Cross-lingual Knowledge Distillation via Flow-based Voice Conversion for Robust Polyglot Text-To-Speech

Dariusz Piotrowski[1][0000−0002−7974−9887]⋆, Renard Korzeniowski[1][0009−0004−8759−803X]⋆, Alessio Falai[1][0000−0002−6574−5607]⋆, Sebastian Cygert[2][0000−0002−4763−8381]†, Kamil Pokora[1][0009−0006−0756−4118], Georgi Tinchev[1][0000−0002−9910−6598], Ziyao Zhang[1][0000−0002−0617−1340], and Kayoko Yanagisawa[1][0000−0002−3444−7287]

[1] Alexa AI, Amazon
{piotrod,korenard,falai,kamipoko,gtinchev,zhaziyao,yakayoko}@amazon.com
[2] Gdańsk University of Technology, Gdańsk, Poland

Abstract. In this work, we introduce a framework for cross-lingual speech synthesis, which involves an upstream Voice Conversion (VC) model and a downstream Text-To-Speech (TTS) model. The proposed framework consists of 4 stages. In the first two stages, we use a VC model to convert utterances in the target locale to the voice of the target speaker. In the third stage, the converted data is combined with the linguistic features and durations from recordings in the target language, which are then used to train a single-speaker acoustic model. Finally, the last stage entails the training of a locale-independent vocoder. Our evaluations show that the proposed paradigm outperforms state-of-the-art approaches which are based on training a large multilingual TTS model. In addition, our experiments demonstrate the robustness of our approach with different model architectures, languages, speakers and amounts of data. Moreover, our solution is especially beneficial in low-resource settings.

Keywords: neural text-to-speech · multilingual synthesis · voice conversion · synthetic data · normalising flows

1 Introduction

Polyglot Text-To-Speech (TTS) systems, which rely on training data from monolingual speakers in multiple language variants (from here on referred to as *locale*), enable speakers in the training corpus to speak any language present in the same training corpus [28,10,14,29]. Such State-Of-The-Art (SOTA) models are able to achieve impressive results in cross-lingual synthesis scenarios, in terms of the high-quality naturalness and accent generated in the synthesised audios. During

---

⋆ Equal contribution.
† Work done while at Amazon.

training, the model learns to decorrelate the embedding spaces of speaker and language conditionings, which enables inference-time-synthesis of {*speaker, language*} pairs unseen during training. Such a training paradigm is the mainstream methodology of building Polyglot TTS systems, which is referred to as *Standard Polyglot* hereinafter. One major drawback of *Standard Polyglot* is that the model needs to have significant capacity to be able to disentangle speaker and language characteristics [29,4]. This means the deployment of such models on low-resource devices is extremely challenging. Moreover, existing studies [29,19] have revealed that this approach can be sensitive to data composition.

In this work, we address the issue of deploying large Polyglot models in computationally-constrained settings, by proposing a framework based on a high-capacity Voice Conversion (VC) model, working in conjunction with a low-capacity, single-speaker, monolingual acoustic model. The key point of this solution is the decoupling of speaker-language disentanglement and TTS tasks, which are realised in a single model with *Standard Polyglot* approaches. Specifically, we shift the demanding task of speaker-language disentanglement to the upstream VC model. In addition to its ability to deliver disentanglement, the robustness of our VC model makes it much less sensitive to the composition of the training corpus, thus easing the burden of scaling the framework to new speakers and locales. Extensive evaluations show that, when compared to the SOTA *Standard Polyglot* approach [14], our framework yields significantly better naturalness and accent similarity and on par speaker similarity. Moreover, the robustness of our framework is confirmed in experiments with varying model architecture, target speaker, target locale, and dataset size. To sum up, the contributions of this work are as follows:

1. We propose a new paradigm for cross-lingual TTS based on an upstream VC model and a downstream monolingual TTS model
2. We demonstrate, through extensive evaluations, that it beats SOTA and is robust with regard to data composition as well as architecture type and size
3. We demonstrate that the proposed approach is especially effective under low-resource constraints, enabling the use of lightweight architectures

2   Related work

The idea of using synthetic data to support TTS tasks was explored before. Some works rely on VC as a data augmentation technique, for improving signal quality in low data regimes [21,5] or for style transfer tasks [2]. Most closely related to our work, authors in [4] presented a system which uses synthetic data for accent transfer tasks, but they have an upstream TTS model, while we explore VC techniques, and their downstream acoustic model and vocoder pair is more heavy-weighted than ours, as their focus is solely on reliability. Other approaches present so-called Accent Conversion (AC) techniques on disentangling speaker and accent [25,30], but they usually work in an in-lingual fashion (e.g. by converting a British English speaker to sound like an American English one); in contrast, this work focuses on more generic cross-lingual applications.

