PAD-VC: A PROSODY-AWARE DECODER FOR ANY-TO-FEW VOICE CONVERSION

Arunava Kr. Kalita[1], *Christian Dittmar*[2], *Paolo Sani*[2], *Frank Zalkow*[2],
Emanuël A. P. Habets[3], *Rusha Patra*[1]
[1]IIIT Guwahati, India
[2]Fraunhofer IIS, Erlangen, Germany
[3]International Audio Laboratories Erlangen†, Germany

ABSTRACT

Voice Conversion (VC) generates synthetic speech from a source speaker recording, preserving linguistic information and applying the voice characteristics of a target speaker. In this paper, we propose PAD-VC, a prosody-aware VC model based on the decoder part of the ForwardTacotron architecture. We train PAD-VC with prosody-related features such as pitch, energy, and voicing confidence and augment those with linguistic features derived from a phoneme posteriorgram representation of the source utterance. This way, we can handle both phonemic information and frame-wise supra-segmental features. During inference time, the source speaker's prosody features are modified to match the prosody statistics of the target speaker. We show that PAD-VC outperforms ForwardTacotron in prosody-cloning for unseen source utterances, achieving higher similarity and naturalness.

1. INTRODUCTION

Voice Conversion (VC) aims at generating synthetic speech signals, combining the linguistic content of utterances spoken by a source speaker with the voice characteristics of a target speaker [1]. Thereby, the goal is to produce speech that sounds natural and exhibits a high similarity to the target speaker's voice. Some VC systems use prosody-related features like phoneme durations and f_0 contour (pitch) along with the linguistic content. We refer to those systems as prosody-aware systems. VC has various real-life applications in speech pathology [1], entertainment [2, 3], and education [4]. In addition, VC can be utilized as a tool to create plausible synthetic speech data to augment training corpora used for speech enhancement and speech recognition.

1.1. Related Work

A variety of methodologies have been proposed for the VC task, as shown in several studies [1, 5]. Early techniques focused on altering voice timbre, often overlooking finer speaker-specific attributes [2]. Usually, these methods relied on parallel training data, where the utterances spoken by the source and target speakers needed to have the same phonetic content. In contrast, more recent VC techniques enable the development of unsupervised systems that can capture and transfer the supra-segmental aspects of human speech such as intonation, stress patterns, and rhythm with basic phonetic content to achieve more convincing voice conversion.

Most VC systems use mel spectrograms as the primary speech representation, encoding all relevant features. However, it is not trivial to disentangle the speech features of interest from there. Moreover, mel spectrograms always have to be converted back to the time domain to yield an audible speech signal. The standard procedure is to use neural vocoders like HiFiGAN [6] or StyleMelGAN [7] for this purpose.

The majority of deep learning architectures employed in VC can be broadly classified into encoder-decoder architectures and Generative Adversarial Networks (GANs). Recent encoder-decoder approaches include AutoVC [8], SpeechSplit [3], and f_0-AutoVC [9]. They are based on the principle that the encoder tries to disentangle the source utterance's linguistic content as a latent from the input feature representation, while the decoder tries to generate a mel spectrogram with the desired target speaker characteristics from the latent. This approach requires a careful bottleneck tuning of the encoder to achieve a desirable disentanglement. GAN-based approaches like StarGAN [10], however, rely on training a generator to convert input speech features from the source to the target speaker's voice, while a discriminator simultaneously distinguishes between natural and synthetic speech representations. Some GAN-based VC methods leverage pre-trained speaker embeddings, such as ECAPA-TDNN and x-vectors [11, 12]. A recent extension [6] of StarGAN is tailored for converting expressive speech. Here, multiple encoders are trained with task-specific losses to capture linguistic content, speaker characteristics, and prosody. In general, these methods can be challenging to train and prone to overfitting. As an alternative to the previous approaches, TTS-based VC systems use textual input to synthesize speech in the target speaker's voice. To this end, an acoustic model is trained to learn phoneme and speaker embeddings, enabling independent swapping of spoken content and speaker identity. Some TTS-based VC methods are prosody-aware, as they transfer prosodic attributes like phoneme-wise duration, pitch, and energy [13] to the target speaker's voice. However, these approaches can be error-prone, as they rely on the extraction of prosody-related features on a phoneme-by-phoneme basis. As an alternative, the use of mid-level linguistic content representations like phoneme posteriorgrams (PPG) has been explored in previous works [14, 15].

