Taco-VC: A Single Speaker Tacotron based Voice Conversion with Limited Data

Roee Levy-Leshem
School of Electrical Engineering
Tel Aviv University
Tel Aviv, Israel
roeelev1@mail.tau.ac.il

Raja Giryes
School of Electrical Engineering
Tel Aviv University
Tel Aviv, Israel
raja@tauex.tau.ac.il

***Abstract*—This paper introduces *Taco-VC*, a novel architecture for voice conversion based on *Tacotron* synthesizer, which is a sequence-to-sequence with attention model. The training of multi-speaker voice conversion systems requires a large number of resources, both in training and corpus size. *Taco-VC* is implemented using a single speaker *Tacotron* synthesizer based on Phonetic PosteriorGrams (*PPGs*) and a single speaker *WaveNet* vocoder conditioned on mel spectrograms. To enhance the converted speech quality, and to overcome over-smoothing, the outputs of *Tacotron* are passed through a novel speech-enhancement network, which is composed of a combination of the phoneme recognition and *Tacotron* networks. Our system is trained just with a single speaker corpus and adapts to new speakers using only a few minutes of training data. Using mid-size public datasets, our method outperforms the baseline in the *VCC* 2018 SPOKE non-parallel voice conversion task and achieves competitive results compared to multi-speaker networks trained on large private datasets.**

***Index Terms*—Voice Conversion, Speech Recognition, Speech Synthesis, Adaptation**

I. Introduction

The purpose of voice conversion (*VC*) is to convert the speech of a source speaker into a given desired target speaker. A successful conversion preserves the linguistic and phonetic characteristics of the source utterance while keeping naturalness and similarity to the target speaker. *VC* can be applied to various applications, such as personalized generated voice in text-to-speech (*TTS*) [1], speaking aid for people with vocal impairments [2] and speaker verification spoofing [3].

A wide range of approaches exists for the *VC* task. Some use a statistical parametric model such as Gaussian mixture models (*GMM*) to capture the acoustic features of the source speaker and create a conversion function that maps to the target speaker [4], [5]. Recently, several deep learning based solutions have been provided and successfully led to a better spectral conversion compared to the traditional GMM-based methods. Various network architectures are employed such as feed-forward deep neural networks [6], [7] recurrent neural networks (*RNN*) [8], [9], generative adversarial networks (*GAN*) [10], [11], and variational autoencoder (*VAE*) [12], [13].

The converted speech of a *VC* system is measured by three main quality parameters: (1) Prosody preservation of the source speech, (2) naturalness, and (3) target similarity. Recent research demonstrates successful prosody preservation when using *VC* based phonetic PosteriorGrams (*PPG*s) [14]. *PPG*s represent the posterior probability of each phonetic class per single frame of speech. The *PPG*s are obtained from speaker-independent automatic speech recognition (*SI-ASR*) network, therefore considered as *SI* features [15]. The quality of the converted speech is profoundly affected by the vocoder used in the speech synthesis system. Recently, *WaveNet* vocoder [16] became highly popular and is broadly used in *VC*, providing high quality converted waveforms [17], [18].

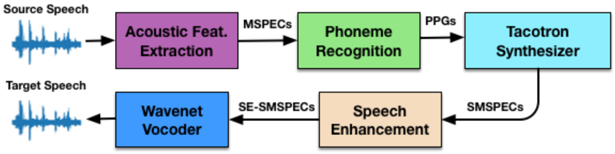

Fig. 1. *Taco-VC* Conversion Process.

TTS research has gained significant progress over the last years, mainly due to the adaptation of sequence-to-sequence (*Seq2Seq*) models such as *Tacotron* [19], [20]. *Seq2Seq* methods are also used for *VC*, among them, the multi-speaker *SCENET* model [21] contains an encoder-decoder with attention, which predicts target *MSPEC*s from source *MSPEC*s and bottleneck features. The *Parrotron* [22] and the work from [23] also describe the usage of *Tacotron* for *VC* purposes, however, they do not use prosody preserved features and require text or phonemes during training.

