INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Controlling formant frequencies with neural text-to-speech for the manipulation of perceived speaker age

Ziya Khan[1†], *Lovisa Wihlborg*[2], *Cassia Valentini-Botinhao*[2,3], *Oliver Watts*[2]

[1]Xperi Inc.
[2]SpeakUnique Ltd.
[3]University of Edinburgh

ziya.khan@xperi.com, lovisa@speakunique.co.uk, cvbotinh@ed.ac.uk, oliver@speakunique.co.uk

Abstract

In this paper, we present a framework for formant-controllable neural text-to-speech. We train a model that predicts formant frequencies which then condition melspectrogram generation. We apply this to manipulate perceived speaker age in an indirect fashion, by modifying the predicted formants in a manner that affects perceived vocal tract length. Our ultimate goal is to allow for the control of perceived ageing in children's text-to-speech voices, since ageing in natural child speech is strongly linked to the growth of a child's vocal tract. However, our experiments indicate that our method shows strong age control capabilities for adult speech as well.

1. Introduction

Speech synthesis has recently seen large advances in quality, due to the leveraging of new deep learning-based approaches and large amounts of high-quality data (see e.g. [1] for a review). However, in various situations there are only limited amounts of data available, but a high-quality text-to-speech (TTS) voice would be beneficial.

This is especially the case for users of Alternative Augmentative Communication (AAC). For this population, it has been found that a personalised synthetic voice is more desirable, both for adults [2] and for children [3]. Furthermore, AAC is reported to benefit children for a range of reasons – including language development, behaviour, and communication skills – and the earlier AAC is introduced, the better [4]. However, it is unlikely that an AAC user will be able record as much audio as is often used to train single-speaker state-of-the-art models (e.g. 24.6 hours in [5] or 34.8 hours in [6]). Additionally, a child using AAC might require a voice that "ages with them", as children's voices undergo rapid changes in short time periods (as opposed to adult voices), and the perceived age of a bespoke synthetic voice created in early childhood would not match the child's age only a few years later. To handle this would require multiple recordings at different stages of their youth.

Additionally, as summarised in [7], working with children's data is challenging for a number of reasons: firstly, the speech of children is acoustically variable, disfluent, and likely to contain articulatory errors; secondly, recording is more difficult as children often have short attention spans, are less fluent readers, and might require a more familiar but less optimal recording environment. These challenges are present not only for TTS, but also for automatic speech recognition (ASR).

A common way of handling the issues is to use adult speech in addition to child speech. For ASR, such methods include transfer learning (as in [8]), data augmentation (of synthetic speech as in [9] or large amounts of adult speech as in [10]), or speaker adaptive acoustic modelling (as in [11]). Similarly in TTS, speaker adaptation, where a model trained on adult speech is later adapted to a smaller corpus of child speech, is a widely applied method (e.g., [12], [13], and [14]).

For the present work, we focus on the issue of modifying the perceived age of a synthetic voice, as a way of bypassing the otherwise regular requirement of data collection in order to create a personal synthetic voice for a child/young person.

To achieve this, we look at manipulating formants (the resonances of the vocal tract; that is, high-energy peaks at the resonant frequencies).The lowest formant (or resonance) is called $\mathcal{F}_1$, the second lowest $\mathcal{F}_2$, and so on. The first two formants are usually sufficient to describe a vowel [15]. Formant manipulation was possible in early speech synthesis systems such as the Klatt synthesiser [16], which was based around explicitly modelled formant frequencies, and where those frequencies could be input to control the output synthesis of specific phonemes. More recent work on formant manipulation is [17], where a DNN-based model was trained with the goal of retaining controllability (which often is lost in DNN-based systems). The authors found that their system was able to be successfully manipulated by a number of interpretable, phonological parameters, while maintaining the high overall quality gained from their DNN-based approach.

For the present purposes, it should be noted that the formants are closely associated with the shape of the vocal tract; variations in the vocal tract length (VTL) causes the formant frequencies to shift in an approximately linear fashion [18]. In this way, the formant frequencies are directly correlated with the size of a body.

