End-to-End Speech Synthesis for the Serbian Language Based on Tacotron

Tijana Nosek[1(✉)], Siniša Suzić[1], Milan Sečujski[1], Vuk Stanojev[1], Darko Pekar[2], and Vlado Delić[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
tijana.nosek@uns.ac.rs
[2] AlfaNum Ltd., Novi Sad, Serbia

Abstract. End-to-end text-to-speech (TTS) systems allow for the generation of high-quality computer-generated speech without relying on expert-created modules. This paper outlines initial efforts to develop a Serbian end-to-end TTS system using the Tacotron architecture. Listening tests revealed that while Tacotron can produce natural-sounding synthesis when properly trained, it is prone to overfitting and requires extensive data to avoid frequent hallucinations and accent errors. The use of a vocoder proved to be crucial in overall speech quality. Although the level of Tacotron training is less critical, it still demonstrates easy overfitting with relatively small databases. Correct accents and the absence of artifacts and hallucinations are extremely important for listeners, and any issues in these areas result in significantly lower ratings. Despite being less expressive, a controllable standard DNN-based TTS with a standard front end receives better grades because it never hallucinates and rarely makes linguistic mistakes. Integrating expert knowledge from existing pipelines can further improve synthesis quality, especially in data-constrained scenarios.

Keywords: Speech Synthesis · Tacotron · Deep Neural Networks · Front End

1 Introduction

Text-to-speech (TTS), also known as speech synthesis, aims to generate natural, expressive, intelligible speech from text mimicking human speech patterns. TTS has broad applications in human communication and has been a long-standing research topic in natural language and speech processing, as well as in artificial intelligence [1]. Over the decades, TTS systems have evolved from concatenative synthesizers, via statistical parametric speech synthesis, to models based on deep neural networks (DNN) [2]. With the development of deep neural networks, TTS systems have evolved from CNN/RNN-based models to transformer-based models, from auto-regressive models to other generative models, from cascaded acoustic models/vocoders to fully end-to-end models [3].

Developing a human-like TTS system requires both signal processing and linguistic background knowledge. In an attempt to bypass the need for linguistic knowledge, TTS systems have moved to end-to-end models that can be trained from scratch on the

A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 219–229, 2025.
https://doi.org/10.1007/978-3-031-77961-9_16

paired data set of < text, speech >. Some end-to-end TTS models like WaveNet [4] and FastSpeech 2 [5] are developed to directly generate waveforms from text. Others, like Tacotron [6], are trained to simplify linguistic and acoustic features converting them into linear-spectrograms, while others like NaturalSpeech [3], Tacotron 2 [7], DeepVoice 3 [8], FastSpeech [9] and FastSpeech 2 [5], predict mel-spectrograms from characters/phonemes. These models are augmented with a neural vocoder to generate waveforms.

End-to-end models do not require alignment information between text and speech and can be scaled with large amounts of acoustic data with transcripts. It can also be easier to adapt the model to new data. End-to-end models can be more robust than models that have separate components for text analysis front-end, acoustic model and vocoder since each component's errors can propagate [6]. Even though these models can produce state-of-the-art results, they suffer from slow training and inference speed, as well as necessity for large amount of high-quality speech corpus required for training, which proves problematic for low resource languages [10].

To the best of the authors' knowledge the results described in this paper present the first attempt to create an end-to-end TTS system in Serbian. There have been attempts in creating end-to-end systems in other South Slavic languages such as Macedonian [11, 12]. The system described in this paper is based on Tacotron 2 architecture. Since there are no datasets in Serbian large enough to enable training the model from scratch, the English model has been adapted using the Serbian speech dataset. To overcome the problem in generation of Serbian accented vowels, the authors propose the usage of previously developed expert based modules for accent prediction in Serbian [13].

The remainder of this paper is structured as follows: in Sect. 2 we will present key components of the model architecture and challenges that occurred during the training; in Sect. 3 we will present the results of subjective tests that have been performed for system evaluation, and in Sect. 4, we will discuss the results we obtained. We will give concluding remarks in Sect. 5.

2 Models and Approaches

In this section an overview of different models used in experiments will be given, as well as the description of data used for creating TTS voices.