In this work, we propose *Taco-VC*, a four stages architecture for high quality, non-parallel, many-to-one *VC*. Its main advantage is that it requires a corpus of only a single speaker for training, and can easily be adapted to other speakers with limited training data. Inspired by the recent success of *TTS* models, we base our *VC* system on the architecture of *Tacotron* [19], which provides high quality and natural speech using a *Seq2Seq* synthesizer with attention mechanism [24], and *WavenNet* vocoder. As can be seen in Fig. 1, Phonetic PosteriorGrams (*PPG*) are extracted from a phoneme recognition (*PR*) model to preserve the prosody of the source speech. Using a single speaker *Tacotron* synthesizer, we synthesize the target mel-spectrograms (*MSPEC*) directly from the *PPG*s. The synthesized *MSPEC*s (*SMSPEC*) pass through a speech enhancement network (*Taco-SE*), which outputs the speech enhanced *SMSPEC*s (*SE-SMSPEC*). Finally, a single speaker *WaveNet* vocoder generates the predicted audio from the *SE-SMPSEC*s. We use the same acoustic features (80-band *MSPEC*s) in our different networks as it leads to a high-quality conversion in terms of similarity to the target speaker [25]. It also allows to train the different networks independently and combine them to generate the final target audio.

The main contributions of this work are: (1) a scheme that relies on a single-speaker *Tacotron* and *WaveNet*, and adapts successfully to other target speakers with limited training data; (2) a novel approach for speech enhancement, which handles over-smoothing and noise using a joint training of the *PR* and *Tacotron* synthesizer without over-parameterization of the model due to weight sharing; (3) a *VC* architecture that uses only public and mid-size data, and outperforms the existing baselines. It also shows competitive results compared to other multi-speaker *VC* networks trained on private and much larger datasets. To the best of our knowledge, *Taco-VC* is the first *VC* system that presents a successful adaptation of single speaker networks to other speakers with limited data.

The paper is organized as follows: Section II describes *Taco-VC* model and its adaptation process to new speakers. Section III reports the experiments and results, showing the advantages of the proposed approach. Section IV concludes the paper.

II. The Voice Conversion Network

Fig. 2 presents the four components of our *VC* system. We provide next details on each of them.

A. *The Phoneme Recognition Network*

We use Phonetic PosteriorGrams as our prosodic preserving features. The *PPG*s are extracted using a *PR* network. This network architecture choice is made with two main goals: (1) Provide the ability to extract *PPG*s at the frame level; (2) Allow joint training with the speech synthesis network. We use a convolutional neural network (*CNN*) based *PR*, which is easy to train as it suffers less from vanishing gradients issues [26] during training compared to *RNN* and it can be integrated with *Tacotron* synthesizer as part of the encoder. The *PPG*s are taken from the last fully connected layer before the *CTC* loss [27], which is employed in our network training.

Fig. 2(1) shows the *Seq2Seq* training process of the *PR* network with the *MSPEC*s as the inputs and the phoneme labels as the targets. This network has the same structure of [28] except of the following changes: (1) We use the *Leaky-ReLU* non-linearity [29] instead of *Maxout* to reduce the number of parameters; (2) We add batch-normalization [30] after each non-linear activation in the convolution layers to increase network stability; (3) For compatibility with *Tacotron* and *WaveNet* networks, the raw audio input of the *PR* network is transformed into *MSPEC*s instead of mel-cepstral coefficients.

The performance of our *PR* network is measured by phoneme error rate (*PER*). It achieves 17.5% *PER* on the core test set, which improves over the 18.2% of the network in [28].

