INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Prosody-controllable gender-ambiguous speech synthesis: a tool for investigating implicit bias in speech perception

Éva Székely[1], *Joakim Gustafson*[1], *Ilaria Torre*[2]

[1]Division of Speech, Music and Hearing, [2]Division of Robotics, Perception and Learning,
KTH Royal Institute of Technology, Stockholm, Sweden

szekely@kth.se, jkgu@kth.se, ilariat@kth.se

Abstract

This paper proposes a novel method to develop gender-ambiguous TTS, which can be used to investigate hidden gender bias in speech perception. Our aim is to provide a tool for researchers to conduct experiments on language use associated with specific genders. Ambiguous voices can also be beneficial for virtual assistants, to help reduce stereotypes and increase acceptance. Our approach uses a multi-speaker embedding in a neural TTS engine, combining two corpora recorded by a male and a female speaker to achieve a gender-ambiguous timbre. We also propose speaker-disentangled prosody control to ensure that the timbre is robust across a range of prosodies and enable more expressive speech. We optimised the output using an SSL-based network trained on hundreds of speakers. We conducted perceptual evaluations on the settings that were judged most ambiguous by the network, which showed that listeners perceived the speech samples as gender-ambiguous, also in prosody-controlled conditions.

Index Terms: speech synthesis, human-computer interaction, gender bias

1. Introduction

Despite decades of research and calls for action, gender biases are still prevalent in today's digital society [1]. These range from male candidates having higher success rates at interviews, to the gender pay gap, to men being considered more persuasive and less "moody" [2, 3]. These stereotypes have been shown to be transferred to artificial agents, including virtual avatars [4], computers [2], and robots [5]. For example, "male" robots were perceived as more suitable for stereotypically male tasks (e.g., repairing technical devices, guarding a house), while "female" robots were perceived as more suitable for stereotypically female tasks (e.g., tasks related to household and care services) [5]. Technology has, at times, made matters worse: for example, image recognition systems perform more poorly for female subjects [6], and gender stereotypes tend to be magnified whenever contextual cues are taken into account for gender prediction [7, 8]. Additionally, the majority of these systems reduce the expression of gender to the traditional view of it being binary and physiological, thus failing to capture the complexities of gender identities [9], and there are clear concerns of the societal and ethical impacts of automatic gender recognition [10]. Also, UNESCO recently published a worrying report suggesting that new technologies such as voice assistants are spreading gender biases [11]. Indeed, gender stereotypes are maintained even when people only hear the voice of a newly met individual [12]. As voice is sometimes the only identifying feature of an artificial agent – or, in the case of robots, the only easily adaptable feature – this raises the question: can gender biases be investigated and reduced with the aid of speech synthesis? Recently, there have been a few calls towards generating "gender-neutral" artificial voices, which could on the one hand give a voice to individuals who are not represented by the existing male/female artificial voices and, on the other hand, reduce voice-induced stereotyping towards the speaking agents [11, 13, 14, 15, 16]. However, there are also concerns that, despite these good intentions, "gender-neutral" voices might still be placed within binary categories by most listeners [17]. Additionally, many "gender-neutral" voice assistants are created only by pitch-shifting [18]; however, only modifying pitch is not enough to achieve a "gender-neutral" voice quality [19, 20]. In this work, we present a method for developing gender-ambiguous TTS that can be used in virtual assistant applications, as well as a tool to investigate implicit gender bias in perceptual experiments. Our approach uses a neural TTS engine with multi-speaker embeddings and utterance-level prosody control to achieve gender-ambiguous timbre. Note that we use the term "gender-ambiguous" (instead of "gender-neutral") to indicate a voice that does not clearly belong to any gender category, and can be perceived as male, female, neutral, or combinations of the above [13, 21].