1.2. Our Contribution

In this paper, we introduce two novel prosody-aware VC approaches based on the ForwardTacotron [1] (FT) architecture. Given the strong performance of FT in conventional TTS settings [16], we adapt it for VC. More precisely, we use an extended version of FT that can process prosody-related features [17–19], hereafter referred to as FT-VC. To our knowledge, this particular architecture has not been used before for VC, although the concept of prosody-cloning has been described in previous works [13].

As a second contribution, we propose PAD-VC, which uses

† A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

[1] `https://github.com/as-ideas/ForwardTacotron`

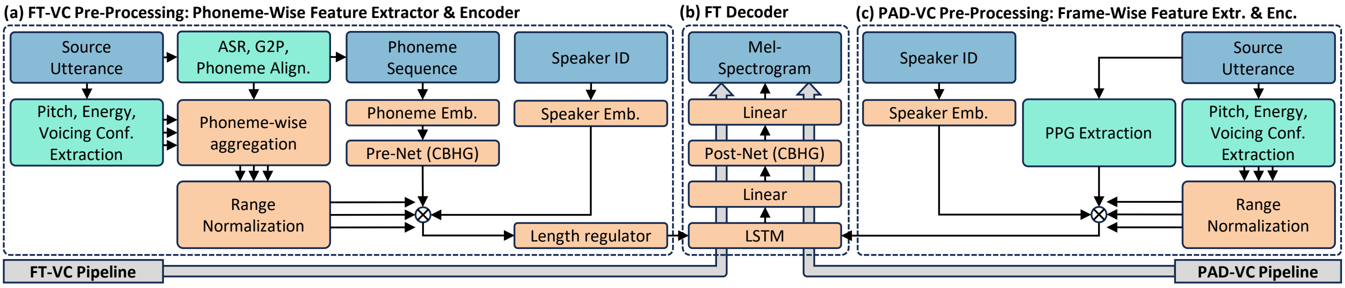

Fig. 1. Overview of the two alternative TTS-based VC approaches explored in this paper. **(a)** Baseline FT-VC pre-Processing **(b)** FT Decoder **(c)** Proposed PAD-VC pre-Processing. As indicated by the light gray arrows, the FT decoder can either be driven by **(a)** or **(c)**, not both concurrently. Blue: Data streams, Cyan: External modules, Apricot: Internal modules, ⊗: Concatenation of data streams.

only the prosody-aware decoder part of FT-VC in conjunction with a Phoneme Posteriorgram (PPG) representation of the spoken content in the source utterance. We will explain in Sec. 2 how PAD-VC can mitigate some inherent issues of FT-VC. While using PPGs for voice conversion has been proposed in other works before [14], we focus on extracting disentangled and interpretable speech features for VC. At the time of writing this manuscript, the paper [15] has come to our attention, which is conceptually very close to PAD-VC. The authors focus on comparing different input representations for PPG extraction w.r.t. pitch-disentanglement and re-synthesis quality. Our work was developed independently, and it remains subject to future work to compare both approaches.

In summary, the main contributions of this paper are as follows: 1) We repurpose FT as the baseline FT-VC and use it to convert unseen source utterances into pre-trained TTS voices. 2) We detail how to explicitly extract interpretable speech features and convert them to match target speaker characteristics. 3) We introduce PAD-VC, a modified decoder-only FT suitable for VC using PPGs and prosody-related features. 4) We compare PAD-VC with the FT-VC approach through subjective listening tests, which focus on the naturalness and speaker similarity of the synthesized speech signals.

