SATTS: Speaker Attractor Text to Speech, Learning to Speak by Learning to Separate

Nabarun Goswami[1], *Tatsuya Harada*[1,2]

[1]The University of Tokyo, Japan
[2]RIKEN, Japan

{nabarungoswami,harada}@mi.t.u-tokyo.ac.jp

Abstract

The mapping of text to speech (TTS) is non-deterministic, letters may be pronounced differently based on context, or phonemes can vary depending on various physiological and stylistic factors like gender, age, accent, emotions, etc. Neural speaker embeddings, trained to identify or verify speakers are typically used to represent and transfer such characteristics from reference speech to synthesized speech. Speech separation on the other hand is the challenging task of separating individual speakers from an overlapping mixed signal of various speakers. Speaker attractors are high-dimensional embedding vectors that pull the time-frequency bins of each speaker's speech towards themselves while repelling those belonging to other speakers. In this work, we explore the possibility of using these powerful speaker attractors for zero-shot speaker adaptation in multi-speaker TTS synthesis and propose speaker attractor text to speech (SATTS). Through various experiments, we show that SATTS can synthesize natural speech from text from an unseen target speaker's reference signal which might have less than ideal recording conditions, i.e. reverberations or mixed with other speakers.

Index Terms: text to speech, speech separation, speaker adaptation, zero-shot, speaker attractor

1. Introduction

With the advancement of various deep learning techniques, TTS systems have improved quite a lot. Most modern TTS systems have two parts, a front end synthesizer and a backend vocoder. The synthesizer takes as input text or phonemes and synthesizes an intermediate representation like mel-spectrogram. Examples of such synthesizers are the Tacotron family of synthesizers [1, 2, 3], Transformer TTS [4], FastSpeech [5], etc. The backend vocoders convert the intermediate representations into speech waveforms. Various kinds of vocoders have been proposed including, but not limited to, Wavenet [6], WaveGlow [7], MelGAN [8], HiFi-GAN [9], etc. Another family of TTS systems work end-to-end, i.e. convert text directly into waveform without going through an intermediate representation, Methods such as EATS [10] and VITS [11] are examples of end-to-end TTS systems.

These methods have pushed the boundaries of quality of synthesized speech, in fact most single speaker methods have naturalness almost at par with real human speech. However, there is still much to be desired in terms of multi-speaker speech synthesis, especially in zero-shot speaker adaptation setting. Multi speaker setting is usually incorporated into the TTS systems in the form of conditioning speaker embedding. This embedding might be from a look up table in the case when using fixed speaker identities [6, 11], trained along with TTS model [12, 13] or in the form of embeddings extracted from a speaker discriminative model such as speaker verification as in SV2TTS [14] and YourTTS [15]. The third method of extracting embeddings from a speaker discriminative model allows for zero shot speaker adaptation without changing the parameters of the TTS system at inference time. While such speaker discriminative embeddings have been shown to work well for multi speaker TTS systems, these embeddings are mainly trained to just capture the broad global attributes of different speakers and being agnostic to the specific features of a particular reference speech sample.

To alleviate this problem, we propose to use speaker attractors [16] as embeddings for zero shot speaker adaptation. Speaker attractors are high dimensional embeddings which are used for pulling the time-frequency embeddings of a speaker closer to itself in the task of speech separation from a mixture of different speakers. The speaker attractors, by nature, are capable of capturing global speaker characteristics as well as information specific to the particular reference speech sample, since they are trained to extract a particular speaker's speech from a mixed or noisy recording. While there are numerous methods for speech separation , the deep clustering [17] and deep attractor [18] based methods are of most relevance to our work. Speaker attractor network [16], an extension of the deep attractor network, which employs metric learning among attractors of same speaker in different mixtures to better capture global speaker features along with specific utterance features.

