HiFi-GAN based Text-to-Speech Synthesis in Serbian

Siniša Suzić
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
sinisa.suzic@uns.ac.rs

Darko Pekar
Alfanum Ltd.
Novi Sad, Serbia
darko.pekar@uns.ac.rs

Milan Sečujski
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
secujski@uns.ac.rs

Tijana Nosek
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
tijana.nosek@uns.ac.rs

Vlado Delić
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
vlado.delic@uns.ac.rs

***Abstract*—In this paper we present a deep neural network based text-to-speech system in the Serbian language, which converts generated acoustic features into a speech signal using the HiFi-GAN vocoder. The HiFi-GAN model was fine-tuned using an existing multi-speaker model trained on an English speech corpus. To overcome the problem of inadequate training data, we introduce a data generation technique based on a guided acoustic neural network, which attempts to minimize the mismatch between data used in HiFi-GAN training and inference. The outputs of the acoustic network are intended to represent a trade-off between original feature trajectories and trajectories generated by the standard text-to-speech system. The results of subjective evaluation through listening tests show that the proposed system produces speech whose quality significantly surpasses the quality of speech generated by the best existing speech synthesis for Serbian, and that its MOS score is very close to the score given to natural speech.**

Keywords—generative adversarial networks, text-to-speech synthesis, vocoder.

I. INTRODUCTION

First widely used text-to-speech (TTS) systems were based on concatenation of speech segments from prerecorded speech corpora [1]. Although such systems could produce speech of good intelligibility, this approach also suffered from drawbacks such as the presence of audible artifacts at concatenation points, the need for large speech corpora including an extremely great number of different phonetic contexts, as well as low flexibility as regards changing speech characteristics. All these drawbacks led to the development of new approaches based on learning from speech data rather than simply reproducing it.

A majority of modern text-to-speech (TTS) systems consists of three basic blocks: text analysis module, acoustic modelling module and a vocoder, connected sequentially [2]. The text analysis module converts raw text into a sequence of linguistic feature vectors. These features can be very simple and include only phoneme identities, but they can also be much more complex and include information related to part-of-speech (POS) tags or intonation. The module for acoustic modelling learns to generate acoustic features from linguistic features by training on speech data, while the vocoder converts obtained acoustic features to speech.

Initially, the use of statistical modelling was restricted to creating models for converting linguistic features into acoustic ones. The most widely used statistical modelling approach was based on using Hidden Markov Models (HMM) and is usually referred to as HMM synthesis [3, 4]. Although such systems enable greater flexibility and produce output with no concatenation artifacts, generated speech is muffled and buzzy. More recently, deep neural networks (DNN) took over the role of HMMs in statistical modelling, firstly in the form of feed-forward networks [5] and later as recurrent neural networks [6, 7]. Deep neural networks were able to generate more natural acoustic feature trajectories compared to HMMs, but the obtained features still needed to be converted to speech by a vocoder such as WORLD [8], which yielded suboptimal results due to inferior vocoder performance.

With further development of neural networks and their tremendous increase in popularity, they were soon used as main components of all building blocks of TTS systems, within architectures such as Deep Voice 1 and 2 [9, 10]. WaveNet, an architecture enabling direct prediction of speech samples from linguistic features, thus avoiding the need for vocoder usage, was also presented [11]. Systems such as Tacotron 1 [12] generate speech directly from the sequence of characters. Tacotron 1 does not generate speech samples directly, but generates spectrograms which are then converted to speech by the Griffin-Lim algorithm [13]. On the other hand, Tacotron 2 [14] uses WaveNet conditioned on mel-spectrograms to produce speech samples directly. Finally, the progress in DNN speech synthesis led to the development of fully end-to-end systems, which enable direct conversion from raw text to speech, without an acoustic model or a vocoder [15].

In this paper we introduce HiFi-GAN [16], a neural vocoder, into the framework of an existing DNN-based text-to-speech synthesis system in the Serbian language [17]. We also

This research was supported by the Science Fund of the Republic of Serbia, #6524560, AI-S-ADAPT, and by the Serbian Ministry of Education, Science and Technological Development, #45103-68/2020-14/200156: "Innovative Scientific and Artistic Research from the Faculty of Technical Sciences Activity Domain".

propose a special neural network, which we refer to as guided acoustic network, for generating training data for Serbian.

