Deep neural network speech synthesis based on adaptation to amateur speech data

Tijana Delić, Siniša Suzić, Milan Sečujski and Vladimir Ostojić

Abstract— **The paper investigates problems related to the automatic creation of personalized text-to-speech (TTS) synthesizers using small amounts of speech data recorded by amateur speakers in home conditions. The personalization of a synthesizer is based on the adaptation of a neural network based model pretrained on a large quantity of high-quality speech data recorded by a professional voice talent. In practice, both the quantity and the quality of target speaker's data used in the adaptation process are significantly inferior to the original training material. This research analyses the quality of synthesis created by adaptation on amateur data with the quality of synthesis created by adaptation on a high-quality speech dataset of the same size. The results of subjective and objective evaluation confirm the usability of the proposed adaptation method for efficient creation of new amateur TTS voices using a limited amount of adaptation data.**

Index Terms— **deep learning; neural networks; speaker adaptation; text-to-speech.**

I. Introduction

ACCORDING to [1] the term "communication" is defined as "systematic process in which people interact with and through symbols to create and interpret meaning". The most common channel through which this interaction is carried out is human speech. However, human speech does not only carry a particular message but a lot of additional information as well, including the speaker's emotional state and attitude towards the listener or the message itself. Furthermore, based on the characteristics of a speaker's voice we are usually able to establish their gender and age, and, even recognize them in case we are familiar with their voice [2-3]. For all these reasons someone's voice can be considered as an intrinsic part of their identity.

This conclusion provides a motivation for the development of methods for simple, fully-automated design of text-to-speech synthesis in the voice of a particular speaker. Such a synthesis would be usable in a number of scenarios, including voice banking, where users who are about to lose their ability to speak, usually due to some medical condition, can re-create their voices synthetically, based on speech recordings they have made previously. A common example of such a situation is total or partial laryngectomy, i.e. surgical removal of the entire larynx or some of its parts due to e.g. laryngeal cancer. Potential stigmatization and social exclusion of people who have lost their ability to speak in such ways has been analyzed in [4] where it has been suggested that their situation could be alleviated by using personalized text-to-speech synthesis (TTS), giving rise to research projects such as [5].

State-of-the-art speech synthesis systems, based on either a concatenative approach [6] or a parametric approach based on neural networks [7], achieve a very high level of speech quality, comparable to natural human voice. However, to achieve such a high level of quality of synthesis, it is necessary to obtain a large quantity of high quality speech data that will be used as basis for synthesis, either as a repository for a concatenative TTS or as training data for a parametric model. It is, thus, well known that speaker selection and database recording are the most critical steps in the process of getting a high-quality TTS, since mistakes made in either of these steps can never be eliminated in later stages of the process. For that reason, much effort is generally invested in finding a speaker with a suitable, resonant voice and good articulation, devoid of any impairments such as disfluency, difficulties in pronouncing specific speech sounds or excessive use of the vocal fry register.

However, in a scenario where the users provide samples of their voices, there are important differences with respect to the previous case: (1) it is usually difficult or impossible to obtain a large quantity of speech data; (2) the suitability of the voice of the target speaker for TTS cannot be guaranteed; and (3) microphone quality and recording conditions are generally inferior. In such circumstances it is impossible to build a TTS of reasonable quality from scratch. However, recent developments in TTS have offered an interesting alternative, namely, the adaptation of an existing speech synthesizer to the target speaker's voice [8]. The classical concatenative approach to TTS, although able to produce speech of high quality, does not possess the necessary flexibility for this, and more recently developed parametric speech synthesis methods have to be used instead. The adaptation of the parametric model to the voice of a new speaker has been a matter of extensive research, and most results are based on the use of hidden Markov models (HMM) [9] or, more recently, deep neural networks (DNN) [8]. Even in the case of adaptation it is still beneficial if the quantity of target speech data is large and its quality is high, but neither of these conditions is a prerequisite. Some researches has specifically focused on the case where the quantity of available speech data is small [10], although no studies have explicitly analyzed the influence of poor sound or speech quality on the speech synthesized by adaptation.

This paper investigates the problems related to automatic creation of personalized speech synthesizers using small amounts of speech data from amateur speakers, recorded in home conditions. We use a DNN-based synthesizer for

Tijana Delić is with the Faculty of technical sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: tijanadelic@ uns.ac.rs).

