

An exploratory eye-tracking study into the time course of sincere and sarcastic speech recognition

Isabella Noor Warner¹, Sueah Kim¹, Thomas Zhao¹, Ryuki Matsuura¹, Seth Wiener¹

¹ Language Acquisition, Processing and Pedagogy Lab, Carnegie Mellon University

iwarner@andrew.cmu.edu, sueahk@andrew.cmu.edu, thomasz2@andrew.cmu.edu,

rmatsuur@andrew.cmu.edu, sethw1@cmu.edu

Abstract

This exploratory study examined how the prosodic characteristics of sincere and sarcastic speech are integrated in real time to understand a speaker's intended meaning. Participants heard a speaker produce a statement ("The book is interesting") followed by another speaker produce a response ("It's so gripping that I can't believe it was a human who wrote it"). Responses were produced with either sincere or sarcastic prosody following previously published acoustic characteristics. While listening to the speech, participants were shown two AI generated cartoon images depicting the sincere or sarcastic meaning. Participants' eye movements were recorded and analyzed. Eye-tracking results showed no difference in the timing of looks to the target sarcastic and sincere images. Additionally, the stimuli's fundamental frequency (F0) and harmonic-to-noise ratio (HNR) significantly predicted looks to targets in line with previous studies on the acoustics of sarcastic and sincere speech. These preliminary findings are discussed within the interactive and modular view of sarcasm processing.

Index Terms: speech recognition, sarcasm, eye-tracking

1. Introduction

Sarcasm and verbal irony are subtypes of communication whose communicative meaning contradict their literal meaning. This is accomplished with a number of speech cues, among which are changes in prosody, that is, acoustic manipulations of intonation and stress patterns. Research suggests that speakers are capable of discerning sarcasm with the aid of prosodic cues, even in testing conditions devoid of communicative context [1]. Even when controlling for semantic cues, speakers still demonstrate a propensity for identifying sarcasm, typically highlighting acoustic cues such as a lower pitch and greater amplitude. However, empirical studies on acoustic cues in sarcastic speech are relatively limited and the results appear to be language specific. Amongst the most commonly researched cue is pitch (F0), though research is conflicted as to whether the "sarcastic voice" is associated with a higher or lower F0. [1] sought to address this gap by examining sarcastic productions of English speech (in contrast to positive humorous, sincere or neutral productions) for patterns in prosodic cues. They found that a reduction in F0 was the most prominent signaller for sarcasm in English speech, followed by F0 variation and reductions in harmonic-to-noise ratio (HNR) [1]. Importantly, these findings

do not hold true across languages. [2] suggested that previously conflicting research on prosodic patterns in sarcastic signalling may have been due to the fact that various studies were performed in differing languages. In a study on six native speakers, they investigated prosodic cues of sarcastic signalling in Cantonese, a language chosen for its complete linguistic separation from English [2]. Results suggested that prosodic cues, in this particular case, F0, are manipulated differently in different languages (see also [3] for Dutch data). Further, [4] demonstrated in a study with 20 native speakers of English and 20 of Cantonese that a given addressee cannot accurately predict sarcasm based on prosody in an unfamiliar language. That is, the specific prosodic patterns for communicating sarcasm seem to be language specific. All of this research suggests a complex relationship between acoustic cues, communicative intent, and perlocutionary effect in sarcastic contexts.

