# SPEAKER GENERATION

*Daisy Stanton  Matt Shannon  Soroosh Mariooryad  RJ Skerry-Ryan*
*Eric Battenberg  Tom Bagby  David Kao*

Google Research, USA

## ABSTRACT

This work explores the task of synthesizing speech in non-existent human-sounding voices. We call this task "speaker generation", and present TacoSpawn, a system that performs competitively at this task. TacoSpawn is a recurrent attention-based text-to-speech model that learns a distribution over a speaker embedding space, which enables sampling of novel and diverse speakers. Our method is easy to implement, and does not require transfer learning from speaker ID systems. We present objective and subjective metrics for evaluating performance on this task, and demonstrate that our proposed objective metrics correlate with human perception of speaker similarity. Audio samples are available on our demo page[1].

***Index Terms***— speaker generation, text-to-speech synthesis

## 1. INTRODUCTION

In recent years, deep neural network-based text-to-speech (TTS) models have been developed to generate high-fidelity audio that can be indistinguishable from genuine speech. A major limitation of most of these models, however, is that they can only synthesize voices from the training set. Expanding the set of speakers for a good-quality TTS model can require recording voice actors in a studio-quality acoustic environment and then re-training or fine-tuning the TTS model, which can be laborious and expensive. This presents a challenge, since the ability to synthesize speech in a rich variety of voices has practical import for applications including audiobook readers, speech-based assistants, character voices for games, and video dubbing. Designing a TTS model that can generate its own novel voices would be transformative, and would also present an attractive privacy-preserving alternative to voice cloning systems, which aim to recreate a human speaker's voice with a small amount of ground truth audio.

In this paper, we present TacoSpawn, a TTS model that, to the best of our knowledge, is the first system designed to directly generate high-quality speech in voices not corresponding to a particular human speaker. We dub this ability "speaker generation". The main contributions of this work are:

- Proposal and analysis of TacoSpawn, a TTS system with a jointly-trained maximum likelihood estimation model that learns a distribution over speaker embeddings.
- Proposal of objective metrics and subjective tests for evaluating speaker generation models.
- Experimental results quantifying the strong performance of our approach.

[1] https://google.github.io/tacotron/publications/speaker_generation

## 2. BACKGROUND AND RELATED WORK

Modern neural text-to-speech (TTS) systems rely on an explicit representation of speaker identity in order to synthesize speech in multiple voices. A good speaker representation uses similar encodings for utterances from the same speaker (even with varying text or prosody), and different encodings for different speakers. When the goal is to synthesize a speaker in the training corpus, this representation may be a simple one-hot encoding of speaker identity [1], or embeddings trained jointly with the rest of the model [2]. When the goal is to synthesize an unseen speaker at test time, few-shot (speaker adaptation) or zero-shot (speaker encoder) approaches are often used. In speaker adaptation, all or part of the TTS model is fine-tuned to a small number of audio samples from the unseen speaker [3, 4, 5, 6]. In the speaker encoder approach, embeddings are inferred by an encoder network embedded directly in the TTS model [7, 8], or using an auxiliary encoder trained on a large amount of audio-only data [4, 9, 10]. The latter may be trained on a speaker-discriminative objective [11, 12] or on a voice-conversion task [13].

Comparatively little research has been devoted to synthesizing speech from truly novel speakers. Tacotron-2D [4] conditions a Tacotron model on an utterance-level d-vector at training time, and can be fed uniformly random points on the unit hypersphere to generate unseen speakers. However, no attempt is made to ensure that these speakers have a distribution similar to the training speakers, and the use of utterance-level d-vectors may limit the extent to which generated speakers can synthesize audio with prosody distinct from that of the reference utterance. Recent work based on a deep Gaussian process mel spectrogram model [14] is similar in spirit to our approach, sampling from a speaker embedding prior to generate new speakers. However, postprocessing of the sampled speaker embeddings is required to ensure that statistics of generated speakers match those of training speakers. In preliminary experiments, we found that approximately marginalizing over speaker embeddings using a variational approach typically performed worse than the TacoSpawn approach described in §3, especially when using a standard normal prior as in that work. It would be interesting to combine the deep Gaussian process mel spectrogram model with TacoSpawn-style speaker modeling and generation.

