Interspeech 2024
1-5 September 2024, Kos, Greece

PitchFlow: adding pitch control to a Flow-matching based TTS model

Tasnima Sadekova, Mikhail Kudinov*, Vadim Popov*, Assel Yermekova, Artem Khrapov*

Huawei Noah's Ark Lab, Moscow, Russia

sadekova.tasnima@huawei.com, popov.vadim1@huawei.com

Abstract

In recent years, there have been various attempts to improve denoising diffusion probabilistic models and make them more suitable for real-world applications. One of the recent advances in this research direction is a flow-matching models framework which has already shown good results in image and speech generation tasks. Despite high quality and generation speed, flow-matching text-to-speech models still have problems with stability and control. To mitigate this issue, we propose two techniques: speaker scoring and pitch guidance allowing to control timbre and pitch contour of the generated speech. We show that the optimal choice of the prior leads to considerable improvement of similarity and a specific design of classifier guidance allows for fine-grained pitch control with high naturalness. Moreover, these techniques may be used to implement a voice conversion system of a competitive quality.

Index Terms: speech synthesis, voice cloning, voice conversion, flow matching, pitch control

1. Introduction

Recently, there has been remarkable progress in the development of text-to-speech (TTS) technology [1, 2, 3, 4, 5]. High speech quality and speaker similarity have been achieved in both single-speaker [1, 3, 5] and multi-speaker mode [6, 7, 8]. Recent models have shown an unprecedented quality in zero-shot voice cloning [9, 10, 11, 12, 13, 14]. Some of these models (e.g. [9, 13]) follow the LLM paradigm [15] and make use of large decoder-only Transformer architectures [16]. Others (e.g. [11, 12, 17, 18]) rely on either Denoising Diffusion Probabilistic Models [19] or more recent Flow-Matching [20] framework. The main advantage of such models is their ability to generate speech non-autoregressively and support content editing and denoising [12].

The most intriguing feature of such models though is their ability to copy not only the speaker's voice, but also the emotions and style even in cross-lingual mode [10, 12, 14]. This capability makes such models an especially appealing choice for speech translation [21] and book reading scenarios [22]. Although recent models trained on larger datasets [23] have shown such emergent capabilities that can be interpreted as accurate prediction of emotions and logical stress from text, the models described above are still incapable of reliable translation of focus and logical stress. As focus and stress are often related to pitch [22], accurate speech translation may require fine-grained pitch control.

There have been many models capable of conditioning on pitch curve [4, 6, 7, 24, 25], but in practice the output pitch may diverge from the conditioning or the output speech can have low sound quality. However, there are successful voice conversion models [26, 27] which can generate speech of high quality and keep the output pitch similar to the conditioning one. In [27], pitch fidelity for speech was implemented via optimal transport technique. In addition, authors used pitch classifier guidance [28] to improve the performance of optimal transport in musical instruments timbre transfer problem. To achieve this goal, gradient-guided inference was used with 4 control functions: (i) loudness; (ii) chromagram; (iii) pitch and (iv) probabilities given by the musical instrument classifier.

In this paper, we follow the same paradigm using specifically designed pitch classifier and a pre-trained speaker classifier to guide the generation process made by a flow-matching model. We show that using a speaker classifier for careful choosing the prior distribution sample used as a starting point for speech synthesis significantly improves speaker similarity. We implement a pitch classifier that can be successfully used during the intermediate steps of inference to significantly improve pitch fidelity. Finally, to demonstrate the quality of our pitch control method we implement a voice conversion system based on our zero-shot voice cloning model.

The paper is structured as follows: in Section 2 we give a brief overview of flow-matching framework for speech generation; we describe our method of pitch and speaker's timbre control in Section 3; our experiments are described in Section 4; we conclude in Section 5.

2. Flow matching speech modeling

Loosely speaking, flow matching [20] is a framework resembling diffusion probabilistic modeling [19] with the forward diffusion trajectories and training objectives specifically chosen to provide models with faster sampling. In what follows we briefly describe the general principles of flow matching and their application to speech synthesis.

