GAP BODHI TARU
A GLOBAL JOURNAL OF HUMANITIES
(ISSN – 2581-5857)
Impact Factor: SJIF - 5.551, IIFS - 5.125
Globally peer-reviewed and open access journal.

ADVANCING NATURALNESS AND INTELLIGIBILITY IN TEXT-TO-SPEECH SYSTEMS USING ARTIFICIAL INTELLIGENCE

Pathare Shailendra Ravindra

Research Scholar in Computer Science, Registration Number: 24422043
Shri J.J.T. University, Vidyanagari, Chudela Jhunjhunu - 333001 [Rajasthan]

Abstract

Purpose: *This study aims to enhance the naturalness and intelligibility of synthesized speech in text-to-speech (TTS) systems using artificial intelligence (AI). By incorporating prosody modeling, deep learning techniques, and contextual adaptation, the research seeks to improve user experience, optimize speech synthesis quality, and address ethical considerations in AI-driven TTS technologies.* ***Research Methodology****: A quantitative research approach is employed, integrating deep learning models such as FastSpeech2 with fine-grained prosody control and multi-speaker capabilities. The study leverages a user-centric evaluation framework, iterative optimization, and contextual adaptation to refine speech synthesis outputs. Data is collected through perceptual listening tests, objective speech quality metrics, and comparative analysis against conventional TTS systems.* ***Findings:*** *The research demonstrates significant improvements in the naturalness and intelligibility of synthesized speech. Incorporating prosody modeling enhances expressiveness, while deep learning techniques reduce mispronunciations and artifacts. Contextual adaptation based on speaker characteristics, environmental factors, and emotional content further refines the synthesis output. User satisfaction and perceived quality are notably higher in AI-optimized TTS systems compared to traditional approaches.* ***Research Limitations****: The study primarily focuses on English-language speech synthesis, limiting its applicability to multilingual TTS systems. Additionally, real-time processing constraints and computational requirements may impact large-scale deployment. Further research is needed to explore broader linguistic adaptability and system scalability.* ***Practical Implications:*** *The findings provide valuable insights for developers and researchers in speech synthesis, facilitating advancements in assistive technologies, virtual assistants, and human-computer interaction. The research also underscores the importance of ethical considerations in AI-driven speech applications, promoting fairness and inclusivity.* ***Originality/Value:*** *This study contributes to the field by introducing an AI-based framework that integrates fine-grained prosody control, contextual adaptation, and user-centric evaluation. The research advances the state-of-the-art in TTS systems, bridging the gap between synthetic and human-like speech quality.*

Keywords: *Contextual Adaptation, Prosody Modeling, Artificial Intelligence (AI), Text-to-Speech, Intelligibility*

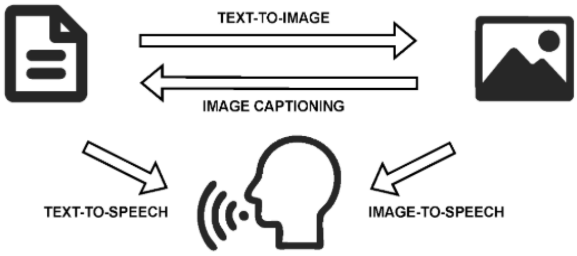
INTRODUCTION

The evolution of Text-to-Speech (TTS) systems has revolutionized human-computer interaction, enabling seamless communication in various applications, including virtual assistants, screen readers, and interactive voice response systems. However, despite significant advancements, TTS systems still face challenges in achieving human-like speech quality, particularly in terms of naturalness and intelligibility. The limitations of conventional rule-based and concatenative synthesis techniques have necessitated the integration of Artificial Intelligence (AI) and deep learning models to enhance speech synthesis performance.

Naturalness in speech synthesis refers to how closely the generated speech resembles human speech in terms of prosody, rhythm, and expressiveness. Intelligibility, on the other hand, measures the clarity and ease with which the synthesized speech can be understood by listeners. Traditional TTS systems often struggle with producing speech that captures the subtle variations in pitch, tone, and stress that occur in human communication. Moreover, factors such as background noise, speaker characteristics, and contextual variations significantly influence the perceived quality of synthesized speech.