Siniša Suzić is with the Faculty of technical sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: sinisa.suzic@ uns.ac.rs).

Milan Sečujski is with the Faculty of technical sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (e-mail: secujski@ uns.ac.rs).

Serbian, originally trained on 3 hours of high-quality speech data, obtained from a professional speaker in a studio, and adapted using speech data from an amateur speaker (10 minutes), recorded in home conditions. This study represents the continuation of our previous work on efficient creation of new TTS voices by using neural network adaptation [10]. However, as opposed to e.g. the research described in [8], the quantity of adaptation data in this research can be considered relatively small. This corresponds to many practical scenarios in which the user interested in obtaining a personalized TTS with as little effort as possible.

This research also addresses another problem of TTS development. Namely, the preparation of speech data for training a parametric model does not require just phonetic annotation, but also some form of prosodic annotation. While phonetic annotation can be done automatically with high accuracy, prosodic annotation generally requires a lot of manual work, and as such, it represents the most time consuming step of the process. Human involvement is also incompatible with the idea of producing a personalized TTS from user voice samples in a fully automated process.

The paper is organized as follows. After the introductory Section I, a brief explanation of regular DNN based TTS modeling is given, and possible ways for its adaptation are presented in Section II. In Section III speech databases used for training are described in more detail. Section IV gives an overview of the experiments and provides an analysis and discussion of their results. Finally, in Section V conclusions are drawn and future research directions are outlined.

II. Obtaining a TTS model by DNN adaptation

DNN-based TTS has been in the focus of the research community in the last decade. Its popularity is due to the fact that it outperforms previously dominant statistical approaches in terms of their ability to generalize, as well as their flexibility in voice modification. The most frequently used and the most intuitive DNN architecture has been proposed in [11], and will be outlined in Section II.A. It has also been successfully adapted and used for synthesis in Serbian [12]. Deep neural network based synthesis in a particular voice can be obtained by different approaches [10, 13, 14], varying in their requirements in terms of target data quantity, and the one used in this research will be briefly explained in Section II.B.

A. A simple speaker-dependent DNN TTS system

A most common DNN-based TTS architecture, which is also used in this research, consists of the duration network and the acoustic network. The first network models phoneme durations and the second one models context-dependent acoustic features. The inputs of both networks are linguistic features such as phonetic context, the number of phones/syllables in the current word, etc. The inputs and outputs of the duration network are phone aligned, while the inputs and outputs of the acoustic one are frame aligned. For this reason, the acoustic input feature vector is extended by additional numeric features, including e.g. the index of the current frame in the state/phone as well as the index of the current state. The target features for the duration network are typically HMM state durations of the phoneme, extracted by forced alignment. On the other hand, acoustic network uses acoustic features extracted from speech recordings by an appropriate vocoder as target features.

B. DNN TTS model adaptation

With a large speech database, starting from randomly initialized weights and biases, it is possible to get the model able to produce speech similar to one from the database. However, it has been shown [10] that the model of speaker A is a better starting point for getting a model of speaker B than a randomly initialized model, in that a smaller quantity of data is sufficient to get a high-quality model B when starting from model A than when starting from a randomly initialized model. A quite obvious reason for this is that the models of any two speakers are closer to each other than a randomly initialized model to any speaker model. The complete procedure is the same as for getting a speaker-dependent model, but in this case the starting model is already trained on a large quantity of high quality data.

III. Database

In [10] it has been stated that the result of the proposed adaptation procedure depends on the quality of recordings used. To investigate this issue deeper, we have established the datasets to be used in this research as follows.

A. Principal training set

The speech database used to train the initial DNN TTS model is a database in Serbian recorded in a professional studio. It contains 3 hours of speech including sentence-medial silent phonetic segments. All the utterances were delivered by a single professional female voice talent. The database was phonetically and prosodically annotated, where the prosodic annotation specified the information related to lexical accent, degree of prosodic stress as well as types/positions of phrase breaks. While phonetic annotation was predominantly automatic (although the remaining errors were corrected manually), prosodic annotation included a significant amount of human effort. The database was recorded with the sample rate of 96 kHz, coded with 16 bits/sample and downsampled to 22 kHz.