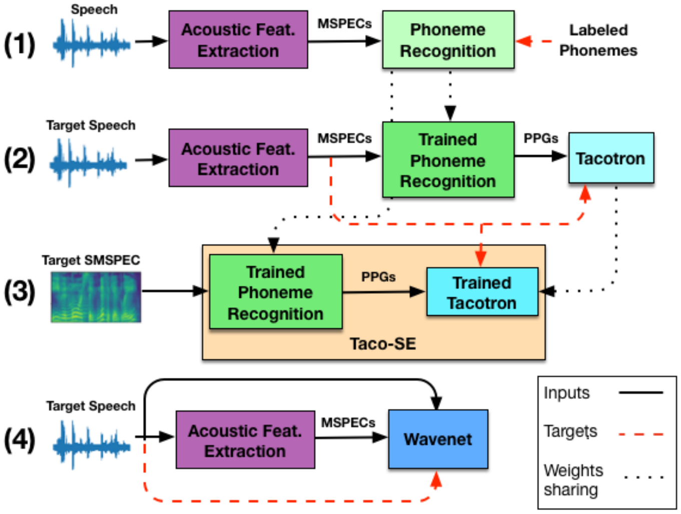

Fig. 2. The training of our model consists of four steps: (1) Phoneme recognition training, (2) *Tacotron* Synthesizer training, (3) Speech enhancement (*Taco-SE*) training, (4) *WaveNet* training.

B. *The Speech Synthesis Network (*Tacotron*)*

Inspired by the success of *Tacotron* in the fields of *TTS*, we propose a single speaker *Tacotron Seq2Seq* model with attention mechanism to predict *MSPEC*s directly from the *PPG*s extracted by the *PR* network for the entire target speech corpus. While *TTS* systems are trained with pairs of $< Text, Audio >$, for *VC* purposes, *Tacotron* is trained with $< PPG, Audio >$ pairs. Fig. 2(2) shows the *Seq2Seq* training of *Tacotron*. The *PPG*s are the single input of the network, while the *MSPEC*s and linear-spectrograms are used as the target.

Our synthesis network has the same structure and loss function as the original *Tacotron* [19] except of the following changes: (1) The Pre-net of the encoder *CBHG* is fed directly with *PPG*s instead of text; (2) While the original *Tacotron* uses teacher forcing mode in the training process, we use linearly decayed scheduled sampling [31] with a final sampling rate of 0.33 for true samples, which helps to increase the quality of the generated *MSPEC*s, especially when adapting the single speaker model to a limited-size train set; (3) As the source utterance length is known, it can be used as the ”stop-token” of the decoder, using the fact that the target utterance has the same length as the source utterance. We have found that constant stop-token helps to get more stable outputs in the generation process.

C. *The Speech Enhancement Network (*Taco-SE*)*

The *PR* network and *Tacotron* are trained separately on different corpora. We have found that the synthesized *MSPEC*s tends to be over-smoothed in the mid-high harmonics. Moreover, the over-smoothing artefacts get worse when adapting *Tacotron*, which is trained on a single speaker speech corpus, to a different speaker with a limited train set.

To address these artefacts, we add another network, *Taco-SE*, which is a concatenated network comprising of the trained

PR ($P(\bullet)$) connected to the trained *Tacotron* ($T(\bullet)$), without over-parameterization of the model due to weights sharing (see Fig. 2(3)). After initialization, *Taco-SE* is trained using only the *Tacotron* loss L_T. As the purpose of *Taco-SE* is to enhance the quality of the *SMSPEC*s, we generate for the entire corpus, using the first two networks, the *SMSPEC* of each utterance. To train the network to increase the quality, we require it to generate the true *MSPEC*, denoted as y, from the *SMSPEC*s, denoted as $\hat{y}$. We also require it to provide this output if y is given as we want the *Taco-SE* to preserve high-quality inputs.

