Prosodic Differences Between Child-Directed and Adult-Directed Speech in Text-to-Speech Generation

Anonymous ACL submission

Abstract

Child-directed speech (CDS), speech produced when addressing children, and adult-directed speech (ADS), speech produced when addressing adults, differ systematically in prosodic characteristics such as pitch and speaking rates. We investigate whether text-to-speech (TTS) models fine-tuned on CDS and ADS reproduce the register-specific acoustic differences observed in human speech. The results showed that generated speech replicated several register-related prosodic patterns observed in human speech, including higher pitch and slower articulation rates in CDS relative to ADS. However, the magnitude of CDS-ADS differences in pitch measures was attenuated in generated speech, while articulation rate differences tended to be exaggerated. Overall, the findings suggest that fine-tuned TTS models can reproduce listener-conditioned register distinctions, while also revealing some limitations.

1 Introduction

Child-directed speech (CDS), the speech register used when speaking to children, is characterized by prosodic patterns distinct from those of adult-directed speech (ADS), the style used when addressing adults. These characteristics include a higher fundamental frequency (F0; the acoustic basis of perceived pitch), a wider F0 range, and a slower speaking rate (Fernald et al., 1989; Soderstrom, 2007; Swanson et al., 1992). These acoustic properties have been proposed to support early language acquisition and emotional bonding by effectively capturing infant attention (Werker and McLeod, 1989), facilitating word segmentation (Thiessen et al., 2005), and conveying emotional intent (Trainor et al., 2000). While previous phonetic and acquisition research has extensively documented these differences in natural human interaction, the rise of generative speech technologies raises a new question: can the prosodic differences between CDS and ADS be learned within a neural text-to-speech (TTS) model?

Modern TTS systems are typically optimized for intelligibility and naturalness in relatively neutral speaking styles (Wang et al., 2017; Shen et al., 2018; Ren et al., 2019; Casanova et al., 2024). Because such systems are generally trained on read speech or audiobooks, they may be less suited to capturing the exaggerated prosody associated with CDS. In this study, we investigate whether TTS models fine-tuned separately on CDS and ADS reproduce the prosodic differences observed in human speech, including differences in pitch-related measures and articulation rate. More broadly, we examine whether generative speech systems can internalize speech register distinctions from relatively limited training data.

1.1 Prior Works

Recent work in neural TTS has explored expressive speaking style modeling, including latent style representations, prosody transfer, and fine-grained style control (Wang et al., 2018; Zhang et al., 2019; Akuzawa et al., 2018; Hsu et al., 2018; Skerry-Ryan et al., 2018; Chen and Rudnicky, 2022; Kim et al., 2022; Min et al., 2021). These studies demonstrate that TTS systems can encode stylistic and prosodic variation independently of content and speaker identity. Other work has addressed the tendency of deterministic TTS systems to produce flattened or "average" prosody by introducing stochastic modeling approaches that improve expressive variability (Hodari et al., 2019; Mayer et al., 2025).

However, prior work has primarily focused on emotion or speaker adaptation rather than listener-conditioned register variation such as CDS versus ADS. In addition, CDS represents a relatively low-resource speech register due to the limited availability of large caregiver-child speech corpora. The present study evaluates whether fine-tuned TTS models can nevertheless reproduce the prosodic

distinctions associated with CDS.

2 Methods

2.1 Data

We obtained CDS and ADS datasets from the ECOLANG multimodal corpus (Gu et al., 2025). The corpus includes speech from British and American English-speaking adults engaged in semi-naturalistic conversations with their child aged 3-4 years (CDS) or with a familiar adult (ADS). The CDS dataset includes 38 speakers (37 females, 1 male) and the ADS dataset includes 31 speakers (18 females, 13 males). The corpus contains ELAN (Max Planck Institute for Psycholinguistics, The Language Archive, 2026) annotations of audiovisual recordings, each of which lasts 35-45 minutes. We used the audio recordings for model training and testing, and the transcripts from the test set as input text for speech generation (see Section 2.3 for details).

