MULTILINGUAL TEXT-TO-SPEECH TRAINING USING CROSS LANGUAGE VOICE CONVERSION AND SELF-SUPERVISED LEARNING OF SPEECH REPRESENTATIONS

Jilong Wu[1]*, Adam Polyak*[1]*, Yaniv Taigman*[1]*, Jason Fong*[2]*, Prabhav Agrawal*[1]*, Qing He*[1]

[1]Facebook AI
[2] The University of Edinburgh

ABSTRACT

State of the art text-to-speech (TTS) models can generate high fidelity monolingual speech, but it is still challenging to synthesize multilingual speech from the same speaker. One major hurdle is for training data. It's hard to find speakers who have native proficiency in several languages. One way of mitigating this issue is by generating polyglot corpus through voice conversion. In this paper, we train such multilingual TTS system through a novel cross-lingual voice conversion model trained with speaker-invariant features extracted from a speech representation model which is pre-trained with 53 languages through self-supervised learning [1]. To further improve the speaker identity shift, we also adopt a speaker similarity loss term during training. We then use this model to convert multilingual multi-speaker speech data to the voice of the target speaker. Through augmenting data from 4 other languages, we train a multilingual TTS system for a native monolingual English speaker which speaks 5 languages(English, French, German, Italian and Spanish). Our system achieves improved mean opinion score (MOS) compared with the baseline of multi-speaker system for all languages, specifically: 3.74 vs 3.62 for Spanish, 3.11 vs 2.71 for German, 3.47 vs 2.84 for Italian, and 2.72 vs 2.41 for French.

Index Terms— multilingual text-to-speech, transfer learning, self-supervised learning, voice conversion.

1. INTRODUCTION

Recent advances in speech synthesis research witnessed the applications of neural network-based models such as WaveRNN [2], Tacotron2 [3] and Transformer-TTS [4] to synthesize natural and intelligible speech with high audio quality. Typically, these models are trained on a monolingual text-audio corpus. However, synthesizing speech from multilingual input data still is a challenging task. A natural way to train such a multilingual TTS system is using training data from either a bilingual or a polyglot speaker. For example, this method [5] presents a Chinese-English TTS system that is trained with recordings from bilingual speakers. However, it is hard to find multilingual speakers with native proficiency in several different languages. To address this problem, one category of approaches is focused on disentangling speaker and language information from a monolingual speech corpus. For instance,this work [6] builds a multilingual TTS system through cross-lingual voice cloning by incorporating an adversarial term in the loss function to enable the model to disentangle the speaker representation. Another paradigm to tackle the same problem of insufficient speaker-language pairs is explicitly generating that data using cross-lingual voice conversion.

In this paper, we propose a multilingual TTS system trained with polyglot speech corpus generated through a novel cross-lingual voice conversion model.To train voice conversion model, we first extract speaker-invariant features from a speech representation model pre-trained through self-supervised learning to leverage large amount of multilingual speech data [1]. Concatenated with speaker and pitch features, those extracted features from the pre-trained model are fed into a HiFiGAN [7] based neural vocoder to generate speech for the target speaker. We also apply multi-task learning with a loss on speaker embedding to ensure speaker identity similarity during training. Details of the proposed training and inference will be discussed in Section 3. Then, to build a multilingual TTS voice, we start with a set of monolingual TTS voices from multiple languages(e.g. Italian, French, Spanish, German). We use the proposed voice conversion model to convert the speaker identities of these non-English languages to the English speaker. Hence, we generate a synthetic dataset containing multilingual data consisting of Italian, French, Spanish and German languages for an English speaker. The synthetic dataset combined with the English speaker's original data is then used to train a single speaker multilingual TTS system with the same model architectures from this work [8]. The whole system of this paper is shown in Figure 1.

2. RELATED WORK

There have been several studies to use cross-lingual voice conversion [9] to build polyglot speech corpus for a multi-

Correspondence to Jilong Wu: jilwu@fb.com

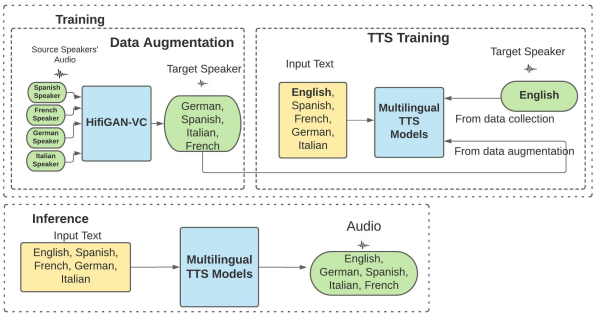

Fig. 1. Multilingual TTS System. Audio data is colored green and input text is colored yellow. Proposed HifiGAN-VC and multilingual TTS models (for the target speaker) are colored blue.

