INTEGRATION OF MULTIPLE CUES IN NATIVE AND NON-NATIVE SPEECH PERCEPTION

Xiaomu Ren and Clara Cohen
Glasgow University Laboratory of Phonetics, University of Glasgow
x.ren.1@research.gla.ac.uk

ABSTRACT

In this visual-world eye-tracking study, native and non-native listeners' integration of prosody and semantics in speech perception was examined. Participants' eye movements were recorded as they listened to English sentences with variations in prosodic accent, target object information, and semantic match. The results revealed a complex three-way interaction between L1, prosody, and verb semantics. Mandarin listeners did not always prioritize semantics over prosody; instead, they relied more on prosody for processing new information. They also exhibited less flexibility in their integration strategy compared to English listeners, employing a similar approach for both repeated and new information. In contrast, English listeners utilized prosodic cues to enhance semantic processing with old information sentences, but they did not consistently give equal weight to different speech cues. Specifically, they used fewer prosodic cues and prioritized semantics for sentences with new information, where verb semantics were beneficial for new nouns.

Keywords: Pitch accent; Verb semantics; Information status; Eye movements; Speech perception

1. INTRODUCTION

Understanding spoken language demands fluency in two skills: perception and comprehension. Low-level perception processes segmental and suprasegmental cues [1], encoding discrete segments, like consonants and vowels [2], along with prosodic information about prominence (stress and pitch accent) [3, 4]. Higher level comprehension processes extract lexical, semantic, and syntactic information, deriving meaning from the incoming speech stream. Listeners must attend to both sources of information, incorporating them in real time.

Native listeners perform this task effortlessly, drawing on a wide range of information to maximize the efficiency of speech perception and minimize adverse interference [5]. They can draw on low-level prosodic cues to make sense of information structure [6,7], disambiguate structural ambiguities [8,9], and predict upcoming referents [7,10,11]; and they can draw on semantic cues from verb to predict the upcoming direct object on hearing a transitive verb but not an intransitive verb [12]. Furthermore, they can combine these levels of information interactively [13], observing that semantic processing of verbs is strengthened when those verbs are highlighted by a prosodic pitch accent.

Non-native listeners can also use high-level and low-level information in speech perception. At high levels, non-native listeners use contextual semantic information to predict the upcoming lexical form [14, 15], and use the morphosyntactic information like article-noun gender agreement to predict the upcoming target nouns [16, 17]. At lower levels, too, non-native listeners draw on prosodic characteristics to judge the naturalness of the pronunciation [18], and build an understanding of information structure [7]. However, non-native speech perception has been claimed to struggle in integrating information across multiple domains [19], relying on semantic information over syntax or prosody [20, 21].

To examine how native and non-native listeners differ in their attention to high-level and low-level information, we ask whether (i) non-native listeners are indeed less attentive to low-level prosodic information, relative to high-level semantic information, and (ii) whether they are less able to combine high-level and low-level information interactively.

To answer this question, we designed a visual-world eye-tracking experiment in which auditory stimuli were presented with manipulations across three types of speech cues: pitch accent, information status, and verb semantics. Pitch accent and information status test listeners' attention to low-level prosodic information, as the expected pitch accent 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 listeners are less attentive to prosody than semantic information, then we expect that non-native listeners will show reduced effects of pitch accent in their ability to identify a target noun, relative to native listeners. We also expect the effects

of prosody in non-native listeners to be smaller than the effects of semantic information. Finally, if non-native listeners are also less able to combine the three types of speech cues interactively, then we expect that the effect of semantic information will be smaller, regardless of pitch accent, whereas native listeners should show increased semantic processing when target nouns are accented, especially when that accent correctly signals new information status in the target noun.

2. EXPERIMENTAL METHOD

2.1. Participants

Participants were 32 English native speakers ("English listeners") and 40 Mandarin native speakers learning English as L2 ("Mandarin listeners"). The Mandarin listeners used English daily with at least intermediate proficiency (IELTS above 5.5). All participants were university students.

2.2. Materials

Participants were presented with a visual display containing four pictures: a target noun (e.g., *cabbage*), a competitor noun sharing minimally the onset and vowel of the first syllable (e.g., *captain*), and two distractor images (e.g., *badger* and *crocus*). Auditory stimuli took the form of two sentences, with participants instructed to click on the noun mentioned in the second sentence.

All sentences were in English, and were designed to manipulate three binary variables: Information Status (Repeated/New), Verb Appropriateness (Appropriate/Inappropriate), and Pitch Accent (Target/Neutral). In New sentences, the competitor was named in the first sentence and the target named in the second sentence (e.g., *Here is a captain. Susan is going to shred the cabbage after school*). In Repeated sentences, the target was named in both sentences (e.g., *Here is a cabbage. Susan is going to shred the cabbage after school*).

