ELTIN JOURNAL:
Journal of English Language Teaching in Indonesia
p-ISSN 2339-1561
e–ISSN 2580-7684

SARCASM IN SYNTHETIC SPEECH: EVALUATING ELEVENLABS' TTS PROSODY FOR PRAGMATICS INSTRUCTION

M. Faisal Afif Zawawi[1*], Evynurul Laily Zen[2], Nurenzia Yannuar[3]
m.faisal.2402218@students.um.ac.id, evynurul.laily.fs@um.ac.id, nurenzia.yannuar.fs@um.ac.id

UNIVERSITAS NEGERI MALANG

Received: April 21, 2026; Accepted: June 10, 2026

ABSTRACT

This study examines how sarcasm is acoustically signaled by text-to-speech (TTS) and human encoders by analyzing key acoustic parameters including mean fundamental frequency (F0), F0 SD, F0 range, mean amplitude, amplitude range, and speech rate. The study compares sarcasm encoding in ElevenLabs, a commercially available advanced TTS model, and native American English speakers. A total of twelve encoders participated in the study, consisting of six TTS voices from ElevenLabs and six native American English speakers recruited through Fiverr. The dataset comprises 240 recordings, with each encoder generating both sincere and sarcastic versions of the same sentences. Acoustic parameters were extracted using *Praat* and computed using descriptive statistics and Linear Mixed Models (LMMs). The results reveal that sarcasm is consistently associated with reduced pitch, pitch variability, and slower speech rate, while amplitude does not differentiate significantly between attitudes. Subsequently, differences between TTS and human encoders are not uniform, which indicates that the TTS model approximates general prosodic patterns but differs in the magnitude of how these features are signaled. Overall, the findings suggest that TTS presents a partial approximation of human sarcastic prosody, capturing general acoustic patterns while remaining limited in expressive variability in sarcastic contexts.

Keywords: Sarcasm, Prosody, Text-to-Speech, Acoustic features, Pragmatics

A. INTRODUCTION

The rapid advancement of artificial intelligence (AI), specifically in text-to-speech (TTS) technology, has offered new opportunities for English as a Foreign Language (EFL) instruction. These technologies are increasingly employed to offer consistent and accessible spoken input for EFL learners. However, while TTS technology exceeds in generating intelligible and clear speech, challenges might persist in teaching more subtle aspects of pragmatic competence. One such challenge is sarcasm–an expression of verbal irony in which speaker conveys critical or negative attitudes through positive utterances (Gibbs, 2000; Haverkate, 1990; Kreuz & Glucksberg, 1989). Unlike humor, which is mainly associated with amusement (Cheang & Pell, 2008), sarcasm heavily relies on prosodic

features such as pitch, intonation, and rhythm instead of its lexical content. For instance, the sentence "You're such a good friend" can either express sarcasm or sincerity depending on how it is delivered. These suprasegmental features function as key cues that enable listeners to interpret speaker's intended meaning.

In the context of Indonesian EFL learners, detecting sarcasm presents additional challenges due to contrasts in linguistic and sociocultural norms. A study has reported that intonation patterns in Indonesian typically do not mark contrastive focus in the same way as English (Putri et al., 2019), and sarcasm in Indonesian contexts is frequently expressed more directly and associated with insult or rudeness (Lubis & Bahri, 2023; Syafruddin et al., 2021). Similarly, a study on Javanese language, the largest regional language in Indonesia, indicates that sarcasm is often delivered through indirect humor and shared cultural knowledge rather than prosody manipulation (Rahardi et al., 2024). Global cross-cultural evidence further suggests that the use and perception of sarcasm differ across sociocultural contexts, where factors such as communicative orientation and regional norms influence both how sarcasm is interpreted and expressed (Dress et al., 2008; Rockwell & Theriot, 2001). Additionally, listeners' perceptions are shaped by contextual familiarity and prior knowledge, which influence whether sarcasm is successfully identified (Blasko et al., 2021). Collectively, these findings indicate that EFL Indonesian learners may rely less on subtle acoustic cues and more on explicit lexical or contextual signals, which make the interpretation of English sarcasm specifically complicated.

Subsequently, TTS technologies have achieved prominence as pedagogical devices in EFL instruction. A growing body of research suggests that TTS can effectively support listening comprehension, pronunciation, and decoding skills, which frequently perform comparably to human instruction in controlled tasks (Al-Jarf, 2022; Bione et al., 2017; Cardoso, 2018). Furthermore, TTS has been reported to lower EFL learners' anxiety, enhance confidence, and offer intelligible and consistent model of spoken English (Amin, 2024; Chiang, 2019; Van Duong, 2022; Oktalia & Drajati, 2018; Önder, 2025; Xiao, 2025). Teachers and learners often perceive TTS as an authentic and reliable source of input, specifically in contexts where access to native speakers is limited. This emerging trend places TTS as a "standard model" of spoken English in many EFL settings.