The key contributions of our work are as follows

- Speaker attractors are extracted from a reference encoder pretrained for speech separation
- We adapt the end-to-end TTS method VITS for the purpose of zero shot multi speaker TTS
- We show through various experiments that SATTS performs at par with a strong baseline system under clean recording conditions and outperforms under varying recording conditions.
- We show, for the first time to the best of our knowledge, the ability to extract speaker embeddings for TTS from mixed reference signals with more than one speakers.

2. Speaker attractor text to speech

In this section we describe our proposed speaker attractor text to speech system. The overall steps of SATTS is similar to SV2TTS [14], in that we first train a model to extract speaker embeddings followed by training of the TTS system with the extracted embeddings. However, in this work, we extract the speaker attractors trained to separate speech from a mixed signal. We utilize these powerful speaker attractors to condition the TTS system. Also unlike SV2TTS, we adapt the end-to-end

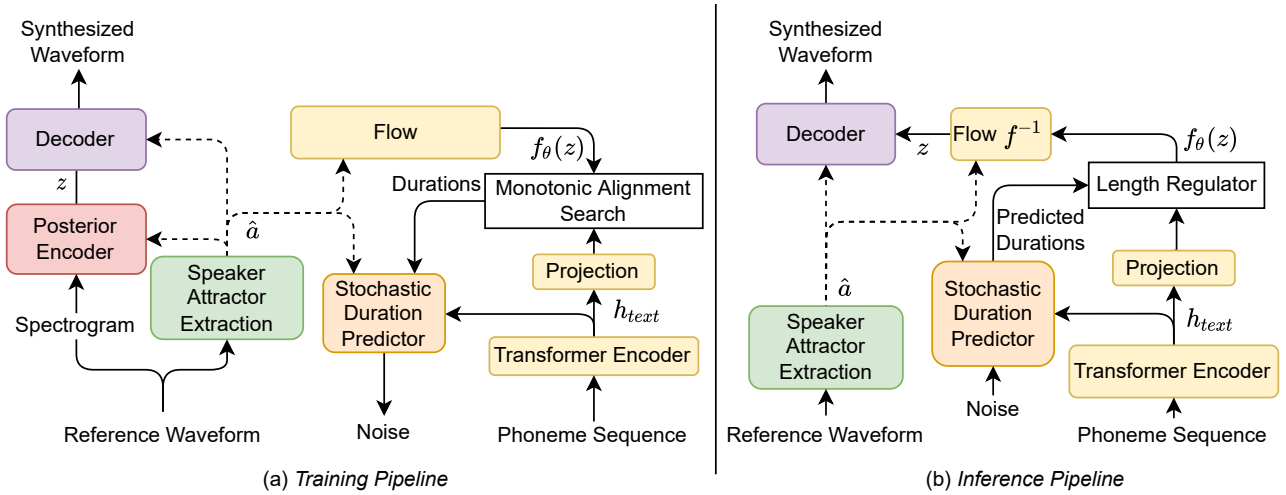

Figure 1: *System overview of SATTS during training and inference.*

VITS TTS system, which is a non-autoregressive conditional variational autoencoder. Fig. 1 shows the system architecture of SATTS. Details about individual components are presented in the following sucsections.

2.1. End-to-end text to speech

The TTS backbone in SATTS is adapted from the VITS architecture [11]. We would like to point out that the use of speaker attractors is not limited to just the end-to-end models and can easily be adapted into any existing multi speaker TTS pipeline.

The VITS architecture at its core is a conditional variational autoencoder. It consists of a posterior encoder, a prior encoder, a stochastic duration predictor and a decoder.

The posterior encoder is a stack of non-causal wavenet residual blocks which takes full scale linear spectrogram as input and produces latent variables as part of a factorized normal distribution, z. To enable multi speaker synthesis, the speaker attractor is incorporated as global conditioning to the residual blocks.

The prior encoder consists of a stack of transformer encoder layers as a text encoder which produces the hidden representation h_{text} also parameterized as a factorized normal distribution. The hidden text representation and the latent variables from the posterior encoder are aligned via a hard monotonic alignment matrix at training time. A normalizing flow f_θ [19] transforms the hidden text representation h_{text} into a more complex distribution of the posterior in an invertible way, $f_\theta^{-1}(f_\theta(z))$, by change of variable on top of the factorized normal distribution. The flow is a stack of affine coupling layers with stacks of wavenet residual blocks. Similar to the posterior encoder, the speaker attractor is incorporated as global conditioning to the residual blocks.