The idea of relying on cross-lingual VC to enable Polyglot TTS has already been investigated as well, but previous works [17,23] mainly present results on legacy VC and TTS models, whereas we do so using modern, SOTA architectures. Moreover, we rely on mel-spectrograms alone for both VC and TTS models, instead of using a diverse set of features, such as phonetic posteriorgrams (PPG), as was done in [23,31]. Other approaches [3] position the VC model for speaker identity conversion after the TTS model, which is not ideal for latency. In addition, some works [23,31] specifically target bilingual scenarios with various restrictions (e.g. in [23] the system is designed for one specific target speaker), whereas our approach can be scaled to any target speaker-language combination. Solutions presented in [26] can also be scaled to any speaker-language pair, but the main differences with our work are the following: (1) we show robustness under varying assumptions, thus making our experimental validation more comprehensive; (2) we specifically focus on computationally-constrained applications; (3) we show that our approach to vocoder training works well in a locale-independent fashion, further highlighting the scalability of the proposed solution.
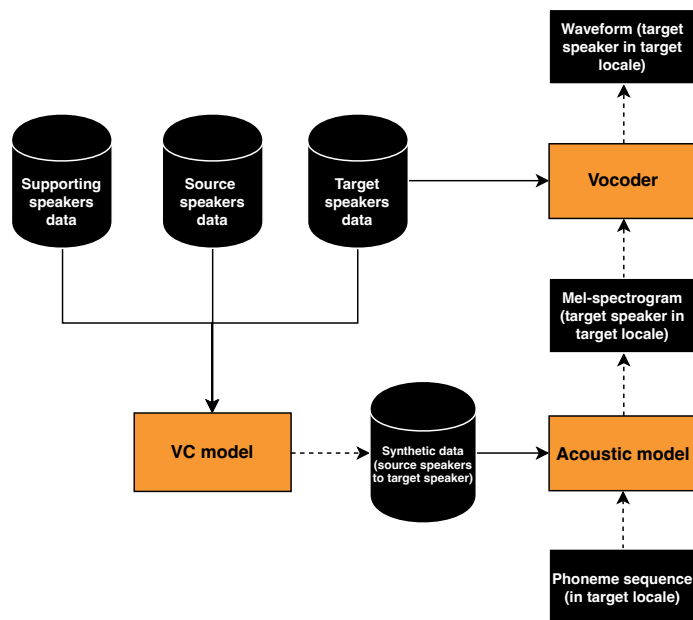
3 Proposed approach

3.1 Framework

Fig. 1: High-level diagram of the proposed framework, where $\rightarrow$ means training, while $\dashrightarrow$ stands for inference.

We propose a 4-stage approach for building a Polyglot TTS system (*VC-based Polyglot*), with the goal of synthesising cross-lingual speech from a *target speaker* in a *target locale*. In addition, we define *source speakers* as the speakers whose native locale is the *target locale*; whereas the locales of *supporting speakers* are not the *target locale*. Our approach consists of the following steps.

1. Train a many-to-many VC model, using training data from *source*, *target* and *supporting speakers*.
2. Convert *source speakers*' identities in the original audio files to that of the *target speaker*, by using the trained VC model from 1, thus creating the synthetic dataset.
3. Use the synthetic dataset from 2 to train a single-speaker, monolingual acoustic model, to produce speech with *target speaker* identity in the *target locale*.
4. Train a speaker-specific locale-independent vocoder using original data from the *target speaker*.

When building the synthetic dataset in 2, we use original phoneme durations (from *source speakers*) instead of re-running forced alignment, as the latter leads to the acoustic model cutting off utterances and producing unnatural prosody. Then, at inference time we discard the upstream VC model and chain the trained acoustic model and vocoder, to output waveforms. Note that 2 and 3 can be repeated for generating synthetic data in multiple *target locales*, without the need to re-train the VC model. Figure 1 illustrates how different components interact with each other at training and inference times.

