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

Same Sentences Different Meanings: Prosodic and Gestural Resolution of Ambiguity in Mandarin Chinese

Jiajun Gao[1], *Yan Gu*[2,3]

[1]The University of Nottingham Ningbo China
[2]University of Essex, [3]University College London, UK
hvyjg1@nottingham.edu.cn, yan.gu@essex.ac.uk

Abstract

Speakers use prosody to resolve ambiguity, but what if prosody cannot make distinctions? We explore (1) how speakers employ prosodic and gestural cues to deal with sentences with ambiguous meanings and (2) what insights the audiovisual resolution of ambiguities offers regarding communicative efficiency and effort. Thirty-two native Chinese speakers were asked to articulate twenty-two ambiguous Mandarin sentences. Half could be semantically differentiated using prosody, and half could not. Firstly, participants articulated all ambiguous sentences spontaneously and provided explanations to a confederate, revealing their dominant interpretations. Secondly, participants articulated the same ambiguous sentences twice, each time guided by a hint suggesting a different meaning. Participants' prosodic cues and gestures were coded and analyzed. The results showed that for ambiguous sentences that can be prosodically distinguished, participants employed various prosodic cues such as pausing, tones, stress, and speaking rates. Additionally, 51.85% of sentences were accompanied by referential (iconic; pointing) gestures, while 17.33% of sentences were accompanied by non-referential (beat; interactional) gestures. However, when prosodic cues were unable to mark ambiguity, participants resorted to more referential gestures (97.30%) but fewer non-referential gestures (1.28%). In conclusion, speakers adopt a multimodal approach to enhance communicative efficiency while there is a trade-off between modalities.

Index terms: prosody, gesture, Chinese, trade-off hypothesis, multimodal ambiguity, communicative efficiency and effort

1. Introduction

Inherent in language communication is the challenge of linguistic ambiguity, where a single expression may give rise to multiple interpretations [1], potentially leading to misunderstandings [2], [3], [4]. While previous research suggests that employing prosodic cues such as pause and stress can assist in dealing with ambiguities [5], [6], [7], relying solely on them is sometimes insufficient to resolve linguistic ambiguities. For instance, in Chinese, the sentence "王先生借了李先生一本书" (wáng-xiān-shēng jiè-le lǐ-xiān-shēng yì-běn-shū) can be interpreted in two ways: either as "Mr. Wang lent a book to Mr. Li" or "Mr. Wang borrowed a book from Mr. Li.". The character "借" here can denote both "lend" and "borrow", but there is no phonemic difference between the two meanings of the character "借". Thus, prosodic cues fail to mark the ambiguity in this instance. Nevertheless, communication is multimodal [8], [9], [10], [11], allowing for the resolution of such ambiguities if the speaker accompanies the statement with corresponding "give" or "receive" gestures. Although efforts to address linguistic ambiguities have extended to body movements that modify languages [12], [13], [14], and despite an increasing number of studies on ambiguity resolution in Chinese, to our best knowledge, no research has incorporated an audiovisual resolution. This study aims to better understand the multimodal resolution of ambiguities in Chinese.

Prosodic cues, including stress, rhythm, and intonation of language [15], [16], [17], are crucial for disambiguation [5], [18], [19]. For instance, variations in word duration [7], pause duration [6], and prosodic contour duration [20] can positively impact listeners' interpretation of ambiguous sentences. Furthermore, the role of prosodic cues extends beyond the immediate sentence being communicated. It can also assist in predicting forthcoming ambiguous structures [19], [20], prompting listeners to attend to the prosodic cues in the subsequent segments, thus facilitating efficient communication.

