Interspeech 2024
1-5 September 2024, Kos, Greece

A Human-in-the-Loop Approach to Improving Cross-Text Prosody Transfer

Himanshu Maurya[1*], *Atli Sigurgeirsson*[1*]

[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

lordzuko.research@gmail.com, s2063518@ed.ac.uk

Abstract

Text-To-Speech (TTS) prosody transfer models can generate varied prosodic renditions, for the same text, by conditioning on a reference utterance. These models are trained with a reference that is identical to the target utterance. But when the reference utterance differs from the target text, as in cross-text prosody transfer, these models struggle to separate prosody from text, resulting in reduced perceived naturalness.

To address this, we propose a Human-in-the-Loop (HitL) approach. HitL users adjust salient correlates of prosody to make the prosody more appropriate for the target text, while maintaining the overall reference prosodic effect. Human adjusted renditions maintain the reference prosody while being rated as more appropriate for the target text 57.8% of the time. Our analysis suggests that limited user effort suffices for these improvements, and that closeness in the latent reference space is not a reliable prosodic similarity metric for the cross-text condition.

Index Terms: Speech synthesis, Prosody, Human-in-the-Loop

1. Introduction

Prosody is used to encode information that is not fully conveyed by linguistic information alone. The same lexical information can be encoded through different prosodic renditions to deliver different meanings and influence particular interpretations [1]. Knowing which rendition is most appropriate at any given time is a non-trivial problem. Typical end-to-end TTS models generate a single *mean* prosodic rendition, reflecting the overall prosodic distribution of the training corpus. Treating prosody in this way yields speech that can be perceived as dull, inexpressive, or inappropriate for the given target text, which negatively affects general comprehension [2].

2. Background

One way of controlling prosody is to learn the prosodic distribution of the training corpus using a prosody-labelled TTS corpus. Prosody labels can be used to jointly train prosody predictors, which can be controlled to generate different prosodic renditions for the same input text [3, 4, 5]. However, few prosody-labelled corpora are generally available. Therefore, much of the recent work on prosody control has shifted to methods that aim to discover the prosodic distribution without supervision. This includes Prosody-Transfer (PT) models [6, 7, 8, e.g.].

Typically, PT models jointly train a *reference encoder* with an underlying acoustic model. This encoder models a highly constrained representation of the target acoustics, which is taken to represent prosody. The prosody representation is then used to condition speech generation in the underlying acoustic model. Under the right conditions, such models are claimed to be able to transfer prosody from any reference to any text [7].

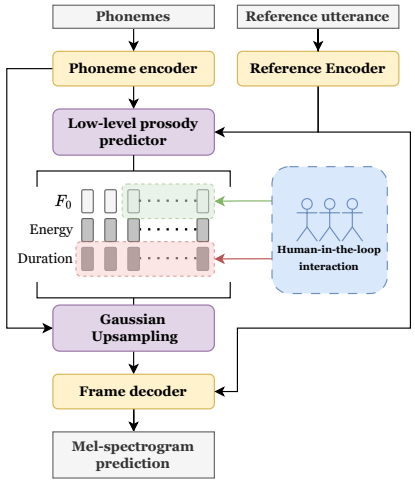


Figure 1: *Overview of the proposed method. Human-in-the-loop participants adjust word- and utterance-level F_0, energy and duration features to improve prosody in cross-text prosody transfer.*

A common PT use case is *cross-text* PT; a reference with the target prosody exists, but the target text does not match the reference text. Under cross-text conditions, however, some PT models have demonstrated a trade-off between perceived naturalness and the overall quality of prosody transfer [9]. This effect is commonly attributed to *acoustic feature entanglement* [6] and *source-speaker leakage* [8]. The fact that PT models are typically not trained on cross-text or cross-speaker samples is believed to give rise to these issues [9]. As a result of the PT training regime, the modelled prosody representations are highly dependent on the reference text, which means that they are not strictly *transferable* to any text [9].