Note that the framework configuration described above is not the only possible option. For example, in 3, we empirically find that adding *supporting speakers* in VC model training does not improve (nor does it degrade) synthesis quality in the *target locale* for downstream models. Also, in 4, early results show that using *source speakers* alongside original data from the *target speaker* is only beneficial for some *target locales*. Moreover, an alternative solution for 4 would be to train the vocoder in a locale-specific fashion, by generating synthetic waveforms w from synthetic mel-spectrograms m using a powerful universal vocoder (such as the one from [7]), and training on (m, w) pairs. We adopt the locale-independent procedure described in 4 because doing so renders more perceptually expressive speech.

3.2   Model description

For our upstream VC model, we initially considered normalising flows [1,12] and Variational Auto-Encoder (VAE) [8] based architectures. However, given that the latter is particularly sensitive to the dimension of the auto-encoder bottleneck [16], which clashes with our requirement for robustness, we decided to adopt the flow-based VC approach. The main reason for selecting an upstream VC model instead of a TTS one is that the former operates on speech-to-speech mappings, which are easier to learn than the more ill-defined text-to-speech ones. This, in turn, leads to more natural synthetic data than their TTS counterparts [1].
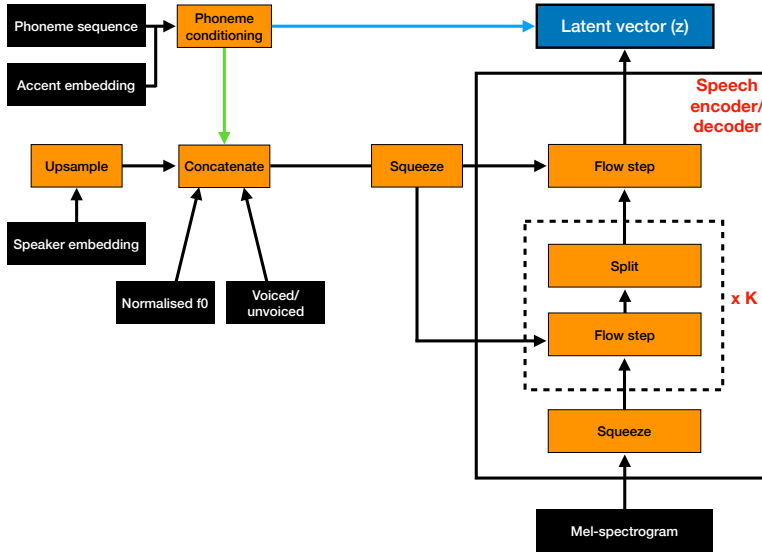
Fig. 2: The many-to-many non-parallel VC model used to generate synthetic datasets. The green line denotes phoneme conditioning, while the blue line the phoneme-based prior.

In particular, the VC model we rely on is the text-conditioned non-parallel many-to-many VC model with fixed prior described in [1] and depicted in Figure 2. The VC model topology is based on Flow-TTS [13] and modified to accept F0 (interpolated and normalised at the utterance level) and V/UV (binary voiced/unvoiced flag) conditionings. Phoneme embeddings are also enriched with accent one-hot encodings, while speaker embeddings are extracted using a pre-trained utterance-level encoder [24]. At training time the model learns to maximise the likelihood

$$\log P_X(x|c) = \log P_Z(z|c) + \log \left| \det \frac{\partial z}{\partial x} \right| \qquad (1)$$

of the prior distribution

$$c \sim N(\mu_p, \sigma_p), \mu_p = 0, \sigma_p = 1 \qquad (2)$$

for all frames regardless of the speech content, where z is the encoded latent vector and x is the input mel-spectrogram. As done in [1], we rely on pre-trained phoneme alignments instead of using an attention mechanism.

As our main acoustic model we chose FastSpeech 2 (FS2) [18], which consists of feed-forward transformer-based encoder with explicit duration, energy, and F0 predictors. In order to reduce latency and model size, we remove the self-attention mechanism in the decoder, thus making it fully-convolutional. To validate the robustness of our acoustic model, we compare it with higher and lower capacity models. The former is based on gated convolutions and recurrent

units (ED) [20], while the latter is LightSpeech (LS) [11], a scaled-down version of FS2 (in two variants, as described in Section 4.3).