Despite the importance of prosodic cues, they may sometimes prove insufficient to fully resolve ambiguities. Several factors contribute to this inadequacy. First, speakers' proficiency with prosodic cues can vary, with age-related declines in sensitivity to speech prosody [21], compounded by cognitive impairments [22] or auditory deficiencies [23]. Second, certain ambiguities may remain unresolved due to inherent linguistic complexity, particularly evident in Chinese sentences. Chinese displays special phonetic features that are less common in Indo-European languages, including the prevalence of homophonic words that share identical pronunciations but convey distinct meanings [24]. For example, the Chinese sentence "他倒了一杯水" (tā dào-le yì-bēi-shuǐ) can mean either "He fills the cup with water." or "He empties the cup.". The character "倒" can signify both pour *into* or pour *out*, creating a lexical ambiguity. In such cases, resolving ambiguity through prosodic cues alone is challenging. However, incorporating a gesture such as "pour into" or "pour out" can effectively disambiguate the sentence. This highlights the importance of gestures in disambiguation.

The above example demonstrates how gestures contribute to communication efficiency. Communicative efficiency refers to the effective transmission of information between communicators with minimal effort [25], [26]. In this context, This effort encompasses the cognitive and physical resources expended by both the speaker and the listener during communication [26], [27]. While prosodic and gestural cues are crucial for disambiguating sentences, their separate or combined effects on communicative efficiency and effort remain unclear. This study examines the effectiveness of audio and visual resolution in clarifying ambiguous Chinese sentences and evaluates how their combination affects communicative efficiency and efforts. By exploring these aspects, we aim to better understand the interplay between

10.21437/SpeechProsody.2024-179

prosody and gestures, and their role in facilitating communication.

While previous studies on disambiguation through gestures covered various ambiguity types and different age cohorts, they have predominately centered around Indo-European languages [27]. Regardless of whether in children or adults, gestures consistently demonstrate their value in facilitating communication [1], [12], [14], [28]-[32]. Furthermore, by addressing ambiguities, gestures improve robots' ability to comprehend human instructions more precisely [33]-[35]. Nevertheless, limited research has investigated the role of gestures in non-Indo-European languages [11], despite the unique linguistic features of languages like Chinese that generate ambiguities less common in Indo-European languages. In Chinese, different interpretations of ambiguous sentences may not be phonetically distinguished due to the absence of discernible phonetic differences. Moreover, Chinese relies more heavily on contextuality than English [37], and an analysis of Chinese may shed light on different patterns of gestural resolution of ambiguity. Thus, studying prosodic and gestural resolution of ambiguity in Chinese can provide valuable insights into communication efficiency and effort. We ask two research questions:

RQ1: How do native Chinese speakers use prosodic and gestural cues to manage ambiguous Chinese sentences?

RQ2: What insights can audiovisual resolution of ambiguity provide on communicative effort and efficiency?

According to the communicative efficiency hypothesis, when speech prosody alone suffices to disambiguate, participants may be less likely to gesture. However, if the ambiguity cannot be resolved solely through prosodic differences, participants may rely more on gestures for clarification.

2. Methodology

2.1 Participants

Thirty-two Chinese-native students (5 males, 27 females) (Mean age = 20.97 years, range 19 - 23 years) from the University of Nottingham Ningbo China participated in this study. The sample size was decided based on G*Power version 3.1 with a power of 0.8 and a medium effect size of 0.5 [39]. Additionally, the researcher appointed one confederate per participant to stimulate participants' communicative intent. These confederates were 32 additional recruits or participants who had previously completed the experiment. All participants reported no hearing or speech impairments. Participants signed an informed consent and were paid for their contribution. The study obtained ethical approval from the University of Nottingham Ningbo China.

2.2 Apparatus and stimuli

The stimuli consisted of 22 ambiguous Chinese sentences adapted from [36], each having two different interpretations. The sentences were evenly divided into two groups based on their disambiguation types. The first group (N = 11) could be disambiguated using prosodic cues such as pauses and stress, while the second group (N = 11) presented challenges in disambiguation solely through prosodic cues. All stimuli were displayed on a MacBook Pro computer screen with a resolution of 2560 × 1600. Each participant completed the experiment in a spacious, well-lit, and quiet room. Both their voices and body movements were recorded using Audacity 3.3.2 (16 bit, 44.1 kHz) and a phone camera (4K at 30 fps), respectively.

