An inclusive approach to creating a palette of synthetic voices for gender diversity

Éva Székely[1], *Maxwell Hope*[2]

[1]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
[2]Department of Linguistics Cognitive Science, University of Delaware, USA

szekely@kth.se, maxhope@udel.edu

Abstract

Mainstream text-to-speech (TTS) technologies predominantly rely on binary, cisgender speech, failing to adequately represent the diversity of gender expansive (e.g., transgender and/or nonbinary) people. This poses challenges, particularly for users of Speech Generating Devices (SGDs) seeking TTS voices that authentically reflect their identity and desired expressive nuances. This paper introduces a novel approach for constructing a palette of controllable gender-expansive TTS voices using recordings from 14 gender-expansive speakers. We employ Constrained PCA to extract gender-independent speaker identity vectors from x-vectors, using acoustic Vocal Tract Length (aVTL) as a known component. The result is applied as a speaker embedding in neural TTS, allowing control over the aVTL and several emergent properties captured as a representation of the vocal space across speakers. In addition to quantitative metrics, we present a community evaluation conducted by nonbinary SGD users.

Index Terms: gender expansive, nonbinary, TTS, gender and speech, diversity, inclusion, speech generating devices, augmentative and alternative communication

1. Introduction

Prior research into the encoding of gender in speech is extensive, especially as it pertains to cisgender male and female speakers. However, recent research has begun to explore gender expansive voice gender encoding such as how transgender, nonbinary, and gender nonconforming people make use of acoustic features. Recent research suggests that nonbinary individuals have an average fundamental frequency (f0) that is in the middle of the average f0s for men and women [1, 2]. Additionally, nonbinary speakers have been found to have brighter voice quality than gender binary speakers [1]. Preliminary work by [3] has shown that nonbinary speakers may manipulate their vocal tract to encode various aspects of their multidimensional gender identity and expression and that the way they encode these aspects of their gender does not pattern like cisgender male and female speakers. An impediment to analyzing gender expansive speech or incorporating it into research methodologies has been the lack of gender expansive corpora, but data collection from transgender and nonbinary speakers has started and is made available for research [4, 5]. Because gender expansive people make use of acoustic cues differently from cisgender individuals, synthetic voices must be able to reflect the broad range of voice gender expressions, as well as prosodic and paralinguistic expressivity, without reinforcing cisgender binary norms.

Recently, there have been attempts towards generating "gender-neutral" artificial voices, [6, 7, 8]. However, research has shown that such synthetic voices might still be placed within binary categories by the majority of listeners [9]. [10] found that gender expansive people perceive of "gender-neutral" voices differently from cisgender people - in particular, their work suggested that gender expansive people may have a unique third voice gender category in addition to the binary male and female voice genders. However, because of relying on cisgender training data, the voices likely lacked actual sociophonetic cues of gender expansive speech. Efforts creating "gender-ambiguous" voices have similarly relied on binary speech data such as a multi-speaker model by [11] and a prosody-controllable method aimed at researching gender bias by [12].

This paper presents a novel method of creating a palette of adjustable gender expansive synthetic voices using an inclusive approach: conducting community research asking nonbinary SGD users about the needs and wants in a synthetic voice, using a corpus recorded by 14 gender expansive people and designing a multi-speaker model that leverages the variety present in the corpus using modifiable gender-independent parameters, and evaluating the TTS with the help of nonbinary SGD users.

2. Inclusive design for synthetic voices

2.1. Community insights on nonbinary voice preferences

Given the limited research on nonbinary SGD users, their needs and desires for voice features remain largely unexplored. In a survey published in [13], nonbinary SGD users indicated that "none [of the pre-set voices on their SGD] captured their gender well" and several participants in the survey commented that "while they had the option to shift the pitch of the synthetic voice, they did not feel this helped them capture their gender sufficiently." Prior to developing the TTS, we surveyed ten nonbinary SGD users to further understand their preferences for voice features in devices. The survey revealed strong community interest in: *Increased Diversity*: 90% desired a wider range of voice gender options. Non-Binary Labeling: 80% preferred voices without traditional male/female labels. *Voice Texture Control*: 70% wanted to adjust breathiness or roughness. *Intonation Control*: 70% sought to modify voice intonation, aiming for a more varied pitch. *Clarity and Resonance*: 80% wished for the ability to enhance voice clarity and resonance. These insights guided our approach to developing a more inclusive and representative TTS system.

