Speaker Adaptation Experiments with Limited Data for End-to-End Text-To-Speech Synthesis using Tacotron2

Ali Raheem Mandeel, Mohammed Salah Al-Radhi, and Tamás Gábor Csapó

***Abstract*—Speech synthesis has the aim of generating human-like speech from text. Nowadays, with end-to-end systems, highly natural synthesized speech can be achieved if a large enough dataset is available from the target speaker. However, often it would be necessary to adapt to a target speaker for whom only a few training samples are available. Limited data speaker adaptation might be a difficult problem due to the overly few training samples. Issues might appear with a limited speaker dataset, such as the irregular allocation of linguistic tokens (i.e., some speech sounds are left out from the synthesized speech). To build lightweight systems, measuring the number of minimum data samples and training epochs is crucial to acquire a reasonable quality. We conducted detailed experiments with four target speakers for adaptive speaker text-to-speech (TTS) synthesis to show the performance of the end-to-end Tacotron2 model and the WaveGlow neural vocoder with an English dataset at several training data samples and training lengths. According to our investigation of objective and subjective evaluations, the Tacotron2 model exhibits good performance in terms of speech quality and similarity for unseen target speakers at 100 sentences of data (pair of text and audio) with a relatively low training time.**

***Index Terms*—speech synthesis, TTS, Tacotron2, WaveGlow, Few-shot.**

I. INTRODUCTION

Speech technology is modern, rapidly developing interdisciplinary field dealing with the artificial intelligence (AI) implementation of any element of the natural human speech chain (such as the speaker, the listener, or even the transmission medium). Several disciplines such as phonetics, machine learning, signal processing, speech acoustics, and cognitive sciences have been used altogether in this field. Based on the available statistical data, it is interesting to say that today we know of just over 7000 living, spoken languages, which poses a serious challenge for speech technology professionals in terms of the uniqueness of each language and the diversity of the linguistic environment [1].

In parallel with the development of infocommunications technology, the need for average users to ensure that language differences, communication features, decrease in the size of devices and the increase in our expectations, as well as other obstacles (e.g., physical resource limitations, functional errors) do not hinder access to certain functions and developments. It is a legitimate user expectation that human speech as a "periphery" should be available for the use of information systems instead of or in addition to peripherals that can be connected to devices (keyboard, mouse, display)[1], [2].

Speech processing (including text-to-speech (TTS) synthesis and automatic speech recognition (ASR)) is beneficial in various fields such as healthcare, security, industry, education, and recreation [3], [4], [5], [6]. Voice disorders in children, such as dysphonia, could be detected early using ASR approaches [3]. The system distinguishes between healthy and pathological sounds using the noisy aware approach. Similarly, ASR might be used to enhance security by catching the target phrases in long sequential sentences or distinct speaker identities. With the advancement of Industry 4.0, machines will become more intelligent, collaborative, and multi-purpose in the future. Laborers can quickly finish the duties by listening to the synthesized speech instructions. Moreover, TTS promises considerable advantages, such that individuals would not need to lose attention by checking the instructions. Furthermore, voiced commands help to have hands-free functions [7].

A. Speech synthesis

Speech synthesis (frequently abbreviated as TTS) has the aim to create natural, human-like voice from written texts. This field has a long history in speech and natural language processing. For a long time, speech synthesis has been a challenging problem. For example, decreasing the TTS model size for real-time synthesis has required much effort and time to enhance the computational complexity. Low resource scenarios, inadequate speakers dataset, robustness problems, expressiveness, and naturalness, have occupied researchers' minds. Much research has been conducted to enhance the quality of synthesized speech in terms of prosody, intelligibility, expressiveness, emotion, robustness, style, naturalness, controllability, etc [8], [9], [10], [11].

A speech synthesizer pipeline basically includes a text analysis module, an acoustic model, and a vocoder (i.e., a speech encoder/decoder module which can decompose the speech signal to a few parameters). A text sequence is converted to linguistic characteristics or phonemes via the text analysis module. Acoustic characteristics are derived from linguistic features or phonemes via acoustic models. At last, vocoders create waveforms based on acoustic/linguistic characteristics. As opposed to the above traditional TTS pipeline, end-to-end

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
E-mail: {alimandeel, malradhi, csapot}@tmit.bme.hu

DOI: 10.36244/ICJ.2022.3.7

TTS systems instantly transform characters or phonemes into synthesized speech, often without a linguistic frontend and without a traditional vocoder.

One of the early text-to-speech technologies was articulatory synthesis. Articulatory synthesis generates speech by mimicking the properties and movements of the human articulators such as the glottis, tongue, lips, and vocal tract in general [12]. Another historical technique is formant synthesis, which generates speech using a reduced source-filter paradigm that is controlled by a set of manually-defined parameters [13]. Concatenative synthesis was introduced with the idea of the concatenation of well chosen speech segments from a database [14]. The database comprises audio clips from complete sentences of recorded syllables from voice actors. Unit selection synthesis has the idea that many of such elements are available, and the algorithms try to find large enough units from the natural speech recordings of several hours. Even though the sound quality generated by these methods can be excellent in intelligibility, this approach has limitations. It consumes many resources (especially memory for storing the units) and often results in a speech with reduced smoothness in prosody (pitch, stress or timing).