The alignment search operation between the prior and posterior distributions is done via Monotonic Alignment Search (MAS) [20], that searches for the alignment which maximizes the Evidence Lower Bound (ELBO) of data parameterized by a normalizing flow.

In conjunction with the prior encoder and MAS, a stochastic duration predictor (SDP) is trained to estimate the duration of h_{text} for each time step. During inference, since the posterior is not available, the predictions from the stochastic duration predictor are used to regulate the length of the hidden text representation before feeding them into the inverse flow. The stochastic duration predictor is a flow based generative model which is trained via a variational lower bound of the log-likelihood of the phoneme duration, which is provided by MAS. For a more in-depth discussion about training the stochastic duration predictor, please refer to [11].

The decoder (G) architecture is akin to the HiFi-GAN [9] generator which consists of a stack of transposed convolution layers followed by multi receptive field receptors. The decoder takes the latent z and upscales it as per the hop size of the spectrogram operation. To train the decoder efficiently and reduce its memory footprint, random fixed length slices are extracted from z. The speaker attractor is linearly transformed and added to the latent variable z.

The proposed SATTS model is trained with the same objective functions as VITS.

$$L_{tts} = L_{recon} + L_{kl} + L_{dur} + L_{adv}(G) + L_{fm}(G), \quad (1)$$

where L_{recon} is the mel-spectrogram reconstruction loss, L_{kl} is the KL-divergence between the prior and posterior distributions following expansion of the prior with MAS, L_{dur} is the negative of the variational lower bound of log-likelihood of the phoneme duration for the stochastic duration predictor and L_{adv} and L_{fm} are the adversarial and discriminator feature matching losses given by a set of multi-period discriminators as in HiFi-GAN. For a more detailed description of the loss terms please refer to [11].

2.2. Speaker attractors

Speaker attractors [16] are points in high dimensional embedding space which pull time-frequency bins belonging to that speaker towards itself in mixed or corrupted signals. Speaker attractors are trained such that they are able to separate speaker sources from a mixture while also being localized in the global embedding space. This is important, as this enables the speaker attractors to be used in cases where the number of sources in the mixture can be different (more or less) than what was used during training, thus generalizing well to unknown number of sources. This property is also the cornerstone of SATTS, as this allows us to use the speaker attractors for capturing a holistic representation of the target speaker from the reference utterance.

The training and inference pipelines of the speaker attractor network (SANET) are shown in Fig. 2. It consists of a temporal

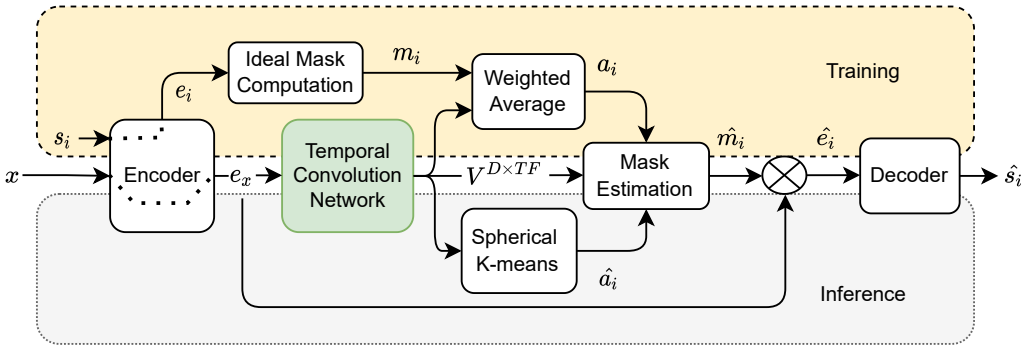

Figure 2: *Overall architecture of Speaker Attractor Network*

convolution network, as the separation backbone, encapsulate by an encoder and a decoder.

