INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Normalization Driven Zero-shot Multi-Speaker Speech Synthesis

Neeraj Kumar[1,2], *Srishti Goel*[1], *Ankur Narang*[1], *Brejesh Lall*[2]

[1]Hike Private Limited, India
[2]Indian Institute of Technology, Delhi, India
neerajku@hike.in, srishtig@hike.in, ankur@hike.in,brejesh@ee.iitd.ac.in

Abstract

In this paper, we present a novel zero-shot multi-speaker speech synthesis approach (ZSM-SS) that leverages the normalization architecture and speaker encoder with non-autoregressive multi-head attention driven encoder-decoder architecture.Given an input text and a reference speech sample of an unseen person, ZSM-SS can generate speech in that person's style in a zero-shot manner. Additionally, we demonstrate how the affine parameters of normalization help in capturing the prosodic features such as energy and fundamental frequency in a disentangled fashion and can be used to generate morphed speech output. We demonstrate the efficacy of our proposed architecture on multi-speaker VCTK[1] and LibriTTS [2] datasets, using multiple quantitative metrics that measure generated speech distortion and MOS, along with speaker embedding analysis of the proposed speaker encoder model.

Index Terms: Speech synthesis, normalization, transfer learning, wav2vec2.0 based speaker encoder, angular softmax

1. Introduction

A lot of exciting developments have been made in speech synthesis systems to synthesize natural sounding human speech. The developments in this area have helped in several applications including audiobook narration, news readers, conversational assistants and engaging user experiences in the virtual worlds.

To realize a natural speech synthesis system, the model has to capture the speaking style of every person. For this prosodic features of speech play an important role. Prosody is a confluence of many phenomena such as paralinguistic information, intonation, stress, and style. Such phenomena are best described by the duration, fundamental frequency and energy of any speech. Multiple efforts are being made to incorporate and control such features into the model to capture and synthesize the speech in a person's speaking style.

High-quality multi-speaker speech synthesis (with prosody consideration) in a zero-shot manner is an interesting and challenging research problem. Present approaches for state-of-the-art TTS (Text to Speech Synthesis) such as Tacotron [3], Fast Speech [4], Fast speech 2 [5] have focused on generating the speaking style of a single speaker. These approaches do not generate audio on multiple speakers. Some of the current approaches [6, 7, 8, 9, 10, 11, 12] have used speaker embedding to capture the identity and speaking style of the person in the speech. Such approaches fail to generate expressive speech as they have not taken the prosodic features and emotions into account and hence have lower quality in generated speech. While some of the approaches [13, 14, 15, 16, 17] rely on prosodic features such as fundamental frequency, duration and energy to generate the expressive speech but these approaches can generate expressive speech for the speakers who are already part of the training. Such approaches are not able to generate expressive speech in a zero-shot manner on multiple speakers.

Some of the recent approaches such as [14] proposed "global style tokens" (GSTs), a bank of embeddings that are jointly trained within Tacotron and learn to model a large range of acoustic expressiveness. Adaspeech[18] uses the acoustic conditioning model and conditional layer normalization in [5] model for incorporating the speaker embedding [19] to adapt the model on custom voices in few shot approach . Such approaches are not able to capture the prosody of unseen reference speech in zero shot manner.

We have proposed a novel zero-shot approach (ZSM-SS) that uses a normalization architecture along with a non-autoregressive transformer-based architecture [5]. ZSM-SS can generate multi-speaker speech output in a zero-shot manner, given an input unseen text and an unseen person's reference speech sample.For normalization, we have proposed two architectures based on convolution and on multi-head attention to learn the prosodic properties in the network through affine parameters. This helps to capture the various affine parameters based on speaker embedding, pitch and energy and generates personalized and temporally smoother speech that incorporate the speaking style of a person.For generating the speaker embedding we have proposed the speaker encoder architecture which uses the XLSR architecture [20] which is pretrained on multiple languages in wav2vec2.0 framework[21]. The speaker encoder leverages the feature encoder of pretrained XLSR to generate the 256-dimensional speaker embedding. Experimental sections have shown the effectiveness of this speaker encoder.

