

LOW-RESOURCE EXPRESSIVE TEXT-TO-SPEECH USING DATA AUGMENTATION

Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, Jaime Lorenzo-Trueba

Amazon Alexa - TTS Research, Cambridge, United Kingdom
{huybrech, thommer, gcomini, perzbart, raahshah, truebaj}@amazon.com

ABSTRACT

While recent neural text-to-speech (TTS) systems perform remarkably well, they typically require a substantial amount of recordings from the target speaker reading in the desired speaking style. In this work, we present a novel 3-step methodology to circumvent the costly operation of recording large amounts of target data in order to build expressive style voices with as little as 15 minutes of such recordings. First, we augment data via voice conversion by leveraging recordings in the desired speaking style from other speakers. Next, we use that synthetic data on top of the available recordings to train a TTS model. Finally, we fine-tune that model to further increase quality. Our evaluations show that the proposed changes bring significant improvements over non-augmented models across many perceived aspects of synthesised speech. We demonstrate the proposed approach on 2 styles (newscaster and conversational), on various speakers, and on both single and multi-speaker models, illustrating the robustness of our approach.¹

Index Terms— Text-to-speech, low-resource, data augmentation, expressive speech

1. INTRODUCTION

Recently, the naturalness and quality of speech generated with neural text-to-speech (TTS) systems has improved significantly [1–6]. However, as concluded in [7], high-quality state-of-the-art models such as Tacotron [2] require at least 10 hours of text and speech pairs. Since collecting data is a costly operation, the need for alternatives is high.

Research that focuses on low-resource TTS tries to mitigate the effects of limited data via multi-speaker modelling and transfer learning [7–13]. By transferring knowledge gained from high-resource speakers, the quality of low-resource systems improves. This is not the only way to tackle the problem of scarce data though.

Data augmentation has been used across fields in machine learning to overcome the aforementioned issue. Especially in computer vision, creating synthetic data to train better models has been thoroughly investigated [14–16]. Data augmentation has also been researched in the speech field, particularly in the field of automatic speech recognition [17–19]. However, despite the large amount of research conducted to create synthetic speech data, only [20] seems to leverage it for

non-expressive TTS end goals. A multi-speaker TTS model is trained, which is used to create synthetic data for the target speaker, which in turn is used to train a new dedicated single-speaker TTS model. This assumes the low-resource multi-speaker model can synthesise speech for the target speaker (with natural prosody and good signal quality), a non-trivial assumption.

A more effective way to augment data in the speech field which bypasses this constraint can be achieved through voice conversion (VC). VC aims to convert an utterance spoken by a source speaker as though it was read by a different target speaker, while maintaining all other features such as linguistic content and prosody. Much research has already been conducted on VC [21–24]. In this paper, we chose to use Copycat [25], known for its robustness and excellent VC quality. Nevertheless, we expect results to be applicable to other high-quality VC techniques.

In this work, we present a novel methodology which results in high-quality TTS models, despite using only a small amount of expressive samples for the target speaker. We propose a 3-step solution involving voice-converted data and fine-tuning. The voice-converted data resembles the target recordings more than any other auxiliary data, regardless of whether this data belongs to another style or speaker. Hence, the model is trained on much more data coming from a similar distribution and can more reliably produce the desired type of speech. The fine-tuning, in turn, allows the model to focus on the actual target space more closely. Our methodology shows significant improvements across many use cases, illustrating its robustness. To the best of our knowledge, our proposed methodology is the first one that uses a VC model to augment data with the aim to train better low-resource TTS models.

2. METHODOLOGY

The proposed methodology consists of 3 steps. In the first step, we create synthetic data by leveraging a VC model [25]. We augment recordings from supporting speakers recorded in the target speaking style to the identity of our target speaker. In the second step, the newly-generated synthetic data is used in addition to the target speaker recordings to train a state-of-the-art TTS model. Finally, fine-tuning of the TTS model is performed to achieve higher quality. We will now describe these steps in more detail.