2.2. Objectives and Hypotheses

Our primary research objective is to enrich the diversity of synthetic voices available by using gender expansive speech as the basis for training a multi-speaker TTS system which is modifiable along multiple acoustic dimensions. The aim is to offer gender expansive SGD users options to reinforce authenticity

and self-expression over having to conform to external standards of gender expression due to the lack of available choices.

Our hypothesis is that, using the variability across 14 gender expansive speakers, we would be able to create TTS modifiable along an acoustic Vocal Tract Length dimension and along various other emergent properties that would prove to be meaningful, creating a wide variety of choice in vocal qualities that people in the GE community report that they may identify with.

2.3. Corpus of gender expansive speakers

We use a corpus of fourteen gender expansive speakers in the mid-Atlantic region of the United States (MAGES Corpus [5]) who were recruited via word-of-mouth and social media to participate in a speech collection process which was IRB-approved by the IRB of the University of Delaware. Speakers recorded 400 sentences, chosen to cover the diphones and triphones of English, from the ModelTalker [14] database. They were required to pass a "screening" set of ten sentences to ensure minimal background noise in their environments and to ensure that they were speaking at a comfortable pace and natural prosody.

3. Multi-speaker TTS with Constrained PCA and aVTL features

3.1. Acoustic Vocal Tract Length feature

To estimate acoustic Vocal Tract Length (aVTL), we use a method proposed by [15]. It correlates with physical vocal tract length to a certain degree, but is more representative of a "performative aspect of gender that is overlaid on physical sex differences in vocal tract length". It represents speakers' ability of shortening and lengthening the vocal tract during articulation (e.g., by spreading the lips or dropping the larynx). We calculate aVTL according to the definition of [15]:

$$\text{aVTL} = \frac{c}{(\Delta F * 2)}, \tag{1}$$

where $c = 34000$ cm/s, the speed of sound in warm moist air. ΔF is calculated using the average spacing method:

$$\Delta F = \text{mean}\left(\frac{f_1}{0.5} + \frac{f_2}{1.5} + \frac{f_3}{2.5}\right) \tag{2}$$

ΔF was calculated using vowels [i], [a], [e], [o], and [u] in stressed syllables, and then each person's average ΔF is to used to compute their aVTL. In the MAGES corpus, aVTL measures ranges between 14.0 and 17.3 cm. The aVTL feature is not used as an explicit acoustic feature in the TTS model, rather, as part of the speaker embedding (see: Sec.3.2). Thus the model learns during training the relationship between aVTL and the spectrogram with the help of learning the variability across speakers with different aVTL features in the corpus.

3.2. Proposed architecture

To make the best use of a gender expansive corpus which is both smaller and more varied than typical multi-speaker corpora used for TTS, as well as to be able to add the aVTL as a modifiable parameter at inference, we propose the following architecture. Speaker embeddings are extracted from the MAGES corpus using a pretrained x-vector model [16], a representation originally developed for speaker recognition, and later successfully used as a speaker embedding in multi-speaker TTS [17].

To reduce the vector space to relevant elements that explain the variability in this corpus, and to be able to add the known component related to the aVTL, we use a constrained PCA on the extracted x-vectors. The constraint is set by the component estimated from each speaker's aVTL feature through regression. PCA is performed on the residuals. The first 8 components extracted from the residuals after removing the projected aVTL capture 43.4% of the total variance in the x-vector data. The output from these components is averaged over each speaker's training data. For training, the 8 resulting speaker-specific features and the aVTL are normalised to a range between -1 (lowest avg. value) to 1 (highest avg. value).

We use a modified Tacotron 2 [18] TTS architecture which allows for features appended to the encoder output which can be controlled on a gradient at inference [12]. The architecture and pre-trained gender-ambiguous model trained on 20 hours of speech data[1] is used as base, from which the existing speaker embedding is dropped and training is reinitialised as a warm start with the new corpus and features. Transfer learning starts with a strong acoustic model that is familiar with the relationship between speech and text. Exchanging the data and continuing the training results in a new model that has no data leakage from the base voice, but benefits from having seen large amounts of data. The model is also expanded to use the 9 normalised features added to the encoder output, as a compact and representative speaker embedding. Where additional features are added to the output of the encoder these are initialised with zero weights in the expanded model. The model was trained for 100k iterations on 4GPUs (batch size 64). The speech signal is decoded from the output using the neural vocoder HiFi-GAN [19], the published model is fine-tuned on the MAGES corpus for 950k iterations. TTS samples are available online[2].