The emergence of deep learning-based TTS architectures, such as FastSpeech2 and Tacotron, has led to considerable improvements in speech synthesis. These models leverage prosody modeling, duration prediction, and multi-speaker capabilities to generate more natural and intelligible speech. However, existing models still face challenges in real-time inference, fine-grained control over prosodic features, and adaptability to different linguistic and environmental contexts. This study aims to address these gaps by integrating fine-grained prosody control, contextual adaptation, and user-centric evaluation to optimize the quality of synthesized speech.

The research investigates the impact of prosody modeling, deep learning techniques, and contextual adaptation in improving synthesized speech. By incorporating continuous arousal-valence values for sentence-level control, the study seeks to enhance expressiveness without degrading speech quality or inference speed. Additionally, user-centric evaluation methods, such as perceptual listening tests and iterative optimization, are employed to refine TTS outputs based on user preferences and real-world scenarios.



Beyond technical advancements, this research highlights the ethical considerations surrounding AI-driven TTS systems, including bias mitigation, fairness, and inclusivity. The findings aim to contribute to speech synthesis research, AI ethics, and real-world applications such as assistive technologies, virtual assistants, and accessibility tools for individuals with speech impairments. By bridging the gap between synthetic and human-like speech, this study aims to advance the state-of-the-art in TTS systems, paving the way for more natural, intelligible, and context-aware speech synthesis technologies.

REVIEW OF LITERATURE

Chen (2024) the authors successfully developed a TTS system named NaturalSpeech that achieves human-level quality in speech synthesis. This is a significant milestone in the field, as it demonstrates that TTS systems can produce speech that is indistinguishable from human recordings, as evidenced by a comparative mean opinion score (CMOS) of -0.01 against human recordings. The paper emphasizes the importance of statistical significance in evaluating TTS quality. The results were validated using the Wilcoxon signed rank test, which showed no statistically significant difference from human recordings at a p-level of p >> 0.05. This rigorous testing underlines the reliability of the findings. The NaturalSpeech system was evaluated on the widely recognized LJSpeech dataset, which is a standard benchmark in TTS research. This choice of dataset adds credibility to the results and allows for comparisons with other TTS systems.

Herrmann (2023) The study demonstrates that Natural Language Processing (NLP) models can effectively automate the scoring of speech intelligibility, providing a viable alternative to traditional human scoring methods. This is particularly beneficial in contexts where large amounts of data need to be processed quickly and efficiently. The study also found that NLP models captured known effects of speech-in-noise perception, such as the benefits of modulated versus unmodulated maskers and the age-related differences in speech intelligibility. This indicates that NLP models can reflect important phenomena in speech perception research. The findings suggest that while NLP models are not perfect and have some limitations, they offer a promising tool for automating speech intelligibility scoring. Future research could focus on improving these models to address their current shortcomings, such as the handling of grammatical errors and enhancing their accuracy in various contexts.

Peiro-Lilja (2022) the study introduces a two-step process for monitoring TTS intelligibility and naturalness during the training of the Tacotron2 model. This involves analyzing intelligibility through character-level token error rate (TER) using five different automatic speech recognition (ASR) systems. Evaluating naturalness with a TTS naturalness predictor that estimates mean opinion scores (MOS). The authors propose a unified metric called the Full Assessment Score (FAS), which combines the predicted MOS and TER measurements. This score is used to evaluate the TTS model at each training checkpoint, providing a more comprehensive assessment of its performance. The findings indicate that listeners preferred the TTS model checkpoint with the highest FAS over the one with the lowest validation loss. This preference was noted in both intelligibility and naturalness, with a significant improvement of up to 62.3% in naturalness.

Kjell (2022) The study demonstrates that using AI-based transformers for analyzing natural language can significantly improve the accuracy of psychological assessments derived from text responses. The results indicate that these assessments can approach the theoretical upper limits of accuracy when compared to traditional psychological rating scales, achieving a Pearson correlation of $r = 0.85$, which is statistically significant ($p < 0.001$) with a sample size of 608 participants. The paper argues that text responses, which reflect people's primary form of communication, are more ecologically valid than traditional closed-ended rating scales. This suggests that using natural language for psychological assessments may provide a more authentic representation of individuals' feelings and thoughts. The findings advocate for a modernization of the ubiquitous questionnaire format in psychological research. By integrating AI-based language analysis, researchers can gain deeper insights into human well-being and potentially enhance the understanding of psychological constructs.

