Mapping Acoustic Cues to Pragmatic Functions: Perceptual Cue Weighting of Prosodic Focus in Mandarin

Wenxi Fei (wen-xi.fei@connect.polyu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR

Yu-Yin Hsu (yu-yin.hsu@polyu.edu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR

Abstract

Understanding how multiple acoustic dimensions are mapped onto linguistic representations is important in speech perception. This study explores how native Mandarin listeners process the communicative intentions of prosodic focus by examining the perceptual weightings of F0, duration, and intensity. Using a Visual World Paradigm, thirty native Mandarin participants listened to re-synthesized audio stimuli and responded to broad-focus or narrow-focus options. Results showed that the acoustic cues significantly influenced focus interpretation, with a greater reliance on F0 than intensity and duration. Eye-tracking data revealed perceptual divergence in the F0 condition, with the divergence of looks occurring at an earlier time window for acoustic processing and later for pragmatic processing. These findings suggest that native listeners effectively map acoustic variations to communicative demands, emphasizing the critical role of F0. The study highlights the temporal dynamics of interpreting prosodic focus, offering insights into language comprehension.

Keywords: Prosodic Focus; Pragmatic Processing; Perceptual Cue Weighting; Mandarin Perception

Introduction

Speech perception is a complex process that involves mapping acoustic signals into meanings in different languages. Traditional models of speech perception have primarily focused on the bottom-up processing of words (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), where listeners assign varying weights to multiple acoustic cues to signal speech categories in human perception, a concept known as *perceptual cue weighting* (Holt & Lotto, 2006; Toscano & McMurray, 2010). However, human language comprehension extends beyond mere word recognition. The interpretation of suprasegmental and pragmatic information is also crucial for successful communication (Cole, 2015; Grice, 1975). Despite this, it remains unclear how listeners infer pragmatic intentions from acoustic signals, particularly in tonal languages where prosodic features play a significant role. This study aims to address this gap by exploring how native Mandarin speakers weigh acoustic cues to perceive prosodic focus. It will shed light on the broader mechanisms of speech perception and the integration of pragmatic information.

Previous models of speech perception have described how listeners map acoustic signals to segmental categories, highlighting the comprehension of segmental components and word meanings (Holt & Lotto, 2006; Toscano & McMurray, 2010). These models have also proposed hypotheses on how and when listeners process multiple types of information. The *sequential processing hypothesis* proposes that the understanding of linguistic meanings occurs later than the processing of sensory and signal information. For example, models such as the TRACE model (McClelland & Elman, 1986) and the auditory sentence processing model (Friederici, 2002, 2011) suggest that initial acoustic processing precedes high-level integration of linguistic information in a sequential manner. In contrast, the *parallel processing hypothesis* suggests that multiple linguistic representations, including phonetic, phonological, syntactic, and semantic information, are processed simultaneously (Gibbs Jr & Colston, 2012; Pulvermüller, Shtyrov, & Hauk, 2009). Extending these discussions, the parallel activation model (Pickering & Strijkers, 2024) challenges these traditional views by proposing a two-stage approach. In the first stage, language-related components are activated simultaneously during speech prediction, while in the second stage, these activations are adjusted based on task demands. Unlike previous models, the parallel activation model allows for dynamic adaptations to different linguistic components based on contexts and task goals. However, despite these theoretical advancements, it remains unclear whether and which of these cognitive models, primarily designed to explain the processing of word meanings, can adequately account for the perception of pragmatic information, especially in the context of prosodic focus.