2. Method

2.1. Synthesis Method

The neural TTS engine Tacotron 2 [22] is modified following [23] by extending the model with an 8-dimensional speaker embedding which is appended to each utterance's encoded text, and passed to the attention and decoder blocks of the model. We use a PyTorch implementation of the system[1]. A two-speaker model is trained on two large publicly available single speaker corpora, LJSpeech [24] containing 16 hours of speech by a female speaker and RyanSpeech [25], containing 9.8 hours of speech by a male speaker. To balance the training data only 9.8 hours of data from LJSpeech is used in training; filenames with last three digits > 115 were removed. To ensure a uniform recording quality between the corpora the samples were joined and cleaned using the Adobe Podcast enhance function[2]. The model was trained for 100k iterations on 4GPUs (batch size 32). The speech signal is decoded from the output using the neural vocoder HiFi-GAN [26], the published model of which is fine-tuned on these corpora for 190k iterations. Any combination of a weighted average of the two speaker embeddings can be used

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

[1]https://github.com/NVIDIA/tacotron2
[2]https://podcast.adobe.com/enhance

10.21437/Interspeech.2023-2086

at inference, the weights do not need to add up to 100%. One drawback of this initial method is that the timbre may be subject to change based on pitch variation, for example resulting in more male-attributable voice qualities at phrase-final falling intonation. In our method we resolve this issue with the introduction of a utterance-level prosody control in Sec. 2.3.

2.2. Speaker Gender Recognition model

We employ a Speech Gender Recognition (SGR) network, trained on Self-Supervised Learning (SSL) representations extracted from a large multi-speaker corpus, as a tool for verifying whether the output of the TTS system can be easily identified as belonging to a male or a female speaker. Since the TTS is trained on a limited number of speakers (in our experiment, only two), an SGR can be helpful to see how the method may generalise to other speakers. However, such a system has limitations in that it is trained on speech that is not ambiguous and hence has only seen negative examples in this context. It also cannot distinguish between cases where timbre is truly ambiguous and where it is inconsistent throughout an utterance. Due to these limitations, we employ it as a development and tuning tool, the results of which should be accompanied by a human listening test. We implemented a similar architecture to the SGR network proposed by [27]. Input is generated from a pretrained wav2vec representation, the row average values of which are used as input to a multilayer perceptron classifier with two fully connected layers of size 512 and ReLu activation. We made the following modifications to this to make it more suitable for the task at hand: (1) As speech representation we use the newer and larger $wav2vec2.0$ representation, which shows strong generalization to unseen data [28] and adjust size of the input layer accordingly to 768. (2) Because it has been reported in literature that using the output of earlier layers of $wav2vec2.0$ is beneficial to downstream tasks [29], we evaluate model performance for representations extracted from the 3rd, 6th, 9th and final layer. (3) We used the LibriTTS [30] dataset for training, specifically, the *train-clean-100* (125 female, 126 male IDs) and *train-clean-360* (439 female, 482 male IDs) subsets, for validation and testing we use *dev-clean* (20 female, 20 male IDs) and *test-clean* (20 female, 20 male IDs) sets, respectively. The same speaker IDs do not appear in more than one set. To balance the gender distribution in all sets, we only included the first 125 and 439 male speaker IDs in the *train-clean-100* and *train-clean-360* sets, respectively. Based on the accuracy on the held out test sample (seeTable 1), the third layer representation from $wav2vec2.0$ was selected as the input for the SGR model.

2.3. Prosody Control

We extend the base synthesis model with an mid-level prosody control method, similar to [31] and [32], in order to stabilise the timbre within a speaker embedding setting, and to add controllable expressivity to the TTS, which can be helpful in experiment designs addressing implicit bias. As additional inputs to the model, mean f0 values over each utterance are normalised, aligning the 1st and the 99th percentile point to -1 and 1 respectively, and appended to the model training input in parallel with the speaker embedding. Normalisation is performed either the whole combined corpus, (Fig. 5 top) or for each speaker individually. This second method aims to disentangle the average f0 from the speaker embedding (Fig. 5 bottom). At inference, pitch can be controlled at utterance- or word-level. The model will sample a natural prosody to match the target that is set, which enables a more fine-grained control without having to specify the exact intonation contour.

