

Asymmetric Discrimination of Pitch Contours among Non-native Tone Listeners

Ratree Wayland¹

Hae Won Kim¹

Rachel Meyer¹

Si Chen²

Department of Linguistics, University of Florida, USA¹

Department of Chinese and Bilinguals Studies,

The Hong Kong Polytechnic University, Hong Kong²

Corresponding author's email: ratree@ufl.edu

Received 5 September 2024; revised 29 October 2024; accepted 30 October 2024

Abstract

A well-established perceptual bias exists in both infants and adults, favoring the shift from a less peripheral vowel to a more peripheral one. However, the perceptual processes and the information on which these processes operate are not yet fully understood. This study explored the role of frequency modulation in the development of perceptual asymmetries in speech tonal stimuli and tested the hypothesis that a shift from a less spectrally dynamic stimulus to a more spectrally dynamic one is more perceptually salient. The effects of memory load, a factor not considered by previous studies, were also explored. Fifteen native speakers of American English discriminated pairs of speech tonal stimuli varying in pitch contour (falling, rising, level) at short (250 ms) and long (1,000 ms) inter-stimulus intervals. Consistent with the spectral dynamics account, a change from a level pitch contour to a falling and a rising pitch contour was easier to detect than the opposite direction. However, inconsistent with the spectral dynamic hypothesis, a change from a falling tone was easier to discriminate than a change from a rising tone, but it resulted in a longer reaction time compared to a level tone. Poorer sensitivity to pitch contour, as opposed to pitch height or average pitch, among non-tonal language speakers may explain the results.

Keywords: perception, asymmetry, bias, tones

1. Introduction

Perceptual asymmetries between two stimuli, such as when stimulus A is more often mistaken for stimulus B than vice versa (directional asymmetry), have been previously explained by response bias, based on the common assumption that perceptual space follows a Euclidean structure (Polka & Bohn, 2003). Findings of perceptual asymmetries across various domains, such as color, line orientation, numbers, geometric shapes, and concepts of countries (Rosch, 1975; Tversky & Gati, 1978), imply that these asymmetries may provide insight into the underlying nature and organization of perceptual space (Polka & Bohn, 2003). For instance, the existence of perceptual asymmetries in vowels suggests that not all vowels are equally salient. Specifically, it is well-documented that a change from a vowel located on the periphery of the perceptual vowel space to a non-peripheral vowel is more challenging to detect compared to the reverse (e.g., Masapollo et al., 2017b; Polka & Bohn, 2003, 2011). Nevertheless, the specific mechanisms and the information guiding these processes have yet to be fully clarified. This study investigated the influence of acoustic variability on perceptual asymmetries by using speech tonal stimuli. The goal was to test the hypothesis that shifts from a more stable pitch contour (such as a level tone) to a more varied pitch contour (such as a falling or rising contour) are more easily detected than the reverse.

2. Directional Perceptual Asymmetries

Perceptual asymmetries (henceforth asymmetries) have been reported for vowels. Polka and Werker (1994) found a strong effect of presentation order on the discrimination of the German /u/-/y/ and /ʊ/-/ɪ/ contrasts in both 6–8-month-old and 10–12-month-old English-learning infants. Both groups performed better when the shift was from /y/ to /u/ (rather than /u/ to /y/), and from /ɪ/ to /ʊ/ (rather than /ʊ/ to /ɪ/). Polka and Werker

(1994) noted that these findings align with the perceptual magnet effect, which had previously been observed in within-category vowel discrimination in both infants and adults (Kuhl, 1991; Kuhl et al., 1992). The results suggested that vowels more typical of the native language (i.e., /u/ and /ʊ/) function as perceptual magnets, unlike less native-like vowels (i.e., /y/ and /ʏ/). However, Polka and Bohn's (1996) finding that both English- and German-learning infants, aged 6-8 months and 12 months, show a preference for changes from /y/ to /u/ (a contrast not present in English) and from /ɛ/ to /æ/ (a contrast not present in German) rather than the reverse, suggests a language-universal bias rather than one based on familiarity with the native language. Subsequent studies on infant and adult vowel perception have led to the proposal of the Natural Referent Vowel (NRV) hypothesis (Polka & Bohn, 2011). The NRV framework posits that shifts from a less peripheral vowel to a more peripheral vowel are easier to distinguish than the reverse. Moreover, the absence of perceptual asymmetries in vowel discrimination among animals (e.g., Hienz et al., 1981) implies that this preference may be specific to humans.

Directional asymmetries have been observed in infants' perception of consonants as well. For instance, Kuhl et al. (2006) reported that both 6-8- and 12-month-old Japanese and English-learning infants found the stimulus change from /ra/ to /la/ more difficult to detect than from /la/ to /ra/. Interestingly, there was no interaction with age. American English-learning infants showed improvement in discrimination over time in both directions, whereas Japanese infants' performance declined with age, also in both directions. In adults, Plauché (2001) found similar asymmetries in consonant place-of-articulation confusions, such as [ki] being mistaken for [ti] more often than the reverse, which is influenced by phonotactic frequency, degraded acoustic cues, and variability in production.

Directional asymmetries have also been reported in the perception of lexical tones by both infants and adults. For instance, Tsao (2008) found that one-year-old Mandarin-learning infants found it easier to detect a change from the background Mandarin T1 (55) to the target Mandarin T3 (213) than the reverse. Similarly, Yeung et al. (2013) reported that 4- and 9-month-old Mandarin-learning infants were better at discriminating Cantonese tones after being familiarized with Cantonese T2 (25) compared to Cantonese T3 (33). Francis and Ciocca (2003) discovered that adult native Cantonese speakers had better tone discrimination when the first syllable had a higher frequency (by approximately 4 Hz) than the second syllable. Lastly, in an ERP study, Politzer-Ahles et al. (2016) found that Mismatch Negativity (MMN) was reduced for both native and non-native Mandarin listeners when Mandarin T3 served as the standard and other Mandarin tones were the deviant, compared to the opposite scenario.

