Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Native and Non-native listeners' Ability in Integrating Prosody and Verb Semantics in Mandarin Speech Comprehension under the Impact of Language-specific Prosodic System

Xiaomu Ren and Clara Cohen

Glasgow University Laboratory of Phonetics, University of Glasgow
x.ren.1@research.gla.ac.uk

Abstract

This visual-world eye-tracking study examined the integration of prosody and verb semantics integration in native and non-native Mandarin speech perception. Native Mandarin and English listeners' eye movements were recorded and they were asked to click on the object named in the second sentence while they listened to Mandarin sentences within a brief discourse context. These sentences varied in three binary factors: prosodic accent, target object information status, and semantic match between verb and target object. Results showed a complex interaction between L1, prosody, and verb semantics. When processing both old and new information, Mandarin listeners interactively combined high-level verb semantic cues with low-level prosodic cues across diverse contexts, such that effects of verb semantics were exaggerated under contrastive focus. English listeners showed a similar pattern with old information as Mandarin listeners, but with new information they showed less integration of different speech cues than Mandarin listeners. These results suggest that non-native listeners can adopt native-like strategies in integrating prosodic and semantic speech cues, but only when the information status of target words is familiar, and hence less taxing to process.

Index Terms: Prosody, Verb semantics, Information structure, Cue integration, Online Mandarin speech comprehension, Eye-tracking paradigm

1. INTRODUCTION

Speech comprehension requires listeners to integrate low-level acoustic-phonetic elements with more high-level linguistic-semantic aspects [1]. Low-level acoustic-phonetic processing handles segmental cues, such as encoding of consonants and vowels [2], as well as suprasegmental cues, such as stress and intonation information [3, 4, 5]. Higher-level comprehension processes focus on the extraction of meaning from lexical, semantic, syntactic and discourse-pragmatic information. The ways in which these sorts of information can be combined during real-time speech comprehension is endlessly varied: listeners can draw on low-level prosodic cues to make sense of information structure [6, 7], disambiguate structural ambiguities [8, 9], and predict upcoming referents [10, 11, 7]. Furthermore, they can combine these levels of information interactively, [12], for example, observed that semantic processing of verbs is strengthened when those verbs are highlighted by a prosodic pitch accent.

Native listeners do all of this effortlessly [13]. What about non-native listeners? Within a single domain, it seems that they also can combine different types of information. For example, at higher levels of comprehension, non-native listeners can utilize morphosyntactic agreement and contextual semantic information to predict upcoming words [14, 15, 16, 17]. At lower levels, too, non-native listeners can integrate different types of information. They can use prosody to evaluate how natural pronunciation sounds, [18], or construct information structures [7]. However, across different levels of information, it is not so clear that non-native speech comprehension can integrate different types of cues so flexibly [19]. Evidence rather suggests that non-native listeners might prefer to privilege some types of information over others, such that they rely more on semantic information, for example, than on prosody [20, 21].

When non-native listeners do rely on low-level pitch accents for prediction, it seems to happen more when their native language aligns closely with their L2's accentuation patterns [22, 23]. When the L1 and L2 patterns do not align, low-level prosodic information may be less useful. For example, English and Mandarin contrast in conveying focus: English utilizes the L + H* pitch accent to indicate contrastive focus, while the H* pitch accent more broadly denotes various types of novel information [24, 25]. Mandarin, by contrast, does not use this sort of accentuation to indicate contrastive focus, but instead employs an extended pitch range and prolonged duration [26, 27, 28]. Mandarin pitch further serves a dual role, conveying both lexical meaning for tones alongside its intonation cues for contrastive focus [29]. This makes the mapping between prosodic features for contrastive focus and their phonological function less clear than in English [30]. Therefore, the language-specific prosodic structure suggests Mandarin and English listeners may employ distinct prosodic perceptual strategies when processing Mandarin [30]. Studies indicate that non-native listeners who prioritize distributed prosodic cues such as F0 over local cues like duration and intensity in their native languages, tend to perform better in non-native prosodic perception than those non-native listeners who prioritize local cues over distributed cues [31]. Earlier research suggested that Mandarin listeners use distributed cues like F0 for Mandarin focus perception, while native English listeners rely on local cues like duration and intensity cues during English speech comprehension [32]. Hence, English listeners may face challenges when dealing with non-native speech perception, as when listening to Mandarin.

