SPECTRAL CUES IN PITCH PERCEPTION IN ENGLISH AND CANTONESE SPEECH AND MUSIC

May Pik Yu Chan, Jianjing Kuang

University of Pennsylvania
pikyu@sas.upenn.edu, kuangj@sas.upenn.edu

ABSTRACT

Listeners integrate multidimensional cues (e.g., spectral shape) during pitch processing in both speech and music. While prior works have shown that this effect applies to tonal and non-tonal speakers, it remains unclear whether listeners' pitch processing strategies differ on different linguistic levels (e.g. tone vs intonation). Furthermore, work has yet to show whether differences in pitch processing lie between natural (both speech and music) vs artificial sounds, or speech vs non-speech sounds. A pitch classification experiment investigated these questions. Listeners were either English or Cantonese speakers listening to stimuli modeled after Cantonese prosody, or English speakers listening to comparable violin stimuli. Results show that all participants integrate spectral cues in pitch perception. However, the effect of musicality on shift but not categoricity differed between Cantonese and English speech conditions. Overall results suggest shared pitch processing for speech and music.

Keywords: Pitch perception, Spectral Slope, Speech and Music, Musicality, Prosodic effects

1. INTRODUCTION

Listeners rely on pitch in both linguistic (e.g. tone, lexical stress, phrasal prosody and intonation) and musical contexts (e.g. melody, harmony). Although fundamental frequency, F0, is understood as an important acoustic property of pitch, various works have shown that pitch perception is driven by multidimensional cues, such as timbral cues in music [1, 2, 3] or spectral slope cues in speech [4, 5, 6], intensity [7].

One core question of the present work is at what level of processing cue integration happens. For example, while work comparing speech and non-speech has shown that integrate spectral slope cues in both sawtooth and speech sounds, listeners are more sensitive to relative differences between spectral slope cues in speech but not non-speech [6]. In other studies focusing on music only, listeners were also found to integrate timbral cues, though works have tended to focus on fully instrumental comparisons [8, 9] or artificially synthesized stimuli [1, 2, 3], none of which have been directly comparable to speech.

Apart from speech vs non-speech, another question within speech is whether there is a processing difference between linguistic vs non-linguistic levels of processing. Prior work on integration comparing speakers' language backgrounds have not found language effects, regardless of whether speakers speak a tonal or non-tonal language [4], or a tonal language with multiple pitch height contrasts [10]. Furthermore, [11] has found that Cantonese participants integrate gender cues to the same extent with respect to lexical tones, though Cantonese musicians and non-musicians process pitch differently when listening to non-linguistic vocalization. We therefore speculate that the degree to which pitch is phonologically defined or specified may affect the extent of cue integration in pitch perception. In other words, when a "speech" mode encompasses both linguistic and non-linguistic speech, it is unclear whether the processing of such "language mode" differs from non-linguistic speech.

On a broader view, differences in processing may also potentially lie in natural vs artificial sounds. While [6] has shown that listeners process speech sounds differently from sawtooth sounds, it remains unclear whether this effect is due to sawtooth sounds not evoking a "speech mode", or because sawtooth sounds are not naturally occurring at all. In this study, we seek to compare speech sounds with violin-sounding stimuli. This is because violin sounds also contrast in "voice quality" with *sul ponticello* and *sul tasto* (playing closer to vs away from the bridge) resulting in differences in spectral slope. These sounds are comparable to "tenser" and "breathier" sounding voicing in speech respectively. Since bowed instruments like the violin are fixed-formant instruments with a clear F1 and F2 structure, it is directly comparable to speech with a stable vowel. The use of violin stimuli would therefore allow us to compare speech vs non-speech, using stimuli that are still natural sounding.