Despite their ubiquity, the perceptual processes underlying asymmetries, the information on which these processes operate, and other constraining factors, have yet to be fully understood. According to the NRV hypothesis, asymmetries arise from listener bias toward vowels produced with extreme articulatory gestures, with corresponding salient acoustic information characterized by the convergence of adjacent formant frequencies (Polka & Bohn, 2011). For example, relative to the neutral vocal tract configuration, the peripheral high front unrounded vowel /i/ is produced with the tongue in the extreme front position, leading to acoustic convergence of the second, third and fourth formants (F2, F3 and F4). On the other hand, the peripheral high back rounded vowel /u/ vowel is produced with an extreme lip rounding gesture, resulting in a convergence between F1 and F2.

Convergence of two or more formant frequencies leads to mutual enhancement of their amplitudes and, thus, a concentration of acoustic energy in a narrow spectral range (e.g., Stevens, 1999). It has been hypothesized

that focalized spectral energy renders prominence to a vowel, leading to its easier detection, encoding, and memorization by the listener (e.g., Masapollo et al., 2017a, b; Masapollo et al., 2019; Schwartz & Escudier, 1989). A listener may rely on two potential acoustic properties to gauge degrees of focalization of a vowel: spectral proximity—the distance between neighboring formants, and spectral dynamics—the steepness of formant slope (Masapollo et al. 2019). Spectral focalization has been found to enhance vowel discrimination, with listeners more accurately detecting changes from vowels with less formant convergence to those with greater formant convergence, compared to the reverse. Additionally, vowels synthesized with closer formant proximity were perceived as better exemplars of their respective vowel category (Masapollo et al., 2017b; Schwartz & Escudier, 1989).

Masapollo et al. (2019) examined the effect of these two acoustic properties on directional asymmetries in vowel perception in a series of five experiments using non-speech tonal analogs. The first experiment examined whether the asymmetric discrimination of English /u/ (less focal) and French /u/ (more focal) observed on speech stimuli in Masapollo et al. (2017b) would also be evident in non-speech tonal analogs. The stimuli were two non-speech tonal analogs of the center frequency of F1 and F2 of the English /u/ and French /u/ from Masapollo, et al. (2017b). Tone 1 was a low-level tone whereas Tone 2 was a high-falling tone. Directional asymmetries were elicited; the listeners were better at detecting a change from the less focal tonal stimuli to the more focal tonal stimuli, compared to the reverse, but the effect was weaker than what was found with speech stimuli in Masapollo et al. (2017b).

The effects of spectral proximity were tested in Experiment 2, but no directional asymmetries were found; a change from a token with the two tonal analogs spaced further apart to a token with them spaced closer together was not any easier to detect, compared to the reverse. In

Experiment 3, the spectral proximity and spectral dynamic explanations for directional asymmetries were directly compared, and unexpectedly, listeners were equally adept at discriminating changes from less dynamic/more proximal tones to more dynamic/less proximal tones, as well as in the reverse direction. However, the authors noted that the null result might be attributed to near ceiling effects.

Experiments 4 and 5 examined the degrees of spectral dynamics in the perceptual asymmetries observed in Experiment 1, using only Tone 2 (high-falling) as the stimulus. Results of Experiment 4 confirmed that the directional asymmetries observed in Experiment 1 were driven by the frequency modulation (steepness of the fall of Tone 2) as opposed to the proximity between the two (low and high-falling) tones. However, the offset and the slope of the tonal stimuli were confounded in this experiment: while their onset values were the same, their offset values, thus their slopes (i.e., spectral change/dynamicity) were different. When the steepness of both sets of stimuli was equated, no directional asymmetries were elicited in Experiment 5.

These results demonstrate that perceptual asymmetries may also occur in non-speech stimuli with analogous acoustic properties of vowels, but to a lesser degree, and that they are driven by the amount of spectral change (spectral dynamicity) in the stimuli. Overall, these results were interpreted as supporting the NRV framework, which suggests that asymmetries in vowel perception are driven by speech-specific processes that respond to the way articulatory gestures influence the acoustic structure of speech (Masapollo et al., 2019, p. 15).

Similar to Masapollo et al. (2019), the current study investigated the influence of stimulus properties on perceptual asymmetries and also tested the spectral dynamics hypothesis, which proposes that detecting a change from a less dynamic stimulus (with a smaller degree of frequency modulation) to a more dynamic one is easier than detecting the reverse.

Only single-component tonal stimuli were used, like those employed in Masapollo et al.'s (2019) Experiments 4 and 5. However, to avoid ceiling effects, we used a smaller amount of frequency modulation (30 Hz slope vs. the 700-1,100 Hz slope used by Masapollo et al., 2019). In addition, some available evidence suggests that pitch encoding may be less accurate in non-speech stimuli than in speech stimuli (Swaminathan et al., 2015). Furthermore, Masapollo et al. (2019) found that asymmetries were weaker in non-speech tonal stimuli than in speech stimuli. Thus, instead of non-speech tonal stimuli, speech tonal stimuli where pitch contours were resynthesized over a naturally produced syllable were used. In addition to falling pitch contours, rising and level pitch contours were also included, thereby extending the evaluation of the spectral dynamics account to more contour types. Lastly, the inter-stimulus interval (ISI) was manipulated (250 ms vs. 1,000 ms) to gauge the effects of memory load on asymmetries. Asymmetries may be more likely to emerge due to auditory decay of pitch information in the longer ISI discrimination condition.

3. Materials and Methods

3.1. Participants

Twenty native speakers of American English were recruited for the study ($M = 21.5$ years, $SD = 2.82$, Range = 18–31 years; 11 males, 9 females). According to the language background questionnaire, fifteen of the participants reported varying degrees of experience with other languages including French, German, Spanish, Italian, and Tagalog. One participant reported one year of Mandarin experience in a classroom setting and four others were heritage speakers of Mandarin, Taiwanese, and Cantonese. In addition, participants were asked to report musical and vocal training as such experience can enhance perceptual sensitivity and aid in the discrimination of tonal stimuli (Wayland et al., 2010; Xie & Myers, 2015). None of the participants reported more than five years of formal musical training. All

participants reported normal hearing with no history of reading, hearing, or speech disorders. Participants volunteered to partake in this experiment and were offered the option to receive an hour of course credit, if applicable, but none chose to accept this offer. All participants provided informed consent prior to their involvement in the study. The study adhered to the principles of the Declaration of Helsinki, and the protocol was approved by the University of Florida Institutional Review Board (UFIRB201903001).