The present study turns to real-time processing of prosodic cues in spoken English through the use of eye-tracking. The primary goal of our exploratory study, therefore, is to determine the time-course of sarcastic and sincere processing and how acoustic cues affect eye movements. We draw on one previous eye-tracking study on sarcastic processing during reading [5]. This study examines whether sarcastic utterances require longer processing time than sincere ones. This theoretical assumption is guided by the two schools of thought on sarcastic processing: modular and interactive. Modular accounts suggest that sarcastic meanings will always be processed slower than sincere counterparts, in line with the traditional standard pragmatic model [6]. Because, following Gricean principles, an interlocutor must first interpret and only afterwards reject the literal meaning of the utterance, the modular account dictates that utterances with a sincere communicative intent (that is, utterances that are meant to be understood in the literal sense) will always be interpreted faster than those with sarcastic or ironic meanings. On the other hand, the interactive account suggests that sarcastic intent can be accessed directly in supported contexts. This account puts emphasis on the importance of situational cues, including visual cues, in aiding perception, arguing that ironic and literal meanings can be assessed in parallel, and in congruence with aiding factors [5]. One particular factor which may aid in this perception is visual cues. Particularly, if the surrounding context of the conversation is visually incongruent with the literal meaning of the utterance, then this may be integrated into listener perception in real time. [5] explored these differing accounts in their eye-tracking study

whose results were inconsistent with the standard pragmatic model, suggesting that surrounding cues may have an impact on listener perception. That study, however, focused on reading times, whereas the present study focuses on listeners' responses to audio and images. In pairing the audio with visual representations of the utterance content, we provide the participants with visual context, which, following the interactive model, would predict that processing time for sarcasm would be equivalent to or possibly even less than sincere processing time. Guided by [1, 2, 4], we believe eye movements will be best predicted by F0 and HNR as these acoustic cues are the primary cues involved in English sarcasm production.

2. Methods

2.1. Participants

100 participants were initially recruited and paid via Prolific. Rather than use the term 'native speaker' (see [20]), we required that all participants had indicated in their Prolific profile that their first, primary, and fluent language was English and that they were born in the U.S. Moreover, all participants were required to confirm that they had no speech or hearing disorders, and were between the ages of 18 and 50. Participants were removed if they did not consent to the study (N = 7), failed the dichotic pitch task screening [7] (N = 13), failed the eye-tracking calibration (N = 32), or had a median frame rate less than 5 Hz (N = 2). This left 46 participants reported here (mean age 37.8 years old).

2.2. Materials

A set of short statement-response sentence pairs was created. Each statement expressed a positive sentiment toward a concrete, visually identifiable topic, e.g., "That book is interesting." These statements contained five to eight syllables and were followed by responses designed to be interpretable as either sincere or sarcastic responses to the aforementioned sentiment, depending solely on the perceived acoustics, e.g., "I think it's so gripping that I can't believe it was a human who wrote it in the first place." Responses contained a three syllable, 440 ms carrier phrase (e.g., "I think it's") followed by the what [1] calls a 'keyphrase' (e.g., "so gripping") in which the acoustics were modified given the intended prosody. All responses contained either 21 or 22 syllables. From this larger set of stimuli, 20 items were selected as the best statement-responses. To promote the sense of a natural dialog, the statements were recorded by a male speaker; the responses were recorded by a female speaker. Recordings were done in quiet rooms using built-in laptop microphones. The responses, which were analyzed for their acoustics, were recorded at 44,100 Hz and saved as wav files. This resulted in 40 total responses (20 items x 2 prosodies).

Images for the visual stimuli were generated using two artificial intelligence image models: ChatGPT 5.0 and Google AI Studio Gemini 2.5 Flash Image. The base prompt described the purpose of the study as an eye-tracking experiment investigating how prosody (sincere vs. sarcastic) influences gaze patterns toward corresponding images. The prompt instructed the models to produce paired depictions of the same sentence; one rendered sincerely and one sarcastically. Each image was required to include a person and an object relevant to the sentence, maintain a consistent minimalist art style, and exclude any text. See our online supplementary materials for

further details including the full prompt and all images used in the study.