¹Samples available on www.amazon.science/low-resource-expressive-text-to-speech-using-data-augmentation

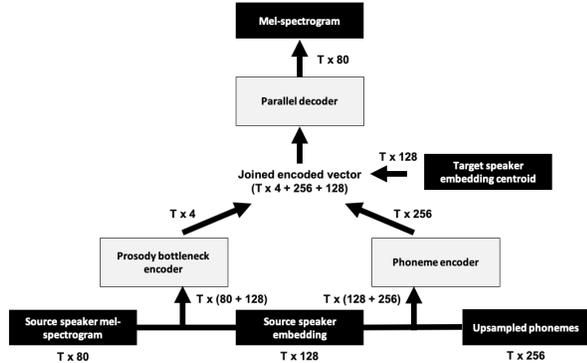


Fig. 1. VC model architecture

2.1. Data augmentation

To create synthetic data for our experiments, we leverage the Copycat model [25] as VC model (Fig. 1). This model is a fine-grained prosody transfer model, aiming to maintain the prosody and linguistic content from a source audio while converting the speaker identity. It consists of three components: (1) a phoneme encoder that learns latent embeddings of the phonemes, (2) a prosody bottleneck encoder which disentangles the prosodic representation from the reference mel-spectrogram, and (3) a parallel decoder which generates a mel-spectrogram given the phoneme embeddings, the prosodic representation, and a provided speaker embedding. For more detailed information about the architecture and the functioning of the model, as well as for more information on the quality one can expect to obtain, we refer the reader to the Copycat paper [25].

We bring one modification to the model, by concatenating the speaker embeddings to the upsampled phonemes before passing the result to the phoneme encoder. This helps encode phonemes according to the speaker identity, and therefore aids the VC process. We find this change to reduce occurrences of speaker leakage.

The VC model is trained on one or more source speakers, plus the target speaker, making it possible to convert expressive samples of a style X from a source speaker S to a target speaker T . Despite the samples not being as perfect as recordings, they still are of great quality and helpful to train better TTS models, as shown in our evaluations.

2.2. TTS model

2.2.1. Architecture

Our TTS system is based on a Tacotron-like [2] structure with an additional variational auto-encoder (VAE) [26] to capture prosody [27, 28] (Fig. 2). Depending on the use case and to show the robustness of our approach across different scenarios, we either train a multi-style single-speaker or a single-style multi-speaker model. When training the latter model,

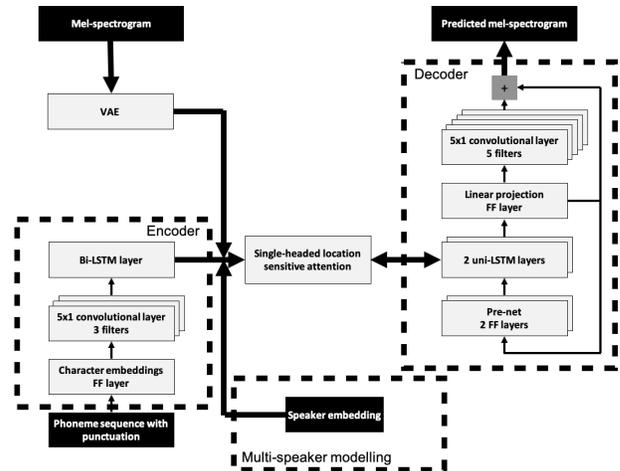


Fig. 2. TTS model architecture

we also add a pre-trained speaker embedding at the decoder level, which is obtained via the speaker classifier described in [25]. The model is trained for 400k steps using a batch size of 32 with a Kullback-Leibler divergence loss for the VAE component, a L1-loss for the predicted mel-spectrograms and a cross-entropy stop-token loss. At inference time, the model is conditioned on the target speaker centroid VAE z-vector computed on the target style recordings. For generating raw audio waveforms from predicted mel-spectrograms, we use a parallel Wavenet universal neural vocoder [6].