lingual TTS system [10, 11]. Among recent works, many papers have shown good performance using phonetic posteriorgram (PPG) as speaker-invariant features for the voice conversion model [12, 13]. However, a PPG based system needs to go through two stages in order to get desirable input features before training the audio synthesizer. It first needs to extract PPG features from target speech using a pre-trained Automatic Speech Recognition(ASR) model. Then it needs to train another model to map from target speech's PPG to source speaker's acoustic features (for example, mel spectrograms, etc). In this paper, instead of using PPG, we use self-supervised speech representations extracted from XLSR-53 [1], a wav2vec2.0-based [14] pre-trained cross-lingual model . Compared with PPG based approach, without training any extra model for feature mapping, our model can train the speech synthesizer directly with features extracted from a cross-lingual representation model. Previous work [15, 16, 17] showed such features enable high-quality audio generation.

3. METHOD

3.1. Multilingual Voice Conversion

The proposed voice conversion model, HiFiGAN-VC, is based on a generative adversarial network architecture similar to [16, 17, 18] and a HiFiGAN neural vocoder [7]. It is conditioned on speaker invariant content representations, spectral features, and, a learned speaker embedding for speaker identity to generate raw speech waveforms. Once trained, voice conversion can be achieved by changing the speaker identity representation to that of the desired speaker. HiFiGAN-VC is illustrated in Figure 2.

We denote the domain of audio samples by $\mathcal{X} \subset \mathbb{R}$. The representation for a raw signal is therefore a sequence of samples $\boldsymbol{x} = (x_1, \ldots, x_T)$, where $x_t \in \mathcal{X}$ for all $1 \leq t \leq T$.

Input Features Given an input, $\boldsymbol{x}$, we use XLSR-53 cross-lingual speech representation [1] as content representations. The representation is extracted via a wav2vec 2.0 [14] encoder, E_{w2v}, trained on raw speech waveforms in multiple languages. Speaker identity is controlled via a speaker embedding $\boldsymbol{v}$ as an additional input. The speaker embeddings are learned during training and stored in a lookup table. Finally, we use the fundamental frequency, $F_0(\boldsymbol{x})$, as a prosodic representation. We use YAAPT [19] to extract pitch contours.

The F_0 values are upsampled and concatenated to the content representation, $E_{w2v}(\boldsymbol{x})$. Then, the speaker embedding $\boldsymbol{v}$ is repeated and concatenated to each time-step of the former concatenation. To summarize, the encoding is given as $E(\boldsymbol{x}) = [E_{w2v}(\boldsymbol{x}), F_0(\boldsymbol{x}), \boldsymbol{v}]$.

Decoder A neural speech synthesizer is used to generate audio output from the extracted representation described above. We decided to use an adapted HiFiGAN [7] architecture to serve as the speech synthesizer. HiFiGAN is reported to produce good audio quality with fast inference speed. Fast inference enables us to efficiently generate large amounts of multilingual data in our data augmentation stage.

HiFiGAN uses a generative adversarial network (GAN) setup to produce raw waveforms. In our work we denote the HiFiGAN generator as G, which is a fully convolutional neural network that takes as input the encoding described above and outputs raw waveforms. The discriminator network D, is composed of two discriminators: (i) a multi-period discriminator (MPD) to handle periodic signals, and, (ii) a multi-scale discriminator (MSD) to handle long-term dependencies. Both the MPD and the MSD consists of multiple sub-discriminators, D_j, operating at different periods and scales accordingly. Specifically, the MPD employs a total of five period discriminators with period hops of $[2, 3, 5, 7, 11]$ and the MSD's three scale discriminators operate at: the original input scale, a $\times 2$ downsampled scale, and a $\times 4$ downsampled scale. Each sub-discriminator D_j minimizes the following loss functions,

$$
\begin{aligned}
L_{adv}(D_j, G) &= \sum_{\boldsymbol{x}} ||1 - D_j(\hat{\boldsymbol{x}})||_2^2, \\
L_D(D_j, G) &= \sum_{\boldsymbol{x}} [||1 - D_j(\boldsymbol{x})||_2^2 + ||D_j(\hat{\boldsymbol{x}})||_2^2],
\end{aligned}
\tag{1}
$$

where $\hat{\boldsymbol{x}} = G(E(\boldsymbol{x}))$, is the generator output.

