



DISCUSSION ABOUT PROSODY AND SYNTHETIC DATA: CONCERN, REASON, AND FUTURE

Gede Satya Hermawan¹, Endry Boeriswati², Miftahulkhairah Anwar³

Universitas Negeri Jakarta¹²³

Email: gede_9906925011@mhs.unj.ac.id¹

ABSTRACT

This study used a literature review as the backbone for the discussion. Talking about three points: concern, reason, and future. As we become more familiar with using artificial intelligence (AI) to create a synthetic environment, we can use it more effectively. With a practical approach that is easy to use, has no boundaries, raises no privacy-breach worries, and is ready to generate. However, despite these benefits, one concern remains: can synthetic data become an alternative to natural data? The main reason for using synthetic data is the manageable time required, unlike using human voices, which involves human-related challenges. However, one question remains: how natural does the prosody of synthetic data sound? Transformer text-to-speech (TTS) models are already publicly available; what is one potential role for them in future language research? Previous literature addresses these questions closely. This paper does not provide a definitive answer but catalyzes future research.

Keywords: literature review; synthetic data; prosody; artificial intelegent; speech

INTRODUCTION

Recent advances in speech technology and artificial intelligence have transformed the way spoken language is created, analyzed, and evaluated. Advances in machine learning, especially in natural language processing and speech-language processing, have enabled these developments. A key driver of these developments is the growing use of synthetic data—artificially generated data designed to mimic natural language behavior. Synthetic data has become an attractive solution for addressing limitations in language corpora, including limited availability, privacy concerns, uneven representation, and the high cost of collecting and annotating spoken language data. Synthetic data are increasingly being used with real-world data to support AI learning and language-related applications (De Wilde et al., 2024). In discourse analysis, prosody remains central. It encompasses aspects of speech beyond individual parts of speech, including intonation, stress, rhythm, timing, pitch variation, duration, and intensity (Ćwiek & Fuchs, 2025). These features contribute to the organization and interpretation of spoken language and are physiologically activated through acoustic parameters such as fundamental frequency

© 2026 Gede Satya Hermawan, Endry Boeriswati, Miftahulkhairah Anwar.

Published by Universitas Negeri Jakarta. This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

(F0), duration, and intensity (Ćwiek & Fuchs, 2025). In speech processing and clustering studies, prosodic information is often represented by changes in the F0 sequence, as modulation reflects key characteristics of natural speech production (Kadowaki et al., 2014). With advances in speech technology, reconstructing dynamic prosodic patterns has become an important task for generating speech that resembles natural human communication. At the same time, synthetic speech data generation has expanded, driven by techniques such as automatic speech recognition (ASR), text-to-speech (TTS), and generation based on large-scale language models. This approach lets researchers create large-scale speech databases with minimal human intervention. However, concerns remain about how well artificially produced speech preserves natural speech characteristics, especially prosodic variation. Studies on speech models show that synthetic data can improve information gain, but it does not automatically guarantee performance equivalent to natural speech. Evidence suggests that using discriminative information in isolation can degrade performance, whereas combining it with real speech can improve efficiency (Mizumoto et al., 2025).

Beyond performance concerns, recent discussions of synthetic data emphasize representativeness, diversity, and quality preservation. Research on metaphorical sentence generation argues that metaphorical narratives should be evaluated not only for their effectiveness but also for parameters such as meaning preservation, style preservation, and distinctiveness from the original narrative (Chim et al., 2025). Furthermore, frequent reliance on artificial outcomes may reduce linguistic diversity over time, affecting the richness of words, vocabulary, and meanings of generated words (Guo et al., 2024). Although these issues have received increasing attention in natural language processing, few studies have examined how synthetic context affects the prosodic behavior and visual representation of synthetic speech, as well as the variability inherent in natural language speech.