2.3. Procedure

The experiment was hosted on Gorilla [8]. Participants were required to use a computer with a web camera and began the experiment with a dichotic pitch task [7], which required that the participant wear wired headphones. This was followed by a 9-point eye-calibration set to reject participants with fewer than four successful points. Participants were given a max of two attempts at calibrating. The eye-tracking task consisted of 20 trials (no fillers; see [9] for discussion on this approach to maximize data points). Half of the targets were sincere and half were sarcastic. Two presentation lists were created in order to counterbalance the presented prosody of each item. On each trial, participants were shown a fixation cross while listening to the statement audio (e.g., "That party was crazy."). Following the statement, two images were shown on screen (placement of sarcastic and sincere targets and competitors were counterbalanced across trials). Figure 1 shows an example slide. After the 1000 ms preview time, the response audio was played ("Last night was so wild..."). Participants were told to click on the image that best matched the perceived audio. Eye movements were continuously recorded using the participant's web camera. In total, the experiment lasted roughly eight minutes.

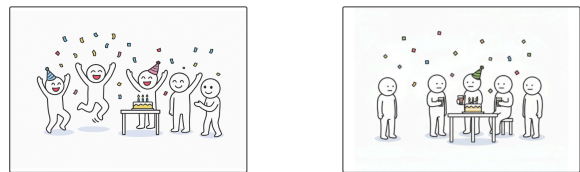


Figure 1: Example display showing sincere (left) and sarcastic (right) images.

2.4. Data analysis

Following [1], we extracted 12 acoustic measures from the stimuli (see online supplementary material for additional information). We employed a deep neural network-based Forced Alignment tool implemented with wav2vec 2.0 [10], as its substantial reliability has been reported across various speech genres (e.g., [11]). F0 and amplitude measures were extracted using the basic functions of Praat [12] with default values for floor and ceiling. Harmonic-to-Noise Ratio (HNR) was computed using PraatSauce [13]. Other measures (speech duration, speech rate, and 1/3 octave spectrogram) were calculated using Python. Crucially, for the purposes of our study, we measured the F0 contour in Hz and voice quality in dB over a series of 50 ms time windows in the keyphrase. F0 and HNR measurements 2.5 SDs beyond mean per prosody type were removed (~2%).

Eye-tracking data were wrangled following [14]. All trials in which participants did not correctly click on the intended image and clicks that were 2.5 SDs beyond mean per participant response time were removed. Further, two items in which over 90% of participants did not correctly identify the intended target (suggesting problematic stimuli) were likewise removed. Finally, only prediction frames were retained and off-screen fixations were removed. This left roughly 50% of the eye-tracking data with a median frame rate of 23.2 Hz,

which is in line with recent web-based eye-tracking studies e.g., [15, 16]. Fixation analyses were carried out with similar aggregate 50 ms time windows within the keyphrase to align with the acoustic data. We examined the first 1000 ms (i.e., 11 consecutive 50 ms time windows) of the keyphrase (adjusted by 200 ms to account for the time needed to launch a saccade [17]). We modeled looks to the target during this window using a generalized linear mixed model (GLMM) with a logit link function, implemented with the lme4 package in R (version 4.5.1). The model predicted looks to the target (1, 0) with fixed effects of scaled F0, scaled HNR, prosody type (two level factor with “sincere” as the reference level) and two-way and three-way interactions with prosody. By-participant and by-item random intercepts were included. Random slopes were not included as the model would not converge.

2.5. Data availability

Stimuli, data, code, and additional study information are available on the Open Science Framework: <https://osf.io/g8wbf/>