Statistical Parametric Speech Synthesis (SPSS) was proposed as an alternative to concatenative and unit selection synthesis [15]. It decomposes speech to acoustic parameters and then uses vocoder algorithms to recover speech from the produced acoustic parameters [15], [16], [17], [18], [19]. The essential benefit of SPSS is its adaptability in terms of voice features, speaking styles, and emotions [15]. Typically, two machine learning techniques have been used in SPSS: hidden Markov-models (HMMs) [15] and deep neural networks (DNNs) [16]. Most recently, novelties in deep learning have allowed the creation of neural network-based speech synthesis, which uses deep neural networks as the machine learning model for TTS. The benefit of DNN-TTS over prior systems is its excellent speech quality (naturalness and intelligibility) and the fact that it demands fewer engineering preprocessing and feature creation. Moreover, being efficient during synthesis is as important as obtaining high-quality synthesized speech.

B. Speaker adaptation for speech synthesis

Speaker adaptation (also called voice cloning or custom voice) is the process of customizing the synthesizer to create a voice for any target speaker. Personalizing a TTS is a popular feature in which the application creates sound employing any target speaker's voice recordings. In this case, the general TTS model is trained with an extensive multi-speaker dataset and then adapted to a target speaker. One aspect of adaptive TTS is an efficacious adaptation setting, which reduces adaptation parameters and data per target speaker. The most likely situation in which speaker adaptation is used is when the dataset for the target speaker is too small for single speaker training, but at least enough for adaptation.

More data will lead to enhanced speech quality [20], but it means at a considerable expense of gathering data. Correspondingly, extra fine-tuning parameters can improve the synthesized speech quality, but it increases memory, and implementation costs [21]. At the same time, we might also suffer from the lack of availability of the target speaker's speech data. Accordingly, creating a TTS model which can work with extremely limited data (a few sentences) would be a solution to these problems.

C. Limited data speaker adaptation

Numerous researchers have investigated the speaker adaptation options for end-to-end speech synthesis with few samples (sentences) and tried to enhance the synthesized speech quality with these models. A study on Tacotron2 proficiency of speaker adaptative speech synthesis was accomplished with a Romanian dataset [22]. Their work concluded that it is sufficient to obtain a speaker's identity (a target speaker's voice attributes) with only one sample of data (i.e., one single sentence from the target speaker). For Spanish and Basque, the performance of the Tacotron2-based system was examined with limited amounts of data [23]. Guided attention was implemented, which provided the system with the explicit duration of the phonemes to reduce lost alignment during the inference process. Otherwise, in non-end-to-end SPSS, many studies have been accomplished for speaker adaptation, such as our previous work [24], which used a Continuous vocoder with limited target speaker data for about 14 minutes.

The speech quality was examined with varying quantities of adaption data using Deep Voice 3, and Griffin-Lim [20]. They studied the combination of two methods: speaker encoding and speaker adaptation. Even with a few adaptation audios, both ways generated adequate results. Speaker adaptation achieved slightly better naturalness than speaker encoding. Another relevant paper investigated three meta-learning variations for sample efficient adaptive TTS [25]. Multi-speaker TTS architecture was updated to allow the cloning of unknown speakers using only a few shot samples (a few training data are emerging). The fine-grained and coarse-grained encoders generate two sorts of embeddings: variable-length and global embeddings. The speaker adaption approach could be improved by using a meta-learning algorithm (Model Agnostic Meta-Learning (MAML)) [25]. Overall, while several efforts have been made to improve limited data speaker adaption models, there is still a gap in investigating the amount of English target data required to achieve adequate speaker identity, similarity and high-quality synthesized speech.

D. End-to-end TTS: Tacotron2

Tacotron2 is an end-to-end neural network architecture for TTS [26]. It consists of a network to convert character sequences of input text to mel spectrogram and a neural vocoder (see next section), which can synthesize the speech (Fig. 1). The sequence-to-sequence architecture can elevate this problem by translating the input text sequences into magnitude spectrograms. As a result, this method reduces the need for sophisticated language and speech information because it uses raw data.

The architecture of the Tacotron2 network mixes long short-term memory (LSTM) and convolutional neural network

(CNN) layers to produce mel spectrogram frames from input character sequences. It employs the Short-Time Fourier Transform (STFT) to compute mel spectrograms. Further, this network has an encoder to transform character sequences to interior features representation and a decoder to change these interior features to spectrograms frames. The advantage of using mel spectrograms as an intermediate value is to enable shorter training of the network part and the vocoder. Also, it emphasizes the distinction of low-frequency speech over high-frequency sounds. These benefits ensure that using spectrograms to synthesize speech patterns is intelligible and natural, therefore we also use Tacotron2 in the current paper.

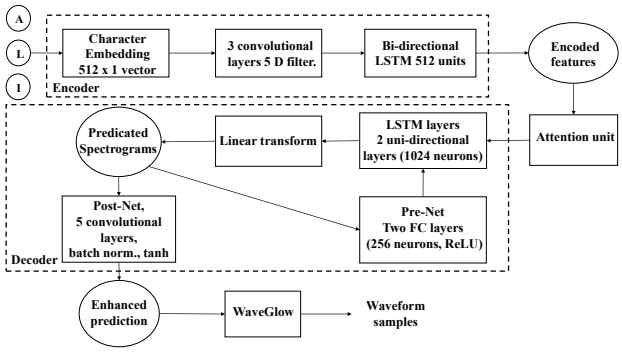

Fig. 1: Tacotron2 architecture.