## 3. MODEL

### 3.1. Multi-speaker Tacotron

The basis of our model is an extension of Tacotron [15] with trainable speaker embeddings to support multiple speakers [2]. The Tacotron model $p_\theta(y|x, s, c)$ autoregressively predicts a mel spectrogram $y = [y_t]_{t=1}^{T}$ given a phoneme sequence $x$, a speaker embedding $s \in \mathbb{R}^D$, speaker metadata $c$ specifying the speaker's locale and gender, and model parameters $\theta$. A trainable *speaker*

*embedding table* $S \in \mathbb{R}^{J \times D}$ specifies the speaker embedding for each of the $J$ *training speakers* which appear in the training corpus.

We train the multi-speaker Tacotron model using maximum likelihood estimation. Throughout we use $i$ to index the $I$ utterances in the training corpus and $j$ to index the $J$ training speakers. The training corpus $(Y, X, C)$ consists of target mel spectrograms $Y = [Y_i]_{i=1}^I$, phoneme sequences $X = [X_i]_{i=1}^I$, and speaker metadata $C = [C_j]_{j=1}^J$. We assume it is known which training speaker $j(i)$ produced each utterance $i$. We learn the speaker embedding table $S = [S_j]_{j=1}^J$ and other parameters $\theta$ to maximize the log likelihood

$$\log p_\theta(Y|X, S, C) = \sum_{i=1}^{I} \log p_\theta(Y_i|X_i, S_{j(i)}, C_{j(i)}) \quad (1)$$

In practice we train using the original deterministic Tacotron teacher-forced $\ell_1$ loss, corresponding to a fixed-variance isotropic Laplace output distribution in the probabilistic formulation presented here.

To synthesize speech for a new phoneme sequence $x$ from training speaker $j$, we recursively generate a mel spectrogram $y = [y_t]_{t=1}^T$ by setting $y_t = \arg\max_{y_t} p(y_t|y_{1:t-1}, x, S_j, C_j)$. We then use a neural vocoder to convert the generated mel spectrogram $y$ to a time-domain waveform $w = \text{vocode}(y)$ [16].

### 3.2. TacoSpawn

Our approach to speaker generation uses the learned speaker embeddings from a multi-speaker Tacotron model as training data for a distribution over speaker embeddings. In this section we describe this distribution and how to use it for speaker generation.

The *speaker embedding prior* $p_\omega(s|c)$ models the distribution over the speaker embedding $s \in \mathbb{R}^D$ given locale and gender metadata $c$ and parameters $\omega$. In §5, we use a mixture of $K$ Gaussians

$$p_\omega(s|c) = \sum_{k=1}^{K} \alpha_{\omega,k}(c)\, \mathcal{N}\left(s; \mu_{\omega,k}(c), \text{diag}(\sigma^2_{\omega,k}(c))\right) \quad (2)$$

where $\omega$ are the parameters of a dense neural net which takes one-hot encodings of the locale and gender metadata $c$ as input, and produces three outputs: mixture component weights $\alpha_\omega(c) \in \mathbb{R}^K$ using a softmax activation, mean vectors $\mu_\omega(c) \in \mathbb{R}^{K \times D}$, and scale vectors $\sigma_\omega(c) \in \mathbb{R}^{K \times D}$ using a softplus activation. Our implementation supports using any subset of $\{\text{locale}, \text{gender}\}$ as the input to the neural net, including the empty set, in which case we have an *unconditional prior* $p_\omega(s|c) = p_\omega(s)$. Despite its simplicity, we find a mixture of Gaussians prior to be effective for high-quality speaker generation.

To train the speaker embedding prior, we use the training speaker embeddings from a multi-speaker Tacotron model as targets. We learn the parameters $\omega$ to maximize the log likelihood

$$\log p_\omega(S|C) = \sum_j \log p_\omega(S_j|C_j) \quad (3)$$

where $S = [S_j]_{j=1}^J$ is the speaker embedding table learned by a multi-speaker Tacotron model and $C = [C_j]_{j=1}^J$ is the locale and gender metadata for each training speaker. We refer to this approach, where a speaker embedding prior is estimated using maximum likelihood on a training speaker embedding table which was itself estimated using maximum likelihood, as **TacoSpawn**.

In practice we train $(\theta, S)$ and $\omega$ at the same time, using a stop-gradient operation on $S$ when optimizing $\log p_\omega(S|C)$ to emulate the separate losses for $(\theta, S)$ and $\omega$ above. Our training objective is

$$L^{\text{TacoSpawn}}(\theta, \omega, S) = \frac{1}{I}\sum_{i=1}^{I} \log p_\theta(Y_i|X_i, S_{j(i)}, C_{j(i)}) + \frac{1}{J}\sum_{j=1}^{J} \log p_\omega(\text{sg}(S_j)|C_j) \quad (4)$$

where sg is the stop-gradient operation and where the first term is replaced by a minibatch approximation in practice.

