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

Jinyoung Jo [1] **Katherine Nguyen** [2] **Sean Choi** [2]

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 investigated whether a text-to-speech (TTS) model fine-tuned on CDS and ADS reproduces the register-related prosodic differences observed in human speech. The results showed that the generated speech replicated several register-specific prosodic patterns observed in human speech, including a higher pitch and a slower articulation rate in CDS relative to ADS. However, the magnitude of CDS-ADS differences in pitch measures was attenuated in the generated speech, while the differences in articulation rate 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 & 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?

Current TTS architectures are typically optimized for intelligibility and naturalness in relatively neutral speaking styles (Casanova et al., 2024; Ren et al., 2019; Shen et al., 2018; Wang et al., 2017). Because these systems are primarily trained on read speech or audiobooks, their baseline data predominantly features standard ADS, making them less suitable for capturing other speech styles, such as the highly exaggerated prosody of CDS. However, given evidence that children attend better to CDS than ADS (Werker & McLeod, 1989), the ability of generative speech systems to appropriately modulate speech registers may become increasingly important as these systems are deployed in child-oriented interactive contexts.

In this study, we investigate whether TTS models fine-tuned separately on ADS and CDS preserve the prosodic characteristics that distinguish CDS from ADS in human speech, including elevated pitch (higher mean, minimum and maximum F0 within an utterance), expanded pitch range (greater F0 range within an utterance) and temporal slowing (slower articulation rate) of CDS. More broadly, this study examines the extent to which generative speech systems can internalize socially meaningful register distinctions from specialized training data.

1.1. Related Work

Recent work in neural TTS synthesis has increasingly focused on whether models can learn and reproduce expressive speaking styles, including variation in pitch, timing, and rhythm. A major line of research has explored latent style representations that enable control and transfer of speaking style without explicit supervision (Akuzawa et al., 2018; Hsu et al., 2018; Wang et al., 2018; Zhang et al., 2019). Related work has further shown that neural TTS systems can perform cross-speaker prosody transfer, i.e., extracting prosodic characteristics from a reference speaker and applying them to a different target voice, and provide control

[1]Department of Linguistics, Stanford University, Stanford, CA, USA [2]Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA. Correspondence to: Sean Choi <sean.choi@scu.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

over expressive variation (Chen & Rudnicky, 2022; Kim et al., 2022; Min et al., 2021; Skerry-Ryan et al., 2018). Together, these studies suggest that TTS systems can encode stylistic and prosodic properties independently of content and speaker identity.

A challenge in expressive TTS concerns the tendency of deterministic systems to produce flattened or "average" prosody that lacks the variability found in natural human speech (Hodari et al., 2019). To address this limitation, recent work has explored stochastic approaches for modeling prosodic features, demonstrating improved variability and more human-like expressive dynamics (Mayer et al., 2025).

Together, these studies demonstrate that neural TTS systems can learn and reproduce stylistic variation in speech. However, most prior work has focused on emotion or speaker adaptation rather than listener-conditioned register variation such as the distinction between CDS and ADS. The present study extends this literature by examining whether fine-tuned TTS models reproduce the prosodic differences associated with CDS and ADS. Further, CDS represents a relatively low-resource register due to the limited availability of large caregiver-child speech corpora with high-quality annotations. Thus, we evaluate whether fine-tuning can nevertheless capture prosodic patterns of CDS.

2. Methods

2.1. Data

We obtained CDS and ADS datasets from the ECOLANG multimodal corpus (Gu et al., 2025). The corpus includes 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.

Given well-established gender-based differences in pitch (Simpson, 2009) and speaking rate (Jacewicz et al., 2009), as well as prosodic differences between British and American English (Grice et al., 2020; 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 conversion 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 between approximately 5 and 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 ut-

terances using input text drawn from the test sets. Prior to sampling this final subset, 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 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 human speech test data 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 & Weenink, 2026), to calculate these measures for individual utterances.