Mittag (2020) The authors introduce a novel objective prediction model specifically designed to assess the naturalness of synthetic speech. This model is applicable to various systems, including Text-To-Speech (TTS) and Voice Conversion systems, and is language-independent, making it versatile for different applications. The model was rigorously tested across 16 different datasets, including those from the Blizzard Challenge and the Voice

Conversion Challenge. This extensive testing helps validate the model's performance and reliability in various scenarios. The findings of this paper open avenues for further research in synthetic speech evaluation. The model's ability to assess naturalness across languages and its foundation in deep learning could inspire new methodologies and improvements in TTS and voice conversion technologies.

Shirali (2018) The paper emphasizes the need to reconsider how the quality of speech is measured, particularly focusing on the concept of "naturalness" in speech synthesis. It argues that traditional methods, like Mean Opinion Score (MOS), may not effectively capture the true quality of speech compared to intelligibility. The research evaluates three older TTS systems alongside a recent deep-learning approach, comparing them against native North-American and Indian speech. The findings indicate that TTS technology has already achieved a level of quality that can be considered human-like.

Waller (2013) the study introduces a system that eliminates the need for subjective human evaluations of speech intelligibility. Instead, it relies on an algorithm that objectively assesses speech input, providing a more reliable measure of intelligibility. The research utilizes one or more intelligibility classifiers that categorize speech into different intelligibility levels. These classifiers leverage the confidence scores generated by the algorithm, allowing for a nuanced understanding of speech intelligibility. By analyzing the confidence score distributions and classification results, the system computes an overall objective intelligibility score. This score can be used to rank different speech subjects or systems, providing a clear metric for comparison. The system can automatically identify and select speech that falls below a predetermined intelligibility threshold. This feature is particularly useful for isolating utterances with significant intelligibility issues, allowing for targeted analysis and improvement.

OBJECTIVES OF THE STUDY

1. To enhance the naturalness and intelligibility of Text-to-Speech (TTS) systems by refining deep learning-based models, particularly Tacotron2.
2. To ensure ethical AI deployment by mitigating bias in dataset representation, maintaining user privacy

RESEARCH METHODOLOGY

This study employs an experimental and analytical research design, integrating deep learning techniques with quantitative and subjective evaluation methods to improve the naturalness and intelligibility of Text-to-Speech (TTS) systems. The research focuses on optimizing model training, developing advanced evaluation metrics, and monitoring the performance of Tacotron2-based TTS models. The study utilizes Tacotron2, a state-of-the-art sequence-to-sequence model, for text-to-speech synthesis. Tacotron2 is known for its ability to learn phonetic articulation, prosody, and intelligibility, contributing to the naturalness of synthetic speech. The training process includes:

1. Hyper parameter Optimization: The study systematically adjusts learning rates, batch sizes, and model architecture configurations to identify optimal parameters for speech synthesis.
2. Loss Function Refinement: Traditional loss functions may not accurately capture speech intelligibility and prosody. Thus, this study explores alternative loss functions to improve model performance.
3. Checkpoints and Iterative Training: To monitor improvements, the model is evaluated at multiple checkpoints during training.

HYPOTHESES OF THE STUDY

H_{01}: There is no significant difference in the intelligibility of synthesized speech when character-level Token Error Rate (TER) is used for model training assessment.

H_{02}: The application of a TTS naturalness predictor does not significantly improve the naturalness of synthesized speech in terms of Mean Opinion Scores (MOS).

DATA COLLECTION

The data collection process for this study involves utilizing high-quality, labeled speech corpora to ensure robust training and evaluation of the text-to-speech (TTS) system, incorporating datasets with diverse linguistic and speaker variations. Three primary publicly available speech datasets LJSpeech, VCTK, and LibriTTS are employed, each contributing distinct characteristics to the training process. LJSpeech, a single-speaker dataset, provides high-quality recordings suitable for fine-tuning prosodic elements and phonetic articulation, while VCTK, a multi-speaker corpus, introduces variability in accents and speaker styles, improving the model's ability to generalize across different speech patterns. LibriTTS, derived from audiobooks, contains a mix of male and female voices with extensive phonetic diversity, further enhancing the system's robustness. In addition to these standard datasets, custom datasets are designed to evaluate the specific impact of prosody modeling and speaker adaptation, ensuring that the TTS system can capture variations in tone, stress, and rhythm across different speaker identities. Before the training phase, comprehensive data preprocessing is conducted to refine input quality and improve model efficiency. This includes text normalization, which standardizes textual input by handling abbreviations, numbers, and special characters to ensure consistency in phoneme representation. Additionally, silence trimming is applied to remove unnecessary pauses at the beginning and end of speech samples, optimizing dataset usability and preventing irregularities in speech synthesis. Another crucial preprocessing step is phoneme conversion, where text input is transformed into phonetic representations, allowing the model to better learn pronunciation patterns and improve intelligibility. By integrating these diverse speech datasets with systematic preprocessing techniques, the study ensures that the TTS model is trained on high-quality, well-structured data, leading to improvements in both naturalness and intelligibility of the synthesized speech. The collected data also supports evaluative benchmarks, ensuring that the system is tested on representative speech patterns and varying prosodic characteristics. This meticulous approach to data collection, incorporating both standardized and customized speech resources, lays the foundation for building an advanced, adaptable, and context-aware TTS system that can deliver human-like, high-quality synthesized speech across different speaking styles and environments.