The research reviewed above suggests that English listeners of Mandarin will, first, rely more on semantic information than prosodic information, and second, will integrate semantic and prosodic information less flexibly than Mandarin listeners. To test these predictions, we designed a visual-world eye-tracking experiment, in which listeners were asked to view arrays of images while listening to sentence pairs, and select the image that was named in the second sentence. These nouns were presented in auditory contexts that were manipulated across three types of speech cues: prosody, information status, and verb semantics. Prosody and information status test listeners' attention to low-

10.21437/SpeechProsody.2024-182

level prosodic information, as the expected prosodic patterns will differ depending on whether a target word is old or new information. Verb semantics tests listeners' attention to high-level semantic information.

If non-native English listeners of Mandarin exhibit a reduced ability in utilizing low-level prosodic cues than Mandarin listeners, we anticipate a diminished impact of prosody on their ability to recognize a target word in comparison to native listeners. Furthermore, when comparing the relative strength of prosodic and semantic cues, English listeners should show a reduced advantage or increased disadvantage of prosodic information over semantic information, relative to Mandarin listeners. Finally, if English listeners are also less able to integrate the high and low level information interactively, then Mandarin listeners should show increased semantic processing when target nouns are accented [12], while English listeners should show reduced evidence of this pattern.

2. EXPERIMENTAL METHOD

2.1. Participants

Participants consisted of 40 Mandarin native listeners and 27 English native listeners learning Mandarin as L2, all over age of 18 and students at the University of Glasgow. The English listeners had at least intermediate Mandarin proficiency, having completed an intermediate course at the university or lived in a Mandarin-speaking environment for over two years.

2.2. Materials

Participants were presented with a visual display containing four pictures: a target noun (e.g. *kōng tiáo*, "air conditioner"), a competitor (e.g. *kōng jiě*, "stewardess"), and two distractors (e.g. *wū guī* , "tortoise", *lán zi*, "basket").

All sentences were in Mandarin; each trial consisted of an instruction composed of two sentences. Each trial was designed to manipulate three binary variables: Information Status (Repeated/New), Verb Appropriateness (Appropriate/Inappropriate), and Prosody (Contrastive/Neutral).

For the Information Status variable, in New sentences, the competitor was named in the first sentence and the target named in the second sentence:

(1) nǐ kě yǐ kàn dào yī xiē kōng jiě. xiàn zài xiǎo yǔ yào qù ān zhuāng kōng tiáo zhè yàng xià tiān jiù biàn dé liáng kuài le.

You can see some stewardesses. Now Xiaoyu is going to install the air conditioner so that the summer will become cool.

In Repeated sentences, the target was named in both sentences:

(2) nǐ kě yǐ kàn dào yī gè kōng tiáo. xiàn zài xiǎo yǔ yào qù ān zhuāng kōng tiáo zhè yàng xià tiān jiù biàn dé liáng kuài le.

You can see an air conditioner. Now Xiaoyu is going to install the air conditioner so that the summer will become cool.

For the Verb Appropriateness variable, in Appropriate sentences, the verb in the second sentence could only apply to the target as a direct object (e.g., *"install the air conditioner"*). In Inappropriate sentences, the verb could not plausibly apply to the target (e.g., *"ask the air conditioner"*).

For the Prosody variable, in Contrastive Focus sentences, a contrastive pitch accent is placed on the underlined target noun (e.g. *"air conditioner"* in sentences like *"Now Xiaoyu is going to install the **air conditioner** so that the summer will become cool"*). The target noun with a contrastive accent was pronounced to be more salient than other words in the sentence, resulting in a higher pitch range on this target noun. In Neutral-accented sentences, the prosody followed a more neutral topic-content melody, with any pitch accent placed on the final adverbial.

These three binary variables created 8 conditions for each of the 40 sentence-pair items. The items were rotated across 8 experimental lists using a Latin square design. Each list included 40 filler sentence pairs, strategically constructed to balance Information status, Verb appropriateness, and Prosody across all trials. Stimuli were presented in pseudo-random order, unique to each participant. All sentences were recorded by a phonetically-trained native Mandarin speaker in a sound-proofed booth.

2.3. Procedure and Analysis

Data were collected through the use of an Eyelink 1000+ eye-tracker, utilizing Experiment Builder software to sample gaze data at a rate of 1000Hz. Each session comprised calibration, two practice trials, and recalibration prior to the main experiment. Gaze data for target and competitor looks were organized into proportions over 50ms windows, spanning from 200 ms before target onset to 1800ms after.