With this in mind, this literature review synthesizes prior research on the relationship between prosody and synthetic information in language research. This review aims to understand how artificial speech data are generated, what prosodic characteristics are commonly examined, how researchers evaluate prosodic measures, and what challenges remain in preserving natural prosodic variability. Thus, this study answers the following research question: How have previous studies investigated prosodic behavior in synthetic speech data, and what opportunities and limitations arise when using synthetic data for speech analysis and modeling? By answering these questions, the article aims to lay a theoretical foundation for research.

METHOD

This paper uses a literature review approach to summarize previous criteria for using all variables, ensuring consistency across them. It does not draw on empirical data from the beginning of all periods; rather, it collects, evaluates, and interprets findings from the published literature.

The literature review uses academic databases and selected databases related to the topic. The findings include the features that were included if the prosody of the utterances was completed by those who performed the task, the model, the syllabus, or the final summary of the speech data. Articles that were not language-specific, lacked a technical perspective, or focused solely on speech analysis without prosody were excluded.

These articles must be included, and thematic synthesis is not used. The purpose of the analysis is to identify common themes across the articles, including (1) the concept of prosody, (2) the overcoming of syntactic data, (3) prosodic analysis, (4) the evaluation of syntactic models, and (5) the model of the capabilities and limitations of syntactic data. The results of each study are then compared and summarized to reach a common understanding of the methods and gaps in the existing literature.

The results are presented descriptively and interpretively, informed by a theoretical overview of analytical methods and behavioral approaches to the misbehavior of summarized speech.

RESULTS AND DISCUSSION

Concern: Can Synthetic Data Preserve Natural Prosodic Variation?

A key issue identified in the reviewed literature is whether synthetic data can adequately capture the variability and complexity of natural speech, especially at the prosodic level. Prosodic factors include linguistic features such as pitch, rhythm, stress, duration, and intonation, which affect the structure and interpretation of spoken communication (Ladd, 2008). Since prosodic behavior results from interactions between linguistic, physical, and acoustic features, capturing these features synthetically remains a major challenge.

The literature shows that synthetic databases have made significant strides in scalability and in supporting language-related tasks; however, improved performance should not automatically be interpreted as evidence of linguistic and natural-language equivalence. Guidelines for synthetic data emphasize evaluating generated data not only for usefulness and model performance but also for representativeness, diversity, and the preservation of meaningful variables (De Wilde et al., 2024). Similarly, evaluation of synthetic text generation should consider not only the method's accuracy but also its validity and relevance to natural data (Chim et al., 2025).

This theme is more evident in language-related research. Mizumoto et al. (2025) questioned the validity of synthetic data for training speech models and showed that improvements from synthetic generation depend on how accurately the generated model represents real speech situations. Similarly, analyses of sign language for accent research have shown that sign language adequately recapitulates acoustic properties but captures only a fraction of the overall speech pattern, suggesting limitations in reproducing natural speech variability (Masson & Carson-Berndsen, 202). These results suggest that speech generation can approximate natural language behavior while facilitating higher-level organization.

Another issue that emerges in recent literature is the loss of racial diversity. Guo et al. (2024) argue that repeated reliance on synthetic output can gradually reduce linguistic differences, as generated features tend to align with strong patterns learned during training. Although this discussion was originally raised in the context of speech production, the issue has become relevant to speech and prosody, as prosodic richness also depends on variability rather than change. Overly standardized metaphors can lead to a failure to appreciate natural variation in pitch, timing, and rhythm.

In summary, the studies reviewed suggest that the problem is not that synthetic data can produce clear speech, but that synthetic creativity can preserve the dynamic, evolving nature of human prosody. As a result, synthetic data can now serve better as a source of additional speech information, rather than as a substitute for direct natural speech.