Additionally, three more terms are added to the loss function. The first one is feature-matching loss [20] which measures the distance between discriminator activations of a real signal and those of the generator output,

$$
L_{fm}(D_j, G) = \sum_{\boldsymbol{x}} \sum_{i=1}^{R} \frac{1}{M_i} ||\psi_i(\boldsymbol{x}) - \psi_i(\hat{\boldsymbol{x}})||_1, \tag{2}
$$

where ψ_i is an operator which extracts the activations of the discriminator i-th layer, M_i is the number of features in layer i, and R is the total number of layers in D_j.

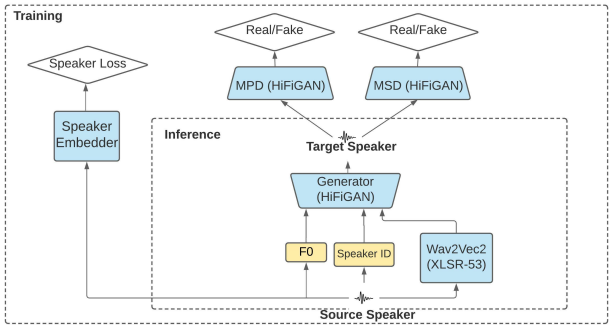

Fig. 2. HiFiGAN-VC. HifiGAN model, speaker embedder model and wav2vec2.0 pretrained model are colored blue. Input features of HiFiGAN-VC are colored yellow.

The second term is a reconstruction term computed between the mel-spectrogram of an input signal and the generated signal,

$$L_{recon}(G) = \sum_{\boldsymbol{x}} ||\phi(\boldsymbol{x}) - \phi(\hat{\boldsymbol{x}})||_1, \tag{3}$$

where ϕ is a spectral operator which computes the mel-spectrogram.

The third term is a speaker similarity term computed between the input signal and a signal converted to the identity of the input signal,

$$L_{spkr}(G) = \sum_{\boldsymbol{x}} 1 - cos(E_{id}(\boldsymbol{x}), E_{id}(\hat{\boldsymbol{y}})), \tag{4}$$

where $\boldsymbol{x}$ is an utterance spoken by a speaker with embedding $\boldsymbol{v}$, $\hat{\boldsymbol{y}} = G(E_{w2v}(\boldsymbol{y}), F_0(\boldsymbol{y}), \boldsymbol{v})$ is the conversion of sample $\boldsymbol{y}$ to the speaker with embedding $\boldsymbol{v}$, E_{id} is a speaker embedding network [21],and, cos is the cosine similarity between two vectors.

In conclusion, discriminator D and generator G minimize the following terms accordingly:

$$\begin{aligned} L_G^{multi}(D,G) =& \sum_{j=1}^{J}[L_{adv}(G,D_j) + \lambda_{fm}L_{fm}(G,D_j)], \\ &+ \lambda_r L_{recon}(G) + \lambda_{spkr}L_{spkr}(G), \\ L_D^{multi}(D,G) =& \sum_{j=1}^{J} L_D(G,D_j), \end{aligned} \tag{5}$$

where D_j is the j-th sub-discriminator of D. We set $\lambda_{fm} = 2$, $\lambda_r = 45$, and, $\lambda_{spkr} = 1$.

Data Augmentation With a trained voice conversion model for the target speaker, we convert our multilingual multi-speaker speech dataset to the target speaker's voice. Formally, we denote the domain of transcription samples by $\mathcal{S} \subset \mathbb{R}$. The representation for a transcription is therefore a sequence of samples $\boldsymbol{s} = (s_1, \ldots, s_T)$, where $s_t \in \mathcal{S}$ for all $1 \leq t \leq T$. Similar to raw audio signals, the length of the transcriptions varies for different samples, thus the number of input samples in the sequence, T, is not fixed. A TTS training dataset is composed of n paired examples of text and audio, $S = \{(\boldsymbol{s}_i, \boldsymbol{x}_i)\}_{i=1}^n$, where $\boldsymbol{s}_i$ is the text transcription of audio $\boldsymbol{x}_i$. Given the desired speaker embedding $\boldsymbol{v}$, we generate a single speaker training dataset, $\tilde{S} = \{(\boldsymbol{s}_i, \hat{\boldsymbol{x}}_i) \mid \hat{\boldsymbol{x}}_i = G(E_{w2v}(\boldsymbol{x}_i), F_0(\boldsymbol{x}_i), \boldsymbol{v})\}_{i=1}^n$.