In Appropriate sentences, the verb in the second sentence could only apply to the target as a direct object (e.g., *shred the cabbage*). In Inappropriate sentences, the verb could not plausibly apply to the target (e.g., *drink the cabbage*).

In Target accented sentences, a contrastive pitch accent was placed on the target noun (e.g., *Susan is going to shred the CABBAGE after school*), while in neutral accented sentences, the prosody was a more neutral topic-content melody, with any pitch accent placed on the final adverbial.

These three binary variables were crossed to create 8 conditions for each item. Forty different sentence-pair items were recorded, and rotated in the 8 conditions across 8 experimental lists in a Latin square design. Each list also contained a further 40 filler sentence pairs. All filler sentences were constructed so as to balance each combination of information status, verb appropriateness, and pitch accent across all trials within each experimental list. Stimuli were presented in pseudo-random order, in a different order for each participant. All sentences were recorded by a phonetically-trained native speaker of English in a soundproofed booth.

2.3. Procedure and Analysis

Data were collected with an Eyelink 1000+ eye-tracker running Experiment Builder software. Gaze data was sampled at 1000Hz. Each experiment began with calibration, two practice trials, and then the system was recalibrated before the main experiment began. Participants completed 8 blocks of 10 trials, with a break and recalibration between each block.

Looks to target and competitor were binned and converted to proportions over 50ms windows, over an interest period stretching from 200 ms before the target onset to 1800ms after target onset. Target advantage (TA) was calculated by subtracting the proportion of looks to competitor from proportion of looks to target. This TA metric formed the dependent variable.

TA was analysed with generalised additive mixed models (GAMMs), using the `mgcv` package (version 1.8.4 [22]) in R (version 4.2.1; [23]). Repeated and new information sentences were analysed separately. For both repeated and new sentences, a simple GAMMs was built, containing parametric effects and difference smooths for each of the three binary main variables: Accent (neutral/target); Verb Appropriateness (appropriate/inappropriate); and L1 (Mandarin/English). All factors were treatment-coded, with default levels set as Neutral Accent, Appropriate Verb, and Mandarin L1. Interaction terms were added progressively as two-way and then three-way interactions, individually considered both as parametric terms, difference smooths, or both. Each subsequent model was tested against the simpler model with a log-likelihood ratio test, as implemented in the package `itsadug` (version 2.4.1 [24]).

3. RESULTS

Figure 1 shows the gaze traces for both repeated and new target nouns, while Table 1 summarises the final models. For repeated information conditions, the optimal model contained all two-way interactions in

the parametric terms, and a three-way interaction in the difference smooths. For new information, the model ended up needing three-way interactions in both parametric terms and smooths.

For repeated information sentences, Mandarin listeners had a reduced TA for Inappropriate verbs ($\beta = -0.068, p < .001$) and Appropriate verbs showed very little effect of Accent ($\beta = 0.001, p = .91$). However, the Accent by Appropriateness interaction reveals that Accent did affect Inappropriate verbs, which suffered more under Target Accent than under Neutral Accent ($\beta = -0.029, p < .01$). English listeners showed higher TA than Mandarin listeners with Appropriate verbs under Neutral Accent ($\beta = 0.214, p < .001$), and unlike Mandarin listeners, English listeners benefited from Target Accent, with both Appropriate and Inappropriate verbs ($\beta = 0.047, p < .001$). English listeners also showed a smaller deficit for Inappropriate verbs than Mandarin listeners ($\beta = 0.042, p < .001$), but the lack of three-way interaction means that, like Mandarin listeners, English listeners also showed a stronger Appropriateness effect under Target Accent than Neutral Accent. The significantly different curvature of the gaze traces indicates that this Appropriateness effect was concentrated in the first 25 time windows for English listeners (Fig 1, top right black lines).

In other words, with repeated nouns, Mandarin listeners were more sensitive to verb Appropriateness than accent, and accenting the target increased the effect of Appropriateness. English listeners shared the heightened effect of Appropriateness under Target Accent, but also showed a stronger sensitivity to Accent than Mandarin listeners.