Flow matching framework assumes that Gaussian noise is added to clean data samples X_1 belonging to data distribution $\mathrm{Law}(X_1)$, so that conditional distribution of noisy data samples $\mathrm{Law}(X_t|X_1)$ is Gaussian with the mean $\mu_t(X_1) = tX_1$ and the covariance matrix $\sigma_t^2(X_1)\,\mathrm{I}$ where I is the identity matrix of the same dimensionality as data X_1 and $\sigma_t(X_1) = 1-(1-\sigma_{min})t$ for some small positive number σ_{min}. It can be shown [20] that such stochastic process with conditional density functions $p_t(x|x_1)$ for $t \in [0, 1]$ can be generated by the following conditional vector field:

$$u_t(x|x_1) = \frac{x_1 - (1-\sigma_{min})x}{1-(1-\sigma_{min})t}\,. \tag{1}$$

Flow matching neural network v_θ estimates this vector field by

*Equal contribution

10.21437/Interspeech.2024-1023

minimizing the Mean Square Error (MSE):

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,X_1,X_t}\|v_\theta(X_t,t) - u_t(X_t|X_1)\|_2^2\,, \quad (2)$$

where t is sampled uniformly from $[0,1]$, X_1 is sampled from data distribution $\mathrm{Law}\,(X_1)$ and X_t is sampled from the conditional distribution $\mathrm{Law}(X_t|X_1)$. Optimizing conditional flow matching objective (2) can be shown to be equivalent to minimizing the MSE error between $v_\theta(X_t,t)$ and *unconditional* vector field $u_t(X_t)$ mapping prior distribution $\mathcal{N}(0,\mathrm{I})$ to the one that is very close to $\mathrm{Law}(X_1)$ (namely, to $\mathrm{Law}(X_1)$ convolved with $\mathcal{N}(0,\sigma_{min}^2\,\mathrm{I})$). Thus, sampling from the trained flow matching model consists of solving the following Ordinary Differential Equation (ODE) on the interval $t \in [0,1]$:

$$\dot{X}_t = v_\theta(X_t,t)\,, \quad (3)$$

where we denote the derivative of X_t with respect to time t by $\dot{X}_t$. This ODE is solved starting from $X_0 \sim \mathcal{N}(0,\mathrm{I})$. It was empirically shown that the trajectories of the ODE (3) are more straight than those of probability flow ODEs corresponding to common diffusion model types like Variance Exploding or Variance Preserving [19], which potentially makes it easier to solve the ODE (3) with sufficient quality and small number of steps.

Recently, flow matching has been employed by the text-to-speech model VoiceBox [12]. VoiceBox consists of a phoneme encoder, a flow matching network and a duration predictor. Given the prompt mel features, VoiceBox generates the target mel features X_1 jointly with those of the input speech prompt X^{prompt}. The VoiceBox flow matching network is also conditioned on the upsampled encoded phoneme sequences z. The prompt phoneme sequence is upsampled using the ground truth durations while the target phoneme sequence is upsampled using the outputs of duration predictor. At training, prompt conditioning X^{prompt} is replaced with zeros with certain probability ω to enable classifier-free guidance. Once the model is trained, we can sample from it by solving the following ODE:

$$\dot{X}_t = (1+\alpha)v_\theta(X_t, X^{prompt}, z, t) - \alpha v_\theta(X_t, 0, z, t) \quad (4)$$

for some classifier-free guidance weight $\alpha > 0$. It was shown empirically [12] that classifier-free guidance improves speaker similarity.

Throughout the remainder of the paper we will skip all additional inputs to the flow matching models in the notation for brevity and denote speech synthesis flow matching models just by $v_\theta(X_t,t)$ where X_t is the noisy mel-spectrogram at the flow matching time step t.

3. Pitch and timbre control

In this section we present techniques that allow to control pitch and timbre of speech generated by the flow matching model PitchFlow we propose.

3.1. Speaker scoring

As other diffusion-related text-to-speech models, PitchFlow is capable of synthesizing diverse speech. It turns out that the prior sample $X_0 \sim \mathcal{N}(0,\mathrm{I})$ has large impact on the timbre of the generated voice. Therefore, it may be useful to be able to increase speaker similarity between the generated speech and the prompt by choosing the optimal prior sample X_0. Next we propose the efficient way of doing it that we call *speaker scoring*.