Among many TTS models that are commercially available recently, ElevenLabs is among the models that is regarded as a top-tier AI voice generator widely (Inworld AI, 2026; Nuzhnyy, 2026), which represents as the contemporary synthetic speech technologies. Recent research has exhibited its capability to generate intelligible and highly natural speech, offering EFL learners to model their pronunciation based on realistic audio input (Mubarok et al., 2025). Additionally, EFL learners have reported positive impressions in using ElevenLabs as a low-pressure speaking partner that increases confidence and enhances fluency (Budianto et al., 2024), while its integration into communicative tasks such as creating podcast has also been revealed to improve speaking performance and lower language anxiety (Amaliah et al., 2025). From a technical aspect, ElevenLabs is capable of producing highly realistic audio across different platforms and contexts, making it compatible for multimodal and interactive learning environments (Chuan et al., 2025; Dewatri et al., 2023). Its utilization in digital storytelling further promotes its capability to generate contextually appropriate and engaging speech output (Kamal & Imran, 2026). Thus, ElevenLabs demonstrates the current state of advanced TTS technology that focuses on usability and naturalness in language learning contexts.

However, despite these advantages, recent studies underline key limitations in TTS technologies, specifically in their production of prosodic variation. Research has revealed that synthetic speech frequently displays prosodic simplification, shortcomings of rhythmic complexity, pitch variability, and voice quality features found in natural human speech (Rotaeta Perez, 2025). These shortcomings are particularly crucial for pragmatic phenomena such as sarcasm that heavily depend on subtle acoustic signals. An experimental study further exhibits that voice quality features such as creaky voice significantly influence the interpretation of speaker stance and sarcastic intent, yet these features usually involve manual adjustments in controllable TTS systems rather than being automatically produced by the TTS systems (Lameris et al., 2024). In addition, while TTS technology can assist basic prosodic learning such as rising intonation patterns, it remains limited in generating more complex or emotionally nuanced intonation (Phuong et al., 2022). This limitation is further shown in recent studies on TTS generation sarcasm speech, where even advanced TTS models only display modest improvements and robotic and unnatural qualities still persist, while human perceptions on the generations often depend on manual adjustment of acoustic features such as energy and pitch (Jiang, 2024; Li et al., 2023).

Insights from acoustic research on human sarcasm further highlight this shortcoming. Studies have constantly demonstrated that sarcastic utterance is associated by particular acoustic cues, which include lower fundamental frequency (F0), reduced pitch variability, slower speech rate, and distinct voice quality features (Anolli et al., 2000; Cheang & Pell, 2008; Rockwell, 2000, 2007; Yang, 2021). These features significantly differ from any other types of utterance such as neutral or sincere speech and play a key role in the perception of sarcasm. However, acoustic parameters of sarcasm vary across languages (Cheang & Pell, 2008, 2009; Loevenbruck et al., 2013; Niebuhr, 2014; Rao, 2013; Yang, 2021), demonstrating that perception and production of sarcasm are shaped by language-specific sociocultural and phonological factors. Such diversity further complicates EFL learners' competence in sarcastic intent detection (Koh et al., 2022), underscoring sarcasm's complexity in both interpretation and production phenomenon.

Given these considerations, a significant gap emerges. While TTS technologies, specifically ElevenLabs, are widely employed and trusted as models of spoken English for EFL instructors and learners (Amaliah et al., 2025; Budianto et al., 2024; Chuan et al., 2025; Dewatri et al., 2023; Kamal & Imran, 2026; Mubarok et al., 2025), it remains unanswered whether they can generate the perceptual and acoustic features of sarcastic speech in a way that approximates human speech. If TTS systems fail to approximate human's nuanced prosodic cues, their usability as pedagogical devices may accidentally provide EFL learners with misleading or incomplete models for pragmatic speech, particularly sarcastic speech.
To address this gap, the present study employs a quantitative approach to compare the acoustic parameters of sincere and sarcastic utterances generated by ElevenLabs, one of the top rated and commercially ready TTS platforms (Inworld AI, 2026; Nuzhnyy, 2026), with those produced by native human speakers. By focusing acoustic aspect, this present study aims to investigate the extent to which TTS system can approximate human-like sarcasm and to explore its implications for pragmatic instruction, particularly in detecting English sarcasm in the Indonesian EFL learning context.

B. METHOD

This study employed a quantitative comparative design (Creswell, 2014) to examine how sarcastic and sincere utterances are acoustically realized by TTS systems and human speakers. The analysis focused on key acoustic parameters as commonly used in sarcasm research (Cheang & Pell, 2008, 2009). A total of twelve encoders were used to generate sarcastic and sincere utterances. First, six synthetic voices from ElevenLabs' TTS models were chosen, which consisted of three male voices: (1) Jarnathan, (2) Liam, and (3) Hale and three female voices: (1) Laura, (2) Adeline, and (3) Eve. These voices were selected because they are available within ElevenLabs' V3 model, which is designed to produce emotionally rich and highly expressive speech (Dabkowski & Staniszewski, 2025). Secondly, mirroring the sample size employed by Cheang and Pell (2008), six native American English speakers, three males and three females, aged 30-52, were selected to record the same set of utterances in both sincere and sarcastic attitudes. These speakers were selected as human encoders via Fiverr, an online freelance platform, where they identified themselves as voice-over artists with years of experience. Their recruitment was based on their linguistic background as U.S. English and their demonstration of experience in voice recording, as shown in their professional profiles throughout the years. All selected speakers provided consent to participate in the study and were compensated for their participation. This balanced design in terms of type of encoders (TTS vs. human) and gender distribution provided a controlled comparison of acoustic parameters across conditions.