With new Information nouns, Mandarin listeners demonstrate a response pattern that was clearly slower than that of English listeners. English listeners reached peak TA at about Time Window 25, while Mandarin listeners did not peak until Time Window 35. For Mandarin listeners, there was little net effect of Appropriateness for Neutral Accent ($\beta = -0.004, p = .7$), but the significantly different curvatures of the gaze traces reveal that TA for Appropriate verbs rose faster and peaked higher than for Inappropriate verbs (Fig 1, bottom left grey lines). With Target Accent, Inappropriate Verbs had significantly lower net TA than Appropriate verbs ($\beta = -0.061, p < .001$). English listeners showed a much larger effect of Appropriateness for Neutral Accent than Mandarin listeners ($\beta = -0.061, p < .001$), but did not share the benefit of Target Accent on Appropriate verbs with Mandarin speakers ($\beta = -0.08, p < .001$). They also showed a smaller effect of Appropriateness under Target Accent than Neutral Accent ($\beta = 0.071, p < .001$), although again the significantly different curvatures of the traces suggest that this reflects a delayed peak in TA for Inappropriate verbs, allowing net TA to 'catch up' to Appropriate verbs (Fig 1, bottom right black lines).

In other words, English listeners showed a larger effect of Appropriateness than Mandarin speakers, but it was slightly reduced under Target Accent. The effect of Appropriateness emerged more slowly in Mandarin speakers, but by contrast to English speakers it was more pronounced under Target Accent than Neutral Accent.

4. DISCUSSION AND CONCLUSION

The study confirms some, but not all of our predictions. First, when processing repeated information sentences, Mandarin listeners were more sensitive to semantics (Appropriateness) than prosody (Accent), while English listeners attended to both. Further, when English listeners listened to repeated information sentences, adding target accent increases the attention to semantics, which was also predicted. This finding replicates a pattern observed in a previous study [13], where placing a contrastive pitch accent on a target noun deepened semantic processing.

However, contrary to prediction, this deepening of semantic processing disappeared for English listeners with new information sentences, while it was present for Mandarin listeners in both contexts. Further, although we predicted that Mandarin listeners would rely more inflexibly on semantics than English listeners, instead Mandarin listeners showed more sensitivity to prosody than semantics with new information, while it was English listeners for whom semantics had the largest effect.

The results suggest that, counter to our predictions, Mandarin listeners are, in fact, capable of integrating speech cues at different levels, and they do not always prioritize semantics over prosody. Nevertheless, we can confirm that Mandarin listeners integrate speech cues *differently* from English listeners, and *less flexibly*. Mandarin speakers showed a similar integration strategy between semantics and prosody across both repeated and new information. The addition of target accent always increased semantic processing, perhaps because the accent rendered the target noun louder and more salient. The accent also aided perception generally in new information sentences, where that prosodic structure was appropriate to the information status of the sentence.

English listeners, by contrast, allowed prosodic accent to deepen semantic processing only with old