3.2. Stimuli

The stimuli used in this study were the same as those used in Guion and Pederson (2007). To prevent the influence of lexical effects on the participant performance, the [wa] syllable, which is a legal but meaningless consonant-vowel sequence in English, was used. The syllable, produced by a male speaker with a duration of 110 ms, was resynthesized using Synthworks software from Scicon (<http://www.sciconrd.com/synthworks.aspx>) to create nine stimuli with varying F0 contours (Table 1, Figure 1). According to the transcription system for Chinese tones, 1 indicates a low tone, 3 indicates a mid tone, and 5 represents a high tone. The onset fundamental frequency (F0) of the three tones varied in equal increments of 30 Hz, ranging from 85 Hz to 115 Hz. Level tones remained constant throughout the syllable, while rising and falling tones were linearly interpolated across the syllable with a slope of 30 Hz.

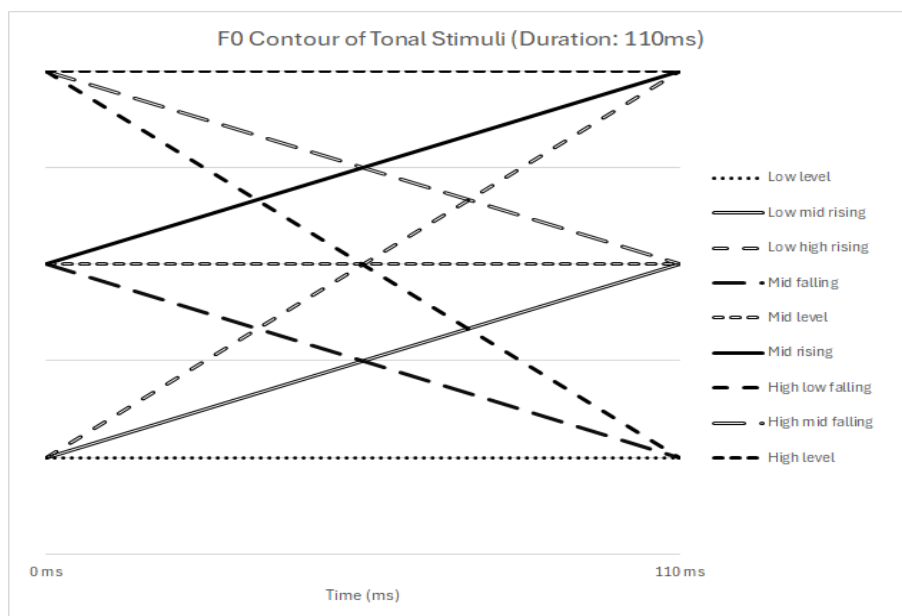
Table 1

The nine tonal stimuli used in the study

Name	Contour	F0 beginning (Hz)	F0 ending (Hz)	Average F0 (Hz)
11	Low-level	85	85	85
13	Low-mid rising	85	115	100
15	Low-high rising	85	145	115
33	Mid-level	115	115	115
31	Mid-low falling	115	85	100
35	Mid-high rising	115	145	115
55	High-level	145	145	145
53	High-mid falling	145	115	130
51	High-low falling	145	85	115

Figure 1

F0 Trajectories for Nine Tonal Stimuli



Participants were presented with all possible pairings of the nine stimuli, including identical pairings, in an AX same-different discrimination task. This resulted in a total of 72 different pairings and 9 identical pairings. The same pairings were included to confirm that participants understood and could complete the task. To prevent response bias, an equal number of ‘different’ and ‘same’ trials were presented to the participants. In total, the participants heard 144 trials: 72 ‘different’ trials and 72 ‘same’ trials (9 same pairings x 8 repetitions). To manipulate memory load, two inter-stimulus intervals (ISI) were used: 250 ms and 1,000 ms.

3.3 Procedure

Participants completed a categorical discrimination task where they were asked to judge whether a pair of tones, separated by an inter-stimulus interval (ISI) of either 250 ms or 1,000 ms, sounded the same or different. All participants were tested on both ISIs, and the order of presentation of the two blocks was counter-balanced across participants. Trials could not be repeated, and the inter-trial interval was set at 3 seconds. Response accuracy and reaction time were recorded for each trial. Participants were tested individually in a quiet phonetics laboratory using a Dell desktop computer, keyboard, and mouse. The experiment was programmed with E-Prime software (version 2.0), which controlled stimulus presentation and collected responses. Stimuli were delivered through headphones at a comfortable listening level, adjusted by the participants during a brief practice session.

A set of 10 practice trials (5 different, 5 same trials) with a 500 ms ISI was presented during the practice session. They were instructed to press ‘0’ on the keyboard for tones they determined to be the same, and ‘1’ for tones they thought to be different. Feedback was immediately provided.

For the experimental session, two blocks of stimuli (144 trials in each block) varying in ISI were presented. All trials in each block were fully

randomized for each participant. The presentation of the two blocks was counterbalanced across participants. As described above, each block consisted of 144 pairs of tones, half of which exhibited a change in stimuli, and half of which did not. Like the practice session, participants were instructed to press '0' on the keyboard for tones they determined to be the same, and '1' for tones they thought to be different. However, unlike the practice session, no feedback was provided. Participants were allowed a short break between blocks. All participants were informally debriefed and interviewed regarding the nature of the stimuli following the completion of the test trials. None of the participants reported that the stimuli were interpreted as meaningful lexical items.