With low-resource embedded devices in mind, we selected MultiBand Mel-GAN [27], which is able to generate high quality speech with only 0.95 GFLOPS, as the vocoder in our pipeline. To address audio artefacts typical to GAN-based vocoders, we standardise input features in the range of $[-1, 1]$. Empirically, we find this solution equivalent to, but more suitable than, commonly used fine-tuning techniques [22,9]. The practical reason for this is that no ground-truth waveforms corresponding to synthetic mel-spectrograms are available.

4   Experiments

4.1   Evaluation setup

For evaluations, we use MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [6] tests with a scale of 0 to 100, through crowd-sourcing platforms, to perceptually assess naturalness, speaker similarity and accent similarity of our syntheses, as done in [29,19]. We opted for MUSHRA tests, rather than more popular Mean Opinion Score (MOS) tests, due to the highly comparative nature of our work. In general, a MUSHRA test explicitly provides the listener with the same utterance synthesised with all of the systems to be compared, anchored by the upper and lower references. This form of testing therefore allows the listener to make direct comparisons when making their judgements, thus resulting in meaningful rank amongst systems.

For each evaluation, we use a test set of 200 utterances, unseen during training. Each of the 60 listeners evaluate a random set of 50 utterances for each evaluation. When evaluating speaker similarity, testers are given a reference system, which consists of *target speaker*'s recordings. Hence, the upper anchor is always equal to 100. Recordings from *source speakers* are instead used as the lower anchor. When evaluating naturalness and accent similarity, *source speakers*' recordings are used as the upper anchor, while the lower anchor is a system based on phoneme-mapping, whereby an input from the *target locale* is translated into the locale of the *target speaker* using hard-coded rules for mapping phones between locales (based on linguistic proximity). The phoneme-mapped input is then fed into a Tacotron 2 [22] model trained on the *target speaker*'s data.

To ensure the integrity of our evaluation results, we apply filtering methods to exclude potential cheaters, i.e., we remove submissions from listeners who scored all systems at extreme or default values for more than 5 times. Moreover, to check the statistical significance of our test results, we use a pairwise two-sided Wilcoxon signed-rank test corrected for multiple comparisons with the Holm-Bonferroni method, as done in [20].

Results in Tables 1 - 5 are reported in terms of average MUSHRA scores[1], along with the Closing The Gap (CTG) percentage, which summarises the im-

---

[1] We use boldface for the highest score per aspect if the gap between baseline and proposed system is statistically significant.

provement of the *VC-based Polyglot* approach with respect to (w.r.t.) *Standard Polyglot*. In particular, CTG is calculated by

$$CTG = \frac{n_v - n_s}{n_s} \times 100, \text{where}$$

$$n_i = 100 - \left(\frac{i - l}{u - l} \times 100 \right), i \in \{s, v\},$$

with s and v representing the mean MUSHRA score for the *Standard* and *VC-based* approaches, respectively, and l and u corresponding to lower and upper anchors. All CTG percentages and Difference in CTG ($DCTG$) scores presented are for $p \leq 0.05$, unless otherwise specified by $^{n.s.}$.

4.2 Data composition

Our default training dataset D_1 consists of 124 speakers in 7 locales[2] (`fr-CA`, `en-GB`, `es-MX`, `en-US`, `es-US`, `fr-FR` and `de-DE`). This dataset totals to roughly 720 hours of studio-quality recordings with 24 kHz sampling rate. We extract 80-dimensional mel-spectrograms with frame length of 50 ms and frame shift of 12.5 ms, which is then used to train the VC model and the *Standard Polyglot* acoustic model. For the experiment on varying dataset size, we also test the proposed approach on a smaller dataset $D_2 \subset D_1$, which consists of data from a subset of speakers in `fr-CA`, `es-MX` and `en-GB` and totals to 150 hours.