Suprasegmental features, such as prosody, are essential in conveying meaning in speech (Ladd, 2008; Pisanski & Bryant, 2019). Perception of prosodic information involves complex interactions between acoustic cues and linguistic contexts. Prosody is often realized through signal variations in fundamental frequency (F0), duration, and intensity (Cole, 2015). For example, in Mandarin production, speakers tend to use extended F0, longer duration, and higher intensity to present prosodic on-focus units in sentences (Xu, 1999). These acoustic features are important for signaling prosodic focus in speech perception, allowing listeners to extract corresponding communicative intentions. This process is defined as signal-based or bottom-up processing (Bishop, 2012). While previous research has shown that listeners can rely on these prosodic cues to interpret communicative intentions (reviewed by Cole (2015)), the mechanism of the relative weighting of these cues and the temporal dynamics remains unexplored.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Recent studies have begun to address these issues in English. For example, Steffman (2021) used a modified Visual World Paradigm to study the influence of prosodic focus on the processing of spectral signals in a word recognition task. The findings suggest that formant cues activate the lexical process at an early time window, and phrasal prosody is integrated later via lexical competition. There is also an early influence of prosodic focus, which shapes the earliest stages of formant use. This study has highlighted the role of prosodic information in processing segmental cues, but does not study the process of prosodic focus. Additionally, Jasmin, Dick, and Tierney (2021) have directly studied how listeners perceive prosodic focus by mapping varying acoustic cues to different focus positions. The results have shown that listeners can map acoustic cues to prosodic information, and different cues were weighed differently in focus perception. Although this study does not specifically address the time course of processing prosody, it presents the importance of understanding how various acoustic dimensions contribute to the perception of prosodic focus.

Overall, despite advances in speech perception research, it remains unclear whether the ability to map acoustic signals to phonemes and word meanings can be extended to the perception of suprasegmental and pragmatic information. The question also arises about whether processing pragmatic information aligns with the processing steps outlined in existing models of speech perception. The present study aims to investigate the perceptual mapping of acoustic signals to intentions of prosodic focus in Mandarin and the time course of this processing, using a modified Visual World Paradigm. Studying the Mandarin process is important because Mandarin listeners rely more on F0 information for word recognition than English listeners (Zeng et al., 2022). Thus, they may employ distinct cue-weighting strategies. These differences offer valuable insights into cross-linguistic variations in prosodic focus perception and processing. Specifically, we investigate the relative contributions of the acoustic cues (i.e., F0, duration, and intensity) in perceiving prosodic focus in Mandarin. The research questions are:

1. How do native listeners weigh prosodic cues when identifying focus intentions in Mandarin?

2. How is pragmatic information conveyed by prosodic focus processed over time?

Methods

Participants

This study recruited 30 college students from The Hong Kong Polytechnic University, aged from 18 to 28 ($M = 22.17$, $SD = 2.90$), with gender counterbalanced. All participants were native speakers of Mandarin Chinese from northern China, right-handed, and had normal hearing and reading abilities. Before the start of data collection, informed consent was obtained from each participant. After they completed the experiment, participants were compensated with about 13 USD.

Materials

Ten unique sentences were created for recording. Since this study did not aim to investigate the effects of tone types, all syllables in the sentences were Mandarin Tone 1. Each sentence was in a Subject-Verb-Object structure, consisting of three words and five syllables, e.g., crow eats watermelon. To ensure consistency in the complexity of these sentences, we measured the log base 10 (log10) word frequency ($M = 3.24$, $SD = 0.71$) and log10 character frequency ($M = 3.92$, $SD = 0.66$) using the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). Twenty college students who did not participate in the main experiments were invited to rate the comprehensiveness of the experimental sentences from 0 to 100. The comprehension scores were all above 80 ($M = 92.74$).

A native female Mandarin speaker recorded each sentence in two versions of focus: a. **Narrow focus on the verb** (crow EATS watermelon), with focus placed on the verb syllable, corresponding to "What does crow do to watermelon?" and b. **Broad focus** (crow eats watermelon), corresponding to "What happened?". In the recorded sentences, on-focus verbs in the narrow-focus condition exhibited significantly higher F0, longer duration, and greater intensity compared to the broad-focus condition; those acoustic values of post-focus units in the narrow-focus condition were significantly lower than those in the broad-focus condition, reported by independent *t*-tests (all ps $< .05$), aligned with the post-focus compression in Mandarin reported in the previous studies (Xu, 1999). To facilitate eye-movement processing, the first two syllables were time-normalized to 600 ms, ensuring the verb onset consistently occurred 600 ms after sentence onset.