These issues with PT models do give rise to questions about whether cross-text prosody transfer can even be achieved *faithfully*. That is, can you transfer prosody to any text such that the resulting rendition is true to the reference prosody *and* is perceived natural for the target text? If the latent reference space models transferable representations of prosody, then a prosody representation that fits that goal should exist. But if it is not retrievable by conditioning on the reference prosody, it becomes a question about how else that representation can be found in the reference space.

[*]These authors contributed equally to this work

10.21437/Interspeech.2024-2055

One possible way to do this would be to directly incorporate human perception in prosody prediction, using feedback from a *Human-in-the-loop* (HitL) [10]. Human feedback has been used for a wide spectrum of tasks in Natural Language Processing (NLP); such as entity extraction and linking [11, 12, 13] and more subjective tasks such as reading comprehension [14] and question-answering [15]. Fine-tuning models on relatively small sets of human-annotated corrections in these tasks has shown improved model performance and robustness [10]. HitL has also been employed in several TTS-specific tasks, such as modelling emotion [16], speaker identity [17], and speaking style [18]. These attempts demonstrate that HitL can, to some extent, be applied to TTS to improve particular perceptual objectives.

An important consideration for any HitL based method is how participants interact with the underlying model. Previous work such as [16, 19] use HitL to explore emotional prosodic variation of neutral utterances. In their work, HitL participants make iterative adjustments in isolated dimensions of a learnt latent prosody space to explore different variations of emotional prosody. HitL interaction with the model is shown to make improvements compared to the baseline model. However, it is a very time-consuming approach, requiring 48 hours to generate 40 samples in [19]. Furthermore, applying HitL directly in a modelled latent space could be highly uninterpretable, as latent dimensions are not guaranteed to encode any perceptually distinct information type or be independent of other latent dimensions [20]. Some methods [20, 21] instead, control prosody in terms of known acoustic correlates of prosody, such as F_0, duration, spectral tilt and energy. There is no established *control-resolution* for HitL based prosody control; [21] suggests using utterance-level control, while [20] controls prosody at the phone-level. Although [20] suggests that phone-level control allows HitL participants to make detailed decisions to improve the output prosody, they also suggest that a higher-level abstraction is perhaps more conducive for a HitL approach.

We investigate whether the prosody representations modelled by a particular PT-model [7] can be improved specifically for the cross-text PT task. The PT model separately models phone-level F_0, energy and duration. These are salient acoustic correlates of prosody, and together they cover the *control space* that our HitL participants interact with. Participants control utterance-level and word-level features through a web-based user interface designed to carry out the experiments. The proposed design enables fewer, interpretable edit iterations when compared to prior HitL work in TTS. We evaluate the human effort involved in achieving the desired prosodic effect for the cross-text PT task. Our results indicate that HitL participants can discover prosodic renditions that are more appropriate for the given target text while maintaining most of the reference prosody effect.

3. Method

3.1. Model Architecture

The proposed HitL method does not require any online supervision and can therefore be implemented using any pre-trained controllable PT-model. Our experiments aim to provide insights into the wider PT-task, instead of any specific model. We therefore choose *Daft-Exprt* [7] as our baseline, since they demonstrate good performance in the cross-text setting. Daft-Exprt is a fully parallel TTS model that builds on the FastSpeech-2 architecture. Daft-Exprt uses a reference-encoder which generates a fixed-size reference embedding. This embedding is used to condition speech generation in the underlying acoustic model to enable prosody transfer. Conditioning is performed by predicting FiLM-parameters [22] for the *low-level prosody predictor* and the frame decoder. The low-level prosody predictor predicts phone-level log- F_0, energy and duration. These predicted values, with encoded phonemes, are used to decode the target Mel-spectrogram. We use HiFi-GAN [23] to convert the predicted Mel-spectrograms to waveform.

3.2. Human-in-the-loop approach

HitL participants interact with Daft-Exprt through the low-level prosody predictor. Given a reference and a target text, an initial cross-text PT sample is generated. Participants are asked to adjust the predicted F_0, energy and duration values to make the prosody more appropriate for the target text, while preserving the prosody of the reference. Manipulating individual phone-level predictions is believed to be too complex for the proposed HitL task [20]. So instead, we devise a control interface for making utterance- and word-level adjustments to prosody. Taking inspiration from existing interfaces for interaction with TTS systems [24, 25], we develop a web-based User Interface (UI), using Streamlit[1], to facilitate the proposed HitL method. Control inputs are received through the UI and corresponding F_0, energy and duration are computed, as explained in Sections 3.2.1-3.2.2. These values are then sent to the acoustic model to complete synthesis.