E. Neural vocoder: WaveGlow

WaveGlow is a neural vocoder, i.e. a generative audio model that samples from distribution to produce waveforms [27]. The number of dimensions in this distribution must match the number of dimensions in the intended output. The samples are acquired from the distribution travel through the flow steps to rebuild the synthesized voice. Twelve coupling layers and twelve invertible 1x1 convolutions exist in the original WaveGlow architecture. Each of the eight layers of dilated convolutions in the affine coupling layer networks has 256 channels for skip links and 512 channels for residual connections [27]. It is a fast model which allows parallel synthesis at 500 kHz on an NVIDIA v100 GPU. Synthesized voices are sharp and close to the real distribution because no Mean squared error (MSE) loss is used for the model training. Moreover, it gives a tractable likelihood of the training data. According to [27], WaveGlow generates high-quality synthesized speech from mel-spectrograms and delivers a quick, efficient voice synthesis. It is faster than early WaveNet versions, therefore we use this neural vocoder in our study.

F. Goal of the current study

This work aims to use as minimal data and parameters as feasible while maintaining good synthesized speech quality during speaker adaptation for end-to-end speech synthesis. We built a TTS model based on the Tacotron2 framework and the WaveGlow neural vocoder. We tested Tacotron2 with five limited data sizes (15, 20, 35, 70, and 100 utterances) from each speaker. We defined three checkpoints (300, 700, and 900) for each target speaker and dataset scenarios. A checkpoint is a moment in the model's state where the current learning rate, weights, and other parameters are stored.

The remainder of the paper is organized in the following manner. Section II describes the experimental design, tools, and dataset. Section III delves more into the experimental results and findings. Finally, the conclusion is presented in Section IV.

II. METHODS

We utilized a multi-speaker dataset with high-quality recordings to train the average end-to-end Tacotron2 model. Then, we adapted this model with four target speakers (two females and two males). Objective and subjective evaluations have been done to test the naturalness and similarity of synthesized speech.

A. End-to-end TTS and neural vocoder

For the Tacotron2, we used the open-source solution offered by NVIDIA (https://github.com/NVIDIA/tacotron2). We employed the official pre-trained WaveGlow vocoder (on the LJ speech dataset with sampling rate 22050 Hz [28]) offered by NVIDIA (https://github.com/NVIDIA/waveglow) with the design of 12 coupling layers, eight dilated convolution layers, 512 residual, and 256 skip connections. This architecture design of WaveGlow is proposed by NVIDIA [27].

B. Speech corpus

In our study, the Tacotron2 model was trained with an English Hi-Fi multi-speaker dataset [29]. This dataset comprises texts from Project Gutenberg and audiobooks from LibriVox. It has approximately 292 hours of speech from ten native English speakers (six females and four males). Every speaker has a minimum of 17 hours of speech sampled at 44.1 kHz in WAV format. Then, we re-sampled it to 22050 Hz to be compatible with the pre-trained WaveGlow vocoder, which was trained with a dataset of 22050 Hz sample rate.

Based on signal-to-noise ratio (SNR) investigation, the Hi-Fi TTS corpus is divided into two categories:

1) The clean subset comprises high-quality audiobooks with adequate audio qualities (at least 40 dB),
2) The other set covers books with fewer SNR (a minimum of 32 dB).

C. Training topology

We used a high-performance NVidia Titan X graphics processing unit (GPU) for training the Tacotron2 model. We used four speakers (two females: Helen Taylor and Sylviamb and two males: Mike Pelton and Tony Oliva) to train the Tacotron2 average model. The total dataset of the four speakers is 88.3 hours (Helen Taylor: 24.3 hours, Sylviamb: 22.2 hours, Mike Pelton: 17.7 hours, and Tony Oliva: 24.1 hours). The dataset from the four Hi-Fi speakers was divided into the training and validation sets for the seen speakers. The validation set consisted primarily of 5% utterances from each speaker.

The Tacotron2 encoder was fed a character sequence, with

each character encoded as a 512-dim character embedding. A spectrogram was created using a 2048 point Fourier transform with Hann windowing, a shift of 16 milliseconds, and a duration of 64 milliseconds. Next, we made a mel spectrogram out of it, with frequency bins of 80 ranging from 125 Hz to 7.6 kHz. We utilized the Adam optimizer [30], a 0.000001 weight decay value, a 0.001 learning rate, hop_length equals 256, iterations_per_checkpoint=1000 (a batch of data has been passed through), and a frame size of 1024. We trained the Tacotron2 model using the four above-mentioned speakers from the Hi-Fi TTS corpus (88.3 hours) until checkpoint 870 000 (870 epochs, an epoch is the number of complete passes through the training dataset). We stopped the training at 870000 based on an investigation of the synthesis quality, which did not improve further after this point. For the whole training procedure, the batch size was set at eight. In addition, we used a pre-trained WaveGlow vocoder provided by NVIDIA (https://github.com/NVIDIA/waveglow).

D. Transfer learning / Speaker adaptation

After that, we adapted the Tacotron2 model, which trained with multi speakers, with four target speakers (two females and two males), namely Female1_other (Maria Kasper), Female2_clean (Cori Samuel), Male1_other (Phil Benson), Male2_clean (John Van Stan) from the Hi-Fi dataset. We used "clean" and "other" sets from the dataset to show the impact of the high-quality and low SNR audios on speaker adaptative synthesized speech quality. We employed two target speakers (female and male/ Female2_clean and Male2_clean) with the "clean" dataset and two target speakers (female and male/ Female1_other and Male1_other) with the "other" dataset. We used the training parameters listed in Table I. The lowest target speaker dataset was 0.48 minutes, and the greatest was 14.32 minutes. During this training, we reduced the batch size to four and iterations_per_checkpoint to 100. The durations of the used datasets are mentioned in the same table. We trained Tacotron2 with three checkpoints (300, 700, and 900). Essential facts to exemplify, we noticed during the training of Tacotron2 that the synthesized speech quality did not be enhanced much over checkpoint 900, or the quality was the same. Therefore, the model reached a stable state, and training it more leads to overfitting.