3. Results

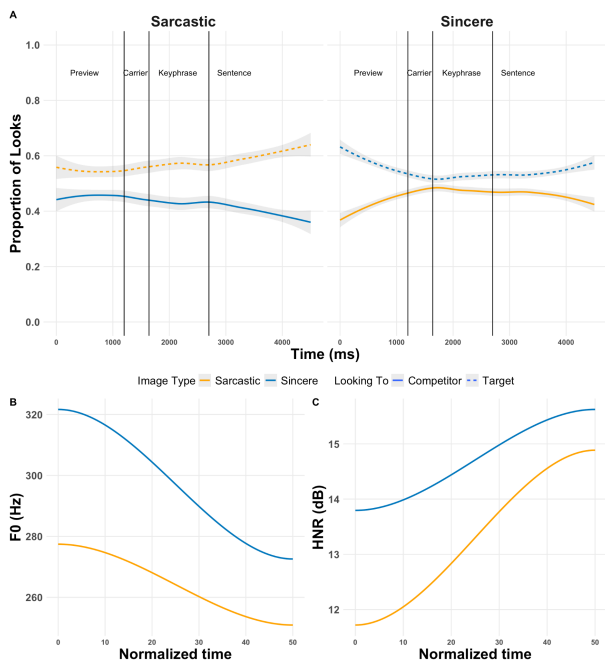


Figure 2: (A) Time course of looks in sarcastic (left) and sincere (right) trials to target (dotted line) and competitor (solid line). Vertical lines show timing of preview, carrier, and keyphrase adjusted by 200 ms. (B) F0 contours and (C) HNR contours over normalized time during keyphrase in sincere (blue) and sarcastic (orange) trials.

Figure 2A shows the time course of looks to the target (dotted line) and competitor (solid line) for sarcastic trials (left plot) and sincere trials (right plot). Any looks during the preview (i.e., first 1000 ms) occurred prior to acoustic information, and are therefore uninformative. The initial bias to the target is somewhat stronger in the sincere trials than the sarcastic trials.

This suggests problematic stimuli in some way. We return to this limitation in the discussion. During the carrier, keyphrase, and remaining sentence, participants continued to fixate on the intended target over time. Minimal competition was observed during the keyphrase and remaining sentence.

Figure 2B shows the keyphrases’ extracted F0 contours and HNR contours (Figure 2C) by prosody type. The figure shows differences between the sincere speech (blue) and sarcastic speech (orange) across normalized time on the x-axis. In both plots, sarcastic speech shows reduced F0 and HNR relative to sincere speech.

The mixed effects model (conditional $R^2 = .20$) revealed a significant effect of F0 ($\beta = .16$, $SE = .05$, $Z = 3.33$, $p < .001$). This effect indicates that as F0 increased, participants were more likely to look to the sincere target. A two-way F0 x sarcastic prosody interaction ($\beta = -.27$, $SE = .07$, $Z = -4.17$, $p < .001$), and a three-way F0 x HNR x sarcastic prosody interaction ($\beta = -.11$, $SE = .04$, $Z = -3.09$, $p = .002$) were found. Both of these interactions indicate that as F0 (and HNR) increased, participants were less likely to look at sarcastic targets. All other coefficients were null at a .05 alpha. This includes a null effect of prosody ($\beta = -.03$, $SE = .23$, $Z = -.12$, $p = .91$) indicating no difference in looks to the sarcastic and sincere targets (when not accounting for acoustics), a null effect of HNR ($\beta = .02$, $SE = .04$, $Z = .69$, $p = .49$), a null two-way HNR x sarcastic prosody interaction ($\beta = -.04$, $SE = .06$, $Z = -.69$, $p = .49$), and a null three way F0 x HNR x sincere prosody interaction ($\beta = .01$, $SE = .04$, $Z = .15$, $p = .88$). Figure 3 plots the model’s log-odds estimates with standard errors.

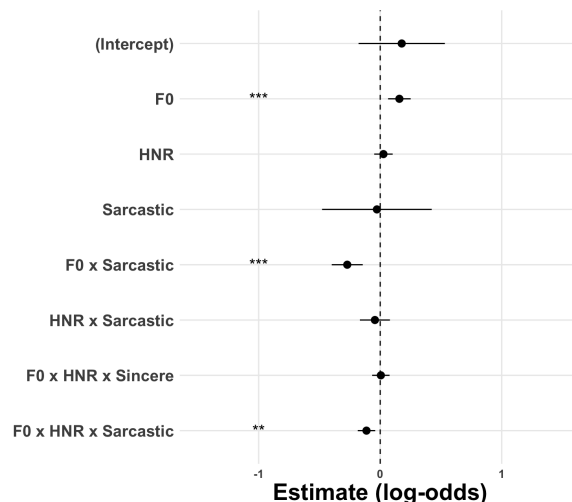


Figure 3: Model output for GLMM model. Points represent model estimates with standard errors. Asterisks represent p-values (**.01, ***.001).