A trained speaker embedding prior can be used to generate new speakers. Given desired locale and gender metadata $c$ for a new speaker, we generate a speaker embedding by sampling $s \sim p_\omega(s|c)$ with temperature one. We synthesize speech from the new speaker as described in §3.1, using $s$ and $c$ in place of $S_j$ and $C_j$.

## 4. EVALUATING SPEAKER GENERATION

To evaluate speaker generation performance, we compare the probability distributions of training and generated speakers' audio using a non-parametric statistical approach. While any audio-based statistic whose value differs between training and generated speakers indicates a flaw in speaker generation, we base our statistics on *d-vectors* [11] to encourage them to be sensitive to speaker identity [17].

### 4.1. Speaker-level d-vectors

Given an evaluation corpus containing additional utterances from the training speakers, we compute a d-vector $V_j^{\text{t}}$ for each training speaker $j$ by averaging utterance-level d-vectors computed on their ground truth audio. We synthesize the evaluation corpus using the training speaker embeddings and similarly compute a d-vector $V_j^{\text{s}}$ for each training speaker $j$. Finally we generate $J$ speakers by sampling $S_j \sim p_\omega(s|C_j)$, synthesize the evaluation corpus using the generated speaker embeddings, and compute a d-vector $V_j^{\text{g}}$ for each speaker $j$. Using the same number of speakers and speaker metadata for training and generated speakers ensures that metrics such as s2s, g2s and g2g described below are directly comparable.

### 4.2. Speaker distance metrics

In this section we propose intuitive objective measures of both speaker generation performance and training speaker fidelity. We use the naming convention $x2y$ to denote the distance from $x$ to $y$, where $x$ and $y$ are one of t (ground truth training speaker audio), s (synthesized training speaker audio) or g (synthesized generated speaker audio).

To evaluate speaker generation performance, writing $d(v_1, v_2) = 1 - \frac{v_1}{\|v_1\|} \cdot \frac{v_2}{\|v_2\|}$ for the cosine distance [18], we compute:

- **s2s**: How close is a typical training speaker to other nearby training speakers, when both are synthesized?

$$\underset{j}{\text{median}} \min_{k \neq j} d(V_j^{\text{s}}, V_k^{\text{s}}) \quad (5)$$

- **g2s**: How close is a typical generated speaker to nearby training speakers, where both are synthesized?

$$\underset{j}{\text{median}} \min_{k \neq j} d(V_j^{\text{g}}, V_k^{\text{s}}) \quad (6)$$

- **g2g**: How close is a typical generated speaker to other nearby generated speakers?

$$\underset{j}{\text{median}} \min_{k \neq j} d(V_j^{\text{g}}, V_k^{\text{g}}) \quad (7)$$

If speaker generation is working perfectly then the statistics of training and generated speakers' synthesized audio will be identical, and so s2s, g2s and g2g will be equal. Comparing g2s to s2s helps detect whether generated speakers sound disproportionately similar to the training speakers (g2s $<$ s2s), as well as whether generated speakers sound disproportionately dissimilar to training speakers (g2s $>$ s2s). This provides a measure of how *realistic* the generated speakers are. Comparing g2g to s2s helps detect whether generated speakers sound disproportionately similar to each other (g2g $<$ s2s). This provides a measure of the *diversity* of the generated speakers.

Speaker generation performance cannot be entirely disentangled from the fidelity with which the model captures training speaker characteristics. We therefore also evaluate *speaker fidelity* using:

- **s2t-same**: How similar is synthesized audio from a typical training speaker to ground truth audio from that speaker?

$$\underset{j}{\text{median}}\, d(V_j^{\text{s}}, V_j^{\text{t}}) \tag{8}$$

- **s2t**: How similar is synthesized audio from a typical training speaker to ground truth audio from other nearby training speakers?

$$\underset{j}{\text{median}} \min_{k \neq j} d(V_j^{\text{s}}, V_k^{\text{t}}) \tag{9}$$

The smaller the value of s2t-same, the closer synthetic speech from a training speaker is to their natural speech.[2] We use s2t as a dataset-specific reference point against which to compare s2t-same. A potential system weakness detected by s2t-same but not by (s2s, g2s, g2g) is where the speaker identity is muddy or indistinct in synthesized audio from both training and generated speakers.