Data from 4 Chinese heritage speakers and 1 English speaker with 1 year of Mandarin experience were excluded since exposure to a tonal language could affect their auditory processing of pitch (e.g., Bent et al. 2006). Additionally, reaction times greater than 3,500 ms (3 SDs above the mean) were removed. From the initial dataset of 4,320 tokens, 80 tokens (1.85%) were removed due to reaction times exceeding 3,500 ms, resulting in a final dataset of 4,240 tokens. Discrimination accuracy rate and reaction time data for the remaining 15 participants were submitted to two generalized linear mixed effects models as dependent variables using the lme4 package in R (Bates et al., 2015).

4. Results

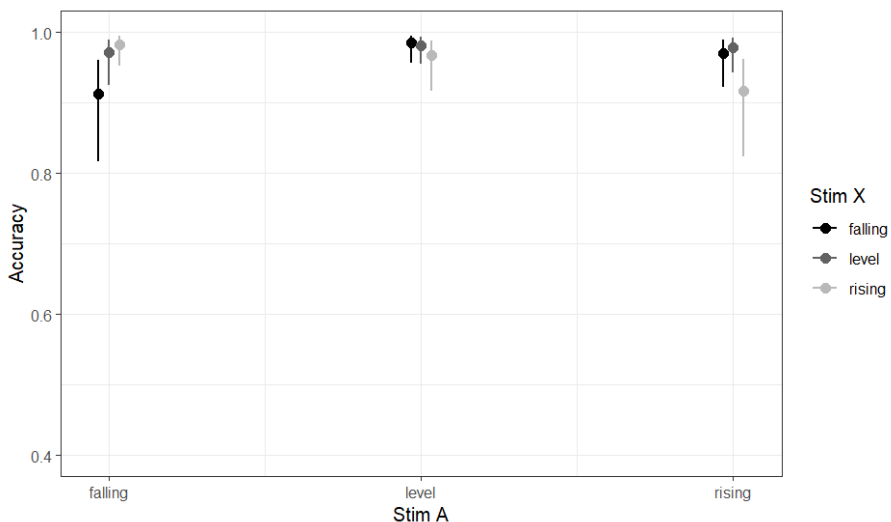
4.1 Accuracy

Mean accuracy rates and standard errors for each combination of first stimulus (Stim A) and second stimulus (Stim X) tones are shown in Figure 2. Stim A categories are on the x-axis and Stim X categories are represented by different shades: falling on the left black level in the center (dark grey), and rising on the right (light grey). Accuracy is generally higher when the first stimulus is level and lower when it is falling or rising, with

variations depending on the second stimulus. In addition, there is a lot more variability when the tone pairs are contour tones in the same direction (e.g., falling-falling or rising-rising).

Figure 2

Means and standard errors of accuracy for tone pairs. The x-axis displays the categories of Stimulus A (Stim A), including falling, level, and rising tones. The accuracy for Stimulus X (Stim X) is represented by filled circles, with falling tones shown in black, level tones in dark gray, and rising tones in light gray.



A generalized linear mixed-effects regression model using the *lmer* function from the *lme4* package (Bates et al., 2015) in R (R Core Team, 2024) was fitted to analyze the effects of inter-stimulus interval (ISI), correct response type (same vs. different), and tone categories (StimA_cat and StimX_cat) and their interactions on discrimination accuracy. Contrasting codings were completed using the *hypr* package (Rabe et al., 2020) and *emmeans* functions (Lent, 2024) from the *BruceR* package (Bao et al. 2024) were used for pair-wise post-hoc analyses.

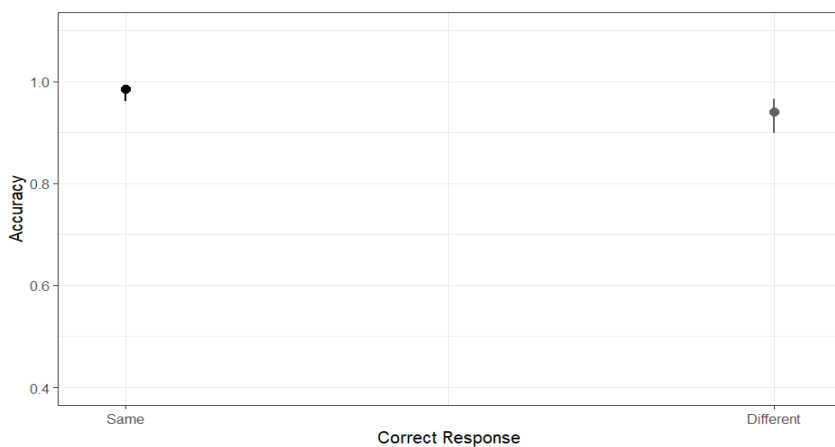
The model included random effects for subjects and tone pairs to account for variability not explained by the fixed effects. The formula for the model is given below:

```
model_tone_acc <- glmer(Accuracy ~ ISI + Correct_response + StimA_cat +  
StimX_cat + ISI:Correct_response + ISI:StimA_cat + ISI:StimX_cat +  
StimA_cat:StimX_cat + (1 | Subject) + (1 | Tone_pairs), data = tone, family =  
binomial, control = glmerControl(optimizer = "bobyqa"))
```

The results indicated that the stimulus interval (ISI) did not significantly affect accuracy, suggesting that memory load did not influence the participants' ability to identify tones correctly (Estimate = -0.041 , $SE = 0.133$, $t = -0.311$, $p = 0.756$). However, accuracy was significantly lower for 'different' trials compared to 'same' trials (Estimate = -1.363 , $SE = 0.423$, $t = -3.222$, $p = 0.001$), indicating that participants found it more challenging to accurately discriminate tone pairs when members of the pair had different pitch contours (see Figure 3).

Figure 3

Mean accuracy for 'same' and 'different' trials.



Interestingly, accuracy significantly improved when the first stimulus (Stimulus A) was a level tone compared to contour tone (falling or rising) (Estimate = 0.576, $SE = 0.277$, $t = 2.079$, $p = 0.038$), indicating that listeners found it easier to discriminate a level tone from a contour tone than vice versa. Additionally, a significant interaction between StimA_cat and StimX_cat was found for falling and rising tones, indicating a greater increase in accuracy when the first stimulus (Stim A) was a falling tone compared to a rising tone (0.914 vs. 0.909), relative to when the second stimulus (Stimulus X) was falling compared to rising (0.913 vs. 0.911) (Estimate = -2.777, $SE = 0.815$, $t = -3.409$, $p = 0.001$). In other words, the difference in accuracy between falling and rising tones was more pronounced when the falling tone was in the first position compared to when it was in the second position.

