Should Robots Sound more like Machines than like Humans? User Expectations Affect the Perception of Prosody in TTS Voices

Ha Eun Shim [1,2], *Paige Tuttösí* [3,4], *Olivia Yung* [1], *Ivan Fong* [1], *Sara Ng* [1], *Angelica Lim* [4], *Yue Wang* [1], *H. Henny Yeung* [1,5]

[1] Department of Linguistics, Simon Fraser University, Canada; [2] Department of Linguistics, University of Massachusetts Amherst, USA; [3] Enchanted Tools, France; [4] School of Computing Science, Simon Fraser University, Canada; [5] Integrated Neuroscience and Cognition Center (UMR 8002), CNRS & Université Paris Cité, France

ha_eun_shim@sfu.ca; haeunshim@umass.edu

Abstract

Prosodic and indexical features of artificial voices, like naturalness, tone, and emotion, are often treated as distinct from intelligibility. Yet prosody can be critical: Consider, "She has to pull Bart" (without commas/prosodic boundaries) versus "She has to pull, Bart" (with commas/prosodic boundaries). Here, we asked how humans disambiguated such sentences while listening to a robot avatar, as prosodic and indexical features in the robot voice were manipulated. Prosodic pitch changes to make the voice either affective or monotonic had little effect on intelligibility. Surprisingly, indexical changes to increase machine-like (and less human-like) voice quality, despite being phonetically irrelevant to this prosodic contrast, *increased* intelligibility, but only for sentences with commas. Results imply that hearing machine-like voices changes listener expectations, biasing or pushing listeners to hear prosodic boundaries (i.e., commas) that are otherwise hard to detect.

Index Terms: speech, prosody, human-computer interaction, voice quality, pitch, text-to-speech (TTS), artificial voice

1. Introduction

1.1. Background

Current work on TTS voice design has largely focused on prosodic (i.e., cues for speech melody) and indexical (i.e., cues that signal speaker identity, gender, accent, etc.) features to improve user experience in human-robot interaction, using measures such as satisfaction, anthropomorphism, perceived naturalness, emotional expressiveness, and trust [1, 2, 3, 4, 5, 6]. However, what remains *less* clear is how these design choices shape *speech understanding*, or intelligibility.

Intelligibility refers to a listener's ability to correctly map a spoken sentence to its intended meaning. Work on TTS voices has historically treated intelligibility as separate from prosodic and indexical factors, using comprehension (e.g., transcription accuracy), as opposed to qualitative, questionnaire-based measures, such as the Mean Opinion Scale (MOS) or Godspeed questionnaires [7, 8]. Indeed, while "extra-linguistic" prosodic and indexical cues, such as voice quality, timbre, pitch variation, speech or disfluency rates, systematically influence how listeners perceive and rate the quality of their conversations or interlocutors, these changes are not often considered in relation to intelligibility [9, 10, 11].

However, prosodic features *can* influence intelligibility under certain circumstances. For example, prosody and intelligibility become related concepts in syntactically ambiguous sentences: While intelligible word by word, sentential meaning can remain ambiguous without accurate prosody. Previous literature shows that prosody can influence language comprehension, especially in resolving syntactic ambiguities [12, 13]. Relative clause attachment is one such example. For instance, with the sentence "Tap the frog with the flower," two interpretations are possible: "Tap // the frog with the flower" and "Tap the frog // with the flower" [13]. The sentence is ambiguous because the prepositional phrase (PP), *with the flower*, can attach either to the noun phrase, the frog, or to the verb phrase, the tapping event. With this ambiguity, the location of a prosodic boundary helps listeners disambiguate the syntactic attachment of the PP.

In this study, we therefore examine how prosodic and indexical cues of TTS voices can be critical for intelligibility, especially for syntactic disambiguation marked by prosodic boundaries. Although modern TTS systems are now very good at mimicking human voices, they still struggle with subtle prosodic cues that listeners use to resolve structure and meaning [14, 15, 16]. Our research question addresses how different voice properties affect prosodic disambiguation of sentence meanings in TTS voices and which aspects of voice design support prosody perception in TTS speech. Specifically, we manipulated two properties: Prosodic pitch changes to make voices sound more monotonic and indexical voice quality changes to make voices sound more machine-like. We used both qualitative, questionnaire-based evaluations and objective measures of intelligibility to identify properties of synthesized voices that better support prosody-driven comprehension.

