Low-dimensional style token control for hyperarticulated speech synthesis

Nishihara, Miku; Wells, Dan; Richmond, Korin; Pine, Aidan

Publisher's version / Version de l'éditeur:

https://doi.org/10.21437/Interspeech.2024-2074

Interspeech 2024, pp. 3385-3389, 2024-09-01

Interspeech 2024
1-5 September 2024, Kos, Greece

Low-dimensional Style Token Control for Hyperarticulated Speech Synthesis

Miku Nishihara[1*], *Dan Wells*[2*], *Korin Richmond*[2], *Aidan Pine*[3]

[1]Department of Computer Science, Nagoya Institute of Technology, Japan
[2]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[3]National Research Council Canada, Canada

m.nishihara.966@stn.nitech.ac.jp, {dan.wells, korin.richmond}@ed.ac.uk,
aidan.pine@nrc-cnrc.gc.ca

Abstract

Global style tokens (GSTs) allow for rich modelling of the variation in a speech corpus and subsequent control of text-to-speech synthesis (TTS). However, certain styles of speech may be marked by variation along multiple dimensions, complicating the interpretation and control of learned style tokens. One example is hyperarticulated or 'clear' speech, for example as directed toward listeners with hearing impairments or language learners in the classroom, which in English is characterised by reduced speaking rate, increased F0, more careful articulation of vowels and plosive consonants, and other factors. We present a method for simplifying control of style tokens by applying principal components analysis (PCA) to GST weights from a TTS system trained on both plain and clear speech. We identify the axes of variation in PCA space with the acoustic correlates of clear speech in English and show that we can synthesise either style by moving along a single dimension in that space.

Index Terms: controllable speech synthesis, speech style embedding, hyperarticulated speech

1. Introduction

One of the core challenges of acoustic modelling in text-to-speech synthesis (TTS) is the one-to-many relationship between text and speech. That is, for any given text input there are numerous possible speech outputs depending on the communicative context. Prior work on style variation focuses on both 'expressive' TTS, aiming to model complex variation in prosody [1] or synthesising emotional speech [2], and modelling non-emotive styles such as whispered speech, or Lombard speech for application in transit systems or other noisy environments [3]. Lombard speech is one example of hyperarticulated speech [4], where speakers adjust their articulatory movements during speech production with the goal of improving intelligibility for the listener. In the case of Lombard speech, this is motivated by a noisy environment surrounding the speaker, but similar articulatory adjustments have also been found in speech directed toward listeners with impaired hearing [5, 6, 7].

We are interested in synthesising hyperarticulated speech for application in computer-assisted language learning [8], specifically in the context of language revitalisation where limited speech data is available [9]. The possible benefits are twofold: high-quality TTS would allow for arbitrary generation of speech examples for students, while hyperarticulation may facilitate learning pronunciations of new words and unfamiliar speech sounds [10]. We also consider that listeners might want explicit control over the degree of hyperarticulation in generated samples, for example moving from very clear and careful speech in early language learning toward a more fluent and conversational style as they gain familiarity with the language.

Previous studies have found multiple acoustic correlates of hyperarticulated speech in English [5, 6], including:

- Decreased speaking rate, both through additional pause insertion and lengthening individual speech sounds.
- Higher mean F0, with higher maximal values leading to sharper utterance-final declinations.
- More consistent release of plosive bursts in all positions.
- More dispersed vowel spaces as measured by F1 and F2.

This presents a challenge for acoustic modelling of a complex multivariate distribution over these factors. While previous work has used jointly-trained 1-hot style embeddings to model plain, whispered and Lombard speech in a single TTS system [3], others have found Global Style Tokens (GSTs) [1] to work better, for example in emotional speech synthesis [2]. In the GST approach, a set of latent style tokens are learned directly from reference audio during training, and combined through an attention mechanism to provide a style embedding which can be used to condition other parts of the TTS model. This has the benefit of additional flexibility from weighted combination of multiple style tokens, and the ability to discover variation in the training data not captured by discrete style labels.

The authors in [1] present some interpretation of style tokens learned with a simple attention mechanism yielding style embeddings through weighted summation of N GSTs each of D dimensions, and control synthesis by conditioning on a single scaled token at a time. However, they also note improved style transfer performance using multi-head attention, where h attention heads each output an embedding of size D/h from N style tokens, which are then concatenated. This in turn increases the dimensionality of style token weight vectors, which ultimately control the synthesis, from N weights to $N \times h$, complicating analysis, interpretation, and explicit control.