For the system taken as a whole, we look for g2s and g2g equal to s2s to indicate effective speaker generation and s2t-same as small as possible relative to s2t to indicate strong training speaker fidelity.

## 5. EXPERIMENTS

### 5.1. Experimental setup

We evaluate TacoSpawn models with 10 Gaussian mixture components, trained on both public and proprietary 24-kHz multi-speaker English datasets:

- **libriclean** (public): All "clean" subsets of the LibriTTS corpus, combined into one (US English, 1230 speakers, ~240.5 hours, mixed-gender). We evaluate on $1\%$ of these utterances.
- **enus1100** (proprietary): A 1100-speaker US English dataset of mixed-gender voices speaking for 30 minutes each, for a total of ~246,000 utterances (~500 hours). We evaluate on $2\%$ of these utterances.
- **en1468** (proprietary): A 1468-speaker mixed-locale English dataset that adds audiobook, voice assistant, and news-reading speakers to enus1100, for a total of ~876,000 utterances, (~717 hours). We evaluate on $1\%$ of these utterances. Many of the speakers have been used in previous research [1, 15, 16].

We use input phoneme sequences produced by a text normalization front-end and lexicon. We use the Adam [19] optimizer for ~300k steps on 32 Google TPUv3 cores, using batch size 256. Except for §5.3.1, we use a non-causal WaveNet GAN vocoder. This has a generator trained to minimize the reverse KL divergence to the reference waveform in a hybrid f-GAN training setup [20, 21, 22] and critics trained using Jensen-Shannon divergence with an $\ell_1$ feature matching loss applied at every hidden layer. The d-vector model used for objective evaluation had 256-dimensional output and was trained on a separate corpus with a speaker-discriminative loss.

[2]Note that d-vectors vary substantially across utterances from the same speaker for both natural and synthesized speech, and so we do not expect s2t-same to be zero even for a perfect multi-speaker speech synthesis system.

| dataset | speaker fidelity | | speaker generation | | |
|---|---|---|---|---|---|
| | s2t-same | s2t | s2s | g2s | g2g |
| libriclean | 0.42 | 0.51 | 0.41 | 0.41 | 0.40 |
| en1468 | 0.27 | 0.33 | 0.17 | 0.18 | 0.17 |
| enus1100 | 0.14 | 0.30 | 0.24 | 0.24 | 0.24 |

**Table 1**. Speaker distance metrics for TacoSpawn models.

### 5.2. Objective evaluation results

#### *5.2.1. Speaker distance metrics*

To evaluate how well the distribution of generated speakers matches that of speakers in the training corpus, we can compute the speaker distance metrics proposed in §4.2. Table 1 shows results for both public and proprietary dataset models. We can see that s2t-same is lower than s2t, providing a useful sanity check that the model successfully captures characteristics of speaker identity for the training speakers. We also see that g2s and g2g almost perfectly match s2s, indicating that generated speakers are statistically as diverse and realistic as training speakers.

As an aside, we note that s2t-same is larger than s2s for many models. While not ideal, this reflects the fact that waveforms output by neural TTS models are still distinguishable from ground truth audio, albeit not necessarily by humans. Specifically, the d-vector model we use to calculate speaker distance metrics may be sensitive to acoustic characteristics such as reverb, channel noise, or speech synthesis artifacts. [4, 23] Since the focus of this work is achieving speaker generation rather than solving speech synthesis, this does not hinder our analysis.

#### *5.2.2. Visualizing speaker distance*

Plotting speaker d-vectors helps visualize these speaker distance relationships. Figure 1 shows t-SNE plots of d-vectors extracted from audio synthesized by our libriclean model for both training and generated speakers. As expected, the d-vectors fall into two gender clusters. We can see that the training and generated speaker distributions are nicely overlapping, and it does not appear that generated speakers clump disproportionately close to training speakers.

#### *5.2.3. Fundamental frequency analysis*

To further examine the perceptual properties captured by speaker distance metrics, we also examined $F_0$ ranges. Following Mitsui et al. [14], we computed the median $F_0$ of utterances spoken by 200 generated and 200 training speakers (100 of each gender).[3] The results, plotted in Figure 2, show that the $F_0$ range of training and generated speakers are equally diverse, clearly cluster into male and female, and are distributionally similar.

[3]We used the Yin [24] extraction algorithm (frame shift=12.5 ms).