The stimulus manipulations were based on those recorded audios. A seven-step continuum was created using WORLD vocoder (Morise, Yokomori, & Ozawa, 2016; Kawahara & Morise, 2024). The seven steps were 0%, 16.67%, 33.33%, 50%, 66.67%, 83.33%, and 100%. Among them, 0% means the F0 contour, duration, or intensity are from the "broad focus" recording (step 1), and 100% means that these acoustic features are from the "verb focus" recording (step 7). The mean F0 (Hz) steps of the verb syllables across ten sentences from step 1 to 7 were 283, 292, 302, 313, 324, 336, 349; duration (ms) steps were 793, 812, 831, 850, 870, 889, 908; intensity (dB) steps were 79.23, 80.02, 80.88, 80.94, 81.03, 81.08, 81.14. The acoustic values of the post-focus syllables followed an opposite trend, with step 1 having the highest value due to post-focus compression in Mandarin (Xu, 1999). To isolate the effects of each acoustic cue, when one cue was manipulated from step 1 to 7, the other two cues were kept at step 4 (50%), which was assumed to be the most ambiguous step (Jasmin et al., 2021). In total, 210 stimuli were created (3 cues $\times$ 7 steps $\times$ 10 sentences), and none of the participants in this study reported the unnaturalness of stimuli.

Procedure

The current study adopted a modified Visual World Paradigm with two response options, following Steffman (2021),

which could facilitate a comparison with the previous results of prosodic processing. The participants sat in front of a 1024×768-pixel computer screen in a sound-attenuated booth and listened to auditory stimuli from the AKG Pro Audio K77 Studio Headphones. Eye movements were recorded from the participants' right eyes using the Eyelink 1000 Plus system, which offers a high spatial resolution ($< 0.5^\circ$) and a sampling rate of 1000 Hz. The experiment was set up using Experiment Builder (2020). All text stimuli were displayed in Simplified Chinese characters, in black on a light gray background, using the Songti SC font (28-point size).

Before the start of the formal experiments, participants were instructed to minimize head movements during the experiments. Each participant was asked to rest his or her head on a chin and forehead support positioned at a distance of 80 cm from the screen. A nine-point calibration and validation procedure was applied to ensure good accuracy and precision of data ($M \leq 0.5^\circ$, $SD \leq 1.0^\circ$). Then, the participants were instructed to do six practice trials following the same procedure as the task trials. In the main experiment, each trial began with a drift correction fixation at the center of the monitor. When participants had passed the experimenter's check on the checkerboard, they would see a black cross fixation (50 pixels × 50 pixels) at the center and hear a beep sound for one second. Consequently, a stimulus audio was played; meanwhile, the response options appeared on the left and right sides of the screen. Participants were asked to decide whether the stimulus responded to a "broad-focus" question or a "narrow-verb-focus" question by pressing the "f" or "j" keys on the keyboard, which corresponded to the spatial position of the response option. The display position of the question options was counterbalanced among participants. Two interest areas (400 pixels × 300 pixels) were defined around the target option, which was slightly larger than the options themselves to ensure that gaze was recorded near the target question. After a keyboard response was recorded, the system automatically moved on to the next trial with a one-second interval.

The whole experiment consisted of 226 trials, that is, 210 task items, 6 practice items, and 10 vigilance items. To avoid fatigue, after every 44 trials were completed, participants could have a rest or continue the experiment when ready. Thus, the total duration of the test session for each participant was approximately 30 minutes, including eye-tracker settings. The system recorded the listeners' responses and the proportion of looks to the two options for analyses.

Data analysis

The data exclusion was made before the formal analyses. First, the vigilance trials were checked and all participants passed the tests. Then, the trials were excluded if response times were shorter than the audio playback duration or absolute standardized residuals exceeding 3.5 standard deviations. For eye-tracking data, trials were excluded if more than 50% of the time bins contained blinks or exclusion-detected samples. After these steps, 98% of trials were kept for response analysis; 76% were kept for eye-tracking analysis.