Using extensive experiments on multi-speaker VCTK [1] and LibriTTS[2] datasets, we show both qualitative and quantitative results along with high-quality output and the capability of our approach to generate speech for a wide variety of unseen speakers. ZSM-SS can also be used as a voice morphing tool by varying the embedding, frequency and energy inputs to the normalization module.

2. ZSM-SS Design

In this section, we present the overall design and architecture of ZSM-SS including speaker encoder, normalization architecture and non-autoregressive multi-head attention based encoder decoder based architecture for zero shot multi-speaker speech synthesis.

2.1. Architecture

Fig. 1 illustrates the architecture used in ZSM-SS. During training, it takes as input: the text-audio pairs of a person along with his(her) reference speech samples. During inference, it takes a zero unseen text-audio pairs along with one reference speech sample on an unseen speaker, to generate speech in that person's speaking style. The normalization architecture is applied both during encoder and decoder stages (Fig. 1) and hence helps in prosody transfer in a zero shot manner.

Feed-Forward Transformer The architecture of Feed-

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-441

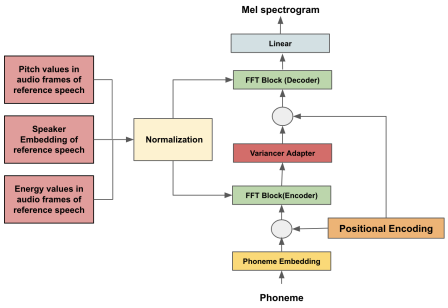

Figure 1: *ZSM-SS : Zero Shot style based text to Speech Generation Architecture*

Forward Transformer(FFT) (Fig. 2) is based on a multi-head self-attention network, and position feed-forward network which consists of two Conv1D and normalization stages. The proposed method stacks multiple FFT blocks with phoneme embedding and position encoding as an input at the encoder side(Fig. 1), and multiple FFT blocks with position encoding and output from variance adapter for the mel-spectrogram generation at the decoder side(Fig. 1).

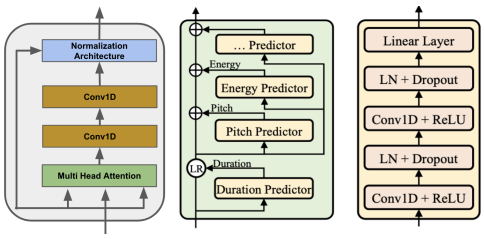

Figure 2: *Left: FFT block , Centre : Variance Adapter[5] , Right : Variance Predictor [5]*

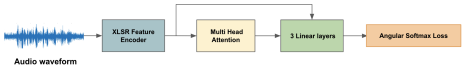

Figure 3: *Speaker Encoder*

Speaker Encoder We have used the pretrained XLSR[20] which is the wav2vec2.0 [21] model and is trained in multiple languages to learn the cross-lingual representation. The speech encoder uses a feature encoder of pretrained XLSR which consists of several blocks containing a temporal convolution followed by layer normalization [22] and a GELU activation function[23]. The multi-head attention layer is used on top of the pretrained feature encoder followed by 3 linear layers. Angular softmax loss[24] is used which aims to achieve smaller maximal intra-class distance than minimal inter-class distance. Angular softmax loss is given by:

$$L_{\text{AL}} = \frac{1}{N}\sum_{i} -log\left(\frac{e^{\|x\|\psi(m\theta_{\text{y}_\text{i},\text{i}})}}{e^{\|x\|\psi(m\theta_{\text{y}_\text{i},\text{i}})} + \sum_{j\neq y_\text{i}} e^{\|x\|cos(m\theta_{\text{y}_\text{i},\text{i}})}}\right) \tag{1}$$

where we define $\psi(m\theta_{\text{y}_\text{i},\text{i}}) = (-1^k)cos(m\theta_{\text{y}_\text{i},\text{i}}) - 2k$ ad $\theta \in [\frac{k\pi}{m}, \frac{k+1\pi}{m}]$ and $k \in [0, m-1]$. $m \geq 1$ is an integer that controls the size of angular margin. We have used $m = 1.35$ for this experiment.