2.3 Procedures

First, in a simple pretest, participants saw each of the ambiguous sentences on a computer screen without hints, requiring them to read the sentence aloud to the confederate and then explain its meaning in their own words. Each sentence was presented independently on a PowerPoint slide. As participants interpreted these 22 sentences intuitively and spontaneously, their explanations revealed participants' dominant interpretations. This process aimed to control for the potential effect of a non-dominant interpretation (less predictable) on prosodic production during the main task. Furthermore, while confederates were not encouraged to give feedback, nods and headshakes were allowed.

In the main task, participants viewed two different slides, each displaying one of two hints for the same sentence suggesting two possible meanings. For instance, the sentence "他倒了一杯水" (tā dào-le yì-bēi-shuǐ) can mean either "he fills the cup with water" or "he empties the cup". One slide showed the target sentence with the hint "往杯子里" (fill the cup) underneath it, while the other slide showed the same sentence with a different hint "水不要了" (empty the cup). Participants verbally expressed the target ambiguous sentence based solely on the hint information, without mentioning the hint itself. Confederates were not encouraged to give feedback but indicated understanding with nods or headshakes. To motivate the communicative intent of speakers, they were told and could see that the confederate would guess and mark down what interpretation the sentence referred to (mean accuracy rate = 94.18%). Speakers were not told to use prosodic or gesture cues. All participants first completed the pre-test, followed by the main task. The sequence of these 22 Chinese sentences was first randomised, creating two counterbalanced versions. The order of the two hints was also counterbalanced.

2.4 Annotations

Speech articulations were annotated in Praat 6.3.10 [40]. From the pretest, participants' dominant interpretation of ambiguous sentences was coded. Most participants shared a similar dominant interpretation for the 22 ambiguous sentences (M = 82.52%, range 53.12% - 100%).

For the main part, firstly, utterance boundaries were automatically detected and manually checked. Secondly, each sentence was given an ID indicating its meaning (according to the hints provided). Thirdly, we coded whether the hint of the sentence aligned with the participant's dominant interpretation (according to information from the pre-test). Fourth, the use of prosodic cues (pausing, lengthening, accented, or different pronunciations) was indicated. Repetition, errors, or disfluencies were noted, and the final best production was used.

For sentences that could be marked by prosody, four types of coding were applied: (1) Binary encoding was used to indicate pause positions in sentences (N = 4) that could be disambiguated by pauses. Pauses were labeled as 'before' if they appeared before the target characters, or as 'nonbefore' if they occurred somewhere else or were absent. (2) One sentence could be disambiguated by two distinct tones ('好', 'hǎo 3', or 'hào 4'). Participants' production of the third or fourth tone was coded, respectively. (3) Five sentences used stress as prosodic cues, with 'stressed' or 'unstressed' labels applied to the target

characters. (4) The last sentence used speech rate for disambiguation, with the target word '多半' either a longer or shorter duration, indicating 'majority' or 'probability'. A Praat script was used to automatically extract pause, tone, intensity, pitch, and duration for the target sentences or items.

Gestures were coded in ELAN 6.5 [41]. The type of the gesture was coded according to iconic, metaphoric, point, beat, and pragmatics [42]. Furthermore, iconic, pointing, and metaphorical gestures were categorized as referential gestures whereas beats and pragmatic gestures were categorised as non-referential gestures [43], [44]. A second person coded 15% of the participants (N=5). The consistency in the presence or absence of a gesture was 98.32%. The overall agreement of gesture functions (referential or non-referential) was 90.05%.