Table 1: *SGR test accuracy of different* $wav2vec2.0$ *layers*

	3rd layer	**6th layer**	**9th layer**	**final layer**
acc.	99.1%	98.2%	94.9%	72.9%

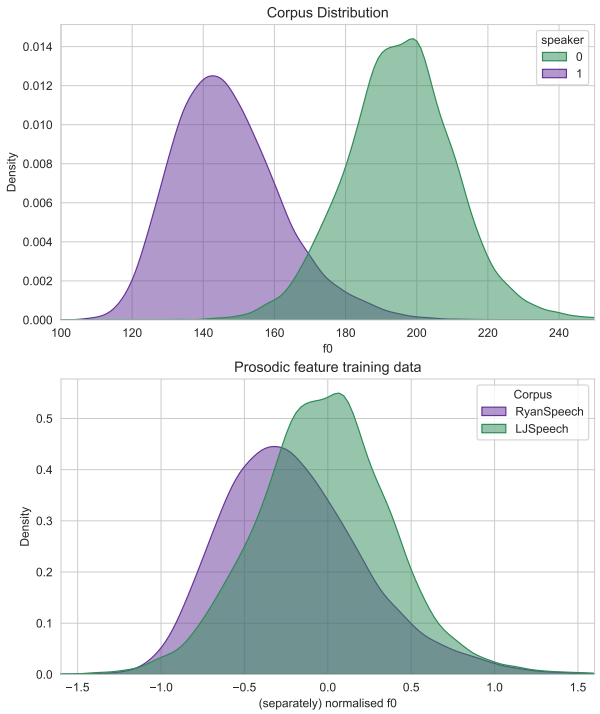

Figure 1: *Corpus distribution of f0 values averaged over each utterance. For prosodic control over the synthesis these can be normalised jointly (upper plot) or separately (lower plot).*

3. Objective evaluation

For the experiments, 30 utterances were selected from the Timit corpus [33] – which consists of emotionally neutral utterances [34] – and were passed through a gender bias detection tool to ensure they were free of textual bias[3]. Each sentence was synthesised on each of the three models with speaker embedding weights varying between 0 and 100% in 5% increments on each speaker and pitch settings (for the two models with pitch control) varying between -0.4 (-20%) to +0.4 (+20%) in 0.2 increments. The synthesised utterances were evaluated using the SGR model; Fig. 5 displays the average probability assigned that the speech is from a female speaker. A first objective comparison is made between the systems by fitting a linear regression on the settings that are deemed ambiguous by SGR. For this we follow the definition set in [27] that if neither gender is assigned a probability greater than 0.6 this can be deemed ambiguous. From all ambiguous settings, we estimate the optimal setting for the male embedding as a linear relation with first only the female embedding setting, and second, based on both the female embedding setting and on the pitch input control setting. The two methods of pitch control are compared using

[3] https://www.appcast.io/gender-bias-decoder

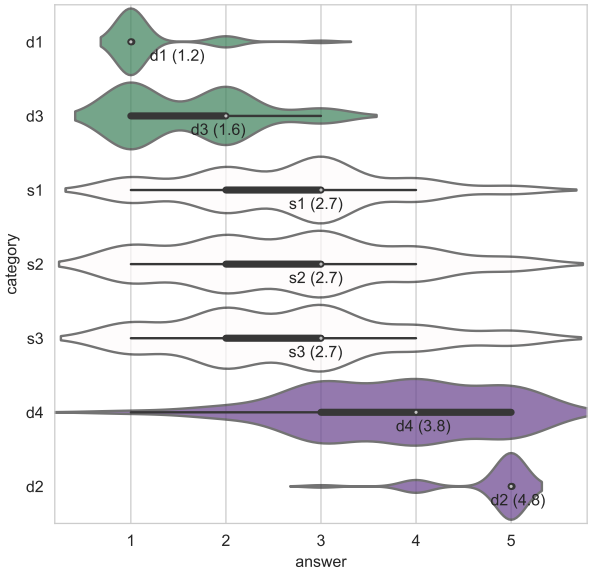

Figure 2: *Evaluation results with distractors (d1-4) and three proposed embedding settings (s1-3) without pitch control. Five point likert scale (1:female, 2:probably female, 3:ambiguous, 4:probably male, 5:male). Category averages are in brackets.*