Reason: Why Researchers Continue Using Synthetic Speech Data

Despite concerns about preserving natural prosodic variation, a literature review shows that synthetic speech datasets continue to attract attention because they address the limitations of collecting and processing natural speech corpora for applications. Creating linguistic datasets is often expensive because of authorship, annotation, and peer-review requirements, as well as legal issues related to privacy and permissions. In response, synthetic datasets have emerged as an alternative way to create scalable, accessible learning resources. The proposal to create synthetic datasets emphasizes that artificially generated datasets can advance artificial intelligence by reducing data collection costs and improving access to limited data sources (De Wilde et al., 2024).

The literature increasingly suggests that synthetic language is useful when domain-specific data is lacking or resources are limited. Mizumoto et al. (2025) reported that synthetic data can support the training of conversational speech models, especially when combined with real-world data sources. Rather than completely replacing natural language, synthetic generation expands the available training space and improves model scalability. Similarly, research on speech synthesis pipelines shows that custom speech datasets can be created from large speech corpora and text-to-speech systems, reducing reliance on manually labeled recordings and yielding measurable improvements in speech recognition performance.

Another reason for the popularity of synthetic language is the level of experimental control it provides. Unlike naturally collected corpora, synthetic speech generation lets researchers manipulate selected linguistic features while holding others constant. This capability is essential for language analysis, as researchers can control individual features such as pronunciation, intonation, lexical content, and prosody. Kozekwa et al. (2022) argue that synthetic speech generation allows pronunciation research to move away from relying on artificially generated error data and demonstrate that generated speech can improve the efficiency of pronunciation error detection.

Synthetic data is also valuable because it can increase the variety and regularity of speech production tasks. Recent studies on speech synthesis across many languages show that the processing pipeline itself can generate diverse accents and voices while maintaining similar acoustic characteristics across languages and locations. Such approaches suggest that synthetic data production can contribute not only to increased data volume but also to greater variability in controlled linguistic resources. However, these studies typically position artificial language as a supplement to real language rather than a replacement.

Overall, the literature reviewed shows that researchers continue to use synthesized speech data due to its scalability, simplicity, manageability, and ability to support research in settings where natural linguistic resources remain limited. The increasing use of synthetic data appears to stem less from the perception that synthetic languages are similar to natural languages than from their usefulness as practical, scalable research resources.

Future: Toward Hybrid and Prosody-Aware Synthetic Data

The reviewed literature suggests that future developments in speech synthesis research should move beyond whether synthetic data can replace natural speech and instead focus on better integrating temporal and spatial information. Although synthetic generation offers great experimental flexibility, previous studies have consistently shown that maintaining natural language flexibility remains challenging. Therefore, future research may benefit from approaches that combine the benefits of synthesis and real-world transformation.

Recent research more strongly supports mixed methods rather than study synthesis alone. Mizumoto et al. (2025) argue that the success of synthetic data in natural language depends not only on increasing data volume but also on maintaining accurate natural language characteristics. Similarly, Senocak et al. (2026) showed that synthetic data can achieve results that are comparable to real data and can even improve them when combined with real examples in a training environment. Their results suggest that synthetic data can best serve as a supplementary source that extends and complements natural data, rather than completely replacing it.

Beyond data structure, future research directions might include closer attention to the prosodic representation of synthetic speech production. Traditional assessments of synthetic language tend to focus on identifying accuracy, skills, or performance on subsequent tasks; However, these comparisons may not fully capture prosodic features. Because prosody reflects changes in pitch, rhythm, duration, stress, and intonation (Ladd, 2008), future research will require assessment models that directly assess whether generated speech retains these suprasegmental properties.

Recent advances in language prosody models point to a possible path forward. ProMode proposes a prosody model that uses acoustic and textual features to represent prosodic information, including pitch (F0), intensity, and duration. This type of model suggests that future synthetic speech patterns may produce dominant prosodic patterns that are independent of the overall truth of the language. Such improvements could make synthetic data more useful for language research and language practice.