2.5 Statistical analysis

The Linear Mixed-Effects models (linear dependent variables) and GLMM models (binary dependent variable) in R were used for data analysis [45]. We investigated whether dominance (alignment of hint with dominant interpretation), prosodic ambiguity (whether prosody can make the distinction), and the interaction between dominance and prosodic ambiguity (IVs) influence prosodic (e.g., speech rate (words per sec), mean pitch (converted to semitone), mean intensity, intensity maximum, and intensity range) and gestural production (referential; nonreferential). Participants and ambiguous sentences were set as random intercepts and prosodic ambiguity was determined as the random slope to the participant.

3. Results

Table 1: *The mean (standard deviation) for prosodic and gestural features of sentences elicited by two different hints.*

Measures	Hint aligns with the dominant interpretations	Hint does not align with the dominant interpretation
Speech Rate	3.89 (1.07)	3.58 (1.01)
Mean Pitch (ST)	25.64 (4.15)	25.61 (4.22)
Mean Intensity (dB)	50.80 (4.66)	58.89 (4.81)
Max Intensity (dB)	70.74 (4.90)	71.24 (4.96)
Intensity Range (dB)	23.97 (6.20)	24.57 (5.98)
Ref gesture (%)	75 (0.43)	74.15 (0.44)
Non-Ref gesture (%)	9.09 (0.29)	9.52 (0.29)

3.1 Prosody

Participants were faster at articulating sentences aligned with their dominant interpretations in comparison to non-dominant interpretations (β = 0.083, p < .001), regardless of whether prosodic cues could resolve ambiguities (Table 1). Non-dominant interpretations had higher mean intensity (β = 0.243, p = 0.027) and maximum intensity (β = 0.439, p = 0.004) than dominant interpretations for sentences that could use prosodic cues to mark ambiguity. However, neither mean intensity (β = 0.101, p = 0.514) nor maximum intensity (β = -0.167, p = 0.437) was significant when ambiguous sentences remained undistinguished by prosodic cues, demonstrating that intensity did not contribute to addressing ambiguities in such instances. There was no significant difference in mean pitch between dominant and non-dominant interpretations (β = 0.054, p = 0.589), irrespective of whether prosody could disambiguate.

Furthermore, controlling for participants' dominant interpretations, we focused on sentences that can use prosody to mark ambiguity. First, for ambiguous sentences disambiguated by the pause, significant influences were found in disambiguation based on two pause positions (48.06% for pausing at position A, 51.94% for 'no pauses at this position', β = 7.23, p < .001). Second, in sentences resolved through two distinct tones, a significant difference was observed in the mean pitch of the two tones when "好" was pronounced in the third tone (M = 23.05 ST, SD = 4.39) and the fourth tone (M = 28.99 ST, SD = 4.69) (β = 0.29, p < .001). Third, when participants resolved ambiguities by stressing characters, stressed characters were articulated with longer duration ($M_{stressed}$ = 0.37 sec, SD = 0.15 vs. $M_{unstressed}$ = 0.29 sec, SD = 0.14, β = 0.092, p < .001), wider intensity range ($M_{stressed}$ = 14.56 dB, SD = 4.81 vs. $M_{unstressed}$ = 11.95 dB, SD = 5.45, β = 2.912, p < .001), higher maximum intensity ($M_{stressed}$ = 67.99 dB, SD = 5.80 vs. $M_{unstressed}$ = 66.21 dB, SD = 5.07, β = 2.151, p < .001), and higher mean pitch ($M_{stressed}$ = 27.36 ST, SD = 6.03 vs. $M_{unstressed}$ = 25.72 ST, SD = 5.04, β = 0.29, p = 0.009). Finally, for one stimulus where the speech rate of target words ("多半") aided in disambiguating, the meaning of "majority" had a longer duration (M = 0.58 sec, SD = 0.15) than the meaning of "probability" (M = 0.47 sec, SD = 0.12), β = 0.094, p = 0.0002.