Given well-established gender-based differences in pitch and speaking rate (Jacewicz et al., 2009), as well as prosodic differences between British and American English (Fernald et al., 1989), we included only British female speakers in our dataset. This filtering process was intended to minimize variation unrelated to register and to avoid mixing speakers with substantially different prosodic characteristics within the data. After filtering, the dataset consisted of 30 speakers in the CDS data and 16 speakers in the ADS data.

2.2 Model

We used Coqui XTTS-v2, a GPT-based transformer decoder with ∼518M parameters (Casanova et al., 2024). XTTS-v2 was selected for its strong zero-shot voice cloning capability, achieving state-of-the-art performance in speaker similarity and naturalness across multiple languages. The model predicts discrete audio tokens from text, with a DVAE handling conversation between raw audio and token representations. At inference time, a short reference audio clip from the target speaker is used to condition the voice cloning, allowing the model to generate speech that mimics the acoustic characteristics of that given speaker. In addition, XTTS-v2 supports speaker fine-tuning, allowing the pretrained model to adapt to a specific speaker's voice characteristics, beyond what zero-shot cloning can achieve alone. This makes it well suited for synthesizing speech in specific registers.

2.3 Procedure

Because the model performs best with training audio files approximately 5-12 seconds in duration, preprocessing of the audio data was necessary. Individual utterances in the corpus were typically shorter than this target range, so adjacent utterances were concatenated to create longer training clips. For both the ADS and CDS datasets, we first extracted audio corresponding to individual utterances based on timestamped ELAN transcripts, and then incrementally concatenated them until the resulting clip fell within the target duration range of 5-12 seconds. During concatenation, 300 ms of silence was inserted between adjacent utterances. The datasets were divided into training (80%) and test (20%) sets. The ADS training set contained 10,538 utterances, resulting in 1,876 concatenated clips, while the CDS training set contained 18,706 utterances, resulting in 2,684 concatenated clips. The test sets consisted of 3,171 ADS utterances and 4,224 CDS utterances.

Fine-tuning was performed using an AdamW optimizer with a learning rate of 5e-6, a weight decay of 1e-2, and a gradient accumulation of 16 steps. These hyperparameters were selected based on the recommended values from Coqui XTTS fine-tuning documentation (erew123, 2024) and adjusted to accommodate the memory constraints of available GPU hardware. The training was conducted with float32 precision, with checkpoints saved every 100 steps. To identify optimal checkpoints, we made perceptual and impressionistic judgments regarding the intelligibility and naturalness of the synthesized speech at regular intervals. Based on these evaluations, CDS checkpoints between 8500 and 9500 steps and ADS checkpoints between 3400 and 4400 steps were identified as producing the most intelligible and natural-sounding speech. Checkpoints within these ranges were included in subsequent prosodic analyses.

At each checkpoint described above, we generated 500 utterances using input text selected from the test sets. Prior to selection, utterances containing text unsuitable for speech synthesis (e.g., symbols such as `<laugh>` indicating speaker laughter or `<unclear>` indicating unintelligible audio) were removed from the transcripts. In addition, short utterances consisting of only one or two words were excluded because such utterances tend to exhibit distinct prosodic characteristics, including reduced pitch variability and slower speaking rates. From

the remaining transcripts, 500 utterances were randomly selected. For each utterance, the corresponding original human recording served as the reference audio to condition the model's voice cloning, allowing the generated speech to retain the target speaker's vocal characteristics.

2.4 Prosody Analysis

We compared the generated speech with the corresponding human speech utterances on the following prosodic measures: mean F0, minimum F0, maximum F0, F0 range (maximum F0 minus minimum F0), and articulation rate (number of syllables divided by speaking time, which excludes pause duration). We used Parselmouth (Jadoul et al., 2018), a Python interface to the Praat software (Boersma and Weenink, 2026), to calculate these measures for individual utterances.