1.2. Current Study

We examined two dimensions of TTS voices that are not typically associated with intelligibility. First, we compared listeners' perception of prosodically ambiguous sentences produced in an affective voice, which had greater pitch variation, and a monotonic voice, which had reduced pitch variation. While affective speech can involve other acoustic correlates beyond pitch, such as intensity and duration, we focused on pitch variation. Note that the degree of pitch change or variation is a phonetic cue that listeners use to mark prosodic boundaries [17, 18, 19].

Second, we examined whether speaker identity, cued by timbre changes in a voice to make it sound more machine-like and less human-like, affected comprehension. Note that changes to voice quality of the type we used here are not directly related to the phonetic cues that humans use to mark prosodic boundaries [20, 21].

2. Method

2.1. Stimuli

We adopted two stimulus types from previous studies: "Direct Address" (DA) and "List" sentences [14, 22]. In both types, the location of prosodic boundaries, represented orthographically by commas, signals different syntactic structures (see Table 1). Each sentence occurred in a pair, contrasted based on punctuation in two conditions: **With Comma** or **Without Comma**. Each member of the pair shared the same lexical sequence, differing in prosodic boundary locations.

We also created pairs of filler stimuli that follow the same sequence of "DA" and "List" sentences. However, instead of contrasting prosodic breaks, filler pairs differed in the tensity of the vowel (e.g./hit/ (heat) vs /hɪt/ (hit)). When perceiving tense and lax vowels, local changes in speech rate can influence which vowel is heard [23, 24], and they are distinct from global speech rate changes that influence evaluative questions about an artificial voice (i.e., impression scores). Fillers were pronounced using the prosodic structures of the With Comma sentences.

Overall, there were six pairs of experimental stimuli for each sentence type ("DA" or "List") and four pairs of tense/lax fillers for each sentence type. Thus, each participant heard 40 unique sentences in the study.

Table 1: *Stimuli examples: List, DA, and Fillers*

Type	**Stimuli***
DA: With Comma	She has to pull, Bart.
DA: Without Comma	She has to pull Bart.
List: With Comma	The mom drove her kids, Eugene, and Adele.
List: Without Comma	The mom drove her kids Eugene and Adele.
Filler (DA): Tense	I felt the heat, man.
Filler (DA): Lax	I felt the hit, man.
Filler (List): Tense	The girl saw a beach, a boat, and a dock.
Filler (List): Lax	The girl saw a beach, a boat, and a duck.

2.2. Voice Condition

We created four voice conditions: Human-like Affective, Human-like Monotonic, Machine-like Affective, and Machine-like Monotonic. Given that open-source TTS systems typically have difficulty making prosodic contrasts, we synthesized recordings of all stimuli using a fine-tuned Matcha-TTS, a neural TTS system designed to be human-like in voice quality [14, 25]. For fine-tuning, a female native English speaker read and recorded sentences contrasting in target prosodic forms, which were used as training input to fine-tune Matcha-TTS to emphasize comma-related prosodic contrasts [14].

The output of the fine-tuned Matcha-TTS constituted the Human-like Affective condition. The other three conditions were derived from this output. To produce Monotonic voices with reduced pitch variation, we used Praat Vocal Toolkit's command *Change pitch median and variation* set to 40% (i.e., a forty percent reduction in pitch height and variation). To generate Machine-like voices with a robotic vocal quality, we used the command *Delay* 0.02s with an *Amplitude* of 0.5 Pa. To generate the Machine-like Monotonic voice, both manipulations were applied to the Human-like Affective voice [26, 27].

2.3. Procedure

To ensure participants could complete the study in a timely fashion (≈30 minutes), each individual only heard a single voice condition. The first phase of the interaction task consisted of a dialogue between a human participant and a robot avatar, during which the participant listened to a robot speak a sentence and confirmed comprehension through a picture-matching exercise, ostensibly in order to "train" the avatar. The second phase involved participants completing a questionnaire (see Table 2).

2.3.1. Robot Training Task

Our task asked participants to provide feedback to the robot avatar to communicate more clearly. In each exchange, they saw two images: a *target* and a *non-target* image. Two practice trials were given to illustrate how these two images represented a potentially ambiguous meaning (as seen in Table 1). In each trial, the avatar would attempt to say a sentence that matched the *target* image, but might have erred in pronunciation (for filler items) and prosody or intonation (for test items), thereby changing the meaning of the sentence and matching the *non-target* image. Written text of the stimulus was never shown to participants. Participants were then asked to provide feedback on whether the avatar's voice matched or mismatched the target image. The feedback was recorded as a button selection.