3.3. Modulating gender-independent emergent components

As a result of the speaker embeddings being constructed through these 9 components (8 emerging from x-vectors and the known component aVTL) we hypothesise that they can both represent speaker identity and provide individual control on a gradient on each feature at inference. We also hypothesise that the emergent features should relate to relevant perceptual speech features as they explain a significant part of the variation in the speaker representation used as a basis.

4. Objective evaluations

To assess speaker consistency across the 14 synthetic voices in the model, we use a speaker similarity score from a state-of-the-art speaker identification model ECAPA-TDNN [20], calculated as cosine-distance of its embedding space. Following [21] we take synthesised samples from two speakers, calculate all pairwise speaker similarity scores over combinations of two utterances and use the average as the metric for speaker similarity. The same approach pairing up different utterances from the same speaker evaluates consistency for a speaker. To evaluate if high observed speaker similarity is inherent to the TTS or a result of similarity between speakers, we also compare with vocoded ground truth samples. To assess the effectiveness of the manipulation of the TTS along the aVTL parameter, we measured the aVTL on synthesised samples from all speakers over the validation set with the normalised aVTL input ranging from -0.5 to 1. For an automatic estimation of the TTS quality, we use a Mean Opinion Score (MOS) prediction system [22].

[1] https://github.com/evaszekely/ambiguous

[2] https://www.speech.kth.se/tts-demos/interspeech2024-inclusive

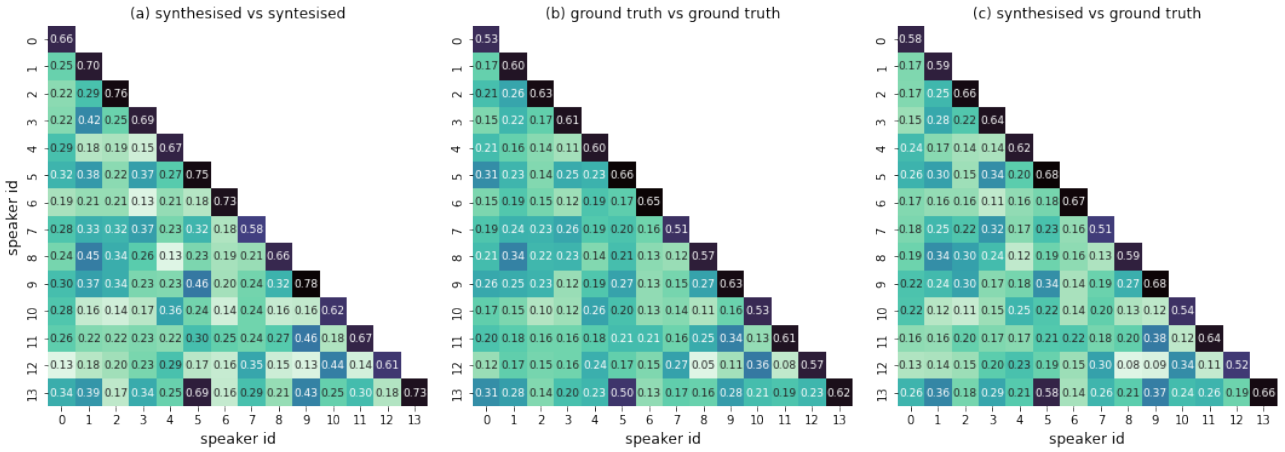

Figure 1: *Speaker similarity calculated as the mean cosine distance of the sample embedding using ECAPA-TDNN speaker identification model . High values demonstrate high similarity between utterances from the same speaker (on diagonal) or utterances from another speaker (off diagonal). Ground truth validation samples are vocoded for comparability.*

5. Community evaluation

5.1. Evaluation Platform: So-to-Speak

In the perceptual evaluation, the interactive platform So-to-Speak [23] [3]. was used. This is an exploratory platform that was developed to evaluate multi-dimensionally controllable TTS. It allow users to interact with hundreds of synthetic speech samples, displaying them on an interactive grid. For this evaluation, the user was able to switch between three grids of 7x7 samples. We displayed the aVTL on the y axis of each grids, while 3 selected components, 1, 2, and 3 were displayed on the x axes of the corresponding grids.