4.3 Experiment designs and results

In this section, we describe the design and results of all experiments we run. Note that in all our evaluations, we compare our *VC-based Polyglot* approach (*VC-based*) to the *Standard Polyglot* approach (*Standard*) described in Section 1. In particular, Table 1 shows the evaluation results using different acoustic model architectures. Then, we further demonstrate the effectiveness of the *VC-based* approach by changing the `en-US` *target speaker* S_1 to `en-GB` speaker S_2 in Table 2, changing the *target locale* to `fr-CA` in Table 3, and changing the dataset to D_2 in Table 4. Finally, Table 5 shows that the *VC-based* approach renders even more superior performance when downsized. As for the training setup, we use the same hyper-parameters as described in [18,11,20]. Note that for the *VC-based Polyglot* approach, we train one single-speaker monolingual acoustic model for each *target locale*.

The goal of Experiment 1 (Table 1) is to assess the quality of the *VC-based* approach against the *Standard* baseline. To this end, we rely on dataset D_1, *target speaker* S_1, *target locale* `es-MX` and FS2 as model architecture. Table 1 shows the superior performance of our proposed approach in all of the evaluated aspects for the FS2 architecture. Replicating the same experiment with ED as acoustic model results in the same conclusions, except for the on-par performance in speaker similarity (more discussions on this in Section 4.4).

---

[2] We use the ISO 639-1 nomenclature to denote locales.

Table 1: MUSHRA and CTG scores for models evaluated in Experiment 1, for dataset D_1, speaker S_1 and locale `es-MX`.

Model	Approach	Naturalness$^\uparrow$	Speaker similarity$^\uparrow$	Accent similarity$^\uparrow$
Upper		82.60	100.00	82.25
FS2	*VC-based*	**69.60**	**66.38**	**71.06**
	Standard	64.08	64.82	66.26
	CTG	29.8 %	4.4 %	30.0 %
ED	*VC-based*	**70.21**	67.86	**70.49**
	Standard	65.24	68.05	66.27
	CTG	28.6 %	$-0.6\,\%^{n.s.}$	26.4 %
Lower		42.62	22.37	20.70

In Experiment 2 (Table 2), we focus on evaluating the robustness of the approach when changing the *target speakers* (S_1 to S_2). Table 2 shows that when we switch to *target speaker* S_2, the *VC-based* approach is still better than the *Standard* one in all aspects other than speaker similarity.

Table 2: MUSHRA and CTG scores for models evaluated in Experiment 2, for dataset D_1, speaker S_2 and locale `es-MX`.

System	Naturalness$^\uparrow$	Speaker similarity$^\uparrow$	Accent similarity$^\uparrow$
Upper	80.24	100.00	82.48
VC-based	**70.34**	67.60	**71.85**
Standard	69.86	68.69	70.54
CTG	4.5 %	$-3.5\,\%^{n.s.}$	10.9 %
Lower	34.47	28.37	19.64

With results from Experiment 3 (Table 3), we show the robustness of our approach when the *target locale* changes. In particular, Table 3 reveals that the *VC-based* approach is still better across all aspects, when changing the *target locale* to `fr-CA`.

Table 3: MUSHRA and *CTG* scores for models evaluated in Experiment 3, for dataset D_1, speaker S_1 and locale `fr-CA`

System	Naturalness$^\uparrow$	Speaker similarity$^\uparrow$	Accent similarity$^\uparrow$
Upper	75.76	100.00	76.52
VC-based	**74.11**	**70.69**	**70.99**
Standard	73.67	67.00	67.60
CTG	24.6 %	11.2 %	38 %
Lower	54.89	50.11	63.71

Furthermore, in Table 4 we show the robustness of the proposed approach when scaling down the size of the dataset. For this purpose, we repeat the experiment but switch to a smaller dataset D_2. As can be seen in Table 4, the proposed approach is preferred across the board.

Table 4: MUSHRA and *CTG* scores for models evaluated in Experiment 4, for dataset D_2, speaker S_1 and locale `fr-CA`.