2. Proposed approach

For the experiments presented here we adopt and modify the FastPitch acoustic model [19]. As described there the model predicts one pitch value for each input token (i.e. phone or letter) and conditions its predictions of mel-spectrograms on those pitch values. This results in a model where it is possible to control the pitch either globally or at the subword level. The implementation we adopt conditions mel-spectrograms also on energy, which is predicted in a similar way. We extend the model to include formant values as control features analogous to pitch and energy in the original model, and experiment with global modifications of the control values in order to manipulate the perceived age of the synthetic speaker.

[†] The first author did this work while a student at the University of Edinburgh, under the supervision of SpeakUnique.

10.21437/Interspeech.2023-820

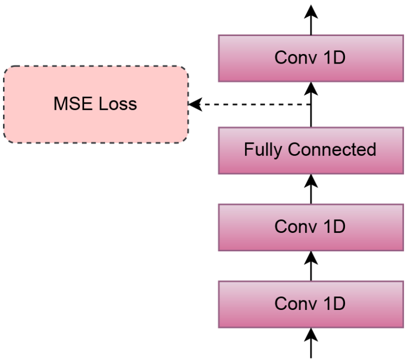

Figure 1: *The architecture of the formant predictor module.*

2.1. Base architecture

A forward pass through the model is as follows: the input text is first passed through an embedding layer which transforms it into a 512-dimensional vector of the same length as the input and then through a cascade of 6 Feed Forward Transformer (FFT) blocks [20]. This representation is then fed to sub-modules to predict pitch and energy projections and duration of our input text representation respectively. These predictions are then summed to give the encoder representation. This representation is then repeated a number of times based on our duration predictor outputs and is then fed to the decoder consisting of 6 FFT blocks and a final projection layer which yields the final mel-spectrogram as output. The encoded representation is dependent on both the pitch and energy projections. This lends some degree of control over the final decoder outputs. A scaling of or an addition to the predicted value of pitch projection leads to a clearly perceptible difference in the pitch of the generated output.

2.2. Postnet

One modification we made to the base architecture was to use a *postnet* based on that used in [5] to refine the output mel-spectrograms. It uses five 1D convolutional layers (each with 512 channels, filter size of 5, and tanh activation function) to postprocess audio output; the postnet's output is then added to the mel-spectrogram as a residual. Both informal listening and objective evaluation using MOSNet [21, 22] (pretrained on the VCC-2018 [23] dataset) showed a slight but not statistically significant improvement when comparing speech produced by the same base architeture with and without the postnet. The postnet is used in the system evaluated in this paper.

2.3. Formant Control

We present here our additions to the FastPitch model to enable explicit modelling of the first four formants[1].

2.3.1. Formant Predictor

The formant predictor module models the first 4 formants of the audio signal. Its architecture is based on the architecture of the pitch predictor and is shown in Figure 1. As in the pitch predictor, it consists of 2 1D convolutional layers with a kernel size of 3, 256 channels and a dropout of 0.1. Next, the output passes through a fully connected/dense layer that instead of predicting one number for each temporal location, as in the pitch predictor, predicts 4 numbers, corresponding to each formant: $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ and $\mathcal{F}_4$. The predicted values of $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ and $\mathcal{F}_4$ are trained using the mean squared error loss function with the targets as the true values of the first 4 formants calculated directly from the audio. The output of the dense layer is then passed into another 1D convolutional layer to project it into the same high-dimensional space as the encoder output and then summed with other outputs as shown in Figure 2.

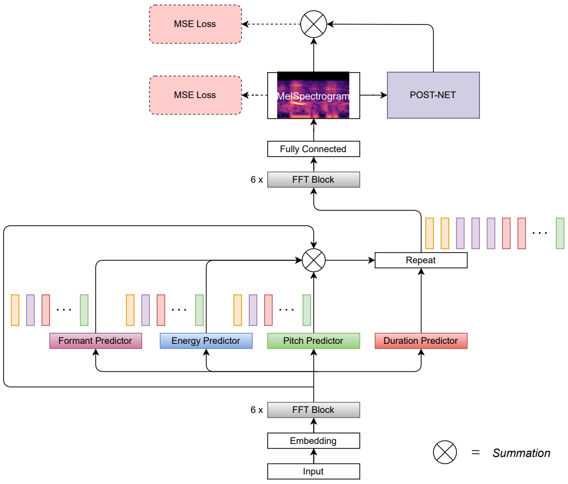

Figure 2: *Modified architecture of FastPitch. We add a postnet and formant predictor module as shown.*

2.3.2. Extraction of Target Formants

To extract formants from training data that can be used as targets when optimising the formant predictor we used Praat [24]. The audio is re-sampled to twice the formant ceiling (set to 5500 Hz based on Praat's recommendation for female speakers). Preprocessing in the form of pre-emphasis and windowing with a Hamming window is carried out on each frame. We then extract the first 4 formants for each time window that best approximate its spectral envelope.