3.2 Gestures

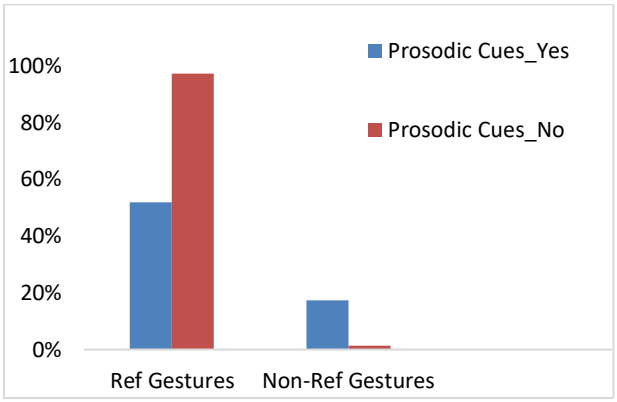

Figure 1: *Participants' gesture performance when articulating ambiguous sentences.*

Overall, there was no significant difference in gesture production between sentences aligned and misaligned with the dominant interpretation (p > 0.05). Importantly, controlling for participants' dominant interpretation, participants were significantly more inclined to gesture when confronted with ambiguous sentences that could not be disambiguated through prosodic cues (M = 98.15%, SD = 0.13, N = 704) compared to sentences that could be disambiguated using prosody (M = 67.05%, SD = 0.47, N = 704) (β = 3.429, p < 0.001). A further analysis according to the referentiality of gestures revealed that such differences were mainly driven by referential gestures, which were more often observed in the prosodically ambiguous condition (M = 97.30%, SD = 0.16) than in the prosodically non-ambiguous condition (M = 51.85%, SD = 0.5) (β = 4.352, p < 0.001) (see Figure 1). For instance, participants were highly likely to produce gestures for the sentence "他倒了一杯水" (tā dào-le yì-bēi-shuǐ) that prosody cannot mark distinctions for different meanings such as "He fills the cup with water" and "He empties the cup" (Figure 2).

Furthermore, the proportion of non-referential gestures (M = 17.33%, SD = 0.39) was higher when prosodic cues effectively resolved ambiguities than when prosody could not resolve ambiguities (M = 1.28%, SD = 0.11), $\beta = 5.029$, $p = 0.003$. Specifically, there were more beats (M = 0.056, SD = 0.231, $\beta = 2.613$, $p < 0.001$) and pragmatic gestures (M = 0.116, SD = 0.321, $\beta = 3.282$, $p < 0.001$) in the prosodic unambiguous sentences compared to the prosodic ambiguous sentences (M = 0.005, SD = 0.075 for beats, M = 0.007, SD = 0.084 for pragmatic gestures).

Figure 2: *Gestures in two interpretations of "他倒了一杯水": (a)"He fills the cup with water"; (b)"He empties the cup".*

4. Discussions

This study examined how speakers use prosody and gesture to resolve ambiguities in Chinese sentences and explored the implications of audiovisual resolution for efficient communication. The findings revealed that, when prosody could address ambiguities, participants employed various prosodic cues but fewer referential gestures compared to sentences with prosodic ambiguity. Consistent with prior studies [7], [19], [22], participants in this research employed pausing, stressed characters with higher mean pitch and maximum intensity to mark ambiguities. Moreover, given the unique tonal systems of the Chinese language [36], [46], participants employed two distinct tones to resolve ambiguities. This highlights the dynamic nature of speech production.

In addition, non-dominant interpretations had higher mean and maximum intensity compared to dominant ones when prosodic cues effectively disambiguated sentences. This indicates that speakers made efforts to emphasize the unmarked interpretation, but only when such information could be effectively conveyed through prosody. However, participants' dominant interpretations had faster speech rates than non-dominant interpretations, irrespective of whether prosodic cues could resolve ambiguities. This is because the duration of words and sentences was longer when speakers articulated less predictable non-dominant meanings [48].