To summarize, *Taco-SE* is trained on the pairs $< y, y >$ and $< \hat{y}, y >$, each with probability 0.5. The first corresponds to retaining the quality by recovering the true target signal given as an input, and the second aims at estimating the target speech signal from a synthesized one with the goal of improving the quality of the network. This leads to the following loss:

$$L_{Taco-SE} = L_T(T(P(y)), y) + L_T(T(P(\hat{y})), y). \qquad (1)$$

As can be seen in Fig. 3, the primary enhancement of *Taco-SE* is being reflected in the mid-higher harmonics (marked by red circles), while in the lower harmonics, there are merely no changes. The *SE-SMSPEC* contains much better-resolved harmonics compare to the *SMSPEC*.

*D. The Vocoder Network (*WaveNet*)*

The conditional *WaveNet* vocoder aims at reconstructing the target raw waveforms from *MSPEC*s. For conditioning the *MSPEC*s, we add local conditioning to the gated units. Since the *MSPEC* is sampled with a lower sampling frequency compares to the raw waveform, we add learnable up-sampling convolutional layers that map it to a new time series with the same resolution of the raw waveform.

We use the implementation and parameters of *WaveNet* from [32]. As Fig. 2(4) shows, for the *WaveNet* training, we use the same single speaker speech corpus used for both *Tacotron* and *Taco-SE*. Also, for local conditioning of *WaveNet*, we use the same *MSPEC*s features that are used for the rest of the networks.

E. System Adaptation

Another aspect of speech synthesis systems in general and *VC* systems, in particular, is the ability to adapt to new speakers given limited training data. *TTS* models are usually trained on large datasets with multiple-speaker support. There are two main strategies for adapting to other target speakers: (1) Using a speaker embedding in multi-speaker systems [33]; and (2) model adjustment by fine-tuning of a multi-speaker *SI* model to a target speaker, which leads to better results in terms of target similarity [34]. Such multi-speaker networks require longer training phases, complex networks with a large number of parameters, and much larger training sets. The usage of model adjustment was also explored for *VC* systems, such as [14], [18] that train a multi-speaker *SI WaveNet* vocoder and adapt it to new target speakers.

We use the model adjustment by the fine-tuning technique for the adaptation process, as done in [18], and explore how a single speaker system will adapt to other speakers with limited data. The trained *Tacotron* is fine-tuned on the new target's training data with linearly decayed scheduled sampling. *Taco-SE* is fine-tuned in the same way as *Tacotron* and uses *SMPSEC*s that are generated for every utterance in the new target training set by the fine-tuned *Tacotron*. *WaveNet* is also fine-tuned on the new target training set. Since the *PR* network is speaker independent, it does not require an adaptation.

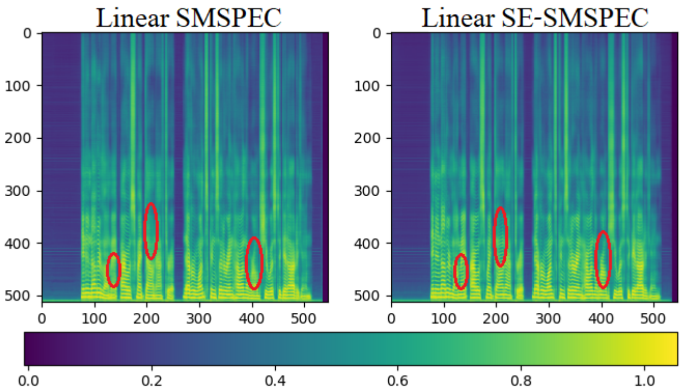

Fig. 3. Linear *SE-SMSPEC* and *SMSPEC* Comparison.

III. Experiments

A. Experimental Setups

The *PR* model is trained using the *TIMIT* corpus [35]. All the 462 speakers training set is used except the SA recordings. The sampling rate of the *TIMIT* is 16 kHz with a 16-bit resolution. For having alignment with the rest of the networks, we up-sampled it to 22050 Hz. *Tacotron*, *Taco-SE*, and *WaveNet* are trained using the public LJ Speech corpus [36], which consists of 13,100 utterances from a single female speaker. The total length of the corpus is approximately 24 hours. All of the utterances are recorded with a sampling rate of 22050 Hz and a 16-bit resolution. The acoustic features used for all of the systems are 80-band *MSPEC*s extracted using Hann windowing of 1024-samples Short Time Fourier Transform, and 256-samples step size. The mel filter-bank base is computed in the range of 125 to 7600 Hz.