Data collection involved the recording of 240 utterances generated by 12 encoders, comprising six voices from ElevenLabs' TTS models and six native English speakers. Each encoder generated 20 utterances based on a set of 10 target sentences, generating each sentence twice: once with a sarcastic attitude and once with sincere attitude. Because the lexical content stayed identical between both attitudes, any observed differences could be primarily attributed to acoustic cues of sarcasm. Recordings of human speakers were conducted in controlled home studio environments and used high-quality condenser microphones and were instructed to generate each utterance as naturally as possible in both attitudes. To support consistent interpretation, each speaker was sent with the full list of target sentences with their corresponding situational contexts. The contexts were developed based on sarcasm scenarios described by Yang (2021) and were adapted to fit the newly generated target sentences employed in the present study. For synthetic voices, the same set of target sentences were inputted into ElevenLabs' TTS feature using prompt-based audio tags which indicated the intended attitude (i.e., [sarcastically] You're such a good friend). The produced audio files were downloaded from the platform with the best audio quality. These procedures were conducted to ensure comparability in recording consistency and quality between synthetic and human encoders.

Acoustic parameters were extracted using *Praat* software (Boersma & Weenink, 2025). A total of 240 audio recordings were analyzed of their six acoustic parameters, including mean F0 (mean pitch height), F0 range (difference between the lowest and the highest pitch values), F0 standard deviation (degree of pitch variability), mean amplitude (mean intensity level), amplitude range (difference between the lowest and the highest intensity values), and speech rate (number of syllables generated per second). First, pitch related extractions were conducted using *Praat*'s filtered autocorrelation method analysis, which is the current recommendation for speech intonation and vocal-fold vibration frequency analyses (Boersma & Weenink, 2025). Next, amplitude extraction was collected using *Praat*'s default

intensity analysis, which computes the root mean square energy of the audio. Lastly for speech rate, because all target sentences were lexically identical across attitudes and encoders, it was calculated by dividing the predetermined syllable count of each sentence by the duration of the corresponding audio file in seconds, yielding in a measure of syllables per second (syll/s).

Acoustic analysis consisted of two steps. First, the extracted acoustic parameters data were subjected to descriptive statistical analysis. Mean values were computed for each acoustic parameter across all conditions. These descriptive measures were presented to identify general patterns of variation between encoders, genders, and attitudes. Following that, each acoustic parameter data was separately analyzed by applying Linear Mixed Effects Models (LMMs) in SPSS (Field, 2018). The fixed effects employed in SPSS were the type of encoder (TTS vs. human), attitude (sarcastic vs. sincere), and gender (male vs. female), alongside their interaction terms. To account for repeated observations and variability across sources, random intercepts for speaker (SPEAKER_ID) and item (SENTENCE_ID) were added in the models. Prior to interpretation, model assumptions were evaluated by assessing the normality of residuals. The result showed that the residuals were approximately normally distributed, which supported the appropriate analysis of the data. Lastly, the results of the analysis were defined based on the Type III tests of fixed effects with the statistical significance of $p < .05$ value. Specific focus on three key effects which are the main effect of ATTITUDE, the interaction between ENCODER and ATTITUDE, and the three-way interaction between ENCODER, GENDER, and ATTITUDE. These effects were evaluated to decide whether acoustic differences between sarcastic and sincere utterances contrasted depending on encoders and genders.

C. FINDINGS AND DISCUSSION

This section presents and discusses findings in relation to the study's two main objectives: identifying acoustic markers of sarcasm and examining differences between human and TTS encoders. The discussion is arranged around key findings regarding the acoustic profile of sarcasm, differences in pitch and temporal realization between TTS and human encoders, gender dependent variation in pitch variability, and the implications of the findings for AI-generated prosody in language learning contexts.

1. Acoustic Profile of Sarcasm

This section reports the acoustic features associated with the expression of sarcastic and sincere attitudes between TTS and human encoder across both genders. Table 1 below reports descriptive statistics of all acoustic measures, revealing consistent differences between utterances produced by sarcastic and sincere attitudes across encoders and genders.

Table 1. Descriptive Statistics of All Acoustic Measures across All Conditions.

Acoustic Measure	Attitude	Male: Human	Male: TTS	Female: Human	Female: TTS
Mean F0 (Hz)	Sarcastic	128.87	113.17	208.96	211.75
	Sincere	149.44	120.19	262.00	228.52
F0 SD (Hz)	Sarcastic	36.98	25.46	51.46	44.37
	Sincere	40.75	31.98	83.32	52.85
F0 Range (Hz)	Sarcastic	144.07	96.67	210.95	169.55
	Sincere	154.07	120.04	358.30	215.76
Mean Amp (dB)	Sarcastic	69.47	72.85	70.67	75.70
	Sincere	69.57	72.24	70.14	75.51
Amp Range (dB)	Sarcastic	49.03	61.64	51.00	71.46
	Sincere	44.65	62.64	49.32	72.40
Speech Rate (syll/s)	Sarcastic	3.81	3.43	3.42	3.20
	Sincere	4.49	3.95	4.20	3.79

Overall, descriptive statistics table above show a consistent pattern in which utterances with sarcastic attitude are generated with lower pitch height, narrower pitch variability, and slower speech rate, whereas amplitude reveal minimal differences across all conditions. To determine whether these observed differences are statistically significant, LMMs were conducted. Table 2 below presents the results of the Type III Tests of Fixed Effects of all acoustic measures.