2.1 Tacotron

Original Tacotron-2 architecture [7] consists of 2 modules: a recurrent sequence-to-sequence network with attention, which is used for predicting mel-spectrograms from an input character sequence, and a WaveNet [4] based vocoder, which generates time-domain waveform samples conditioned on the predicted mel-spectrograms. In all of our experiments WaveNet based vocoder is replaced by more efficient and better quality HiFi-GAN vocoder [14].

The mel-spectrogram predicting network consists of encoder and decoder with attention. The encoder consists of character-embedding layer, 3 convolution layers and bidirectional LSTM layer. The encoder output is passed through attention network and its output is further passed to an autoregressive decoder network producing mel-spectrograms as output. In our experiments we used the implementation presented in [15].

Since the training of Tacotron model is data intensive and there is not enough material in Serbian to train the model from scratch, the idea was to use Tacotron model already trained on LJSpeech dataset and adapt it to the Serbian database. The main change was the introduction of set of characters for Serbian language. We used Latin characters, with the exception of digraphs LJ, NJ and DŽ, which are conventionally treated as single letters, and replaced by Q, W and X respectively for convenience.

Although initial experiments showed promising results producing intelligible and good quality speech, we noticed its problems, most notably those related to generating appropriate accents. In order to mitigate these problems we extended the initial set of characters defined for Serbian to cover accents types representative for Serbian. The prediction of accents for the Serbian language was performed by the TTS front-end module, based on high-quality expert system using dictionaries and morpho-syntactic rules [13]. The description of accent used is given in Sect. 2.1.1.

We tried two different approaches for including accent information in system training. In first one a digit was added after each vowel to indicate a certain accent type or the absence of accent (e.g. *točak* would be represented as *to2ča0k*). In the second approach each accented vowel was presented by a different diacritic, (e.g. *točak* would be represented as *tòčak*). More details about accent types in Serbian are given in the following section.

2.2 A Note on Serbian Orthography

The Serbian language exhibits almost ideal phonemic orthography i.e. an orthography in which the graphemes correspond consistently to the phonemes of the language. An ideal correspondence between graphemes and phonemes would imply that each word is pronounced exactly as it is written, and hence that in a text-to-speech system explicit grapheme-to-phoneme conversion methods, based on dictionaries and/or conversion rules, are largely unnecessary, since the spelling of a word unambiguously and transparently indicates its pronunciation. However, neither of the two alphabets used for Serbian (Cyrillic and Latin) distinguishes between short and long vowels or rising and falling tones in Serbian, which is why a written vowel character (e.g. “e”) can stand for any of the 6 possible cases – a non-stressed short vowel (/e/), a stressed vowel with short falling accent: (/ȅ/), short rising accent (/è/), long falling accent (/ê/), a long rising accent (/é/), as well as post-accent long vowel (/ē/). A difference between the accents can imply a difference between word meanings, which is why pitch accents should be considered as relevant to the phonemic inventory. Marking differently accented vowels in the text (with digit suffixes from 0 to 5 or with different diacritics) can be compared to the use of explicit phonetic transcriptions in TTS systems for languages with non-phonemic orthography, and in this research it was carried out in order to help the system establish relationships between words and their pronunciations more easily under conditions of data sparsity.

2.3 Standard TTS with Neural Vocoder

Standard Serbian TTS consists of three blocks: front-end, which performs text normalization and produces a set of linguistic features, a DNN based block, which predicts some acoustic features using linguistic features as inputs, and a vocoder. The initial system based on the usage of deterministic WORLD vocoder is introduced in [16], while the system using neural HiFi-GAN vocoder is presented in [17].

The DNN block for acoustic feature prediction consists of two neural networks [16], one for prediction of phoneme durations and the other which predicts vocoder features based on input linguistic features and outputs of duration prediction network. Both networks consist of 3 feed-forward layers and one LSTM layer. This block was further improved by enabling multi-speaker training and applying target speaker adaptation as presented in [18].