2. METHODOLOGY

In the following sections, we first briefly describe the prosody-aware FT architecture and its adaptation for VC. Second, we detail the proposed PAD-VC architecture and explain the differences in design compared to FT-VC. Figure 1 provides an overview of the data flow at inference time in both pipelines. Note that all external modules depicted in Fig. 1 are to be interpreted as pre-trained and fixed.

2.1. Prosody-Aware ForwardTacotron

ForwardTacotron is a non-autoregressive acoustic model for TTS, consisting of an encoder, a length regulator, and a decoder part. The main purpose of the encoder part is to transform given phoneme sequences and speaker identity codes into internal vector representations. In addition, scalar values for f_0, energy, and voicing confidence are concatenated to this hidden representation (see Sec. 3.1 for details). The length regulator then resamples the resulting vectors in a non-equidistant fashion to the temporal dimension of the target mel spectrogram by means of nearest neighbor interpolation (i.e., replication over several frames). Finally, the FT decoder shown in Fig. 1(b) converts this coarse internal representation into a plausible mel spectrogram with the desired spoken content, prosody, and target voice timbre.

2.2. ForwardTacotron Voice Conversion (FT-VC)

In a TTS-based VC scenario, the phoneme sequence describing the linguistic content in the source utterance must be known. To this end, we use an Automatic Speech Recognition (ASR) system, whose transcription is run through Grapheme-to-Phoneme (G2P) conversion. FT-VC requires the length regulator to have duration information for each phoneme. We use a pre-trained phoneme aligner [20] to estimate those durations given the G2P output and the source utterance. The typical result of speech-phoneme alignment is illustrated in Fig. 2(b). Note that we grouped the cascade of ASR, G2P, and phoneme aligner into a single, external module in Fig. 1(a) to avoid clutter. For the same reason, we did not draw the necessary connection from the aligner to the length regulator. The role of the range normalization module is explained in Sec. 3.1.

In practice, we observe three main problems with FT-VC: 1) The ASR and G2P are language-dependent, and any error they make propagates into the downstream processing. 2) The phoneme aligner can exhibit inaccuracies when processing utterances of unseen speakers, especially for expressive speech. While fine-tuning on the source utterance can help [13], we want to avoid retraining and have only fixed modules in the pipeline. 3) Through the phoneme-wise aggregation, frame-wise prosody features are first collapsed into an average value across the duration of the corresponding phoneme, only to be expanded back to their original temporal extent later on by the length regulator.

2.3. Prosody-Aware Decoder Voice Conversion (PAD-VC)

The proposed PAD-VC approach mitigates the aforementioned problems by making use of frame-wise prosody features as well as frame-wise content representations in the form of PPGs (see Sec. 3.2). As indicated by its name and depicted in Fig. 1(c), the PAD-VC architecture can be interpreted as a truncated, decoder-only FT-VC. The range normalization and the speaker embedding layer are retained from the FT-VC architecture. The rationale behind using PPGs is to have an interpretable mid-level representation of the spoken content that is not as rigid as discrete symbolic phoneme sequences but a probability of phoneme occurrences over time. It also captures fine pronunciation details in the source utterance through gradual transitions and possibly overlapping phoneme activations. We illustrate the correspondence between PPGs and phoneme sequences in Fig. 2. Not having to use ASR and G2P simplifies the complexity and reduces susceptibility to error-propagation of the pipeline. Not working with phoneme sequences also renders the phoneme aligner and length regulator mechanisms obsolete. This way, we avoid the process of first aggregating and then expanding the frame-wise features again. Instead, both prosody-related and PPG features are resampled by linear interpolation to the temporal dimension of the target mel spectrogram. In summary, PAD-VC is a streamlined method to derive a similarly powerful internal representation as FT-VC to be subjected to FT decoding for any-to-few VC.