If participants thought the avatar spoke correctly, they would simply repeat the sentence. If not, alternatively, they would correct it by repeating the sentence with the accurate pronunciation or intonation. Participants were able to listen to their recorded feedback at will; however, they had only one additional attempt to re-record feedback, which was kept as the final recording. An example trial is shown in Figure 1.

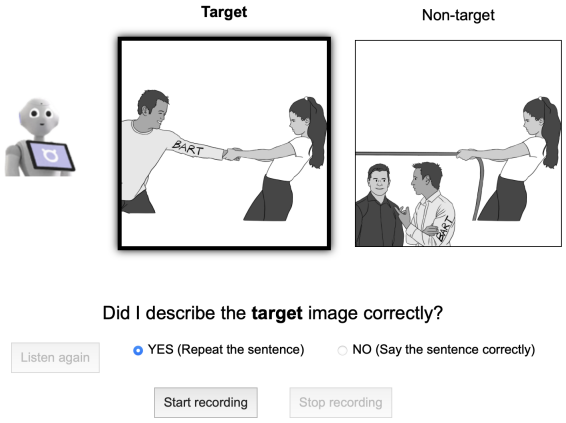

Figure 1: *Participants evaluated the robot avatar producing the sentence, "She has to pull Bart," as either matching (correct) or not matching the target image (incorrect).*

2.3.2. Questionnaires

We first asked two questions about participants' familiarity with AI, and then evaluated their perception of the robot's voice with six questions. For the latter part, we followed a previous study [4], whose short version of the MOS questionnaire, MOS-X2

Table 2: *Questionnaire*

Questions: Rated from 0 (negative) to 10 (positive)
Intelligible: Please rate the extent to which it was easy or difficult to understand what the voice was saying
Natural: How natural (pleasantly human-like) was the sound of the voice?
Prosodic: To what extent were the elements of timing, pitch, and emphasis appropriate for the messages?
Sociable: To what extent was the tone of voice socially and emotionally appropriate for the messages?
Competent: Please rate your impression of the robot (Incompetent or Competent)
Intelligent: Please rate your impression of the robot (Unintelligent or Intelligent)

[7], contained four items asking whether the voice sounded intelligible, natural, prosodic(ally appropriate), and sociable, paired with two items from the perceived intelligence dimension of the Godspeed questionnaire [8]. We used the same format as [4], which standardized responses on an 11-point scale (0–10) across MOS-X2 and Godspeed items (see Table 2).

2.4. Participants

160 participants were recruited from Prolific. All participants were native English speakers residing in the United States or Canada and were between 19 and 30 years old ($M_{\text{age}} = 25.34$, $SD = 3.18$). The sample included 91 females and 68 males (one participant preferred not to indicate gender).

We recorded 40 speakers in each Voice Condition. The sample size was informed by prior studies that manipulated pitch and measured comprehension, mimicking the affective versus monotonic condition in our study [28, 29]. They showed that degrading pitch cues yielded significant accuracy differences in interpreting ambiguous syntactic distinctions. Because power estimation for mixed-effects models is difficult without pilot or simulated data, we used G*Power to approximate the sample size based on the smallest effect reported in [28] (partial $\eta^2 = .103$). G*Power suggested that a similar ANOVA with four groups would require n = 154 to reach 95% power ($\alpha = 0.05$), which we adopted as a conservative benchmark.

3. Results

Our primary analysis on the robot training task was preregistered[1]. We noted there that we would also analyze questionnaires. While the present questionnaire analysis is different from the preregistration, it follows the general logic of our preregistered analysis and is described in the following section.

3.1. Robot training task data

In each trial, we measured whether a participant decided that the robot avatar's voice mismatched (0) or matched (1) the target image. These responses were entered into a mixed-effects logistic regression with several predictors: Whether the target sentence was actually a match or not (**Picture Match**, dummy coded); **Voice Condition** (which of the four voices was heard; contrast coded as below); **Comma Condition** (With or Without Comma; dummy coded); and the full set of interactions between these three variables.

[1]Preregistration: https://osf.io/8ka5y

Voice Condition was coded such that the first **affectiveness contrast** evaluated intonation (the two affective versus the two monotonic conditions), the second **Human-like/Machine-like contrast** examined the effect of voice quality (the two human-like versus machine-like conditions), and the third **congruency contrast** explored whether human expectations about how intonation and voice quality were congruent (Human-like Affective and Machine-like Monotonic conditions versus Human-like Monotonic and Machine-like Affective conditions).