We evaluate our system on the VCC2018 SPOKE task [37], which is a non-parallel *VC* task. It has an English speech dataset, containing two males (VCC2TM1, VCC2TM2) and two females (VCC2TF1, VCC2TF2) target speakers and two males (VCC2SM3, VCC2SM4) and two females (VCC2SF3, VCC2SF4) source speakers. Each speaker has the same 81 content utterances for training, and 35 utterances for testing. The whole training set is approximately 5 minutes of speech per target speaker. All of the utterances are recorded with a sampling rate of 22050 Hz and a 16-bit resolution.

Tacotron and *Taco-SE* are trained with a batch size of 5 and optimized using Adam optimizer with a linearly decayed learning rate with an initial value of 0.002 for *Tacotron* and 0.0005 for *Taco-SE*. We use reduction factor $r = 3$ as it leads to the best attention alignment. To adapt the different networks,

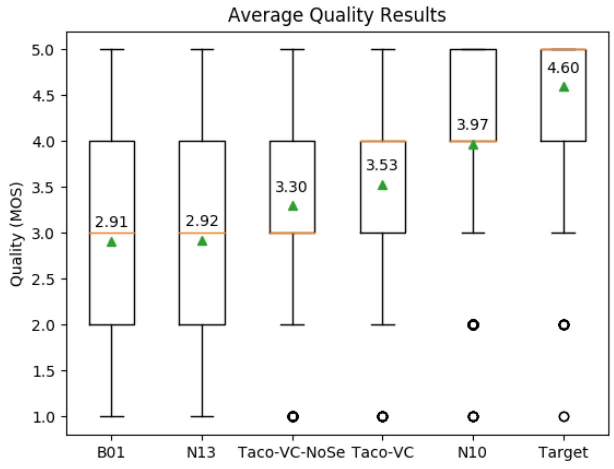

Fig. 4. Total Average Quality (Naturalness) MOS of the five evaluated networks and target speech. The triangle value is the mean. The bold line is the median.

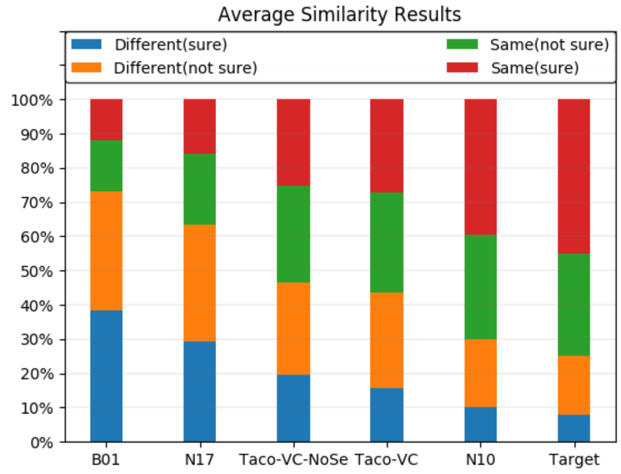

Fig. 5. MOS of target similarity of the evaluated five networks and target speech.

we fine-tuned the trained *Tacotron* and *Taco-SE* for each of the target speakers for another 10,000 steps, using linearly decayed scheduled sampling as for the initial training. *WaveNet* is fine-tuned with another 20,000 steps.