Table 2. Type III Tests of Fixed Effects[a] of All Acoustic Measures.

Acoustic Measure	Statistics	Attitude	Encoder * Attitude	Encoder * Gender * Attitude
Mean F0	*F* (1, 232)	23.967	6.271	1.305
	p	.000	.013	.254
F0 SD	*F* (1, 232)	18.953	3.146	5.047
	p	.000	.077	.026
F0 Range	*F* (1, 232)	25.229	3.774	6.422
	p	.000	.053	.012
Mean Amplitude	*F* (1, 232)	.523	.047	.381
	p	.470	.828	.537
Amplitude range	*F* (1, 232)	.458	1.740	.207
	p	.499	.188	.649
Speech Rate	*F* (1, 232)	60.938	1.140	.001
	p	.000	.287	.976

The analysis above showed a significant main effect for ATTITUDE on mean F0, $F(1, 232) = 23.967$, $p = .000$ (< .05), F0 SD, $F(1, 232) = 18.953$, $p = .000$ (< .05), F0 range, $F(1, 232) = 25.229$, $p = .000$ (< .05), and speech rate, $F(1, 232) = 60.938$, $p = .000$ (< .05). These results demonstrate that across genders and encoder types, sarcastic utterances consistently displayed lower pitch, less dynamic pitch movement, and slower temporal delivery compared to sincere utterances. Collectively, these converging cues indicate that sarcasm is realized through a relatively constrained or prosodically flattened stance, rather than through isolated acoustic cues, which aligns with previous established descriptions of human sarcastic speech (Anolli et al., 2000; Cheang & Pell, 2008; Rockwell, 2000, 2007). Rather

than functioning independently, these features emerge to work collectively in cueing a sarcastic attitude.

In contrast, two acoustic measures showed non-significant main effects for ATTITUDE which are mean amplitude, $F(1, 232) = .523$, $p = .470$ ($> .05$), and amplitude range, $F(1, 232) = .458$, $p = .499$ ($> .05$), demonstrating that sarcastic and sincere attitudes did not significantly vary in overall loudness level and variability across the dataset. This finding suggests that loudness is not a reliable acoustic feature in differentiating sincerity from sarcasm, which supports prior studies in reinforcing that sarcasm heavily relies more on temporal and pitch adjustment rather than intensity (Cheang & Pell 2008; Rockwell, 2007).

2. Reduced Pitch Height Modulation in TTS Compared to Human Speech

After establishing that one of consistent acoustic features of sarcasm is pitch lowering, the analysis further investigates whether this pattern is equally realized across TTS and human encoders. Notably, the result on acoustic measure of mean F0 on Table 2 show a significant statistical interaction between ENCODER and ATTITUDE, $F(1, 232) = 6.271$, $p = .013$ ($< .05$), demonstrating that the magnitude of pitch height reduction significantly differs between TTS and human encoders.

Descriptively, Table 1 shows that both TTS and human encoders follow the same directional pattern, where utterances delivered with sarcastic attitudes are generated with lower average mean F0 compared to the sincere ones. However, the magnitude of this decrease is prominently different. Moreover, the three-way interaction of mean F0 shown on Table 2 between ENCODER, GENDER, and ATTITUDE yielded a non-statistically significant result, $F(1, 232) = 1.305$, $p = .254$ ($> .05$), further signifying the encoder-related difference is consistent across both female and male voices. This difference is displayed by how human female encoders employed a reduction from 262 Hz to 208.96 Hz (≈53 Hz), while TTS female encoders only reduce from 229.52 Hz to 208.96 Hz (≈17 Hz). Similarly, the pattern is also observed in male encoders, where human encoders decrease their pitch from 149.44 Hz to 128.87 Hz (≈20 Hz), whereas TTS encoders only show a slight reduction from 120.52 Hz to 113.41 Hz (≈7 Hz).

These findings align with prior claims that synthetic speech models can generate general prosodic features but might vary in their fine-tuned production due to prosodic simplification (Jiang, 2024; Li et al., 2023; Rotaeta Perez, 2025). Presumably, pitch lowering might be encoded in TTS systems as a general stylistic cue, while human encoders employ it more dynamically as a pragmatic strategy for cueing sarcastic intent. Therefore, these findings suggest that while TTS encoders successfully exhibit the general direction of sarcastic prosody, they do not fully generate the magnitude of pitch height modulation observed in human sarcastic speech.

3. Gender-dependent Variation in Pitch Variability

Beyond examination of pitch height, the analysis further tests whether pitch variation is similarly realized between TTS and human encoders across genders. Unlike mean F0, the findings show a more complex pattern in which encoder differences appear mainly when gender is considered. As established in the previous analysis, SD and F0 range reported in Table 2 show significant main effects for ATTITUDE, demonstrating that sarcastic utterances are steadily generated with narrower pitch variability and range in comparison to

sincere utterances. However, unlike mean F0, the two-way interaction between ENCODER and ATTITUDE is not significant statistically for both F0 SD, $F(1, 232) = 3.146$, $p = .077$ ($> .05$), and F0 range, $F(1, 232) = 3.774$, $p = .053$ ($> .05$). These findings indicate that both human and TTS encoders display a similar pattern in narrowing pitch variability when delivering sarcasm at a global level. This pattern is significant given previous considerations that TTS models frequently lack pitch variability and rhythmic complexity found in natural speech (Rotaeta Perez, 2025).