TABLE I
TARGET SPEAKERS' DATA FOR THE EXPERIMENTS.

Speakers	Dataset (sentences)	Duration (minutes)	Dataset division (training / validation)	Checkpoints
Female1_other	15	0.73	15 / 1	300, 700, 900
	20	1.72	20 / 2	
	35	0.91	35 / 2	
	70	3.84	70 / 4	
	100	5.94	100 / 5	
Female2_clean	15	0.48	15 / 1	300, 700, 900
	20	0.76	20 / 2	
	35	1.47	35 / 2	
	70	2.82	70 / 4	
	100	3.97	100 / 5	
Male1_other	15	0.7	15 / 1	300, 700, 900
	20	0.87	20 / 2	
	35	1.7	35 / 2	
	70	3.56	70 / 4	
	100	5.1	100 / 5	
Male2_clean	15	2.12	15 / 1	300, 700, 900
	20	2.9	20 / 2	
	35	5.25	35 / 2	
	70	10.29	70 / 4	
	100	14.32	100 / 5	

III. RESULTS

We compare the Tacotron2 model performance on different sizes of training data from target speakers ranging from 15 to 100 utterances and three training periods to demonstrate the training efficiency. Both objective and subjective evaluations are carried out. Using an attention-based model, we aim to achieve good alignment with minimal data and a short training period.

A. Objective evaluation

1) MCD (dB):
Mel cepstral distortion (MCD) is a metric that measures the similarity between the spectra of two sounds [31]. The lower the MCD value between synthesized and natural mel cepstral sequences, the more similar a synthetic voice is to a natural one (Eq. 1). x and y are the Mel-cepstrum of the original and synthetic voice waveforms, respectively. M is the order of Mel-cepstrum. The dynamic time warping algorithm was used before making the comparison because sequences are not aligned. We used an open-source DTW-MCD implementation (https://github.com/jasminsternkopf/mel_cepstral_distance).

$$MCD = \frac{10}{\log 10}\sqrt{\sum_{m=1}^{M}\left(x(m) - y(m)\right)^2} \qquad (1)$$

Table II details the MCD results of the synthesized speech sentences that we obtained for the four target speakers. We noticed that the minimum MCD values for the target speakers Female1_other (8.51) and Male2_clean (10.6) are obtained by the checkpoint-900 at 70 and 35 samples. At the same time, we obtained the lowest MCD value for the target speaker at checkpoint 900 with only 35 samples. Also, we noticed that the speaker Female2_clean did not offer common tendencies with increasing data and training periods. We believe the reason for this unusual behavior is because of the smaller dataset duration for this speaker compared to others (see Table I).
Fig. 2 shows the average MCD values of the four target speakers. We can conclude for each data sample of 35, 70, and 100, the MCD values decreased as the training got higher. For example, the MCD values of the 35 samples (12.46 / checkpoint-300, 11.82 / checkpoint-700, and 11.26/ checkpoint-900). Otherwise, the average MCD values did not decline as the data raised from 20 to 100 samples. Overall, the MCD metric did not reflect the expected patterns in certain circumstances as limited data and training periods increased.

2) Encoder-Decoder alignment graphs analysis:
The attention mechanism function is considered as a

TABLE II
THE RESULT OF THE MCD BASED ON THE LIMITED DATA.

Speaker	Dataset	MCD checkpoint-300	MCD checkpoint-700	MCD checkpoint-900
Female1_other	15	10.73	10.43	10.79
	20	9.52	9.65	9.66
	35	9.67	9.35	9.00
	70	9.33	8.65	**8.51**
	100	10.26	9.16	8.91
Female2_clean	15	12.59	14.58	13.22
	20	**12.42**	13.41	13.33
	35	15.52	14.71	13.01
	70	18.18	15.96	15.31
	100	16.42	15.87	14.77
Male1_other	15	16.77	15.2	16.91
	20	12.87	12.73	14.24
	35	13.45	**12.15**	12.45
	70	12.27	12.88	12.3
	100	13.03	12.71	12.87
Male2_clean	15	13.31	11.32	11.38
	20	11.64	11.1	11.4
	35	11.21	11.08	**10.6**
	70	11.58	13.73	10.64
	100	11.66	11.03	11.1

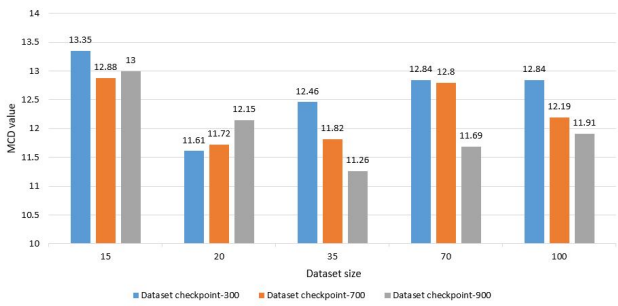

Fig. 2: The average MCD of the four target speakers.