2.4 HiFi-GAN Vocoder

A HiFi-GAN vocoder initially presented in [14] is a neural vocoder based on generative adversarial networks (GAN) [19]. A generative adversarial network typically comprises two main components: a discriminator and a generator. The generator produces data that mimics the statistical properties of the training dataset, while the discriminator's role is to determine whether a given sample is real or synthetic. HiFi-GAN, however, includes one generator and two types of discriminators. The generator in HiFi-GAN is a fully convolutional network that utilizes transposed convolutions and takes mel-spectrograms as input. The multi-period discriminator (MPD) consists of several sub-discriminators, each processing equidistant samples from the input speech, i.e. operating on a different sampling interval. This design allows the MPD to identify periodic patterns in the speech, working under the assumption that speech can be decomposed into sinusoidal components. Meanwhile, the multi-scale discriminator (MSD) analyzes consecutive samples from the input speech.

The process of adapting HiFi-GAN vocoder to standard Serbian TTS is described in [17]. The model is adapted from universal HiFi-GAN model trained on English data. This model was not trained directly on spectrograms extracted from natural speech but on data produced by specific guided acoustic network. In this way the model is better adapted to the outputs of a standard Serbian TTS system.

For the purposes of Tacotron based system the corresponding vocoder was also trained (finetuned). This vocoder was trained on mel-spectrograms produced by Tacotron by using text from original training dataset as Tacotron input. The target samples represent natural speech.

2.5 Training

All systems presented in the following subsections were trained using a Serbian speech corpus of a single female voice talent. This corpus was recorded in a professional studio and contains around 1.5 h of speech (including silent segments within utterances).

For the purposes of Tacotron training we used the same parameters as presented in the implementation given in [15], while the HiFi-GAN vocoders were trained using same hyper-parameter values given by the authors of original paper [14].

Both Tacotron and HiFI-GAN models were adapted by using starting models which were trained on LJSpeech dataset [20], which contains approximately 24 h of speech in English.

3 Experiments

For the evaluation and comparison of the selected models, several listening tests were performed, which will be presented in detail in following subsections. In each test 20 native Serbian speakers were included. Participants were instructed to use headphones to clearly hear even subtle differences in synthesized speech. None of the sentences used in tests were seen during the training of the models.

3.1 MUSHRA Test

The Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test consists of 10 sets of utterances. Each set contains five utterances for grading and a clearly marked reference utterance. All utterances have the same linguistic content. The reference utterance contains natural speech of the target speaker. Among the five utterances for grading, one is identical to the reference utterance (hidden reference), while the other four are synthesized using different synthesizers. One of these four synthesized utterances is generated using the standard TTS model described in Sect. 2.2 (referred to as *st_TTS*), while the other three are synthesized by models based on Tacotron, described in Sect. 2.1. The first one generated by the model trained with a database not containing annotated accents (referred to as *TAC_noAcc*). The second one is the output of the model trained with accent information carried by a digit suffix, detailed in Sect. 2.1 (referred to as *TAC_Acc*), and the third one is the output of the model trained with accent information introduced through different diacritics (i.e. different characters) for each accented vowel, detailed in Sect. 2.1 (referred to as *TAC_Acc1*).

Listeners were asked to grade each of the five utterances by moving a slider on a scale from 0 to 100, allowing for very fine gradation of the quality of synthesized speech. The reference utterance served as an example of how natural speech should sound, and the same utterance was included among the five utterances for grading to verify if the listeners could identify and correctly rate it with a score of 100 or close enough.

The results (Fig. 1) showed that the reference utterance was graded almost 100, with an average score of 94.5. The lowest grade was given to *TAC_noAcc* (41.3), followed by *TAC_Acc* (50.9). The *st_TTS* and TAC_Acc1 received much better grades, with average scores of 64.8 and 69.6, respectively.