Statistical analyses of the CDS-ADS differences were conducted in R (R Core Team, 2024) using mixed-effects linear regression models implemented in the *lmerTest* package (Kuznetsova et al., 2017). Separate models were fitted for each measure mentioned above. Each model included REGISTER (sum-coded with CDS as 1 and ADS as -1), VOICE TYPE (sum-coded with generated speech as 1, human speech as -1), and their interaction as fixed effects. To account for variability across different data subsets, the models also included a random intercept for DATA SUBSET, encompassing the human baseline and each model training checkpoint (e.g., 3400, 3500, ... for ADS; 8500, 8600, ... for CDS), nested within REGISTER. Post-hoc analyses were conducted using the *emmeans* package (Lenth, 2024) to perform pairwise comparisons of REGISTER levels within each VOICE TYPE condition.

3 Results

3.1 F0

Figure 1 shows mean F0, minimum F0, maximum F0, and F0 range values for generated ADS and CDS across training checkpoints, alongside the corresponding human speech test data. Overall, generated speech reproduced several register-related pitch patterns observed in human speech. However, across measures, the magnitude of the CDS-ADS distinction was attenuated in generated speech relative to human speech. In addition, F0 values were generally higher in generated speech than in human speech, regardless of register. We present the results for each measure separately below.

For mean F0, generated speech exhibited overall higher values than human speech, as indicated by the significant main effect of VOICE TYPE ($\beta = 1.94$, $SE = 0.60$, $t = 3.22$, $p < .01$). Post-hoc comparisons further showed that CDS had significantly higher mean F0 values than ADS in both generated speech ($\beta = 2.96$, $SE = 0.70$, $z = 4.25$, $p < .001$) and human speech ($\beta = 8.68$, $SE = 2.31$, $z = 3.76$, $p < .001$). However, the significant interaction between REGISTER and VOICE TYPE indicated that the CDS-ADS difference was smaller in generated speech than in human speech ($\beta = -1.43$, $SE = 0.60$, $t = -2.37$, $p < .05$), suggesting attenuation of the register effect in generated speech.

A similar pattern was observed for minimum F0. Overall, minimum F0 values did not significantly differ between generated and human speech ($\beta = -1.10$, $SE = 1.05$, $t = -1.05$, $p = .31$). Importantly, CDS exhibited significantly higher minimum F0 values than ADS in both generated speech ($\beta = 4.01$, $SE = 1.21$, $z = 3.32$, $p < .001$) and human speech ($\beta = 13.99$, $SE = 4.01$, $z = 3.49$, $p < .001$). However, the interaction between REGISTER and VOICE TYPE showed that the CDS-ADS difference was significantly smaller in generated speech than in human speech ($\beta = -2.49$, $SE = 1.05$, $t = -2.38$, $p < .05$).

For maximum F0, generated speech exhibited overall higher values than human speech ($\beta = 10.84$, $SE = 1.70$, $t = 6.39$, $p < .001$). No significant CDS-ADS difference was observed in either generated speech ($\beta = -0.37$, $SE = 1.96$, $z = -0.19$, $p = .85$) or human speech ($\beta = -6.34$, $SE = 6.49$, $z = -0.98$, $p = .33$), nor was the main effect of REGISTER significant ($\beta = -1.68$, $SE = 1.70$, $t = -0.99$, $p = .33$). The interaction between REGISTER and VOICE TYPE was not significant, either ($\beta = 1.50$, $SE = 1.70$, $t = 0.88$, $p = .39$). Thus, although this result does not replicate previous findings that maximum F0 is higher in CDS (e.g., Fernald et al., 1989), the generated speech nevertheless reproduced the absence of a CDS-ADS distinction observed in the current human speech data.

Similar to mean F0 and maximum F0, generated speech exhibited larger F0 ranges than human speech ($\beta = 11.94$, $SE = 1.79$, $t = 6.67$, $p < .001$). Contrary to the pattern typically reported in the literature (e.g., Fernald et al., 1989), CDS exhibited significantly *smaller* F0 ranges than

ADS within both generated speech ($\beta = -4.38$, $SE = 2.07$, $z = -2.12$, $p < .05$) and human speech ($\beta = -20.33$, $SE = 6.86$, $z = -2.97$, $p < .01$). Because F0 range is calculated as maximum F0 minus minimum F0, this pattern is consistent with the finding reported above that minimum F0 was higher in CDS whereas maximum F0 did not reliably differ between CDS and ADS. Importantly, although the direction of the CDS effect differed from findings in previous research, the generated speech nevertheless replicated the pattern observed in the human speech data in the current study. However, the significant interaction between REGISTER and VOICE TYPE ($\beta = 3.99$, $SE = 1.79$, $t = 2.23$, $p < .05$) indicated that the CDS-ADS difference in F0 range was significantly smaller in generated speech than in human speech.