duration model in learning the time alignment between the input text row (encoder) and the outcome acoustic sequence (decoder). The success of an end-to-end model relies on attention alignment. The attention mechanism's irregular and inexact alignment results in word repetitions, mispronunciations, and skipping [32]. The attention alignment graph is a way of showing the quality of end-to-end TTS models, as it shows how well the decoder attends to encoder input [32]. The encoder gradually receives the input and generates status vectors. It examines all status vectors and sequentially generates audio frames. The sloping line appears when audio frames are developed by concentrating on the proper input characters. In other words, the inclination of the diagonal line is an indicator of the quality of the produced speech. The clear near the diagonal line is the sign or the criteria that the alignment graph should meet and be considered as sufficient alignment. The figures (Fig. 3 and Fig. 4) depict the process of attention learning for the speaker Male1_other and the Female1_other for the synthesized sentences *"it is a curious little church,"* and *"in his own mind or that of the public,"* respectively utilizing 15, 20, 35, 70, and 100 samples of training data at three checkpoints (300, 700, and 900). These figures show how the text-to-spectrogram prediction network improved learning attention during the training process and increased data samples. With the smallest amount of training data (15 sentences), clearly the alignment path is not accurate (attention failures), indicating that this is not enough for proper training, even with a high checkpoint number. With training data of at least 20 samples, and after checkpoint 300 of training, the Tacotron2 model began to pick up on alignment. Despite that, the attention line alignment for the data samples of 35 and 70 at checkpoint 300 showed irregular behavior for the speaker Male1_other – this tells us that in general, checkpoint 300 is not enough but the network should be trained longer. In conclusion, according to the encoder-decoder alignment graphs objective evaluation, at least 20 sentences and a checkpoint of 700 shows a good alignment (close to a straight line). for these two sample cases. As we noticed, 15 sentences of data (at several training times) and checkpoint 300 (at various datasets) are insufficient to obtain a proper alignment. Therefore, at this stage of objective evaluation, we nominated 20 sentences/checkpoint 700 to be the minimum threshold that encoder-decoder alignment makes an acceptable outcome.

B. *Subjective evaluation*

In order to determine which proposed version is closer to natural speech, we conducted an online MUSHRA-like test [33]. Our aim was to compare the natural sentences with the synthesized sentences depending on the training data size (number of sentences: 15 / 35 / 70 / 100) and training time (checkpoint: 300 / 700 / 900). In the test, the listeners had to rate the naturalness of each stimulus in a randomized order relative to the reference (which was the natural sentence), from 0 (very unnatural) to 100 (very natural). As a lower anchor, we used the fewest data size and training (15 samples, checkpoint 300) because of its poor quality. We chose three sentences from the test set of the four speakers used in the adaptation experiments. The variants appeared in randomized order (different for each listener). The samples can be found at https://aliraheem.github.io/infocommunications_journal_2022/. With this test, we experimented the following seven variants for the four target speakers:

(a) natural voices,
(b) synthesized voices at 15 sentences/checkpoint-300,
(c) synthesized voices at 35 sentences/checkpoint-900,
(d) synthesized voices at 70 sentences/checkpoint-900,
(e) synthesized voices at 100 sentences/checkpoint-300,
(f) synthesized voices at 100 sentences/checkpoint-700,
(g) synthesized voices at 100 sentences/checkpoint-900.

As a result, we will have an opportunity to observe the fixed 100 sentences at different checkpoints (300, 700, and 900) and with fixed checkpoint 900 at a different number of sentences used as adaptation data (35, 70, and 100). We fitted the four target speakers and three sentences from each of them. Thus, we have 84 sentences (4 speakers x 3 sentences x 7 variants) to compare. Twenty-one subjects (18 in quiet rooms, 3 in a noisy environment; ten females, 11 males; 19- 48 years old;

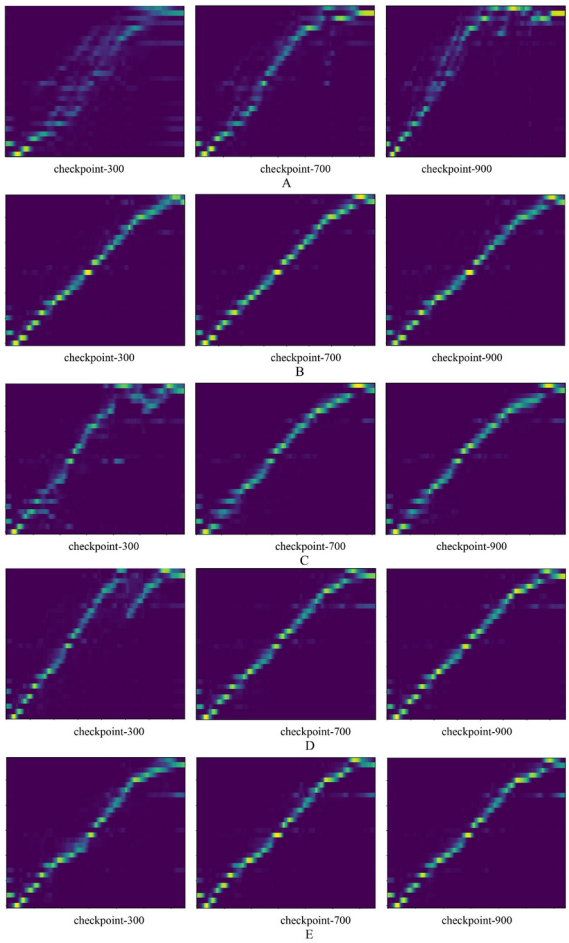

Fig. 3: The Attention alignment graph of the Male1 other speaker. The horizontal axis denotes the decoder time-steps of the creation speech sequence with a max of 175 frames. The vertical axis represents encoder time-steps a most of the 20 phonemes. A= 15 sentences, B= 20 sentences, C= 35 sentences, D= 70 sentences, and E= 100 sentences.