Participants finished 8 blocks, each containing 10 trials, with breaks and recalibration interspersed between blocks. The participants' eye movements were recorded while they listened to sentences within a concise discourse context. Subsequently, within each trial, they were prompted to click on the object named in the second sentence.

Target advantage (TA) was computed by deducting competitor looks from target looks, constituting the dependent variable. The analysis of TA involved the use of generalised additive mixed models (GAMMs) with the `mgcv` package (version 1.8.4 [33]) in R (version 4.2.1; [34]). Repeated and new information sentences were analysed separately using simple GAMMs, encompassing parametric effects and difference smooths for each of the three binary main variables: Prosody (neutral/contrastive); Verb Appropriateness (appropriate/inappropriate); and L1 (Mandarin/English). Factors were treatment-coded, defaulting to Neutral Prosody, Appropriate Verb, and English L1. All models included three-way interactions between L1, Prosody, and Appropriateness, allowing for a direct examination of the prediction concerning potential differences in the interactive combination of Prosody and Appropriateness between English and Mandarin listeners.

3. Results

Figure 1 displays gaze traces for repeated and new target nouns, while Table 1 summarizes final models. Both conditions show three-way interactions in both parametric terms and smooths, with only the parametric terms are discussed here.

3.1. Repeated information

English listeners had a positive Prosody effect ($\beta = 0.056, p < .001$), indicating that a contrastive focus facilitated TA, while a negative effect of Appropriateness indicated that inappropriate verb semantics preceding the target noun reduced TA

($\beta = -0.038, p < .01$). These two effects were of very similar magnitude. Prosody and Appropriateness also interacted among English listeners ($\beta = -0.104, p < .001$), such that the reduction in TA associated with inappropriate verbs was exaggerated under contrastive focus.

Mandarin listeners consistently exhibited higher overall TA than English listeners ($\beta = 0.129, p < .001$). Furthermore, L1 did not interact with either Prosody ($\beta = -0.030, p = .11$) or Appropriateness effect ($\beta = -0.036, p = .05$), suggesting that Mandarin listeners benefited from the contrastive pitch accent and suffered from inappropriate verbs in similar ways to English listeners. The three-way interaction ($\beta = 0.041, p = .12$) indicates that Mandarin listeners did not differ from English listeners' in their two-way interaction between Prosody and Appropriateness: a stronger Appropriateness effect under contrastive prosody than neutral prosody.

Thus, for repeated information sentences, both groups were sensitive to low-level prosodic and high-level semantic information, and combined them interactively in similar ways.

3.2. New information

With new information sentences, English listeners again exhibited a negative effect of Appropriateness ($\beta = -0.15, p < 0.001$), suggesting that, in neutral prosody, an inappropriate verb led to a decrease in TA. They also showed a significant negative effect of Prosody, which went in the opposite direction from its effect with repeated information: here, contrastive focus reduced TA ($\beta = -0.036, p < .05$). Note, too, that, the size of the Appropriateness effect was three times larger than the size of the Prosodic effect (-0.15 TA units, compared to -0.036 TA units), suggesting that English listeners were more influenced by semantics than prosody. No interaction between accent and appropriateness was observed among English listeners ($\beta = 0.039, p = .07$).

Mandarin listeners again exhibited an overall higher TA than English listeners in processing new information sentences ($\beta = 0.079, p < .05$). The effect of Appropriateness did not differ significantly from English listeners ($\beta = -0.013, p = .51$), indicating that both groups showed reduced TA with inappropriate verbs. The effect of Prosody, however, did interact significantly with L1 ($\beta = 0.089, p < .001$), such that the overall negative effect of Prosody with English listeners was reversed to a positive effect for Mandarin listeners: that is, contrastive focus increased TA. Finally, although English listeners showed no interaction between Prosody and Appropriateness, a three-way interaction between those variables and L1 indicated that Mandarin listeners did ($\beta = -0.108, p < .001$). For Mandarin listeners, the negative appropriateness effect was magnified by the presence of contrastive focus prosody in a new information context in the same way it had been in the old information context.

In other words, English listeners remained sensitive to both Prosody and Appropriateness in processing new information sentences. Furthermore, they were three times as sensitive to Appropriateness as they were to Prosody. However, although the role of contrastive focus in deepening semantic processing persisted in Mandarin listeners, it disappeared for English listeners.