We seek to answer the following questions: First, whether linguistic structure affects listeners' degree of cue integration in pitch perception. While prior work has shown that listeners' experience with tonal languages does not have effects on cue integration [6, 4, 10], it remains unclear whether the prosodic structure of the stimuli plays a role. Specifically, since previous work included stimuli modeled off English prosody, it is possible that, for Mandarin and Cantonese listeners, the stimuli based on foreign prosody do not trigger listeners' "linguistic mode" thereby dampening potential language effects. Therefore, in this study, we test whether the effect is transferable to stimuli of a syllable-timed prosodic structure that could be mapped to tonal categories that may affect listeners' pitch perception strategies. Second, whether musical stimuli showing within-instrument differences are as integrative as speech. And third, as various works have shown that listeners' musical competence affects pitch processing in both linguistic [12, 13, 14, 15] and non-linguistic tasks [16, 17] as well as integration strategies [4], we also test whether the effects of musicality remain across both languages and modalities. A pitch classification experiment was performed to test these questions.

2. METHODS

A total of 183 participants participated in a pitch classification experiment. 88 were native English speakers, 47 of which participated in the speech condition, 41 of which participated in the music condition. 95 participants were native Cantonese speakers and participated in the speech condition.

2.1. Pitch classification experiment

To test whether listeners of different backgrounds integrate spectral cues in pitch perception differently in speech and music modalities, a pitch classification experiment following the methods in [6] was done. Stimuli involving two pitch contours were resynthesized from a native Cantonese female talker's production of the pseudoword "ma-ma-ma", or from a violin's recording of the note sequence G3 (196 Hz), D4 (294 Hz), G3 (196 Hz). Unlike the stimuli in [6], the production of the pseudoword was in Cantonese T4-T1-T4 (low falling - high level - low falling, as in 麻媽麻) tones, which matched closely to the pitch interval of the violin note sequence. By extension, the tempo of our current stimuli includes three syllables being comparably long in duration, contrasting stimuli from [6] where lexical stress in the second syllable is significantly longer than the first and third unstressed syllables. Both recordings were made in the same recording environment and were around 700 ms.

Resynthesis was done in Praat, where the recordings were first normalized to 700 ms. The pitch tier from the Violin recording was extracted. It was then hand-corrected to smooth out transitions between notes, and to create a little pitch declination on the two G3 notes, referencing the low falling tone regions of the Cantonese recordings. An eleven-step continuum, each 0.35 semitones apart, was then created based on the manipulated pitch tier. Step 6 corresponded to the original pitch, and steps 1-5 were lower in F0, while steps 7-11 were higher in F0. Each stimulus consisted of two pitch peaks (in other words, 6 syllables, or 6 violin notes). The first pitch peak has a consistent F0 range, and only included the pitch peak at Step 6; the second pitch peak included pitch peaks from the 11-step continuum.

Spectral slopes of the sound sources were manipulated by inverse filtering of an LPC (burg) object in Praat, after resampling both sound files to 20k Hz. The original sound source was used as the "tenser" voice quality stimuli source, and a "breathier" version of the sound source was created by using the Filter (de-emphasis) command, which decreases the spectral slope by 6 dB/octave. The pitch tiers were then replaced by the manipulated pitch tiers described above, and the stimuli were filtered using the original LPC objects of the speech and violin recordings. The manipulated pitch peaks were then concatenated with the baseline (step 6) pitch peak in all "tenser" or "breathier" spectral slope permutations (TB, BT, TT, BB). A 500 ms gap was created between the two pitch peaks to avoid listeners comparing the last note of the first peak and the first note of the second peak, instead of listening to the overall contour. 88 stimuli were created in total (2 sources (speech/violin) x 4 spectral slope permutations x 1 baseline F0 x 11 F0 steps.)

Participants completed the experiment using Qualtrics. Cantonese-speaking participants only completed the speech condition, and English participants completed either the speech or the violin condition. A between-subject design was used to avoid training effects and to allow for sufficient repetitions of trials. Listeners listened to one concatenated pitch peak pair per trial and were asked to judge whether the second pitch peak was higher in pitch than the first. For English-speaking participants, each stimulus was repeated 5 times; and for Cantonese-speaking participants, each stimulus was repeated 3 times. All trials were randomized.