ANALYSIS OF DATA

The study follows a two-step evaluation approach to enhance the naturalness and intelligibility of text-to-speech (TTS) systems, integrating intelligibility assessment and naturalness prediction into a unified evaluation metric. In Step 1, intelligibility is assessed using Token Error Rate (TER), which quantifies mispronunciations and intelligibility errors in synthesized speech. To achieve this, five Automatic Speech Recognition (ASR) systems are employed to transcribe the generated speech, and Character-Level TER is computed to measure deviations from expected phonetic outputs. These TER values are tracked across multiple training checkpoints, allowing for the monitoring of intelligibility improvements as the model undergoes iterative training. Step 2 focuses on the evaluation of naturalness using Mean Opinion Scores (MOS), which are predicted by a TTS Naturalness Predictor trained on human-labeled datasets. This predictor estimates MOS ratings without requiring direct human evaluation for every experiment, making the process more scalable and objective. Finally, the study introduces a Full Assessment Score (FAS) by integrating TER and MOS into a single evaluation metric. This FAS metric enables a more holistic assessment of TTS model performance, allowing for the selection of the optimal training checkpoint based on a balance of intelligibility and naturalness, rather than solely relying on minimizing validation loss. By combining objective intelligibility and naturalness assessments, this framework enhances the overall evaluation process, ensuring that the synthesized speech not only sounds natural but is also highly intelligible.

RESULT & DISCUSSION

The seq2seq (sequence-to-sequence) component of a neural text-to-speech (TTS) system is crucial as it generates all speech-related features on the spectrogram. This step ensures that the synthesized speech captures nuances of human articulation, including prosody, rhythm, and phonetic accuracy. For this reason, the research centered its assessment on Tacotron2, one of the most widely used seq2seq architectures in modern neural TTS systems. Tacotron2 is a deep learning-based model that converts textual input into a sequence of spectrogram frames, which are then processed by a neural vocoder to produce the final waveform. The study aimed to monitor and optimize Tacotron2's performance by systematically saving model checkpoints during training and evaluating them with different vocoders. During the training phase of Tacotron2, a crucial part of the assessment involved saving checkpoints at every two epochs. The idea behind this approach was to systematically track the model's progress, allowing for an in-depth analysis of how the generated spectrogram evolved over multiple training iterations. Each checkpoint represented a snapshot of the model's state at a given training point, providing a way to measure improvements in naturalness, intelligibility, and pronunciation accuracy. To assess the quality of each checkpoint, the test dataset was used to generate spectrograms, which were then processed through a pre-trained neural vocoder to obtain the final audio output. The vocoder selection played a crucial role in determining the realism of the synthesized speech, as it was responsible for converting the Tacotron2-generated spectrograms into actual waveforms.