B. Reference adaptation set

To study the influence of the quality of adaptation speech data on the quality of speech synthesized by a TTS obtained by adaptation, a segment of another high-quality speech database, delivered by another professional female voice talent, was used as a reference adaptation set. The segment in question was chosen randomly, without taking into account phonetic coverage. This database was also recorded in a professional studio with the same sampling frequency and bit depth, and the segment in question contains 10 minutes of speech (including sentence-medial silences). This set of utterances has also been phonetically and prosodically annotated in the same way as the principal training set.

C. Amateur adaptation set

Finally, to analyze the effect of poor quality of recorded speech on the quality of speech synthesis obtained by DNN adaptation, the same set of utterances as in III.B was re-recorded under conditions that could reasonably be expected from an average user who wishes to submit his or her voice samples to obtain a personalized TTS. Specifically, it was

recorded in a relatively reverberant home environment, using an average desktop computer with an integrated sound card, which also acted as a noise source, and a standard desktop microphone. The utterances were recorded with the sample rate of 44.1 kHz, coded with 16 bits/sample and downsampled to 22 kHz. The set has been delivered by a female amateur speaker, one of the authors of the paper. Some common post-processing was carried out on the amateur adaptation set, including noise cancellation, normalization and compression.

As the idea of the paper was to analyze the effect of the limited amount and low quality of adaptation speech data in a fully automated process of TTS adaptation, in the principal experiment no manual annotation of recorded speech was allowed, either phonetic or prosodic. However, the lack of human intervention in phonetic and prosodic annotation was circumvented by having the speaker deliver relatively short utterances from the principal training set, copying the original prosody as far as possible. This allowed the use of original phonetic and prosodic annotation of the principal training set, at the cost of a relatively small number of phonetic and prosodic errors which remained uncorrected.

IV. Experiments

The DNN based TTS system used in this research was built using the *Merlin* toolkit [11] with some modifications, as well as the CNTK framework [15].

The initial model, trained on the principal database, consists of two hybrid networks (just one layer with recurrence), both with 4 hidden layers and 1024 neurons per layer. All hidden layers use tangent hyperbolic activation, but while the first three contain simple feed-forward neurons, in the 4th hidden layer long short-term memory (LSTM) neurons are used. For the output layer, linear activation is used. The entire network is trained by the Nesterov optimizer, using back propagation with L2 regularization and mean squared error cost function.

The network uses 737 binary linguistic features as inputs. The HMM state durations were extracted by a procedure based on forced alignment, as proposed by *Merlin* toolkit, while the acoustic features were extracted by the WORLD vocoder [16].

Further training on adaptation datasets was carried out using the same values of the parameters, except for the learning rate, which was set to 0.003, i.e. to a value 3 times higher than in the training of the initial model.

The results of all experiments were evaluated objectively and subjectively. In the objective evaluation, the predicted features were compared with those extracted from the original recordings. The features used include mel-cepstral distance (MCD), band aperiodicity mean square error (BAP), root mean square error for F0 (F0 RMSE), correlation of F0 (F0 CORR), percentage of frames with incorrectly predicted voicing (VUV), as well as root mean square error and correlation for HMM state durations (DUR RMSE and DUR CORR). The objective measures were calculated on 15 randomly chosen sentences, not used during the training. Since objective evaluation of synthesized speech is known to be unreliable, it is usually complemented by subjective listening tests. The subjective MOS test carried out within this research included 16 amateur listeners, who were asked to grade the similarity of voices in pairs of utterances, using grades from 1 (different speakers) to 5 (identical speaker). In each pair, one of the sentences was the original recording while the other was synthesized either with acoustic features extracted from the original recordings (referred to as copy synthesis) or with acoustic features predicted by the neural network (referred to as synthesis). The 25 pairs of utterances, presented to the listeners in a random order, included the following:

- 5 pairs comparing an original recording from the reference set with corresponding copy synthesis;
- 5 pairs comparing an original recording from the reference set with synthesis by the model obtained by adaptation on the reference set;
- 5 pairs comparing an original recording from the amateur set with corresponding copy synthesis;
- 5 pairs comparing an original recording from the amateur set with synthesis by the model obtained by adaptation on the amateur set, using the original prosodic annotation;
- 5 pairs comparing an original recording from the amateur set with synthesis by the model obtained by adaptation on the amateur set, using manually corrected prosodic annotation.