System	Naturalness$^\uparrow$	Speaker similarity$^\uparrow$	Accent similarity$^\uparrow$
Upper	78.64	100.00	80.27
VC-based	**71.07**	**72.43**	**75.84**
Standard	66.66	68.57	73.97
CTG	36.8 %	12.3 %	29.9 %
Lower	61.55	44.80	44.30

Finally, to demonstrate the effectiveness of our proposed approach when scaling down model capacity, in Experiment 5 we repeat the previous experiments (with both `es-MX` and `fr-CA` as *target locales*), but use a "slimmed down" version of FS2 [18], i.e. LS [11]. To further test it in a more extreme scenario, we again downsize LS, thus producing LS-S, by reducing the size of all hidden layers from 256 to 192 and removing one convolutional layer from both

the encoder and the decoder. Table 5 reports the results in terms of the Difference in CTG ($DCTG$), which indicates how much more we close the gap with the smaller models w.r.t. the bigger ones. The $DCTG$ score is computed as $DCTG(CTG_s, CTG_b) = CTG_s - CTG_b$ percentage points (p.p.), where CTG_s is the CTG score for smaller models and CTG_b is the one for bigger models. Each cell of the table indicates the CTG score for a specific model in the *target locale*, representing the improvements gained from using the proposed approach. Overall, we can see that the *VC-based* approach outperforms the *Standard* baseline by a wider margin when lightweight architectures are employed.

Table 5: $(D)CTG$ scores in Experiment 5 for 4 MUSHRA tests, dataset D_1, speaker S_1 and locales `es-MX` and `fr-CA`.

Locale	Model	Naturalness$^\uparrow$	Speaker similarity$^\uparrow$	Accent similarity$^\uparrow$
es-MX	FS2	32.87 %	−2.71 %	25.42 %
	LS	42.27 %	−1.91 %	41.24 %
	$DCTG$	9.40 p.p.	0.80 p.p.$^{n.s.}$	15.82 p.p.
fr-CA	FS2	28.89 %	4.37 %	16.42 %
	LS	28.23 %	4.98 %	45.19 %
	$DCTG$	−0.66 p.p.	0.61 p.p.	28.77 p.p.
es-MX	FS2	27.55 %	−0.98 %	30.90 %
	LS-S	38.33 %	−0.23 %	35.08 %
	$DCTG$	10.78 p.p.	0.75 p.p.$^{n.s.}$	4.18 p.p.
fr-CA	FS2	30.49 %	−0.44 %	25.97 %
	LS-S	39.49 %	−0.64 %	35.54 %
	$DCTG$	9.00 p.p.	−0.20 p.p.$^{n.s.}$	9.57 p.p.

4.4   Discussion

As presented in Section 4.3, the proposed solution is preferred overall, while the only aspect where there is a non-statistically significant gap between *VC-based Polyglot* and *Standard Polyglot* (where the latter got more votes) is speaker similarity. We attribute this to the choice of the upstream VC model. In particular, although flow-based models are not widely-known to suffer from the speaker-accent trade-off, where speaker similarity degrades as accent similarity improves (and vice-versa), we believe that more powerful density-based models, such as those based on diffusion [15], have the potential of closing this gap in terms of speaker similarity.

We also hypothesise that the increase in naturalness scores can be attributed to the fact that in the *VC-based Polyglot* models the generated prosody follows that of the source speaker(s). In particular, many prosodic features are language-specific and inheriting them from source speakers (i.e. native speakers of the target language) shows to be beneficial. Future work will focus on assessing the impact of prosody transfer in this cross-lingual scenario.

Furthermore, we observe that the proposed approach is robust to variations in data composition: when comparing the data generated using VC models trained on vastly different datasets, there is no statistically significant preference. In comparison, previous works with the *Standard Polyglot* approach explicitly report how the mixture of training data has a significant impact on final performance [29].

5   Conclusions

In this paper, we proposed a novel paradigm for building highly robust Polyglot TTS systems. The introduction of a powerful upstream VC model successfully lifted the burden of learning disentangled speaker-language representations from the downstream acoustic model. This, in turn, allowed for effective use of high-performance architectures targeting low-powered devices. In summary, our proposed approach is able to simultaneously (1) outperform the current SOTA Polyglot TTS systems in terms of naturalness and accent similarity, while maintaining on-par performance in speaker similarity, (2) improve model robustness w.r.t. training data composition and (3) enable the use of lightweight architectures for speech synthesis on resource-constrained devices. Such results are backed by controlled experiments using different model architectures, languages, speakers and dataset sizes.

For future work, we will test how well our proposed approach scales in more diverse scenarios. In addition, we believe that by employing more advanced VC techniques, there would be further improvements in the downstream task, especially in terms of speaker similarity.