Window and hop lengths were set to ∼46.4 ms and ∼11.6 ms (1024 and 256 samples at 22050 Hz sampling rate) so that the formant framerate matched that of the mel-spectrograms used. Similar to how pitch is treated, the outputs of the duration predictor are used to average the target formant values over the frames of each character in the input, such that the target formant values used in training are at the appropriate rate (that of the input phones/characters). In segments determined to be unvoiced (based on the corresponding pitch target matrix), both pitch and formant values were set to zero, meaning that unvoiced segments are handled by their own row of the pitch and formant embedding matrices. Similar to pitch, energy, and duration, it is important to standardise formant values using statistics determined on all the training data. These were computed independently for each of the 4 formants. The addition of the formant predictor results in an additional loss term when training the model.

[1] Samples and code available at: `https://github.com/ziafkhan/FastPitch`

2.3.3. *Inference: formant manipulation*

The approach we implemented allows phone-level control of individual formants; however, here we only experiment with global modifications to formant values, with the goal of altering the perceived age of the synthetic voice.

After we generate the formant values from the formant predictor, we can modify those values by either scaling them with constants, or by adding or removing a constant value. As there are too many degrees of freedom available for the modification of formants (all formants could potentially be modified independently of one another), we add two constraints to the possible modifications:

- all formants can be scaled by any constant value between 0.8 and 1.2. These values were chosen based on empirical evidence, as values lower or higher resulted in unintelligible audio.
- individual formants cannot be scaled independently; that is, if $\mathcal{F}_1$ is scaled by a constant factor k, all other formants will be scaled by the same value.

The authors in [25] estimate the formant frequency by modelling the vocal tract as a tube closed at one end. It is open at the lips and closed at the vocal folds. The resulting formula for the n^{th} formant frequency ($\mathcal{F}_n$) is given by:

$$\mathcal{F}_n = \frac{(2n-1)c}{4L} \tag{1}$$

where c is the speed of sound, L is the VTL. Using this equation, given that we have a mean formant frequency, we can estimate the VTL of a body. To get a more accurate estimate of VTL, we can use multiple formant means and averaging the resultant VTLs.

A change in the VTL can thus be implemented by modifying the individual formant frequencies. For example, to scale the L by a fixed constant k, the modified equation would be:

$$k * L = \frac{c}{16}\left(\frac{1}{\mathcal{F}_1/k} + \frac{3}{\mathcal{F}_2/k} + \frac{5}{\mathcal{F}_3/k} + \frac{7}{\mathcal{F}_4/k}\right) \tag{2}$$

Apart from the scaling shown above, it is clear that a constant reduction in the formant frequencies $\mathcal{F}_1$ through $\mathcal{F}_4$ would also result in a larger estimate VTL and vice versa. Specifically, multiplying $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ and $\mathcal{F}_4$ by a factor of k will result in a VTL estimate scaled by a factor of $\frac{1}{k}$.

3. Experiments

3.1. Data

In initial experiments we tried training our system on a large single-speaker database. This consists of 13,100 recordings ranging from 1 to 10 seconds (nearly 24 hours in total) from a single female speaker [26]. These initial experiments showed that the within-dataset formant variation is insufficient to learn a formant-controllable system as the resulting model allowed for very little global formants modification before the output audio became unintelligible. [17] addressed similar issues by using data augmentation; in the current work we took an alternative approach and supplemented the single speaker data with the VCTK dataset [27], a multi-speaker dataset consisting of 110 English speakers with different accents. For each speaker there are approximately 400 sentences. We removed 3 speakers: p280, p315 and s5. This meant we now had a total of 107 distinct speakers in our dataset containing $\sim$41 hours of total spoken data spread across 43,752 files.