4. DISCUSSION AND CONCLUSION

The results of this study aligned well with some but not all of our predictions. First, we predicted that English listeners, as non-native listeners, would be less adept at using prosodic information than Mandarin listeners, who are native listeners. Second, we anticipated that English listeners would rely more on high-level semantic information than on low-level prosodic information. Third, we predicted that English listeners would integrate these two types of cues less interactively than Mandarin listeners in the context of Mandarin speech processing and comprehension.

When processing repeated information sentences, none of these predictions were confirmed. Rather, English listeners adopted a perceptual strategy similar to Mandarin listeners. Both groups showed similar effect sizes for Prosody and Appropriateness, and the use of contrastive focus magnified the effect of inappropriate semantics. This observation replicates the findings [12] for native listeners, and extends them to non-native listeners. Using contrastive prosody to highlight a target noun deepens semantic processing, showing an interactive combination of high and low level cues that we did not predict for non-native listeners.

In processing new information, however, English listeners behaved closer to our predictions, diverging in their strategies from native Mandarin listeners. They displayed a greater sensitivity to verb appropriateness than pitch accent, indicating their inclination as non-native Mandarin listeners to be more affected by high-level semantic information than low-level prosodic cues. English listeners also showed an opposite effect of low-level prosodic cues from Mandarin listeners. They did not benefit from the contrastive prosody signaling new information status; instead, they were adversely affected. Moreover, English listeners showed no interaction between these effects, suggesting no interactive integration of semantics and prosody in new information sentences. In contrast, Mandarin listeners, adopting a similar strategy to the repeated information context, demonstrated an interaction between the effects of verb appropriateness and pitch accent, with the target accent heightening challenges associated with processing inappropriate verb semantics.

Thus, the results reveal that English listeners, at least when processing sentences with repeated information, effectively integrate speech cues at different levels. Further, they do not always prioritize semantics over prosody, instead demonstrating the capability to utilize both high- and low-level cues. However, it is evident that English listeners do integrate speech cues differently and less flexibly compared to Mandarin listeners. When confronted with new information, English listeners exhibit a diminished ability to integrate these cues, and started to rely more on semantics than prosody.

These findings form an interesting contrast with those of a previous study [35], which was a similar experiment using English stimuli. Their results indicated that non-native Mandarin listeners processing English sentences still derived benefits from contrastive focus prosody even in new information contexts. The divergence may reflect differences between Mandarin and English listeners' use of prosodic accent for contrastive focus. English listeners tend to rely more on local prosodic cues like duration and intensity [31]. This can cause more challenges in online L2 prosodic comprehension for English listeners than for Mandarin listeners, who rely more on distributed prosodic cues like F0 [31]. Furthermore, Mandarin achieves contrastive focus by extending pitch range and prolonging duration, potentially distorting pronunciation of lexical tones [36]. When the nouns were repeated, non-native listeners could handle this distortion, since the target word was already highly activated by the earlier mention; but when the nouns