Normalization Architecture The proposed normalization architecture consists of a general framework that is applied at the encoder and decoder side to capture the speaker and prosodic features with various learnable parameters such as γ (scale)and β(bias). The various learnable parameter corresponding to speaker embedding from speaker encoder, pitch and energy are computed through two proposed approaches: based on convolution network and multi-head attention network. This helps in adjusting the bias and scale of the normalized features to learn the required properties of speech signals including prosody. Given an input x with batch size k, $B\{x_1 \cdots x_k\}$, each sample x_i whose flatten format is $\{x_{i,1} \cdots x_{i,m}\}$, layer normalization is defined by (Equation 3).

1.Convolution based Normalization The three audio features(speaker embedding, fundamental frequency and energy of the reference speech sample) are passed into the convolution layer to generate the affine parameters (Fig. 4). The parameter ρ is used to combine these parameters (Equation 4). The value of ρ is constrained to the range of [0, 1] simply by imposing bounds at the parameter update step. We employ a residual connection (x = z + Sublayer(z)) around convolution layers(Sublayer(z)), followed by layer normalization with learnable scale and bias [22], The other part of this equation has layer normalization (Equation 3) having γ_{SE} and β_{SE} coming from speaker embedding.ρ helps the model to learn the speaker embedding information as a scale and shift in the normalization in proportion to the learnable scale and shift from the first part of Equation 4. The learnable scale and shift generated from the first part of Equation 4 will help models to learn the various factors and correlation and higher-order statistics in text and speech for speech style. In the case of pitch and energy, we have multiplied the respective scale and shift so that it learns both the nuances rather than in proportion(Equation 5).

$$\mu_\text{i} = \frac{1}{m}\sum_{j=1}^{m} x_{ij}, {\sigma_\text{i}}^2 = \frac{1}{m}\sum_{j=1}^{m}(x_{ij} - \mu_\text{i})^2 \tag{2}$$

$$\hat{x}_{i,j} = \frac{x_{ij} - \mu_\text{i}}{\sqrt{\sigma_i^2 + \epsilon}} \tag{3}$$

$$\hat{y}_{ij} = \rho(\gamma_{\text{LN}}\hat{x}_{ij} + \beta_{\text{LN}}) + (1-\rho)(\gamma_{\text{SE}}\hat{x}_{ij} + \beta_{\text{SE}}) \tag{4}$$

$$Output = \gamma_{\text{energy}}(\gamma_{\text{pitch}}\hat{y}_{ij} + \beta_{\text{pitch}}) + \beta_{\text{energy}} \tag{5}$$

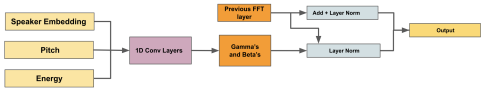

Figure 4: *Convolution based normalization in proposed ZSM-SS architecture*

2.Multi Head Attention Based Normalization In this architecture (Fig. 5), we have concatenated the speaker embedding (256 dimensional vector), frequency and energy of the reference speech sample to generate a tensor of size (batches* audio-frames * 258). This is then fed it into the multi-head attention network [25] to generate the affine parameters. These affine parameters (Equation 6) are used to bias $\beta_{\text{attention}}$ and scale $\gamma_{\text{attention}}$ the output feature map coming from previous FFT block (Fig. 2).

$$Output = \rho(\gamma_{\text{LN}}\hat{x}_{ij} + \beta_{\text{LN}}) + (1-\rho)(\gamma_{\text{attention}}\hat{x}_{ij} + \beta_{\text{attention}}) \tag{6}$$

The speaker embeddings, frequency and energy are concatenated and passed to the linear layer independently to become query, key and values of multi-head attention layer. The multi-head attention performs the scaled dot-product attention(Equation 8) with h heads by linearly projecting the query,

key and values h times with different learned projections to d_k, d_k and d_v dimensions, respectively in parallel. The multi-head attention [25] equation is given by:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1 \cdots \text{head}_\text{h})W^O \quad (7)$$

where $\text{head}_i = \text{Attention}(\text{QW}_i^Q, KW_i^K, VW_i^V)$ and the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_\text{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_\text{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_\text{model} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_\text{model}}$.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (8)$$

In this work, we have used $h = 2$ parallel attention heads. $d_k = d_v = d_\text{model}/h = 129$.

