllables with high pitch level (T55) as longer. On the other hand, individuals with more accurate pitch tracking (i.e. higher SRC) experience syllables with a high pitch (T55) as shorter than other syllables with other f0 dynamics, and contour tone syllables as longer than level tone ones. These findings suggest that pitch-tracking accuracy can modulate listeners' duration perception. Listeners with more accurate pitch tracking exhibit a wider range of perceptual adjustments than listeners with lower SRC. These findings are consistent with the pitch-specific encoding hypothesis laid out in the Introduction.

We did not find support for the general neural encoding robustness hypothesis. That is, general neural encoding indices, such as RMS and SNR, did not significantly modulate tone-duration interaction, suggesting that the effect of pitch dynamics on perceived duration might not be related to individual variability in general auditory encoding robustness.

Given that perception of syllable duration is primarily cued by the sonorous portion of the target syllable, we had hypothesized that listeners with more accurate pitch tracking would have more veridical duration perception. This prediction is not borne out. Instead, we found that listeners with more accurate pitch-tracking (a higher SRC score) exhibit different types of perceived duration adjustments compared to those with less accurate pitch-tracking in the FFR. These findings suggest that the perceived duration effect might require explanations at both the early sensory encoding subcortical level and the higher-order cortical processes. Further examination of the listeners' cortical responses to syllables with different pitch dynamics is needed to elucidate the different roles each level of neural representation plays in how pitch dynamics influence duration perception.

Our study joins the recent line of emerging research using individual differences in FFR to predict human performance [e.g., (Reis, Heald, Veillette, Van Hedger, & Nusbaum, 2021)]. In particular, we successfully use FFR to derive behaviorally relevant neural measures of pitch-tracking accuracy to capture individual differences in the tone-duration interaction in early speech processing. We thus demonstrate how variability in early sensory encoding can support and account for the heterogeneity in speech perception at the individual level.

Acknowledgments

This work was supported in part by the National Science Foundation (BCS1827409). We thank Robert McAllister, Jinghua Ou, and Vivienne Jinwen Zhang and for their assistance in collecting the data. Authors are listed in alphabetical order.

References

Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage*, *79*, 201–212.

Boersma, P., & Weenink, D. (2017). Praat. *Software Package*. Retrieved from `http://www.praat.org/`

Gandour, J. (1977). On the interaction between tone and vowel length: Evidence from thai dialects. *Phonetica*, *34*(1), 54–65.

Gordon, M. (2001). A typology of contour tone restrictions. *Studies in Language*, *25*, 405-444.

Gussenhoven, C., & Zhou, W. (2013). Revisiting pitch slope and height effects on perceived duration. In *Interspeech 2013* (pp. 1365–1369). ISCA.

Krizman, J., & Kraus, N. (2019). Analyzing the FFR: A tutorial for decoding the richness of auditory function. *Hearing research*, *382*, 107779.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge MA: MIT Press.

Liu, F., Maggu, A. R., Lau, J. C., & Wong, P. C. (2015). Brainstem encoding of speech and musical stimuli in congenital amusia: evidence from Cantonese speakers. *Frontiers in human neuroscience*, *8*, 1029.

Munson, B., Schellinger, S. K., & Carlson, K. U. (2012). Measuring speech-sound learning using visual analog scaling. *Perspectives on Language Learning and Education*, *19*(1), 19–30.

Reis, K. S., Heald, S. L., Veillette, J. P., Van Hedger, S. C., & Nusbaum, H. C. (2021). Individual differences in human frequency-following response predict pitch labeling ability. *Scientific Reports*, *11*(1), 14290.

Wang, W., Lehiste, I., Chuang, C.-K., & Darnovsky, N. (1976). Perception of vowel duration. *Journal of the Acoustical Society of America*, *60*, S92.

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature neuroscience*, *10*(4), 420–422.

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of Acoustical Society of America*, *111*, 149–174.

Yu, A. C. L. (2010). Tonal effects on perceived vowel duration. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 151–168). Mouton de Gruyter.

Yu, A. C. L., Lee, H., & Lee, J. (2014). Variability in perceived duration: pitch dynamics and vowel quality. In *Fourth international symposium on tonal aspects of languages.*

Zhang, J. (2001). *The effects of duration and sonority on contour tone distribution: Typological survey and formal analysis* (Unpublished doctoral dissertation). University of California, Los Angeles, Los Angeles.