While the encoder and decoder can be any function which can extract a time frequency representation of audio waveform, it has been shown that a data driven pre-trained encoder decoder performs best for speech separation [21]. The encoder comprises of a single convolution layer followed by rectified linear unit activation, while the decoder is a single transposed convolution layer without any activation. The time domain waveforms, are processed with overlapping windows of 16 samples with a hop of 8 samples producing a sequence of frames with a F dimensional vector per frame, e. This form of time frequency representation is quite useful because the encoding and decoding process does not have to worry about phase reconstruction [22]. The encoder and decoder are first pretrained without the TCN.

The separation backbone TCN is an adapted version of the TCN in Conv-TasNet [23]. It takes the time frequency input embeddings, e_x, of the mixture signal x, and produces D dimensional vectors for each time frequency (TF) bin, $V^{D\times TF}$

During training, the time frequency representation (e_i) of the C mixing sources, (s_i), are used to compute an ideal ratio mask [24], m_i. These ideal masks are then used for weighted averaging of $V^{D\times TF}$ to produce the ideal attractors, a_i, which lie on the unit sphere in $\mathbb{R}^D$, for each source in the mixture.

$$\boldsymbol{a}_i = \frac{\boldsymbol{V}\cdot(\boldsymbol{w}\odot\boldsymbol{m}_i)}{\|\boldsymbol{V}\cdot(\boldsymbol{w}\odot\boldsymbol{m}_i)\|_2}, \tag{2}$$

where, $\boldsymbol{w} = \boldsymbol{e}_x/\|\boldsymbol{e}_x\|_1$ is the weight which ensures that low energy regions (silence) do not affect the attractor formation.

Following this, the cosine distance between the vectors of V are computed to each attractor followed by a C way softmax over the distances to decide the assignment of each time frequency bin to the closest attractor. This operation produces the estimated masks, $\hat{m}_i$ which are then applied to e_x and decoded by the decoder to produce the source estimates, $\hat{s}_i$.

The SANET model is trained with a combination of three objective functions, a reconstruction loss, a contrastive circle loss and a compactness loss,

$$L = L_{recon} + L_{circle} + L_{compact} \tag{3}$$

The reconstruction loss is the scale invariant signal to distortion ratio between the sources and estimates. The contrastive circle loss drives the attractors to be localized in the global embedding space. And, the compactness loss leads to a compact distribution of embedding vectors of the same speaker. For detailed description of each loss term, please refer to [16].

During inference, the attractors ($\hat{a}_i$) are estimated from $V^{D\times TF}$ by means of Sperical K-means clustering [25], which uses cosine distance instead of Euclidean distance, which ensures there is no mismatch between training and inference time.

For the purpose of extracting the speaker attractors of the reference waveforms for SATTS training, we set $K = 1$ for spherical K-means clustering.

3. Experiments

In our experiments, we compare SATTS with a SV2TTS baseline system where we swap out the speaker attractor extraction with a speaker encoder trained for the speaker verification task, similar to [14]. The baseline speaker encoder is prepared with a ResNet backbone and trained with Angular Prototypical [26] loss function.

Both the speaker encoder of SV2TTS and SANET are trained on the English *train* subset of the Commonvoice v6.1 dataset [27], which is a large scale crowdsourced speech dataset, with 66k speakers and a variety of accents and recording conditions. We resample the speech dataset to 16kHz sampling rate and use a speaker embedding dimension $D = 128$ for both the methods.

For training the SANET, we created the mixture by adding two randomly chosen utterances from the dataset and scaling them with random gains r and $(1-r)$ with $0.25 \leq r \leq 0.75$

To reduce the training time, similar to [15], we initialized the TTS model weights from a single speaker model trained on LJSpeech dataset [28] for 1 million steps, followed by multi-speaker training on the *train-clean-100* subset of the LibriTTS dataset [29], at 22050Hz. To generate the spectrograms for the posterior encoder, a 1024 point short time Fourier transform (STFT) with sliding windows of 1024 samples and 75% overlap is used. The input to the text encoder is the IPA phonetic transcription of the text. For training the decoder we use segments of 32 frames and 80 mel bands.