The idea for the first set of experiments is to compare model adaptation when the target database is recorded in a professional studio by a professional speaker (reference adaptation dataset) with the case when the target database is recoded in a home environment by an amateur speaker (amateur adaptation set). For both datasets the same phonetic and prosodic annotation is used, the one made for and corresponding to the reference adaptation dataset.

The objective results (Table I) show that the adaptation

TABLE I
COMPARISON OF OBJECTIVE MEASURES OBTAINED IN THE SYNTHESIS OF TEST SENTENCES FROM MODELS TRAINED ON REFERENCE AND AMATEUR DATASETS

	Reference dataset	Amateur dataset
MCD (dB)	4.62	5.25
BAP (dB)	0.26	0.30
F0 RMSE (Hz)	18.72	29.21
F0 CORR	0.80	0.45
VUV (%)	4.04	7.39
DUR RMSE (frame/phoneme)	5.11	5.66
DUR CORR	0.85	0.81

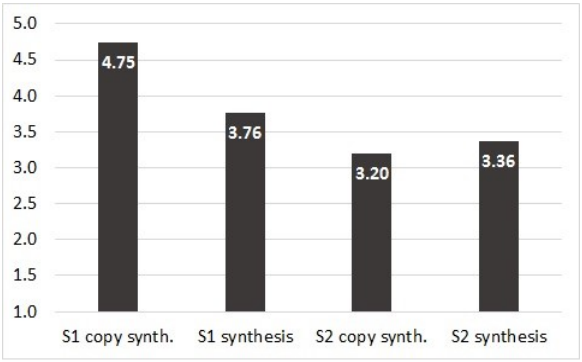

Fig 1. Subjective grades – comparison of copy synthesis and synthesis for the reference dataset (S1) and the amateur dataset (S2)

TABLE II
COMPARISON OF OBJECTIVE MEASURES OBTAINED IN THE SYNTHESIS OF TEST SENTENCES FROM MODELS TRAINED ON AMATEUR DATASET WITH ORIGINAL AND ADJUSTED ANNOTATION

	Original PA	Adjusted PA
MCD (dB)	5.25	5.21
BAP (dB)	0.30	0.30
F0 RMSE (Hz)	29.21	29.06
F0 CORR	0.45	0.46
VUV (%)	7.39	7.02
DUR RMSE (frame/phoneme)	5.66	5.41
DUR CORR	0.81	0.83

with the amateur dataset achieved significantly inferior objective measures. This is especially visible in case of f0 and can be attributed to the inconsistency of phonation and frequent vocal fry in the amateur adaptation set.

In the subjective test, listeners have graded the copy synthesis with the reference dataset with 4.75, while the copy synthesis with the amateur dataset obtained a much lower average grade of 3.20. Interestingly, in case of amateur adaptation set actual synthesis was graded better than copy synthesis. This somewhat surprising outcome may indicate that the quality of the amateur adaptation set was indeed so low that even the values of acoustic parameters robustly predicted by the neural network yielded a higher quality of synthesis than the acoustic parameters directly obtained from the adaptation set. In other words, the neural network, in terms of consistency of acoustic features, outperformed the speaker herself. On the other hand, the grades for actual synthesis for the two datasets obtained much closer grades. Regardless of the fact that the annotation has been originally made for the reference dataset, and regardless of the difference in the quality of both speakers and databases, the average overall grades differ less than 0.5 in favor of the reference dataset.

One additional experiment was carried out, in which the prosodic and phonetic annotation of the amateur dataset was done manually, i.e. the reference annotation was adjusted to the target speaker. Thereafter, training and the synthesis were repeated with thus corrected annotations, and compared to the results obtained with the original ones. Interestingly, the objective results (Table II) show that there is no significant difference between the results of these two setups. The subjective test confirms this finding, even giving a slightly better grade to the experiment with the original annotation – 3.28 average grade for corrected, and 3.36 for original annotation. This should be attributed to a general high variability of grades between listeners, higher than any actual statistically significant difference in the quality.