Despite the results, it was observed that the interaction between ENCODER, GENDER, and ATTITUDE is statistically significant for both F0 SD, $F(1, 232) = 5.047$, $p = .026$ ($< .05$), and F0 range, $F(1, 232) = 6.422$, $p = .012$ ($< .05$), demonstrating that the two-way pattern shown above is not consistent across both genders. Specifically, the difference between TTS and human encoders are more pronounced in female voices. For instance, human female encoders substantially reduce values in F0 range from 358.30 Hz to 210.95 Hz (≈147 Hz), while TTS female encoders display a much smaller decrease from 215.76 Hz to 169.55 Hz (≈46 Hz). A similar pattern is also observed in F0 SD values, where human female encoders variability value reduces from 83.32 Hz to 51.46 Hz (≈31.86 Hz), whereas TTS female encoders only slightly decrease from 52.85 Hz to 44.37 Hz (≈8.48 Hz).

In comparison, however, male voices show a smaller and more comparable differences between TTS and human encoders. As reflected in Table 1, human male encoders reduce value in F0 SD from 40.75 Hz in sincere utterances to 36.98 Hz in sarcastic utterances (≈3.77 Hz). Similarly, TTS male encoders also exhibit identical reduction in value in F0 SD from 31.98 Hz in sincere ones to 25.46 Hz in sarcastic ones (≈6.52 Hz). Similar pattern also displayed in F0 range vales, where human male encoders exhibit narrowing range value from 154.07 Hz to 144.07 Hz (≈10 Hz), which comparably mimicked by TTS male encoders where they reduce their range value from 120.04 Hz to 96.67 Hz (≈23.37 Hz).

These results indicate that limitations in TTS systems appear less in generating variability reduction itself than capturing its gender sensitive magnitude, specifically for female TTS encoders. Possibly, greater pitch dynamism employed by human female sarcastic speech may cause a more complex prosodic target for current TTS systems to model. This also may underscore potential data training or modeling biases in how expressive variability is represented across gendered voices in TTS systems. Thus, this present dataset suggests that while both encoders display the same general direction toward narrower pitch variation, TTS could not mimic the magnitude of pitch variability employed by human encoders, specifically TTS female encoders.

4. Comparable Temporal Realization in TTS and Human Speech

The findings in speech rate measure present one of the strongest converging patterns in the present study. Unlike the encoder differences observed in pitch-based parameters, temporal modulation displayed strong similarity between human and TTS encoders. As reported in Table 2, it was observed that for speech rate, there are no significant interaction effects of both ENCODER and ATTITUDE, $F(1, 232) = 1.140$, $p = .287$ ($> .05$), and ENCODER, GENDER, and ATTITUDE, $F(1, 232) = .001$, $p = .976$ ($> .05$), demonstrating that TTS model closely approximates human temporal adjustment in generating sarcastic utterances for both male and female voices. This convergence is prominent because it indicates that, unlike pitch adjustment, slower speech temporal delivery might be generated with comparable consistency by human and TTS encoders.

This finding is specifically noteworthy, as prior studies on sarcasm's speech synthesis mainly focused on energy- and pitch-related features (Jiang, 2024; Li et al., 2023) and limited attention given to speech rate as one of the parameters in examining sarcastic prosody. This demonstrates that temporal aspect of prosody might be more robustly modeled by recent TTS models compared to fine-grained pitch adjustment, possibly because durational modulation is computationally simpler than expressive intonational adjustment. These results also underline speech rate as an underexplored dimensions but potentially significant feature in evaluating sarcasm in speech synthesis.

5. Implications for AI-generated Prosody in Language Learning Contexts

To further contextualize these findings within the broader landscape of AI in language learning, previous studies have steadily underlined that AI as pedagogical devices utilize most efficiently as complementary and supportive devices rather than full replacement from real human input. Across conceptual, and empirical studies, AI has been shown to imitate aspects of human communicative processes, enhance learner engagement and personalization, and aid language development through low anxiety and interactive environments (Aini & Basthomi, 2025; Floris et al., 2024; Hastomo et al., 2026; Hastomo et al., 2025a; Hastomo et al., 2025b; Hastomo et al., 2025c). Subsequently, these studies also underscore the significance of careful instructional integration, specifically in addressing problems regarding ethical use, pedagogical alignment, and overreliance. In relation to the present study, these insights should strengthen the idea that ElevenLabs' TTS products should be interpreted not as a full replacement for human prosodic modeling, but as a valuable pedagogical device, yet noticed as inherently limited approximation, specifically for complex pragmatic phenomena such as sarcasm.