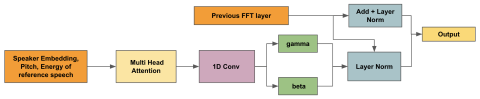

Figure 5: *Multi-Head Attention Based Normalization in proposed ZSM-SS architecture*

Variance Adapter The variance adapter consists of three **variance predictors** namely: the duration predictor, pitch predictor and energy predictor. The variance predictor is used to predict the prosodic features of speech such as duration, fundamental frequency and energy. For the duration predictor, the output is the length of each phoneme in the logarithmic scale and is optimized with mean square error loss. For pitch and energy predictor, the output is the frame-level fundamental frequency and energy of melspectrogram respectively and use mean absolute error to optimize.

3. Experiments

Datasets We train and evaluate the model on two datasets namely VCTK [1] and LibriTTS multi-speaker dataset [2]. We have used 44 hours of speech with 108 speakers of the VCTK dataset and 586 hours of speech with 2456 speakers of the LibriTTS dataset.

Training and Preprocesing Steps We convert the text sequence into the phoneme sequence [26, 3] using open-source grapheme-to phoneme tool[27] We extract the phoneme duration with MFA [27], an open-source system for speech-text alignment to improve the alignment accuracy. We extracted the pitch contour, F_0 using PyWorldVocoder tool[28]. and quantized each frame to 256 possible values. We transfer the raw waveform into melspectrograms by setting the frame size and hop size to 1024 and 256 with respect to the sample rate of 22050 Hz. The proposed speaker encoder uses the raw audio waveform sampled at 16000Hz and normalized to zero mean and unit variance and generates the 256 dimensional speaker embedding. The pretrained feature encoder of XLSR model is taken from [29].

We have trained the model on 4 V100 GPU based machines for around 3 days with the batch size of 64 for around $300k$ steps.The Adam optimizer is used with $\beta 1 = 0.9$, $\beta 2 = 0.98$, $\epsilon = 10e - 9$.

Model Configuration We have used 4 feed forward transformer blocks at the phoneme encoding stage and at the output mel-spectrogram decoder stage. The dimension of phoneme embedding and hidden layer of self attention is set to 256 in every FFT block. The number of attention heads is set to 2. The output linear layer converts the 256-dimensional hidden states into 80-dimensional mel-spectrograms. The size of the phoneme vocabulary is 76, including punctuations.

The speaker encoder uses 512 dimensional feature encoder of pretrained XLSR model which consists of several blocks containing a temporal convolution followed by layer normalization [22] and a GELU activation function[23]. The feature encoder contains seven blocks and the temporal convolutions [30] in each block have 512 channels with strides (5,2,2,2,2,2,2) and kernel widths (10,3,3,3,3,2,2). This results in an encoder output frequency of 49 Hz with a stride of about 20ms between each sample, and a receptive field of 400 input samples or 25ms of audio. The multi-head attention layer uses 2 attention heads with dropout 0.1 after feature encoder. 3 linear layers are used after multi head attention layer with dimensions($512 \rightarrow 256 \rightarrow n_c$) respectively where n_c is the number of speakers.

The Convolution based normalization architecture feeds 256 dimensional speaker embedding into 1D convolution layer to generate affine parameters. The fundamental frequency and energy of the reference speech are fed to 1D convolution layers each to reduce the channel length from max frames of speech signal in the dataset to 512. The affine parameters are then calculated by adding 1D convolution layer to generate 256 channel output respectively.

In multi-head attention based normalization architecture, the 256 dimensional speaker embedding is replicated along the time frame of the mel spectrogram and then concatenated with frequency and energy features to generate 258 dimensional feature vectors for all time steps (audio frames). It is then fed to multi head attention with the number of heads set to 6. The generated feature map is then fed to 1D convolution to generate 256 channel output which is added with output of layer normalization using the learnable parameter ρ.

The Variance predictor consists of 2 blocks of Conv1D, ReLU, layer normalization and dropout layer. The kernel sizes of the 1D-convolution is set to 3, with input/output sizes of 256/256 for both layers and the dropout rate is set to 0.5.The generated melspectrogram is optimised with mean square error loss. The pretrained Wave Glow architecture waveglow is used as a vocoder to generate the speech at 22050 Hz.