To quantify participants' cue integration extents, a categoricity and a shift score were calculated per participant. The categoricity score was calculated by the mean 'the second peak is higher' response rates in steps 7-11 minus that of steps 1-5. Participants with more categorical responses will have a higher score, and those with less categorical responses will have a lower score. A shift score was calculated by summing the distance between the TB/BT condition and the baseline conditions. In other words, we take the absolute value of average responses from the TB minus BB, TB minus TT, TT minus BT, and BB minus BT. We then divide the summed value by two, as there are two baseline conditions. This allows us to scale the shift score from 0 to 1, such that it is comparable to the categoricity score. A high score would suggest that participants integrate spectral cues more, resulting in more shift from the baseline. A low score would suggest that the participant attends to F0 more, thus shifting less from the baseline.

2.2. Musicality test

We use a continuous musicality score as a proxy for listeners' musical competence, which can capture both the musical proficiency achieved from musical training and from innate musical disposition. Listeners completed the Montreal Battery of Musical Abilities (MBEMA) [18] after the pitch classification task, also administered on Qualtrics in one setting. The task includes a melody, rhythm and memory task. The melody and rhythm task asks listeners to distinguish whether two melodic phrases, that may differ slightly in the melody or the rhythm, were the same or different, The memory task asks participants to identify whether a melodic phrase was novel or repeated from the former two tasks. Each task includes 20 questions, resulting in 60 questions total. The percentage of participants' overall correct responses is used as a measure of participants' musical competence.

3. RESULTS

The proportion of "the second peak is higher" responses faceted by experiment condition is shown in figure 1. Across all three conditions, listeners integrate spectral cues in the same direction. When listeners hear tenser-breathier stimuli for example, listeners are biased towards thinking that the second pitch peak is higher in pitch, even in cases when the F0 is not necessarily higher. The opposite direction holds for the breathier-tenser condition. Among the baseline conditions, the tenser-tenser condition biases participants towards a higher second pitch peak compared to the breathier-breathier condition. Across the three experiments, participants had a mean musicality score of 50.66 (SD = 7.86), and their mean melody, rhythm, and memory scores were 17.03 (SD = 2.42), 16.57 (SD = 3.30) and 17.06 (SD = 3.28) respectively.

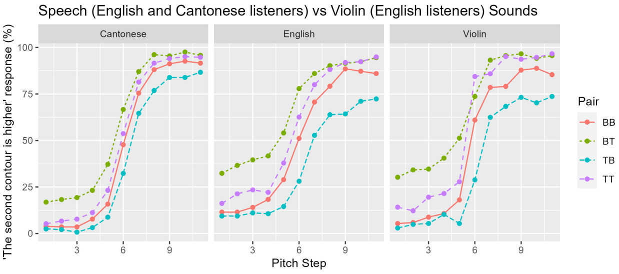

Figure 1: Percent the second contour is higher response rates

We fitted a mixed effect logistic regression model to evaluate the effect of spectral condition (4-factor levels), experiment condition (3-factor levels, reference: English), and pitch step (11 numeric steps) on participants' responses. The reference levels of the spectral condition were rotated to achieve pairwise level comparison. By participant random intercepts were included. Results showed a significant main effect of F0 (β = -0.609, SE = 0.007, $z = -93.024$, $p < 0.0001$), and that responses from the Cantonese condition differed significantly from the English condition (β = 0.229, SE = 0.074, $z = 3.114$, $p = 0.002$), but the Violin condition did not differ significantly from the English condition (β = 0.003, SE = 0.086, $z = 0.039$, $p = 0.969$). The main effect of spectral slope differed significantly from each other across all pairs, reaching p<0.0001 even after applying Bonferroni correction for multiple testing,