3.2.1. Word-level Control

Daft-Exprt predicts phone-level F_0, energy and duration. Therefore, word-level control inputs have to be mapped to corresponding phone-level adjustments. We treat F_0 and energy adjustments in the same way, while treating duration adjustments slightly differently. First, the acoustic model predicts all phone-level F_0 and energy values $v_1, v_2, \ldots, v_n$ (representing either phone-level F_0 or energy predictions henceforth). We then compute a word-level feature mean, K_w, and a per-phone scaling factor, s_i for each input word $w = p_1, p_2, ..., p_n$:

$$K_w = \frac{1}{|w|}\sum v_i, \quad s_i = \frac{K_w}{v_i}, \quad i \in 1, \ldots, n$$

The user can then suggest changing the initial mean feature value K_w to a new one, K'_w. This results in per-phone changes of that word in such a way that:

$$v'_i = \frac{K'_w}{s_i}, \quad i \in 1, \ldots, n$$

This form of control therefore results in equal adjustments to all phones p_i in the word, proportional to their originally predicted value v_i. Note that in case of F_0, voiceless phones are not modified and do not contribute to the computation of the value of K_w. This control mechanism is presented to the user in the form of a *slider*. The control slider is initially set to K_w, allowing for separately increasing or decreasing F_0 and energy values for each word in the input text. F_0 and energy values far outside the known training distribution can result in unintelligible output. We therefore determine a suitable range for both F_0 and energy modifications. From our initial experimentation, we found that limiting the F_0 range to $\pm 3\sigma$ and energy to $\pm 1.5\sigma$ resulted in a good control range for this task. We therefore limit

[1] https://streamlit.io/

F_0 and energy control such that no edited phone-level F_0 or energy value falls outside this range.

This control mechanism is counterintuitive for duration, since, by its design, it would set the initial duration of every word to its phone-duration mean. Therefore, we resort to a simpler approach for duration control. Participants can equally scale the duration of each phone in the word by a constant within the range $[0, 2]$. That is, each word can be made up to twice as long in duration. This design technically allows participants to set the duration of any word to 0 seconds, but no HitL participant did. Like with F_0 and energy, word-level duration is controlled using a slider UI element.

3.2.2. Utterance-level control

We also provide utterance-level controls for F_0, energy and duration. Users also interact with these controls through slider UI elements. We believe that applying global changes to prosody may help participants better control emotion, expression, and overall tone of the output. This is similar to the control scheme used in [26]. Again, we treat F_0 and energy control in the same way. When a HitL participant submits an utterance-level control input, we compute the appropriate word-level control input for each word and apply it. We determine the utterance-level control-range such that any resulting word-level change remains within the statistical per-phone ranges described in Section 3.2.1. The utterance-level duration control simply allows participants to scale all phone durations equally within the range $[0, 2]$.

4. Experiments

4.1. Baseline Model

The baseline Daft-Exprt model is trained on the 2013 Blizzard Challenge corpus [27], which is an expressive single-speaker book narration corpus. We use only the segmented split of the corpus, which comprises 52 hours from 55 different books. Utterances shorter than 0.3 seconds and longer than 15 seconds are excluded. The remaining, approximately 40,000 utterances ($\approx$ 40 hours), are used to train the baseline PT-model. We follow the original Daft-Exprt preprocessing procedure [7] and extract alignments, using the Montreal forced aligner [28], and phone-level F_0 and energy estimations.

The model is trained for 24 hours, distributed over 4 NVIDIA Tesla V100-SXM2-16GB GPUs, using a batch size of 96 samples. We train a HiFi-GAN [23] vocoder using the same corpus as is used to train the acoustic model. The vocoder is trained for 12 hours on a single NVIDIA Tesla V100-SXM2-16GB GPU using a batch size of 8 samples.