The rest of the paper is organized as follows. Section 2 presents an overview of the development of neural vocoders in general, including a more detailed presentation of the HiFi-GAN vocoder. The proposed system and procedure for data preparation and training are described in Section 3. The experiments are described in section 4 together with the results of subjective evaluation of the quality of synthesized speech. Finally, Section 5 concludes the paper and outlines the directions of future research.

II. Neural Vocoders

WaveNet [11], the first neural vocoder, represents an autoregressive convolutional neural network which predicts speech samples from linguistic features. It has later been adapted in order to use spectrograms instead of linguistic features as inputs [9, 10]. The main drawback of vocoders obtained in such a way is inference time, since they are capable of producing only a single speech sample in one forward pass. Architectures attempting to solve this problem are presented in [18, 19]. Another group of neural vocoders are flow-based vocoders [20, 21]. Models of this type estimate the conditional distribution of a speech signal conditioned on acoustic features by applying an invertible transformation, or flow, to a latent variable. The latent variable is generated by a simple probabilistic distribution such as a Gaussian distribution with zero mean and unit variance.

At the moment, generative adversarial networks (GAN) [22] are probably the most popular generative models, initially suggested for image generation, but successfully applied to other domains including audio and speech generation [23, 24, 25]. Although vocoders of this type are computationally efficient, the quality of produced speech is lower compared to speech obtained by autoregressive and flow-based approaches. Further research eventually resulted in the HiFi-GAN vocoder [16], which is both computationally efficient and capable of producing speech of quality comparable to other types of neural vocoders.

A. HiFi-GAN vocoder

A generative adversarial network generally consists of two modules, a discriminator and a generator. The generator creates data with the same statistics as in the training set, while the discriminator tries to distinguish whether a given sample is real or generated. On the other hand, HiFi-GAN consists of one generator and two discriminators. The generator is a fully convolutional network which uses transposed convolutions and mel-spectrograms as inputs. Multi-period discriminator (MPD) consists of a number of sub-discriminators which use equidistant samples from input speech (each discriminator uses a different sampling period). In such a way, MPD attempts to detect periodic patterns in speech, assuming that it can be decomposed into sinusoidal signals. On the other hand, multi-scale discriminator (MSD) uses consecutive samples from input speech.

With such an architecture, the loss function represents a weighted sum of the following 3 terms:

- GAN loss, in which standard GAN loss from [22] is replaced by least squares error function;
- mel-spectrogram loss, which represents the L_1 distance between mel-spectrograms extracted from original speech and generated speech respectively;
- feature matching loss, which represents the distance between discriminator features obtained for original speech and generated speech.

III. Proposed System

In this paper we introduce the HiFi-GAN vocoder in the pipeline of the Serbian DNN-based TTS system presented in [17]. Firstly, it should be noted that, since HiFi-GAN uses mel-spectrograms as inputs to produce speech samples, its "knowledge" is actually mostly language independent, which implies that mel-spectrograms extracted from Serbian speech could be used as inputs to a model trained for another language to produce output speech in Serbian. There could be some contexts which are specific for certain language, but these should be successfully learned during model tuning. With that in mind, in this research we utilize a universal model trained on a multi-speaker speech corpus in English and fine-tune it to the voice of a Serbian speaker. However, if the HiFi-GAN model is simply tuned using original speech in Serbian, this would produce suboptimal results in TTS inference stage. Namely, features produced by standard DNN TTS are smoothened so there would be a mismatch between data used in HiFi-GAN training and inference. On the other hand, it would also be possible to use a trained standard TTS in Serbian to generate the entire training database, but this would also introduce data inconsistency. Namely, HiFi-GAN model would be fine-tuned using features which are different than those used for training the universal base model. For this reason, we introduce a guided acoustic network for generation of training data. The outputs of this network should represent a trade-off between original feature trajectories and smoothed trajectories generated by the standard TTS system. This network is described in the following subsection.

A. Guided acoustic network

The architecture of this network is based on the architecture of the network used for acoustic data prediction in standard TTS and is shown in Fig. 1. This network predicts acoustic features extracted by the vocoder: mel-generalized cepstral coefficients (MGC), fundamental frequency and aperiodicities. The principal component of the input to the network consists of linguistic features. These are extended by positional features, which provide information related to the position of the current frame in a phone. However, in order to provide additional information to the network, acoustic features corresponding to the middle frame of the current phone are added to the input, which could be considered as additional guide for the network. The target features are mean-var normalized. The linguistic part of the input is normalized using min-max normalization, while mid-phone acoustic features are also mean-var normalized. The information about phone durations is taken from the original corpus.