For response data, we analyzed whether and how they were affected by different acoustic continuum steps and acoustic conditions, as well as the perceptual cue weightings of each acoustic condition. Bayesian logistic mixed-effects regression models were conducted using the *brms* package of R (Bürkner, 2017). The dependent variable was the selection of "broad-focus" questions and "verb-focus" questions, with the "broad focus" option coded as 0 and "verb focus" option as 1. The proportions of choosing "verb focus" by participant were also computed for visualization. The main effects were the continuum step and the acoustic condition. The continuum step was scaled and centered before analysis. The condition factor was sum coded, condition 1 meaning the comparison between F0 and the sum of duration and intensity, while condition 2 means the comparison between duration and the sum of F0 and intensity. The MCMC diagnosis was conducted to determine the random effect, the final model only included the random intercept of participants. After computing the posteriors, the Highest Density Interval (HDI) of posterior distributions was computed using *bayestestR* package (Makowski, Ben-Shachar, & Lüdecke, 2019) to present the uncertainty characterization of posterior distributions as Credible Interval (*CI*). If the *CI* of an effect does not include 0, it means the effect is credible. For the perceptual cue weightings, the Bayesian models were run separately in the three acoustic conditions to get the step estimates and thus show the perceptual sensitivity of each acoustic cue.

For eye-tracking data, eye fixations were time-binned by 20 ms intervals and extracted using Data Viewer (2020). The time window of analysis was selected from the onset of the whole sentences (-600 ms) to 1400 ms after the onset of target verbs, which was the very cues for the "verb-focus" option, based on the previous studies on sentence context and focus intention perception (Gao et al., 2024; Steffman, 2021). Given that both auditory stimuli are linguistically plausible, this study used fixation patterns to analyze cognitive preferences. The dependent variable was the proportion of fixations to the "verb focus" interest area. The proportion measures were normalized for one target over another; a positive preference value corresponded to a preference for "verb focus", while a negative preference value corresponded to a preference for "broad focus".

To explore the time course of processing each acoustic condition, we first analyzed whether there were divergences in processing different steps of each acoustic information using visualization. If the visualizations showed obvious divergences in one acoustic cue, the comparisons of step pairs were then conducted using Generalized Additive Mixed Models (GAMMs) to identify the exact time window for processing different levels of information. The model was fitted via the *bam()* function in the *mgcv* package (Wood, 2017) and the time windows of divergences were reported by the *itsadug* package (van Rij, Wieling, Baayen, & van Rijn, 2022). The dependent variable was a log-transformed standardized preference measure, using the transformation

Table 1: Bayesian model output of main effects for responses with a random intercept of participant.

	Estimate	L-95%*CI*	U-95%*CI*	$\hat{R}$	ESS
intercept	1.22	0.74	1.72	1.003	761
step	0.72	0.64	0.80	1.001	6652
condition 1 (F0 vs. *M* of dur and int)	-0.36	-0.46	-0.25	1.001	5835
condition 2 (dur vs. *M* of F0 and int)	0.15	0.05	0.25	1.001	6116
step : condition 1	1.19	1.06	1.31	1.001	5782
step : condition 2	-0.64	-0.75	-0.53	1.001	6252