Future studies could expand assessment criteria beyond functional intelligibility and accuracy to include measures of physical flexibility. Measures may include pitch accuracy, word rate, breath distribution, rhythmic organization, duration measures, and vowel variation. In this sense, the future of synthesized languages may not depend on generating completely artificial languages, but on creating languages that preserve enough language diversity and processes to ensure reliable language analysis.

In summary, the reviewed studies suggest that the future of synthesized speech depends on developing mixed, process-oriented systems that enable synthetic production to complement natural speech while preserving the flexibility inherent in human communication.

CONCLUSION

The discussion shows that synthetic speech datasets can expand speech research and support the development of speech technology; however, their use still depends on appropriate model parameters and evaluation criteria. In speech and phonetic analysis,

synthetic production does not operate in isolation; it requires predefined conditions for speech structure and acoustic features, as well as validation steps, to ensure that the resulting output represents natural speech. This concern is particularly important for prosodic events, where changes in pitch, rhythm, timing, and duration significantly affect verbal communication.

Meanwhile, the reviewed literature suggests that synthetic datasets are likely to continue gaining popularity in speech research due to their scalability and availability. Future developments may include standardized reference systems and more robust integration of synthetic and real speech data. Rather than completely replacing natural language, synthetic data can serve as a complementary resource that extends existing data and enables accurate exploration of linguistic phenomena.

These developments may be particularly important for oral language research because language is often difficult to preserve and is recorded only. In such cases, synthetic data may offer the opportunity to reconstruct, translate, or simulate oral language situations that cannot be achieved through direct observation. However, the utility of such reconstruction will still depend on how well the synthetic production preserves the linguistic variation and prosody found in natural human speech.

BIBLIOGRAPHY

- Barkat-Defradas, M., Raymond, M., & Suire, A. (2021). Vocal preferences in humans: A systematic review. In B. Weiss, A. Trouvain, & M. Barkat-Defradas (Eds.), *Voice attractiveness: Studies on sexy, likable, and charismatic speakers* (pp. 55–80). Springer. https://doi.org/10.1007/978-981-15-6627-1_4
- Chim, J., Ive, J., & Liakata, M. (2025). Evaluating synthetic data generation from user-generated text. *Computational Linguistics*, 51(1), 191–248. https://doi.org/10.1162/coli_a_00540
- De Wilde, P., Arora, P., Buarque de Lima Neto, F., Chin, Y., Thinyane, M., Stinckwich, S., Fournier-Tombs, E., & Marwala, T. (2024). Recommendations on the use of synthetic data to train AI models. United Nations University.
- Guo, Y., Shang, G., Vazirgiannis, M., & Clavel, C. (2024). The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3589–3604).
- Karl, A. L., Fernandes, G. S., Pires, L. A., Serpa, Y. R., & Caminha, C. (2025). Synthetic AI data pipeline for domain-specific speech-to-text solutions, in *Proceedings of STIL 2025*.
- Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., & Kostek, B. (2022). Computer-assisted pronunciation training—Speech synthesis is almost all you need. *Computer Speech & Language*, 76, Article 101390. <https://doi.org/10.1016/j.csl.2022.101390>
- Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press.
- Masson, M., & Carson-Berndsen, J. (2024). Investigating the use of synthetic speech data for the analysis of Spanish-accented English pronunciation patterns in ASR. In *Proceedings of Synthetic Data's Transformative Role in Foundational Speech Models* (pp. 81–85).



- Mizumoto, T., Kojima, A., Fujita, Y., Liu, L., & Sudo, Y. (2025). Is synthetic data truly effective for training speech language models? In Proceedings of Interspeech 2025 (pp. 1808–1812).
- Senocak, A., Park, S., Oh, T.-H., & Chung, J. S. (2026). How far can we go with synthetic data for audio-visual sound source localization? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2026).
- Dua, K., Mittal, P., Gupta, R., & Patel, H. L. (2025). SpeechWeave: Diverse multilingual synthetic text & audio data generation pipeline for training text to speech models, in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Industry Track (pp. 718–737).