4. Discussion and conclusion

Our exploratory study analyzed eye movements to assess how the acoustics of sarcastic and sincere speech is used in real-time speech recognition. We presented participants with AI generated visuals representing either the sincere or sarcastic interpretation of an utterance (see Figure 1 for an example). Stimuli differed in the recorded keyphrases’ F0 and HNR as previously described for sincere and sarcastic English speech [1, 2]. We recorded different acoustics for each

prosody type (see Figure 2B and 2C) and presented these recordings to participants while recording their eye movements. We found effects of F0 and HNR on sarcastic prosody recognition during our keyphrase. The eye-tracking results showed that an increase in F0 alone (and together with an increase in HNR) led to fewer looks to sarcastic targets; an increase in F0 alone led to more looks to sincere targets. This finding is in line with previous research that suggests F0 is the primary prosodic indicator of English sarcasm and HNR serves as a secondary indicator [1, 2]. This finding also corroborates a previous eye-tracking study that shows the acoustic cues involved in emotional prosody cues can affect real-time gaze behavior [25]. Here we show novel evidence of real-time use of F0 and HNR to understand sentence-level prosody involving sarcastic and sincere speech. Participants reliably converged upon the intended interpretation over the course of the keyphrase and prior to the completion of the sentence (see Figure 2A). Whereas auditory input alone may not have necessarily led to a sincere or sarcastic expectation, the addition of a congruent (or incongruent) image potentially helped establish the speaker’s intended meaning. This exploratory finding may suggest that if visual and auditory information support the sarcastic interpretation, the listener does not need to first access the literal meaning, i.e., sarcastic intent can be accessed directly within supported contexts [5, 18, 19].

Beyond this broad claim that the acoustics of sarcastic and sincere prosody contribute to speech recognition, our results are fairly inconclusive. Whereas our participants looked to the intended target with minimal competition from the prosody competitor, our exploratory results are limited by a number of shortcomings with our design.

First, our design resulted in a bias towards certain sincere targets (Figure 2A, right panel). It is unclear why such a bias was present. We do not think the bias is enough to dismiss these data as uninformative—perhaps these images were intrinsically more interesting—but given that participants had not yet heard anything, this initial bias suggests a design flaw, despite our best efforts to normalize the images. Whereas the eye-fixation patterns observed in the keyphrase window showed no difference between looks to sarcastic and sincere targets, this interpretation of the data assumes no prior bias towards a target or competitor. Our data (Figure 2A) suggest otherwise and our observed null effect may actually reflect "recovery" from the initial bias (see [24] for discussion). Further studies will need to clarify this point. Therefore, our results cannot address the question we set out to explore: do prosodic cues and the added visual stimulation provide the necessary supporting context for participants to analyze the communicative intent of the utterance in an interactive manner or is processing still modular?

Second, our analysis was conducted on a small sample of participants (N = 46) and with relatively limited data after our stringent data wrangling steps. We also note that, despite norming our stimuli, certain trials resulted in incorrect target identification by more than 90% of the participants. Whether this was due to the design of the utterance, the image, the audio, or some combination of the three, is unclear. Further, our research design inherently reduced the naturalness of speech in the stimuli in favor of uniformity. For example, we only manipulated prosodic cues during the keyphrase window. In natural speech, both sincere and sarcastic productions would likely feature prosodic differences across the entire utterance, rather than two or three words. We sacrificed these

more natural speech contours in favor of a uniform time-window during which we could analyze eye-tracking data. Similarly, by accounting for semantic cues and standardizing our lexical choices, we reduce the naturalness of sarcastic and sincere speech. Related to this point, the terms "sarcastic" and "sincere" are not absolutes. One can be sincerely sarcastic or sarcastically sincere. The difficulties in determining the characteristics of a "sarcastic" prosody may lie in separating it from the syntactic-lexical tier and the situational context.