To understand the effect of musicality on listener responses, figure 2 shows participant responses after they were grouped into musicality quartiles within their respective experiment. Participants with the highest musicality scores are in the group 4 facet, who tends to have the most categorical responses with less differences between spectral conditions compared to other groups. By contrast, listeners in group 1, who have the lowest musicality scores in their experiment condition, show the

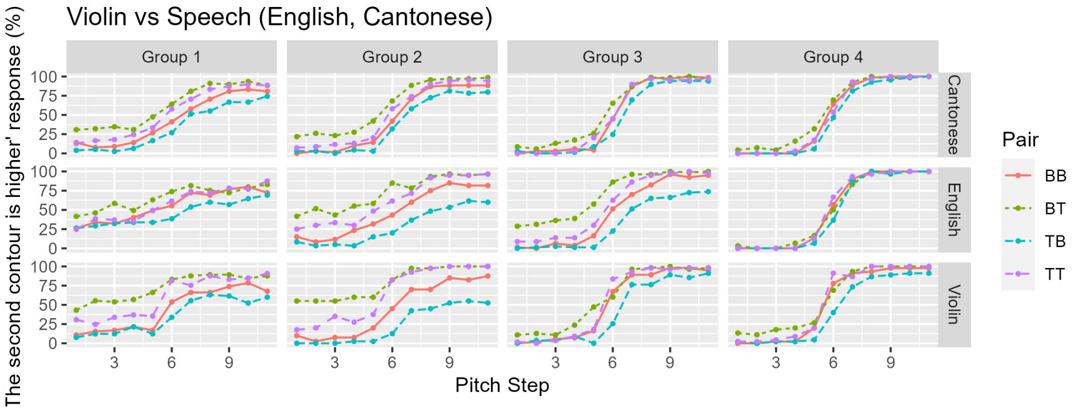

Figure 2: Responses faceted by musicality quantiles within each experiment

least categorical responses with more shift between spectral slope conditions. The difference between groups appears to be least robust in the Cantonese condition, where participants even in group one, appear to have relatively categorical responses.

We model participants' categoricity scores and shift scores by their musicality scores. We ran two linear regressions with the same model specifications, predicting either categoricity or shift. Main effects of musicality (numeric) and experiment condition (3-factor levels, reference: English) were included, as well as their interaction term.

Results for the categoricity model observed a significant main effect of musicality (β = 1.32, SD = 0.27, t = 4.84, p < 0.0001), suggesting that listeners with higher musicality have more categorical responses. No other terms were significant. Results for the shift model found no significant main effect of musicality, unlike results in [6, 4]. However, there is a significant interaction effect of musicality and the Cantonese experiment condition (β = -0.81, SD = 0.41, t = -1.98, p = 0.049). The results suggest that there is a stronger negative correlation between shift and musicality for Cantonese, which was significantly differing from English responses. Figure 3 illustrates the effect of musicality on categoricity and shift.

4. DISCUSSIONS

In this study, we aim to understand (1) whether linguistic level affects listener's pitch perception strategies, (2) whether speech vs non-speech mode results in different pitch processing strategies, and (3) whether the effect of musicality is consistent across conditions.

We test our first question by using stimuli modeled after Cantonese tones and comparing responses from English and Cantonese speakers. Our results show that listeners of both languages attend to spectral cues in similar directions during pitch perception. Listeners are biased towards thinking tenser sounding stimuli is higher in pitch and vice versa. However, our results show that when the stimuli are modeled after Cantonese prosody, responses from Cantonese speakers do differ from that of English speakers. Moreover, the effect of musicality on shift is more apparent for Cantonese speakers than English speakers, though categoricity patterns similarly. These results suggest that linguistic levels may play a minor role in affecting pitch perception, though cue integration is overall consistent in speech processing by speakers of different languages. We also show that listeners of the violin condition exhibit similar responses compared to the speech responses. This contrasts with results consisting of sawtooth sounds in [6], suggesting that listeners integrate spectral cues in pitch perception similarly in natural sounding stimuli, and not only in speech. Overall results suggest that cue integration in pitch perception spans multiple general and linguistic domains.