2.2.2. Fine-tuning

Once the model is trained on the entire dataset, we resume training the model in the exact same setting (i.e. same architecture, same losses, etc.), but only on the non-synthetic data for the target speaker in the target speaking style. This fine-tuning training is run for an additional 4k steps. Doing so, we are able to improve both the signal and segmental quality, as well as the expressiveness, without it being detrimental for stability/overfitting. The rationale behind this idea is that fine-tuning guides the model to focus on the space that matters the most. Unlike many existing low-resource TTS fine-tuning techniques [7, 10, 13], the target data is here already present in the so called pre-training step, making our fine-tuning step more of a refinement step.

3. EXPERIMENTS

3.1. Experimental settings

We make use of professionally recorded American English internal datasets. To demonstrate the efficacy of the proposed methodology, we conducted experiments in two different data scenarios, for each of which we test on different expressive speaking styles (i.e. newscaster and conversational). For both scenarios, only 30 min of target recordings are used. We explain the settings on the basis of the notations of Table 1.

Notation	Description
$S_{i,j,k}$	Dataset of real speaker i in style j consisting of k hours
$S_{i,j,k}^*$	Dataset of synthetic/voice-converted speaker i in style j consisting of k hours
$VC(\{S_{i,j,k}\})$	Voice conversion model trained on a set of datasets $S_{i,j,k}$
$TTS(\{S_{i,j,k}^*\})$	TTS model trained on a set of datasets $S_{i,j,k}^*$
$FT(TTS, S_{i,j,k})$	Fine-tuning model TTS on dataset $S_{i,j,k}$

Table 1. Notation used to describe experimental settings.

In the **single-speaker TTS - newscaster style** scenario, we have access to a large amount of target speaker data in a different speaking style (i.e. neutral). However, we have only one other different supporting speaker. We train next 2 pairs of VC & multi-style single-speaker TTS models, one for target speaker *Female 1* and one for target speaker *Female 2*:

1. $VC_1(S_{1,neutral,20h}, S_{1,news,0.5h}, S_{2,neutral,20h}, S_{2,news,7h})$
 \hookrightarrow Convert $S_{2,news,7h}$ to $S_{1,news,7h}^*$
 $\hookrightarrow TTS_1(S_{1,neutral,20h}, S_{1,news,0.5h}, S_{1,news,7h}^*)$
 $\hookrightarrow FT(TTS_1, S_{1,news,0.5h})$
2. $VC_2(S_{1,neutral,20h}, S_{1,news,4h}, S_{2,neutral,20h}, S_{2,news,0.5h})$
 \hookrightarrow Convert $S_{1,news,4h}$ to $S_{2,news,4h}^*$
 $\hookrightarrow TTS_2(S_{2,neutral,20h}, S_{2,news,0.5h}, S_{2,news,4h}^*)$
 $\hookrightarrow FT(TTS_2, S_{2,news,0.5h})$

In the **multi-speaker TTS - conversational style** scenario, we have a large amount of conversational data across various speakers (split evenly between male and female). However, we have no supporting data from the target speaker in other speaking styles. We train one single VC model, and one single-style multi-speaker TTS model for each of our 8 target speakers S_{1-8} , using 10 supporting speakers S_{9-18} :

- $VC(S_{1-8,conv,0.5h}, S_{9-15,conv,5h}, S_{16-18,conv,1.5h})$
 $\hookrightarrow \forall i \in [1, 8]:$ Convert $S_{x,news,5h}$ to $S_{i,conv,5h}^*$, with x chosen from $S_{9-15,conv,5h}$
 $\hookrightarrow \forall i \in [1, 8]: TTS_i(S_{i,conv,0.5h}, S_{9-18,conv,\Sigma}, S_{i,conv,5h}^*)$
 $\hookrightarrow \forall i \in [1, 8]: FT(TTS_i, S_{i,conv,0.5h})$

Experiments, omitted here due to a lack of space, found that converting data from more than one 5h-supporting speaker doesn't increase quality.