3.1. Analysis of Speaker embedding generated by Speaker Encoder

We have applied t-SNE and cosine similarity on 256 embedding of test speakers to show that the embeddings generated by speaker encoder lie closer in the embedding space for same speaker and far for different speaker. In Figure 6, the male(6 to 10) and female speakers(0 to 5) are closer to each other respectively and far from male and female speakers.Right part of Figure 6 shows that the higher similarity of embedding on actual and generated speech for same speaker.

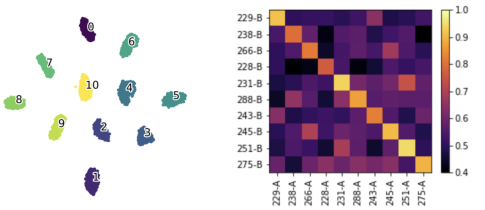

Figure 6: *Left:t-SNE visualization of speaker embeddings generated from speaker encoder model. Cluster id 0 to 5 refers to female speakers and 6 to 10 refers to male speakers, Right: Cross Similarity*

3.2. Audio Quality

Twenty samples of speakers with different accents are taken for VCTK testset and Twenty samples of english speaking speakers from LibriTTS are used to evaluate the generated samples in terms of naturalness(how the synthesized voices sound natural like human), similarity(how the synthesized voices sound similar to the reference speaker in terms of speaking style). The text content is kept consistent among different systems so that all testers only examine the audio quality without other interference factors. We conduct human evaluations with MOS (mean opinion score) for naturalness and SMOS(similarity MOS) for similarity and average the MOS and SMOS scores of multiple adapted speakers as the final scores as shown in Table 1.The NVC[10] has the MOS and SMOS score of 3.12 ± 0.33 and 3.01 ± 0.8 respectively which used 128 dimensional speaker embedding in Transformer based TTS in few shot approach. The generated outputs from ZSM-SS architecture are given in [1]

Apart from subjective evaluation, we have used the metrics namely Gross Pitch Error(GPE) [31], Voicing Decision Error(VDE) [31], F0 Frame Error(FFE) [32], Mel Cepstral Distortion(MCD) [33] which are used in audio signal processing to measure the prosody of the signal. Table 2 shows that the convolution-based normalization has lower errors compared to multi-head attention based normalization. [13] has lower MCD(10.87) due to the use of autoregressive(tacotron) based encoders to capture the pitch and speaker embedding whereas ZSM-SS uses pitch, energy and speaker embedding of reference speech though proposed normalization methods.

Table 1: *MOS and SMOS score on ZSM-SS with 95% confidence interval for VCTK and LibriTTS dataset. GT - ground Truth , GTmel - groung truth mel spectrogram with waveglow as vocoder, Conv - Convolution based normalization with waveglow , Attention - Multi head attention based normalization with waveglow*

Metric	Method	VCTK↑	LibriTTS↑
MOS	GT	3.85 ± 0.14	3.70 ± 0.08
	GTmel	3.80 ± 0.04	3.62 ± 0.13
	Conv	3.41 ± 0.42	3.19 ± 0.10
	Attention	3.38 ± 0.24	3.15 ± 0.09
SMOS	GT	3.91 ± 0.04	3.80 ± 0.24
	GTmel	3.83 ± 0.12	3.69 ± 0.21
	Conv	3.55 ± 0.16	3.28 ± 0.14
	Attention	3.50 ± 0.12	3.24 ± 0.08

Table 2: *Quantitative metrics on ZSM-SS on zero-shot approach. Conv: Convolution based normalization in ZSM-SS , Attention : Multi head attention based normalization in ZSM-SS, 1- VCTK dataset, 2- LibriTTS dataset*

Method	MCD↓	GPE↓	VDE↓	FFE↓
Conv-1	**12.70**	**27.75**	**17.01**	**34.43**
Attention-1	14.31	30.13	19.59	37.46
Conv-2	**11.21**	**30.49**	**19.91**	**38.28**
Attention-2	15.17	32.51	21.06	41.64

3.3. Ablation Study

Analysis of Normalization architecture We have done the ablation study on Convolution-based normalization framework with normalization in ZSM-SS with a zero-shot approach on the VCTK dataset. The base model(BM) with pitch and energy in the normalization stage with normalization without speaker embedding of reference speech has not shown very good results as the information of speaker identity is missing in the architecture. We have then used speaker embedding in the normalization steps with the base model and do not incorporate pitch and energy values of the reference unseen speaker. The quality of output degrades as the variance predictor is not able to predict the required duration, frequency, and energy values. The addition of pitch values along with speaker embedding of reference speech helps in improving the speech quality. Table 3 shows the better MOS and SMOS with the addition of normalization architecture.