4.2. Edited Speech Corpus

In our proposed method, participants first listen to the reference utterance used for the cross-text PT. They are then asked to make adjustments to F_0, energy, and duration to make the resulting prosody more appropriate for the target text. Participants are asked to make these adjustments while maintaining the perceived naturalness of the output and the overall prosodic effect of the reference utterance. After making the adjustments, the *edited* sample and the initially synthesised cross-text PT sample are saved. We also ask participants to indicate how confident they are (*"low"/"high"*) in their suggestions.

To account for any possible individual bias, we ask all of our participants to make adjustments to the same list of cross-text PT samples. We choose 5 reference utterances on the basis that each one would likely yield a perceptually distinct effect in the PT output. We create four target texts for the task that would typically elicit a particular prosodic effect. The target texts are all short (5 words or fewer) to keep the HitL process as simple as possible. As a result, the target texts are notably shorter than the text read for the reference utterances we use. The 5 references and 4 target texts create $5 \times 4 = 20$ reference-target pairs which are used to generate the 20 cross-text PT samples that each HitL participant modifies. A full list of references and target texts is available on our demo page[2].

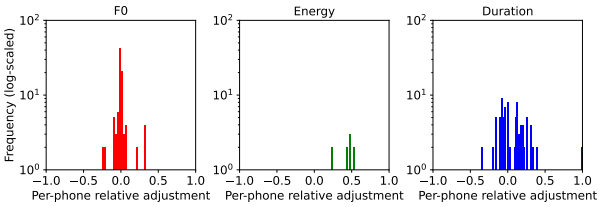


Figure 2: *The distributions of phone-level control-inputs for F_0, energy and duration (unchanged values omitted). These results suggest that energy control was substantially less important to our HitL participants than F_0 and duration control.*

We recruit 33 HitL participants with a diverse language background to take part in our study. Their participation resulted in $5 \times 4 \times 33 = 660$ *original* / *edited* sample pairs. Participants spent on average 123.5 ± 138.0 seconds modifying each sample, resulting in on average 7.0 ± 7.1 individual operations per utterance. Of the 660 pairs, 193 sample pairs were unmodified. In these cases, the participants deemed the *original* prosodic rendition appropriate enough for the target text. These identical pairs and 47 additional unusable pairs were removed from the final set. The resulting 420 pairs are used in our subjective and objective experiments. Of the 420 pairs, the participants indicated a high confidence in their performance for 82.6% and a low confidence for 17.4%. Interestingly, as shown in Figure 2, participants made around $3\times$ fewer edits to predicted energy values when compared to F_0 and duration. This could mean that HitL participants did not find energy-control to be as useful or important to complete the task.

4.3. Evaluation

We evaluate perceived naturalness, quality of prosody transfer, and prosodic appropriateness. We use a 5-point Likert scale to evaluate perceived naturalness in a standard Mean Opinion Score (**MOS**) survey. We follow the **MUSHRA-like** design proposed in [6] to evaluate prosody transfer. We also include a sample that is generated using a random reference utterance. A random reference is unlikely to be informative about the target prosody, so we view this randomly generated sample as an anchor in our design. Participants first listen to the reference utterance used to generate the *original* cross-text PT sample. Then they listen to the *original* sample, the *edited* sample, and the random anchor. They then indicate, on a scale from 0 to 100, how prosodically similar these samples are to the reference. We use a simple **A/B** preference design to evaluate the prosodic preference between *original* and *edited* samples. Participants are asked to indicate this preference in terms of how appropriate the prosodic rendition is for the synthesised text.

[2]https://linktr.ee/anonymousconferenceuser

We evaluate all the 420 pairs described in Section 4.2. This results in 840 MOS questions, 420 MUSHRA-like screens, and 420 A/B preference questions. We recruited 68 native UK/US participants through Prolific[3]. Each sample is evaluated by at least 3 different raters, and at most 5. Table 1 summarises our subjective evaluation results.