3.2 MOS Test

The Mean Opinion Score (MOS) test consists of 18 utterances with different linguistic content. One third of the utterances are synthesized with *st_TTS*, another third with *TAC_noAcc*, and the rest with *TAC_Acc1*. Among the six utterances produced by *TAC_noAcc*, half of them contain at least one incorrectly accented word, while in the rest all words are correctly accented. Among the six utterances produced by *TAC_Acc1*,

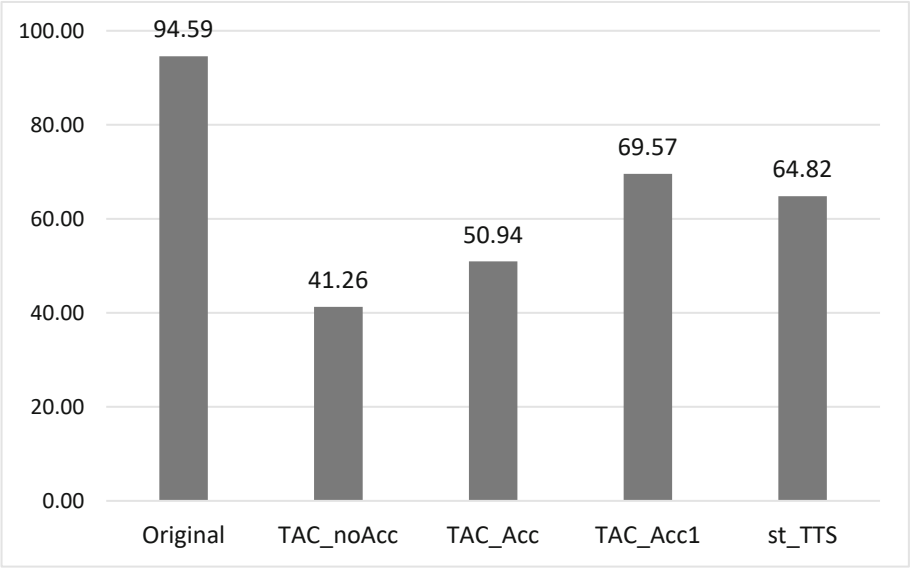

Fig. 1. Results of MUSHRA test – grades of 1–100 scale for quality of different synthesizers.

half of them contain hallucinations at the end of the utterance, while the rest are free of hallucinations (randomly synthesized non-existent phonemes).

Listeners were asked to grade each utterance on a 1–5 scale in terms of speech quality, i.e., its naturalness and intelligibility. A grade of 1 indicates unnatural and/or unintelligible speech.

The *st_TTS* received the highest average grade, 4.7, followed by *TAC_Acc1* with 4.1, and *TAC_noAcc* received the lowest grade, 3.0 (Fig. 2). However, when graded separately, the utterances produced by *TAC_Acc1* without hallucinations had an average grade of 4.6, almost as high as *st_TTS*, while those with hallucinations were graded 3.5 on average. Similarly, the utterances produced by *TAC_noAcc* with correctly accented words had an average grade of 3.6, while those with incorrect accents were graded 2.5 on average. The presence of hallucinations and incorrect accents in synthesized speech significantly lowered the perceived quality, resulting in grades lower by over 1 point.

3.3 Preference Test

In the preference test, there were 14 pairs of utterances. Each pair contained two utterances with the same linguistic content but produced by different synthesizers. All utterances are produced by models based on Tacotron. Eight pairs of utterances were used to analyze the impact of training the Tacotron model for different numbers of epochs, while the remaining pairs focused on the importance of adapting the HFG-based vocoder to the target speaker. In the first eight pairs, one utterance was produced by a less-trained model, while the other was produced by a more-trained model, with both utterances in each pair produced by the same vocoder. Two out of the eight pairs were produced by models trained with accent information, with one model trained for 250 epochs and the

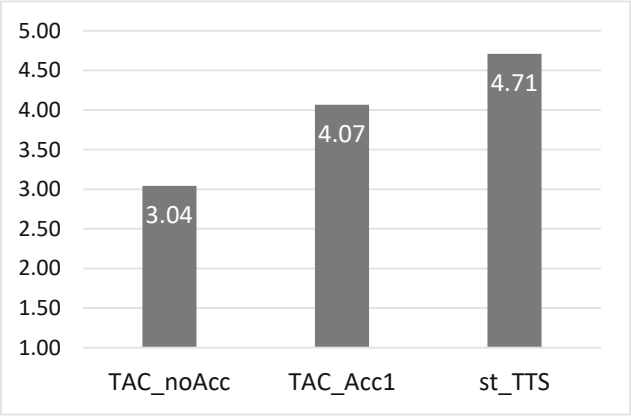

Fig. 2. Results of MOS test – grades for 1–5 scale for quality of different synthesizers.

other for 100 epochs. The rest of the pairs were produced by models trained without accent information, trained for 100, 300, 500, and 900 epochs. Each pair of models was compared. In the last six pairs of utterances, each pair contained one utterance produced with a universal model of HFG and the other with an HFG model trained for the specific Tacotron model used in both utterances.