Third, all of our acoustic analysis measurements were automated. We acknowledge that human intervention in the form of precise annotation and textgrid creation would likely aid in the specificity and reliability of the results. In future research, we hope to address these limitations and conduct a more refined analysis. To that end, our time window (1000 ms) is also too large. We combined multiple time bins to increase statistical power. A more refined analysis (with more data) and specific hypotheses is needed to fully examine the dynamics of speech recognition. Other future research avenues include broadening the number of productions to include humorous, positive, neutral, or other contexts, and exploring these interactions in other languages. Given that prosodic cue manipulation patterns are dependent on the language, we anticipate this to be a ripe area for future research.

An accidental but noteworthy finding is the likelihood that the AI generated images were in some way flawed. Despite the growing consensus that AI-driven tools can benefit linguistics research (e.g., [21, 22, 23]), we found that using AI to create the visual stimuli for our visual world paradigm study was somewhat problematic. The limitation could be, in part, due to the specific task (see prompt details on OSF) and the complexity of subtle prosody differences. We acknowledge as well that our stimuli would have been difficult for humans to design. A more straightforward image creation prompt would potentially yield better results for experimenters.

In summary, we found that looks to the target images occurred prior to the end of the speaker’s utterance, suggesting incoming acoustics are processed in real-time to arrive at the intended prosody. Increases in F0 led to more looks to sincere targets; decreases in F0 and HNR led to fewer looks to sarcastic targets. This supports the claim that F0 is the primary cue for English sarcasm and HNR is a secondary cue [1, 2]. We also found participants looked to the sincere and sarcastic target with similar time courses and did not display a measurable delay for sarcastic as opposed to sincere utterances. Yet, due to the breadth of design flaws, our results are inconclusive such that we are unable to say whether our results support an interactive or modular view of sarcasm processing. Future research in this domain is needed.

5. Acknowledgements

We thank Jay Dasilva for his help creating the stimuli and the anonymous reviewers for providing critical feedback.

6. References

- [1] H. S. Cheang and M. D. Pell, "The sound of sarcasm," *Speech Comm.*, vol. 50, no. 5, pp. 366–381, 2008.