linear regression models in- and excluding the pitch input settings. For the joint normalisation model, inclusion of the pitch input factor explains a significant amount of variance among the ambiguous SGR results ($\Delta R^2 0.21, p < 0.01$). For the separate normalisation model, the pitch input factor explains a smaller amount of variance ($\Delta R^2 0.10, p < 0.01$), while the contribution of the embedding weight has greater explanatory power (in both relative and absolute terms of R^2) based on which we conclude that this model provides the most stable outcomes both with and without f0 modification (see Fig. 5).

4. Perceptual evaluation

4.1. Evaluation setup

We carried out a perceptual evaluation with the separately normalised controllable model. Three different f/m settings (chosen to be relatively far apart from each other) were selected from those determined as ambiguous by the SGR model to be evaluated through a perceptual experiment. Specifically, the settings (s1) f:30/m:32, (s2) f:50/m:46 and (s3) f:65/m:57 were selected, where the value of m is based on regression results excluding the impact of possible f0 input shifts. For the perceptual evaluation, 8 sentences out of the original 30 were synthesised at each setting. To evaluate the stability of the model when prompting for a change in pitch, the sentences were also synthesised with pitch input settings of +0.2 and -0.2, resulting in a total of 72 stimuli. A set of 26 distractors were created for the evaluation, using the same 8 utterances in the following conditions: (d1) 7 utterances with only female embedding f:100/m:0, (d2) 7 utterances with only male embedding f:0/m:100, (d3) 6 utterances with predominant female embedding f:70/m:30 and (d4) 6 utterances with predominant male embedding f:30/m:70. For distractors d3 and d4 two utterances were created with pitch settings -0.2, 0 and +0.2, respectively. All samples were presented to each participant and evaluated on a 5-point scale: 1-female, 2-probably

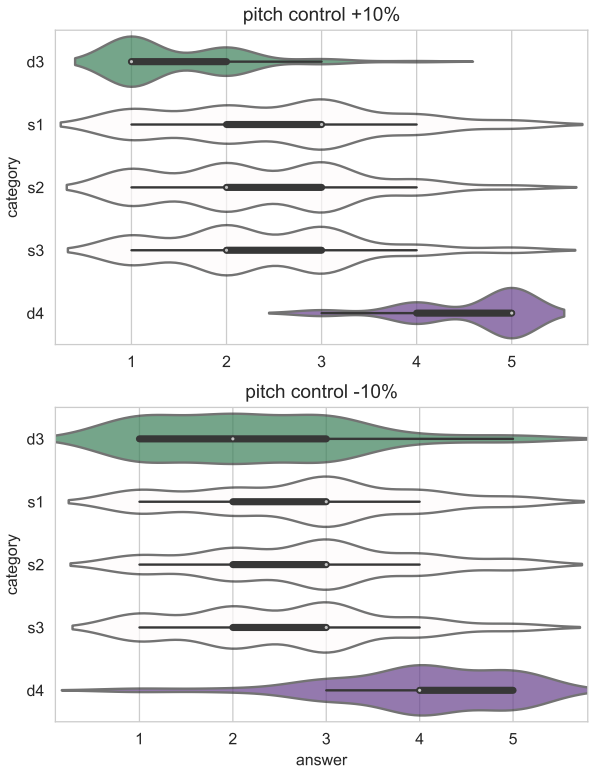

Figure 3: *Evaluation results when pitch input feature is shifted by 10% upwards or downwards.*

female, 3-ambiguous, 4-probably male, 5-male. The evaluation samples can be listened to under https://www.speech.kth.se/tts-demos/gender-ambiguous.