Statistical analyses 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 dependent variable (mean F0, minimum F0, maximum F0, F0 range, and articulation rate). 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, along with a random intercept for each CHECKPOINT nested within REGISTER to account for variability associated with different stages of model training. 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. Detailed statistical results are summarized in Table 1.

Generated speech exhibited higher values than human speech for most F0 measures: significant main effects of VOICE TYPE were observed for mean F0 and maximum F0, with generated speech showing higher values than human speech. In addition, generated speech exhibited a greater F0 range than human speech.

Crucially, the pairwise comparisons of REGISTER (CDS vs. ADS) within each VOICE TYPE showed that generated speech largely reproduced the register-related pitch patterns observed in the human speech data. Mean F0 and minimum F0 were significantly higher in CDS than ADS for both generated and human speech. Maximum F0 did not differ significantly between registers in either voice type, and F0 range was significantly smaller in CDS than ADS in both generated and human speech. Although this pattern differs from previous reports of higher maximum F0 and wider F0 ranges in CDS (e.g., Fernald et al., 1989), the generated speech nevertheless replicated the pattern observed in the human speech data in the present study.

However, significant VOICE TYPE × REGISTER interactions were observed for mean F0, minimum F0, and F0 range. In all three cases, the CDS–ADS difference was smaller in generated speech than in human speech, indicating attenuation of register-related pitch distinctions in the generated speech.

Taken together, these results indicate that the generated speech reproduced the direction of several register-related pitch differences observed in human speech, while generally reducing the magnitude of those differences.

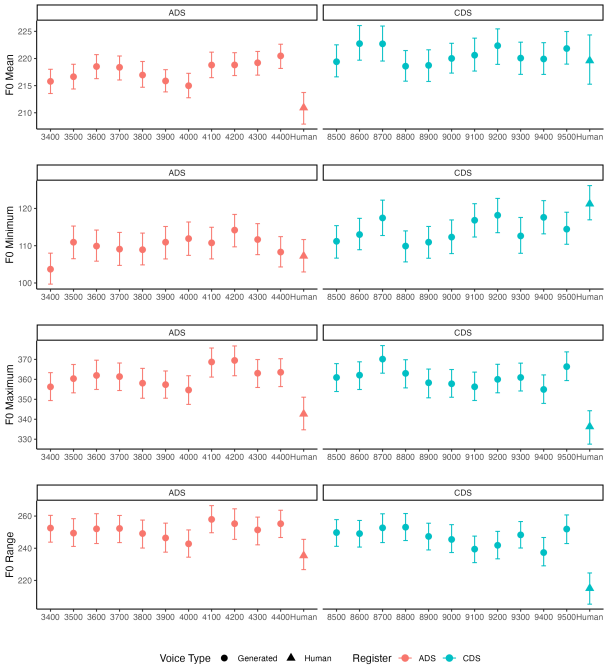

Figure 1. Mean F0, minimum F0, maximum F0, and F0 range values for generated ADS and CDS across training checkpoints, compared with the corresponding human speech test data. Each measure was first calculated per utterance, and the resulting utterance-level values were then averaged within each training checkpoint. Points represent these averaged values, and error bars indicate 95% confidence intervals.

Table 1. Summary of mixed-effects model results for F0 measures and articulation rate.