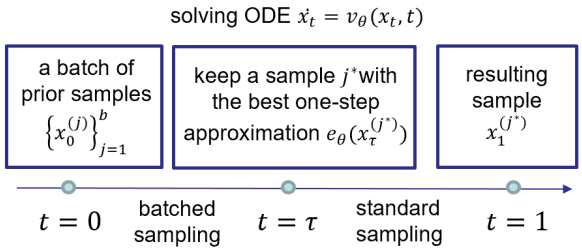

Figure 1: *Speaker scoring scheme.*

The basic idea is illustrated in Figure 1. First, we take a batch of prior samples $\{X_0^{(j)}\}_{j=1}^b$. Then we solve the ODE (3) up to time $t=\tau \in (0,1)$ by estimating vector field with the neural network v_θ in batch mode. At time $t=\tau$ we calculate some approximation of the generated mel-spectrogram $e_\theta(X_\tau^{(j)},\tau)$ for every $j=1,..,b$ and choose the best one in terms of speaker similarity to the input speech prompt. Speaker similarity can be estimated, for example, by converting mel estimations $e_\theta(X_\tau^{(j)},\tau)$ into raw audio by a vocoder and running a speaker verification network. After the optimal noisy sample $X_\tau^{(j*)}$ is chosen, we can continue solving the ODE (3) on the interval $[\tau,1]$ with the initial condition $X_\tau = X_\tau^{(j*)}$ in a standard mode, i.e. with batch size = 1.

The efficiency of the proposed algorithm depends on the values of τ and b: it is preferable to have rather small τ and b to reduce number of neural network evaluations in batch mode and batch size respectively. As for clean data estimation $e_\theta(X_t,t)$ we propose to use the following efficient formula requiring only one neural network evaluation in batch mode:

$$e_\theta(x,t) = (1-\sigma_{\min})x + (1-(1-\sigma_{min})t)v_\theta(x,t). \quad (5)$$

The formulae (1-2) imply that

$$v_{\theta^*}(x,t) = \mathbb{E}\left[\frac{X_1 - (1-\sigma_{min})x}{1-(1-\sigma_{min})t}\Big|X_t = x\right] \quad (6)$$

for the optimally trained neural network v_{θ^*}, thus

$$e_{\theta^*}(x,t) = \mathbb{E}[X_1|X_t = x]\,, \quad (7)$$

which explains why the formula (5) can be used to estimate clean data X_1 given its noisy version $X_t = x$. Also, since $\sigma_{min} \approx 0$, we have $e_\theta(x,t) \approx x + (1-t)v_\theta(x,t)$, so $e_\theta(x,t)$ can be seen as the solution of the ODE (3) with the initial condition $X_t = x$ in one step of the Euler method. It was found out empirically that for PitchFlow fair estimations $e_\theta(x,\tau)$ of clean mel-spectrograms could be obtained on rather early flow matching steps – for $\tau = 0.25$.

3.2. Pitch guidance

Classifier guidance [19] is a popular method of controllable generation with diffusion models. While it is widely used in conditional image generation, this method is not as popular for speech synthesis.

Controllability of a pre-trained diffusion model is achieved by exploiting a classifier $p_{\phi_t}(y|X_t)$ trained to predict the target label on a noisy sample X_t. Then gradients $\nabla_x \log p_{\phi_t}(y|X_t)$

are used to guide the diffusion backward pass towards the class label y.

In this work, we propose to apply this technique to pitch control in a flow-based speech generation model as shown in Figure 2. For this purpose, the frame level pitch is quantized in log-scale into 50 bins, one per frame. The classifier $c_\phi(p|X_t)$ is trained on noisy mel-spectrograms X_t with cross-entropy loss. The ground-truth pitch values were extracted with Praat $partselmouth$[1] library with autocorrelation method [29]. During sampling at each time step t we calculate the probability that X_t matches the target pitch contour $\hat{p}$ and solve the following ODE:

$$\dot{X}_t = v_\theta(X_t, t) + \alpha_t \nabla_x \log c_\phi(\hat{p}|X_t) \tag{8}$$

for some guidance weights α_t. In this setting, synthesized speech has a pitch contour which is very close to the target one thus providing fine-grained pitch manipulation, such as adding accented words. Combined with copying style from the reference audio and manipulations with input phoneme durations, this approach provides almost full control of speech prosody.

Furthermore, the proposed method provides voice conversion capabilities to PitchFlow model. To implement voice conversion feature, it is sufficient to take the following steps:

1. Run an aligner[2] model on the source recording S_s to extract phoneme sequence T_s and phoneme durations D_s;
2. Extract pitches P_s and P_t from the source and target recordings S_s and S_t and calculate mean-variance statistics (μ_s, σ_s^2) and (μ_t, σ_t^2). Re-normalize source pitch as $\hat{P}_s = \frac{P_s - \mu_s}{\sigma_s}\sigma_t + \mu_t$;
3. Generate speech from phonemes T_s using prompt S_t, durations D_s and pitch guidance with target $\hat{P}_s$.