5.2. Qualitative evaluation method

Three nonbinary SGD users were recruited via email from a previously curated pool of nonbinary SGD users who were interested in SGD research to participate in a brief, anonymous trial of the system. The participants listened to all fourteen base voices and then chose their top three to use to test the manipulation of acoustic vocal tract length and three selected emergent acoustic features (components). We refrained from assigning descriptive labels, as it was part of the aim of the evaluation to let listeners infer what type of dimensions each component may represent for them (e.g., acoustic-prosodic, expressive or gender-expression related subjective dimensions). After testing, they filled out a survey assessing various aspects of the system.

6. Results

6.1. Objective evaluations

We evaluate speaker consistency on the 88 utterances from the validation set and their synthesised versions using the embedding for that speaker. High values indicate high speaker consistency, and vice versa for low speaker consistency, as shown in Fig.1(a). Fig.1(b) shows the same metrics for natural speech, we observe that the high speaker similarity between speakers 5 and 13 in the TTS metrics can be attributed to ground truth similarity. Fig.1(c) comparing the ground truth with the synthesised samples, indicates how truthfully the model represents speaker identities.

[3] https://github.com/evaszekely/So_To_Speak

We synthesised the validation set with the aVTL feature adjusted from original to -0.5 and +1.0 for each speaker. Results in Fig. 2 show that the adjustment of the aVTL input results in a measurable shift in that direction.

We evaluate synthesis quality on the validation set, and modulate aVTL and one of the components 1 to 3 in three steps in both directions from the original speaker embedding. Average auto-MOS scores are reported in Fig. 3; significant differences in quality to the results for the unadjusted speaker embedding (middle of each plot) are marked with *. Results show that adjustments of aVTL up to 0.5 in either direction (50% of total range within the corpus) can be done without significant change in quality. When modulating aVTL together with one of the components, changes up to 0.33 in any combination of directions do not significantly degrade quality (Fig. 3).

6.2. Community evaluation

Nonbinary SGD user 1: This SGD-user indicated at the beginning of the interview that they are diagnosed with a hearing impairment. They expressed that they could see how such a fine-grained choice between voice identities and qualities would be useful, but *"the differences [between voices] are really subtle."* They reported that because of their experience being hard-of-hearing, *"it's a bit overwhelming to choose between so many voices."* Despite the overwhelming experience, they reported still being able to detect differences in the voices along the various dimensions and felt that these voices better captured their identity and expression compared to pre-set SGD voices.

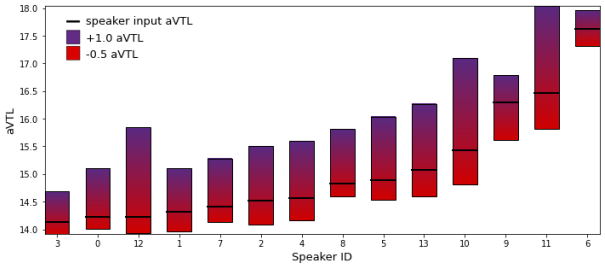

Figure 2: *aVTL measurements on TTS samples where aVTL input feature is adjusted by -0.5, 0.0 and +1.0.*