3.2. Metrics and evaluation procedure

Throughout the evaluations, we focus on 4 metrics to assess our models: (1) *Signal quality* (How good is the signal quality of the samples?), (2) *Style adequacy* (Do the samples match the desired style?), (3) *Naturalness* (Do the samples sound natural?), and (4) *Speaker similarity* (Do the samples sound like the target speaker?). For style adequacy and speaker similarity, we provide the listeners a reference sample. We evalu-

ate our models via MUSHRA tests [29]. For each newscaster target speaker, 200 test samples have been held out and 200 testers have been recruited who evaluate each 10 test screens. For the conversational use case, 50 test utterances from each of the 8 target speakers are evaluated, with each test screen being rated 30 times. The evaluations are carried out using crowd-sourcing platforms. To check for statistical differences, we perform paired t-tests using the Holm-Bonferroni correction and a p-value of 0.05.

3.3. Results

Single-speaker TTS - Newscaster style

Component ablation - In our first experiment, we show the (combined) improvements obtained by the two contributions of this paper: 1) using voice-converted data, and 2) fine-tuning the model on the target data. We train 4 types of models for the ablation study: (B) a baseline TTS model, ($B+FT$) the same model as (B) but further fine-tuned on the newscaster recordings, ($B+VC$) the same model as (B) but including the voice-converted data in the training process, and ($B+VC+FT$) the same model as ($B+VC$) but further fine-tuned on the newscaster recordings.

Listener responses from the ablation study are shown in Table 2. These results show that the signal quality and the style adequacy of our methodology ($B+VC+FT$) significantly outperforms all other models for both speakers, except for the *Female 2* style adequacy, for which the improvement is not statistically different. Furthermore, we observe that the separate use of synthetic data and fine-tuning brings improvements in the majority of cases. However, the combination of both approaches brings the largest improvements. We hypothesise that including synthetic data helps the model to get a better picture of the type of data it has to produce, as it sees much more target data, despite it not being all real data. The fine-tuning then helps the model to focus on the actual target space more closely after having seen a lot of neutral and synthetic data.

	Female 1		Female 2	
	Sig. Q.	Style A.	Sig. Q.	Style A.
Base (B)	69.9±1.1	65.7±1.4	67.0±1.0	68.7±1.3
$B+FT$	69.7±1.0	66.2±1.3	69.3±1.0	70.5±1.3
$B+VC$	70.2±1.1	66.6±1.4	69.4±1.0	70.0±1.3
$B+VC+FT$	72.9±1.0	67.9±1.3	71.2±1.0	70.0±1.3

Table 2. Ablation study on usage of VC data and fine-tuning, showing average + 95% CI MUSHRA scores for both signal quality and style adequacy in the single-speaker setting on 2 female speakers. An underlined value signifies a statistical difference between (B) and ($B+VC+FT$).

Data ablation - To understand how this methodology is affected by the amount of data from the target speaker in the target speaking style and compares to upper anchors, we performed further evaluations between different data reduced (DR) and non-data reduced ($non-DR$) scenarios. These evaluations compare 5 systems: ($Recs$) recordings, ($non-DR$) TTS

models trained on neutral data and the full/non-data reduced newscaster dataset (i.e. 4h for *Female 1*, 7h for *Female 2*), (*DR+VC+FT*) TTS models trained on neutral data, a reduced newscaster dataset (i.e. 45, 30 and 15 min), synthetic data generated via VC and fine-tuned for 4k steps, (*DR*) TTS models trained on neutral data and a reduced newscaster dataset (i.e. 45, 30 and 15 min), and (*Neutral*) a neutral TTS model which has been trained on the neutral dataset only.

It is expected that (*DR*) performs worse than (*non-DR*), as (*DR*) is the same model as (*non-DR*) but trained on less data (i.e. 5x to 28x less). We anticipate that (*DR+VC+FT*), generated via our methodology, reduces (and potentially exceeds) the gap between (*DR*) and (*non-DR*). (*DR+VC+FT*) observes the same amount of recorded data as (*DR*), but makes additional use of the synthetic data and TTS model fine-tuning. (*Recs*) is given as an upper anchor, while (*Neutral*) is given as a lower anchor for the style adequacy metric. Every column in Table 3 is made of one *Female 1* and one *Female 2* MUSHRA evaluation.