3. DATASET AND FEATURES

For this paper, we use a multi-speaker English dataset comprising recordings from four distinct speakers: ‘female1’ [21] (22.89 hours), ‘male1’ [22] (5.43 hours), ‘female2’ (2.22 hours), and ‘male2’ (2.14 hours). The first two datasets comprise recordings from native English speakers with a British accent. The latter two datasets are proprietary and feature recordings in the English language by professional voice actors with an audible German accent. Each dataset includes pairs of weakly aligned phoneme annotations and corresponding speech recordings. As a pre-processing step, the recordings are downsampled to 22,050 Hz, high-pass filtered to remove DC offsets, and normalized w.r.t. maximum absolute amplitude. As shown in Fig. 1(b), we use mel spectrograms as the target speech representation. They are extracted with 80 bands, using a hop size of 256 samples and a block size of 1024 samples.

3.1. Prosodic Features

As shown in Figs. 1(a) and (c), we extract prosody features in a frame-wise manner from the source utterance. They comprise energy (computed as the L2-norm of mel spectrogram frames), pitch (f_0 in Hertz), and voicing confidence (saliency of the pitch estimate). We use CREPE [23] as a pitch extractor. Obviously, those features are not necessarily speaker-agnostic. Pitch, for example, varies among individuals due to factors like age, gender, and physiological traits. Thus, we normalize the pitch features of the source utterance to zero mean and unit variance using the source speaker’s statistics. We subsequently adapt them by scaling and translation to match the statistics of the target speaker. Consistent with previous work, pitch estimates below a voicing confidence threshold of 0.5 are disregarded in the calculation of the mean and variance. The range normalization module in both FT-VC and PAD-VC is responsible for this adaption.

3.2. Frame-Wise vs. Phoneme-aligned Features

The subsequent feature processing steps differ between FT-VC and PAD-VC. In FT-VC, we aggregate the frame-level features to the phoneme level. As a pre-requisite, the given weak alignments between phoneme transcriptions and mel spectrograms need to be refined into a more accurate phoneme-wise alignment, which specifies the temporal correspondences between mel frames and individual phonemes in the transcription. For this task, we utilize the aligner model [20]. It uses the Connectionist Temporal Classification (CTC) paradigm [24] to temporally align phoneme sequences with mel spectrograms. In contrast, for PAD-VC training, we use frame-wise features and simply resample them by means of linear interpolation along the time-axis to match the temporal dimension of the target mel spectrograms. While these steps may seem like a coarse approximation, they work surprisingly well in practice since the recurrent units in the FT decoder still provide enough modeling capacity to compensate for slight inaccuracies at the frame level.

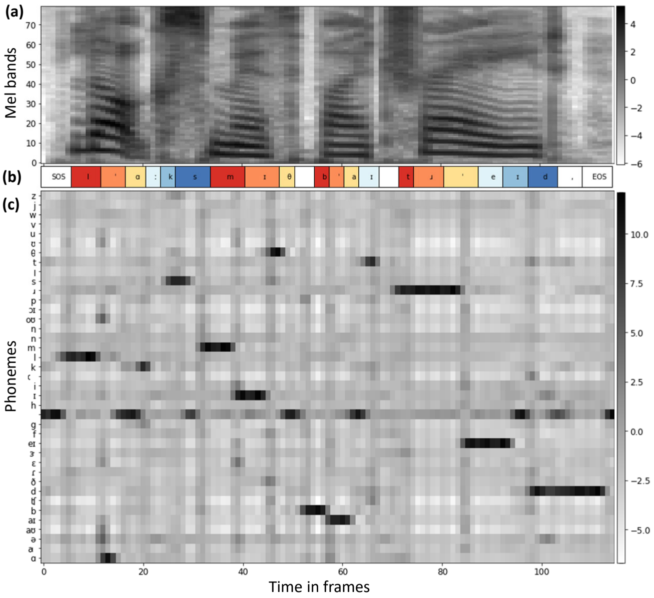

Fig. 2. Different input representations for the source utterance **“Locksmith by trade.”** **(a)** Mel spectrogram, **(b)** time-aligned phoneme sequence, **(c)** phoneme posteriorgram. In some segments, a high agreement between the PPG and the aligned phoneme sequence can be observed, e.g., for the IPA symbol **[m]**, active between frame 30 and 40.