4.2 Reaction Times

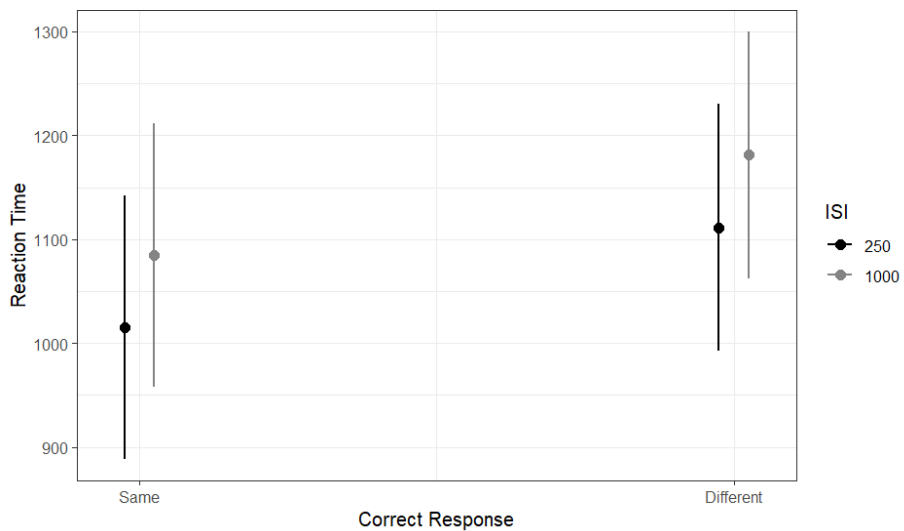
The second linear mixed-effects model tested the effects of ISI, response type (same vs. different), StimA_cat (first stimulus), StimX_cat (second stimulus), and their interactions on the continuous outcome variable reaction time. It also accounted for random intercepts for subjects, trials, and tone pairs. The formula of the model is given below:

```
model_tone_rt <- lmer(reaction_time~ISI + Correct_response + Accuracy +  
StimA_cat + StimX_cat+ISI:Accuracy + Correct_response:StimX_cat +  
Accuracy:StimA_cat + Accuracy:StimX_cat + StimA_cat:StimX_cat + (1|Subject)  
+ (1|Tone_pairs), data = tone, control = lmerControl(optimizer = "bobyqa"))
```

Figure 4 shows reaction times for the ‘same’ and the ‘different’ trials for the two ISIs. The results indicated that reaction times were significantly longer when the ISI was 1000 ms compared to 250 ms (Estimate = 69.72, $SE = 23.93$, $t = 2.914$, $p < 0.004$).

Figure 4

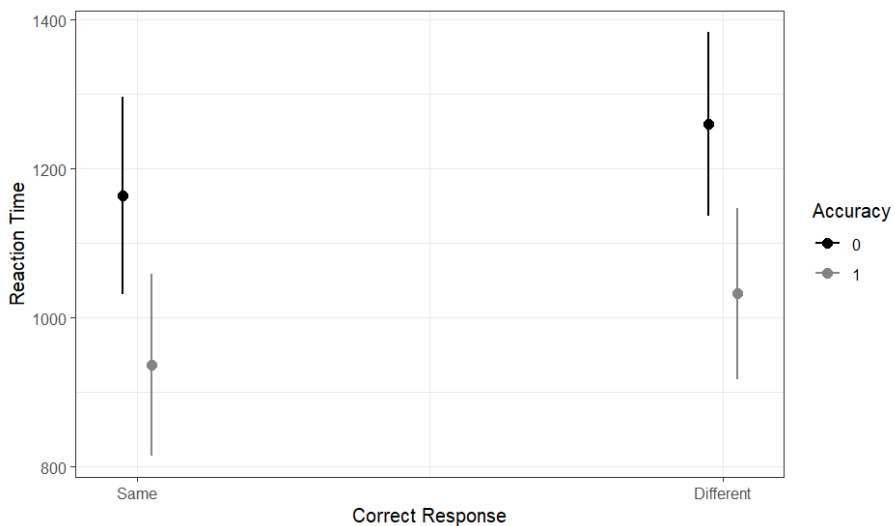
Mean reaction times in milliseconds (ms) and standard errors for ‘same’ (black) and ‘different’ (gray) responses across short (250 ms) and long (1000 ms) inter-stimulus intervals (ISI).



Reaction times were also significantly longer for ‘different’ trials compared to ‘same’ trials (Estimate = 96.11, $SE = 27.28$, $t = 3.523$, $p < 0.002$), indicating that participants were slower to identify trials as ‘different’ rather than the ‘same’. Interestingly, reaction times were also significantly faster for ‘correct’ compared to ‘incorrect’ responses (Estimate = -227.67 , $SE = 26.167$, $t = -8.703$, $p < 0.001$), as illustrated in Figure 5.

Figure 5

Mean reaction times in milliseconds (ms) and standard errors for ‘correct’ (1, indicated in gray) and ‘incorrect’ (0, indicated in black) responses across different stimuli.



In addition, reaction times were significantly faster when the first stimulus (Stim A) was a level tone than when it was a contour tone (falling or rising), with an estimate of -78.78 ($SE = 33.38$, $t = -2.360$, $p = 0.019$). Furthermore, reaction times were significantly slower when the first stimulus (Stim A) was a falling tone compared to when it was a rising tone (Estimate = 94.14 , $SE = 35.99$, $t = 2.616$, $p = 0.01$). These results are illustrated in Figure 6.

Figure 6

Mean reaction times in milliseconds (ms) for ‘same’ and ‘different’ trials when the first stimulus (Stim A) is categorized as falling (left, black), level (center, dark gray), and rising (right, light gray). The data points represent the accuracy of responses, with standard errors indicated.