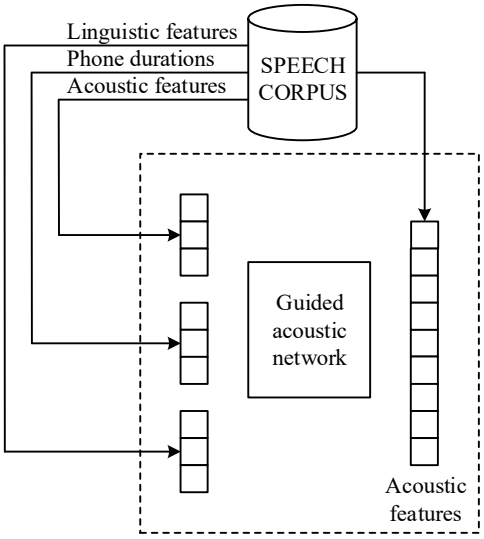

Fig 1. Guided acoustic network training

Once this network is trained, it is used to generate acoustic features for all the training data. In the inference process original phone durations are used as well as original mid-phone features. These features are then fed to the vocoder and generated speech is used in extraction of mel-spectrograms fed to the input to the HiFi-GAN model during its training, as shown in Fig 2. Target speech in HiFi-GAN training is the original speech from the training database.

B. *Data augmentation*

Since the availability of speech data in Serbian is limited, we also resort to data augmentation techniques:

- speech rate acceleration by a factor of up to 1.5% (chosen randomly from a set of predefined values);
- phase perturbation (FFT of original speech is calculated, its phase is changed, and the inverse FFT is then applied to the new values to obtain speech samples).

It should be noted that data obtained in this way is used only in HiFi-GAN training.

IV. Experimental results

The speech samples used in subjective evaluation of the quality of speech synthesis were obtained based on a Serbian speech corpus of a female voice talent. The corpus was recorded in professional studio and contains around 3 hours of speech (including silent segments within utterances). For the purposes of the experiments all files were resampled to 22.05 kHz. By applying data augmentation techniques, the content of the database was extended to approximately 25 hours of speech.

The standard TTS system consists of one neural network for predicting durations represented as numbers of frames in HMM obtained state-level alignments and one network for predicting acoustic features used by the WORLD vocoder. Acoustic features include 30 MGCs, fundamental frequency and two aperiodicity coefficients. We also calculate the first and second order derivatives of the aforesaid features, i.e. delta features, as well as the information about the voicing of individual frames, which adds up to a target feature vector dimension of 100. Final acoustic features are obtained by applying Maximum Likelihood Parameter Generation algorithm [26] to the acoustic features and their derivatives. The inputs to both networks include 743 binary linguistic features, and the inputs to the acoustic network are extended by 9 positional features. Both networks have the same architecture: 3 feed-forward layers with 1,024 neurons which used the ReLU activation function and one LSTM layer with 1024 neurons.

The guided acoustic network has the same architecture as the standard acoustic network and uses the same target features. When adding acoustic features to the inputs derivatives are not used. As a starting HiFi-GAN model we use the universal model provided by the authors of original HiFi-GAN paper, which is trained on the LJSpeech dataset [27]. This corpus contains approximately 44 hours of speech. All the training parameters were the same as in [16] and the model