Table 3 shows improvements for both speakers for the proposed model (*DR+VC+FT*) over the baseline model (*DR*), on both metrics and across 11 out of the 12 MUSHRA evaluations. Out of the 11, 8 are statistically significant improvements. For *Female 1*, we notice that the gap between (*DR*) and (*non-DR*) is often much smaller than the equivalent gap for *Female 2*. Hence, the improvements shown with our models (*DR+VC+FT*) are sometimes able to not only improve over (*DR*) but also (*non-DR*). For *Female 2*, (*DR+VC+FT*) never reaches the quality of (*non-DR*), but the gap between (*DR*) and (*non-DR*) almost always gets vastly reduced. While our methodology works remarkably well for all data scenarios, for the (lowest-resource) 15-minute *Female 2* scenario, it is even the reason why a newscaster model can be built in the first place. Without the use of synthetic data and fine-tuning, the model is only able to generate neutral speech, as can be seen by the style adequacy score that is similar to the one of the neutral TTS model.

#min news recs	Signal quality			Style adequacy		
	45	30	15	45	30	15
Female 1						
Recs	76.9	77.8	78.0	74.7	78.9	78.1
non-DR (4h)	71.4	71.7	71.9	69.7	72.7	68.8
DR+VC+FT	<u>71.9</u>	<u>73.0</u>	69.0	68.3	<u>72.6</u>	63.9
DR	69.4	69.2	68.9	67.3	69.0	64.0
Neutral	X	X	X	61.8	63.9	60.7
Female 2						
Recs	75.0	74.0	71.7	79.2	77.2	76.7
non-DR (7h)	71.0	70.2	69.2	72.4	71.0	71.4
DR+VC+FT	68.6	66.8	67.5	69.2	66.7	<u>67.7</u>
DR	58.8	62.7	66.5	66.9	66.4	58.1
Neutral	X	X	X	61.5	61.4	56.0

Table 3. Average MUSHRA scores for signal quality and style adequacy in the single-speaker setting, showing the performance of our methodology (*DR + VC + FT*) in different data reduced scenarios and for 2 female speakers. An underlined value signifies a statistical difference between (*DR*) and (*DR + VC + FT*).

Multi-speaker TTS - Conversational style

Finally, we demonstrate our methodology in a different scenario, to illustrate its robustness. The target speakers have different amounts of recordings available, enabling us to understand the comparative effects of data reduction (i.e. how does reducing a voice to 30 min of speech data compared to having 45 min, 1.5h or 5h of recordings available?). We compare here two multi-speaker models: (*non-DR*) trained using all data available for all supporting and target speakers, and (*DR+VC+FT*) trained for each target speaker using only 30 min of their recordings plus all data from the supporting speakers with the proposed approach.

Table 4 shows that given just 30 min of speech data the proposed methodology is found to be significantly better than voices trained on the full amount of data for the target speaker. This level of quality is achieved for all 8 speakers evaluated, and this even when 90% of reduction in data was applied (i.e. from 5h of data to 30 min). In addition, we observe that this is achieved without degrading perceived speaker similarity. We believe results are even better for this single-style multi-speaker setting, as synthetic data becomes increasingly valuable in the absence of supporting neutral target speaker data.

		Naturalness	Speaker similarity
SMALL - Reducing 45 min to 30 min			
4 speakers	Recs	78.80±0.73	68.10±0.65
	non-DR	51.78±0.76	65.61±0.67
	DR+VC+FT	<u>56.86±0.72</u>	<u>66.80±0.66</u>
MEDIUM - Reducing 1.5h to 30 min			
2 speakers	Recs	79.86±1.02	69.78±0.89
	non-DR	56.58±1.04	67.45±0.91
	DR+VC+FT	<u>58.72±1.02</u>	67.41±0.90
LARGE - Reducing 5h to 30 min			
2 speakers	Recs	81.65±1.00	67.83±0.94
	non-DR	55.16±1.06	66.25±0.95
	DR+VC+FT	<u>59.39±0.97</u>	<u>66.72±0.93</u>

Table 4. Average + 95% CI MUSHRA scores for naturalness and speaker similarity in the multi-speaker setting, showing the performance of our methodology (*DR + VC + FT*) in different data reduced scenarios and for 8 speakers (evenly split between male and female). An underlined value signifies a statistical difference between (*non-DR*) and (*DR+VC+FT*).