As our ultimate goal is to vary the age of child voices, we also experimented with adding child speech to our training database. However, due to the inherent difficulties in handling child speech already mentioned in Section 1 we were not able to successfully make use of [28] or [29]. We had more success when adding the dataset described in [30] – consisting of 1.5 hours of speech recorded by a 7 year old girl – to our training set. More extensive exploration of the impact of using larger quantities of child speech with the model we describe is left for future work.

All audio was down-sampled to 22,050 Hz for the work described here. To maintain a constant loudness throughout and across datasets, loudness normalization to -23 LUFS was carried out over all data, using the pyloudnorm [31] library based on the EBU-R128 standard [32].

3.2. Model training

Inputs to the model were produced by processing the text transcripts associated with the training data as follows. For words present in the CMUDict pronunciation dictionary [33], the corresponding phonetic sequence was used as input; for missing words, we used characters as the input instead. Mel-spectrograms were extracted from the audio part of the data to serve as target outputs for the model during training.

We trained a model of the type described in Section 2 for 400 epochs, with an effective batch size of 128, split across 4 NVIDIA Titan X GPUs. The learning rate scheduling was set to increase during the first 1000 ”warmup” steps linearly, and then to start decreasing exponentially. The vocoder in all experiments described in this paper was a pre-trained WaveGlow model [34]. It is a universal vocoder capable of inverting mel-spectrograms from multiple speakers into audio samples.

3.3. Experiment 1: Ageing effect of VTL adjustment

In this experiment, we sought to establish whether VTL modification can lead to the perception of modified speaker age in a pair-wise comparison.

16 different texts were synthesized from the trained model at three different VTL multipliers (0.95, 1.00, and 1.05), creating a total of 48 audio samples. Note that a VTL multiplier of 0.95 implies that the formants were multiplied by the number $\frac{1}{0.95} = 1.05$. Everything apart from the VTL multiplier was held constant across the audio samples of a single text. In each case, the speaker chosen for the audio was the speaker from the LJSpeech dataset. This was because it was the largest single speaker dataset we used, and also the speaker for whom the pre-trained vocoder gave the best results.

A listening test was implemented (using Qualtrics XM) where 38 participants (sourced from Prolific Academic) were asked to say in which of two audio samples the speaker sounded older. The total number of participants were 38 (a minimum number of 30 participants was determined from [35]). Each participant was paid £4.75 for their time spent in completing the survey.

Three participants were excluded due to having completed the survey in a too short time, and also on the basis of MUSHRA-specific criteria in the subsequent experiment (see Section 3.4). The remaining 35 participants' judgements of 10 different audio pairs were retained. Each pair consisted of the same text being spoken using two different VTL multipliers.

In 82% of trials (287 out of 350) listeners thought the speech generated with the higher VTL multiplier sounded like the older speaker. The effect was significant under a two-tailed Binomial test using a significance level of 0.05.

3.4. Experiment 2: Continuous control of age

In this experiment we wished to determine whether listeners' ratings of synthetic speaker age on a continuous scale can be controlled by adjusting VTL and mean F0 in combination. Table 1 shows the 5 different VTL multipliers used in this experiment and the corresponding pitch multipliers used (as well as formant multipliers implied by VTL adjustment).

VTL	Pitch	Formant
0.90	1.11	1.11
0.95	1.05	1.05
1.00	1.00	1.00
1.05	0.95	0.95
1.10	0.91	0.91

Table 1: *VTL multipliers used and corresponding multipliers for pitch and formants.*

As in Experiment 1, the LJSpeech speaker was used for synthesis in all cases. Participants were asked to complete a MUSHRA style test, where we gave a reference audio and told the participants that the age of the speaker in the reference was 30. This reference had a VTL multiplier of 1.0. The participants were then asked to estimate the ages of the speakers in the 5 samples presented to them. There were a total of 8 such groups of 5 audio samples presented to every participant. The participants were told that one of the samples was the same as the reference and needed to be given a perceived age of 30. The participants were not allowed to move further in the questionnaire until at least one of the 5 audios was marked with an age of 30. While many participants did not find the correct reference a few times, 3 of the participants never correctly marked the reference, and their responses were not included in the analysis. The distribution of responses received is shown as a box-plot in Figure 3. The strong correlation that can be seen between control values and perceived age in Figure 3 indicates that perceived age can be controlled in a continuous fashion using global formant and pitch modifications.