In [2], utterances from a labelled emotional speech corpus are projected into the GST weight space of such a model, and emotion centroids are extracted to provide a set of reference weights for later speech synthesis. While this approach improved upon a baseline system using joint 1-hot emotion embeddings, it does not provide a principled way to navigate the discovered style embedding space. In [11], the authors investigate correlations between the dimensions of a style embedding space and four acoustic measures: F0 mean and standard deviation, spectral tilt and speech rate. They are then able to make adjustments directly in the style embedding space based on regression models for each feature, providing an interpretable method of control in terms of prominent prosodic features. However, this approach relies on a known set of easily-measurable acoustic correlates, which is not the case for some aspects of hyperar-

*Equal contribution

10.21437/Interspeech.2024-2074

ticulated speech such as careful articulation of plosive releases.

To address these challenges, we present a method for low-dimensional control of GST synthesis by projecting multi-head attention token weights using principal components analysis (PCA). For this, we train a FastSpeech 2 [12] system with a GST style embedding layer on the Alba speech corpus [13], which includes both normal speech and speech directed to a hearing-impaired listener, and then fit a PCA transform of the GST weights extracted from utterances in both styles. We find that the resulting PCA space effectively aggregates multiple acoustic factors distinguishing plain and hyperarticulated speech, so that we can synthesise in one style or the other by moving along a single dimension. We verify this controllability through a subjective listening test wherein participants rank samples according to how carefully they perceive the speaker to be talking. Finally, we provide acoustic analyses for two prominent aspects of hyperarticulated speech, namely F0 and duration, comparing the outputs of our TTS system against natural speech.[1]

2. Data and TTS model specifications

The Alba speech corpus comprises around 4 hours of normal read speech (denoted *Plain*) and 20 minutes each of parallel texts read in three additional styles: fast, computer-directed and clear speech [13]. All utterances were read by a Scottish female voice talent with specific instructions for each style. For the clear speech partition (*clear_h* in the original corpus description, henceforth *Clear*), the voice talent was instructed to speak *"as if talking to somebody with a hearing impairment"* [14]; this data collection strategy is consistent with previous phonetic studies on clear speech phenomena [6].

We selected a FastSpeech 2 architecture for its ease of training with limited data [9] and explicit control of F0 and duration through its variance adaptors. We used the ESPnet2 implementation [15], which optionally includes a GST layer with multi-head attention as described in [1]. The resulting style embedding is added to hidden representations from the text encoder before the variance adaptors, so that symbol-level duration and F0 predictions are conditioned on style information from the target audio. To train the duration predictor, we used Kaldi [16] to extract forced alignments for each utterance, using the Edinburgh-accented Unisyn pronunciation lexicon [17].

We used 4,768 utterances for training, including 314 parallel texts read in both *Plain* and *Clear* styles (628 utterances total), and reserved 60 parallel utterances per style (120 total) for evaluation. All audio was downsampled from the original 48 kHz to 22.05 kHz. Our FastSpeech 2 GST system was trained for 200 epochs, taking reference audio matching the target Mel spectrogram to train the GST layer. At synthesis time, we instead pass explicit GST weight values, synthesising without any input audio. We also trained a HiFi-GAN vocoder with the default V1 configuration [18] for 2.5 M steps on the same training utterances, using the implementation from [19] as expected by ESPnet2. To remove over-smoothing artefacts at synthesis time [20], we find it sufficient to add a small amount of Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.01)$ to the predicted Mel spectrograms, rather than fine-tuning the vocoder.

3. PCA analysis of GST weights

Since our aim is to capture the systematic variation between the *Plain* and *Clear* speaking styles, we applied PCA to the GST

[1] Audio samples: `https://edin.ac/clear-speech-gst`

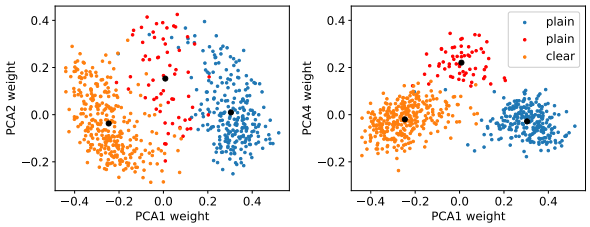

Figure 1: *Initial PCA projections of GST weights for parallel utterances, labelled by style. Black circles mark cluster means.*