Speakers indeed used multimodal cues to mark ambiguities in Chinese sentences. Even when prosodic cues alone were adequate for disambiguation, they still exhibited a high proportion of referential gestures (51.85%). Additionally, speakers also produced a notable proportion of non-referential gestures (17.33%). Interestingly, these non-referential gestures, such as beats and pragmatic gestures, coincided with prosodic prominence, as corroborated by [47].

Furthermore, participants exhibited a significantly higher frequency of gestures (98.15%) in cases where prosodic cues were insufficient for resolving ambiguities, compared to instances where prosody successfully functioned. This suggests a stronger tendency to use a multimodal approach for disambiguation [13], [30], [49] in communication [9], [10], [11]. These findings align with the trade-off hypothesis between resolving ambiguities and achieving communicative efficiency, indicating a balance between competing goals in communication and the manual efforts individuals exert.

Additionally, there was a decrease in the occurrence of non-referential gestures when prosodic cues were ineffective in disambiguating. This suggests that different types of gestures compete in gesture production, with non-referential gestures being less prioritized in resolving ambiguity. These gestures were more frequently produced in prosodically unambiguous sentences where the coupling of prosodic prominence and beat gestures demonstrated a parallel between prosody and gesture, employing both channels simultaneously.

This study lays the foundation for future research. For example, it is unknown how speakers' use of prosody and gesture to resolve ambiguity may differ in more naturalistic conversation. It is also interesting to examine the respective roles of these cues in disambiguation during comprehension.

5. Conclusion

This is the first study focused on the audiovisual resolution of ambiguities in Chinese sentences, revealing the diverse multimodal strategies participants employed for effective communication. Participants used pauses, tone variations, stressed characters, and speech rate adjustments, alongside gestures, to disambiguate prosodic unambiguous sentences. Conversely, they relied more on referential gestures to clarify prosodically ambiguous ones. In sum, speakers adopt a multimodal approach to achieve communicative efficiency, while there is a trade-off between modalities.

6. Acknowledgements

We thank all participants in this study. The work was supported by The National Social Science Fund of China (20BYY179).

7. References

[1] E. Biau, L. A. Fromont, and S. Soto-Faraco, "Beat gestures and syntactic parsing: an ERP study," *Language Learning*, vol. 68, pp. 102–126, Jun. 2018.

[2] T. A. Harley, *The psychology of language: From data to theory*. Hove: Psychology Press, 2013.

[3] T. A. Harley, *Talking the talk: Language, psychology and science*. Hove, East Sussex Philadelphia: Psychology Press, 2017.

[4] P. Warren, *Introducing psycholinguistics*. Cambridge: Cambridge University Press, 2013.

[5] J. D. Fodor, "Psycholinguistics cannot escape prosody," in *Speech Prosody 2002*, New York, USA, Apr. 11-13, 2002.

[6] A. Gollrad, E. Sommerfeld, and F. Kügler, "Prosodic cue weighting in disambiguation: Case ambiguity in German," in *Speech Prosody 2010-Fifth International Conference*, Chicago, USA, May 2010.

[7] S. Wiener, S. R. Speer, and C. Shank, "Effects of frequency, repetition and prosodic location on ambiguous Mandarin word production," in *Speech Prosody 2012*, Shanghai, China, pp. 528–531, Jan. 2012.

[8] T. Chen, and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

[9] J. P. Higham, and E. A. Hebets, "An introduction to multimodal communication," *Behavioral Ecology and Sociobiology*, vol. 67, pp. 1381–1388, Jul. 2013.

[10] J. Holler, and S. C. Levinson, "Multimodal language processing in human communication," *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 639–652, Aug. 2019.

[11] G. Vigliocco, P. Perniss, and D. Vinso, "Language as a multimodal phenomenon: implications for language learning, processing and evolution," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1651, p. 20130292, Sep. 2014.