3.5. Discussion

Informal listening shows that increasing VTL gives the impression that the speech originates from a larger-bodied person: this holds given the correlation between VTL and speaker size [36]. The results of both experiments presented here suggest that listeners correlate speaker size with age even for adult voices where speakers would be assumed to be physically fully grown. This is perhaps partially an artefact of how we designed our experiment, and future work should include evaluation of natural samples against synthetic ones, for both child and adult voices. Regardless, we would assume that the previously mentioned correlations hold even more strongly and with more justification for children's voices, as a portion of the ageing of a child's voice can certainly be attributed to the growth of the child's vocal tract.

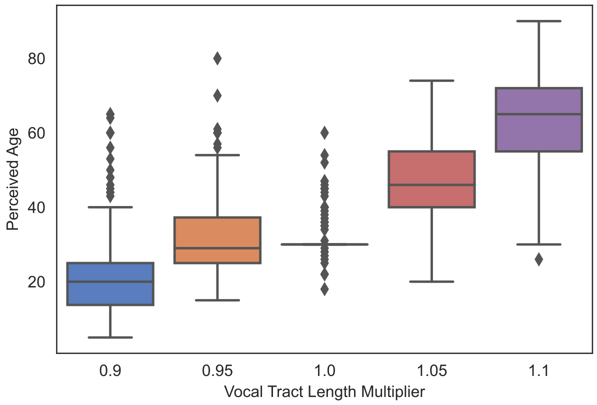

Figure 3: *Response distribution on the MUSHRA test. Boxes show median values and interquartile ranges (IQR); whiskers extend an extra 1.5 IQR; dots show outliers.*

4. Conclusions

This paper focused on establishing a process to modify the perceived age of synthetic speech. This was achieved using explicit modelling of formants $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$, $\mathcal{F}_4$, whose frequency can thus be modified at inference time. We showed that this effectively amounts to adjusting the VTL estimate and thus adjusting the perceived size of the body of the speaker. The adjustment in body size would then lead to an alteration of perceived age of the synthetic voice.

We successfully trained a model to realistically modify speaker age, and realised that one requirement of such a system is to have available a variety of formant values from which to learn different effects of formant predictions. Hence, we used a large dataset constructed from multiple smaller datasets with different speakers.

In Experiment 1, we found that the listeners showed a strong perception of a difference in age after the modification of predicted formant values, despite the pitch values being unchanged. This implies that the formant values, even when disentangled from pitch, can cause a perception of ageing. Finally, in Experiment 2, we saw a strong relation between perceived age and a scaling of pitch and formant modification, with the perceived age varying consistently with the pitch and formant modification. This leads to our conclusion that our model is an effective and controllable way to induce an ageing effect in synthetic speech.

5. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, "Building personalised synthetic voices for individuals with severe speech impairment," *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.

[3] M. E. Begnum, D. Meen, and T. Nordgård, "A strategy for producing high-quality Norwegian synthetic child voices."

[4] K. Drager, J. Light, and D. McNaughton, "Effects of AAC interventions on communication and language for young children with complex communication needs," *Journal of pediatric rehabilitation medicine*, vol. 3, no. 4, pp. 303–310, 2010.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous,

Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions,” in *Proc. ICASSP*, 2018, pp. 4779–4783.

[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[7] C. Terblanche, M. Harty, M. Pascoe, and B. V. Tucker, “A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative evidence,” *Applied Sciences*, vol. 12, no. 5623, 2022.

[8] P. G. Shivakumar and P. Georgiou, “Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations,” *Computer speech & language*, vol. 63, p. 101077, 2020.

[9] T. Hasija, V. Kadyan, and K. Guleria, “Out domain data augmentation on Punjabi children speech recognition using Tacotron,” in *Journal of Physics: Conference Series*, vol. 1950, no. 1, 2021, p. 012044.