Listeners are asked to choose the better, i.e. the more natural sounding utterance between the two in each pair, but they are also allowed to choose "no preference" as well.

The results presented in Fig. 3, show that listeners slightly prefer utterances generated by the Tacotron model trained for a longer time. There is also preference in favor of using HFG model adapted to target speaker compared to using universal HFG model as show in Fig. 4. However, in either case the differences are not significant.

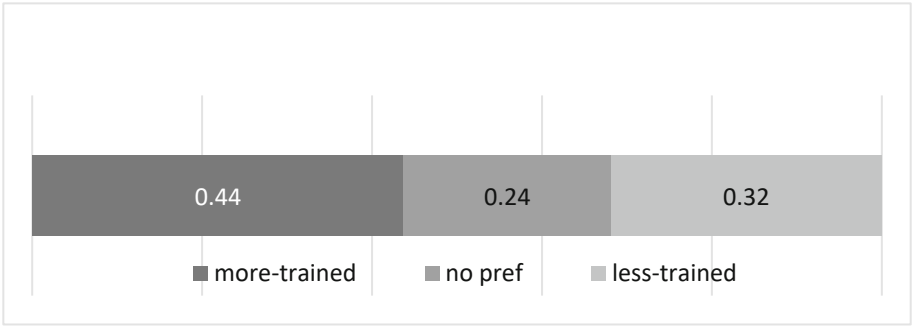

Fig. 3. Results of preference test – more or less trained Tacotron models.

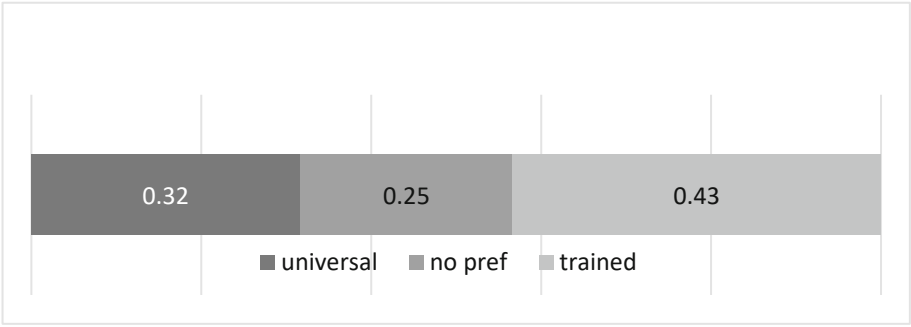

Fig. 4. Results of preference test – universal vs trained HFG model.

4 Discussion

The listening tests provided clear insights into how people perceive the quality of different speech synthesizers and what their main objections are. The MOS test showed that people perceive standard TTS as being of very high quality, grading it 4.7 out of 5 on average. Tacotron-based synthesizers received significantly lower grades, but a more detailed analysis reveals some key conclusions.

Firstly, the differences in grades between the two Tacotron-based models (3.0 and 4.1) indicate that training the same model with and without explicit information about accents in Serbian is crucial for improving the model. This is likely due to the system's inability to properly handle ambiguous vowel characters without sufficient data. When comparing synthesis from the same model trained without accent information, it received a grade of 3.6 for utterances with correct accents and 2.5 for utterances with incorrect accents. The presence of incorrect accents in Serbian not only impairs the naturalness of the synthesis but can also render speech unintelligible or change the meaning of an utterance. It is thus not surprising that the most significant objections from listeners are related to incorrect accents. This problem is largely mitigated by providing accent information during both training and synthesis.