The maximal random-effects structure of this model contained an intercept of Item (also classified as either a "DA" or "List" sentence type) with a slope of Voice Condition, and an intercept of Participant with slopes of Picture Match and Comma Condition. Since this maximal model did not converge, we further reduced the model by removing one slope at a time. The maximal converged model had no slope on the item-intercept and only the slope of Picture Match on the participant-intercept. Results, evaluated using Type III Likelihood Ratio Tests (LRTs), are shown in Table 3. Critically, there was a 3-way interaction between Picture Match, Voice Condition, and Comma. We focused on this most complex interaction to understand the lower-order effects in the model.

Table 3: *Statistical results from omnibus models*

Model results for robot training data

Effect	χ^2	**p-value**
Picture Match	318.13	$<$ .001**
Voice Condition	7.57	.056 ˆ
Comma	225.80	$<$ .001**
Picture Match × Voice Condition	4.97	.17
Picture Match × Comma	458.47	$<$ .001**
Voice Condition × Comma	7.68	.053 ˆ
Picture Match × Voice Cond × Comma	9.93	.019 *

Model results for the questionnaire data

Effect	χ^2	**p-value**
Voice Condition	6.24	.10
Question Item	210.35	$<$ .001**
AI Familiarity	88.02	$<$ .001**
Voice Condition × Question Item	14.03	.52

Note: ** $p <$.001; * $p < .05$; ˆ $p < .10$

As shown in Figure 2, there was a strong effect of Picture Match: Participants rated target sentences as matching the picture when this was the case and vice versa (i.e., a measure of accuracy). This accuracy effect also interacted with the factor of Comma, such that participants rated the avatar's voice as not often marking prosodic boundaries, and were therefore more often correct when this was indeed the case in the Without Comma condition (i.e., the Picture Match × Comma effect). Crucially, this effect interacted with Voice Condition, where contrast analysis showed that Human-like differed from Machine-like voices ($\beta = -0.51$, $SE = 0.22$, $z = -2.29$, $p = .022$), and Incongruent from Congruent voices to a smaller degree ($\beta = 0.44$, $SE = 0.22$, $z = 1.99$, $p = .046$). Overall, Machine-like voices showed higher accuracy in the With Comma condition.

3.2. Questionnaire data

Our analysis of questionnaire data involved a composite score from the two AI Familiarity questions, as well as the four question items from MOS-X2 and two question items from God-

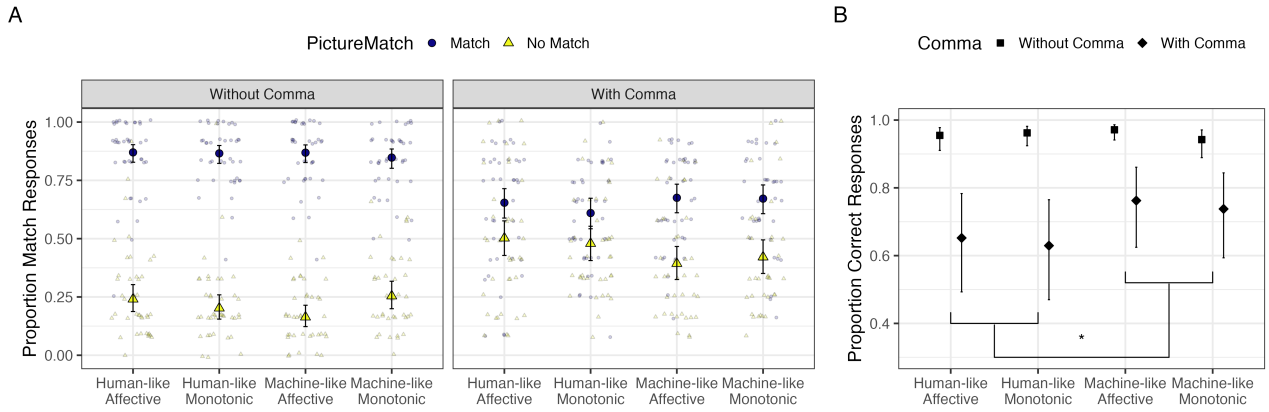

Figure 2: *Small dots represent participant data, while large dots and error bars represent model estimates and 95% CIs.* ***A:*** *Results showing the interaction between Picture Match, Comma, and Voice Condition.* ***B:*** *Model estimates of correct responses across Voice Condition and Comma conditions. Brackets show the distinct patterns of machine-like versus human-like voices (*p < .05).*

speed. Those latter six questions (Table 2) were on an 11-point Likert scale, such that higher ratings always indicated better voice evaluations. They were predicted using a cumulative link model (ordinal package in R [30]) by the composite AI-Familiarity score, Voice Condition, Question Item (the six questions in Table 2), and the interaction of Voice Condition and Question Item.