3.2. Multilingual TTS System

Our TTS system uses the same system described in [8] which combines multi-rate attention based acoustic and prosody models with an adapted WaveRNN-based neural vocoder. The linguistic frontend adopts the supersegmental International Phonetic Alphabet (IPA) representation [22].The linguistic features are rolled out based on the phone level duration predictions from the prosody model. Then they are up-sampled by repetition to generate frame level features which are used by the acoustic model. Additionally other categories of input features are extracted: sentence level features like language ID and speaker ID; phrase level features such as intonation; word level features such as word embedding; syllable level features such as syllable type; and phone level features such as articulation diacritics. These features are consumed by the acoustic model to predict pitch features, periodicity features and mel-frequency cepstrum features (MFCC). Both acoustic and prosody models are trained using the Mean Square Error (MSE) loss function as described in the paper [8]. For the last step, we use an adapted version of WaveRNN neural vocoder to condition on the predicted acoustic features and synthesize speech for the target speaker.

4. EXPERIMENTS

4.1. Datasets

Both HiFiGAN-VC and TTS models are trained with our internal dataset, which was recorded in a voice production studio by contracted professional voice talents. Each voice talent speaks only one language. Our voice data are from voice talents speaking English, German, Italian, French and Spanish. We use 24kHz recorded audio from 8 speakers: 3 English speakers (32 hours, 13 hours, 13 hours), 1 French speaker(28.5 hours), 1 German speaker (8.5 hours), 1 Spanish speaker (23.7 hours) and 3 Italian speakers (9.8 hours, 9.7 hours, 12.3 hours).

In this work, we train a multilingual TTS system using one of the English speakers' voice which is able to synthesize audio in 4 other languages: Spanish, German, French, and Italian. For internal dataset, we only have target speaker's data speaking English. To obtain data for other 4 languages, we first train a HiFiGAN-VC with internal data from 8 speakers covering targeted 4 languages, as described above. Then

we use audio dataset from 4 speakers who are monolingual and speak native German, Spanish, Italian and French respectively and convert their speakers' voice to the identity of our target English speaker. We now have a set of data with target speaker's voice obtained from voice conversion in 4 languages: Spanish (8 hours), French (4 hours), German (7 hours) and Italian (9 hours) and data from paid recording in English (32 hours).

Table 1. Speaker Identities Equal-Error-Rate(%)

Models	ES	DE	IT	FR
Base_multi	44.0	29.3	60.0	40.5
Base_mono	2.0	11.1	9.5	4.0
Proposed	2.0	14.5	18.3	19.0

4.2. Evaluation

We evaluate the multilingual TTS system against three baselines. The first baseline is a TTS system trained only with the target speaker's recorded monolingual data (Base_mono). The second baseline (Base_multi) is a multilingual multi-speaker TTS system trained with 8 speakers' data described in *Datasets*. The training data has the same set of language-speaker pairs as those used in the proposed voice conversion model. This baseline model relies on transfer learning to enable target English speaker to speak multiple languages. A third baseline (Base_vc) is set up as the following: we first train separate monolingual TTS models for each target language. Then we use those TTS models to synthesize multilingual audio samples with native speakers' voice of that language. We then use the proposed voice conversion model to convert those audio samples to the target English speaker's voice. This baseline is expected to deliver high quality because the TTS system is trained in their native language. However, this baseline system is undesirable for practical applications because it's computationally expensive to run a TTS followed by a voice conversion model during inference time. All the TTS systems above are using the same spectrum, prosody, and neural vocoder models.

4.2.1. Subjective evaluation

For evaluation, we conducted a mean opinion score (MOS) study which used test set utterances excluded from the training data. We used a crowdsourcing platform which recruited the following number of participants for each test set language: 277 Spanish, 39 French, 290 Italian, and 39 German. Each rater rated randomly assigned 30 samples. For each MOS test, the participants rated each sample between 1-5 (1:bad - 5:excellent). We asked participants to rate the audio samples in terms of overall naturalness and intelligibility. As you can see from Figure 3, for each language, our proposed model achieves higher MOS than the multi-speaker and mono-speaker baselines. Even compared with Base_vc whose TTS models are trained with native speakers, the scores are comparable. A selection of samples can be found on the webpage of this paper[1].