We train each TTS model on 4 Nvidia A100 GPUs with 80GB of memory, with a batch size of 108 per GPU for a total batch size of 432. We used AdamW optimizer with $\beta_1 = 0.8, \beta_2 = 0.99$ and weight decay $\lambda = 0.01$. The learning rate decay is reduced by a factor of $0.9991/8$ every epoch with an initial learning rate of $2e^{-4}$. We use mixed precison training [30] and train the models for a further 40k iterations.

We perform inference on a total of 21 unseen speakers, 10 from *test-clean* subset of the LibriTTS and 11 from VCTK dataset [31]. There are 12 female and 9 male speakers. We randomly draw 55 test sentences from the *test-clean* subset of the LibriTTS dataset, with a constraint of at least 20 words per sentence. 5 utterances were synthesized per speaker. As ground truth, we randomly select 5 audios for each of the test speakers. We set the noise scaling parameters of both the prior encoder

and the stochastic duration predictor (SDP) to 0.333 for all experiments, except that in Sec.3.3, which uses 1.333 for the SDP. Please refer to the 'demo' folder in the attached multimedia files to listen to samples.

To evaluate our methods, we use crowd sourced subjective tests to evaluate the mean opinion score (MOS) [32] on a scale of 1-5 with intervals of 0.5. Similar to [14], we also evaluate the speaker similarity in terms of MOS (sim-MOS). Each sample received one vote and all evaluations done independently without directly comparing any of the methods.

3.1. Reference with clean recording conditions

Table 1 shows the performance of SATTS in comparison with SV2TTS under clean recording conditions. All reference samples were taken from the dataset and resampled to 16kHz for extraction of the speaker embeddings and attractors followed by TTS inference. We observe that SV2TTS has a slight advantage over SATTS in terms of sim-MOS especially for the VCTK dataset, which consists of the most variance in terms of accents. The naturalness MOS is at par or slightly better for SATTS in samples from both the datasets.

Table 1: *Comparison of MOS and sim-MOS between SV2TTS and SATTS outputs for clean reference signals.*

	MOS		sim-MOS	
	LibriTTS	VCTK	LibriTTS	VCTK
GT	4.13 ± 0.22	4.09 ± 0.15	3.88 ± 0.29	3.83 ± 0.27
SV2TTS	3.82 ± 0.27	**3.83 ± 0.21**	**3.68 ± 0.34**	**3.68 ± 0.28**
SATTS	**3.95 ± 0.21**	3.79 ± 0.25	3.66 ± 0.27	3.45 ± 0.32

3.2. Reference with varying recording conditions

Table 2: *Comparison of MOS and sim-MOS between SV2TTS and SATTS outputs when the reference signal is convolved with a random RIR, simulating different recording conditions.*

	MOS		sim-MOS	
	LibriTTS	VCTK	LibriTTS	VCTK
GT-RIR	4.16 ± 0.22	3.91 ± 0.22	3.54 ± 0.36	3.50 ± 0.31
SV2TTS	3.91 ± 0.17	3.96 ± 0.20	3.38 ± 0.36	3.50 ± 0.34
SATTS	**3.96 ± 0.20**	**4.06 ± 0.21**	**3.47 ± 0.34**	**3.62 ± 0.31**

To evaluate the synthesis performance of the proposed method under different recording and reverberation conditions, for each of the test reference samples and ground truths, we randomly sample one simulated room impulse responses from the *simulated_rirs* subset of the Room Impulse Response and Noise dataset [33]. We perform the exact same evaluation as with the cleanly recorded samples described above. All samples are compared with the clean version of the reference for the subjective evaluation. Table 2 show the performance comparison under this condition. We can see that SATTS performs slightly better than SV2TTS in both the datasets. Though, it must be noted that SV2TTS also works quite well as the verification training is done on the Commonvoice dataset which has a variety of recording conditions.