For subjective evaluation, we use the mean opinion score (*MOS*) of naturalness and target similarity. Both evaluations are conducted using the Amazon Mechanical Turk framework. We compare our test utterances to the published, submitted test utterances of the VCC2018. We also do an ablation study by removing the *Taco-SE* network. The tested models are[1]:

- B01 - The baseline system of VCC2018 is a vocoder-free system based on a *GMM* conversion model [38].
- N10 - The best system in both the similarity and naturalness scores of VCC2018 [18]. It uses a *DBLSTM* conversion model that converts *STRAIGHT* extracted spectral features and F_0. The vocoder is a speaker-dependent multi-speaker *WaveNet*. The networks are trained using iFlytek large private datasets. As we do not have access to this large corpus, this work has an inherit advantage.
- N17 - The second-best system in the similarity score of VCC2018 [39]. The conversion model is *DNN* based encoder-decoder trained on a parallel training set generated by *TTS* from the non-parallel corpus. The vocoder is a speaker-dependent multi-speaker *WaveNet*.
- N13 - The second-best system in the naturalness score of the VCC2018.
- *Taco-VC* - Our proposed method, including *Taco-SE* network.
- *Taco-VC-NoSe* - Our proposed method without *Taco-SE* network.

B. Naturalness Evaluation

In the naturalness evaluation, human subjects rate the quality of the different converted utterances. In each assignment, subjects rate six different utterances with the same content speech - N10, N13, B01, *Taco-VC*, *Taco-VC-NoSe*, and the original target. The quality rate is on a scale of 1 (Bad - Completely unnatural speech) to 5 (Excellent - Completely natural speech). The number of evaluation utterances is ten conversions per source with a total of 40 per target, and a total of 160 utterances per system. Every utterance gets ten votes. The utterances are presented in random order. Total of 128 different evaluators participated in the experiment with an average of 74 utterances ranks.

Fig. 4 shows the average *MOS* for naturalness averaged on all pairs. The results indicate a significant effect of the *Taco-SE* on the quality scores. The quality *MOS* results indicate that in terms of subjective quality evaluation, *Taco-VC* outperforms the baseline and gets the same median as N10, though using only a single speaker baseline. The quality gap between *Taco-VC* and N10 can also be explained by the relatively high *PER* of the *PR* network.

C. Target Similarity Evaluation

In the target similarity evaluation, subjects rate the similarity of the different converted utterances to the target speaker utterances. The reference target utterance is chosen by random selection from the training set. In each assignment, subjects rate six different test utterances with the same content speech - N10, N17, B01, *Taco-VC*, *Taco-VC-NoSe*, and the original target. The similarity rate is on a scale of 1 (Different - absolutely sure), to 4 (Same - absolutely sure). We use the same utterances as in the naturalness evaluation. Total of 165 different evaluators participated in the experiment with an average of 57 utterances ranks.

Figure 5 shows the *MOS* distribution for target similarity averaged on all pairs. For *Taco-VC*, almost 60% are ranked as similar to the target, while the baseline (B01) has less than 30%. The real target utterances get the rank of 75%. Note that the impact of *Taco-SE* on the similarity score is minor compared to the naturalness case.

[1] Audio samples - https://roee058.github.io/Taco-VC/

IV. Conclusion

This work presents *Taco-VC*, a *VC* system comprised of *PR* network, *Tacotron* synthesizer, and *WaveNet* vocoder. It has the advantage that it can produce a high-quality speech conversion by just being trained on a single speaker large corpus and then be adapted to new speakers only using a small amount of data. We also introduce the speech enhancement network *Taco-SE*, which might be of interest by itself, and describe how to enhance the synthesized mel-spectrograms only using the trained networks. We show in the *MOS* experiments that our architecture, using public, single speaker training set, can adapt to other targets with limited training sets, and provide competitive results compared to multi-speaker *VC* systems trained on private and much larger datasets. We believe that the high error rate of the *PR* network has a significant impact on the converted speech. As future work, we suggest adding more acoustic features to the generated *PPG*s, or extract *PPG*s from other speech recognition networks with lower error rates. Another possible future research direction is applying the *Taco-SE* architecture (with a corresponding *WaveNet* for denoising [40]) to speech denoising tasks.