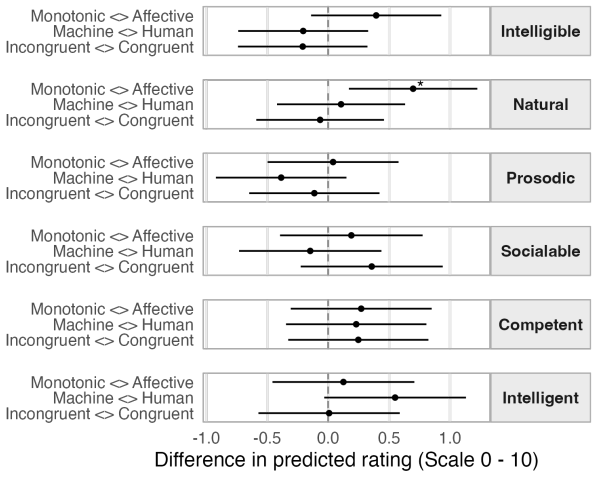

Figure 3: *Model estimates of how ratings on questionnaire items changed for each Voice Condition contrast. Error bars indicate 95% CIs (* p < .01).*

Results are shown in Table 3: AI Familiarity predicted more positive responses, and question items differed in mean rating. Critically, there was no main effect or interaction of Voice Condition (all p's > .10). Figure 3 shows (uncorrected) post-hoc comparisons of affectiveness, human/machine, and congruency contrasts on question items. Affective voices were rated as more natural ($p < .01$), while human-like voices were marginally rated as sounding smarter than machine-like ones ($p = .064$).

4. Discussion

We asked how prosodic and indexical cues impact the comprehension of syntactically ambiguous sentences in TTS voices. Our results revealed that understanding TTS voices that produced syntactic ambiguities—which can be resolved through prosody—was very challenging for human users: Listeners often missed prosodic boundaries (i.e., commas). Pitch changes to produce affective versus monotonic intonation did not impact intelligibility, although did improve participants' evaluative ratings of naturalness. Surprisingly, however, intelligibility increased when the voices sounded *more* machine-like (Figure 2), which participants also rated as marginally "less intelligent" (Figure 3), likely because voice quality altered listeners' expectations about what TTS voices are capable of articulating.

The locus of this effect remains unclear. One possibility is that participants were biased to assume robotic TTS voices are not making prosodic boundaries, hence the greater accuracy in the Without Comma condition. Crucially, a drop in accuracy in the With Comma condition was weaker for Machine-like voices, suggesting that participants assumed that the robotic voices *intended* to insert a prosodic boundary, even though phonetic cues to prosodic boundaries, including pause duration and pitch contours, were identical to Human-like voices.

A non-mutually exclusive alternative is that participants listened more carefully when hearing a "less intelligent" Machine-like voice, perhaps enhancing scrutiny of prosodic cues in our task. However, we note that participants did not show a clear affective advantage for Machine-like voices, nor did they show a clear machine-like advantage in the Without Comma condition (albeit ceiling performance in our current dataset). Future work will need to examine measures of listening effort to adjudicate among these possibilities.

Our study is not without limitations. Work on TTS voices typically uses within-subjects designs so that participants can compare voice manipulations. Future studies will thus include different avatars, permitting such comparisons, while still maintaining consistent expectations for individual avatars.

5. Conclusion

We showed that indexical aspects of TTS voices matter, likely by modifying users' expectations about an avatar's vocal capabilities, shaping how closely they listen to and/or interpret speech. TTS voices that better support prosodic interpretation are essential, not only for improving user experience, but also for reducing the risk of critical miscommunications.

6. Generative AI Use Disclosure

All authors are responsible and accountable for the work and content of the paper. We disclose that generative AI was only used for editing the manuscript for minor revisions. It was not used for producing a significant part of the manuscript.

7. References

[1] C. McGinn and I. Torre, "Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 211–221.