From a local pedagogical perspective, these results are specifically relevant in the context of Indonesian EFL learners, where sarcasm interpretation is already complex due to contrasts in prosodic usage and sociocultural norms (Lubis & Bahri, 2023; Putri et al., 2019; Rahardi et al., 2024; Syafruddin et al., 2021). Since learners might depend more on contextual or explicit signals rather than subtle acoustic cues (Koh et al., 2022), the capability of TTS model to generate key prosodic features such as pitch lowering and slower speech rate, indicates that it can be utilized as accessible models for familiarizing contrasts between sincere and sarcastic utterances. However, given that TTS output may display low to moderate pitch height and variability adjustments, learners may not be thoroughly exposed to the range of variation present in authentic sarcastic speech.

Crucially, these findings do not signify that TTS fully presents misleading input. Instead, they promote a more balanced interpretation: TTS models approximate essential acoustic cues of sarcasm but do not thoroughly capture the variability and complexity of human sarcastic prosody. This aligns with previous studies highlighting both the advantages of TTS in offering intelligible and consistent input (Bione et al., 2017; Mubarok et al., 2025; Oktalia & Drajati, 2018) and its limitations in generating nuanced pragmatic features (Phuong et al., 2022; Rotaeta Perez, 2025). Therefore, TTS should be positioned as a complementary pedagogical device, specifically efficient in contexts with limited access to native speakers input, rather than as a complete replacement for authentic spoken sarcasm input.

D. CONCLUSION

This present study investigated how sarcasm is acoustically produced in TTS, particularly ElevenLabs TTS model and human encoders by examining key prosodic parameters, including pitch height, pitch variation, amplitude, and speech rate, across 240 produced by the twelve encoders. The results reveal that sarcasm is consistently signaled by reduced pitch height, narrower pitch variability, and slower speech rate, while amplitude does not display a key role in distinguishing sarcasm from sincerity. These findings support prior studies that sarcastic intent is primarily signaled through pitch and temporal adjustment, which form a clear acoustic profile measures between TTS and human encoders. However, this measure is not uniformly generated by the TTS model. While the TTS model capture the general direction of sarcastic prosody, it generates a weak to moderate magnitude of pitch adjustment and reduced variability, specifically in female TTS encoders, while speech rate is robustly and consistently modeled for sarcastic utterances. Some limitations of the present study lie in its relatively limited number of encoders and controlled dataset, which might not fully capture the variability of authentic human sarcastic speech across broader communicative contexts.

All in all, the findings reflect that ElevenLabs' TTS model presents a partial approximation of sarcastic human prosody, effectively exhibiting general patterns but lacking the magnitude of modulation found in authentic human sarcasm. This underlines the significance of treating the TTS model as a complementary device rather than full replacement model, specifically for complex pragmatic phenomena such as sarcasm. These findings contribute to a more nuanced understanding of synthetic speech by displaying that its ability varies across acoustic features, and they highlight the need for more detailed and further studies on how such differences affect sarcasm interpretation and language learning, especially for pragmatic instruction in differentiating between sincere and sarcastic utterances. Future studies are recommended to investigate sarcastic prosody in more interactional and naturalistic contexts, specifically covering spontaneous sarcastic speech to better capture the variability present in real life communication. Further research should also expand this work by integrating more diverse and larger datasets, as well as examine how these prosodic differences affect listener comprehension and interpretation of sarcasm in language learning contexts.

E. REFERENCES

Aini, N., & Basthomi, Y. (2025). Integration of Artificial Intelligence (AI) in learning English writing in higher education. *Journal of Learning for Development*, *12*(2), 364–371. https://doi.org/10.56059/jl4d.v12i2.1596

Al-Jarf, R. (2022). Text-To-Speech software for promoting EFL freshman students' decoding skills and pronunciation accuracy. *Journal of Computer Science and Technology Studies*, *4*(2), 19–30. https://doi.org/10.32996/jcsts.2022.4.2.4

Amaliah, S., Wahyuni, I. Y., Murdia, M., & Wahid, A. (2025). AI-assisted Podcast Creation in EFL Learning: Enhancing Speaking Fluency and Reducing Foreign Language Anxiety among Gen Z Learners. *Klasikal Journal of Education Language Teaching and Science*, *7*(2), 1174–1189. https://doi.org/10.52208/klasikal.v7i2.1566

Amin, E. A. (2024). EFL Students' Perception of Using AI Text-to-Speech Apps in Learning Pronunciation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4746800

Anolli, L., Ciceri, R., & Infantino, M. G. (2000). Irony as a game of implicitness: acoustic profiles of ironic communication. *Journal of Psycholinguistic Research*, *29*(3), 275–311. https://doi.org/10.1023/a:1005100221723

Bione, T., Grimshaw, J., & Cardoso, W. (2017). An evaluation of TTS as a pedagogical tool for pronunciation instruction: the 'foreign' language context. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 56-61). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.689

Blasko, D. G., Kazmerski, V. A., & Dawood, S. S. (2021). Saying what you don't mean: A cross-cultural study of perceptions of sarcasm. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *75*(2), 114–119. https://doi.org/10.1037/cep0000258

Boersma, P., & Weenink, D. (2025). Praat: doing phonetics by computer [Computer program]. Version 6.4.39, retrieved 13 July 2025 from https://praat.org

Budianto, L., Zulaiha, S., Ar-Rahma, A. P., & Addiyaullami, N. D. (2024). Students' responses on ElevenLabs AI for English-speaking learning. In *Proceeding International Seminar on Islamic Education and Peace*, 4, 80–93.