4. CONCLUSION

We propose a novel methodology to build high-quality expressive TTS models, while having only little expressive recordings for the target speaker. The methodology consists of: 1) generating synthetic data via a VC model, 2) incorporating this converted data in training a Tacotron-like TTS model, and 3) fine-tuning this TTS model on the target data. The proposed approach brings improvements in signal quality, naturalness and style adequacy, whilst suffering no drop in speaker similarity. Whilst many VC papers exist, as far as we are aware, there has been no work on the use of VC-created data to train TTS models. The successful use of such data to train TTS models means a reduced cost of recording data, as well as a faster and easier expansion of TTS voices.

5. REFERENCES

- [1] Aaron van den Oord et al., “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [2] Yuxuan Wang et al., “Tacotron: Towards end-to-end speech synthesis,” *arXiv preprint arXiv:1703.10135*, 2017.
- [3] Jose Sotelo et al., “Char2wav: End-to-end speech synthesis,” 2017.
- [4] RJ Skerry-Ryan et al., “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” *arXiv preprint arXiv:1803.09047*, 2018.
- [5] Nal Kalchbrenner et al., “Efficient neural audio synthesis,” *arXiv preprint arXiv:1802.08435*, 2018.
- [6] Aaron Oord et al., “Parallel wavenet: Fast high-fidelity speech synthesis,” in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.
- [7] Yu-An Chung et al., “Semi-supervised training for improving data efficiency in end-to-end speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6940–6944.
- [8] Andrew Gibiansky et al., “Deep voice 2: Multi-speaker neural text-to-speech,” in *Advances in neural information processing systems*, 2017, pp. 2962–2970.
- [9] Ye Jia et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Advances in neural information processing systems*, 2018, pp. 4480–4490.
- [10] Noé Tits et al., “Exploring transfer learning for low resource emotional tts,” in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2019, pp. 52–60.
- [11] Javier Latorre et al., “Effect of data reduction on sequence-to-sequence neural tts,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7075–7079.
- [12] Yuan-Jui Chen et al., “End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning,” in *Interspeech*, 2019, pp. 2075–2079.
- [13] Haitong Zhang et al., “Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages,” *arXiv preprint arXiv:2008.04549*, 2020.
- [14] Luis Perez et al., “The effectiveness of data augmentation in image classification using deep learning,” *arXiv preprint arXiv:1712.04621*, 2017.
- [15] Agnieszka Mikołajczyk et al., “Data augmentation for improving deep learning in image classification problem,” in *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE, 2018, pp. 117–122.
- [16] Connor Shorten et al., “A survey on image data augmentation for deep learning,” vol. 6, no. 1, pp. 60, 2019.
- [17] Tom Ko et al., “Audio augmentation for speech recognition,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [18] Tom Ko et al., “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.
- [19] Daniel S Park et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [20] Jin Xu et al., “Lrspeech: Extremely low-resource speech synthesis and recognition,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2802–2812.
- [21] Seyed Hamidreza Mohammadi et al., “An overview of voice conversion systems,” *Speech Communication*, vol. 88, pp. 65–82, 2017.
- [22] Chin-Cheng Hsu et al., “Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks,” *arXiv preprint arXiv:1704.00849*, 2017.
- [23] Jaime Lorenzo-Trueba et al., “The voice conversion challenge 2018: Promoting development of parallel and non-parallel methods,” *arXiv preprint arXiv:1804.04262*, 2018.
- [24] Takuhiro Kaneko et al., “Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.
- [25] Sri Karlapati et al., “Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech,” *arXiv preprint arXiv:2004.14617*, 2020.
- [26] Diederik P Kingma et al., “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [27] Ya-Jie Zhang et al., “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.
- [28] Shubhi Tyagi et al., “Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection,” *arXiv preprint arXiv:1912.00955*, 2019.
- [29] RB ITU-R, “1534-1, method for the subjective assessment of intermediate quality levels of coding systems (mushra),” *International Telecommunication Union*, 2003.