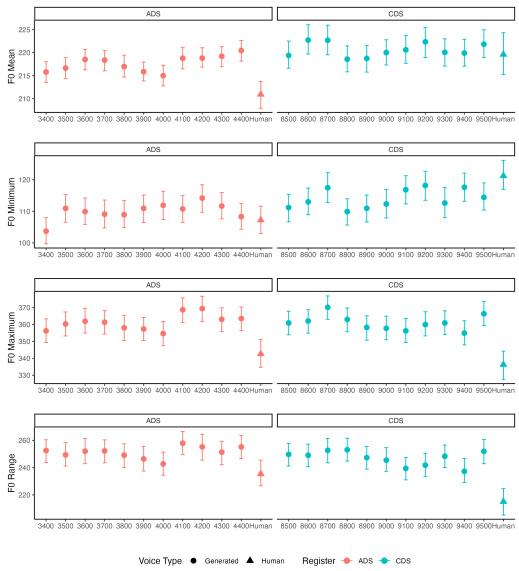

Figure 1: Mean F0, minimum F0, maximum F0, and F0 range values for generated ADS and CDS across training checkpoints, compared with the corresponding human speech test data. Points represent mean values, and error bars indicate 95% confidence intervals.

Overall, these results indicate that the generated speech reproduced several register-related pitch patterns observed in human speech. However, across measures, the magnitude of the CDS-ADS distinction was generally attenuated in generated speech relative to human speech.

3.2 Articulation Rate

Figure 2 shows articulation rate values for generated ADS and CDS across training checkpoints, alongside the corresponding human speech test data. No significant overall difference was observed between generated and human speech ($\beta = -0.01$, $SE = 0.02$, $t = -0.48$, $p = .64$). Crucially, CDS exhibited significantly slower articulation rates than ADS in both generated speech ($\beta = -0.35$, $SE = 0.03$, $z = -13.96$, $p < .001$) and human speech ($\beta = -0.17$, $SE = 0.08$, $z = -2.07$, $p < .05$), indicating that the generated speech reproduced the register-related articulation rate pattern observed in the human speech data. Although the interaction between REGISTER and VOICE TYPE did not reach conventional levels of significance ($\beta = -0.05$, $SE = 0.02$, $t = -2.06$, $p = .053$), the larger CDS-ADS difference observed in generated speech suggests a tendency toward an exaggerated register effect relative to the human speech data.

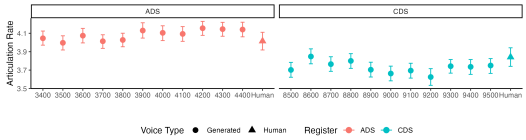

Figure 2: Articulation rates for generated ADS and CDS across training checkpoints, compared with the corresponding human speech test data. Points represent mean articulation rates, and error bars indicate 95% confidence intervals.

4 Discussion

We found that the generated speech replicated many of the register-related prosodic patterns observed in the human speech data across F0 measures and articulation rate. However, the magnitude of these differences was attenuated for F0 measures and exaggerated for articulation rate relative to human speech. In addition, generated speech generally exhibited higher F0 values than human speech regardless of register.

More broadly, these findings demonstrate that neural speech generation systems can learn macro-level register distinctions even when trained on relatively limited, semi-naturalistic corpora. The consistent directional contrasts across pitch and temporal metrics indicate that standard fine-tuning paradigms are sensitive to the global acoustic structures that characterize social registers. This suggests that the systematic variations driven by interactive contexts, such as the distinct prosodic profile of CDS, can be internalized by generative models even under low-resource training conditions, although some limitations remain.