3.3. Reference with overlapping speakers

We evaluate the performance of SATTS in zero shot speaker adaptation when the reference sample consists of a mix of more than one speaker. We added a distractor speech sample from a different unseen speaker from the LibriTTS dataset to all the test reference signals, and extracted the target speaker's attractor by setting $K = 2$ for the spherical K-means clustering. Since extraction of speaker embedding from mixed speech is not possible for SV2TTS, we only evaluate SATTS. Table 3 shows the performance comparison of the clean, RIR and mix conditions. We can clearly see that SATTS works in these challenging and different settings without compromising on the naturalness and speaker similarity. While, it might seem odd that the clean reference is not the best performing one, this is expected, since the mixed condition matches the training of SANET better than the clean condition.

Table 3: *Comparison of MOS and sim-MOS for SATTS outputs based on the conditions of reference signal, clean speech, convolved with random RIR and mixed with another speech.*

	MOS		sim-MOS	
	LibriTTS	VCTK	LibriTTS	VCTK
Clean	3.95 ± 0.21	3.79 ± 0.25	**3.66 ± 0.27**	3.45 ± 0.32
RIR	**3.96 ± 0.20**	4.06 ± 0.21	3.47 ± 0.34	**3.62 ± 0.31**
Mix	3.84 ± 0.26	**4.09 ± 0.20**	3.60 ± 0.32	3.40 ± 0.31

3.4. Speech recognition results

Table 4: *Comparison word error rate (WER) based on a off-the-shelf ASR model from SpeechBrain (Lower is Better)*

	SV2TTS-clean	SV2TTS-rir	SATTS-clean	SATTS-rir
WER	8.65	8.74	**6.30**	6.34

We also compared the intelligibility in terms of automatic speech recognition (ASR) performance using an off-shelf ASR model from SpeechBrain [34]. Table 4 shows that SATTS achieves much better word error rates (WER) compared to the baseline SV2TTS method for all evaluations samples from both LibriTTS and VCTK dataset demonstrating the superiority of speaker attractors for TTS.

4. Conclusion

In this work, we propose speaker attractor text to speech which utilizes speaker attractors trained for speech separation and transfer the learning to train an end-to-end text to speech synthesis system. Through subjective evaluations, we show the robustness of SATTS over various challenging conditions for the reference signal. To the best of our knowledge, this is the first work to utilize transfer learning from speech separation to text to speech synthesis, and the ability to extract target speaker's attractor from a signal with more than one speakers speaking simultaneously could be quite useful in real world applications.

5. Acknowledgements

This work was partially supported by JST AIP Acceleration Research JPMJCR20U3, Moonshot R&D Grant Number JPMJPS2011, JSPS KAKENHI Grant Number JP19H01115, and JP20H05556 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] 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 *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5709–5713.

[4] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network." AAAI Press, 2019.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[7] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[8] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[9] 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.

[10] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *International Conference on Learning Representations*, 2021.

[11] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[12] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *Advances in neural information processing systems*, vol. 30, 2017.

[13] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[14] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[15] E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. Gölge, and M. Antonelli Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone," *arXiv e-prints*, p. arXiv:2112.02418, Dec. 2021.

[16] F. Jiang and Z. Duan, "Speaker attractor network: Generalizing speech separation to unseen numbers of sources," *IEEE Signal Processing Letters*, vol. 27, pp. 1859–1863, 2020.

[17] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[18] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[19] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[20] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[21] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 31–35.

[22] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "Phasenet: Discretized phase modeling with deep neural networks for audio source separation." in *Interspeech*, 2018, pp. 2713–2717.

[23] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 266–270.

[25] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine learning*, vol. 42, no. 1, pp. 143–175, 2001.

[26] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

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

[29] 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," 2019.

[30] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *International Conference on Learning Representations*, 2018.

[31] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[32] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, vol. 22, 1996.

[33] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[34] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.