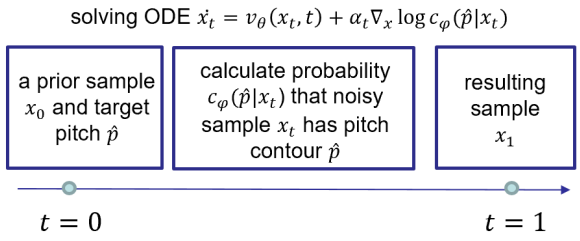

Figure 2: *Pitch guidance scheme.*

However, such pitch normalization method is not always robust [30], the possible reasons are inaccuracy of the pitch extraction algorithm or non-robustness of speaker statistics for short audios. To tackle this problem, we find the optimal pitch shift during inference. We apply an approach similar to the speaker scoring method described in Section 3.1, i.e. we prepare a batch of target pitch values $\hat{P}_s^{(j)} = \{\hat{P}_s c_j\}_{j=1}^{7}$, where $c_j \in [0.85; 1.15]$ is a coefficient taken with step 0.05. Then the ODE (8) is solved up to time $t = \tau_{pg}$ in batch mode and raw audios are synthesized from the approximation of mel-spectrogram $e_\theta(X_{\tau_{pg}}^{(j)}, \tau_{pg})$ for every j as in Section 3.1. Only one target pitch variant $\hat{P}_s^* = \hat{P}_s^{(j*)}$ is selected for further sampling based on speaker similarity obtained from the speaker verification model. Our experiments have shown that $\tau_{pg} = 0.125$ is enough to achieve a speaker similarity improvement in pitch-guided sampling.

[1]https://parselmouth.readthedocs.io/

[2]https://montreal-forced-aligner.readthedocs.io/

Table 1: *Zero-shot voice cloning, subjective evaluation. MOS scores are given with 95% confidence intervals.*

MOS	Naturalness	Similarity
PitchFlow-32x8	$\mathbf{4.49 \pm 0.09}$	$\mathbf{4.02 \pm 0.07}$
PitchFlow-32x1	4.46 ± 0.09	3.91 ± 0.08
VALL-E [9]	4.14 ± 0.12	3.88 ± 0.09
YourTTS [7]	4.30 ± 0.12	3.56 ± 0.08
Ground Truth	4.59 ± 0.12	4.34 ± 0.07

4. Experiments

Our model, *PitchFlow*, consists of Duration predictor, Phoneme encoder and Flow-based decoder. The decoder architecture is similar to the one proposed in [12] except that we don't use skip-connections between Transformer blocks to reduce computational burden at training. Duration predictor has the same architecture as in [4]. The phoneme encoder produces embeddings of size 128. It includes 3 layers of 1-d convolutions with ReLU activations and kernel size 5 followed by 8 layers of Transformer encoder blocks with 8 heads and $4 * 128$ hidden size of feed-forward layer of Transformer block. The model was trained on LibriTTS [31] and a part of Librilight [32] datasets. The training set included about 6 mln recordings. The model was trained for about 1.2 mln iterations with batch size 32 with Adam optimizer on 8 NVIDIA V100 GPUs. Classifier-free guidance parameters ω and α were set to 0.3 and 0.7 respectively.

The pitch classifier model c_ϕ consists of 4 convolutional blocks with residual connections followed by a projection layer. Each of the blocks is represented by 2 stacks with a convolutional layer, layer normalization and GeLU activation. The kernel size is set to 5 and hidden size to 256. The classifier is trained on the LibriTTS dataset for 750k steps with batch size 22 and Adam optimizer on 1 NVIDIA V100 GPU.

The universal Hifi-GAN [33] vocoder was used to transform mel-spectrograms into waveforms. It was trained on the LibriTTS dataset with sample rate 16kHz.

4.1. Zero-shot voice cloning with speaker scoring

We test speaker scoring method proposed in Section 3.1 on zero-shot voice cloning task with emotional prompts from 10 English speakers from ESD dataset [34]. We chose $\tau = 0.25$ as it provided good clean data estimation according to the formula (5) and $b = 8$. To calculate speaker similarity between prompt and generated speech estimation at time $t = \tau$, we convert mel-spectrograms to raw audio signals with HiFi-GAN and use WavLM-Base for Speaker Verification model [35] to calculate cosine similarity. Sampling from PitchFlow was done by solving ODEs with 32 steps of the explicit midpoint method. With such parameters, speaker scoring increased sampling time by as little as 50%.