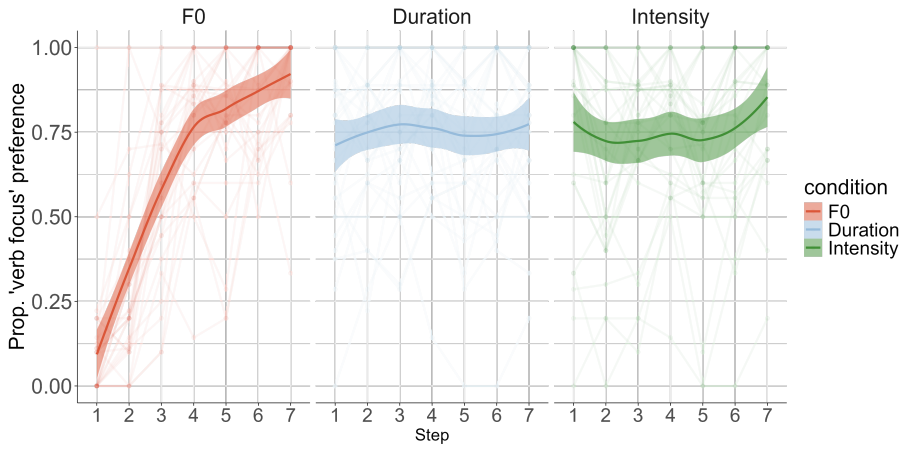

Figure 1: Response patterns by condition. Shade areas represent 95% Bayesian confidence intervals. Background lines show the individual patterns.

method as (Reinisch & Sjerps, 2013): $logit = ln\left(\frac{y+0.5}{n-y+0.5}\right)$, n denotes the total number of samples in a given time window and y denotes the number of samples in a given interest area. The GAMMs were run in each acoustic condition, with normalized continuum steps and smooth terms for time bins as the fixed effects. Random smooths were included for participant-level and item-level variability. Factor smooth interactions (*bs='fs'*) were specified.

Results and Discussion

The present study examined how native Mandarin listeners map the perceptual weightings of F0, duration, and intensity to prosodic focus when they were asked to identify pragmatic intentions. We found that listeners can weigh prosodic cues to different communicative intentions conveyed by prosodic focus in Mandarin, with F0 playing a critical role. Eye-tracking data showed that observable perceptual divergence occurred only in the F0 condition. Early divergence of looks mainly suggests initial processing of signal differences, while late divergence largely indicates late processing of pragmatic information.

Responses

Bayesian logistic mixed-effects models' results (Table 1) showed that all effects were credible, including the continuum step, acoustic condition, and their interaction. In particular, the continuum step credibly impacted focus-interpretation responses in the interaction with the F0 condition. Figure 1 shows that, as the continuum step increased, "verb focus" responses increased, which was also most visible in the F0 condition. Based on the model estimates, the results of perceptual cue weightings showed that listeners had a greater reliance on F0 ($B = 1.78, 95\%CI = [1.61, 1.96]$) than intensity ($B = 0.19, 95\%CI = [0.06, 0.32]$) and duration ($B = 0.08, 95\%CI = [-0.04, 0.20]$).

These response results have revealed that there is a notable rise in "verb focus" responses as the continuum step of acoustic features increases. It answers the question of whether the variations across multiple acoustic dimensions can be mapped onto language functions. The results suggest that pragmatic functions of prosodic focus can be signaled by multiple acoustic dimensions as the *perceptual cue weighting* at the segmental level, such as perceiving phonemes (Holt & Lotto, 2006; Toscano & McMurray, 2010). Our results also extend Jasmin et al. (2021) on simply identifying the on-focus units into the perception of pragmatic functions by asking the participants to map the answers to the question options with focus meanings embedded. The findings about the communicative intention comprehension on prosodic focus can be an addition to Gao et al. (2024), which has studied suggestion and warning in speech. Additionally, this perceptual cue weighting of the intentions of prosodic focus was