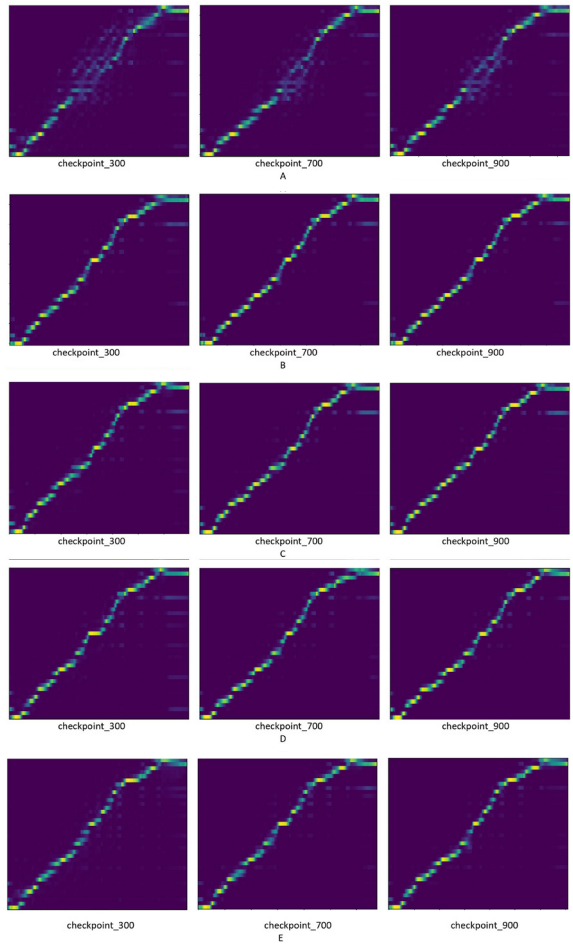

Fig. 4: The Attention alignment graph of the Female1 other speaker. The horizontal axis denotes the decoder time-steps of the creation speech sequence with a max of 175 frames. The vertical axis represents encoder time-steps a most of the 20 phonemes. A= 15 sentences, B= 20 sentences, C= 35 sentences, D= 70 sentences, and E= 100 sentences.

none of them were native English) volunteered to do the test. The test duration was 13.5 minutes, on average.

Fig. 5 presents the average naturalness scores for the different categories (data samples and checkpoints). The natural utterances obtained 88% out of 100% from the subjects. We plotted the '100 samples/checkpoint-900' two times - this way, it is easy to compare the effect of data size visually and checkpoint size. Increasing the data samples from 35 to 70 with fixing checkpoint 900 exhibited a remarkable increase in synthesized speech naturalness (35 samples= 38%, 70 samples= 45%). On the other hand, increasing the data to 100 showed a slight improvement in the synthesized speech naturalness (47%) above the naturalness of data of 70 sentences. Additionally, fixing the data samples at 100 samples and increasing the training periods (checkpoints= 300, 700, and 900) demonstrated the same behavior as the previous case by increasing the speech naturalness (checkpoint 300= 33%, checkpoint 700= 39%, and checkpoint 900= 47%). Therefore, increasing the training period creates more natural speech. For example, comparing the samples 70 and 100, we notice the naturalness of synthesized speech at 70 samples at checkpoint 900 received more scores than 100 samples at checkpoint 700. Similarly, in the case of 35 samples at checkpoint 900 has higher rates than 100 samples at checkpoint 300.

Next, Fig. 6 displays the speech naturalness speaker by speaker. We noticed that increasing the data samples from 35 to 100 with fixing checkpoint 900 showed that the 100 data samples' speech naturalness is more than 35 with all four target speakers. Nevertheless, the 70 samples' speech naturalness did not show the same behavior for all target speakers. Meanwhile, increasing the training period from checkpoint 300 to checkpoint 900 with 100 samples keeps the same scenario with the average case by gaining more speech naturalness.

Furthermore, it appears that increasing the sample size from 70 to 100 decreases the naturalness of speech for "other" data speech, whereas the opposite is true for "clean" data

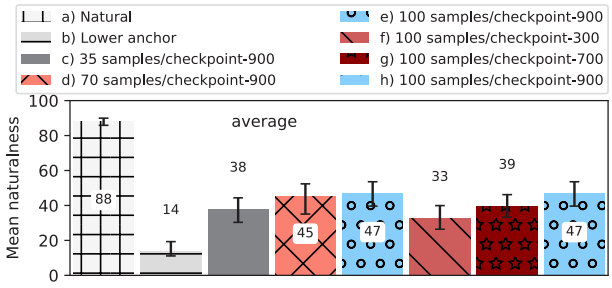

Fig. 5: Average naturalness ratings of the four speakers' speech.

speech. Similarly, using checkpoint 900 only slightly improves performance compared to using checkpoint 700 for "other" data speech. In contrast, the improvement for "clean" data speech (between options f and g) is much more noticeable.