Cardoso, W. (2018). Learning L2 pronunciation with a text-to-speech synthesizer. In P. Taalas, J. Jalkanen, L. Bradley & S. Thouësny (Eds), *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018* (pp. 16-21). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.806

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication*, *50*(5), 366–381. https://doi.org/10.1016/j.specom.2007.11.003

Cheang, H. S., & Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America*, *126*(3), 1394–1405. https://doi.org/10.1121/1.3177275

Chiang, H. (2019). A Comparison between Teacher-Led and Online Text-to-Speech Dictation for Students' Vocabulary Performance. *English Language Teaching*, *12*(3), 77. https://doi.org/10.5539/elt.v12n3p77

Chuan, C. W., Pau, L. C., Hui, O. Y., & Yunus, M. M. (2025). Leveraging AI-Generated Audio in The Metaverse for Enhanced English Listening Skills: An Innovative Approach. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.14854932

Creswell, J.W. (2014) Research Design: Qualitative, Quantitative and Mixed Methods Approaches. Sage Publications Ltd.

Dabkowski, P., & Staniszewski, M. (2025, June 3). *Introducing Eleven v3 (alpha)*. ElevenLabs. Retrieved April 2, 2026, from https://elevenlabs.io/blog/eleven-v3

Dewatri, R. A. F., Al Aqthar, A. Z., Pradana, H., Anugerah, B., & Nurcahyo, W. H. (2023). Potential tools to support learning: OpenAI and ElevenLabs integration. *ODELIA: Southeast Asia Journal on Open and Distance Learning, 1*(2), 59–69.

Dress, M. L., Kreuz, R. J., Link, K. E., & Caucci, G. M. (2008). Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, *27*(1), 71–85. https://doi.org/10.1177/0261927x07309512

Field, A.P. (2018) Discovering Statistics Using IBM SPSS Statistics. 5th Edition, Sage, Newbury Park.

Floris, F. D., Widiati, U., Renandya, W. A., & Basthomi, Y. (2024). Artificial intelligence in English Language Teaching: Fostering joint enterprise in Online communities. *JEES (Journal of English Educators Society)*, *9*(1), 12–21. https://doi.org/10.21070/jees.v9i1.1825

Gibbs, R. W., Jr. (2000). Irony in talk among friends. *Metaphor and Symbol, 15*(1-2), 5–27. https://doi.org/10.1207/S15327868MS151&2_2

Haverkate, H. (1990). A speech act analysis of irony. *Journal of Pragmatics, 14*(1), 77–109. https://doi.org/10.1016/0378-2166(90)90065-L

Hastomo, T., Sari, A. S., Widiati, U., Ivone, F. M., Zen, E. L., & Andianto, A. (2025a). Exploring EFL teachers' strategies in employing AI chatbots in writing instruction to enhance student engagement. *World Journal of English Language*, *15*(7), 93. https://doi.org/10.5430/wjel.v15n7p93

Hastomo, T., Sari, A. S., Widiati, U., Ivone, F. M., Zen, E. L., & Kholid, M. F. N. (2025b). Does Student Engagement with Chatbots Enhance English Proficiency? *ELOPE English Language Overseas Perspectives and Enquiries*, *22*(1), 93–109. https://doi.org/10.4312/elope.22.1.93-109

Hastomo, T., Widiati, U., Ivone, F. M., & Zen, E. L. (2026). Integrating Generative AI in Primary English Material Design: Insights from Indonesian Teachers' Perceptions. *Revija Za Elementarno Izobraževanje*, *19*(1). https://doi.org/10.18690/rei5409

Hastomo, T., Widiati, U., Ivone, F. M., Zen, E. L., Hasbi, M., & Khulel, B. (2025c). AI-powered conversational agents and intercultural learning: Insights from Indonesian EFL students. *Intercultural Communication Education*, *8*(1), 103217. https://doi.org/10.29140/ice.v8n1.103127

Inworld AI. (2026, February 13). *Best AI Voice Generators for Realistic, Low-Latency TTS (2026 Comparison + Benchmarks)*. Retrieved April 9, 2026, from https://inworld.ai/resources/best-ai-voice-generators?utm_source=chatgpt.com

Jiang, W. (2024). *Synthesis of sarcastic speech: Research on adjusting pitch and energy at keyword level using FastSpeech 2* (Master's thesis). University of Groningen – Campus Fryslan

Kamal, A., & Imran, M. C. (2026). Integrating GenAI in creating digital storytelling as STEAM-based English tenses material on social media. *IDEAS: Journal of Language Teaching and Learning, Linguistics and Literature, 14*(1), 225–241. https://doi.org/10.24256/ideas.v14i1.8481

Koh, J., Lee, S., & Lee, J. M. (2022). L2 pragmatic comprehension of aural sarcasm: Tone, context, and literal meaning. *System*, *105*, 102724. https://doi.org/10.1016/j.system.2022.102724

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General, 118*(4), 374–386. https://doi.org/10.1037/0096-3445.118.4.374

Lameris, H., Szekely, E., & Gustafson, J. (2024). The Role of Creaky Voice in Turn Taking and the Perception of Speaker Stance: Experiments Using Controllable TTS. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16058–16065, Torino, Italia. ELRA and ICCL.