Two methods for incorporating accent information were used: one involving adding accents as additional characters, so that combinations of subsequent characters (vowel + accent) provided full information. In the other approach we adopted, different characters were used for each possible vowel/accent combination, thus providing full information with just one character, although this increased data sparsity. To analyze performance, we synthesized 50 utterances with each of the three Tacotron-based models: the one without accent information (*TAC_noAcc*), the one with accents given as separate characters (*TAC_Acc*), and the third model with different characters for each vowel/accent combination (TAC_Acc1). *TAC_noAcc* produced utterances with at least one incorrectly accented word in 74% of utterances, *TAC_Acc* in 6%, and TAC_Acc1 only in 4% of all utterances. These results suggest that the proposed approaches utilizing accent predictions significantly reduce the problem of incorrect accents even with a relatively small training dataset.

Another problematic aspect of Tacotron-based models, especially when insufficient data is used for training, is the occurrence of hallucinations. These are manifested as randomly synthesized non-existent phonemes, usually at the end of an utterance, or by repeating the last phoneme from the input sentence. While hallucinations do not greatly impact overall intelligibility and naturalness, they are extremely annoying and negatively affect people's perception of the synthesizer's quality. By examining 50 utterances we conclude that hallucinations occur in 56%, 74% and 94% of them, in *TAC_Acc, TAC_noAcc* and TAC_Acc1, respectively. Additionally, in about 10% of utterances, the synthesis is completely unusable as the system fails to produce anything intelligible. The MOS test showed that people rated TAC_Acc1 synthesis at 4.6 when there were no hallucinations, but 3.5 when hallucinations were present. The hallucination problem can only be reduced by providing more training data in case of this model/architecture.

Although the *st_TTS* was graded as the best in the MOS test, likely due to the absence of any hallucinations and incorrect accents, owing to its front-end module and high controllability, listeners gave a slight advantage to *TAC_Acc1* in the MUSHRA test. A more detailed analysis of MUSHRA results shows that natural speech received a grade of 94.6, which is expected, while the next highest grade was 69.6. This significant gap indicates that synthesized speech is still easily distinguishable from natural speech, especially when directly compared with the same utterances produced by natural speakers. The lower grades for *TAC_Acc* and *TAC_noAcc* can be attributed to the more frequent occurrences of incorrect accents and hallucinations, as previously discussed.

However, the slightly lower grade for *st_TTS* compared to TAC_Acc1 (64.8 vs. 69.6) can be explained by the more lively or dynamic synthesis produced by the Tacotron-based model. Although *st_TTS*, when heard alone without any artifacts, hallucinations, or mistakes, sounds very good (receiving a grade of 4.7 out of 5 in the MOS test), hearing it together with Tacotron-based synthesis with the same linguistic content can highlight its lack of expressiveness (Fig. 5).

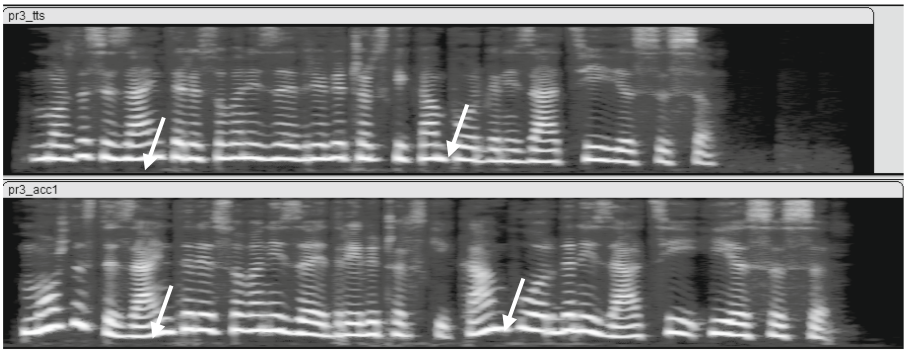

Fig. 5. Spectrograms of utterances with the same linguistic content produced by different synthesizers (the upper one produced by *st_TTS*, the lower one produced by *TAC_Acc1*).