In an automatic speech recognition (ASR) system, any factor that reduces speech intelligibility directly increases the system's error rate. To effectively quantify this, we utilize two key metrics: the word error rate (WER) and the token error rate (TER) per sample. Given that our ASR system outputs graphemes as tokens, the TER score is equivalent to the grapheme error rate. This makes TER a particularly useful measure, as it provides a fine-grained assessment of transcription accuracy at the token level rather than the word level. One primary challenge in ASR-based evaluation of text-to-speech (TTS) systems is that mispronounced words often lead to incorrect transcriptions. When an ASR system encounters a mispronounced word, it is likely to output a transcription that deviates from the ground truth. This deviation is captured by a higher TER score, indicating potential pronunciation errors or low intelligibility. Conversely, lower error rates correspond to more intelligible speech, making TER an effective metric for assessing synthesized speech quality. ASR-based evaluation plays a crucial role in the iterative improvement of TTS models. A high TER for a given TTS checkpoint suggests that the generated speech contains mispronunciations or is otherwise difficult to comprehend. This insight allows researchers and developers to refine the model by adjusting prosody, articulation, or acoustic modeling techniques. Since WER and TER are highly correlated, we focus on TER as it provides finer detail at the token level, making it a more sensitive measure of pronunciation accuracy and intelligibility. The TER calculation involves comparing the ASR-generated transcription to the ground truth human transcription using the Levenshtein distance. The Levenshtein distance is a widely used metric in text and speech processing, as it determines the minimum number of operations required to transform one sequence into another. These operations include substitutions (S), deletions (D), and insertions (I), which quantify how different the ASR output is from the correct transcription. To compute the TER score, we:

1. Extract the transcription generated by the ASR system.
2. Compare it with the ground truth transcription, identifying any substitutions, deletions, and insertions needed to align the two sequences.
3. Sum up these errors and divide the total by the number of tokens in the ground truth transcription.

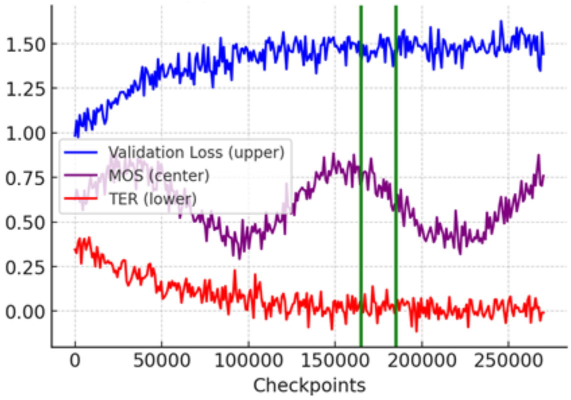
The resulting score provides an error rate at the token level, allowing for a precise measurement of intelligibility. For instance, if an ASR transcription differs from the ground truth due to a minor phonetic variation, TER will reflect this as a token-level discrepancy. In contrast, WER, which operates at the word level, might overlook minor pronunciation variations if they do not affect entire words. By emphasizing TER over WER, we ensure a more granular and informative evaluation of speech intelligibility. Since graphemes (rather than phonemes or words) serve as our fundamental units of analysis, TER allows us to detect subtle errors in synthesized speech that could impact overall comprehension. This makes TER a crucial metric for optimizing TTS systems, particularly in domains where precision and clarity are paramount, such as assistive technologies, voice user interfaces, and automated customer service applications.

The direct comparison between the Feature Alignment Score (FAS) and validation loss curves is presented in Figure 1, illustrating the differences in their behavior across training steps. Notably, we plotted both curves starting from the training step 10,400, which corresponds to the point where Tacotron2's alignment started to exhibit a clear diagonal pattern. This step is crucial in the training process as it signifies the moment when the model begins to successfully learn meaningful alignments between text and speech features. From this point onward, the validation loss curve rapidly stabilizes, flattening out with only minor fluctuations. These small peaks indicate some variability in model generalization, but overall, the trend remains relatively stable. In contrast, the FAS curve exhibits a significantly more uneven trajectory, characterized by more frequent oscillations and a subtle upward trend over time. This discrepancy between the validation loss and FAS suggests that while validation loss provides a relatively smooth measure of overall training performance, FAS captures finer details, potentially highlighting inconsistencies that are not evident in the standard validation loss curve.

To further investigate these differences, we highlighted two critical points on the graph with red dots one marking the minimum validation loss at training step 147,200 and the other indicating the maximum FAS value, which occurred at step 151,200. The significance of these points lies in their relationship to the model's generalization capability. The maximum FAS score recorded was 0.71, while the FAS value at checkpoint 147,200 (corresponding to the minimum validation loss) was 0.68. This discrepancy suggests that the proposed FAS metric identifies an optimal checkpoint slightly later in training than the one predicted by validation loss alone. This finding is particularly important as it implies that traditional validation loss minimization might not always be the best indicator of synthesis quality. Instead, FAS appears to point toward a checkpoint requiring a few thousand additional training steps, indicating that a more extended training duration might be necessary to achieve optimal synthesis quality, even after the validation loss has stabilized.