References

[1] J. Latorre, V. Wan, and K. Yanagisawa, "Voice expression conversion with factorised HMM-TTS models," in *Proc. Interspeech*, pp. 1514–1518, 2014.

[2] D. Erro, I. Hernáez, A. Alonso, D. García-Lorenzo, and E. Navas, "Personalized synthetic voices for speaking impaired: Website and app," in *Proc. Interspeech*, pp. 1251–1254, 2015.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, and C. Hanilçi, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, pp. 2037–2041, 2015.

[4] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[5] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] L.-h. Chen, Z.-h. Ling, L.-j. Liu, and L.-r. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE International Conference on Audio, Speech, and Lang. Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[7] Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using input-to-output highway networks," *IEICE Trans. on Information and Systems*, vol. 100, no. 8, pp. 1925–1928, 2017.

[8] M. V. Ramos, A. W. Black, R. F. Astudillo, I. Trancoso, and N. Fonseca, "Segment level voice conversion with recurrent neural networks," in *Proc. Interspeech*, pp. 3414–3418, 2017.

[9] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, no. 1, pp. 4869–4873, 2015.

[10] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. on Audio Speech and Lang. Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[11] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, pp. 2100–2104, 2018.

[12] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and D-Vectors," *Proc. ICASSP*, pp. 5274–5278, 2018.

[13] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. on Audio Speech and Lang. Processing*, vol. 27, no. 9, pp. 1432–1443, 2019.

[14] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *Proc. Interspeech*, pp. 1978–1982, 2018.

[15] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *Proc. ICME*, pp. 1–6, 2016.

[16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, and O. Vinyals, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[17] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, pp. 1138–1142, 2017.

[18] L.-J. Liu, Z.-H. Ling, and L.-R. Dai, "WaveNet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, pp. 1983–1987, 2018.

[19] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, and R. J. Weiss, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, pp. 4006–4010, 2017.

[20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, and N. Jaitly, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, pp. 4779–4783, 2018.

[21] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. on Audio, Speech, and Lang. Processing*, pp. 631–644, 2019.

[22] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," *Proc. Interspeech*, pp. 4115–4119, 2019.

[23] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source Tacotron and WaveNet," *Proc. Interspeech*, pp. 1298–1302, 2019.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, pp. 1–15, 2014.

[25] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based WaveNet vocoder," in *Proc. Interspeech*, pp. 1993–1997, 2018.

[26] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[27] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," *JMLR Workshop and Conference Proceedings*, vol. 32, no. 1, pp. 1764–1772, 2014.

[28] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, and C. L. Y. Bengio, "Towards end-to-end speech recognition with deep convolutional neural networks," in *Proc. Interspeech*, 2016.

[29] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc ICML*, vol. 30, 2013.

[30] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[31] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc NIPS*, pp. 1171–1179, 2015.

[32] R. Yamamoto, M. Andrews, M. Petrochuk, W. Hycbrom, O. Vishnepolski, M. Cooper, and K. Chen, "r9y9/wavenet vocoder: v0.1.1 release."

[33] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," *Proc. ICML*, pp. 3683–3691, 2018.

[34] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Proc. NIPS*, pp. 10040–10050, 2018.

[35] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[36] Keith Ito, "The LJ speech dataset," 2017.

[37] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, and F. Villavicencio, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *Proc. Odyssey*, pp. 195–202, 2018.

[38] K. Kobayashi and T. Toda, "sprocket : Open-source voice conversion software," in *Proc. Odyssey*, pp. 203–210, 2018.

[39] Y.-c. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The Nu non-parallel voice conversion system for the voice conversion challenge 2018," in *Proc. Odyssey*, no. June, pp. 211–218, 2018.

[40] D. Rethage, J. Pons, and X. Serra, "A WaveNet for speech denoising," in *Proc. ICASSP*, pp. 5069–5073, 2018.