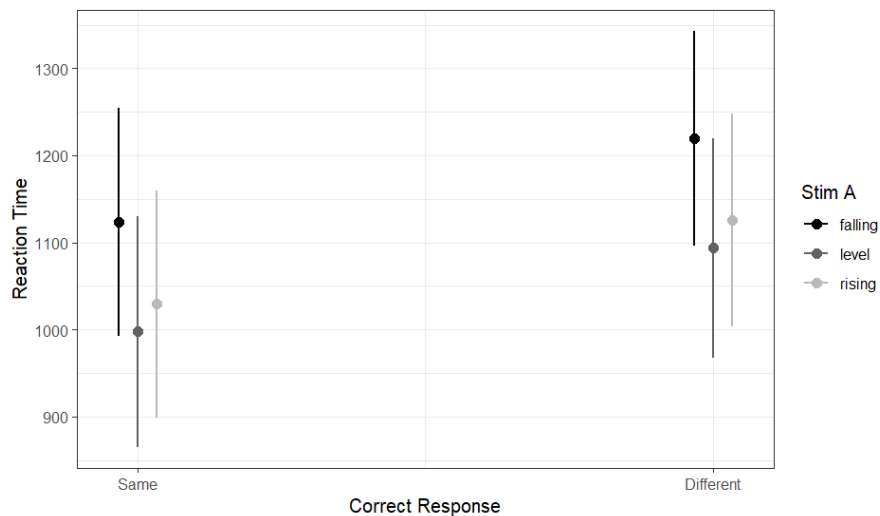


Figure 7 shows the reaction times for the ‘same’ and ‘different’ trials when the second stimulus (Stim X) is falling (left), level (center), and rising (right). The results showed a significant interaction between Correct Response and Stim X (Estimate = 174.321, $SE = 67.609$, $t = 2.578$, $p = 0.010$). When the second tone was falling, reaction times were significantly slower for ‘different’ trials compared to ‘same’ trials, with a difference of –218 ms. There was no difference in reaction time between the ‘same’ and ‘different’ trials when the second tone was rising or level.

Figure 7
Mean reaction times in milliseconds (ms) for ‘same’ and ‘different’ trials when the second stimulus (Stim X) is categorized as falling (left, black), level (center, dark gray), and rising (right, light gray).

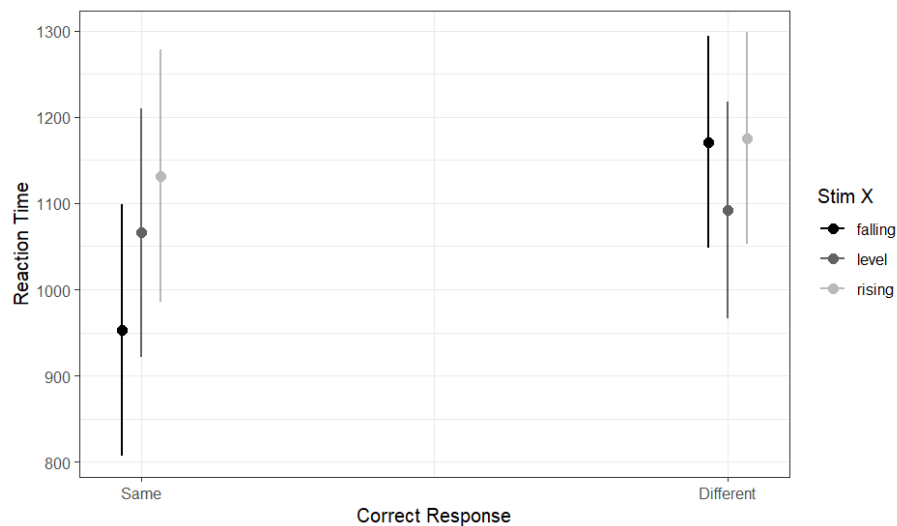
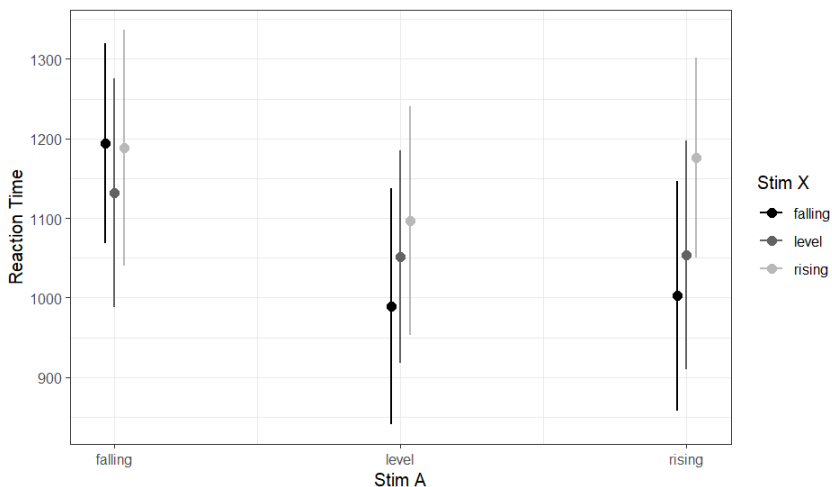


Figure 8 illustrates the reaction times for all pairings of the three pitch contours for Stim A and Stim X.

Figure 8

Mean reaction times in milliseconds (ms) for all pairings of the first = Stim A and the second = Stim X.



The results indicated that participants took longer to respond when the second stimulus (Stim X) was rising compared to falling (Estimate = -91.51, $SE = 41.89$, $t = -2.185$, $p = 0.03$), but there was no difference between level versus contour tones. Lastly, a significant interaction between StimA_cat and StimX_cat was found, indicating that reaction times were significantly longer when Stimulus A (first stimulus) was falling as opposed to rising (94 ms) compared to when Stimulus X (second stimulus) was falling rather than rising (-92 ms) (Estimate = 179.33, $SE = 67.09$, $t = 2.673$, $p < 0.01$).