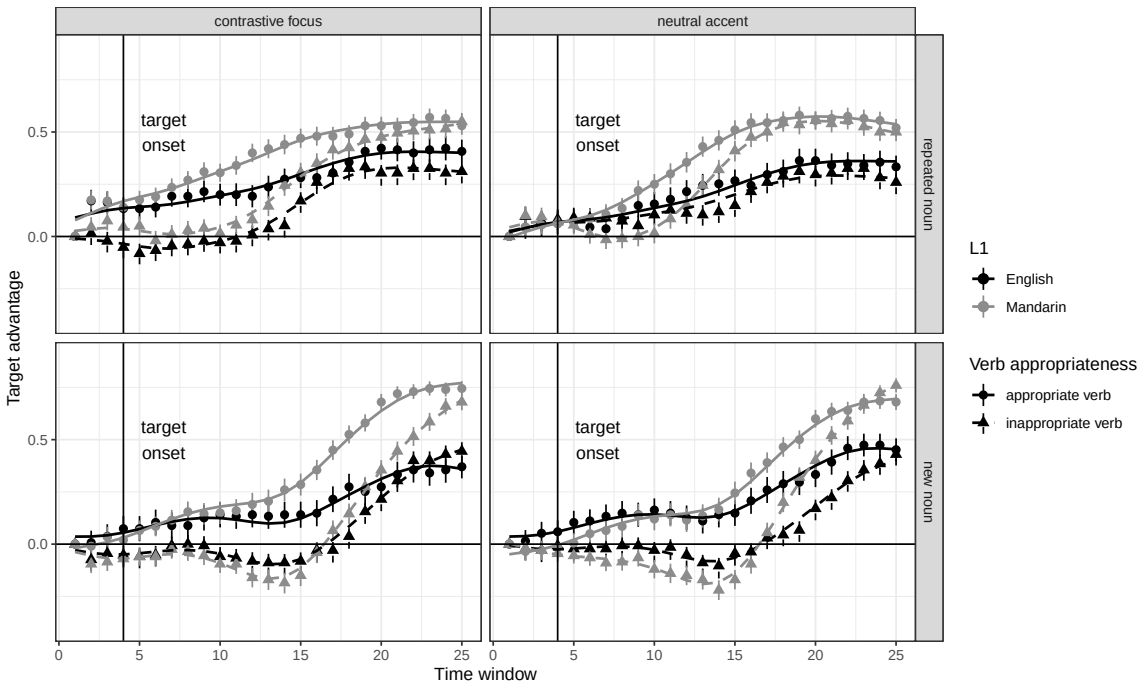

Figure 1: *Gaze traces with repeated and new nouns, showing target advantage (TA) across English (black line) and Mandarin (grey line) listeners, for appropriate (dotted circles, connected by solid lines) and inappropriate (solid triangle, connected by dashed lines) verbs semantics. Neutral accent is on the right, while Contrastive focus is on the left. Contrastive focus is the unexpected prosody for repeated information (top row), and the expected prosody for new target nouns (bottom row).*

Table 1: *GAMM models for sentences with repeated (left) and new (right) information. Levels for Prosody (abbreviated* pro*) are Neutral (abbreviated* neu*) and Contrastive focus (target accent, abbreviated* tar*). Levels for Appropriateness (*App*) are Appropriate (reference,* a*) and Inappropriate (*i*). Levels for L1 are English (reference,* eng*) and Mandarin (*man*). Difference smooths are calculated on an 8-level factor representing the interaction between Acc, App, and L1.*

Repeated information					**New information**				
Parametric	Est.	SE	t	p	*Parametric*	Est.	SE	t	p
Intercept	0.214	0.029	7.46	$< .001$	Intercept	0.202	0.030	6.67	$< .001$
Pro=tar	0.056	0.014	3.88	$< .001$	Pro=tar	-0.036	0.015	-2.39	$< .05$
App=i	-0.038	0.014	-2.64	$< .01$	App=i	-0.15	0.015	-9.97	$< .001$
L1=man	0.129	0.037	3.52	$< .001$	L1=man	0.079	0.038	2.06	$< .05$
Pro=tar:App=i	-0.104	0.020	-5.11	$< .001$	Pro=tar:App=i	0.039	0.021	1.84	.07
Pro=tar:L1=man	-0.030	0.019	-1.61	.11	Pro=tar:L1=man	0.089	0.019	4.60	$< .001$
App=i:L1=man	-0.036	0.019	-1.95	.05	App=i:L1=man	-0.013	0.019	-0.65	.51
Pro=tar:App=i:L1=man	0.041	0.026	1.55	0.12	Pro=tar:App=i:L1=man	-0.108	0.027	-3.92	$< .001$
Difference smooths		edf	F	p	*Difference smooths*		edf	F	p
Window		4.41	8.55	$< .001$	Window		6.86	16.93	$< .001$
Win:Pro=tar,App=a,L1=eng		1.00	0.37	.55	Win:Pro=tar,App=a,L1=eng		1.29	1.79	.10
Win:Pro=neu,App=i,L1=eng		1.01	3.23	.07	Win:Pro=neu,App=i,L1=eng		3.26	6.27	$< .001$
Win:Pro=tar,App=i,L1=eng		3.77	4.32	$< .001$	Win:Pro=tar,App=i,L1=eng		3.34	7.20	$< .001$
Win:Pro=neu,App=a,L1=man		3.62	5.78	$< .001$	Win:Pro=neu,App=a,L1=man		1.00	14.78	$< .001$
Win:Pro=tar,App=a,L1=man		1.96	1.80	.12	Win:Pro=tar,App=a,L1=man		1.00	19.19	$< .001$
Win:Pro=neu,App=i,L1=man		5.63	9.92	$< .001$	Win:Pro=neu,App=i,L1=man		5.14	32.93	$< .001$
Win:Pro=tar,App=i,L1=man		4.22	6.94	$< .001$	Win:Pro=tar,App=i,L1=man		4.75	22.34	$< .001$
Window, by subj		302.85	5.77	$< .001$	Window, by subj		292.12	5.04	$< .001$

were new, the distorted tones may have increased processing difficulty, because it's a speculation that the tones are the reason.