4.2. Perceptual evaluation results

Native English speakers were recruited using Prolific [35]. 18 participants identified as female and 17 identified as male. Two participants were removed for failing attention checks. The results (see Fig. 2), show that each of the three settings resulted in ambiguous speech whether evaluated on the mean or median rating, while the distractors are also consistently rated according to the setting. There was no significant difference between the ratings given by the male or female participants ($p = 0.23$).

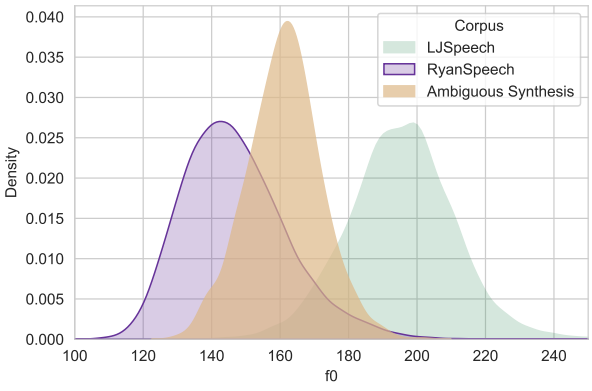

Figure 4: *Range of average f0 in each corpus and within the ambiguous range for the new model with f0 control up to 10%.*

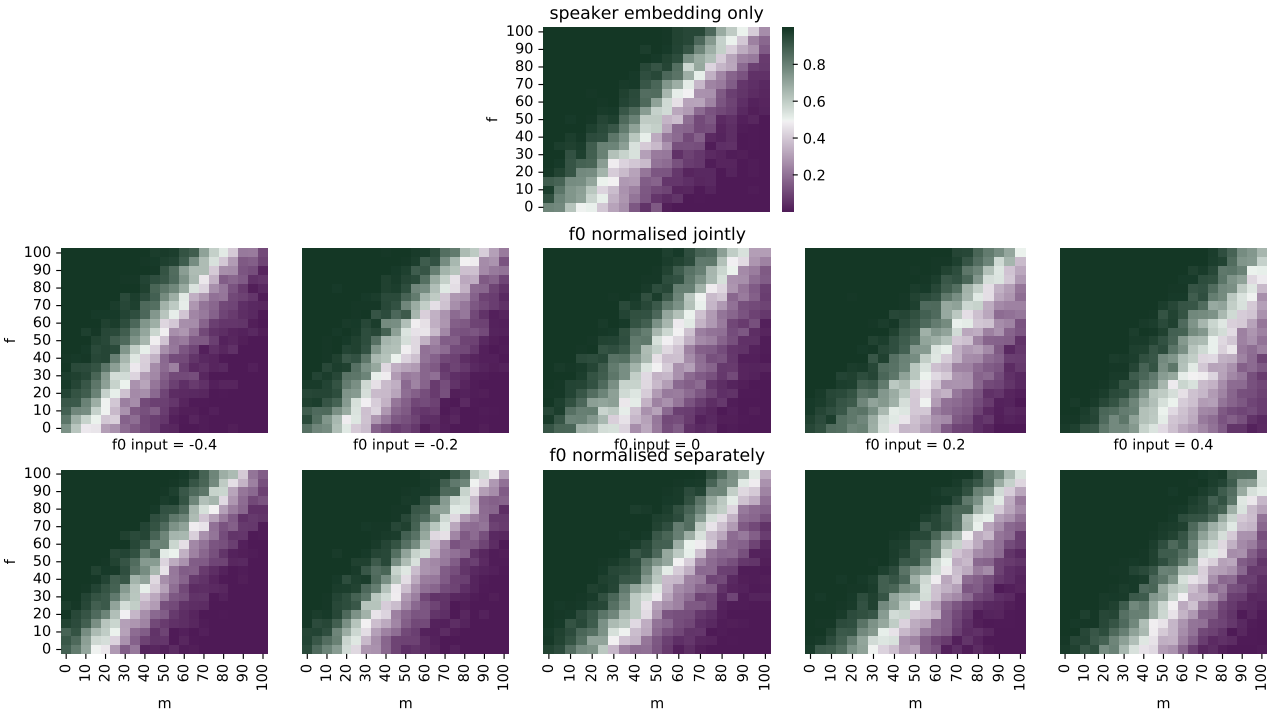

Figure 5: *Heatmap for SGR evaluation on synthesis of 30 utterances; probability assigned that speaker is female is displayed: evaluated as female in green, evaluated as male in purple. Top is system without f0 control, mid is system with f0 input features jointly normalised, bottom with f0 input normalisation per speaker.*