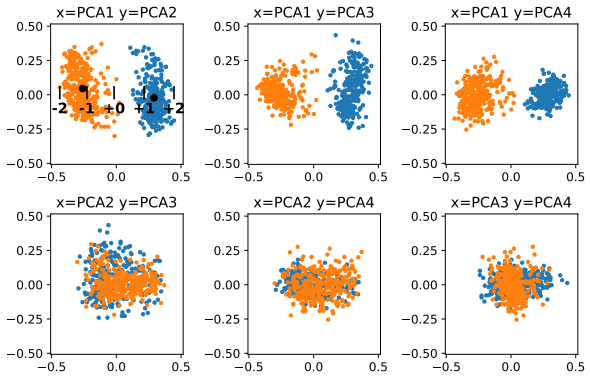

Figure 2: *PCA projections of GST weights for combinations of PCA1–4, after removing the Plain utterances in the red cluster in Figure 1, showing Plain (blue) and Clear (orange) styles. The top left plot shows cluster means (black dots) and control steps along PCA1 (indexed -2, -1, 0, +1, +2) used for synthesis.*

weights of the combined set of parallel *Clear* and *Plain* training utterances. We reasoned that matched text and an equal balance of utterances would best represent the distribution of GST weights across the two styles. The first observations from initial application of PCA are presented in Figure 1. While the *Clear* utterances form a single cluster, we note that the *Plain* utterances are split into two distinct clusters, separated most apparently in the PCA4 direction. This suggests that the Alba speaker used (at least) two distinct styles when reading *Plain* utterances, which we confirmed both by listening to natural speech recordings from each cluster and synthesising utterances using GST weights recovered from their respective mean points (black).

To avoid mixing two different *Plain* styles, we removed the 66 utterances corresponding to the red points and re-ran the PCA. Figure 2 shows the updated result, and Figure 3 shows the variance explained by each principal component. We see that PCA1–4 account for around 80% of the variance in the data, with PCA1 alone covering nearly 50%. This is visible in the top row plots in Figure 2, where the data are now distributed in two distinct clusters along PCA1, whereas there is significantly more overlap for all dimensions beyond PCA1. This all indicates that PCA1 largely divides the two styles, so we chose this as our control dimension. To evaluate this, we chose 7 points along this axis, 5 of which are shown in the top left plot of Figure 2: *0* is set to the global mean point, ± 2 are the max and min observed PCA1 values respectively and ± 1 are the midpoints between those. In addition, to explore potential for style extrapolation, we chose two further points beyond the observed range of GST values, ± 3, which are half as far again from the *0* point as ± 2. We use these values for PCA1 and set all other

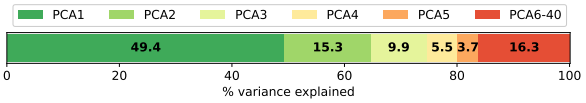

Figure 3: *Percentage variance explained, for PCA in Figure 2*

PCs to their global mean values, then apply the inverse PCA transformation to retrieve new GST weight values for synthesis. In this way, we have effectively reduced 40 GST dimensions (4 heads $\times$ 10 tokens) to a single one, and could extend to 4 control parameters covering 80% of the variation in our data.

4. Subjective evaluation

We evaluated the overall audio quality of our trained system and the extent to which we can control how clearly the model speaks through a subjective listening test. We recruited 30 native English speakers from the UK to take part in the two-part test. The test took around 15 minutes for participants to complete on average, and listeners were paid £3.50 for their responses.

4.1. Audio quality

In the audio quality test, we compared our GST model against a baseline FastSpeech 2 model using jointly-trained 1-hot style embeddings corresponding to the *Plain* and *Clear* class labels in our subset of the Alba corpus, trained for the same number of epochs as the GST system. This model should be less able to model variation in the data than a GST system, reducing variation within each labelled class to the mean. We also presented four GST voices, using the mean PCA1 values for *Plain* and *Clear* data and the out-of-distribution points at the ± 3 points on that axis. Finally, we included natural recordings from both *Plain* and *Clear* utterances. We split participants across two blocks of 32 questions each, with each model synthesising 4 randomly-selected utterances which do not overlap between blocks (so each system is evaluated on 8 utterances overall), for a total of 120 ratings per system. Participants were instructed to rate each on a scale from 1–5, and to *"Ignore how the person is talking, for example how fast or slow, or how carefully, and only focus on the recording quality when making your decision."*