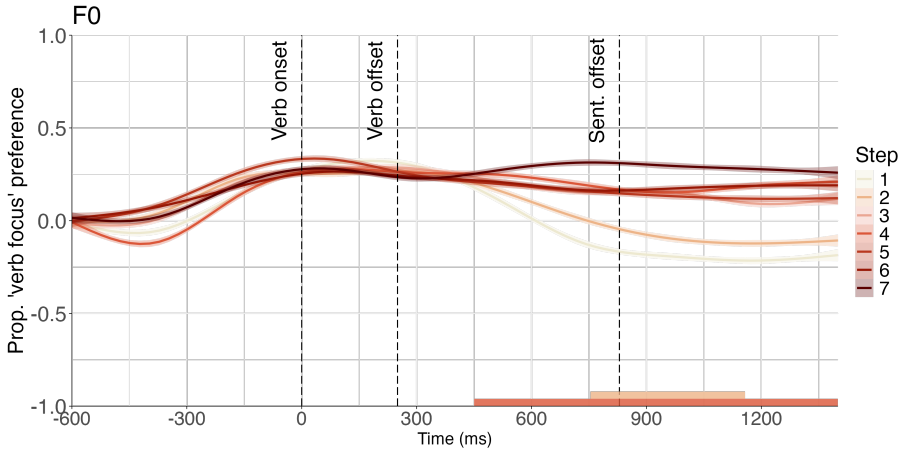

Figure 2: Eye movement data of the F0 manipulation condition. The colored horizontal lines indicate the significant time window reported by GAMMs between step pairs. "Verb offset" and "Sent. offset" represent mean offsets of verbs and sentences.

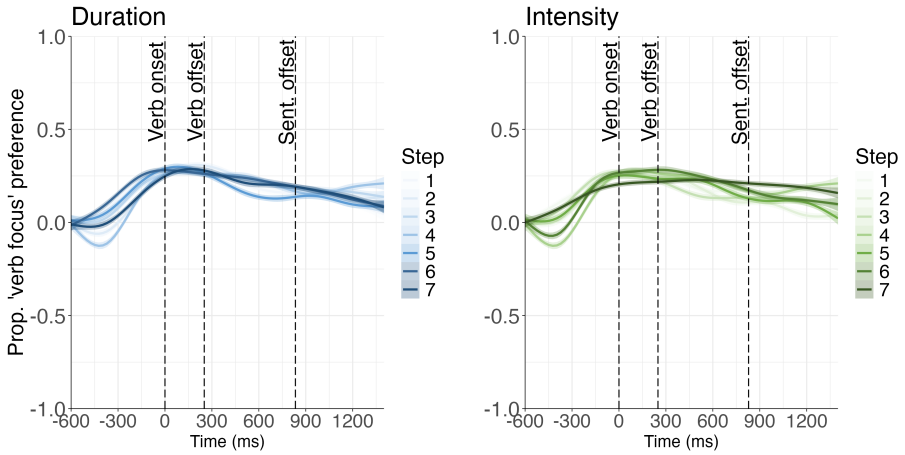

Figure 3: Eye movement data of the duration and intensity manipulation conditions.

most salient in the F0 condition, and further analysis indicates that F0 holds a significantly higher estimate of perceptual cue weighting compared to intensity and duration. This finding echoes the previous results about the weighting of each acoustic cue in Mandarin prosody from word-level perception (Zeng et al., 2022).

Eye-tracking

Eye-tracking data showed that the robust perceptual divergence for different steps appeared only in the F0 condition (Figure 2). Though there were no such obvious divergences in the duration and intensity conditions, it seemed that the processing of focus intention was more sensitive to intensity than duration (Figure 3), in line with the estimates of perceptual cue weightings.

We selected two pairs in the F0 condition for targeted analysis. The 1-7 pair was most acoustically different from each other (ΔF0 = 66 Hz), and it might also contain the processing of pragmatic information. Thus, this pair had the highest possibility of having the processing divergence at an early time window and a late one, representing the perception of both signal differences and pragmatic intention differences, respectively. The 2-3 pair was the acoustically ambiguous one (ΔF0 = 10 Hz) but showed the most robust pairwise difference in response and visualization results (Figure 1 and 2). Thus, it was used to check whether the pragmatic processing happened even with minimal evidence from acoustic signals. In step comparisons shown in Figure 4, the significant divergences between step 1 and 7 appeared from around 451 ms. For the comparison between step 2 and 3, there are significant divergences from about 754 ms to 1158 ms after the onset of verb syllables. The step comparison between the acoustically ambiguous pair indicates that the pragmatic information was not involved in the early processing together with the acoustic information, but occurred independently later.