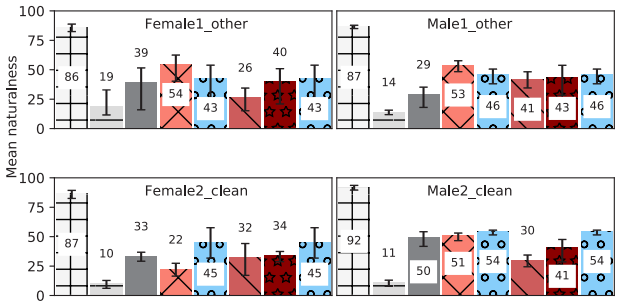

Fig. 6: Naturalness ratings of the four speakers' speech naturalness (speaker by speaker). A higher value indicates a higher level of overall quality. Errorbars represent the bootstrapped 95 percent confidence intervals.

IV. Conclusion

We investigated the minimum dataset and training period required and experimented with the Tacotron2 end-to-end TTS and WaveGlow neural vocoder to construct a TTS model with an unseen target speaker's dataset. First, we trained a general model with a multispeaker dataset of 88.3 hours, after which we applied speaker adaptation. Four target speakers (two females and two males) with two kinds of audio qualities (clean of SNR at least 40 dB and other of SNR equal to 30 dB) were used for the speaker adaptation. We conducted objective and subjective evaluation experiments. Based on our evaluations, the Tacotron2 model admirably produces synthesized speech quality and resemblance with 100 sentences of data (at least five minutes) with a relatively short training period (checkpoints 900) for both speakers of both genders. We did not find a direct relation between the SNR of the adaptation audio dataset and the quality of the synthesized speech between the four speakers' suggested data to build the system (100 sentences). These outcomes can be beneficial to building applications with personalized text-to-speech synthesis, e.g., in speech communication aids for the speaking impaired.

Acknowledgment

The research was partially sponsored by the APH-ALARM project (contract 2019-2.1.2-NEMZ-2020-00012), funded by the European Commission and the National Research, Development and Innovation Office of Hungary and supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. The research reported in this publication, carried out by the Department of Telecommunications and Media Informatics Budapest University of Technology and Economic and IdomSoft Ltd., was supported by the Ministry of Innovation and Technology and the National Research, Development and Innovation Office within the framework of the National Laboratory of Infocommunication and Information Technology. Tamás Gábor Csapó's research was supported by the Bolyai János Research Fellowship of the Hungarian Academy of Sciences and by the ÚNKP-21-5 (identifier: ÚNKP-21-5-BME-352) New National Excellence Program of the Ministry for Innovation and Technology from the source of the National, Research, Development and Innovation Fund. The Titan X GPU used was donated by NVIDIA Corporation. We would like to thank the subjects for participating in the listening test.

References

[1] G. Németh,"Why is speech technology important and what is it good for? – Hungarian successes firsthand." in Hungarian: Miért fontos és mire jó a beszédtechnológia?– magyar sikerek első kézből," Infocommunications Journal, vol. 70, pp. 12–16, 2015.

[2] G. Németh, G. Olaszy, K. Vicsi, and T. Fegyó, "Speaking machines?! in Hungarian: "Beszélgető gépek?!," HÍRADÁSTECHNIKA: HÍRKÖZLÉS-INFORMATIKA, vol. 64, pp. 53–57, 2009.

[3] Miklós Gábriel Tulics and Klára Vicsi, "Automatic classification possibilities of the voices of children with dysphonia", Infocommunications Journal, Vol. X, No 3, September 2018, pp. 30-36. DOI: 10.36244/ICJ.2018.3.5

[4] Masakazu Kanazawa, Atsushi Ito, Kazuyuki Yamasawa, Takehiko Kasahara, Yuya Kiryu and Fubito Toyama, "Method to Predict Confidential Words in Japanese Judicial Precedents Using Neural Networks With Part-of-Speech Tags", Infocommunications Journal, Vol. XII, No 1, March 2020, pp. 17-25. DOI: 10.36244/ICJ.2020.1.3

[5] G. Du, M. Chen, C. Liu, B. Zhang and P. Zhang, "Online Robot Teaching With Natural Human–Robot Interaction," in IEEE Transactions on Industrial Electronics, vol. 65, no. 12, pp. 9571-9581, Dec. 2018, DOI: 10.1109/TIE.2018.2823667.

[6] P. -S. Chiu, J. -W. Chang, M. -C. Lee, C. -H. Chen and D. -S. Lee, "Enabling Intelligent Environment by the Design of Emotionally Aware Virtual Assistant: A Case of Smart Campus," in IEEE Access, vol. 8, pp. 62032-62041, 2020, DOI: 10.1109/ACCESS.2020.2984383.

[7] Jean D. Hallewell Haslwanter, Michael Heiml, and Josef Wolfartsberger. Lost in translation: machine translation and text-to-speech in industry 4.0. In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '19). Association for Computing Machinery, USA, pp. 333–342, 2019, DOI: 10.1145/3316782.3322746.

[8] G. C. Tamás, Z. Csaba, and N. Géza, "A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System," Infocommunications Journal, vol. 65, no. 1, pp. 32–37, 2010.

[9] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech." Advances in Neural Information Processing Systems, vol. 32, pp. 3165–3174, 2019.

[10] G. C. Tamás and G. Németh, "A novel irregular voice model for HMM-based speech synthesis," in ISCA 8th Speech Synthesis Worksop (SSW8), 2013, pp. 229–234.

[11] A. Łańcucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6588–6592, DOI: 10.1109/ICASSP39728.2021.9413889.

[12] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," in 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, pp. 121–126, 2001.