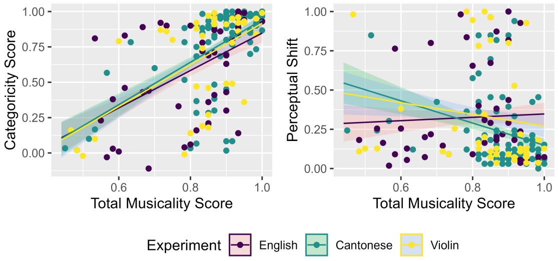

Figure 3: The relationship between musicality and categoricity and shift scores

5. REFERENCES

[1] E. J. Allen and A. J. Oxenham, “Symmetric interactions and interference between pitch and timbre,” *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1371–1379, 2014.

[2] P. G. Singh and I. J. Hirsh, “Influence of spectral locus and f 0 changes on the pitch and timbre of complex tones,” *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2650–2661, 1992.

[3] C. M. Warrier and R. J. Zatorre, “Influence of tonal context and timbral variation on perception of pitch,” *Perception & Psychophysics*, vol. 64, no. 2, pp. 198–207, 2002.

[4] A. Cui and J. Kuang, “The effects of musicality and language background on cue integration in pitch perception,” *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4086–4096, 2019.

[5] J. Kuang and M. Liberman, “The effect of spectral slope on pitch perception,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] ——, “Integrating voice quality cues in the pitch perception of speech and non-speech utterances,” *Frontiers in psychology*, vol. 9, p. 2147, 2018.

[7] J. G. Neuhoff and M. K. McBeath, “The doppler illusion: The influence of dynamic intensity change on perceived pitch.” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 4, p. 970, 1996.

[8] C. L. Krumhansl and P. Iverson, “Perceptual interactions between musical pitch and timbre.” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 3, p. 739, 1992.

[9] J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg, “The dependency of timbre on fundamental frequency,” *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2946–2957, 2003.

[10] M. P. Y. Chan and J. Kuang, “Spectral cue integration in pitch perception by Cantonese listeners,” 2021, 1st International Conference on Tone and Intonation.

[11] W. Lai and J. Kuang, “The effect of speaker gender on Cantonese tone perception,” *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 4119–4132, 2020.

[12] J. A. Alexander, P. C. Wong, and A. R. Bradlow, “Lexical tone perception in musicians and non-musicians,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[13] D. Behne, Y. Wang, M. H. Ro, A.-K. Hoff, H. A. Knutsen, and M. Schmidt, “Effects of musical experience on linguistic pitch perception training,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3168–3168, 2006.

[14] T. L. Gottfried and D. Riester, “Relation of pitch glide perception and Mandarin tone identification,” *Journal of the Acoustical Society of America*, vol. 108, no. 5, p. 2604, 2000.

[15] C.-Y. Lee and T.-H. Hung, “Identification of Mandarin tones by English-speaking musicians and nonmusicians,” *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3235–3248, 2008.

[16] A. Houtsma, N. Durlach, and D. Horowitz, “Comparative learning of pitch and loudness identification,” *The Journal of the Acoustical Society of America*, vol. 81, no. 1, pp. 129–132, 1987.

[17] R. Wayland, E. Herrera, and E. Kaan, “Effects of musical experience and training on pitch contour perception,” *Journal of Phonetics*, vol. 38, no. 4, pp. 654–662, 2010.

[18] I. Peretz, N. Gosselin, Y. Nan, E. Caron-Caplette, S. E. Trehub, and R. Béland, “A novel tool for evaluating children’s musical abilities across age and culture,” *Frontiers in systems neuroscience*, vol. 7, p. 30, 2013.