VALL-E [3] [9] and YourTTS [4] [7] were chosen as baselines. We generated 10 out of unique 20 sentences for each of 10 ESD speakers with each model under comparison resulting in 100 synthesized speech samples for each model. We measured objective metrics such as Word Error Rate (WER) and cosine similarity as well as subjective metrics – 5-point scale Mean Opinion Scores

[3]https://github.com/lifeiteng/vall-e

[4]https://github.com/coqui-ai/TTS

Table 2: *Zero-shot voice cloning, objective evaluation. Cosine similarity is given with 95% confidence intervals.*

	Cosine similarity	WER
PitchFlow-32x8	$\mathbf{0.910 \pm 0.008}$	10.7%
PitchFlow-32x1	0.887 ± 0.010	**9.7%**
VALL-E [9]	0.900 ± 0.008	14.8%
YourTTS [7]	0.906 ± 0.009	14.1%
Ground Truth	0.910 ± 0.007	—

(MOS) for speech naturalness and speaker similarity. Whisper[5] model was used to compute WER and cosine similarity was calculated with the WavLM model. Subjective listening tests were conducted on the platform Appen. Every synthesized speech sample was evaluated by 3 assessors (out of unique 54 workers) to obtain naturalness scores and 5 assessors (out of unique 112 workers) to obtain similarity scores. The results of subjective and objective evaluation are given in Tables 1 and 2 respectively.

Baseline models VALL-E and YourTTS suffer from pronunciation problems when trying to copy style of highly emotional speech prompts of ESD speakers which is reflected in higher WER scores than those of PitchFlow models in Table 2. Also, this table demonstrates that samples produced by PitchFlow with speaker scoring denoted as *PitchFlow-32x8* are comparable with those generated by VALL-E and YourTTS and better than the ones generated without this technique (denoted as *PitchFlow-32x1* and similar to the VoiceBox model) in terms of cosine speaker similarity. Which is more important, *PitchFlow-32x8* outperforms *PitchFlow-32x1* and other baselines in terms of speaker similarity through subjective evaluation as shown in Table 1 while maintaining high naturalness scores of around 4.5. These results prove that speaker scoring is a reliable method that helps to deal with negative effects that starting prior samples may have on speaker similarity with the input prompt.

4.2. Zero-shot voice conversion with pitch guidance

To evaluate PitchFlow zero-shot conversion capability we prepared a subjective test. We took 3 sentences for 5 ESD speakers with 2 different emotions as source samples and 10 VCTK speakers for target voice. 160 random audios were selected for the final test.

Two models were selected as the baselines: (1) BNE-PPG-VC[6] [26], any-to-any voice conversion system with pitch encoder and pretrained speaker encoder; (2) YourTTS, a hybrid model capable of solving both cloning and conversion tasks. As for our models, we evaluated PitchFlow with pitch guidance with and without speaker verification model. Pitch guidance weight was selected according to the linear schedule $\alpha_t = 5 \cdot (1 - t)$ and sampling was performed with 32 steps of the Euler solver. The test was conducted on Appen platform. Every speech sample was assessed 3 times on naturalness (35 unique workers) and 5 times on similarity (109 unique workers) to obtain 5-point MOS scores.

Results are presented in Table 3. All models achieve very close speaker similarity values. Additional pitch shifting with the speaker verification model *PitchFlow + SV* improves base *PitchFlow* similarity score. Our model outperforms YourTTS baseline in terms of naturalness, working on par with BNE-PPG-VC.

[5]https://github.com/openai/whisper

[6]https://github.com/liusongxiang/ppg-vc

Table 3: *Zero-shot voice conversion, subjective evaluation. MOS scores are given with 95% confidence intervals.*

MOS	Naturalness	Similarity
PitchFlow	4.06 ± 0.05	3.38 ± 0.1
PitchFlow + SV	4.09 ± 0.05	$\mathbf{3.47 \pm 0.1}$
BNE-PPG-VC [26]	4.03 ± 0.05	3.41 ± 0.1
YourTTS [7]	3.93 ± 0.06	3.41 ± 0.1

Table 4: *Normalized cross-correlation (NCC) with 95% confidence intervals.*

PitchFlow	BNE-PPG-VC [26]	YourTTS [7]
$\mathbf{0.90 \pm 0.02}$	0.73 ± 0.02	0.66 ± 0.02

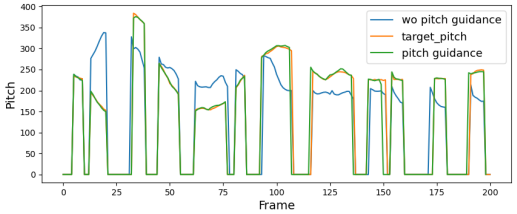

Figure 3: *Pitch guidance with manually changed target pitch.*

To demonstrate that pitch guidance allows to preserve style of a source sample we measure an objective metric - normalized cross-correlation (NCC) between pitch contours extracted from the source speaker audio and the synthesized speech. 150 emotional samples were evaluated with this metric and results are presented in Table 4. Despite the fact that BNE-PPG-VC model shows a high score, PitchFlow outperforms it in terms of NCC in the pitch transfer task.