The results have aligned with Steffman (2021) that the process of prosodic focus usually happens at a late stage of language comprehension, which then extends the previ-

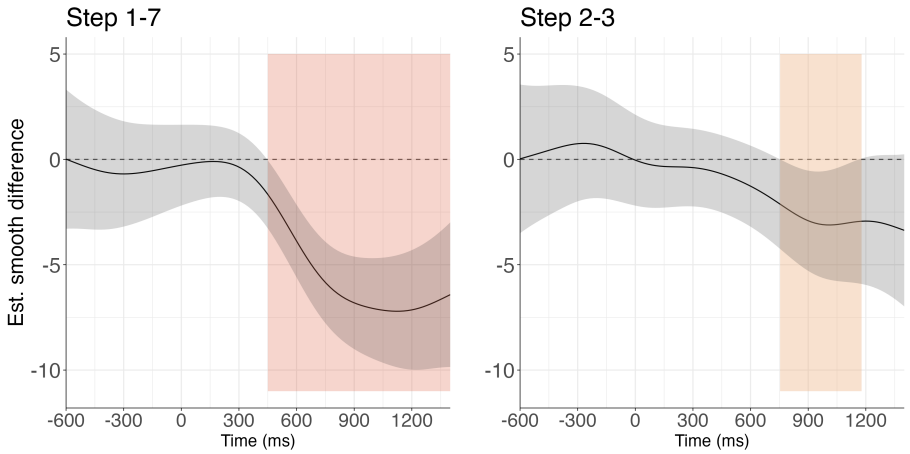

Figure 4: Comparisons of step 1-7 and step 2-3 of F0. The colored rectangles indicate the significant time window reported by GAMMs between step pairs, corresponding to the colored horizontal lines in Figure 2.

ous speech perception models on word recognition to the pragmatic level. The separate stages of processing acoustic and pragmatic information shown in the present findings are more likely to be explained by the *sequential processing hypothesis* (Friederici, 2011; Grice, 1975; Pickering & Garrod, 2013). Nevertheless, since the traditional *parallel processing hypothesis* does not include the description at the pragmatic level (Gibbs Jr & Colston, 2012; Pulvermüller et al., 2009), it could also be possible that our findings present an additive stage of pragmatic processing with high cognitive load after the parallel processing of the target which carries other language-related information, i.e., phonological, semantic, etc. Thus, further studies are needed to examine these possibilities by exploring the role of pragmatic function in language comprehension.

Besides, the results showed that the overall time window for processing the two stages of information was about 200 ms later than Steffman (2021), in which the acoustic processing starts from 270 ms and the prosodic-focus processing starts from 482 ms. This delayed process indicates that listeners require more processing time to integrate acoustic signals with communicative intentions compared to the word recognition task. This finding can be understood through two complementary theoretical frameworks. The parallel activation model (Pickering & Strijkers, 2024) suggests that specific task demands and cognitive load lead listeners to prioritize sentence and discourse context over acoustic differences. According to the task-specific prediction stage of this model, listeners may quickly reallocate their attention to the high-demanding task of inferring pragmatic functions rather than focusing on the early sensory detection of acoustic signals, which happens at about 50-300 ms. This model also aligns with the *associated view* of word recognition (Kim, Tremblay, & Cho, 2024), which proposes that real-time processing of early acoustic information becomes less continuous when early cues are less informative than later cues. This explains our finding of similar look patterns across different F0 steps during the initial time window. As early acoustic cues provide limited information for focus perception, listeners shift toward processing higher-level pragmatic information, resulting in delayed perceptual divergence. This processing shift, combined with the effort required for pragmatic information processing, suggests a complex interaction between acoustic and pragmatic cue processing, cognitive load, and attention allocation (Lialiou, Harris, Grice, & Schumacher, 2024). Future studies are needed to develop a more comprehensive model of prosodic processing that accounts for these factors.