References

1. Bilinski, P., Merritt, T., Ezzerg, A., Pokora, K., Cygert, S., Yanagisawa, K., Barra-Chicote, R., Korzekwa, D.: Creating New Voices using Normalizing Flows. In: Proc. Interspeech 2022. pp. 2958–2962 (2022). https://doi.org/10.21437/Interspeech.2022-10195
2. Comini, G., Huybrechts, G., Ribeiro, M.S., Gabryś, A., Lorenzo-Trueba, J.: Low-data? No problem: low-resource, language-agnostic conversational text-to-speech via F0-conditioned data augmentation. In: Proc. Interspeech 2022. pp. 1946–1950. https://doi.org/10.21437/Interspeech.2022-10338
3. Ellinas, N., Vamvoukakis, G., Markopoulos, K., Maniati, G., Kakoulidis, P., Sung, J.S., Hwang, I., Raptis, S., Chalamandaris, A., Tsiakoulis, P.: Cross-lingual text-to-speech with flow-based voice conversion for improved pronunciation (2022). https://doi.org/10.48550/ARXIV.2210.17264, https://arxiv.org/abs/2210.17264

4. Finkelstein, L., Zen, H., Casagrande, N., an Chan, C., Jia, Y., Kenter, T., Petelin, A., Shen, J., Wan, V., Zhang, Y., Wu, Y., Clark, R.: Training Text-To-Speech Systems From Synthetic Data: A Practical Approach For Accent Transfer Tasks. In: Proc. Interspeech 2022. pp. 4571–4575. https://doi.org/10.21437/Interspeech.2022-10115

5. Hwang, M.J., Yamamoto, R., Song, E., Kim, J.M.: Tts-by-tts: Tts-driven data augmentation for fast and high-quality speech synthesis. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6598–6602 (2021). https://doi.org/10.1109/ICASSP39728.2021.9414408

6. ITUR Recommendation: Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra). International Telecommunications Union, Geneva (2001)

7. Jiao, Y., Gabryś, A., Tinchev, G., Putrycz, B., Korzekwa, D., Klimkov, V.: Universal neural vocoding with parallel wavenet. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6044–6048 (2021). https://doi.org/10.1109/ICASSP39728.2021.9414444

8. Karlapati, S., Karanasou, P., Łajszczak, M., Ammar Abbas, S., Moinet, A., Makarov, P., Li, R., van Korlaar, A., Slangen, S., Drugman, T.: CopyCat2: A Single Model for Multi-Speaker TTS and Many-to-Many Fine-Grained Prosody Transfer. In: Proc. Interspeech 2022. pp. 3363–3367 (2022). https://doi.org/10.21437/Interspeech.2022-367

9. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 17022–17033. Curran Associates, Inc. (2020)

10. Latorre, J., Bailleul, C., Morrill, T., Conkie, A., Stylianou, Y.: Combining speakers of multiple languages to improve quality of neural voices. In: Proc. 11th ISCA Speech Synthesis Workshop (SSW 11). pp. 37–42 (2021). https://doi.org/10.21437/SSW.2021-7

11. Luo, R., Tan, X., Wang, R., Qin, T., Li, J., Zhao, S., Chen, E., Liu, T.Y.: Lightspeech: Lightweight and fast text to speech with neural architecture search. In: ICASSP 2021. pp. 5699–5703

12. Merritt, T., Ezzerg, A., Biliński, P., Proszewska, M., Pokora, K., Barra-Chicote, R., Korzekwa, D.: Text-free non-parallel many-to-many voice conversion using normalising flow. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6782–6786 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746368

13. Miao, C., Liang, S., Chen, M., Ma, J., Wang, S., Xiao, J.: Flow-tts: A non-autoregressive network for text to speech based on flow. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7209–7213 (2020). https://doi.org/10.1109/ICASSP40776.2020.9054484

14. Nachmani, E., Wolf, L.: Unsupervised polyglot text-to-speech. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7055–7059 (2019). https://doi.org/10.1109/ICASSP.2019.8683519

15. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M.S., Wei, J.: Diffusion-based voice conversion with fast maximum likelihood sampling scheme. In: International Conference on Learning Representations (2022)