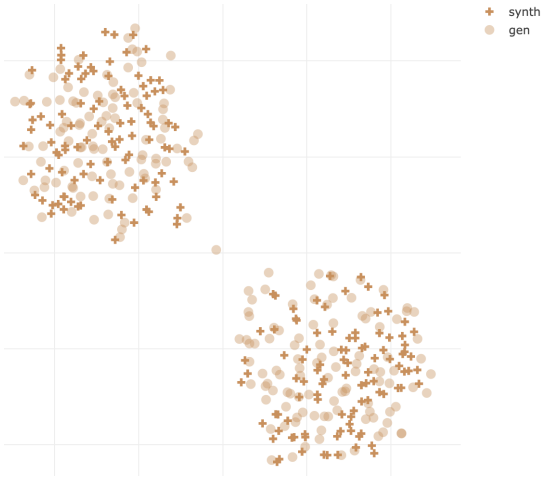


**Fig. 1**. t-SNE plot of speaker-level d-vectors for 100 training (synth) and 100 generated (gen) speakers for audio synthesized using the libriclean TacoSpawn model.

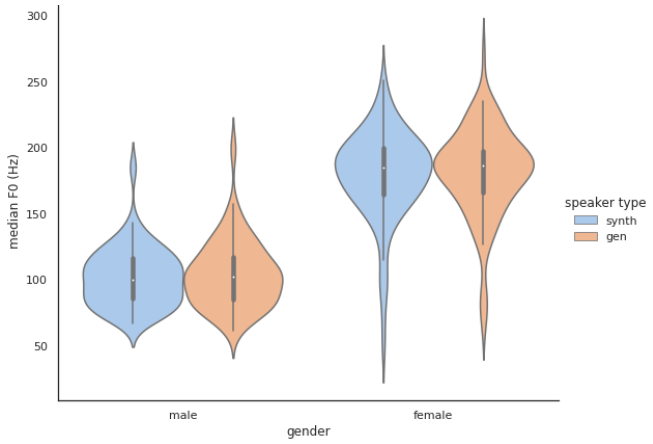


**Fig. 2**. Median $F_0$ of 200 training (synth) and 200 generated (gen) speakers for audio synthesized using the libriclean TacoSpawn model, split by gender.

### 5.3. Subjective evaluation results

#### *5.3.1. Subjective speaker similarity*

To examine the degree to which the metrics proposed in §4.2 capture speaker identity characteristics, we asked a pool of crowdsourced human reviewers to determine whether 1294 utterance pairs were uttered by the same speaker (yes or no). Each utterance was 3-5 seconds long, and either ground truth audio (t), synthesized training speaker audio (s), or synthesized generated speaker audio (g). Transcripts used for evaluation were unseen during training and typically different within a pair. The TacoSpawn model was trained on the enus1100 dataset and used a WaveRNN [25] vocoder.[4] We included ~220 of each of the six possible utterance pair types (g2g, g2s, s2s, g2t, s2t, t2t), and drew pairs from the entire range of d-vector scores. Figure 3 shows the correlation between d-vector cosine similarity and the average human ratings, and indicates that the d-vector–based metric captures some perceptual notion of speaker identity.

[4]This earlier experiment is the only result using a WaveRNN vocoder.

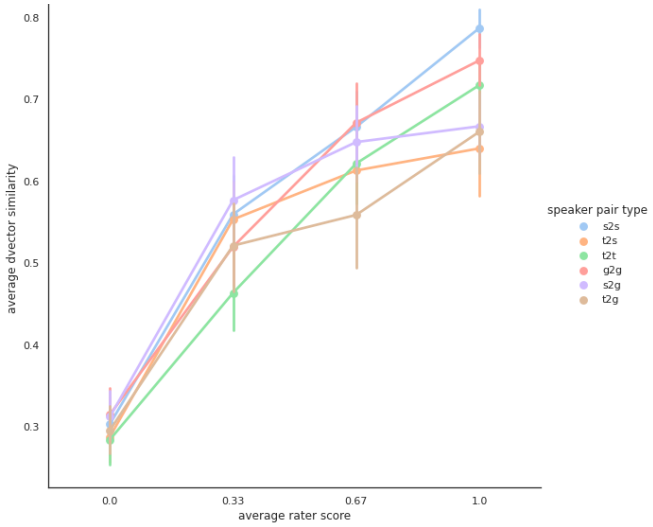


**Fig. 3**. Average correlation between d-vector cosine similarity ($y$ axis) and human "same speaker" ratings ($x$ axis) of pairs of US English utterances, broken down by speaker pair type. The $x$ axis ticks reflect the average of 3 boolean ratings per utterance pair.