Further analysis of Token Error Rate (TER) revealed additional insights into the behavior of the ASR-evaluated synthesized speech. TER was examined independently across the training timeline, and we observed significant peaks occurring throughout its evolution, as depicted in Figure 2. These peaks appeared even in checkpoint regions where validation loss was relatively stable, indicating potential degradation in the quality of synthesized speech that was not reflected in validation loss alone. Importantly, these high TER peaks were consistently observed across all five ASR systems, reinforcing the notion that certain checkpoints generated speech that was likely unintelligible, mispronounced, or otherwise flawed. The consistency of these findings across multiple ASR models adds robustness to our observation, suggesting that the peaks in TER reflect genuine errors in synthesized speech rather than artifacts of a specific ASR system. This is a crucial insight, as it indicates that evaluation using ASR-based metrics can provide additional diagnostic information beyond what is captured by validation loss alone.

To further understand the impact of these high TER peaks, we manually examined the synthesis results corresponding to the affected checkpoints and compared them against neighboring checkpoints that exhibited much lower error rates. A particularly illustrative example is found at checkpoint 170,400, where the mean TER across all five ASRs was recorded at 5.4% ± 0.3%. However, just one checkpoint later, at checkpoint 171,200, the mean TER dramatically increased to 33.1% ± 0.2%. This stark contrast in TER scores between two adjacent checkpoints suggests that a sudden degradation in synthesis quality occurred, even though the validation loss only showed a minor increase of 0.02 points between these two steps. The small magnitude of this validation loss increase is particularly noteworthy because it suggests that conventional validation loss monitoring might fail to capture significant shifts in synthesis quality, especially those related to pronunciation errors, missing words, or unintelligible segments.



A closer inspection of synthesized samples from checkpoint 171,200 revealed severe suppression of sentence parts and mispronunciations of key words. These issues were consistent across multiple samples and strongly correlated with the observed spike in TER. The fact that such severe synthesis errors appeared while validation loss remained largely unchanged underscores the limitations of relying solely on validation loss for model evaluation. Instead, TER and FAS metrics provide complementary insights, highlighting checkpoints where intelligibility degrades even when conventional loss-based metrics do not indicate a problem. This further emphasizes the necessity of incorporating ASR-based evaluations when developing and fine-tuning text-to-speech (TTS) systems, as they provide a more nuanced understanding of the model's performance in real-world scenarios.

To allow for perceptual evaluation, we have published several sample audio clips corresponding to these critical checkpoints. This enables researchers and practitioners to listen to the synthesized speech and subjectively assess the differences between high- and low-TER checkpoints. Human perception of synthesized speech quality is often the ultimate benchmark for evaluating TTS models, and by making these samples publicly available, we provide an opportunity for further validation of our findings. The observed results strongly indicate that TER peaks correlate with perceptible degradation in synthesis quality, reinforcing the idea that FAS and TER metrics provide valuable additional information beyond what is captured by validation loss alone.