We compare the ratings for the three settings with the paired ratings where the utterance received a shift in the pitch input factor of 0.2 (+ and -10% respectively, Fig. 3). Based on a Wilcoxon signed-rank test we do not reject the null hypothesis that the evaluation is the same for the lower pitch controlled condition ($p = 0.24$). For the condition with the higher pitch control, the hypothesis is rejected ($p < 0.01$). However, as the score stays within the ambiguous range, we can conclude that up to 10% change of pitch control does not affect the perception of the voice as ambiguous. We evaluate the range of pitch values the voice can generate while maintaining ambiguous voice characteristics. We compare the average pitch values attained on the synthesised samples from the subset of original 30 utterances with a input shift of maximum 10%, where these are evaluated as ambiguous based on SGR to the distribution of utterance average pitch values of the original corpora. Measured on the 90 percent IPR, the new voice has a range of 35.1 Hz on average utterance pitch while maintaining ambiguity, while the corpora have a total range of 50.4 Hz (f) and 49.9 Hz (m) on average utterance pitch (see Fig. 4).

5. Discussion

We demonstrated good results using 10-hour long sections of two TTS corpora. The use of the SGR network trained on hundreds of speakers ensures that the method itself will generalise to other corpora, but it is possible that the timbre and voice quality of the particular speakers affects how ambiguous a TTS voice created with this method can be. The method itself can be easily adapted for multiple speaker embeddings in the TTS model, however ,with multiple speakers, the relationship between timbre, f0 range and perceived gender-ambiguity can become more complex and harder to disentangle. This expressive ambiguous TTS voice can be applied to a variety of perceptual and human-computer interaction domains. For example, it could be used to synthesise utterances containing phrases that are considered "gendered language", in different interactive contexts, and examine how they impact listeners' judgements or behavior. It could also be presented on female- or male-appearing robots to see whether the ambiguous voice moderates any stereotypes induced by the robot's appearance. Using the prosody control in the TTS, bias related to prosodic characteristics of speech such as uptalk [36] could be investigated. Future work involves extending this method to spontaneous speech corpora [37], to allow for investigating effects of disfluencies and discourse markers [38] on bias, and adding control of voice quality features such as creak [39] and spectral tilt.

6. Conclusions

Our study presents a prosody-controllable method for generating gender-ambiguous neural TTS that is capable of producing expressive speech with a diverse pitch range. The method includes employing the improved version of a previously proposed SSL-based SGR network that was trained on hundreds of speakers, to find speaker embedding settings that are potentially ambiguous. Our perceptual evaluation showed that listeners rated the speech samples synthesised with these settings as gender-ambiguous, including those with upward and downward pitch control. The code is available open source [4], and we hope that TTS voices created with our method will become a valuable tool for researchers to examine gender bias in speech perception, and to investigate how social interaction is affected by stereotypes and implicit bias.

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

7. References

[1] A. Danielescu, "Eschewing gender stereotypes in voice assistants to promote inclusion," in *Proc. CUI*, 2020, pp. 1–3.

[2] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proc. SIGCHI*, 1994, pp. 72–78.

[3] I. M. Latu, M. S. Mast, and T. L. Stewart, "Gender biases in (inter) action: The role of interviewers' and applicants' implicit and explicit stereotypes in predicting women's job interview outcomes," *Psychology of Women Quarterly*, vol. 39, no. 4, pp. 539–552, 2015.

[4] S. Brahnam and A. De Angeli, "Gender affordances of conversational agents," *Interacting with Computers*, vol. 24, no. 3, pp. 139–153, 2012.

[5] F. Eyssel and F. Hegel, "(s) he's got the look: Gender stereotyping of robots," *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012.