The results from this part of the test are shown in Figure 4. Between *Plain* and *Clear* groups, we can see that *Plain* utterances tend to score slightly higher than *Clear* ones. *Clear* natural speech samples are rated significantly higher than all *Clear* synthesised utterances (according to a Mann-Whitney U test with Bonferroni correction for repeated pairwise comparisons, at the 95% confidence level), whereas for *Plain* TTS, only the *GST +3* voice is significantly worse than natural speech. More specifically, the *Clear GST -3* voice appears to rate lowest overall, and is rated significantly worse than *Plain 1-hot* and *GST mean*, although not worse than the other *Clear* TTS samples. *Plain GST +3* also scores significantly worse than *Plain GST mean*. This further indicates that extending beyond the range of GST weight values seen during training can introduce artefacts in the generated speech which reduce overall perceived quality.

Finally, although the *Plain 1-hot* voice appears to have a longer tail of low ratings compared to other *Plain* TTS systems, they are not significantly different from each other. This runs counter to our expectation that the *1-hot* model should be less able to model variation in the data, for example the two different (unlabelled) styles of *Plain* speech noted in Section 3, which will both contribute to a single average *Plain* representation in this model, landing at a point somewhere between the

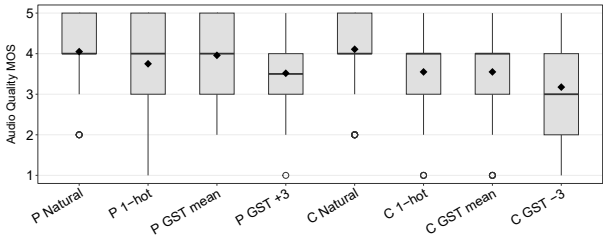

Figure 4: *Box plot of audio quality MOS results, grouped into* Plain *(P) and* Clear *(C) utterances. Solid bars indicate median quality ratings, diamonds indicate mean scores, circles outliers.*

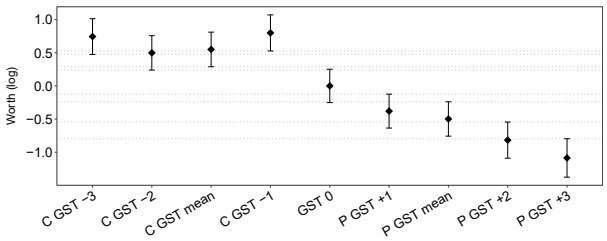

Figure 5: *Clearness AB results for utterances synthesised at different points along the PCA1 axis. Worth values and 95% confidence intervals from a Bradley-Terry model of system rankings.*

two groups. It seems that this intermediate representation is nonetheless able to produce acceptable audio quality, even if the style might be somewhat inconsistent.

4.2. Style controllability

To assess the controllability of *Clear* and *Plain* speech styles, we synthesised test utterances using GST weights at 9 different points along the PCA1 axis: the 7 points in the interval $[-3, 3]$ as defined in Section 3, plus the mean point for each style. We then asked listeners to compare pairs of utterances at different points, and to select which one *"sounds most like the person is trying to speak clearly or carefully,"* focusing on articulation and ignoring raw audio quality. There are 36 combinations of these 9 points along the PCA1 axis, with each voice appearing in 8 pairs, synthesising a different utterance each time. Listeners heard one example of every voice match-up, generating 240 comparisons per voice.

The results for this AB clearness test are shown in Figure 5. Following [21], we fitted a Bradley-Terry model [22] to paired comparison responses (using the `PlackettLuce` *R* package [23]). This model assigns worth values to each voice based on how often it is preferred over others, which we can use to compare any pair of systems and determine if there is a significant difference along the axis of interest. Given our question formulation, asking listeners to select the most clearly-spoken of two systems, we set the origin point along PCA1 *GST 0* as a reference point (worth $:= 0$), and expect *Clear* and *Plain* voices to have higher and lower worth values respectively.

As expected, we see that synthesising at points further toward the *Plain* end of PCA1 produces samples which are consistently judged less clearly-spoken. Considering the 95% confidence intervals around calculated worth values, *Plain GST mean* is rated significantly less clear than the global mean *GST 0*, and the extreme *GST +3* is significantly less clear than both; there is no significant difference between the intermediate points, however. For *Clear* speech, we again see some appar-