Measure	Voice Type Effect	Register Effect	Voice Type × Register Interaction
Mean F0	Generated > Human $\beta = 1.94$, $SE = 0.60$, $p < .01$	Generated: CDS > ADS $\beta = 2.96$, $SE = 0.70$, $p < .001$ Human: CDS > ADS $\beta = 8.68$, $SE = 2.31$, $p < .001$	Attenuation in generated speech $\beta = -1.43$, $SE = 0.60$, $p < .05$
Minimum F0	Not significant $\beta = -1.10$, $SE = 1.05$, $p = .31$	Generated: CDS > ADS $\beta = 4.01$, $SE = 1.21$, $p < .001$ Human: CDS > ADS $\beta = 13.99$, $SE = 4.01$, $p < .001$	Attenuation in generated speech $\beta = -2.49$, $SE = 1.05$, $p < .05$
Maximum F0	Generated > Human $\beta = 10.84$, $SE = 1.70$, $p < .001$	Generated: Not significant $\beta = -0.37$, $SE = 1.96$, $p = .85$ Human: Not significant $\beta = -6.34$, $SE = 6.49$, $p = .33$	Not significant $\beta = 1.50$, $SE = 1.70$, $p = .39$
F0 Range	Generated > Human $\beta = 11.94$, $SE = 1.79$, $p < .001$	Generated: CDS < ADS $\beta = -4.38$, $SE = 2.07$, $p < .05$ Human: CDS < ADS $\beta = -20.33$, $SE = 6.86$, $p < .01$	Attenuation in generated speech $\beta = 3.99$, $SE = 1.79$, $p < .05$
Articulation Rate	Not significant $\beta = -0.01$, $SE = 0.02$, $p = .64$	Generated: CDS < ADS $\beta = -0.35$, $SE = 0.03$, $p < .001$ Human: CDS < ADS $\beta = -0.17$, $SE = 0.08$, $p < .05$	Trend toward exaggeration in generated speech $\beta = -0.05$, $SE = 0.02$, $p = .053$

3.2. Articulation Rate

Figure 2 shows articulation rates for generated ADS and CDS across checkpoints, alongside the corresponding human speech test data; statistical results are presented in Table 1. Crucially, CDS exhibited significantly slower articulation rates than ADS in both generated and human speech, indicating that the generated speech reproduced the register-related temporal pattern observed in the human speech data. Although the REGISTER × VOICE TYPE interaction did not reach significance, the larger CDS–ADS difference in generated speech suggests an exaggerated register effect relative to the human speech data. No significant overall difference was observed between generated and human speech.

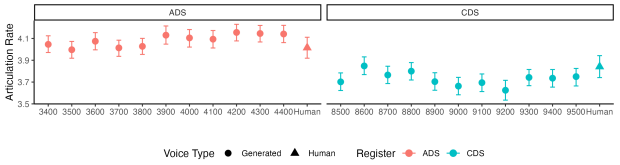

Figure 2. Articulation rates for generated ADS and CDS across training checkpoints, compared with the corresponding human speech test data. Articulation rate was first calculated per utterance, and the resulting utterance-level values were then averaged within each training checkpoint. Points represent these averaged values, 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. 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 internalize 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 patterns that characterize social registers. These findings suggest that systematic variations driven by communicative contexts, such as the distinct prosodic profile of CDS, can be internalized even under low-resource training conditions, although some limitations remain.

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. Moreover, the corpus contains different speakers for ADS and CDS, making it difficult to disentangle register effects from speaker-specific characteristics. Future work using larger datasets is needed to examine whether the findings generalize beyond the current corpus; constructing such datasets is challenging, as relatively few corpora provide large amounts of speech data with time-aligned transcripts from the same speakers across both ADS and CDS.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Akuzawa, K., Iwasawa, Y., and Matsuo, Y. Expressive speech synthesis via modeling expressions with variational autoencoder. In *Interspeech 2018*, pp. 3067–3071, 2018. doi: 10.21437/Interspeech.2018-1113.

Boersma, P. and Weenink, D. Praat: doing phonetics by computer [computer program], 2026. URL https://praat.org. Retrieved 24 April 2026.

Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., and Weber, J. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pp. 4978–4982, 2024. doi: 10.21437/Interspeech.2024-2016.

Chen, L.-W. and Rudnicky, A. 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)*, pp. 7907–7911, 2022. doi: 10.1109/ICASSP43922.2022.9747747.

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

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., and Fukui, I. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16(3):477–501, 1989. doi: 10.1017/S0305000900010679.