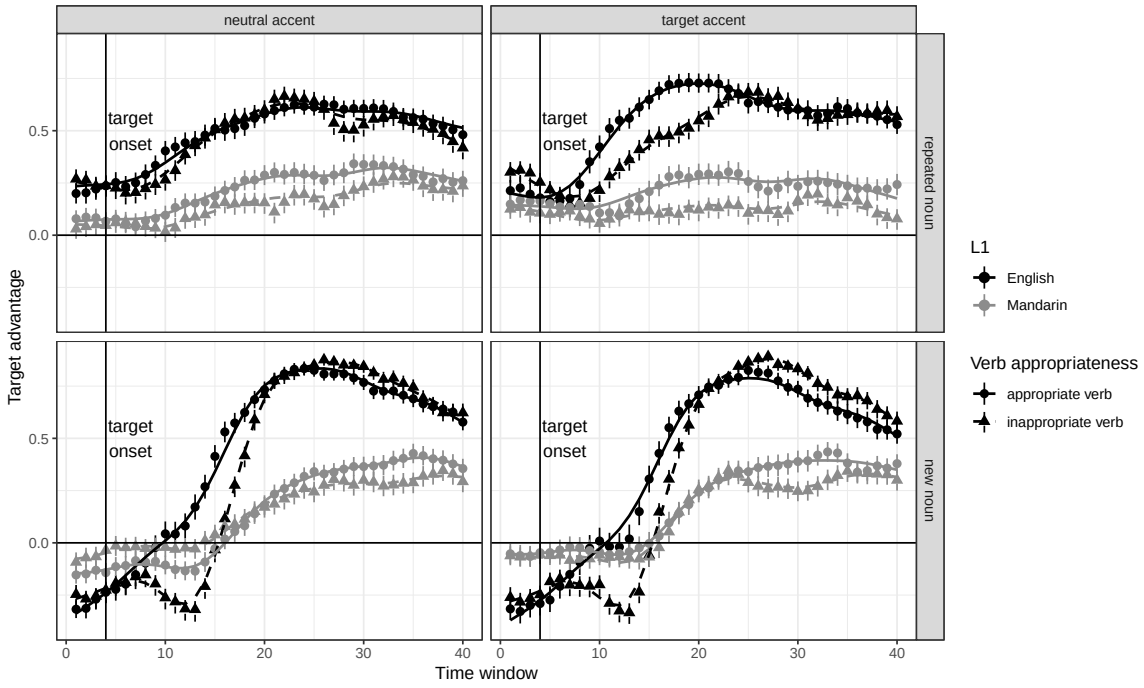

Figure 1: Gaze traces with repeated nouns, showing target advantage (TA) across English (black line) and Mandarin (grey line) listeners, for appropriate (dotted circles, connected by solid lines) and inappropriate (solid triangle, connected by dashed lines) verbs semantics. Neutral accent is on the left, while Target Accent is on the right. Target accent 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 Accent (abbreviated *acc*) are Target (neutral, abbreviated *neu*) and target (abbreviated *tar*). Levels for Appropriateness (*App*) are Appropriate (reference, *a*) and Inappropriate (*i*). Levels for L1 are Mandarin (reference, *ma*) and English (*eng*). Difference smooths are calculated on an 8-level factor representing the interaction between Acc, App, and L1.

Repeated information

Parametric	Est.	SE	*t*	*p*
Intercept	0.240	0.030	7.85	$< .001$
Acc=tar	0.001	0.009	0.11	.91
App=i	-0.068	0.009	-7.71	$< .001$
L1=eng	0.214	0.043	4.95	$< .001$
Acc=tar:App=i	-0.029	0.011	-2.74	$< .01$
Acc=tar:L1=eng	0.047	0.011	4.36	$< .001$
App=i:L1=eng	0.042	0.011	3.88	$< .001$

Difference smooths	edf	*F*	*p*
Window	6.65	11.25	$< .001$
Win:Acc=tar,App=a,L1=eng	6.82	8.03	$< .001$
Win:Acc=neu,App=i,L1=eng	5.56	2.73	$< .05$
Win:Acc=tar,App=i,L1=eng	6.55	5.26	$< .001$
Win:Acc=neu,App=a,L1=man	1.02	0.82	.36
Win:Acc=tar,App=a,L1=man	1.01	6.34	$< .05$
Win:Acc=neu,App=i,L1=man	3.08	3.61	$< .05$
Win:Acc=tar,App=i,L1=man	3.22	6.97	$< .001$
Window, by subj	434.00	8.38	$< .001$

New information

Parametric	Est.	SE	*t*	*p*
Intercept	0.154	0.029	5.31	$< .001$
Acc=tar	0.036	0.010	3.72	$< .001$
App=i	-0.004	0.10	-0.39	.70
L1=eng	0.260	0.043	6.02	$< .001$
Acc=tar:App=i	-0.041	0.014	-2.93	$< .005$
Acc=tar:L1=eng	-0.080	0.014	-5.59	$< .001$
App=i:L1=eng	-0.061	0.014	-4.28	$< .001$
Acc=tar:App=i:L1=eng	0.071	0.021	3.40	$< .001$

Difference smooths	edf	*F*	*p*
Window	8.51	63.01	$< .001$
Win:Acc=tar,App=a,L1=eng	1.00	1.20	.27
Win:Acc=neu,App=i,L1=eng	8.63	32.96	$< .001$
Win:Acc=tar,App=i,L1=eng	8.63	31.26	$< .001$
Win:Acc=neu,App=a,L1=man	6.39	18.92	$< .001$
Win:Acc=tar,App=a,L1=man	6.31	19.56	$< .001$
Win:Acc=neu,App=i,L1=man	6.67	21.94	$< .001$
Win:Acc=tar,App=i,L1=man	6.49	20.17	$< .001$
Window, by subj	445.41	8.12	$< .001$

information sentences, where the accent was inappropriate for the information structure. Perhaps this was because the prosodic pattern, by virtue of its unexpectedness, served to direct their attention to the target noun. In new information sentences, the prosodic pattern was expected, and so almost invisi-

ble to English listeners.

5. REFERENCES

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

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

[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, pp. 292–314, 2002, place: Netherlands Publisher: Elsevier Science.

[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, pp. B63–B72, 2006, place: Netherlands Publisher: Elsevier Science.

[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, pp. 367–392, 2006, place: United Kingdom Publisher: Kingston Press.

[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] M. Arai and F. Keller, "The use of verb-specific information for prediction in sentence processing," *Language and Cognitive Processes*, vol. 28, no. 4, pp. 525–560, 2013.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[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 âinterfaceâ 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. Gruter, 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, Feb. 2020, publisher: Routledge.

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

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

[24] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2022, r package version 2.4.1.