Finally, as regards the results of the preference test, the lack of a clear difference between Tacotron models trained for more or fewer epochs can be explained by Tacotron's tendency to overfit easily, although the more trained versions were slightly

favored. Additionally, the authors confirmed that training for more epochs did not reduce the percentage of produced hallucinations nor did it improve accent learning.

Another conclusion from the preference test is that there is no significant difference between using a trained or universal HFG-based vocoder, although the trained one had a slight advantage. The authors find it more significant to use the trained version of the vocoder. The reason is the occurrence of artifacts and slight buzzing when using the universal HFG-based vocoder, but these issues were probably not prominent or annoying in the short and few examples that listeners heard during the test.

5 Conclusion

In this paper, we present a TTS (Text-to-Speech) system for end-to-end synthesis in Serbian, based on the Tacotron architecture. Due to the lack of a large, high-quality speech database in Serbian, the system was created by adapting a pre-trained English model. Initial experiments revealed issues with appropriately generating accents in Serbian. To address this, the authors proposed two methods involving modules for accent prediction from text. The approach using different symbols for each accented vowel produced better results. Although the Tacotron-based system can outperform the current best Serbian synthesizer, which uses separate front-end and DNNs, in some contexts, errors typical of sequence-to-sequence models, such as hallucinations and repetitions, significantly decrease the overall performance of the system.

Future work will include attempts to overcome data sparsity problems, especially with accents, by augmenting the training set using TTS-generated data. The authors also plan to explore newer architectures.

Acknowledgments. This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7449, Multimodal multilingual human-machine speech communication, AI-SPEAK.

References

1. Tan, X., Qin, T., Soong, F., Liu, T.Y.: A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561 (2021)
2. Delić, V., et al.: Speech technology progress based on new machine learning paradigm. Comput. Intell. Neurosci. **2019**(1), 4368036 (2019)
3. Tan, X., et al.: Naturalspeech: end-to-end text-to-speech synthesis with human-level quality. IEEE Trans. Pattern Anal. Mach. Intell. (2024)
4. Van Den Oord, A., et al.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 12 (2016)
5. Ren, Y., et al.: Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558 (2020)
6. Wang, Y., et al.: Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)
7. Shen, J., et al.: Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: 2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), pp. 4779–4783. IEEE (2018)

8. Ping, W., et al.: Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)
9. Ren, Y., et al.: Fastspeech: fast, robust and controllable text to speech. Adv. Neural Inf. Proc. Syst. **32** (2019)
10. Mu, Z., Yang, X., Dong, Y.: Review of end-to-end speech synthesis technology based on deep learning. arXiv preprint arXiv:2104.09995 (2021)
11. Mishev, K., Karovska Ristovska, A., Trajanov, D., Eftimov, T., Simjanoska, M.: MAKEDONKA: applied deep learning model for text-to-speech synthesis in Macedonian language. Appl. Sci. **10**(19), 6882 (2020)
12. Sofronievski, B., et al.: Macedonian speech synthesis for assistive technology applications. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 1183–1187. IEEE (2022)
13. Secujski, M.S.: Obtaining prosodic information from text in Serbian language. In: EUROCON 2005-The International Conference on Computer as a Tool, vol. 2, pp. 1654–1657. IEEE (2005)
14. Kong, J., Kim, J., Bae, J.: Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. Adv. Neural. Inf. Process. Syst. **33**, 17022–17033 (2020)
15. NVIDIA. Tacotron 2. GitHub repository, https://github.com/NVIDIA/tacotron2. Accessed 23 May 2024
16. Delić, T., Sečujski, M., Suzić, S.: A review of Serbian parametric speech synthesis based on deep neural networks. Telfor J. **9**(1), 32–37 (2017)
17. Suzić, S., Pekar, D., Sečujski, M., Nosek, T., Delić, V.: HiFi-GAN based Text-to-Speech Synthesis in Serbian. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 2231–2235. IEEE (2022)
18. Secujski, M., Pekar, D., Suzic, S., Smirnov, A., Nosek, T.V.: Speaker/style-dependent neural network speech synthesis based on speaker/style embedding. J. Univers. Comput. Sci. **26**(4), 434–453 (2020)
19. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
20. Keith Ito. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset. Accessed 23 July 2024