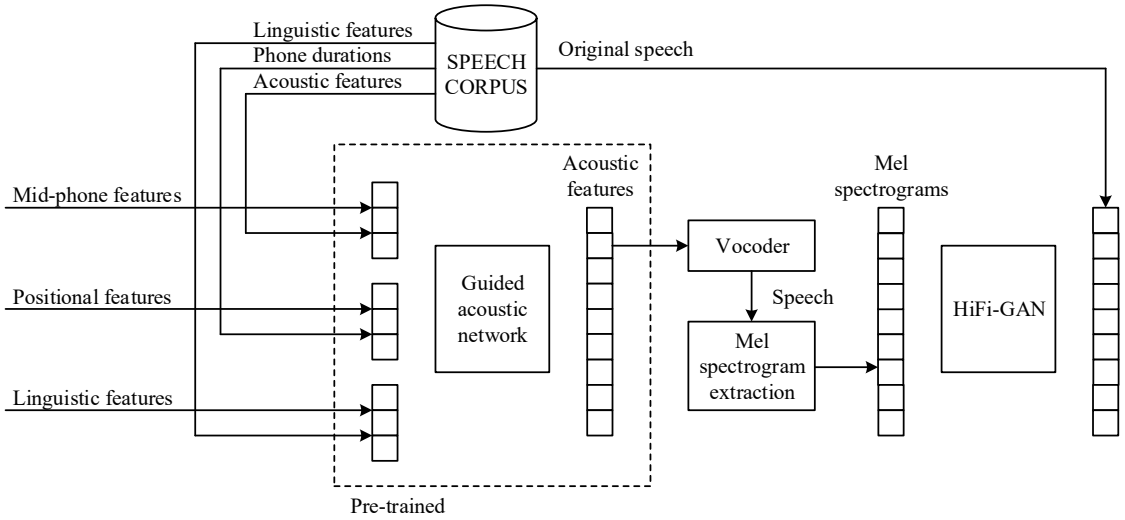

Fig 2. The HiFi-GAN training pipeline

was fine-tuned for additional 60,000 steps. The mel-spectrogram loss for the validation set is shown in Fig. 3.

A. Subjective evaluation

The quality of the generated speech is evaluated through two subjective tests. In both tests the input to the HiFi-GAN model were the mel-spectrograms extracted from the speech generated by the standard TTS. Generated sentences were not used in either standard TTS or HiFi-GAN training.

The first test was the preference test, and it consisted of 15 tasks. In each task two utterances with the same lexical content were given, one generated by the standard TTS and the other generated by the HiFi-GAN based system. The subjects were required to select the utterance with better overall speech quality. The "no-preference" choice was also offered. The test was performed by 20 native Serbian speakers. The results presented in Fig. 4 show a clear preference for the HiFi-GAN based system. Namely, 71.67 % of utterances generated by the HiFi-GAN vocoder were preferred over utterances generated by WORLD. In only 12.66 % of cases WORLD generated files were considered as better and in 15.67 % of cases no preference over either of the generated utterances was stated.

The second test was a Mean Opinion Score (MOS) test. The subjects were presented with 30 utterances in a random order, whereof 10 original utterances from the corpus, as well as 10 corresponding utterances generated by HiFi-GAN and WORLD based systems each. The subjects were required to score each utterance with a mark from 1 to 5, where 5 represented natural speech, and 1 represented speech of very low quality. A total of 20 subjects took part in this test and the results are given in Fig. 5. It can be seen that the HiFi-GAN based system significantly outperforms the WORLD based system and that the MOS scores obtained by HiFi-GAN are very close to the ones obtained for natural speech (4.41 vs 4.53). These results are very similar to MOS values obtained in original HiFi-GAN paper [16] (4.36 vs 4.35).

V. Conclusion and Future Work

In this paper we present a high-quality TTS system for the Serbian language based on the usage of HiFi-GAN vocoder. We have shown that the HiFi-GAN vocoder can be successfully fine-tuned for the Serbian language using universal models trained for English. The quality of speech synthesized using HiFi-GAN has been shown to surpass the quality of speech synthesized using WORLD, achieving MOS grades that are quite close to those given to natural speech in the listening tests.

As to the direction of our future research, by using simple data augmentation techniques described in the paper, we also intend to investigate the possibility of training models fine-tuned to speakers with a very small quantity of available speech data as well as with a reduced number of linguistic features. This is expected to accelerate the process of speech data generation and the creation of new voices.