16. Qian, K., Zhang, Y., Chang, S., Yang, X., Hasegawa-Johnson, M.: AutoVC: Zero-shot voice style transfer with only autoencoder loss. Proceedings of Machine Learning Research, vol. 97, pp. 5210–5219. PMLR, Long Beach, California, USA (09–15 Jun 2019)

17. Ramani, B., Actlin Jeeva, M.P., Vijayalakshmi, P., Nagarajan, T.: Voice conversion-based multilingual to polyglot speech synthesizer for indian languages. In: 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013). pp. 1–4 (2013). https://doi.org/10.1109/TENCON.2013.6719019

18. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: Fastspeech 2: Fast and high-quality end-to-end text to speech. In: International Conference on Learning Representations (2021)
19. Sanchez, A., Falai, A., Zhang, Z., Angelini, O., Yanagisawa, K.: Unify and Conquer: How Phonetic Feature Representation Affects Polyglot Text-To-Speech (TTS). In: Proc. Interspeech 2022. pp. 2963–2967 (2022). https://doi.org/10. 21437/Interspeech.2022-233
20. Shah, R., Pokora, K., Ezzerg, A., Klimkov, V., Huybrechts, G., Putrycz, B., Korzekwa, D., Merritt, T.: Non-Autoregressive TTS with Explicit Duration Modelling for Low-Resource Highly Expressive Speech. In: Proc. 11th ISCA Speech Synthesis Workshop (SSW 11). pp. 96–101 (2021). https://doi.org/10.21437/SSW.2021-17
21. Sharma, M., Kenter, T., Clark, R.: StrawNet: Self-Training WaveNet for TTS in Low-Data Regimes. In: Proc. Interspeech 2020. pp. 3550–3554 (2020). https: //doi.org/10.21437/Interspeech.2020-1437
22. Shen, J., et al.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779–4783. https://doi.org/10.1109/ICASSP.2018. 8461368
23. Sun, L., Wang, H., Kang, S., Li, K., Meng, H.: Personalized, cross-lingual tts using phonetic posteriorgrams. In: Interspeech 2016. pp. 322–326 (2016). https: //doi.org/10.21437/Interspeech.2016-1043
24. Wan, L., Wang, Q., Papir, A., Moreno, I.L.: Generalized end-to-end loss for speaker verification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4879–4883 (2018). https://doi.org/10.1109/ICASSP. 2018.8462665
25. Wang, Z., Ge, W., Wang, X., Yang, S., Gan, W., Chen, H., Li, H., Xie, L., Li, X.: Accent and speaker disentanglement in many-to-many voice conversion. In: 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 1–5 (2021). https://doi.org/10.1109/ISCSLP49672.2021.9362120
26. Wu, J., Polyak, A., Taigman, Y., Fong, J., Agrawal, P., He, Q.: Multilingual text-to-speech training using cross language voice conversion and self-supervised learning of speech representations. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8017–8021 (2022). https: //doi.org/10.1109/ICASSP43922.2022.9746282
27. Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., Xie, L.: Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In: 2021 IEEE Spoken Language Technology Workshop (SLT). pp. 492–498 (2021). https://doi.org/10. 1109/SLT48900.2021.9383551
28. Zhang, Y., Weiss, R.J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R., Jia, Y., Rosenberg, A., Ramabhadran, B.: Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In: Proc. Interspeech 2019. pp. 2080–2084 (2019). https://doi.org/10.21437/Interspeech. 2019-2668
29. Zhang, Z., Falai, A., Sanchez, A., Angelini, O., Yanagisawa, K.: Mix and Match: An Empirical Study on Training Corpus Composition for Polyglot Text-To-Speech (TTS). In: Proc. Interspeech 2022. pp. 2353–2357 (2022). https://doi.org/10. 21437/Interspeech.2022-242
30. Zhao, G., Sonsaat, S., Levis, J., Chukharev-Hudilainen, E., Gutierrez-Osuna, R.: Accent conversion using phonetic posteriorgrams. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5314–5318. https://doi.org/10.1109/ICASSP.2018.8462258
31. Zhao, S., Nguyen, T.H., Wang, H., Ma, B.: Towards Natural Bilingual and Code-Switched Speech Synthesis Based on Mix of Monolingual Recordings and Cross-Lingual Voice Conversion. In: Proc. Interspeech 2020. pp. 2927–2931 (2020). https: //doi.org/10.21437/Interspeech.2020-1163