In sum, non-native listeners' ability to utilize low-level prosodic cues, and integrate them with high-level semantic information, is reduced relative to native listeners. This pattern, however, only emerges under certain conditions—here, when lexical recognition is made more difficulty by the introduction of new information in the discourse. When the task of processing sentences is less challenging, non-native listeners can in-

tegrate low-level and high-level information similarly to native listeners. Future investigation could enrich the knowledge of how multiple speech cues are utilized and integrated under both tonal and non-tonal language contexts, particularly focusing on listeners' performance when processing sentences with new information. This direction would deepen insights into listeners' adaptive strategies and cognitive flexibility required to adapt to different linguistic systems.

5. References

[1] M. v. d. Ven, M. Ernestus, and R. Schreuder, "Predicting acoustically reduced words in spontaneous speech: The role of semantic/syntactic and acoustic cues in context," *Laboratory Phonology*, vol. 3, no. 2, pp. 455–481, 2012.

[2] X. Tong, C. McBride, C.-Y. Lee, J. Zhang, L. Shuai, U. Maurer, and K. K. H. Chung, "Segmental and suprasegmental features in speech perception in Cantonese-speaking second graders: An ERP study: Speech perception in Cantonese children: An ERP study," *Psychophysiology*, vol. 51, no. 11, pp. 1158–1168, 2014.

[3] D. Dahan, "Prosody and language comprehension," *WIREs Cognitive Science*, vol. 6, no. 5, pp. 441–452, 2015.

[4] K. A. Wenrich, L. S. Davidson, and R. M. Uchanski, "Segmental and Suprasegmental Perception in Children Using Hearing Aids," *Journal of the American Academy of Audiology*, vol. 28, no. 10, pp. 901–912, 2017.

[5] A. Cutler and A. Jesse, "Word Stress in Speech Perception," in *The Handbook of Speech Perception*. John Wiley & Sons, Ltd, 2021, pp. 239–265.

[6] D. Dahan, M. K. Tanenhaus, and C. G. Chambers, "Accent and reference resolution in spoken-language comprehension." *Journal of Memory and Language*, vol. 47, no. 2, pp. 292–314, 2002.

[7] M. Perdomo and E. Kaan, "Prosodic cues in second-language speech processing: A visual world eye-tracking study," *Second Language Research*, vol. 37, no. 2, pp. 349–375, 2021.

[8] A. Weber, M. Grice, and M. W. Crocker, "The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements," *Cognition*, vol. 99, no. 2, pp. B63–B72, 2006.

[9] C. Nakamura, J. A. Harris, and S.-A. Jun, "Integrating prosody in anticipatory language processing: how listeners adapt to unconventional prosodic cues," *Language, Cognition and Neuroscience*, vol. 37, no. 5, pp. 624–647, 2022.

[10] A. Weber, B. Braun, and M. W. Crocker, "Finding Referents in Time: Eye-Tracking Evidence for the Role of Contrastive Accents," *Language and Speech*, vol. 49, no. 3, pp. 367–392, 2006.

[11] K. Ito and S. R. Speer, "Anticipatory effects of intonation: Eye movements during instructed visual search," *Journal of Memory and Language*, no. 58(2), pp. 541–573, 2008.

[12] L. Wang, M. Bastiaansen, Y. Yang, and P. Hagoort, "The influence of information structure on the depth of semantic processing: How focus and pitch accent determine the size of the N400 effect," *Neuropsychologia*, vol. 49, no. 5, pp. 813–820, 2011.

[13] A. Cutler, *Native Listening: Language Experience and the Recognition of Spoken Words*. The MIT Press, 2012.

[14] H. Hopp, "Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability," *Second Language Research*, vol. 29, no. 1, pp. 33–56, 2013.