5. Discussion

This study aimed to test the stimulus dynamics hypothesis, which was originally proposed to explain the widespread directional asymmetry patterns observed in vowel perception among both infants and adults. The Natural Referent Vowel (NRV) framework suggests that vowel perception asymmetries stem from a speech-specific bias toward peripheral vowels, which are typically produced with more exaggerated articulatory gestures and exhibit greater acoustic convergence of spectral energies (i.e., closer proximity between formants) compared to non-peripheral vowels (Masapollo et al., 2019). By this account, the convergence of spectral energies gives a peripheral vowel its characteristically well-defined spectral peaks and increases its auditory salience to which listeners are highly attuned (Masapollo et al., 2019). Alternatively, instead of proximity between vowel-adjacent formants, asymmetries may reflect a general auditory processing bias favoring a greater spectral change inherent in peripheral vowels (Masapollo et al., 2019).

According to this stimulus dynamics account, a peripheral vowel is more auditorily salient due to a more drastic shift in its spectral energy resulting from its extreme articulatory postures. This hypothesis was supported by the findings indicating that a change from a non-speech tonal analog of a vowel's second formant with less frequency modulation (i.e., a shallower falling slope) to one with greater frequency movement (i.e., steeper slope) was easier to detect than the reverse. Also consistent with the stimulus dynamic account, no asymmetries were found for the discrimination of flat tones differing in degrees of proximity or of falling tones differing in onset frequency but matched in slope. However, the stimulus dynamics hypothesis could not explain why no asymmetries were found for the discrimination of falling tones versus flat tones. It is possible that a ceiling effect contributed to the absence of directional perceptual asymmetries for this contrast.

In the present research, we extended the stimulus dynamics account to the discrimination of speech tonal stimuli among native speakers of American English. Specifically, we compared discrimination of level, falling, and rising tones resynthesized over a naturally produced, nonsense syllable [wa]. To prevent a ceiling effect, the slope for the rising and falling tones was much shallower (30 Hz versus 700-1,100 Hz) than that of Masapollo et al. (2019). In addition, the role of memory load, a factor not considered in previous studies, on the emergence of directional asymmetries was also investigated by comparing discrimination under a relatively shorter and a longer ISI condition. All possible pairings of the three tonal contours were investigated.

Unlike Masapollo et al.'s (2019) study, our listeners found a change from a flat tone to both a falling and a rising tone more discriminable compared to the reverse. This is consistent with the stimulus dynamics account, which posits that asymmetries are driven by a dynamic spectral change in the stimuli, such that a change from a less dynamic stimulus to a more dynamic stimulus is more salient than the reverse. Based on the stimulus dynamics account, this finding would imply that native English speakers are more perceptually attuned to falling and rising pitch contours than to level pitch contours. However, this contradicts the well-established observation that native speakers of non-tonal languages are more sensitive to pitch height and/or average pitch rather than pitch slope (Gandour & Harshman, 1978; Jongman et al., 2017). An alternative explanation for the observed asymmetries is that level tones serve as better reference points than contour tones for native American English speakers, as pitch height is more easily encoded than pitch contour. According to Rosch (1975), to be a “reference point” within a category, a stimulus must be shown to be one which other stimuli are seen “in relation to” (p. 532, original emphasis). Given the absence of tonal categories in their phonological representation, a psycho-acoustically salient flat tone with only F0 height information

might act as a strong anchor against which other pitch contours can be compared.

Asymmetries also emerged for the discrimination of contour tones. Accuracy increased when the first tone was falling compared to rising, suggesting that a falling tone may be a better reference than a rising tone. Given that the slope for both contour types was physically matched at 30 Hz, this result is inconsistent with the stimulus dynamics account because no asymmetry should emerge. Jongman et al. (2017) found that native English speakers show greater sensitivity to a rising contour than to a falling contour. However, a shallower slope of 15 Hz was used in the aforementioned studies instead of the 30 Hz used in the current study. Another possibility is that discrimination is easier when transitioning from a less salient to a more salient pitch contour. Since native English speakers are more sensitive to rising contours, they may find it easier to detect a change from the less salient falling contour to the more salient rising contour, rather than the reverse. However, the finding by Wayland et al. (2015) that a change from a rising contour to a falling contour, both with a slope of 15 Hz, was more discriminable than the reverse among both Mandarin and English listeners is inconsistent with this interpretation. Whether the difference in slope (15 Hz vs. 30 Hz) is responsible for the discrepancy remains to be resolved. Nonetheless, this finding suggests that asymmetries do not arise from degrees of spectral change between the two stimuli, but rather from the relative salience with which one is encoded by the listeners.

Neural discrimination of a change from the less focal, English-like synthetic /u/ vowel to a more focal, French-like /u/ deviant demonstrated in Zhao et al. (2019) may be reinterpreted to support this experience-dependent bias favoring stimuli with well-encoded spectral properties. Specifically, familiarity with the spectral properties of English /u/ allowed for better detection of the deviant French /u/ among native English

listeners. That is, regardless of degrees of spectral movement, a comparison between the first and the second stimulus is better made if the perceived salient characteristics of one (or both) stimuli are well encoded. Which acoustic properties are perceptually salient may vary across listeners, stimulus types, listening conditions, and listening tasks. Thus, better understanding of these factors and the role they play in the elicitation of asymmetries should be the focus of future research.