The proposed technique can be used in text-to-speech systems with fine-grained pitch control – for instance, for supporting logical stress in generated speech. Figure 3 illustrates PitchFlow's ability to exactly follow the target pitch, modified manually. More examples can be found on demo page `https://pitch-flow.github.io` along with the audios used in subjective tests.

5. Conclusion

We have proposed a flow-based text-to-speech model with pitch control capabilities based on pitch classifier guidance, and have shown that speaker similarity of such models can be improved by means of careful choosing the prior distribution sample based on speaker classifier score. Our model has achieved state-of-the-art results in zero-shot text to speech and has performed on a competitive level in voice conversion mode. Combining the model with a reliable pitch generation model will be a promising direction of future research.

6. References

[1] J. Shen, R. Pang *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2017.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural Speech Synthesis with Transformer Network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 07 2019, pp. 6706–6713.

[3] J. Kim, S. Kim *et al.*, "Glow-TTS: a generative flow for text-to-speech via monotonic alignment search," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20, 2020.

[4] Y. Ren, Y. Ruan, X. Tan, T. Qin *et al.*, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 3171–3180.

[5] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8599–8608.

[6] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He, and S. Zhao, "DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021," pp. 80–86, 10 2021.

[7] E. Casanova, J. Weber *et al.*, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162. PMLR, 2022.

[8] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, 2021.

[9] C. Wang, S. Chen *et al.*, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," *ArXiv*, 2023.

[10] Z. Zhang, L. Zhou *et al.*, "Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling," *ArXiv*, 2023.

[11] K. Shen, Z. Ju *et al.*, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *The Twelfth International Conference on Learning Representations*, 2024.

[12] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W. Hsu, "Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, December 10 - 16*, 2023.

[13] Z. Jiang, Y. Ren *et al.*, "Mega-TTS: Zero-shot text-to-speech at scale with intrinsic inductive bias," *ArXiv*, 2023.

[14] Z. Jiang, J. Liu, Y. Ren, J. He, Z. Ye, S. Ji, Q. Yang, C. Zhang, P. Wei, C. Wang, X. Yin, Z. MA, and Z. Zhao, "Mega-TTS 2: Boosting prompting mechanisms for zero-shot speech synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[15] T. B. Brown, B. Mann, N. Ryder *et al.*, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Curran Associates Inc., 2020.

[16] A. Vaswani, N. Shazeer *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[17] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-tts: A fast tts architecture with conditional flow matching," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 341–11 345.

[18] S. Kim, K. Shih, J. F. Santos, E. Bakhturina, M. Desta, R. Valle, S. Yoon, B. Catanzaro *et al.*, "P-flow: A fast and data-efficient zero-shot tts through speech prompting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," in *International Conference on Learning Representations*, 2021.

[20] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[21] Seamless Communication, "SeamlessM4T: Massively Multilingual & Multimodal Machine Translation," *ArXiv*, 2023.

[22] D. Diatlova and V. Shutov, "EmoSpeech: guiding FastSpeech2 towards Emotional Text to Speech," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[23] M. Łajszczak, G. Cámbara *et al.*, "BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data," *ArXiv*, 2024.

[24] A. Łańcucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[25] Z. Shaheen, T. Sadekova *et al.*, "Exploiting Emotion Information in Speaker Embeddings for Expressive Text-to-Speech," in *Proc. INTERSPEECH 2023*, 2023, pp. 2038–2042.

[26] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[27] V. Popov, A. Amatov *et al.*, "Optimal Transport in Diffusion Modeling for Conversion Tasks in Audio Domain," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[28] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[29] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the Institute of Phonetic Sciences*, 1993.

[30] D. T. Chappell and J. H. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, vol. 2. IEEE, 1998, pp. 885–888.

[31] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech*, 2019.

[32] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-Light: A Benchmark for ASR with Limited or No Supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[33] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, virtual*, 2020.

[34] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[35] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.