Le, P. N., Vu, H. M. L., & Tran, M. N. (2022). Improving EFL Students' Intonation in Text Using Shadowing Technique with The Implementation of Google Text-to-Speech. *AsiaCALL Online Journal*, 13(1), 93121.EOI:http://eoi.citefactor.org/10.11251/acoj.13.01.006

Li, Z., Gao, X., Nayak, S., & Coler, M. (2023). SarcasticSpeech: Speech Synthesis for Sarcasm in Low-Resource Scenarios. In *Proceedings 12th ISCA Speech Synthesis Workshop (SSW2023)* (pp. 242-243). ISCA.

Lœvenbruck, H., Jannet, M.A.B., D'Imperio, M., Spini, M., Champagne-Lavau, M. (2013) Prosodic cues of sarcastic speech in French: slower, higher, wider.In *Proceedings Interspeech 2013*, 3537-3541, doi: 10.21437/Interspeech.2013-761

Lubis, F. K., & Bahri, S. (2023). Sarcasm in Indonesian television show Pesbukers. *Randwick International of Social Science Journal*, *4*(1), 91–99. https://doi.org/10.47175/rissj.v4i1.626

Mubarok, A. F., Salim, M. A., & Deliany, Z. (2025). AI Based Applications to Foster Pronunciation Accuracy. *Essence Journal of English Language Teaching Linguistics and Literature*, *2*(1). https://doi.org/10.33367/essence.v2i1.7479

Niebuhr, O. (2014) "A little more ironic" Voice quality and segmental reduction differences between sarcastic and neutral utterances. in *Proceedings Speech Prosody 2014*, 608-612, doi: 10.21437/SpeechProsody.2014-110

Nuzhnyy, S. (2026, April 8). *Best Text-to-Speech AI 2026: Top picks & In-depth reviews*. Retrieved April 9, 2026, from https://aimlapi.com/blog/best-text-to-speech-ai?utm_source=chatgpt.com

Oktalia, D., & Drajatı, N. A. (2018). English teachers' perceptions of text to speech software and Google site in an EFL Classroom: What English teachers really think and know. *International Journal of Education and Development Using Information and Communication Technology*, *14*(3), 183–192. http://files.eric.ed.gov/fulltext/EJ1201565.pdf

Önder, F. (2025). The effectiveness of online Text-To-Speech tools in improving EFL students' pronunciation. *Journal of Educational Studies and Multidisciplinary Approaches*, *5*(1). https://doi.org/10.51383/jesma.2025.116

Putri, S., GE, H., Hart, A., Yip, V., & Chen, A. (2019). The effect of explicit training on comprehension of English focus-to-prosody mapping by Indonesian learners of English. In *Proceedings of the 19th International Congress of Phonetic Sciences*, 1937-1941.

Rahardi, R. K., Handoko, H., Rahmat, W., & Setyaningsih, Y. (2024). Javanese silly gags on daily communication on social media: Pragmatic Meanings and Functions approach. *JURNAL ARBITRER*, *11*(1), 49–59. https://doi.org/10.25077/ar.11.1.49-59.2024

Rao, R. (2013). Prosodic Consequences of Sarcasm Versus Sincerity in Mexican Spanish. *Concentric : Studies in Linguistics,* *39*(2), 33-59. https://doi.org/10.6241/concentric.ling.39.2.02

Rockwell, P. (2000). Lower, Slower, Louder: Vocal Cues of Sarcasm. *Journal of Psycholinguistic Research*, *29*(5), 483–495. https://doi.org/10.1023/a:1005120109296

Rockwell, P., & Theriot, E. M. (2001). Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, *18*(1), 44–52. https://doi.org/10.1080/08824090109384781

Rockwell, P. (2007). Vocal features of Conversational Sarcasm: A comparison of methods. *Journal of Psycholinguistic Research*, *36*(5), 361–369. https://doi.org/10.1007/s10936-006-9049-0

Rotaeta Perez, P. (2025). Can AI speak like us? An analysis of Accent Replication in Speech Synthesis. In *Independent Degree Project at Undergraduate Level* [Thesis].

Syafruddin, S., Thaba, A., Rahim, A. R., Munirah, M., & Syahruddin, S. (2021). Indonesian people's sarcasm culture: an ethnolinguistic research. *Linguistics and Culture Review*, *5*(1), 160–179. https://doi.org/10.21744/lingcure.v5n1.1150

Van Duong, T. (2022). The Effects of using Online Text-To Speech Tools on EFL Students' Perceptions in Learning Pronunciation. *International Journal of Science and Management Studies (IJSMS)*, 270–276. https://doi.org/10.51386/25815946/ijsms-v5i4p129

Xiao, Y. (2025). The impact of AI-driven speech recognition on EFL listening comprehension, flow experience, and anxiety: a randomized controlled trial. *Humanities and Social Sciences Communications*, *12*(1). https://doi.org/10.1057/s41599-025-04672-8

Yang, S. (2021). Listener's ratings and acoustic analyses of voice qualities associated with English and Korean sarcastic utterances. *Speech Communication*, *129*, 1–6. https://doi.org/10.1016/j.specom.2021.02.002