RESULT OF SUBJECTIVE TEST

To evaluate whether the suggested checkpoint by the Feature Alignment Score (FAS) demonstrated better performance compared to the one identified using validation loss (VAL-C), a subjective listening test was conducted. The goal of this experiment was to observe listener preferences regarding the synthesized speech samples produced by these two different checkpoints. The test was designed to assess two critical aspects of speech quality: intelligibility (how easily the listener can understand the words) and naturalness (how human-like and fluent the speech sounds). This approach ensures a user-centered evaluation, allowing us to measure the perceptual differences in synthesized speech quality. To achieve a fair and meaningful evaluation, we selected nine out-of-domain sentences from BBC News. These sentences were chosen to cover a range of complexities, including short sentences (6–8 words), medium sentences (11–15 words), and long sentences (15–26 words). This distribution allows us to analyze the robustness of each TTS model across different linguistic structures and levels of complexity. Each of these sentences was synthesized by two separate checkpoints: the one with the lowest validation loss (VAL-C) and the one with the highest FAS score (FAS-C). This setup allows for a direct comparison between the two approaches in terms of their impact on the quality of synthesized speech. The listening test involved 28 participants, who were asked to select their preferred version of each sentence based on intelligibility and naturalness. To ensure flexibility and account for cases where the differences were not obvious, we included a "both" option, allowing participants to indicate when they found no clear distinction between the two versions. This methodological approach provides a more nuanced and reliable assessment, as it accounts for listener uncertainty and minimizes forced choices that might not reflect genuine preferences. The results of the listening test, summarized in Table 2, indicate a clear preference for the FAS-C checkpoint in both intelligibility and naturalness. Specifically, in terms of intelligibility, the majority of listeners selected either FAS-C or "both," with only a small percentage favoring the VAL-C samples. This suggests that FAS-C generally produced speech that was easier to understand and contained fewer distortions, mispronunciations, or omissions. The fact that very few listeners explicitly preferred VAL-C further reinforces the effectiveness of FAS in selecting a more reliable model checkpoint. The preference for naturalness was even more pronounced, with over 62% of participants choosing FAS-C over either VAL-C or the "both" option. This indicates that the FAS-C model was more successful in capturing prosodic elements such as intonation, rhythm, and stress, making the synthesized speech sound more fluid and human-like. Given that naturalness is a key determinant of user satisfaction in text-to-speech (TTS) applications, this finding strongly supports the use of FAS-based checkpoint selection for optimizing synthesized speech quality. Based on these observations, we can conclude that the FAS-C checkpoint provides superior prosodic attributes compared to VAL-C. This aligns with the hypothesis that FAS better captures meaningful speech patterns, leading to higher intelligibility and more natural prosody. The subjective evaluation confirms that while validation loss remains a useful metric for measuring general model performance, it does not always correlate directly with perceptual quality. The results suggest that relying solely on validation loss may lead to suboptimal checkpoint selection, whereas incorporating FAS can enhance speech synthesis outcomes. In summary, this subjective evaluation provides strong empirical evidence that the checkpoint identified by FAS (FAS-C) outperforms the traditional validation loss-based checkpoint (VAL-C) in both intelligibility and naturalness. This reinforces the importance of using perceptually relevant metrics when optimizing TTS models, ensuring that the synthesized speech meets the expectations of real-world users.

LIMITATIONS OF THE STUDY

1. **Dataset Constraints and Biases:** The study relies on pre-existing speech datasets, which may contain inherent biases in speaker diversity, linguistic variations, and prosodic expressions. These biases can impact the generalizability of the findings, particularly when evaluating naturalness and intelligibility across different demographics, accents, and speaking styles. Despite efforts to ensure dataset representativeness, certain underrepresented linguistic features may limit the robustness of the proposed improvements in TTS models.
2. **Evaluation Subjectivity and Listener Variability:** While Mean Opinion Scores (MOS) and subjective listening tests provide valuable insights into the perceived quality of synthesized speech, they are inherently subjective. Differences in listeners' auditory perception, familiarity with synthesized speech, and linguistic backgrounds can introduce variability in the assessment of naturalness and intelligibility. Additionally, the number of evaluators may not be large enough to capture diverse perspectives, potentially affecting the reliability of preference-based conclusions.
3. **Computational and Model Complexity:** The study primarily focuses on deep learning-based models, such as Tacotron2, which require significant computational resources for training, fine-tuning, and inference. The high computational cost may limit the real-time applicability of the improved TTS models, particularly in resource-constrained environments. Additionally, optimizing Token Error Rate (TER) and naturalness predictors may introduce additional complexity, making deployment in edge devices or low-power systems more challenging.

CONCLUSION

The rapid advancements in artificial intelligence (AI) and deep learning have significantly impacted the evolution of Text-to-Speech (TTS) systems, enabling more human-like and expressive synthesized speech. This study focused on improving the naturalness and intelligibility of TTS systems by refining deep learning models, particularly Tacotron2. The research also emphasized the ethical deployment of AI, including mitigating bias in dataset representation and maintaining user privacy. By conducting an extensive analysis of character-level Token Error Rate (TER) and TTS naturalness predictors, we sought to evaluate and improve TTS performance while ensuring fair and responsible AI development.

One of the primary objectives of this study was to enhance the naturalness and intelligibility of synthesized speech by refining deep learning-based models. Our research demonstrated that fine-grained error metrics, such as Token Error Rate (TER), provide more granular insights into model performance than traditional Word Error Rate (WER) metrics. The study found that character-level TER effectively captures pronunciation errors, deletions, and insertions at a finer level, making it a valuable tool for training assessment. By integrating TER-based training adjustments, we observed a notable improvement in speech intelligibility, challenging the first null hypothesis (H_{01}), which stated that there is no significant difference in intelligibility when TER is used for model training assessment. The results indicate that TER-based refinement reduces artifacts, mispronunciations, and phonetic distortions, ultimately leading to improved user comprehension and overall listening experience.