[12] S. Kita, "Pointing: A foundational building block of human communication," in S. Kita (ed) *Pointing*. Brandon, VT: Psychology Press, pp. 9–16, 2003.

[13] M. Khalili, R. Rahmany, and A. A. Zarei, "The Effect of Using Gesture on Resolving Lexical Ambiguity in L2," *Journal of Language Teaching & Research*, vol. 5, no. 5, pp. 1139–1146, Sep. 2014.

[14] W. G. Smith, and C. L. H. Kam, "Children's use of gesture in ambiguous pronoun interpretation," *Journal of Child Language*, vol. 42, no. 3, pp. 591–617, Feb. 2015.

[15] C. Y. Tseng, S. H. Pin, Y. Lee, H. M. Wang, and Y. C. Che, "Fluent speech prosody: Framework and modeling," *Speech Communication*, vol. 46, no. 3-4, pp. 284–309, Jul. 2005.

[16] A. Wennerstrom, *The music of everyday speech: Prosody and discourse analysis*. Oxford: Oxford University Press, 2001.

[17] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, Jul. 2011.

[18] A. J. Schafer, S. R. Speer, and P. Warren, "Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task," in J. C. Trueswell and M. K. Tanenhaus (eds) *Approaches to Studying World-situated Language Use*. Cambridge: MIT Press, pp. 209–225, 2005

[19] J. Snedeker, and J. Trueswell, "Using prosody to avoid ambiguity: Effects of speaker awareness and referential context," *Journal of Memory and Language*, vol. 48, no. 1, pp. 103–130, Jan. 2003.

[20] Y. Lamekina, and L. Meyer, "Entrainment to speech prosody influences subsequent sentence comprehension," *Language, Cognition and Neuroscience*, vol. 38, no. 3, pp. 263–276, Aug. 2023.

[21] B. Z. Keller, "Ageing and speech prosody," in *Speech Prosody 2006*, Dresden, Germany, pp. 1–5, May 2-5, 2006.

[22] J. J. Diehl, L. Bennetto, D. Watson, C. Gunlogson, and J. Mcdonough, "Resolving ambiguity: A psycholinguistic approach to understanding prosody processing in high-functioning autism," *Brain and Language*, vol. 106, no. 2, pp. 144–152, Aug. 2008.

[23] T. M. Hopyan-Misakyan, K. A. Gordon, M. Dennis, and B. C. Papsi, "Recognition of affective speech prosody and facial affect in deaf children with unilateral right cochlear implants," *Child Neuropsychology*, vol. 15, no. 2, pp. 136–146, Nov. 2009.

[24] J. Grzybek, "Polysemy, homonymy and other sources of ambiguity in the language of Chinese contracts," *Czasopisma Naukowe/Journals*, vol. 1, pp. 207–215, Dec. 2009.

[25] B. Grzyb, S. L. Frank, and G. Vigliocco, "Communicative efficiency in multimodal language," *Preprint*, 2022. Available: https://doi.org/10.31234/osf.io/a9wt3

[26] M. Rasenberg, W. Pouw, A. Özyürek, and M. Dingemanse, "The multimodal nature of communicative efficiency in social interaction," *Scientific Reports*, vol. 12, no. 1, pp. 19111, Nov. 2022.

[27] J. Henrich, S. J. Heine, and A. Norenzay, "The weirdest people in the world," *Behavioral and Brain Sciences*, vol. 33, no. 2-3, pp. 61–83, Jun. 2010.

[28] A. Brown, and M. Kamiya, "Gesture in the resolution of syntactic ambiguity: Negation and quantification in English," in *UK-CLC 2016 Conference Proceedings*, Bangor, Gwynedd, Jul. 18-21, 2016.

[29] H. Holle, C. Obermeier, M. Schmidt-Kassow, A. D. Friederici, J. Ward, and T. C. Gunter, "Gesture facilitates the syntactic analysis of speech," *Frontiers in Psychology*, vol. 3, p. 74, Mar. 2012.