3.3. Phoneme Posteriorgram

For PAD-VC, a speaker-independent ASR module is used to extract the PPG from the source utterance. A PPG consists of a temporal sequence of probability vectors, where each vector corresponds to a probability distribution over phoneme symbols, following the International Phonetic Alphabet (IPA), as shown in Fig. 2(c). We use a pre-trained model [2] as a PPG extractor. It is based on Wav2Vec [25] and the CTC paradigm [24]. The raw PPG output of this system also contains entries for non-required symbols, such as padding, unknown, word boundary, and sentence delimiter symbols, which often have high probabilities. We post-process the PPG and remove these entries. If the highest probability for a given frame corresponds to one of the removed entries, we add its probability to this frame’s value of the probability-maximizing phoneme from the previous frame, effectively prolonging the activity of the high-probability phonemes in the PPG. A similar methodology has been found to produce valid results [15, 26]. This representation is used as a speaker-agnostic encoding of the linguistic content of a source utterance.

4. EXPERIMENTS

4.1. Experimental Setup

Both FT-VC and the proposed PAD-VC model are trained on an NVIDIA 1660 Ti GPU with the datasets described in Sec. 3 for 300k steps, using a batch size of 16. A learning rate scheduler is employed with the Adam optimizer using an initial value of 10^{-4}, which decreases to 10^{-5} at a predetermined step count.

[2] `https://github.com/ASR-project/Multilingual-PR`

4.2. Inference

In this section, we provide a detailed explanation of how inference is performed with both FT-VC and PAD-VC in our experiments. Initially, in FT-VC, prosody features are extracted from the source utterance. In addition, we apply the popular ASR system Whisper [3] to transcribe the utterances and convert their transcripts to phoneme sequences by means of a British English pronunciation dictionary. Later, we use the phoneme sequence, the modified prosody features, and the desired speaker identifier as input to the FT decoder to synthesize the source utterance in the target speaker's voice. For PAD-VC, apart from the change in linguistic content representation, the same prosody features used as in FT-VC are extracted. Unlike in FT-VC, they are not subjected to phoneme-wise aggregation but still undergo the same scale-and-shift adaption. For both FT-VC and PAD-VC, the mel spectrograms predicted by the FT decoder are converted into time-domain speech signals using a pre-trained StyleMelGAN [7] neural vocoder.

4.3. Listening Tests

We used audio excerpts from the Expresso dataset [4] as source utterances for listening tests. The dataset comprises four speakers, each contributing to two distinct categories: spontaneous conversational and read speech. Our source utterances are drawn from the read speech category with a neutral reading style. While selecting the samples, we ensured that the source speakers were distinct from our target speakers and that the utterances covered a wide range of phonetic content variations. The listening test was conducted in a MUSHRA-like test environment [27] for two aspects: speaker similarity and speech naturalness. Participants were asked to assess the voice-converted audio samples generated by both FT-VC and PAD-VC in comparison to a copy-synthesis reference. The listening tests included intra-gender and inter-gender conversions, i.e., the source and the target speakers were chosen for male-to-male (M2M), male-to-female (M2F), female-to-male (F2M), and female-to-female (F2F) conversions. In the evaluation, we summarized the listening test scores for these gender conversion types into a single result. However, PAD-VC exhibits similar performance across different gender conversion types, which illustrates the robustness of the proposed method.