[13] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," the Journal of the Acoustical Society of America, vol. 67, no. 3, pp. 971–995, 1980, DOI: 10.1121/1.383940.

[14] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech communication, vol. 9, no. 5-6, pp. 453–467, 1990, DOI: 10.21437/eurospeech.1989-172.

[15] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," speech communication, vol. 51, no. 11, pp. 1039–1064, 2009, DOI: 10.1016/j.specom.2009.04.004.

[16] H. Ze, A. Senior and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7962-7966, DOI: 10.1109/ICASSP.2013.6639215.

[17] S. King, "An introduction to statistical parametric speech synthesis," Sadhana, vol. 36, no. 5, pp. 837–852, 2011, DOI: 10.1007/s12046-011-0048-y.

[18] T. G. Csapó, G. Németh, M. Cernak and P. N. Garner, "Modeling unvoiced sounds in statistical parametric speech synthesis with a continuous vocoder," in EUSIPCO, IEEE, 2016, pp. 1338-1342, DOI: 10.1109/EUSIPCO.2016.7760466.

[19] M. S. Al-Radhi, O. Abdo, T. G. Csapó, S. Abdou, G. Németh, and M. Fashal, "A continuous vocoder for statistical parametric speech synthesis and its evaluation using an audio-visual phonetically annotated Arabic corpus," Computer Speech and Language, vol. 60, 2020, DOI: 10.1016/j.csl.2019.101025.

[20] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," Advances in Neural Information Processing Systems, vol. 31, pp. 10 040–10 050, 2018.

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

[22] G. Săracu and A. Stan, "An analysis of the data efficiency in Tacotron2 speech synthesis system," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2021, pp. 172–176, DOI: 10.1109/SpeD53181.2021.9587411.

[23] V. García, I. Hernáez, and E. Navas, "Evaluation of Tacotron based Synthesizers for Spanish and Basque," Applied Sciences, vol. 12, no. 3, p. 1686, 2022, DOI: 10.3390/app12031686.

[24] A. R. Mandeel, M. S. Al-Radhi, and T. G. Csapó, "Speaker Adaptation with Continuous Vocoder-Based DNN-TTS," in International Conference on Speech and Computer (SPECOM), Springer, vol. 12997, pp. 407–416, 2021, DOI: 10.1007/978-3-030-87802-3_37.

[25] S. -F. Huang, C. -J. Lin, D. -R. Liu, Y. -C. Chen and H. -y. Lee, "Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1558–1571, 2022, DOI: 10.1109/TASLP.2022.3167258.

[26] J. Shen etal., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in ICASSP, IEEE, 2018, pp. 4779–4783, DOI: 10.1109/ICASSP.2018.8461368.

[27] R. Prenger, R. Valle and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in ICASSP, IEEE, 2019, pp. 3617-3621, DOI: 10.1109/ICASSP.2019.8683143.

[28] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[29] Bakhturina, E., Lavrukhin, V., Ginsburg, B., Zhang, Y. (2021) Hi-Fi Multi-Speaker English TTS Dataset. Proc. Interspeech 2021, pp. 2776–2780, DOI: 10.21437/Interspeech.2021-1599.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in ICLR, USA, 2015.

[31] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion." in SLTU, 2008, pp. 63–68.

[32] X. Zhu, Y. Zhang, S. Yang, L. Xue and L. Xie, "Pre-Alignment Guided Attention for Improving Training Efficiency and Model Stability in End-to-End Speech Synthesis," in IEEE Access, vol. 7, pp. 65 955–65 964, 2019, DOI: 10.1109/ACCESS.2019.2914149.

[33] "ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality," 2001.

Ali Raheem Mandeel was born in Iraq. He received his M.Sc degree in computer engineering from the University of Missouri, Columbia, the US, in 2016. Currently, he is pursuing his Ph.D. in computer engineering in the Department of Telecommunications and Media Informatics at the Budapest University of Technology and Economics, Budapest, Hungary. His research interests include deep machine learning, signal processing, and speech synthesis.

Dr. Mohammed Salah Al-Radhi received a B.Sc. degree in Computer Engineering at Basra University in 2007, and a M.Sc. degree in Communication Systems Engineering at Portsmouth University, UK which was achieved with first-class honours in 2012 and awarded the MSc top student certificate in 2013. He received his Ph.D. from the Faculty of Electrical Engineering and Informatics at the Speech Technology and Smart Interactions Laboratory in Budapest University of Technology and Economics, Hungary in 2020, where he obtained it with honour (100%) and summa cum laude. Since October 2020, he has been a postdoctoral researcher in the Department of Telecommunications and Media Informatics, BME, Budapest. He served as a reviewer for several top-tier journals and conferences proceedings. His main interests are Artificial Intelligence and Machine Learning in speech signal processing and voice conversion.

Dr. Tamás Gábor Csapó obtained his Ph.D. in computer science & speech synthesis from Budapest University of Technology and Economics (BME), Hungary in 2014. He was a Fulbright scholar at Indiana University, USA in 2014, where he started to deal with ultrasound imaging of the tongue. In 2016, he joined the MTA-ELTE Lingual Articulation Research Group, focusing on investigating the Hungarian articulation during speech production. His research interests include Silent Speech Interfaces, ultrasound-based articulatory-to-acoustic mapping and articulatory-to-acoustic inversion, speech analysis and synthesis, and deep learning methods applied for speech technologies. Currently, he is a research fellow at BME.