Regarding the ISI effects, the participants took longer to discriminate a stimulus change under the longer ISI condition (1000 ms) compared to the shorter ISI (250 ms), but their discrimination accuracy rates were comparable. Longer reaction times were also observed for ‘different’ trials compared to ‘same’ trials and for ‘incorrect’ responses compared to ‘correct’ responses. The ISI effects have been interpreted to reflect different modes or levels of processing. Specifically, it has been demonstrated that a longer ISI activates a higher, phonological level of processing, while a shorter ISI engages lower phonetic and auditory processing levels. A phonetic mode of processing refers to a language-general perception approach, where phones are distinguished purely based on their phonetic differences, without being influenced by the listener’s prior linguistic experience. In contrast, phonological processing is shaped by the listener’s prior experience with the phonological system of their language. Similar to the phonetic mode, the auditory mode of processing is a language-general approach to speech perception. However, unlike the phonetic mode, auditory-level discrimination relies on the acoustic variability of the stimuli rather than their phonetically relevant categories (Wayland & Li, 2008). According to the literature, the 250 ms ISI likely engaged the auditory mode of processing, whereas the 1000 ms ISI probably activated processing at the phonological level. Since pitch is not used to differentiate lexical meaning in English, contrastive pitch categories are likely absent from its phonological inventory. Therefore, the faster reaction time observed in the shorter ISI

condition was likely due to the reduced demand on listeners' working memory.

Shorter reaction times were observed when the first tone was a level tone as opposed to a contour tone (falling or rising). Considering this together with the result on accuracy, this finding further suggests that a level tone is a better reference tone than a contour tone for native English speakers.

6. Conclusions

Asymmetries in auditory perception of tonal contours in speech among native speakers of a non-tonal language emerged from a processing bias favoring salient stimulus properties encoded by the listeners, with little effect of processing demand imposed on working memory during processing. Future research involving different stimulus characteristics and groups of listeners, particularly native speakers of tonal languages under varying listening conditions, would offer further insights into the underlying processing mechanisms, constraints, and the specific information these processes rely on.

References

- Bao, H.-W.-S. (2024). *bruceR: Broadly useful convenient and efficient R functions* (Version 2024.6) [Computer software]. <https://CRAN.R-project.org/package=bruceR>
- Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Bent, T., Bradlow, A. R., & Wright, B. A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 32(1), 97.

- Francis, A. L., & Ciocca, V. (2003). Stimulus presentation order and the perception of lexical tones in Cantonese. *The Journal of the Acoustical Society of America*, 114(3), 1611-1621.
- Gandour, J. T., & Harshman, R. A. (1978). Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and Speech*, 21(1), 1-33.
- Guion, S. G., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. *Language experience in second language speech learning*, 57-77.
- Hienz, R. D., Sachs, M. B., & Sinnott, J. M. (1981). Discrimination of steady - state vowels by blackbirds and pigeons. *The Journal of the Acoustical Society of America*, 70(3), 699-706.
- Jongman, A., Qin, Z., Zhang, J., & Sereno, J. A. (2017). Just noticeable differences for pitch direction, height, and slope for Mandarin and English listeners. *The Journal of the Acoustical Society of America*, 142(2), EL163-EL169.
- Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2), 93-107.
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2), F13-F21.
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044), 606-608.
- Lenth, R. (2024). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.10.5, <https://doi.org/10.32614/CRAN.package.emmeans>

- Masapollo, M., Polka, L., & Ménard, L. (2017a). A universal bias in adult vowel perception—By ear or by eye. *Cognition*, 166, 358-370.
- Masapollo, M., Polka, L., Molnar, M., & Ménard, L. (2017b). Directional asymmetries reveal a universal bias in adult vowel perception. *The Journal of the Acoustical Society of America*, 141(4), 2857-2869.
- Masapollo, M., Zhao, T. C., Franklin, L., & Morgan, J. L. (2019). Asymmetric discrimination of nonspeech tonal analogs of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 45(2), 285.
- Plauché, M. C. (2001). Acoustic cues in the directionality of stop consonant confusions. University of California, Berkeley. <https://escholarship.org/uc/item/8hz6k4c2>
- Politzer-Ahles, S., Schluter, K., Wu, K., & Almeida, D. (2016). Asymmetries in the perception of Mandarin tones: Evidence from mismatch negativity. *Journal of Experimental Psychology: Human Perception and Performance*, 42(10), 1547.
- Polka, L., & Bohn, O. S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *The Journal of the Acoustical Society of America*, 100(1), 577-592.
- Polka, L., & Bohn, O. S. (2003). Asymmetries in vowel perception. *Speech Communication*, 41(1), 221-231.
- Polka, L., & Bohn, O. S. (2011). Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics*, 39(4), 467-478.
- Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2), 421.
- R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>

- Rabe, M. M., Vasishth, S., Hohenstein, S., Kliegl, R., & Schad, D. J. (2020).
hypr: An R package for hypothesis-driven contrast coding. *Journal of Open Source Software*, 5(48), 2134.
- Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, 7(4), 532-547.
- Schwartz, J. L., & Escudier, P. (1989). A strong evidence for the existence of
a large-scale integrated spectral representation in vowel perception.
Speech Communication, 8(3), 235-259.
- Swaminathan, J., Krishnan, A., & Gandour, J. T. (2008). Pitch encoding in
speech and nonspeech contexts in the human auditory brainstem.
Neuroreport, 19(11), 1163.
- Stevens, K. N. (1999). Acoustic phonetics. MIT Press.
- Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4), 327-352.
- Tsao, F. M. (2008). The effect of acoustical similarity on lexical-tone
perception of one-year-old Mandarin-learning infants. *中華心理學刊*,
50(2), 111-124.
- Wayland, R. P., & Li, B. (2008). Effects of two training procedures in cross-
language perception of tones. *Journal of Phonetics*, 36(2), 250-267.
- Wayland, R., Herrera, E., & Kaan, E. (2010). Effects of musical experience
and training on pitch contour perception. *Journal of Phonetics*,
38(4), 654-662.
- Xie, X., & Myers, E. (2015). The impact of musical training and tone language
experience on talker identification. *The Journal of the Acoustical
Society of America*, 137(1), 419-432.
- Yeung, H. H., Chen, K. H., & Werker, J. F. (2013). When does native language
input affect phonetic perception? *The precocious case of lexical
tone. Journal of Memory and Language*, 68(2), 123-139.
- Zhao, T. C., Masapollo, M., Polka, L., Ménard, L., & Kuhl, P. K. (2019). Effects
of formant proximity and stimulus prototypicality on the neural
discrimination of vowels: Evidence from the auditory frequency-
following response. *Brain and Language*, 194, 77-83.