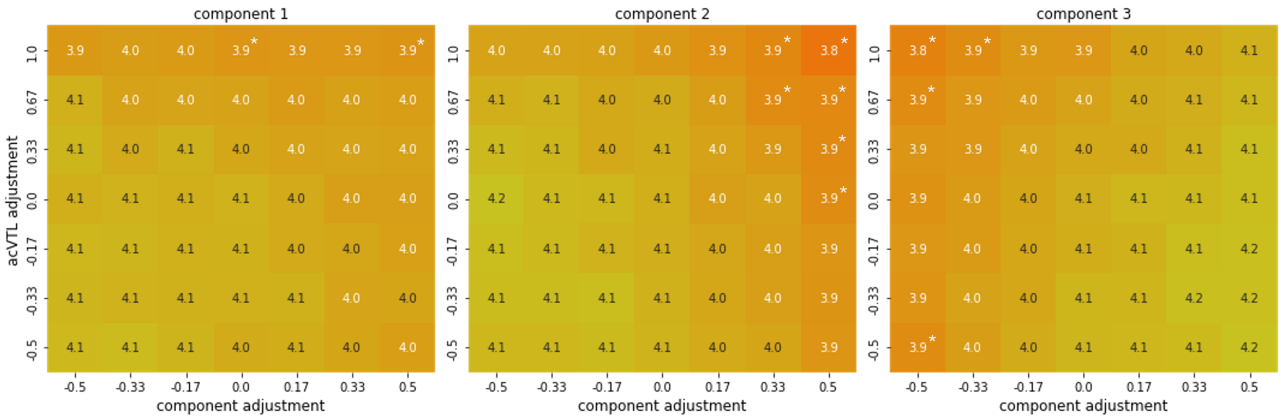

Figure 3: *Consistency on quality when varying aVTL and Component values away from the speaker embedding setting, average of 88 observations per cell. Significant differences from the midpoint with no adjustment ($p<0.05$) are marked with *.*

Nonbinary SGD user 2: This participant reported that they were pleasantly surprised with how good the voice quality was, saying, *"The [voice] quality was a lot better than I expected. [The voices] were surprisingly easy to understand."* They expressed that the base voices didn't fit them exactly, but *"because of the ability to manipulate the different features, I was able to get a voice I liked more."* and they said that if there had been more options for the base voices, they might be able to find their "favorite voice". They expressed that they enjoyed the mechanism of manipulation saying, *"I love the idea of this sort of manipulation of voices. I could easily spend weeks trying to perfect voices if I had this option. This allows for so many missing options that aren't currently available. I need more than pitch shifting in order to be comfortable. I don't [currently] have more than pitch shifting."*

Nonbinary SGD user 3: This SGD user reported that being able to shift more aspects of the voice besides pitch was helpful, *"I liked that not only pitch but things like emphasis and hoarseness could be manipulated."* They shared that if they and another person chose the same base voice, they could still use the manipulation of features to achieve unique ways of speaking, *"It felt like even if, say, I and another person were using the same base voice, the ways in which we used it could be combined with our specific way of speaking to make us possible to distinguish from one another just by voice, which is not a thing you normally expect from this kind of voicebank."* They also enjoyed the different acoustic features but expressed the potential need for more explicit labels describing them: *"I loved the specificity of the manipulation and how granular it was, you really had so many options! But while the effort not to label it in ways that felt gendered was nice, I think for people with less tech savvy or with cognitive issues it could be quite confusing and some simple labeling would be helpful, even if there are aspects of the labels that come off as gendered."*

All three nonbinary SGD users said they would want voices modeled after gender expansive speakers like these ones on their own SGD. Two of the users said they would want the ability to manipulate these acoustic features on their own SGD. All three participants indicated that component 1 seemed to correlate to breathiness from their perception and component 2 captured a dimension that could be described as more gravelly or "hoarseness". Component 3 was associated with "queerness" by user 2 who noted that this component fulfilled their desire for voices to encode aspects of gender as well as sexuality.

7. Discussion

TTS is typically evaluated using online listening tests rating speech samples on various scales [24]. Because of the wide variety the proposed multi-speaker, multi-dimensionally controllable model is capable of producing, combined with the specificity of the use-case, we found it meaningful to replace traditional listening tests with a combination of objective metrics and a more extensive qualitative evaluation involving fewer individuals. The evaluation has highlighted that considering the diverse community of nonbinary SGD users, a highly customisable system is needed. For some, an optimal experience might be achieved through a streamlined selection of voice options or a guided system for choosing voices. For others, true inclusivity entails a more engaged process, involving them to contribute directly to voice design decisions, even the ones involving development [25]. Inspired by the nuanced use of vocal tract manipulations by gender expansive individuals, it is our hope that a synthetic voice modifiable along these dimensions would potentially be capable of approximating gender expansive individuals' ability to gradiently encode identity through voice [3]. To assess this, a more extensive evaluation is needed.

8. Conclusions

The voice palette created on the MAGES corpus with the proposed method of using constrained PCA and aVTL in speaker embeddings was evaluated with objective metrics which have shown that: (1) the TTS reflects the voice identity of the original speakers and does not show significant leakage; (2) synthesis with varied aVTL settings demonstrates measurable changes in the samples for all speakers in the corpus, albeit to a varying degree; (3) the adjustment of the aVTL and three evaluated components only results in quality degradation in the case of extreme values. The qualitative evaluation revealed that participants welcomed the variety represented by the gender expansive voices and they found the modifiable features useful in various ways, but the need for simplification and labeling was also expressed. Future work involves expanding this model to use spontaneous conversational speech, to better model expressive and pragmatic functions in conversational settings. We are also planing to adapt the So-to-Speak evaluation interface, to create a platform optimised to allow users to navigate and customise the voice palette.