Conclusion

This paper investigates the perceptual weighting of prosodic cues in Mandarin, highlighting the critical role of F0 in interpreting the communicative intentions behind prosodic focus. Eye-tracking data revealed that the robust perceptual divergence occurred only in the F0 condition. The early divergence in the acoustically-different pair indicates both early processing of signal differences and later integration of pragmatic intentions. The late divergence in the acoustically-ambiguous pair suggests that pragmatic processing can occur even with limited acoustic evidence. The findings generally suggest a complex process of mapping acoustic cues to pragmatic functions with the considerations of cognitive load, attention allocation, and task-specific mechanisms. By studying how prosodic cues are processed in Mandarin, this study offers implications for the cognitive processes underlying prosodic focus interpretation and theories of language comprehension. These insights can further inform the development of more accurate speech-processing technologies and targeted clinical interventions for language disorders. Future studies can consider individual factors and other speech perception tasks to further clarify language processing mechanisms.

Acknowledgments

This project is supported by the General Research Fund of the Research Grants Council of Hong Kong (GRF: 15612922). We would also like to thank the anonymous reviewers for their feedback and suggestions.

References

Bishop, J. (2012). Information structural expectations in the perception of prosodic prominence. *Prosody and Meaning*, *25*, 239.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

Cai, Q., & Brysbaert, M. (2010). Subtlex-ch: Chinese word and character frequencies based on film subtitles. *PloS one*, *5*(6), e10729.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1-2), 1–31.

Data Viewer. (2020). *Sr research experiment builder.* SR Research Ltd.

Experiment Builder. (2020). *Sr research experiment builder.* SR Research Ltd.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, *6*(2), 78–84.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357–1392.

Gao, P., Jiang, Z., Yang, Y., Zheng, Y., Feng, G., & Li, X. (2024). Temporal neural dynamics of understanding communicative intentions from speech prosody. *NeuroImage*, *299*, 120830.

Gibbs Jr, R. W., & Colston, H. L. (2012). *Interpreting figurative meaning*. Cambridge University Press.

Grice, H. P. (1975). Logic and conversation. *Syntax and semantics*, *3*, 43–58.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

Jasmin, K., Dick, F., & Tierney, A. T. (2021). The multidimensional battery of prosody perception (mbopp). *Wellcome Open Research*, *5*, 4.

Kawahara, H., & Morise, M. (2024). Interactive tools for making vocoder-based signal processing accessible: Flexible manipulation of speech attributes for explorational research and education. *Acoustical Science and Technology*, *45*(1), 48–51.

Kim, H., Tremblay, A., & Cho, T. (2024). Perceptual cue weighting matters in real-time integration of acoustic information during spoken word recognition. *Cognitive Science*, *48*(12), e70026.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Lialiou, M., Harris, J., Grice, M., & Schumacher, P. B. (2024). Attention allocation to deviants with intonational rises and falls: Evidence from pupillometry. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive psychology*, *18*(1), 1–86.

Morise, M., Yokomori, F., & Ozawa, K. (2016). World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, *99*(7), 1877–1884.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329–347.

Pickering, M. J., & Strijkers, K. (2024). Language production and prediction in a parallel activation model. *Topics in Cognitive Science*.

Pisanski, K., & Bryant, G. A. (2019). The evolution of voice perception. *The oxford handbook of voice studies*, 269–300.

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain and language*, *110*(2), 81–94.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116.

Steffman, J. (2021). Prosodic prominence effects in the processing of spectral cues. *Language, Cognition and Neuroscience*, *36*(5), 586–611.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434–464.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2022). *itsadug: Interpreting time series and autocorrelated data using gamms.* (R package version 2.4.1)

Wood, S. (2017). *Generalized additive models: An introduction with r* (2nd ed.). Chapman and Hall/CRC.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0contours. *Journal of phonetics*, *27*(1), 55–105.

Zeng, Z., Liu, L., Tuninetti, A., Peter, V., Tsao, F.-M., & Mattock, K. (2022). English and mandarin native speakers' cue-weighting of lexical stress: Results from mmn and ldn. *Brain and language*, *232*, 105151.