[10] Y. Q. X. W. K. Evanini and D. Suendermann-Oeft, “Improving DNN-based automatic recognition of non-native children’s speech with adult speech,” in *5th Workshop on Child Computer Interaction*, 2016, pp. 40–44.

[11] M. Gerosa, D. Giuliani, and F. Brugnara, “Towards age-independent acoustic modeling,” *Speech Communication*, vol. 51, no. 6, pp. 499–509, 2009.

[12] O. Watts, J. Yamagishi, S. King, and K. Berkling, “Synthesis of child speech with HMM adaptation and voice conversion,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1005–1016, 2009.

[13] R. Karhila, D. R. Sanand, M. Kurimo, and P. Smit, “Creating synthetic voices for children by adapting adult average voice using stacked transformations and VTLN,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4501–4504.

[14] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, “A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis,” *IEEE Access*, vol. 10, pp. 47 628–47 642, 2022.

[15] K. Paliwal, W. Ainsworth, and D. Lindsay, “A study of two-formant models for vowel identification,” *Speech Communication*, vol. 2, no. 4, pp. 295–303, 1983.

[16] D. H. Klatt, “Software for a cascade/parallel formant synthesizer,” *the Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[17] G. T. D. Beck, U. Wennberg, Z. Malisz, and G. E. Henter, “Wavebender GAN: An architecture for phonetically meaningful speech manipulation,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6187–6191.

[18] M. Gales and S. Young, *Application of Hidden Markov Models in Speech Recognition*. Foundations and Trends in Signal Processing, 2008, vol. 1, no. 3, pp. 195–304.

[19] A. Łańcucki, “Fastpitch: Parallel text-to-speech with pitch prediction,” in *Proc. ICASSP*, 2021, pp. 6588–6592.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. NIPS*, vol. 30, 2017, p. 5998–6008.

[21] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “MOSNet: Deep learning-based objective assessment for voice conversion,” in *Proc. Interspeech*, 2019, pp. 1541–1545.

[22] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Y. Lin, and H.-y. Lee, “Utilizing self-supervised representations for MOS prediction,” in *Proc. Interspeech*, 2021, pp. 2781–2785.

[23] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods,” in *Proc. Odyssey*, 2018, pp. 195–202.

[24] P. Boersma and V. Van Heuven, “Speak and unSpeak with PRAAT,” *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[25] M. D. Groll, V. S. McKenna, S. Hablani, and C. E. Stepp, “Formant-estimated vocal tract length and extrinsic laryngeal muscle activation during modulation of vocal effort in healthy speakers,” *J. Speech Lang. Hear. Res.*, vol. 63, no. 5, pp. 1395–1403, 2020.

[26] K. Ito and L. Johnson, “The LJ Speech Dataset,” 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[27] C. Veaux, J. Yamagishi, and K. MacDonald, “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2019. [Online]. Available: https://doi.org/10.7488/ds/1994

[28] A. Batliner, M. Blomberg, S. D’Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, “The PF-STAR Children’s Speech Corpus,” in *Proceedings of the 9th European Conference on Speech Communication and Technology*, Bonn, 2005, pp. 2761–2764.

[29] K. Shobaki, J. Hosom, and R. A. Cole, “The OGI kids^2 speech corpus and recognizers,” in *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, 2000, pp. 258–261.

[30] O. Watts, J. Yamagishi, S. King, and K. Berkling, “Synthesis of child speech with hmm adaptation and voice conversion,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 1005–1016, Jul. 2010.

[31] C. J. Steinmetz and J. D. Reiss, “pyloudnorm: A simple yet flexible loudness meter in python,” in *150th AES Convention*, 2021.

[32] EBU, *R128 Loudness normalisation and permitted maximum level os signals*, 2020, [accessed 12-April-2022]. [Online]. Available: https://tech.ebu.ch/docs/r/r128.pdf

[33] CMU, 2022. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[34] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” 2018. [Online]. Available: https://arxiv.org/abs/1811.00002

[35] M. Wester, C. Valentini-Botinhao, and G. Henter, “Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations,” in *Proc. Interspeech*, 2015, pp. 3476–3480.

[36] S. Dusan, “Estimation of speaker’s height and vocal tract length from speech signal,” in *Proc. Interspeech*, 2005, pp. 1989–1992.