- [2] H. S. Cheang and M. D. Pell, "Acoustic markers of sarcasm in Cantonese and English," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1394–1405, 2009.
- [3] N. G. Jansen and A. Chen, "Prosodic encoding of sarcasm at the sentence level in Dutch," in *Proc. 10th Int. Conf. Speech Prosody*, 2020, pp. 409–413.
- [4] H. S. Cheang and M. D. Pell, "Recognizing sarcasm without language: A cross-linguistic study of English and Cantonese," *Pragmatics & Cognition*, vol. 19, no. 2, pp. 203–223, 2011, doi: 10.1075/pc.19.2.02che.
- [5] A. Turcan and R. Filik, "An eye-tracking investigation of written sarcasm comprehension: The roles of familiarity and context," *J. Exp. Psychol.: Learn. Mem. Cogn.*, vol. 42, no. 12, pp. 1867–1893, 2016, doi: 10.1037/xlm0000285.
- [6] H. P. Grice, "Logic and conversation," in *Syntax and Semantics, vol. 3: Speech Acts*, J. Morgan and P. Cole, Eds. New York, NY: Academic Press, 1975, pp. 41–58.
- [7] A. E. Milne *et al.*, "An online headphone screening test based on dichotic pitch," *Behav. Res. Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.
- [8] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, 2020.
- [9] Y. Lin, H. Rohde, and S. Wiener, "More participants, fewer trials: A silver lining of moving eye-tracking experiments online," in *Proc. 35th Annu. Human Sentence Processing Conf.*, 2022.
- [10] A. Baeovski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: <http://arxiv.org/abs/2006.11477>
- [11] R. Matsuura, S. Suzuki, K. Takizawa, M. Saeki, and Y. Matsuyama, "Gauging the validity of machine learning-based temporal feature annotation to measure fluency in speech automatically," *Res. Methods Appl. Linguist.*, vol. 4, no. 1, pp. 1–23, 2025, doi: 10.1016/j.rmal.2024.100177.
- [12] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*, 2024. [Online]. Available: <http://www.praat.org/>
- [13] J. Kirby and R. Puggaard-Rode, *PraatSauce: Praat-based tools for spectral processing*, Version 1.0.4, 2025. GitHub repository: [kirbyj/praa sauce](https://github.com/kirbyj/praa sauce).
- [14] A. A. Bramlett and S. Wiener, "The art of wrangling: Working with web-based visual world paradigm eye-tracking data in language research," *Ling. Approaches Biling.*, vol. 15, no. 4, pp. 538–570, 2025.
- [15] A. A. Bramlett and S. Wiener, "Focus (on) replication: Focus processing in L1 and L2 English using the fidelity, refinement, and exploratory extension (FiREE) replication framework," *Res. Methods Appl. Linguist.*, vol. 4, no. 3, 100275, 2025.
- [16] A. A. Bramlett and S. Wiener, "Individual differences modulate prediction of Italian words based on lexical stress: A close replication and LASSO extension of Sulpizio and McQueen (2012)," *J. Cult. Cogn. Sci.*, vol. 9, no. 1, pp. 55–81, 2025.
- [17] E. Matin, K. C. Shao, and K. R. Boff, "Saccadic overhead: Information-processing time with and without saccades," *Percept. Psychophys.*, vol. 53, no. 4, pp. 372–380, 1993.
- [18] A. N. Katz, D. G. Blasko, and V. A. Kazmerski, "Saying what you don't mean: Social influences on sarcastic language processing," *Current Dir in Psych Sci*, vol. 13, no. 5, pp. 186–189, 2004.
- [19] J. Woodland and D. Voyer, "Context and intonation in the perception of sarcasm," *Metaphor and Symbol*, vol. 26, no. 3, pp. 227–239, 2011.
- [20] B. Brown, B. Tuszmagambet, V. Rahming, C. Y. Tu, M. B. DeSalvo, and S. Wiener, "Searching for the 'native' speaker: A preregistered conceptual replication and extension of Reid, Trofimovich, and O'Brien (2019)," *App. Psycholing.*, vol. 44, no. 4, pp. 475–494, 2023.
- [21] S. Suzuki, H. Takatsu, R. Matsuura, M. Koyama, M. Saeki, and Y. Matsuyama, "Feedforwarding diagnostic language assessment: Artificial intelligence-(AI-) driven weakness identification and contextualised feedback for second language speaking," *Lang. Testing*, vol. 42, no. 4, pp. 476–507, 2025.
- [22] M. Malik, M. K., Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Apps.*, vol. 80, no. 6, pp. 9411–9457, 2021.
- [23] E. E. Jang, and Y. Sawaki, "Advancing language assessment for teaching and learning in the era of the artificial intelligence (AI) revolution: Promises and challenges," *Lang. Testing*, vol. 42, no. 4, pp. 361–368, 2025.
- [24] J. S. Magnuson, "Fixations in the visual world paradigm: where, when, why?" *J. of Cultural Cog. Sci.*, vol. 3, no. 2, pp. 113–139, 2019.
- [25] S. Paulmann, D. Titone, and M. D. Pell, "How emotional prosody guides your way: Evidence from eye movements," *Speech Comm.*, vol. 54, no. 1, pp. 92–107, 2012.