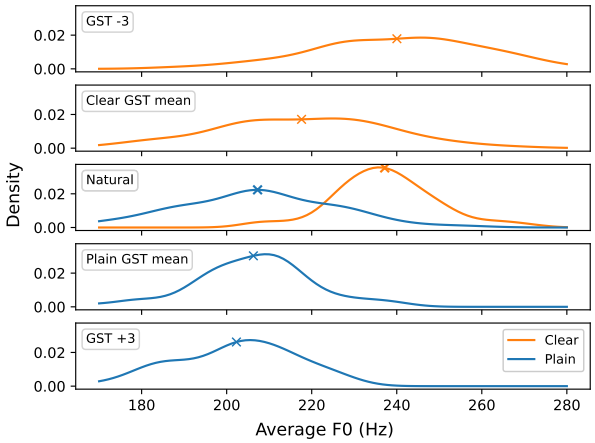

Figure 6: *Density estimation of average F0 values for test utterances in Plain and Clear styles. Crosses indicate overall mean.*

ent difficulty in fine-grained synthesis control: while samples from all *Clear* points along PCA1 are rated significantly more clearly-spoken than *GST 0*, there are no significant differences between them internally. We attribute this to a possible misinterpretation of our task instructions, where participants may have conflated two senses of the word "clear": the *Clear* speech style we intended them to judge vs. how easy it was to understand the samples through any audio quality artefacts. In that case, as more extreme GST values reduce overall audio quality, even as their articulations may become more *Clear*, we would expect a balancing effect on final worth values as seen here.

5. Acoustic correlates of clear speech

In addition to subjective evaluation, we can measure the extent to which our synthesised samples match the expected acoustic characteristics of the target speech styles.

5.1. F0

In the natural speech recordings from the Alba corpus, we observe a general tendency toward higher mean F0 in *Clear* speech compared to *Plain*. This is evident in the centre plot of Figure 6, which shows the distribution of average F0 values calculated per utterance in the test set (extracted using the PYIN algorithm [24]). While the *Plain GST mean* achieves a mean F0 value close to natural utterances, the overall distribution of average F0 values per utterance is more closely matched at the *GST +3* point. For *Clear* speech, we see that the *Clear GST mean* weights have a lower average F0 than natural utterances, and that the distribution of F0 values is much wider, largely overlapping with natural *Plain* utterances. At the *Clear GST -3* point, the average F0 across utterances is very close to the natural mean, but the distribution of values remains wider, and extends beyond the upper limits in the reference recordings. While we indeed found *Clear* TTS samples to sound noticeably higher pitched than *Plain*, these measurements suggest a lack of adequate control and realisation of natural F0 variation between the two styles. We speculate that this may also contribute to the lower audio quality ratings for *Clear* TTS in our listening test.

5.2. Duration

We measure phone durations based on forced alignments for 30 natural speech test utterances each from *Plain* and *Clear* styles, and from TTS-predicted durations for the same utterances using *GST mean* weights, all manually-corrected (this effort prevents us from comparing more synthesised samples, e.g. from *GST* ± 3). *Clear* speech tends to be elongated relative to *Plain* in two ways: by increasing the number and duration of pauses between words, and by increasing the duration of individual phone articulations. Across the 30 parallel test utterances, there are 17 short pauses in the *Plain* style with average duration $140.8\,\mathrm{ms}$, and 66 in *Clear* averaging $268.4\,\mathrm{ms}$. Without a principled way to predict the placement of short pauses for TTS, however, we synthesised all listening test stimuli without including explicit pauses between words. Table 1 shows the average durations of broad phone classes in the two speech styles for both natural speech and synthesised samples, and the relative increase in duration in *Clear* over *Plain* speech. We see that while our TTS system closely matches the duration of nearly all phone classes in *Plain* speech, it tends not to elongate them to the full extent when synthesising with the *Clear GST mean* weights, although the increased durations are still noticeable.

Table 1: *Average duration (ms) per phone class in Plain and Clear styles for natural speech and TTS samples synthesised using mean GST style embeddings. C/P denotes relative duration in Clear style compared to Plain.*

	Natural			Synthesised		
Phone class	Plain	Clear	C/P	Plain	Clear	C/P
Vowel	79.7	115.2	1.45	82.3	99.9	1.21
Plosive	94.7	140.5	1.48	96.6	120.2	1.24
Fricative	91.7	124.9	1.36	85.0	103.5	1.22
Nasal	81.8	110.0	1.35	76.5	94.9	1.24
Approximant	73.2	100.3	1.37	60.1	85.8	1.43
Affricate	106.7	117.9	1.11	127.4	123.0	0.97

6. Conclusion

In this paper, we presented an approach for controlling synthesis of plain and hyperarticulated (or 'clear') speech styles by a TTS system equipped with global style tokens. Rather than synthesising from reference audio or directly manipulating the 40-dimensional GST weights used by our system, we used PCA to project the learned GST weights into a lower-dimensional representation capturing most of the acoustic variation which characterises the two styles. In this way, we were able to synthesise recognisably hyperarticulated and plain speech by moving along a single dimension in the PCA space to generate GST weights, according to a subjective listening test. However, our attempts to control the degree of hyperarticulation showed limited success: moving beyond the range of GST values observed during training introduces audible acoustic artefacts, which we believe impacts listener judgements and limits our ability to evaluate the controllability of our system.