[15] S. De Deyne, D. J. Navarro, and G. Storms, "Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations," *Behavior Research Methods*, vol. 45, no. 2, pp. 480–498, 2013.

[16] A. Foucart, E. Ruiz-Tada, and A. Costa, "Anticipation processes in L2 speech comprehension: Evidence from ERPs and lexical recognition task," *Bilingualism: Language and Cognition*, vol. 19, no. 1, pp. 213–219, 2016.

[17] H. Hopp, "Semantics and morphosyntax in predictive L2 sentence processing," *International Review of Applied Linguistics in Language Teaching*, vol. 53, no. 3, pp. 277–306, 2015.

[18] C. Tsurutani, K. Tsukada, and S. Ishihara, "Comparison of Native and Non-native Perception of L2 Japanese Speech Varying in Prosodic Characteristics," Queensland, Australia, 2010, pp. 122–125.

[19] A. Sorace, "Pinning down the concept of "interface" in bilinguals," *Linguistic Approaches to Bilingualism*, vol. 1, pp. 1–33, 2011.

[20] H. Clahsen and C. Felser, "Grammatical processing in language learners," *Applied Psycholinguistics*, vol. 27, no. 1, pp. 3–42, 2006.

[21] T. Grüter, E. Lau, and W. Ling, "How classifiers facilitate predictive processing in L1 and L2 Chinese: the role of semantic and grammatical cues," *Language, Cognition and Neuroscience*, vol. 35, no. 2, pp. 221–234, 2020.

[22] J. Klassen, "Second Language Acquisition of Focus Prosody in English and Spanish," Unpublished Thesis, McGill University, Montreal, Quebec, Canada, 2015.

[23] A. Foltz, "Using prosody to predict upcoming referents in the L1 and L2," *Studies in Second Language Acquisition*, vol. 43, no. 4, pp. 753–780, 2021.

[24] J. C. Wells, "Chapter 3 Tonicity: where does the nucleus go? The old and the new," in *English Intonation: An Introduction*. Cambridge: University of Cambridge Press, 2006, pp. 109–111.

[25] L. M. Morett, S. H. Fraundorf, and J. C. McPartland, "Eye see what you're saying: Contrastive use of beat gesture and pitch accent affects online interpretation of spoken discourse." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 47, no. 9, pp. 1494–1526, 2021.

[26] Y. Xu, "Effects of tone and focus on the formation and alignment of f0contours," *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[27] Y.-C. Lee, T. Wang, and M. Liberman, "Production and Perception of Tone 3 Focus in Mandarin Chinese," *Frontiers in Psychology*, vol. 7, 2016.

[28] A. Yang and A. Chen, "The developmental path to adult-like prosodic focus-marking in Mandarin Chinese-speaking children," *First Language*, vol. 38, no. 1, pp. 26–46, 2018.

[29] J. Zhang, S. Duanmu, and Y. Chen, "China and Siberia," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds. Oxford University Press, 2020, pp. 331–343.

[30] P. Tang, I. Yuen, K. Demuth, and N. X. Rattanasone, "The acquisition of contrastive focus during online sentence-comprehension by children learning Mandarin Chinese." *Developmental Psychology*, vol. 59, no. 5, pp. 845–861, 2023.

[31] O. Scharenborg, S. Kakouros, B. Post, and F. Meunier, "Cross-linguistic Influences on Sentence Accent Detection in Background Noise," *Language and Speech*, vol. 63, no. 1, pp. 3–30, 2020.

[32] Y. Liu and J. Ning, "The perception of Mandarin focus intonation by native English speakers," in *6th International Symposium on Tonal Aspects of Languages (TAL 2018)*. ISCA, 2018, pp. 232–236.

[33] S. Wood, "Thin-plate regression splines," *Journal of the Royal Statistical Society (B)*, vol. 65, no. 1, pp. 95–114, 2003.

[34] R. C. Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[35] X. Ren and C. Cohen, "Integration of Multiple Cues in Native and Non-Native Speech Perception," in *Proceedings of the 20th International Congress of Phonetic Sciences*, R. Skarnitzl and J. Volin, Eds. Prague Congress Center, Czech Republic: GUARANT International spol.s r.o., 2023, pp. 97–101.

[36] C. Shih, "Tonal Effects on Intonation," in *Proceedings of International Symposium on Tonal Aspects of Languages: Emphasis on Tone Lan- guages*, Beijing, 2004, pp. 63–168.