V. CONCLUSION

In this paper a method for the fast creation of new TTS voice based on speech data recorded by amateur speakers has been presented. The method is based on the adaptation of a neural network previously trained on a large quantity of high-quality speech data. Results obtained by both objective and subjective evaluation show that the quality of a TTS voice obtained by adapting on amateur data is somewhat inferior, but still comparable to the quality of a TTS voice obtained by adapting on the same amount of data recorded by a professional voice talent in a studio environment.

Our further research will include a more systematic evaluation of a greater number of TTS voices. Furthermore, due to the fact that the lack of prosodic annotation for each amateur dataset is bypassed by prosody copying, another research direction concerns the introduction of an objective measure of the similarity between the prosodic features of the recorded utterance and the original one. Introducing such a measure in the recording procedure for amateur speakers is expected to increase the compatibility of obtained speech data with the existing prosodic annotation and ultimately result in a better quality of synthesis.

ACKNOWLEDGMENT

The research was conducted within the project “Development of Dialogue Systems for Serbian and Other South Slavic Languages” (TR32035) and “Interdisciplinary Research into the Quality of Verbal Communication” (OI178027), financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia as well as the EUREKA project DANSPLAT (E!9944).

REFERENCES

[1] J. T. Wood, *Communication Mosaics: An Introduction to the Field of Communication,* 6th ed. Boston, USA: Cengage Learning Inc, 2009.
[2] P. Rose, *Forensic Speaker Identification,* 1st ed, London: CRC Press, UK, 2002
[3] M. Tiwari, M. Tiwari, “Voice – How Humans communicate”, *J Nat Sci Bio Med,* vol. 3, no. 1, pp. 3-11, Jan, 2012,
[4] J. Martl, E. Zackova, B. Repova, “Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis”, *Disabil. Rehabil. Assist.. Tehnol, vol. 14, no. 4, pp. 1-11, Apr, 2017*
[5] Z. Hanzlıccek, J. Matoussek, “Voice conservation: Towards creating a speech-aid system for total laryngectomees,” in Beyond AI: Interdisciplinary Aspects of Artificial Intelligence. pp. 55–59, Pilsen, 2011,
[6] A. J. Hunt, A.W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, In Proc. Of ICASSP-96, vol. 1, pp. 373-376, Atlanta, USA, 1996
[7] H. Zen, A. Senior, M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc. IEEE ICASSP, Vancouver, Canada, pp. 7962–7966, 2013
[8] Z Wu, P. Swietojanski, C. Veaux, S. Renals, S. King, “A study of speaker adaptation for DNN-based speech synthesis”, In Proc. INTERSPEECH, pp. 879-883, Dresden, Germany, 2015.
[9] J. Yamagishi, T. Kobayashi, Y. Nakano, K.Ogata, J. Isogai, “Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm”, *IEEE/ACM Trans. Audio, Speech, Language Process,* vol. 17, no. 1, pp. 66-83, Jan, 2009
[10] T. Delić, S. Suzić, M. Sečujski, D. Pekar, “Rapid Development of New TTS Voices by Neural Network Adaptation”, INFOTEH-JAHORINA, Jahorina, BiH, 2018
[11] Z. Wu, O. Watts, S. King, "Merlin: An Open Source Neural Network Speech Synthesis System", in Proc. 9th ISCA Speech Synthesis Workshop (SSW9), Sunnyvale, USA, September 2016
[12] T. Delić, M. Sečujski and S. Suzić, “A review of Serbian parametric speech synthesis based on deep neural networks”, Telfor Journal, Belgrade, ISSN: 1821-3251, Vol. 9, No. 1, pp. 32–37, 2017.
[13] Y. Fan, Y. Qian, F. K. Soong, and L. He, “Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis,” In ICASSP, Brisbane, Australia, April, 2015.
[14] N. Hojo, Y. Ijima, H. Mizuno, "An Investigation of DNN-Based Speech Synthesis Using Speaker Codes", Interspeech, San Francisco, USA, September 2016.
[15] F. Seide, A. Agarwal, “CNTK: Microsoft's Open-Source Deep-Learning Toolkit,” KDD '16 Proceedings of the 22nd ACM SIGKDD, San Francisco, California, USA, 2016.
[16] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE T. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877-1884, 2016.