Grice, M., German, J. S., and Warren, P. Intonation systems across varieties of english. In Gussenhhoven, C. and Chen, A. (eds.), *The Oxford Handbook of Language Prosody*. Oxford University Press, 12 2020. ISBN 9780198832232. doi: 10.1093/oxfordhb/9780198832232.013.18. URL https://doi.org/10.1093/oxfordhb/9780198832232.013.18.

Gu, Y., Donnellan, E., Grzyb, B., Brekelmans, G., Murgiano, M., Brieke, R., Perniss, P., and Vigliocco, G. The ECOLANG Multimodal Corpus of adult-child and adult-adult Language. *Scientific Data*, 12(1): 89, January 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04405-1. URL https://doi.org/10.1038/s41597-025-04405-1.

Hodari, Z., Watts, O., and King, S. Using generative modelling to produce varied intonation for speech synthesis. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pp. 239–244. ISCA, 2019. doi: 10.21437/SSW.2019-43.

Hsu, W.-N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., Cao, Y., Jia, Y., Chen, Z., Shen, J., Nguyen, P., and Pang, R. Hierarchical generative modeling for controllable speech synthesis, 2018. URL https://arxiv.org/abs/1810.07217.

Jacewicz, E., Fox, R. A., O'Neill, C., and Salmons, J. Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21(2):233–256, 2009. doi: 10.1017/S0954394509990093.

Jadoul, Y., Thompson, B., and de Boer, B. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, 2018. doi: https://doi.org/10.1016/j.wocn.2018.07.001.

Kim, C., Um, S., Yoon, H., and Kang, H.-G. Fluenttts: Text-dependent fine-grained style control for multi-style tts. In *Interspeech 2022*, pp. 4561–4565, 2022. doi: 10.21437/Interspeech.2022-988.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017. doi: 10.18637/jss.v082.i13.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lenth, R. V. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024. URL https://CRAN.R-project.org/package=emmeans. R package version 1.10.5.

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

Mayer, P., Lux, F., Pérez-González-de Martos, A., Elizarova, A., Vanderlyn, L., Väth, D., and Vu, N. T. Investigating stochastic methods for prosody modeling in speech synthesis. In *Interspeech 2025*, pp. 439–443, 2025. doi: 10.21437/Interspeech.2025-1940.

Min, D., Lee, D. B., Yang, E., and Hwang, S. J. 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*, pp. 7748–7759. PMLR, 2021. URL https://proceedings.mlr.press/v139/min21b.html.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech: Fast, Robust and Controllable Text to Speech. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f63f65b503e22cb970527f23c9ad7db1-Paper.pdf.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018. doi: 10.1109/ICASSP.2018.8461368.

Simpson, A. P. Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2):621–640, 2009. doi: https://doi.org/10.1111/j.1749-818X.2009.00125.x. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2009.00125.x.

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R., Clark, R., and Saurous, R. A. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4693–4702. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/skerry-ryan18a.html.

Soderstrom, M. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532, 2007. ISSN 0273-2297. doi: https://doi.org/10.1016/j.dr.2007.06.002. URL https://www.sciencedirect.com/science/article/pii/S0273229707000202.

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

Thiessen, E. D., Hill, E. A., and Saffran, J. R. Infant-directed speech facilitates word segmentation. *Infancy*, 7(1):53–71, 2005. doi: https://doi.org/10.1207/s15327078in0701_5. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15327078in0701_5.

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

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. Tacotron: Towards End-to-End Speech Synthesis. In *Interspeech 2017*, pp. 4006–4010, 2017. doi: 10.21437/Interspeech.2017-1452.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., and Saurous, R. A. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5180–5189. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/wang18h.html.

Werker, J. F. and McLeod, P. J. 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, 1989.

Zhang, Y.-J., Pan, S., He, L., and Ling, Z.-H. 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)*, pp. 6945–6949, 2019. doi: 10.1109/ICASSP.2019.8683623.