[30] E. Kidd, and J. Holler, "Children's use of gesture to resolve lexical ambiguity," *Developmental Science*, vol. 12, no. 6, pp. 903–913, Nov. 2009.

[31] T. Okahisa, and A. Shirose, "Influence of hand gestures on prosodic disambiguation of syntactically ambiguous phrases," *Acoustical Science and Technology*, vol. 39, no. 2, pp. 171–174, Mar. 2018.

[32] W. Q. Yow, "Monolingual and bilingual preschoolers' use of gestures to interpret ambiguous pronouns," *Journal of Child Language*, vol. 42, no. 6, pp. 1394–1407, Nov. 2015.

[33] N. Botting, N. Riches, M. Gaynor, and G. Morg, "Gesture production and comprehension in children with specific language impairment," *British Journal of Developmental Psychology*, vol. 28, no. 1, pp. 51–69, Dec. 2010.

[34] C. Scholl, and S. Mcroy, "Using gestures to resolve lexical ambiguity in storytelling with humanoid robots," *Dialogue & Discourse*, vol. 10, no. 1, pp. 20–33, Feb. 2019.

[35] D. Weerakoon, V. Subbaraju, N. Karumpulli, T. Tran, Q. Xu, U. X. Tan, J. H. Lim, and A. Misra, "Gesture enhanced comprehension of ambiguous human-to-robot instructions," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, Virtual Event, Netherlands, pp. 251–259, Oct. 25-29, 2020.

[36] B. R. Huang, and W. Li, *Xiandai hanyu*. Beijing: Beijing Book Co, 2012.

[37] D. A. Watkins, and J. B. Biggs, *The Chinese learner: Cultural, psychological, and contextual influences*. Hong Kong: Hong Kong University Press, 1996.

[38] S. Kita, "Cross-cultural variation of speech-accompanying gesture: A review," *Language and Cognitive Processes*, vol. 24, no. 2, pp. 145–167, Jan. 2009.

[39] Z. Field, J. Miles, and A. Field, *Discovering statistics using R. Discovering Statistics Using R*. London: Sage Publications Ltd, 2012.

[40] P. Boersma, and V. Van Heuve, "Speak and unSpeak with PRAAT," *Glot International*, vol. 5, no. 9/10, pp. 341–347, Jan. 2001.

[41] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A professional framework for multimodality research," in *5th international conference on language resources and evaluation (LREC 2006)*, Genoa, Italy, pp. 1556–1559, May 24-26, 2006.

[42] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[43] M. Graziano, E. Nicoladis, and P. Marentette, "How referential gestures align with speech: Evidence from monolingual and bilingual speakers," *Language Learning*, vol. 70, no. 1, pp. 266–304, Aug. 2020.

[44] I. Vila-Gimenez, and P. Prieto, "The value of non-referential gestures: A systematic review of their cognitive and linguistic effects in children's language development," *Children*, vol. 8, no. 2, p. 148, Feb. 2021.

[45] V. A. Brown, "An introduction to linear mixed-effects modeling in R," *Advances in Methods and Practices in Psychological Science*, vol. 4, no. 1, Mar. 2021.

[46] A. Jongman, Y. Wang, C. Moore, and J. Sereno, "Perception and production of mandarin tone," in P. Li, L. H. Tan, E. Bates, and O. J. L. Tzeng (eds) *Handbook of East Asian Psycholinguistics*. Cambridge: Cambridge University Press, pp. 209–217, 2006.

[47] E. Krahmer, and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, Oct. 2007.

[48] S. Seyfarth, "Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation," *Cognition*, vol. 133, no. 1, pp. 140–155, Oct. 2014.

[49] J. Holler, and G. Beattie, "Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener," *Gesture*, vol. 3, no. 2, pp. 127–154, Dec. 2003.