In the first test, we evaluated the speech naturalness. The participants were instructed to rate the outputs for the two different approaches on a scale from 1 to 5, where 1 indicates a perception of the output as being very unnatural, and 5 indicates a perception of the output as being very natural. In the second test, we evaluated the speaker similarity to assess how closely the synthesized speech resembles the characteristics of the target speaker. In VC, a central objective is to ensure that the converted speech closely mimics the target speaker's characteristics, such as pitch, timbre, intonation, and accent. To evaluate this property, the participants were again instructed to assess the outputs from our two distinct approaches using a scale from 1 to 5. A score of 1 suggests that the synthesized sample's speaker characteristics are very dissimilar to the reference's, while a score of 5 indicates they are very similar.

5. RESULTS AND DISCUSSION

In total, 14 listeners participated in both tests, including eight males and six females, with no known hearing impairments and an average age of 31.4 years. Both listening tests were taken independently. We had to exclude two listeners from the statistical evaluation since they had given the copy-synthesis references a lower rating than the VC examples.

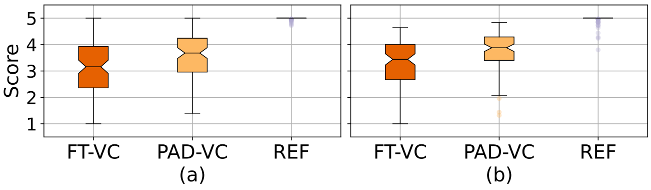

Fig. 3. Boxplots for the results of **(a)** the speech naturalness test and **(b)** the speaker similarity test, where our FT-VC and PAD-VC approaches were evaluated along with the reference (REF) obtained by copy synthesis.

The results for the two assessments are depicted in Fig. 3. Specifically, Fig. 3(a) shows the boxplots related to speech naturalness, and Fig. 3(b) shows the speaker similarity results, respectively. It is apparent from the plot that the proposed PAD-VC system performs better than the baseline FT-VC for both speaker similarity and speech naturalness tests. As expected, the copy-synthesis reference is preferred by the listeners, which indicates that there is room for further improvement [5]. It is important to note that the source samples are unseen and exhibit distinctive speaker characteristics in comparison to the training data (e.g., American accents). Nevertheless, the proposed PAD-VC system effectively transfers those utterances with their prosodic nuances to the target speakers' voices. One reason for the poorer performance of the FT-VC system is its reliance on phoneme durations, where possible misalignments lead to a reduction in the clarity of linguistic content.

6. CONCLUSION

We proposed PAD-VC to perform prosody-aware voice conversion with a modified FT decoder. To this end, we combine a phonetic content representation with the prosody-related features of the source speaker. A simple yet effective shift-and-scale technique normalizes the source utterance's f_0 and energy to the pre-calculated target speaker range. In subjective listening tests, we compared the capabilities of PAD-VC versus the baseline FT-VC approach for any-to-few VC. This study lays the groundwork for further TTS-based VC research, inviting future exploration to address its complexities and nuances. Furthermore, we believe it is promising to investigate the suitability of PAD-VC for additional scenarios like any-to-any VC and singing VC.

7. ACKNOWLEDGEMENT

The author conducted this work during his stay in Fraunhofer IIS, Erlangen, with the TTS group. This internship was made possible through the generous support and funding provided by IGSTC. This research was partially supported by the Free State of Bavaria in the DSAI project and by the Fraunhofer-Zukunftsstiftung.

[3] `https://github.com/openai/whisper`

[4] `https://speechbot.github.io/expresso`

[5] We provide all samples used in the listening tests on the accompanying website: `https://audiolabs-erlangen.de/resources/NLUI/2024-PAD-VC`

8. REFERENCES

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 285–288 vol.1, 1998.

[3] C. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, "Speechsplit2.0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, 2022, pp. 5243–5247.