5 Limitations

A limitation of this study concerns the relatively small number of speakers available in the corpus used in the present study. We maintained speaker-disjoint training and test sets, and the limited speaker pool resulted in a small human test set. Consequently, the acoustic characteristics of the test data may reflect idiosyncratic properties of those speakers rather than broader register-level patterns. Moreover, the corpus contains different speakers for ADS and CDS, making it difficult to disentangle register-related effects from speaker-specific acoustic characteristics. Future work using larger speaker populations is needed to examine whether the present findings generalize beyond the current corpus, although constructing such datasets remains challenging because relatively few corpora provide large amounts of speech data with time-aligned transcripts from the same speakers across both ADS and CDS conditions.

6 Ethical Considerations

This study uses speech data from a corpus consisting of adult-child interactions. Because generative speech technologies may be misused for deceptive voice imitation, we emphasize that the present work is intended solely to investigate whether TTS systems can learn prosodic distinctions between CDS and ADS.

References

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. In *Interspeech 2018*, pages 3067–3071.

Paul Boersma and David Weenink. 2026. Praat: doing phonetics by computer [computer program]. Retrieved 24 April 2026.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

Li-Wei Chen and Alexander Rudnicky. 2022. Fine-grained style control in transformer-based text-to-speech synthesis. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7907–7911.

erew123. 2024. Xtts model finetuning guide (simple version). https://github.com/erew123/alltalk_tts/wiki/XTTS-Model-Finetuning-Guide-(Simple-Version). Accessed: 2026.

Anne Fernald, Traute Taeschner, Judy Dunn, Mechthild Papousek, Bénédicte de Boysson-Bardies, and Ikuko Fukui. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16(3):477–501.

Yan Gu, Ed Donnellan, Beata Grzyb, Gwen Brekelmans, Margherita Murgiano, Ricarda Brieke, Pamela Perniss, and Gabriella Vigliocco. 2025. The ECOLANG Multimodal Corpus of adult-child and adult-adult Language. *Scientific Data*, 12(1):89.

Zack Hodari, Oliver Watts, and Simon King. 2019. Using generative modelling to produce varied intonation for speech synthesis. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 239–244. ISCA.

Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. 2018. Hierarchical generative modeling for controllable speech synthesis. *Preprint*, arXiv:1810.07217.

Ewa Jacewicz, Robert A. Fox, Courtney O'Neill, and Joseph Salmons. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21(2):233–256.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Changhwan Kim, Seyun Um, Hyungchan Yoon, and Hong-Goo Kang. 2022. Fluenttts: Text-dependent fine-grained style control for multi-style tts. In *Interspeech 2022*, pages 4561–4565.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Russell V. Lenth. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.5.

Max Planck Institute for Psycholinguistics, The Language Archive. 2026. ELAN (version 7.1) [computer software]. Retrieved from https://archive.mpi.nl/tla/elan.

Paul Mayer, Florian Lux, Alejandro Pérez-González-de Martos, Angelina Elizarova, Lindsey Vanderlyn, Dirk Väth, and Ngoc Thang Vu. 2025. Investigating stochastic methods for prosody modeling in speech synthesis. In *Interspeech 2025*, pages 439–443.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. Meta-stylespeech : Multi-speaker adaptive text-to-speech generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7748–7759. PMLR.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4693–4702. PMLR.

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Lori A. Swanson, Laurence B. Leonard, and Jack Gandour. 1992. Vowel Duration in Mothers' Speech to Young Children. *Journal of Speech, Language, and Hearing Research*, 35(3):617–625. _eprint: https://pubs.asha.org/doi/pdf/10.1044/jshr.3503.617.

Erik D. Thiessen, Emily A. Hill, and Jenny R. Saffran. 2005. Infant-directed speech facilitates word segmentation. *Infancy*, 7(1):53–71.

Laurel J. Trainor, Cynthia M. Austin, and Renée N. Desjardins. 2000. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, 11(3):188–195.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Interspeech 2017*, pages 4006–4010.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5180–5189. PMLR.

Janet F Werker and Peter J McLeod. 1989. Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 43(2):230–246.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949.