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. of the 1st Conference on Fairness, Accountability and Transparency*, vol. 81, 2018, pp. 77–91.

[7] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints," *Proc. EMNLP*, pp. 2979–2989, 2017.

[8] G. I. Melsión, I. Torre, E. Vidal, and I. Leite, "Using explainability to help children understand gender bias in ai," in *Interaction design and children*, 2021, pp. 87–99.

[9] F. Hamidi, M. K. Scheuerman, and S. M. Branham, "Gender Recognition or Gender reductionism? The social implications of Automatic Gender Recognition systems," in *Proc. SIGCHI*. ACM Press, 2018, pp. 1–13.

[10] O. Keyes, "The misgendering machines: Trans/HCI implications of automatic gender recognition," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–22, Nov. 2018.

[11] M. West, R. Kraut, and H. Ei Chew, "I'd blush if I could: closing gender divides in digital skills through education," Tech. Rep., 2019. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[12] S. J. Ko, C. M. Judd, and I. V. Blair, "What the voice reveals: Within-and between-category stereotyping on the basis of voice," *Personality and Social Psychology Bulletin*, vol. 32, no. 6, pp. 806–819, 2006.

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

[14] 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.

[15] J. Cambre and C. Kulkarni, "One voice fits all? social implications and research challenges of designing voices for smart devices," vol. 3, no. CSCW, pp. 1–19, 2019.

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

[17] 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.

[18] S. Tolmeijer, N. Zierau, A. Janson, J. S. Wahdatehagh, J. M. M. Leimeister, and A. Bernstein, "Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution," in *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–7.

[19] Y. Leung, J. Oates, and S. P. Chan, "Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 2, pp. 266–297, 2018.

[20] M. Hope and J. Lilley, "Cues for perception of gender in synthetic voices and the role of identity." in *Proc. Interspeech*, 2020, pp. 4143–4147.

[21] M. T. Elian, S. Bao, S. Masuko, and T. Yamanaka, "Designing gender ambiguous voice agents-effects of gender ambiguous voice agents on usability of voice user interfaces," *International Journal of Affective Engineering*, vol. 22, no. 1, pp. 53–62, 2023.

[22] 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.

[23] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*, 2020, pp. 6189–6193.

[24] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[25] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," in *Proc. Interspeech*, 2021.

[26] 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.

[27] D. Rizhinashvili, A. H. Sham, and G. Anbarjafari, "Gender neutralisation for unbiased speech synthesising," *Electronics*, vol. 11, no. 10, p. 1594, 2022.

[28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[29] S. Wang, G. E. Henter, J. Gustafson, and É. Székely, "A comparative study of self-supervised speech representations in read and spontaneous TTS," in *Proc. ICASSP SASB*, 2023.

[30] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for Text-to-Speech," 2019.

[31] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *Proc. Interspeech*, pp. 4432–4436, 2020.

[32] A. Kirkland, M. Włodarczak, J. Gustafson, and É. Székely, "Perception of smiling voice in spontaneous speech synthesis," in *Proc. SSW11*, 2021, pp. 26–28.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[34] C. Busso, S. Lee, and S. S. Narayanan, "Using neutral speech models for emotional speech analysis." in *Proc. Interspeech*, 2007, pp. 2225–2228.

[35] "Prolific," Oxford, UK, 2014, accessed: 24.02.2023. [Online]. Available: https://www.prolific.co

[36] A. Ritchart and A. Arvaniti, "The form and use of uptalk in Southern Californian English," in *Proc. Speech Prosody*, 2014, pp. 331–335.

[37] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector," in *Proc. ICASSP*, 2019.

[38] S. Wang, J. Gustafson, and É. Székely, "Evaluating sampling-based filler insertion with spontaneous tts," in *Proc. LREC*, 2022, pp. 1960–1969.

[39] H. Lameris, M. Włodarczak, J. Gustafson, and Székely, "Neural speech synthesis with controllable creaky voice style," in *Proc. ICPhS*, 2023.