Table 3: *Mean Opinion Score for the naturalness(N) and similarity(S) of ZSM-SS. BM is Base Model without normalization method. SE is speaker embedding in normalization, P and E are pitch and energy values in the normalization network.*

Method	MOS↑	SMOS↑
BM+P+E	2.98 ± 0.15	2.90 ± 0.11
BM+SE	3.12 ± 0.15	3.11 ± 0.02
BM+SE+P	3.20 ± 0.07	3.30 ± 0.09
ZSM-SS	$\mathbf{3.41 \pm 0.42}$	$\mathbf{3.55 \pm 0.16}$

3.4. Extension of Proposed Method

Voice morphing We can independently tune the speaker embedding, fundamental frequency and energy of the reference speech which are fed into the normalization steps to generate the morphed speech. Figure 7 shows that independently modulating the pitch and energy values leads to the voice morphing. This has a lot of applications in the virtual world, the gaming industry, voice modulation, etc.

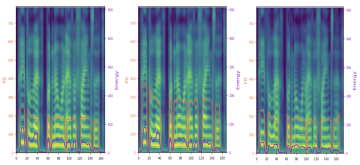

Figure 7: *Left: Synthesized samples on a reference speaker and text, Centre: Pitch is modulated by increasing the F0 to 1.25F0 keeping the energy values constant on same reference speaker and text Right: Energy values are reduced from E to 0.5E keeping the pitch values constant on same reference speaker and text*

4. Conclusions

In this paper, we have proposed a novel zero-shot approach (ZSM-SS) that uses a normalization architecture and speaker encoder along with a non-autoregressive feed-forward transformer-based architecture. ZSM-SS can generate multi-speaker speech output in a zero-shot manner, given an unseen input text and an unseen person's reference speech sample. For normalization, we have proposed two architectures: one based on convolution and other based on multi-head attention to capture the prosodic properties in the network through affine parameters. We have also proposed the speaker encoder model based on pretrained wav2vec2.0 to generate 256 dimensional embedding. Using extensive experiments on multi-speaker datasets(VCTK and LibriTTS), we have shown both qualitative and quantitative results along with high-quality of audio output. ZSM-SS can also be used as a voice morphing tool by varying the embedding, frequency and energy inputs to the normalization module.

[1]Generated audios : `https://sites.google.com/view/interspeech/home`

5. References

[1] C. Veaux, J. Yamagishi, and K. Macdonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[3] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," 04 2018, pp. 4779–4783.

[4] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," 05 2019.

[5] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," 06 2020.

[6] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voice synthesis for in-the-wild speakers via a phonological loop," 07 2017.

[7] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," 02 2018.

[8] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," 06 2018.

[9] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "Multispeech: Multi-speaker text to speech with transformer," 06 2020.

[10] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," 02 2018.

[11] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," 05 2017.

[12] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," 10 2017.

[13] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," 03 2018.

[14] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," 03 2018.

[15] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," 05 2019, pp. 6945–6949.

[16] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," 03 2017.

[17] G. Sun, Y. Zhang, R. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," 02 2020.

[18] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, sheng zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=Drynvt7gg4L

[19] L. Wan, Q. Wang, A. Papir, and I. Moreno, "Generalized end-to-end loss for speaker verification," 10 2017.

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[22] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[23] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2020.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," 2018.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[26] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, J. Chen, J. Chen, Z. Chen, M. Chrzanowski, A. Coates, G. Diamos, K. Ding, N. Du, E. Elsen, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," 12 2015.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *INTERSPEECH*, 2017.

[28] PyWORLD, "https://github.com/jeremycchsu/python-wrapper-for-world-vocoder," 2019.

[29] XLSR, "https://github.com/pytorch/fairseq/tree/master/examples/wav2vec," 10 2020.

[30] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," 2017.

[31] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[32] Wei Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3969–3972.

[33] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128 vol.1.