Table 1: *Main subjective results broken down by HITL participant confidence. The highlighted row includes high-confidence results of the proposed method.*

Participant confidence		MOS	MUSHRA like	A/B
Low	*original*	3.1 ± 0.7	61.2 ± 13.4	50.7%
	edited	2.7 ± 0.8	53.0 ± 18.1	49.3%
High	*original*	3.2 ± 0.7	58.8 ± 13.1	40.4%
	edited	$\mathbf{3.0 \pm 0.8}$	$\mathbf{55.0 \pm 14.3}$	**59.6%**
Total	*original*	3.2 ± 0.7	59.2 ± 13.2	42.1%
	edited	3.0 ± 0.7	54.7 ± 13.2	57.8%

The *edited* samples are perceived to be slight, although not significantly, less prosodically similar to the reference. The *edited* samples are also perceived as slightly less natural than the *original* samples. However, when HitL participants indicate high confidence in their efforts, this difference decreases. Furthermore, in these cases, the evaluators indicate a high preference for the *edited* prosodic renditions at a rate of 59.6%. We hypothesised that an improvement in the prosodic renditions of cross-text PT samples would consequently translate into an improvement in perceived naturalness. Our results do not support that claim. However, we hypothesise that the drop in perceived naturalness can largely be explained by artefacts, audible discontinuities and interruptions introduced by our HitL control procedure.

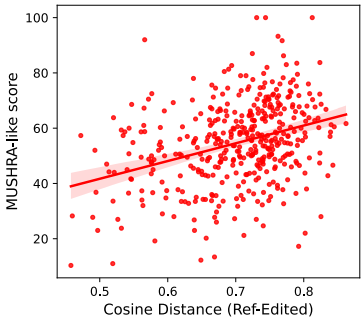


Figure 3: *MUSHRA-like scores of edited cross-text PT samples plotted against the cosine distance between the original reference embedding and the embedded edited cross-text PT sample. A linear regression model is fitted to this data, the shaded red area indicates the 95% confidence interval.*

We investigate the relationship between the reference embeddings and embedded synthesised samples. We speculated that synthesised PT samples that are rated as prosodically similar to the reference would be, once embedded, close to the reference embedding in the latent space. However, we observe the opposite as shown in Figure 3. Edited cross-text PT samples that are found to be prosodically similar to the reference tend to be further away from the reference embedding in the latent space. A similar trend was not found for unedited cross-text PT samples. We believe these results demonstrate two things: 1) participants can identify a *"prosodic intent"* from the reference and *faithfully* modify the cross-text PT sample with regard to the identified intent and the target text, and 2) Closeness to a reference embedding is not a reliable metric for prosodic similarity for cross-text PT.

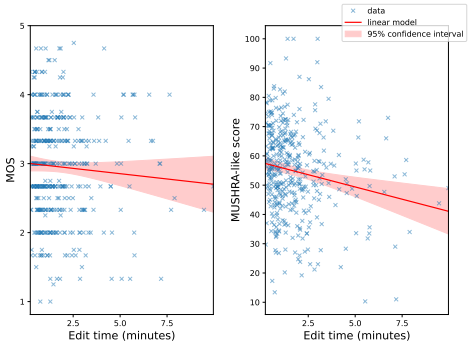


Figure 4: *The relationship between HitL effort and the quality of the output. Left: the line of best fit for perceived naturalness as a function of HitL effort, Right: the same but for perceived quality of prosody transfer.*

We finally study the relationship between HitL participant effort and the quality of the output. Here, we take the time spent editing a sample as an indicator of a participant's effort. One might assume that perceived naturalness would be positively correlated with the overall effort of HitL participants. However, this is not the case, as demonstrated by Figure 4. We fit a linear regression model to our MOS results as a function of the time taken to make the modifications. Contrary to our assumption, HitL effort negatively correlates with perceived naturalness. There are many plausible reasons for this result, but it underscores the importance of a robust user interface for the proposed method.

5. Conclusions

Our findings suggest that HitL participants can make meaningful changes to cross-text PT samples. Participants can make the output prosody more appropriate for the target text while preserving the reference prosody. However, this does not lead to an improvement in perceived naturalness. The results shown in Figure 4 also indicate that more interaction with the model leads to a decrease in naturalness. A notable outcome, highlighted in Figure 3, suggests that a reference embedding predicted by a PT model may yield perceptually different prosody, dependent on the target text.