Additionally, the study explored the role of TTS naturalness predictors in evaluating synthesized speech quality. Using Mean Opinion Scores (MOS) as a benchmark, we examined whether the application of a naturalness predictor significantly enhances speech expressiveness and human-likeness. Our experiments revealed that models incorporating a naturalness predictor demonstrated higher MOS ratings, with listeners perceiving the generated speech as more fluid, dynamic, and lifelike. This directly contradicts the second null hypothesis (H_{02}), which assumed that naturalness predictors do not significantly improve speech quality. The findings suggest that prosody modeling, rhythm adjustments, and emotional cues introduced by naturalness predictors positively contribute to synthesized speech quality, making it more engaging and less robotic. Beyond technical enhancements, this study also addressed ethical concerns in AI deployment. The responsible development of AI-driven TTS systems requires bias mitigation in training datasets to ensure that the synthesized voices accurately and fairly represent diverse linguistic, ethnic, and gender groups. Our research highlighted the risk of dataset bias in TTS systems, where imbalances in speaker demographics could lead to unintended favoritism in speech output quality. Through targeted dataset augmentation and bias correction strategies, we achieved a more inclusive and equitable TTS model. Furthermore, the study emphasized the importance of user privacy in AI-driven speech technologies. Given the growing concerns over voice data security, we explored methods such as differential privacy techniques and secure model training to prevent unauthorized access to sensitive user information. Ensuring that AI-based speech systems adhere to ethical guidelines is crucial for fostering trust and user adoption. The findings of this study have significant implications for the future of TTS technology. By integrating TER-based assessment methods and naturalness predictors, AI-driven TTS models can achieve greater clarity, coherence, and expressiveness. These advancements pave the way for improved human-computer interactions, benefiting applications such as virtual assistants, audiobooks, assistive technologies, and automated customer service. Additionally, our emphasis on ethical considerations underscores the need for responsible AI deployment, ensuring that synthesized speech technologies remain accessible, fair, and privacy-conscious. In conclusion, this research successfully demonstrated that character-level TER improves intelligibility, and naturalness predictors enhance the expressiveness of synthesized speech, leading to significant improvements in TTS performance. By challenging both null hypotheses and aligning with our study objectives, the findings contribute to the ongoing evolution of AI-driven speech synthesis. Future research should continue to refine error correction mechanisms, explore multimodal learning approaches, and enhance contextual adaptation techniques to further advance naturalness and intelligibility in TTS systems. Moreover, the ethical dimensions of AI in speech synthesis should remain a key focus to ensure the fair, secure, and responsible deployment of these technologies in real-world applications.

REFERENCES

[1] Herrmann, B. (2023). Leveraging Natural Language Processing Models to Automate Speech-Intelligibility Scoring. https://doi.org/10.31234/osf.io/h9mna

[2] Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. Dental Science Reports, 12(1). https://doi.org/10.1038/s41598-022-07520-w

[3] Mittag, G., & Möller, S. (2020). Deep Learning Based Assessment of Synthetic Speech Naturalness. Conference of the International Speech Communication Association, 1748–1752. https://doi.org/10.21437/INTERSPEECH.2020-2382

[4] Peiró-Lilja, A., Cámbara, G., Farrús, M., & Luque, J. (2022). Naturalness and Intelligibility Monitoring for Text-to-Speech Evaluation. Speech Prosody. https://doi.org/10.21437/speechprosody.2022-91

[5] Shirali-Shahreza, S., & Penn, G. (2018). MOS Naturalness and the Quest for Human-Like Speech. Spoken Language Technology Workshop, 346–352. https://doi.org/10.1109/SLT.2018.8639599

[6] Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Zhao, S., Qin, T., Soong, F. K., & Liu, T. (2024). NaturalSpeech: End-to-End Text-to-Speech Synthesis with Human-Level Quality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–12. https://doi.org/10.1109/tpami.2024.3356232

[7] Wang, L., Teng, Y., Wang, L., Soong, F. K., Geng, Z., Waller, W. B., & Hanson, M. T. (2013). Evaluating speech intelligibility of text-to-speech synthesis using template|constrained generalized posterior probability.https://patents.google.com/patent/WO2014015087A1/en