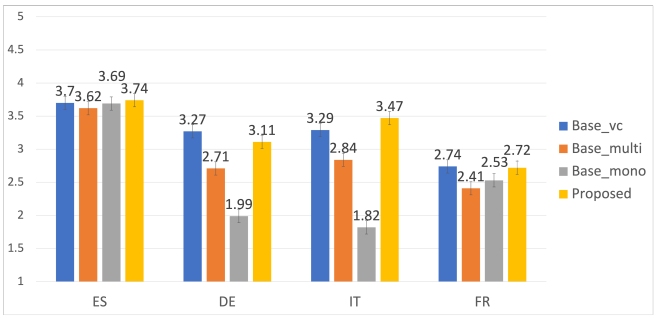

Fig. 3. MOS results described in the section of subjective evaluation. 95% confidence intervals are depicted as black lines. For each language, from left to right: Base_vc (colored blue), Base_multi (colored orange), Base_mono (colored grey), Proposed model (colored yellow). Languages evaluated are Spanish(ES), German(DE), Italian(IT) and French(FR).

4.2.2. Objective evaluation

Besides MOS studies, we also conducted an objective evaluation on speaker similarities. We enrolled the 5 speakers from *Data Augmentation* with a trained speaker embedding network [21]. We calculated the Equal-Error-Rate (EER) based on speaker embeddings' multi-class similarity matrix. The results are shown in table 1. As you can see, there is a noticeable improvement compared with Base_multi which is trained with multi-speaker's data. Also the proposed model narrows the gap between the TTS model(Base_mono) trained with data purely from recordings of the target speaker.

5. CONCLUSIONS

In conclusion, we propose a new way to train multilingual TTS with data augmentation through voice conversion. The voice conversation model is trained with wav2vec2.0 representations and uses HiFiGAN as the audio synthesizer. Our MOS study shows that our proposed multilingual model achieves better quality compared to the multi-speaker or single-speaker baselines. It also has comparable quality with baseline TTS models trained with native speakers' dataset. For future work, we plan to explore other audio synthesizer architectures to further improve the quality of the voice conversion model.

[1] https://multilingual-tts-data-aug.github.io/multilingual_tts_data_aug/

6. REFERENCES

[1] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[2] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj 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)*, pages 4779–4783. IEEE, 2018.

[4] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713, 2019.

[5] Huaiping Ming, Yanfeng Lu, Zhengchen Zhang, and Minghui Dong. A light-weight method of building an lstm-rnn-based bilingual tts system. In *2017 International Conference on Asian Language Processing (IALP)*, pages 201–205. IEEE, 2017.

[6] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*, 2019.

[7] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020.

[8] Qing He, Zhiping Xiu, Thilo Koehler, and Jilong Wu. Multi-rate attention architecture for fast streamable text-to-speech spectrum modeling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5689–5693. IEEE, 2021.

[9] Sai Sirisha Rallabandi and Suryakanth V Gangashetty. An approach to cross-lingual voice conversion. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2019.

[10] P Vijayalakshmi, B Ramani, MP Actlin Jeeva, and T Nagarajan. A multilingual to polyglot speech synthesizer for indian languages using a voice-converted polyglot speech corpus. *Circuits, Systems, and Signal Processing*, 37(5):2142–2163, 2018.

[11] B Ramani, MP Actlin Jeeva, P Vijayalakshmi, and T Nagarajan. Cross-lingual voice conversion-based polyglot speech synthesizer for indian languages. In *Fifteenth annual conference of the international speech communication association*, 2014.

[12] Qinghua Sun and Kenji Nagamatsu. Building multi lingual tts using cross lingual voice conversion. *arXiv preprint arXiv:2012.14039*, 2020.

[13] Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, and Bin Ma. Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion. *arXiv preprint arXiv:2010.08136*, 2020.

[14] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[15] Adam Polyak, Lior Wolf, and Yaniv Taigman. TTS skins: Speaker conversion via asr. *INTERSPEECH*, 2020.

[16] Adam Polyak et al. Unsupervised Cross-Domain Singing Voice Conversion. In *INTERSPEECH*, 2020.

[17] Adam Polyak et al. High fidelity speech regeneration with application to speech enhancement. *ICASSP*, 2021.

[18] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. *INTERSPEECH*, 2021.

[19] K. Kasi and S. A. Zahorian. Yet another algorithm for pitch tracking. *ICASSP*, 2002.

[20] Anders Boesen Lindbo Larsen et al. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016.

[21] Li Wan et al. Generalized end-to-end loss for speaker verification. *ICASSP*, 2018.

[22] International Phonetic Association. The international phonetic alphabet, 2015. `https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf`.