We make simplifying assumptions to reduce overall HitL effort. For example, we do not control pause durations or change word-level F_0 contours in a nuanced manner. Participants therefore had limited control over the full range of prosodic renditions. Despite this, we believe that the current work demonstrates that HitL participants can identify more appropriate prosodic renditions for the cross-text PT task, while preserving the reference prosody. Moreover, our results indicate that under cross-text conditions, *closeness* to the reference embedding is not a reliable metric for measuring prosodic similarity.

[3]https://www.prolific.com/

6. References

[1] D. Wilson and T. Wharton, "Relevance and prosody," *Journal of pragmatics*, vol. 38, no. 10, pp. 1559–1579, 2006.

[2] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," *Proc. Interspeech 2018*, pp. 2838–2842, 2018.

[3] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. M. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5734–5738, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225067708

[4] G. E. Henter, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *ArXiv*, vol. abs/1807.11470, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:51871648

[5] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3331–3340.

[6] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80, 10–15 Jul 2018, pp. 4693–4702.

[7] J. Zaïdi, H. Seut'e, B. van Niekerk, and M.-A. Carbonneau, "Daft-exprt: Cross-speaker prosody transfer on any text for expressive speech synthesis," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:247997035

[8] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," *Proc. Interspeech 2020*, pp. 4387–4391, 2020.

[9] A. T. Sigurgeirsson and S. King, "Do prosody transfer models transfer prosody*f*," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X22001790

[11] A. L. Gentile, D. Gruhl, P. Ristoski, and S. Welch, "Explore and exploit. dictionary expansion with human-in-the-loop," in *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*. Springer, 2019, pp. 131–145.

[12] S. Zhang, L. He, E. Dragut, and S. Vucetic, "How to invest my time: Lessons from human-in-the-loop entity extraction," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2305–2313.

[13] J.-C. Klie, R. E. de Castilho, and I. Gurevych, "From zero to hero: Human-in-the-loop entity linking in low resource domains," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 6982–6993.

[14] M. Bartolo, A. Roberts, J. Welbl, S. Riedel, and P. Stenetorp, "Beat the ai: Investigating adversarial human annotation for reading comprehension," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 662–678, 2020.

[15] E. Wallace, P. Rodriguez, S. Feng, I. Yamada, and J. Boyd-Graber, "Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 387–401, 2019.

[16] P. M. C. Harrison, R. Marjieh, F. Adolfi, P. van Rijn, M. Anglada-Tort, O. Tchernichovski, P. Larrouy-Maestri, and N. Jacoby, "Gibbs sampling with people," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[17] K. Udagawa, Y. Saito, and H. Saruwatari, "Human-in-the-loop speaker adaptation for dnn-based multi-speaker tts," in *Interspeech*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249889440

[18] T. Cornille, F. Wang, and J. Bekker, "Interactive multi-level prosody control for expressive speech synthesis," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8312–8316, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249437286

[19] P. van Rijn, S. Mertes, D. Schiller, P. M. C. Harrison, P. Larrouy-Maestri, E. Andr'e, and N. Jacoby, "Exploring emotional prototypes in a high dimensional tts latent space," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233739719

[20] D. S. R. Mohan, Q. Hu, T. H. Teh, A. Torresquintero, C. G. R. Wallis, M. Staib, L. Foglianti, J. Gao, and S. King, "Ctrl-p: Temporal control of prosodic variation for speech synthesis," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235446341

[21] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221703703

[22] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI Conference on Artificial Intelligence*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:19119291

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[24] D. Kondo and M. Morise, "Human-in-the-loop speech-design system and its evaluation," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 608–612.

[25] N. Tits, K. El Haddad, and T. Dutoit, "Ice-talk 2: Interface for controllable expressive tts with perceptual assessment tool," *Software Impacts*, vol. 8, p. 100055, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665963821000038

[26] A. T. Sigurgeirsson and S. King, "Using a large language model to control speaking style for expressive tts," *arXiv preprint arXiv:2305.10321*, 2023.

[27] S. King and V. Karaiskos, "The blizzard challenge 2013," 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:166265879

[28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.