9. Acknowledgements

This research was supported by the Swedish Research Council projects Connected (VR-2019-05003), Perception of speaker stance (VR-2020-02396), the Riksbankens Jubileumsfond project CAPTivating (P20-0298).

10. References

[1] B. LeAnn and P. Claire, "Bright voice quality and fundamental frequency variation in non-binary speakers." *Journal of Voice: Official Journal of the Voice Foundation*, pp. S0892–1997, 2022.

[2] M. Schmid and E. Bradley, "Vocal pitch and intonation characteristics of those who are gender non-binary," in *Proc. ICPhS*, 2019, pp. 2685–2689.

[3] M. Hope, C. Ward, and J. Lilley, "Nonbinary American English speakers encode gender in vowel acoustics," in *Proc. Interspeech*, 2023, pp. 4713–4717.

[4] D. V. Dolquist and B. Munson, "A palette of transmasculine voices. retrieved from the data repository for the university of minnesota." [Online]. Available: https://doi.org/10.13020/0fas-n510.

[5] M. Hope, "The Mid-Atlantic Gender Expansive Speech (MAGES) Corpus." [Online]. Available: https://maxwell-hope.com/mages-corpus/

[6] S. J. Sutton, "Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity," in *Proc. of CUI*, 2020, pp. 1–8.

[7] J. Carpenter, "Why project Q is more than the world's first nonbinary voice for technology," *Interactions*, vol. 26, no. 6, pp. 56–59, 2019.

[8] C. Yu, C. Fu, R. Chen, and A. Tapus, "First attempt of gender-free speech style transfer for genderless robot," in *Proc. HRI*, 2022, pp. 1110–1113.

[9] S. Mooshammer and K. Etzrodt, "Gender ambiguity in voice-based assistants: Gender perception and influences of context," *Human-Machine Communication*, vol. 5, no. 1, p. 2, 2022.

[10] M. Hope and J. Lilley, "Gender expansive listeners utilize a non-binary, multidimensional conception of gender to inform voice gender perception," *Brain and Language*, vol. 224, p. 105049, 2022.

[11] K. Markopoulos, G. Maniati, G. Vamvoukakis, N. Ellinas, G. Vardaxoglou, P. Kakoulidis, J. Oh, G. Jho, I. Hwang, A. Chalamandaris *et al.*, "Generating multilingual gender-ambiguous text-to-speech voices," in *Proc. Interspeech*, 2023, pp. 621–625.

[12] É. Székely, J. Gustafson, and I. Torre, "Prosody-controllable gender-ambiguous speech synthesis: a tool for investigating implicit bias in speech perception," in *Proc. Interspeech*, 2023, pp. 1234–1238.

[13] M. Hope and J. Lilley, "Differences in sibilant perception between gender expansive and cisgender individuals," *Seminars in Speech and Language*, vol. 44, pp. 61–75, 2023.

[14] H. T. Bunnell, J. Lilley, and K. McGrath, "The modeltalker project: A web-based voice banking pipeline for als/mnd patients." in *Proc. Interspeech*, 2017, pp. 4032–4033.

[15] K. Johnson, "The δf method of vocal tract length normalization for vowels," *Laboratory Phonology*, vol. 11, no. 1, 2020.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[17] D. Stanton, M. Shannon, S. Mariooryad, R. Skerry-Ryan, E. Battenberg, T. Bagby, and D. Kao, "Speaker generation," in *Proc. ICASSP*, 2022, pp. 7897–7901.

[18] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[19] 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.

[20] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[21] S. Wang and É. Székely, "Evaluating text-to-speech synthesis from a large discrete token-based speech language model," in *Proc. LREC-COLING*, 2024, pp. 6464–6474.

[22] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022.

[23] É. Székely, S. Wang, and J. Gustafson, "So-to-Speak: an exploratory platform for investigating the interplay between style and prosody in TTS," in *Proc. Interspeech*, 2023, pp. 2016–2017.

[24] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[25] G. Pullin, J. Treviranus, R. Patel, and J. Higginbotham, "Designing interaction, voice, and inclusion in aac research," *Augmentative and Alternative Communication*, vol. 33, no. 3, pp. 139–148, 2017.