We also measured some of the acoustic correlates of hyperarticulated speech in English as they occur in natural utterances and our TTS samples, namely raised F0 and longer phone durations. Again, we found that our model approaches but does not match the full extent of variation in clear speech along these dimensions. Motivated by potential applications in computer-assisted language learning, especially in the context of language revitalisation, further work should include language-specific analyses of how hyperarticulated speech is expressed, and whether the style is adequately modelled using standard TTS architectures.

7. Acknowledgements

We would like to thank Professor Keiichi Tokuda and Associate Professor Yoshihiko Nankaku of Nagoya Institute of Technology, for their participation in discussions of this research. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1), the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences, and the National Research Council of Canada's Ideation Fund: 'Small Teams – Big Ideas'.

8. References

[1] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[2] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "An Effective Style Token Weight Control Technique for End-to-End Emotional Speech Synthesis," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1383–1387, 2019.

[3] Q. Hu, T. Bleisch, P. Petkov, T. Raitio, E. Marchi, and V. Lakshminarasimhan, "Whispered and Lombard Neural Speech Synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 454–461.

[4] J. Šimko, Beňuš, and M. Vainio, "Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 151–162, 2016.

[5] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking Clearly for the Hard of Hearing II," *Journal of Speech, Language, and Hearing Research*, vol. 29, no. 4, pp. 434–446, 1986.

[6] R. Smiljanić and A. R. Bradlow, "Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes," *Language and Linguistics Compass*, vol. 3, no. 1, pp. 236–264, 2009.

[7] V. Hazan and R. Baker, "Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?" in *Proc. DiSS-LPSS Joint Workshop (DiSS 2010)*, 2010, pp. 7–10.

[8] Z. Handley and M.-J. Hamel, "Establishing a Methodology for Benchmarking Speech Synthesis for Computer-Assisted Language Learning (CALL)," *Language Learning & Technology*, vol. 9, no. 3, pp. 99–120, 2005.

[9] A. Pine, D. Wells, N. Brinklow, P. Littell, and K. Richmond, "Requirements and motivations of low-resource speech synthesis for language revitalization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 7346–7359.

[10] G. Piazza, M. Kalashnikova, and C. D. Martin, "Phonetic accommodation in non-native directed speech supports L2 word learning and pronunciation," *Scientific Reports*, vol. 13, no. 1, p. 21282, 2023.

[11] J. Šimko, T. Törö, M. Vainio, and A. Suni, "Prosody Under Control: Controlling Prosody in Text-to-Speech Synthesis by Adjustments in Latent Reference Space," in *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023.

[12] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[13] C. Valentini-Botinhao and J. Yamagishi, "Alba speech corpus," 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3270

[14] C. Valentini-Botinhao, M. Wester, J. Yamagishi, M. Toman, M. Pucher, and D. Schabus, "Non linear time compression of clear and normal speech at high rates," 2019, arXiv:1901.07239 [eess].

[15] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the Edge of TTS Research," 2021, arXiv:2110.07840 [cs, eess].

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[17] S. Fitt, "Unisyn Lexicon (v1.3)," 2008. [Online]. Available: https://www.cstr.ed.ac.uk/projects/unisyn/

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

[19] T. Hayashi, "kan-bayashi/ParallelWaveGAN," original-date: 2019-10-29. [Online]. Available: https://github.com/kan-bayashi/ParallelWaveGAN

[20] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting Over-Smoothness in Text to Speech," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 8197–8213.

[21] S. Shirali-Shahreza and G. Penn, "MOS Naturalness and the Quest for Human-Like Speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 346–352.

[22] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[23] H. L. Turner, J. van Etten, D. Firth, and I. Kosmidis, "Modelling rankings in R: the PlackettLuce package," *Computational Statistics*, vol. 35, no. 3, pp. 1027–1057, 2020.

[24] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.