[2] F. Alonso Martin, M. Malfaz, A. Castro-González, J. C. Castillo, and M. A. Salichs, "Four-features evaluation of text to speech systems for three social robots," *Electronics*, vol. 9, no. 2, 2020.

[3] A. K. Coyne and C. McGinn, "The effect of human prosody on comprehension of TTS robot speech," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023, pp. 1816–1822.

[4] J. Miniota, S. Wang, J. Beskow, J. Gustafson, É. Székely, and A. Pereiral, "Hi robot, it's not what you say, it's how you say it," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 307–314.

[5] P. Tuttösí, E. Hughson, A. Matsufuji, C. Zhang, and A. Lim, "Read the room: Adapting a robot's voice to ambient and social contexts," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3998–4005.

[6] P. Tuttösí, S. Mehta, Z. Syvenky, B. Burkanova, G. E. Henter, and A. Lim, "Emojivoice: Towards long-term controllable expressivity in robot speech," in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2025, pp. 657–664.

[7] J. R. Lewis, "Investigating MOS-X ratings of synthetic and human voices," *Voice Interaction Design*, vol. 2, no. 1, p. 22, 2018.

[8] C. Bartneck, E. Croft, and D. Kulic, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots," in *Proc. Metrics for HRI workshop, technical report*, vol. 471, 2008, pp. 37–44.

[9] S. Lev-Ari and B. Keysar, "Why don't we believe non-native speakers? The influence of accent on credibility," *Journal of experimental social psychology*, vol. 46, no. 6, pp. 1093–1096, 2010.

[10] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[11] C. Nass and K. M. Lee, "Does computer-generated speech manifest personality? An experimental test of similarity-attraction," in *Proc. of the SIGCHI conference on Human Factors in Computing Systems*, 2000, pp. 329–336.

[12] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 905–945, 2010.

[13] 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, 2003.

[14] H. E. Shim, O. Yung, P. Tuttösí, B. Kwan, A. Lim, Y. Wang, and H. H. Yeung, "Generating consistent prosodic patterns from open-source TTS systems," in *Proc. INTERSPEECH 2025*, 2025, pp. 5383–5387.

[15] C. Pouw, A. Alishahi, and W. Zuidema, "A linguistically motivated analysis of intonational phrasing in text-to-speech systems: Revealing gaps in syntactic sensitivity," in *Proc. of the 29th Conference on Computational Natural Language Learning*, 2025, pp. 126–140.

[16] M. Domínguez, M. Farrús, and L. Wanner, "The information structure–prosody interface in text-to-speech technologies. an empirical perspective," *Corpus Linguistics and Linguistic Theory*, vol. 18, no. 2, pp. 419–445, 2022.

[17] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[18] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, 2010.

[19] S.-A. Jun, *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press, 2006, vol. 1.

[20] J. Kreiman and D. Sidtis, *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, 2011.

[21] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[22] N. Dehé, *Parentheticals in spoken English: The syntax-prosody relation*. Cambridge University Press, 2014.

[23] R. S. Newman and J. R. Sawusch, "Perceptual normalization for speaking rate: Effects of temporal distance," *Perception & psychophysics*, vol. 58, no. 4, pp. 540–560, 1996.

[24] P. Tuttösí, H. H. Yeung, Y. Wang, J.-J. Aucouturier, and A. Lim, "You sound a little tense: L2 tailored clear TTS using durational vowel properties," in *13th edition of the Speech Synthesis Workshop*, 2025, pp. 228–235.

[25] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 341–11 345.

[26] S. Wilson and R. K. Moore, "Robot, alien and cartoon voices: Implications for speech-enabled systems," in *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, 2017, pp. 40–44.

[27] G. Zellou, M. Cohn, and A. Pycha, "Listener beliefs and perceptual learning: differences between device and human guises," *Language*, vol. 99, no. 4, pp. 692–725, 2023.

[28] V. Sharpe, D. Fogerty, and D.-B. den Ouden, "The role of fundamental frequency and temporal envelope in processing sentences with temporary syntactic ambiguities," *Language and Speech*, vol. 60, no. 3, pp. 399–426, 2017.

[29] A. Goto, "The effects of prosodic cues on auditory sentence processing: An analysis focusing on early and late closure," *LET Journal of Central Japan*, vol. 27, pp. 11–21, 2016.

[30] R. H. B. Christensen, *ordinal—Regression Models for Ordinal Data*, 2025, r package version 2025.12-29. [Online]. Available: https://CRAN.R-project.org/package=ordinal