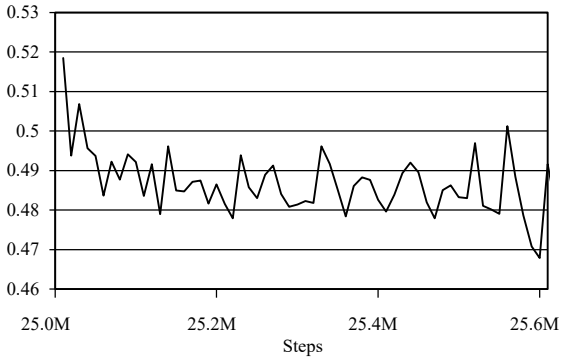

Fig 3. Smoothened mel-spectrogram loss for the validation set

Fig. 4. Results of the preference test

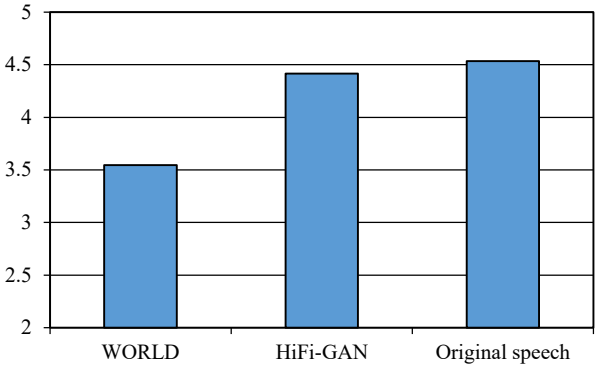

Fig. 5. Results of the MOS test

References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in Proc. ICASSP 1996, vol. 1, pp. 373–376, 1996.

[2] X. Tan, T. Qin, F. Soong, and T. Y. Liu, "A survey on neural speech synthesis," arXiv preprint arXiv:2106.15561, 2021.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in Proc. 6th European Conference on Speech Communication and Technology, 1999.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. ICASSP 2000, vol. 3, pp. 1315–1318, 2000.

[5] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks", in Proc. ICASSP 2013, pp. 7962–7966, 2013.

[6] Y. Fan, Y. Qian, F. L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in Proc. 15th ISCA Conf., 2014.

[7] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis", in Proc. ICASSP 2015, pp. 4470–4474, 2015.

[8] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Inf. Syst., vol. 99, no. 7, pp. 1877–1884, 2016.

[9] S. Ö. Arık et al., “Deep voice: Real-time neural text-to-speech,” in Proc. ICML, pp. 195–204, 2017.

[10] A. Gibiansky et al., “Deep voice 2: Multi-speaker neural text-to-speech”, Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[11] A. van Den Oord et al., “WaveNet: A generative model for raw audio,” in Proc. SSW 2016, vol. 125, no. 2, 2016.

[12] Y. Wang, et al., “Tacotron: Towards end-to-end speech synthesis,” arXiv preprint arXiv:1703.10135, 2017.

[13] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” IEEE Trans. Acoust. Speech Signal Process., vol. 32, no. 2, pp. 236– 243, 1984.

[14] J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in Proc. ICASSP 2018, pp. 4779–4783, 2018.

[15] Y. Ren et al., “Fastspeech 2: Fast and high-quality end-to-end text to speech”, in Proc. Int. Conf. on Learning Representations, 2021.

[16] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 17022–17033, 2020.

[17] T. Delić, M. Sečujski, and S. Suzić, “A review of Serbian parametric speech synthesis based on deep neural networks,” Telfor Journal, vol. 9, no. 1, pp. 32–37, 2017.

[18] S. Mehri et al., “SampleRNN: An unconditional end-to-end neural audio generation model,” arXiv preprint arXiv:1612.07837, 2016.

[19] N. Kalchbrenner et al., “Efficient neural audio synthesis,” in Proc. ICML 2018, pp. 2410-2419, 2018.

[20] A. van de n Oord et al., “Parallel WaveNet: Fast high-fidelity speech synthesis,” in Proc. ICML 2018, pp. 3918–3926, 2018.

[21] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in Proc. ICASSP 2019, pp. 3617–3621, 2019.

[22] I. Goodfellow et al., “Generative adversarial networks,” Adv. Neural Inf. Process. Syst., vol. 3, no.11, 2014.

[23] C.Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” arXiv preprint arXiv:1802.04208, 2018.

[24] K. Kumar et al., “Melgan: Generative adversarial networks for conditional waveform synthesis,” Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[25] M. Bińkowski et al., “High fidelity speech synthesis with adversarial networks”, arXiv prepr int arXiv:1909.11646, 2019.

[26] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, “Speech parameter generation algorithms for HMM-based speech synthesis,” in Proc. ICASSP 2000, vol. 3, pp. 1315 - 1318.

[27] K. Ito and L. Johnson, “The LJ Speech Dataset,” available at URL h ttps://keithito.com/LJ-Speech-Dat aset/, retrieved 05.01.2022.