[4] A. Pérez, G. G. Díaz-Munío, A. Giménez, J. A. Silvestre-Cerdà, A. Sanchis, J. Civera, M. Jiménez, C. Turró, and A. Juan, "Towards cross-lingual voice cloning in higher education," *Engineering Applications of Artificial Intelligence*, vol. 105, p. 104413, 2021.

[5] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied Sciences*, vol. 13, no. 5, p. 3100, 2023.

[6] S. Ghosh, A. Das, Y. Sinha, I. Siegert, T. Polzehl, and S. Stober, "Emo-StarGAN: A semi-supervised any-to-many non-parallel emotion-preserving voice conversion," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*. ISCA, 2023.

[7] A. Mustafa, N. Pia, and G. Fuchs, "StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 6034–6038.

[8] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, vol. 97, 2019, pp. 5210–5219.

[9] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Virtual Barcelona: IEEE, 2020.

[10] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*, 2021.

[11] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 912–919.

[12] Ü. E. Gaznepoglu and N. Peters, "Deep learning-based F0 synthesis for speaker anonymization," in *Proc. of the European Signal Processing Conf. (EUSIPCO)*. Helsinki, Finland: IEEE, 2023, pp. 291–295.

[13] F. Lux, J. Koch, and N. T. Vu, "Exact prosody cloning in zero-shot multispeaker text-to-speech," in *Proc. of the Spoken Language Technology Workshop (SLT)*. Doha, Qatar: IEEE, 2022, pp. 962–969.

[14] L. Sun, K. Li, H. Wang, S. Kang, and H. M. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. of the Int. Conf. on Multimedia and Expo, ICME*, Seattle, WA, USA, 2016, pp. 1–6.

[15] C. Churchwell, M. Morrison, and B. Pardo, "High-fidelity neural phonetic posteriorgrams," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*, Seoul, Korea, 2024.

[16] P. Govalkar, A. Mustafa, N. Pia, J. Bauer, M. Yurt, Y. Özer, and C. Dittmar, "A lightweight neural TTS system for high-quality German speech synthesis," in *ITG Conf. on Speech Communication*, 2021, pp. 39–43.

[17] D. V. Sang and L. X. Thu, "FastTacotron: A fast, robust and controllable method for speech synthesis," in *Proc. of the Int. Conf. on Multimedia Analysis and Pattern Recognition (MAPR)*, Hanoi, Vietnam, 2021.

[18] F. Zalkow, P. Sani, M. Fast, J. Bauer, M. Joshaghani, K. Kayyar, E. A. P. Habets, and C. Dittmar, "The AudioLabs system for the blizzard challenge 2023," in *Blizzard Challenge Workshop*, 2023, pp. 63–68.

[19] J. Bauer, F. Zalkow, M. Müller, and C. Dittmar, "Evaluating the impact of prosody feature normalization on the controllability of pitch in speech synthesis," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, Regensburg, Germany, 2024, pp. 9–16.

[20] F. Zalkow, P. Govalkar, M. Müller, E. A. P. Habets, and C. Dittmar, "Evaluating speech–phoneme alignment and its impact on neural text-to-speech synthesis," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi multi-speaker English TTS dataset," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*, Brno, Czech Republic, 2021.

[22] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H. Hain, X. S. Wang, and M. Garcia, "TC-STAR: Specifications of language resources and evaluation for speech synthesis," in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006, pp. 311–314.

[23] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 161–165.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the Int. Conf. on Machine Learning (ICML)*, New York, NY, USA, 2006, pp. 369–376.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[26] T. tom Dieck, P. A. Pérez-Toro, T. Arias, E. Nöth, and P. Klumpp, "Wav2Vec behind the scenes: How end2end models learn phonetics," in *Proc. of the Annual Conf. of the Int. Speech Communication Association (Interspeech)*. Incheon, Korea: ISCA, 2022, pp. 5130–5134.

[27] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS.1534 (MUSHRA)," in *Proc. of the Web Audio Conf.*, Paris, France, 2015.