| data | locale | spkrs | training speakers | generated speakers |
|---|---|---|---|---|
| libriclean | us | 200 | $3.37 \pm 0.14$ | $3.54 \pm 0.14$ |
| en1468 | au | 164 | $3.30 \pm 0.15$ | $3.03 \pm 0.14$ |
| en1468 | us | 300 | $3.68 \pm 0.11$ | $3.62 \pm 0.11$ |
| en1468 | gb | 212 | $3.69 \pm 0.12$ | $3.51 \pm 0.13$ |

**Table 2**. Speech naturalness mean opinion score (MOS) of utterances synthesized by systems trained on the libriclean and en1468 datasets, broken down by locale. Scores are shown for both training and generated speakers with 95% confidence intervals.

#### *5.3.2. Speech naturalness*

We measured the naturalness of speaker generation models in three English locales using subjective listening tests. For each model and locale, we randomly selected a fixed number of gender-balanced training and generated speakers, and synthesized one utterance per speaker. A pool of human reviewers rated the naturalness of these utterances on a scale from 1 (bad) to 5 (excellent) in increments of 0.5 [26, Figure A.2]. Each utterance received three independent ratings, and reviewers rated no more than ~25 utterances each. Results are shown in Table 2. Mean opinion scores (MOS) for training and generated speakers are similar, indicating that TacoSpawn voices achieve quality comparable to speakers in the training corpus. We encourage readers to visit our demo page to listen to audio samples.

## 6. ACKNOWLEDGMENTS

The authors would like to thank Rif A. Saurous, Ye Jia, Chun-an Chan, Rob Clark, Heiga Zen, and Yu Zhang for their helpful feedback and discussion.

## 7. REFERENCES


[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, [Online], Available at https://arxiv.org/abs/1609.03499.

[2] Sercan Arık, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in Neural Information Processing Systems*, 2017, pp. 2962–2970.

[3] Sercan Arık, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018.

[4] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, 2018.

[5] Henry B Moss, Vatsal Aggarwal, Nishant Prateek, Javier González, and Roberto Barra-Chicote, "Boffin TTS: Few-shot speaker adaptation by Bayesian optimization," in *Proc. ICASSP*, 2020, pp. 7639–7643.

[6] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu, "Adaspeech: Adaptive text to speech for custom voice," in *Proc. ICLR*, 2021.

[7] Seungwoo Choi, Seungju Han, Dongyoung Kim, and Sungjoo Ha, "Attentron: Few-shot text-to-speech utilizing attention-based variable-length embedding," in *Proc. Interspeech*, 2020, pp. 2007–2011.

[8] Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti, "SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model," in *Proc. Interspeech*, 2021, pp. 3645–3649.

[9] Erica Cooper, Jeff Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. ICASSP*, 2020, pp. 6184–6188.

[10] Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee, "Investigating on incorporating pre-trained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *Proc. ICASSP*, 2021.

[11] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014, pp. 4052–4056.

[12] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[13] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *Proc. Interspeech*, 2019.

[14] Kentaro Mitsui, Tomoki Koriyama, and Hiroshi Saruwatari, "Deep Gaussian process based multi-speaker speech synthesis with latent speaker representation," *Speech Communication*, vol. 132, pp. 132–145, 2021.

[15] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[16] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R.J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[17] Shuai Wang, Yanmin Qian, and Kai Yu, "What does the speaker embedding encode?," in *Proc. Interspeech*, 2017, pp. 1497–1501.

[18] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[19] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[20] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 271–279.

[21] Ben Poole, Alexander A. Alemi, Jascha Sohl-Dickstein, and Anelia Angelova, "Improved generator objectives for GANs," in *Proc. NIPS Workshop on Adversarial Training*, 2016.

[22] Matt Shannon, "Properties of f-divergences and f-GAN training," Tech. Rep., Google, 2020, Available at https://arxiv.org/abs/2009.00757.

[23] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz, "Translatotron 2: Robust direct speech-to-speech translation," 2021.

[24] Alain de Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[25] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, et al., "Efficient neural audio synthesis," in *Proc. ICML*, 2018, vol. 80, pp. 2410–2419.

[26] Eric Battenberg, Soroosh Mariooryad, Daisy Stanton, R.J. Skerry-Ryan, Matt Shannon, David Kao, and Tom Bagby, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," 2019, [Online], Available at https://arxiv.org/abs/1906.03402.